Clustering and k-Means View in Big Data And Data Mining

G. NasrinFathima[1] R.Durga[2] R.Geetha[3]
[1,2]Research Scholar [3]Assistant Professor
[1,2,3]Department of computer science engineering
[1,2,3]Jamal Mohamed College, Trichirapalli, TN, India

Abstract— Partitioning a large set of objects into homogeneous clusters is a fundamental operation in data mining and big data. The k-means algorithm is best suited for implementing this operation because of its efficiency in clustering large data sets. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. The purpose of this paper is to analysis of k-means algorithm in big data and data mining.

Key words: Big data, Data Mining, Clustering, K-means algorithm.

I. INTRODUCTION

Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and bio-medical sciences[1]. Data mining can be defined as the process of selecting, exploring and modeling large amounts of data to uncover previously unknown patterns. Data Clustering is considered an interesting approach for finding similarities in data and putting similar data into groups. Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups .The idea of data grouping, or clustering, is simple in its nature and is close to the human way of thinking; whenever we are presented with a large amount of data, we usually tend to summarize this huge number of data into a small number of groups or categories in order to further facilitate its analysis.Cluster analysis is one of the major data analysis methods and the k-means clustering algorithm is widely used for many practical applications Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings.

II. RELATED WORK

Some researchers have been improved clustering algorithms. Some were presented new algorithms. And some other studied and compared clustering algorithms. In this section, we will review previous studies that presented influence of different factors on efficiency of a number of k-means clustering algorithm and results were compared.

Big Data applications are to explore the large volumes of data and extract useful information or knowledge for future actions (Rajaraman and Ullman, 2011). In many situations, the knowledge extraction process has to be very efficient and close to real-time because storing all observed data is nearly infeasible. For example, the Square Kilometer Array (SKA) (Dewdney *et al.* 2009) in Radio Astronomy consists of 1,000 to 1,500 15-meter dishes in a central 5km area. It provides 100 times more sensitive vision than any existing radio telescopes, answering fundamental questions about the Universe. However, with a 40 gigabytes(GB)/second data volume, the data generated from the SKA is exceptionally large.[1].

MadjidKhalilian[2] with using of divide and conquer method improves the K-Means algorithm for use in high-dimension data sets. Rui[3] presented the survey of clustering algorithms for data sets including in statistics, computer science, and machine learning, and explained their applications in some benchmark data sets, the traveling salesman problem, and bioinformatics and also subjects like adjacent measures and evaluating clustering were discussed.HE Ling[4] provided a detailed survey of current clustering algorithms in data mining at first, then it makes a comparison among them, presented their scores (merits), and identified the problems to be solved and the new directions in the future according to the application requirements in multimedia domain. Treshansky[5] presented a survey of clustering algorithms and paid particular attention to those algorithms that require less amount of knowledge about the domain being clustered.

Data mining plays an important role in IT as it discovers knowledge from historical data of various domains. For instance data mining can be used to mine medical data as Healthcare domain produces huge amount of data aboutpatients, diseases, diagnosis, and medicine and so on. By applying data mining techniques in Healthcare domain, the administrators can improve the QoS (Quality of Service) by discovering latent potentially useful trends required by medical diagnosis [6].Abe et al. [7] proposed an integrated time-series data mining environment for mining huge amount of medical data for extracting more valuable rule-sets.

III. PROPOSED WORK

With the rising of data sharing websites, such as Facebook and Flickr, there is a dramatic growth in the number of data.For example, Facebook reports about 6 billion new photo every month and 72 hours of video are uploaded to YouTube every minute. One of major data mining tasks is to unsupervised categorize the large-scale data [Biswas and Jacobs, 2012;Lee and Grauman, 2009; Dueck and Frey, 2007; Cai et al.,2011], which is useful for many information retrieval and classification applications.

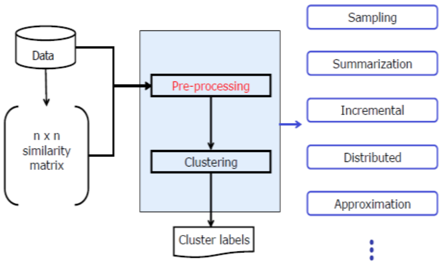

Fig.1: Clustering process on Big data

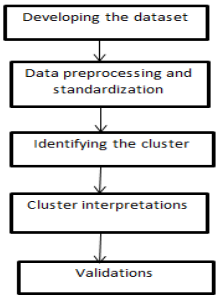

Fig. 2: Clustering process on Data Mining

Sinceclusters arenotpredefined, adomain expertisoften required tointerpret themeaning ofthecreated clusters. Here we are going to find k-means algorithm how used in big data and data mining.

