International Journal of Advanced Engineering and Innovative Technology (IJAEIT)
ISSN No 2348-7208 Special Issue on "Emerging Technology for Innovative India"

Big Data And K-Means Clustering

[1]Vishakha Mehendale,[2]Shital Dhamal
[12] *Department Of Computer Engineering, Mumbai University,*
[1]*SES'SFOE College of Engineering,*
Diksal, Raigad, Maharashtra, India
[2]*Lokmanya Tilak College of Engineering,*
Koparkhairne, Navi Maharashtra, India.
[1]vishu.m20@gmail.com
[2]dhamalsk@gmail.com

Abstract- Nowadays, with evolving technologies Big data has its significant impact on our economy. Current economy is directly or indirectly related to large data which is in high volume, velocity, variety and veracity. This data is being generated at tremendous rate from the sources like Internet, mobile devices, social media, and geo-spatial devices, sensors. Analyzing Big data with in specified time is the challenge being faced by researchers and data mining communities.Clustering algorithms have developed a powerful meta-learning tool which can precisely analyze the volume of data produced by modern applications. Clustering methods, we can use to easily visualize analysis of Big data also it helps in decision making. Clustering technique is used in many application as marketing, insurance, surveillance, fraud detection. This paper contains an overview of big data along with clustering method for analysis of Big data.

Keywords- Big data,Clustering methods, unsupervised learning.

1. INTRODUCTION

Technology is being evolved day by day and has played vital role on every aspect of our economy. We Can also say that economy is driven by different emerging technologies. Nowadays we live in digital world where everything is digitized. With increased digitization, large amount of structured and unstructured data being created and stored. The data is being produced by various sources like internet, online transactions, mobile devices, social media, digital images, videos and audios. Many governments has observed the phenomenon of Big data and initiated exploiting big data in many areas such as science and engineering, healthcare, national security.

Analyzing data provides an enterprise with significant competitive advantages. But current volume of big data sets are complicated to managed and processed by traditional relational database management systems as well as warehousing.

Mainly these large datasets are stored in data centers. But to store such large data, retrieve it and analyze is the key issue. This paper also emphasizes the problem of analyzing Big data within predefined parameters. The clustering of Big data in compact format can help to analyze and manage it. Clustering is an unsupervised technique used to classify large data sets into correlative groups.

2. WHAT IS BIG DATA

Despite the realization that "Big Data" holds the key to many new researches, there is no consistent definition of Big Data. Till now, it has been described only in terms of its promises and features (volume, velocity, variety, value, veracity). Given below are few definitions by leading experts and consulting companies:

- The IDC definition of Big Data (rather strict and conservative): "A new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis" [11].
- A simple definition by Jason Bloomberg [12]: "Big Data: a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques." This is also in accordance with the definition given by Jim Gray in his seminal book [13].
- The Gartner definition of Big Data that is termed as 3 parts definition: "Big data is high- volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." [9].

3. CHARACTERISTICS OF BIG DATA

Big Data is characterized by the following 4 Vs:

• Volume - the vast amount of data generated every secondthat are larger than what the conventional relationaldatabase infrastructures can cope with.
• Velocity - the frequency at which new data is generated,captured, and shared.
• Variety - the increasingly different types of data (fromfinancial data to social media feeds, from photos to sensordata, from video capture to voice recordings) that no longerfits into neat, easy to consume structures.
• Veracity - the disarrayed data (Facebook posts with hashtags, abbreviations, typos, and colloquial speech)

4. APPLICATIONS OF BIG DATA

4.1 Big data in healthcare

*Big Da*ta plays a vital role to improve the quality and efficiency of healthcare delivery. Big Data applications are expected to have higher impact when data from various healthcare areas, such as clinical, administrative, financial, or outcome data, can be integrated. Healthcare organizations are leveraging big data technology to capture all of the information about a patient to get a more complete view for insight into care coordination and outcomes-based reimbursement models, population health management, and patient engagement. Although profit is not and should not be a primary motivator, it is vitally important for healthcare organizations to acquire the available tools, infrastructure, and techniques to leverage big data effectively or else risk losing potentially millions of dollars in revenue and profits.

