An Appropriate Big Data Clusters with K-Means Method

Mr. Pravin Anil Tak[1], Dr. S. V. Gumaste[2], Prof. S. A. Kahate[3]

[1] M.E. II[ND] Year Student, [2]Professor, [3]Assistant Professor
Computer Engineering Department, Sharadchandra Pawar College of Engineering, Dumberwadi, Otur, Tal- Junnar, Dist- Pune, Maharashtra, India

ABSTRACT

The problems related to big data are increasing now days since the arrival of data becomes faster and existing techniques are not capable to handle such big data. Big data shows various attributes due to that complexity and problems towards mining big data get enhanced. A clustering method is used to map big data amongst various clusters according to its properties and further applying statistical method and k-means algorithm to get fast real extracted results.

Keyword: - *Clustering, K-Means algorithm, Iterative Method, Big data.*

1. INTRODUCTION

Big Data defines to datasets whose sizes (large in volume) are beyond the ability of typical database software tools to capture, store, manage and analyze. There is no predefined definition of how big a dataset should be in order to be considered Big Data. New technological tools have to be in place to manage and process this Big Data phenomenon. Now a day's Big Data technologies as a new generation of technologies and architectures designed to extract value or pattern economically from very large volumes of a wide variety of data by enabling high velocity capture, discovery and analysis. Big data is data that exceeds the processing capacity of traditional database systems. The data is too big, moves too fast, or does not fit the structures of existing database architectures. To gain value from these data, there must be an alternative way to process it.

1.1 Characteristics of Big Data-

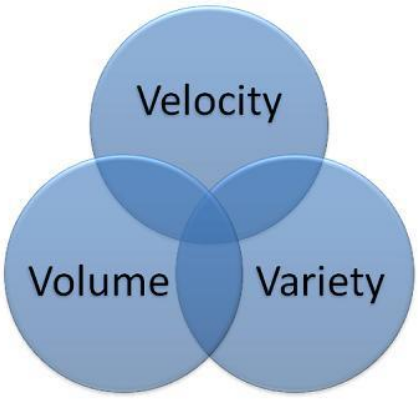


Fig 1- The Three V's of Big Data

Basically there are three main attributes of big data are as Volume, Varity and Velocity. Since volume defines a large and huge data that cannot handle within a specified time parameter. Varity defines the difference in structure and data types of data, due to this complex structure have evolved. Velocity shows the speeds of coming data get increased according to time since data size changes as time get change.

2. LITERATURE SURVEY

The studies of past historical concerns are necessary to acquire the detail knowledge about topic. Since here is the discussion about few of the papers has been carried out to define different views of author related to big data and its mining to extract useful pattern from it.

Yang Song defines the concept of storage mining that meets the data analytics in this cloud based data center infrastructures imposes remarkable challenges to IT management operations, e.g., due to virtualization techniques and more stringent requirements for cost and efficiency. They was defined the Information Lifecycle Management (ILM) with Storage Tiering that reduces the cost solution that focuses on moving less accessed storage volumes (e.g., virtual disks in storage virtualization) to low-end storage devices in order to free up space on high-end devices with superb performance, e.g., good results within specified time. They had defined there are two main challenges for big data mining as big data storage and scalable machine learning. The defined scalable machine learning framework uses Hadoop & Cassandra cluster to implement predictive analytics.

Shuliang Wang et all Spatial Data Mining in the Context of Big Data in that they had explained Spatial data plays a primary role in big data, attracting human interests. Spatial data mining confronts much difficulty when attracting the value hidden in spatial big data. The techniques to discover knowledge from spatial big data may help data to become intelligent. Spatial data is closely related to human daily life, permeated all walks of life. The number, size and complexity are all in sharp increasing. A large amount of data has been stored in the spatial database and warehouse in types of text, graphics, images and multimedia. Big data is voluminous and it grows quickly, but it has very low density in value, which means there is a lot of junk data. Big data processing is to implement the transitions: from data to information, from information to knowledge and from knowledge to wisdom. Big data processing includes object superposition, target buffer, spatial data cleaning, spatial data analysis, spatial data mining, spatial feature extraction, image segmentation, and image classification.