A. K-means algorithm:

The k-means algorithm (MacQueen 1967, Anderberg1973) is built upon four basic operations: (1) selection ofthe initial k means for k clusters, (2) calculation of thedissimilarity between an object and the mean of a cluster,(3) allocation of an object to the cluster whose mean isnearest to the object, (4) Re-calculation of the mean of acluster from the objects allocated to it so that the intracluster dissimilarity is minimised. Except for the firstoperation, the other three operations are repeatedlyperformed in the algorithm until the algorithm converges.The essence of the algorithm is to minimise the costfunction

$$E = \sum_{l=1}^{k}\sum_{i=1}^{n} y_{i,l}\, d(X_i, Q_l) \qquad (1)$$

where n is the number of objects in a data set **X**, Xi Î **X**, Qlis the mean of cluster l, and yi,l is an element of a partitionmatrix **Ynx l** as in (Hand 1981). dis a dissimilaritymeasure usually defined by the squared Euclideandistance.There exist a few variants of the k-means algorithmwhich differ in selection of the initial k means,dissimilarity calculations and strategies to calculate clustermeans (Anderberg 1973, Bobrowski and Bezdek 1991).The sophisticated variants of the k-means algorithminclude the well-known ISODATA algorithm (Ball andHall 1967) and the fuzzy k-means algorithms (Ruspini1969, 1973).Most k-means type algorithms have been provedconvergent (MacQueen 1967, Bezdek 1980, Selim andIsmail 1984). The k-means algorithm has the followingimportant properties.

(1) It is efficient in processing large data sets. The computational complexity of the algorithm isO(tkmn), where m is the number of attributes, n isthe number of objects, k is the number of clusters,and t is the number of iterations over the whole dataset. Usually, k, m, t <<n. In clustering large datasets the k-means algorithm is much faster than thehierarchical clustering algorithms whose generalcomputational complexity is O(n2) (Murtagh 1992).

(2) It often terminates at a local optimum (MacQueen1967, Selim and Ismail 1984). To find out theglobal optimum, techniques such as deterministicannealing (Kirkpatrick et al. 1983, Rose et al. 1990)and genetic algorithms (Goldberg 1989, Murthyand Chowdhury 1996) can be incorporated with thek-means algorithm.

(3) It works only on numeric values because itminimises a cost function by calculating the meansof clusters.

(4) The clusters have convex shapes (Anderberg 1973).Therefore, it is difficult to use the k-meansalgorithm to discover clusters with non-convexshapes.

B. K-means algorithm in Big data:

The k-means (Lloyd) algorithm, an intuitive way to explore the structure of a large data set. The idea is to view the observations in an N variable data set as a region in N dimensional space and to see if the points form themselves into clusters according to some method of measuring distance. To apply the k-means algorithm one takes a guess at the number of clusters (i.e. select a value for k) and picks k points (maybe randomly) to be the initial center of the clusters. The algorithm then proceeds by iterating through two steps:

(1) Assign each point to the cluster to which it is closest

(2) Use the points in a cluster at the mth step to compute the new center of the cluster for the (m +1)th step

Eventually, the algorithm will settle on k final clusters and terminate.

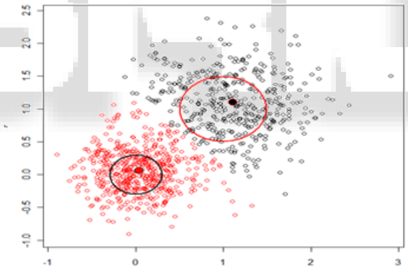

Fig. 3: k-means on artificialBig data

The k-means clustering on an artificial 2-dimensional data set shows the data come from two different normal distributions, one centered at (0,0) and the other at (1,1). The large circles are show the points within one standard deviation of the true means. The smaller colored circles are the calculated centers of the clusters. For this artificial example, the two clusters do a pretty good job of describing the data. However, for real data the situation will generally not be so clear cut. There is no guarantee that clusters found will be globally optimal, and of course, since the choice of k was arbitrary, there is no reason to believe that the clusters really mean anything.

For example, The first thing to notice about the data is that apparently nobody in the three states involved makes between $175,000 and $300,000. As probably guessed, this is an artifact of how the data were coded. Values of income higher than $175,000 were recorded as state means: hence, the three green lines. The choice of 4 clusters appears to be reasonable and k-means provides some insight. The horizontal clusters indicate that people group together more by income than by age, and the black dots which mark the centers of the clusters confirm that people generally get wealthier as they age. Time taken 0.4 seconds to run in R.