4.2 Big Data in Finance

Big Data is seen as a years-long journey forfinancial organizations. Financial markets have undergone a remarkable transformation over the past two decades due to advances in technology, including faster and cheaper computers, greater connectivity among market participants, and tremendous amounts of data. Financial services, in particular, have widely adopted big data to inform better investment decisions with consistent returns. The sector is beginning to build out road maps of where Big Data could deliver the most value within this broader set of technology investments.

4.3 Big Data in Retail

The exponential growth of retail channels and the increasing use of social media are empowering consumers. With the information assets readily available online, consumers are now better able to compare products, services and prices. When consumers interact with companies publically through social media, they have greater power to influence other customers or damage a brand. In order for retailers to capitalize on these and other changes in the industry, they need ways to collect, manage and analyze a tremendous volume, variety, velocity and veracity of data. If retailers succeed in addressing the challenges of big data, they can use this data to generate valuable insights for personalizing marketing and improving the effectiveness of marketing campaigns, and removing inefficiencies in distribution and operations.

4.4 Government

Big Data can give governmental organizationsaccess to data on a much larger extent than ever before.Governments are dealing with increasing volumes of data that have high variety of structures and helps them to increase the tax collection. Governments additionally have high potential for improving their utilization of so-called dark data; data that is available somewhere in the system but not actively used. The continuous digitalization of governmental services and communication with citizens willfurther accelerate the growth of data and big datatechnologies will play in future an important role forefficient and customer centric services
.

4.5 E. Biometrics

It has become increasingly obvious thatapplications of Big Data are expanding immensely.Very large-scale biometric systems are becomingmainstream in nationwide identity cards and mobile secure payment methods. As with other Big Data systems, biometric systems contend with the “four V” challenges that involve the effective managing of the complex life cycle and operations of identity information—despite the immense enrollment database size (volume) and rapid transactionresponse-time (velocity) requirements using potentially noisy, fraudulent (veracity), and multiple (variety) biometric identifiers. Biometric systems also provide a rich case study involving how these issues manifest and are addressed in a unique, domain-specific way. We believe that by virtue of dealing with some of the most critical entities,namely identity and entitlement, biometric systems are likely to emerge as among the most critical of the Big Datasystems.

4.6 Agriculture

To feed the world's rapidly-expanding populationin the coming decades, agriculture must produce more. Big data holds one of the keys for farmers, but it's also a weapon that could be used against them. One of the most important technologies in agriculture nowadays is to create agriculture big data. It may be created naturally, if field sensing technology is distributed widely and measured data are shared on cloud storage services. However commercial storage services are not sustainable. Robust storage service to record permanently such important data is needed.

4.7 Smart cities

Cities are focusing on sustainable economic development and high quality of life, with wise management of natural resources. These applications will allow people to have better services, better customer experiences, and also be healthier. Big data is certainly enriching our experiences of how cities function, and it is offering many new interaction and more informed decision-making with respectto our knowledge of how best to interact in cities.

5. ARCHITECTURE OF BIG DATA STORAGE TECHNIQUES INCLUDES:

- Multiple clustered network attached storage(NAS) alsocalled as scale-out NAS Clustered NAS employsstorage devices attached to a network. Groups of storage devices attached to different networks are then clusteredtogether.
- Object-based storage system distribute set of objectsover a distributed storage system.
  - Hadoop is used to process unstructured and semistructured data.It uses the map reduce paradigm tolocate all relevant data then select only the datadirectly answering the query. Hadoop works well inscale-out NAS environment. Hadoop is adistributed file system and data processing engine that is designed to handle extremely high volumesof data in any structure.

Hadoop has two components:

- The Hadoop distributed file system (HDFS), whichsupports data in structured relational form, inunstructured form, and in any form in between .
- The MapReduceprogramming paradigm formanaging applications on multiple distributedservers.

- NoSql, MoongDB, and Terra Store processstructured big data.

- NoSql data is characterized by being BASE - Basically Available, Soft state(changeable), andEventually consistent rather than the traditional dbdata characteristics of Atomicity, Consistency,Isolation and Durability(ACID).

- NoSQL focuses on a schema-less architecture (i.e., the data structure is not predefined). In contrast, traditional relation DBs require the schema to be defined before the database is built and populated.