Hui Chen et all explained the concept of Parallel Mining Frequent Patterns over Big Transactional Data in Extended MapReduce. They had explained the extended mapreduce framework in that a parallel algorithm is presented for mining the frequent patterns in a big transactional data. According to MapReduce frame, the big data is firstly split into a lot of data subfiles, and the subfiles are concurrently processed by a cluster consisting of hundreds or thousands computers. In order to improve the performance of proposed method, the insignificant patterns in the data subfile are efficiently located and pruned by probability analysis. The simulation result proves that the method is efficient and scalable, and can be used to efficiently mine frequent patterns in big data.

Xindong Wu et all was defines a three tier big data processing framework. The research challenges form a three tier structure and center around the "Big Data mining platform"(Tier I), which focuses on low-level data accessing and computing. Challenges on information sharing and privacy, and Big Data application domains and knowledge form Tier II, which concentrates on high-level semantics, application domain knowledge, and user privacy issues. The outmost circle shows Tier III challenges on actual mining algorithms.

3. HACE THEOREM

The HACE theorem defines an efficient way to handle the big data by considering it attributes. HACE Theorem Statement defined as Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data.

These characteristics make it a major challenge for discovering useful knowledge from the Big Data. The fundamental characteristic of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This is because different information collectors prefer their own schemata or protocols for data recording, and the nature of different applications also results in diverse data representations. An autonomous data source with distributed controls is one of characteristic of Big Data applications. Being autonomous, each data source is able to generate and collect information without involving (or relying on) any centralized control. While the volume of the Big Data increases, so do the complexity and the relationships underneath the data. In an early stage of data information systems, the concentrating on finding best feature values to represent each observation. As the data comes from various autonomous sources followed by their own schemata defines a new level of complexity that cannot be easily track out and doesn't show any information. Since evolving relationships are get enhanced and affect on entire process of mining big data.

By considering the challenges evolved due to these characteristics of big data can be easily achieved since define a solution in such a way that a heterogeneity, decentralized control and complex and newly evolving relationships are get handled easily. To mine big data properly means accurate and within approximated time, the handling of data by considering these attribute gives better results.

The HACE equation for big data is

Big Data = Heterogeneous + Autonomous sources with distributed and decentralized control + Complex + Evolving Relationships.

4. K-MEANS ALGORITHM

The clustering is a good way to classify and define data according to its similarity and dissimilarity. There are various methods of clustering basically here the method of partitioning has used. The K-means is a one of partitioning algorithm in which partitioning of objects has done, which is depends on the factor of similarity that means clusters are formed in such a way that objects in the same cluster having high similarity but dissimilar to objects in other cluster.

K-Means is one of the simplest algorithm used for clustering. K-means partitions "n" observations/objects into k clusters in which each observation/object belongs to the one cluster having similar attributes but dissimilar to object in other cluster.

a. Algorithm K-Means-

Input:

k: the number of clusters,

D: a data set containing n objects.

Output:

A group or set of k clusters.

Method:

1) Arbitrarily choose k objects from D as the initial centers;

2) repeat

3) assign/reassign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

4) update the cluster means, that is calculate the mean value of the objects for each cluster;

5) until no change;

In this category of clustering the k-means is the simplest and easy method to cluster the data. The similarity measurement for this method is carried out by distance measurement. It is most common to calculate the dissimilarity between two patterns using a distance measure define on the feature space. The most popular metric for distance measurement is Euclidean distance.