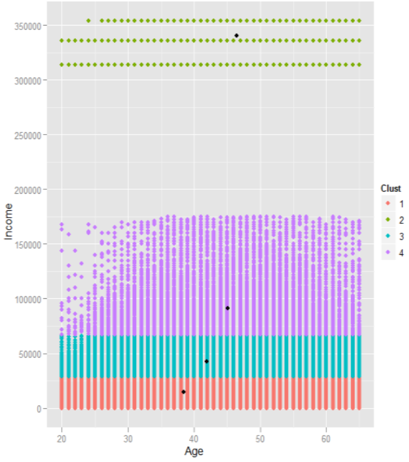

Fig. 4: K-means analysis on Large Census Data [8]

C. K-means algorithm in Data mining:

K-means is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached.The algorithm then proceeds by iterating through following steps:

(1) The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.

(2) 2.For each data point: Calculate the distance from the data point to each cluster.

(3) If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.

(4) Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.

For example, Government can group its crop variety based on common features. Government does not have any predefined for this label. Based on the outcome of the grouping they will target production and average yield campaigns to the different groups for a particular type of scheme.

The information they have about the farmers include survey number, crop name and variety.The data has been preserved from records in Agriculture Department, Perambalur. The collected data has been entered and analyzed using weka(machine learning). Time taken to run is 0.06 seconds.

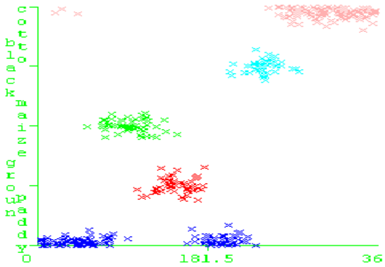

Fig. 4: Cluster visualizations using K-means algorithm

IV. CONCLUSION

The study shows that term Big Data literally concerns about data volumes and Clustering is an exploratory technique; used in every scientific field that collects data.Clustering is essential for "Big Data" problem and challenges are scalability, very large number of clusters, heterogeneous data, streaming data, validity.

K-mean algorithm has biggest advantage of clustering large data sets and its performance increases as number of clusters increases. But its use is limited to numeric values.

K-Means algorithm is faster than other clustering algorithm and also produces quality clusters when using in large data.

The most attractive property of the k-means algorithm in data mining is its efficiency in clustering large data sets. However, that it only works on numeric data limits its use in many data mining applications because of the involvement of categorical data.

REFERENCES

[1] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data mining with Big Data"

[2] MadjidKhalilian, Norwati Mustapha, MD NasirSuliman, MD Ali Mamat, "A Novel K-MeansBased Clustering Algorithm for HighDimensional", International multi conference of Engineers and Computer Scientists, 2010.

[3] RuiXu, Wunsch, D., II, Dept. of Electr.&Comput. Eng., Univ. of Missouri-Rolla, Rolla,MO, USA, "Survey of clustering algorithms", IEE Transaction on Neural Networks, 2005.

[4] HE Ling WU Ling-da, CAI Yi-chao(College ofInformation System & Management ,NationalUniversity of Defense Technology, ChangshaHunan 410073,China), "Survey of ClusteringAlgorithms in Data Mining", 2007.

[5] Treshansky, Allyn, McGraw, Robert M,"Overview of clustering algorithms", Enablingtechnology for Simulation Science", 2001.

[6] M. Ilayaraja Department of Computer Science & Engineering Alagappa University Karaikudi, India(2013). Mining Medical Data to Identify Frequent Diseases using Apriori Algorithm.IEEE.

[7] HidenaoAbe AND Hideto Yokoi (n.d). Developing an Integrated Time-Series Data Mining Environment for Medical Data Mining. *IEEE*.p1-6.

[8] Referred from the website: http://blog.revolutionanalytics.com/

[9] Manish Verma, MaulySrivastava, NehaChack, Atul Kumar Diswar, Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining," *International Journal of Engineering Reserch and Applications (IJERA)*, Vol. 2, Issue 3, pp.1379-1384, 2012.

[10] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011

[11] .Mr. A. B. Devale and Dr. R. V. Kulkarni "A REVIEW OF DATA MINING TECHNIQUES IN INSURANCE SECTOR" GOLDEN RESEARCH THOUGHTS VOL -1 , ISS - 7 [JAN 2012]

[12] Mr.A. B. Devale and Dr. R. V. Kulkarni "APPLICATIONS OF DATA MINING TECHNIQUES IN LIFE INSURANCE" IJDKP) Vol.2, ISSUE-4, July 2012

[13] 8. The estimates of the productivity and production of various crops were made based on the crop cutting experiment carried out with the joint efforts of the Departments of Economics and Statistics, Agriculture and Horticulture and Plantation crops, Retrieved from

[14] http://agritech.tnau.ac.in/pdf/2012/Season%20& %20Crop%20Report%202012.pdf

IJSRD