  - Data are structured
  - Limited in scope
  - Designed around ACID principles

- MoongDB and Terra Store are both NoSql relatedproducts used for document-oriented applicationssuch as storage and searching of whole invoicesrather than individual data fields from the invoice.

The focus is on supporting redundancy, distributedarchitectures, and parallel processing

Apache Avro: designed for communication betweenHadoop nodes through data serialization

Cassandra and Hbase: a non-relational database designed for use with Hadoop

Hive: a query language similar to SQL (HiveQL) but compatible with Hadoop

Mahout: an AI tool designed for machine learning;that is, to assist with filtering data for analysis andexploration

Pig Latin: A data-flow language and execution framework for parallel computation

*ZooKeeper:*Keeps all the parts coordinated and working together

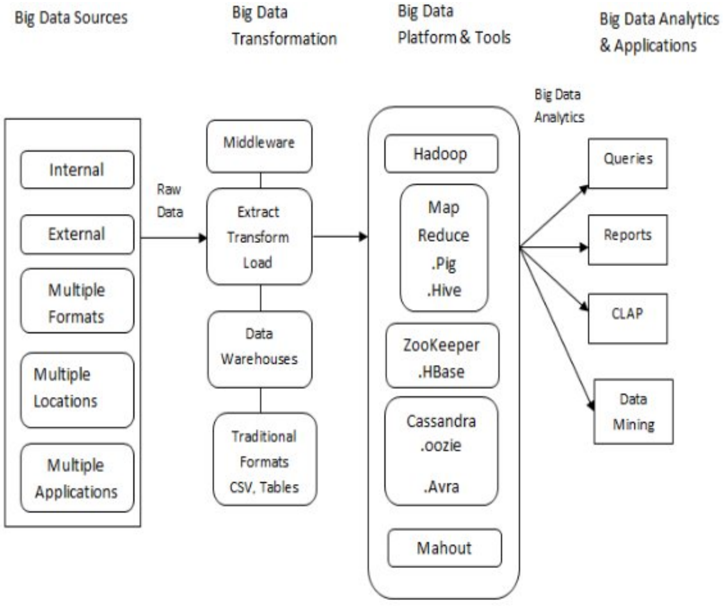

Fig 1. Big Data Architecture and its storage technologies

6. CLUSTERING METHODS

t can be concluded that Big data is different from the data being stored in traditional warehouses. The data stored there first needs to be cleansed, documented and even trusted. Moreover it should fit the basic structure of that warehouse to be stored but this is not the case with Big data it not only handles the data being stored in traditional warehouses but also the data not suitable to be stored in those warehouses. Thus there is requirement of entirely new frameworks and methods to deal in Big Data. Currently, the main challenges identified for the IT Professionals in handling Big data are:

i.)The designing of such systems which would be able to handle such large amount of data efficiently and effectively.
ii.)The second challenge is to filter the most important data from all the data collected by the organization.

Clustering in Big data is required to identify the existing patterns which are not clear in first glance. The properties of big data pose some challenge against adopting traditional clustering methods[2]:

A. **Type of dataset**: The collected data in the real world contains both numeric and categorical attributes. Clustering algorithms work effectively either on purely numeric data or on categorical data; most of them perform poorly on mixed categorical and numerical data types.

B. **Size of dataset**: The size of the dataset has effect on both the time-efficiency of clustering and the clustering quality (indicated by the precision). Some clustering methods are more efficient than others when the data size is small, and vice versa.

C. **Handling outliers/ noisy data**: Data from real applications suffers from noisy data which pertains to faults and misreported readings from sensors. Noise (very high or low values) makes it difficult to cluster an object thereby affecting the results of clustering. A successful algorithm must be able to handle outliers/noisy data.

D. **Time Complexity**: Most of the clustering methods must be repeated several times to improve the clustering quality. Therefore if the process takes too long, then it can become impractical for applications that handle big data.

E. **Stability**: Stability corresponds to the ability of an algorithm to generate the same partition of the data irrespective of the order in which the data are presented to the algorithm. That is, the results of clustering should not depend on the order of data.

F. **High dimensionality**: “Curse of dimensionality”, a term coined by Richard E. Bellman is relevant here. As the number of dimensions increases, the data become increasingly sparse, so the distance measurement between pairs of points becomes meaningless and the average density of points anywhere in the data is likely to be low. Therefore, algorithms which partition data based on the concept of proximity may not be fruitful in such situations.

G. **Cluster shape[8]**: A good clustering algorithm should be able to handle real data and their wide variety of data types, which will produce clusters of arbitrary shape. Many algorithms are able to identify only convex shaped clusters.

7. CLUSTERING ALGORITHMS FOR LARGE DATASETS

Clustering[2] is an unsupervised technique used to classify large datasets in to correlative groups. No predefined class label exists for the data points or instances. Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups and the groups are called as clusters[2].

The properties of clustering algorithms to be considered for their comparison from point of view of utility in Big Data analysis include[3]:

- Type of attributes algorithm can handle
- Scalability to large datasets
- Ability to work with high dimensional data
- Ability to find clusters of irregular shape
- Handling outliers
- Time complexity
- Data order dependency
- Labeling or assignment (hard or strict vs. soft or fuzzy)
- Reliance on a priori knowledge and user defined parameters
- Interpretability of results

7.1 Partitioning Methods:

a). K-MEANS

Partitioning [10] algorithms spilt the data into several subsets. The reason of splitting the data into several subsets is that it is computationally not feasible to check every possible subset; there are certain greedy probing schemes used in the form of iterative inflation. Categorically, this means different relocation schemes that iteratively reassign points between the k clusters. Relocation algorithms gradually improve clusters. In other words, the partitioning algorithms divide data objects into a number of partitions, where each partition represents a cluster.

The simplest and most fundamental version of cluster analysis is partitioning, which organizes the objects of a set into several exclusive groups or clusters. Formally, given a data set, D of n objects, and k, the number of clusters to form, a partition algorithm organizes the objects into k partitions ($k \leq n$), where each partition represents a cluster. K-Means[5][2] is a one of partitioning algorithm which defines efficient way to cluster analysis. The K-means algorithm is simplest unsupervised learning algorithm that solves the well known clustering problem. The procedure defines a simple and easy way to classify a given data set through a certain number of clusters (consider/assume k clusters) fixed a priori. The main idea is to evaluate k centroids, one for each cluster and such centroids must be placed in a cunning way because of different location causes different outcomes. So, the better way is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Let consider a data set, D contains n objects in Euclidean space. Partitioning methods distribute the objects in D into k clusters, C1, C2,……………..Ck, that is Ci< D and Ci ∩ Cj = Ø for (1 ≤ i, j ≤ k). An objective function is used to assess the partitioning quality so that objects within a cluster are similar to one another but dissimilar to objects in other clusters. This is the objective function aims foe high intracluster similarity and low intercluster similarity. A centroid based partitioning technique uses the centroid of a cluster, Ci , to represent that cluster. Conceptually, the centroid of a cluster is its center point. The centroid can be defined in various ways such as by the mean or medoid of the objects or points assigned to the cluster. The difference between an object p & ci, the representative of the cluster, is measured by dist(p,ci), where dist(x,y) is the Euclidean distance between two points x and y. The quality of cluster ci can be measured within cluster variation, which is the sum of squared error between all objects in ci and the centroid ci, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2$$

Where E is the sum of the squared error for all objects in the data set; p is the point in space representing a given object; and ci is the centroid of cluster Ci here both p and ci are multidimensional. In other words, for each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This objective function tries to make the resulting k clusters as compact and as separate as possible.

Advantages-

1. Easy to understand and implement.
2. Produce more dense clusters than the hierarchical method especially when clusters are spherical.
3. For large number of variables, K-means algorithm may be faster than hierarchical clustering, when k is small.
4. Efficient in processing large datasets.

Drawbacks-

1 Poor at handling noisy data and outliers.

2 Works only on numeric data.

3 Empty cluster generation problem.

4 Random initial cluster center problem.

5 Not suitable for non-spherical clusters.

6 User has to provide the value of k.

b) K-medoids

This algorithm is very similar to the K-means algorithm. It varies from the k-means algorithm mainly in its representation of the different groups or clusters. Each cluster is represented by the most centric object in the cluster, rather than by the implicit mean that may not belong to the cluster. For every point in the cluster: add up all distances to the other points in the cluster. Point in the cluster for which this distance is the smallest, becomes the medoid or centroid. Both k-means and k-medoids require the user to specify K, the number of clusters. Two well-known types of k-medoids clustering [7] are the PAM.(Partitioning AroundMedoids) and CLARA (Clustering Large Applications) Algorithm- - Starts from an initial set of medoids and clusters are generated by points which are close to respective medoids. - The algorithm iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering. The K-medoids method [6] is more robust than the K-means algorithm. In the presence of noise and outliers, a medoid is less influenced by outliers or other extreme values than a mean. However, its processing is more costly than the K-means method.

7.2 COMPARISON

Results of CLUSTERING analysis based on portioning method is given in the table I.

Method Name	Algorithm	Time Complexity	Dataset Size	Dataset Type	Cluster Shape	Handle Outlier

TABLE I Shows Result
(I= No of Iterations, K= No of Clusters, n= No of Objects)

Partitioning	K- Means	O(lkn)	Huge	Numeric	Spherical	No
	K-Medoids	O(n^2I)	Small	Categorical	Spherical	Yes

8 CONCLUSION

Big data is the “new” business and social sciencefrontier. The amount of information and knowledge that can be extracted from the digital universe iscontinuing to expand as users come up with new waysto massage and process data. Moreover, it has becomeclear that “more data is not just more data”, but that is data with large “Volume, Velocity and Variety ”.The conclusion tends to define clustering is a major factor needed while big data management has been carried out since k means algorithm basically a way to form efficient clusters. Through clustering large amount of data can be easily handled and with the help of that heterogeneous huge data can be manage and finally gives a better results. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters).

In general, the empirical study allows us to draw the following conclusions for big data No clustering algorithm performs well for all the evaluationcriteria, and future work should be dedicated toaccordingly address the drawbacks of each clustering.

REFERENCES

[1] An Architecture for Big Data Analytics by Joseph O. Chan Communications of the IIMA ©2013 1 2013 Volume 13 Issue 2

[2] Clustering methods for Big data analysis by KeshavSanse, Meena Sharma International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 3, March 2015

[3] Adapting k-means for Clustering in Big Data by Mugdha Jain and ChakradharVerma International Journal of Computer Applications (0975 – 8887) Volume 101– No.1, September 2014

[4] Survey of Various Clustering Techniques for Big Data in Data Mining byKosha Kothari, Ompriya Kale © 2014 IJIRT | Volume 1 Issue 7 | ISSN: 2349-6002

[5] K-Means Clustering Method for Big Data Mining 1Pravin Anil Tak, 2Dr. S. V. Gumaste, 3Prof. S. A. Kahate International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-6)

[6] LiorRokach, OdedMaimon Chapter 15 CLUSTERING METHODS “DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK”

[7] MihaelAnkerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander, "OPTICS: Ordering Points To Identify the Clustering Structure" Proc. ACM SIGMOD"99 Int. Conf. on Management of Data, Philadelphia PA, 1999

[8] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil A. Zomaya, S. Foufou, and A. Bouras ,“A Survey of Clustering Algorithms for Big Data:Taxanomy& Empirical Analysis”. IEEE Transactions on Emerging Topics in Computing, JUNE 2014.

[9] Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s, By Svetlana Sicular, Gartner, Inc. 27 March 2013. [online] http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-datadefinition-consists-of-three- parts-not-to-be-confused-with-three-vs/.

[10] Cheng-Ru Lin, Chen, Ming-SyanSyan , “Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging” IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 2,pp.145-159, 2005.

[11]Extracting Value from Chaos, By Gantz, J. and Reinsel, D. IDC IVIEW June 2011. [online] http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

[12]The Big Data Long Tail.Blog post by Bloomberg, Jason.On January 17, 2013. [online] http://www.devx.com/blog/the-big-data-long-tail.html.

[13] The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Hey, T. ,Tansley, S. and Tolle, K.. Microsoft Corporation, October 2009. ISBN 978-0-9825442-0-4.

[14]S.Sangeetha et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3269-3274