b. Flow Chart-

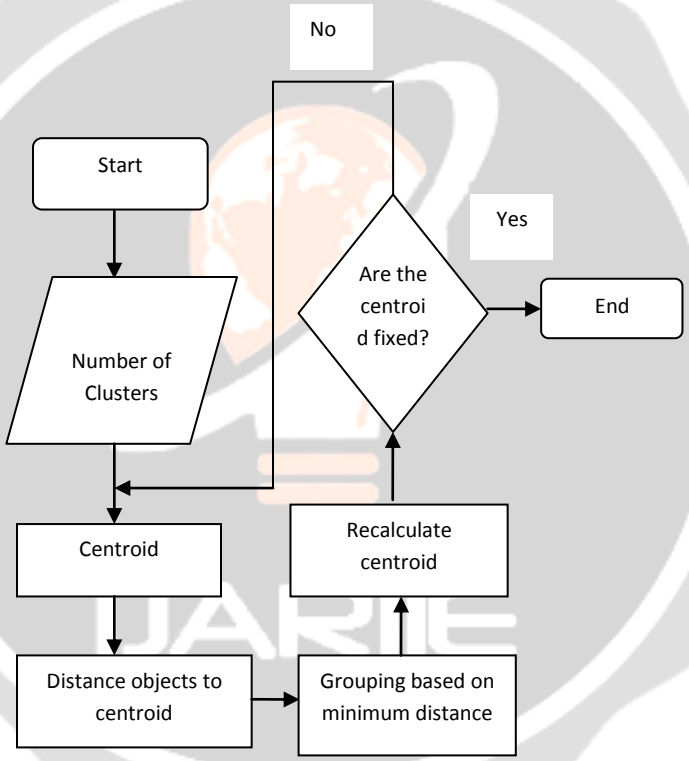


Fig 2 -Flow Chart

5. PROPOSED SYSTEM

Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work.

The proposed system defines main five sections or module as K-Means clustering module, Data mining module, Heterogeneity classification module, Homogeneity classification module and calculate fitness module respectively. An initially input set has given to k-means clustering module to form clusters of objects such that

similar objects should be enclosed in a one cluster. In second stage data mining trends applied to mine data efficiently from the existing clustered data set.

In this system there are two classification trends has set such as heterogeneity and homogeneity. Firstly heterogeneity classification has been carried out since the difference in structure and attributes of objects defined. Then in next step homogeneity factor and classification done to define similarity within the objects. Finally the calculation of fitness factor has calculated since define or put exact or most similar data set as output according to user request that comes from various clusters. In proposed system to build a stream-based Big Data analytic framework for fast response and real-time decision making. The key challenges and research issues includes a designing Big Data sampling mechanisms to reduce Big Data volumes to a manageable size for processing and building prediction models from Big Data streams.

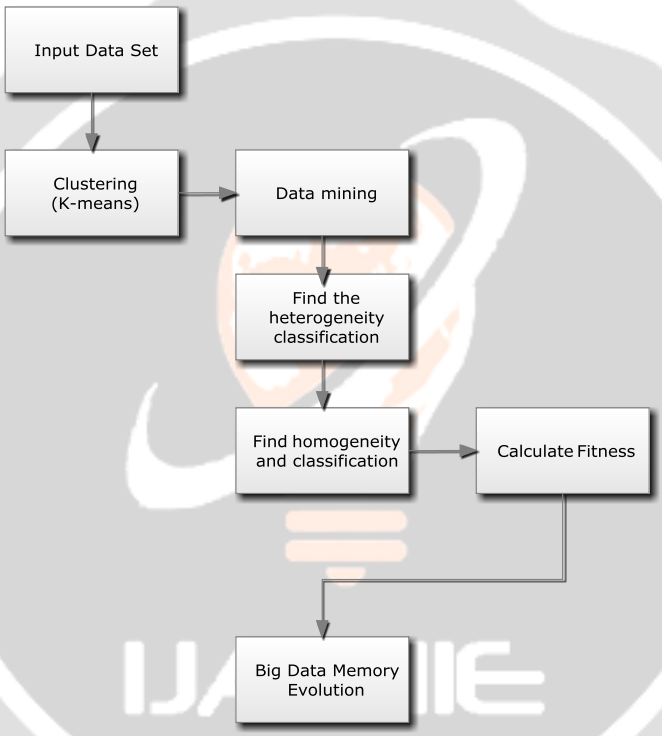


Fig 3 - Proposed Framework

6. RECURSIVE findK() FUNCTION

Searching is the process of looking for a particular value in a collection. For example, a program that maintains the membership list for a club might need to look up the information about a particular member. This involves some form of search process. A explanation of something that refers to itself is called a recursive definition. At first glance it looks like the function expects an initial standard deviation; but in fact, it is used such that on the first run, that parameter is initialized to -1. Therefore, it will calculate the initial standard deviation based on the entire sample. Within it there is a variable, tolerance, which is used to indicate how close the smaller samples standard deviation should be to the last standard deviation to continue splitting the sample. This value is chosen arbitrarily, and it is likely the experiment results could have been better had attempted to let the program "learn" a good value to use.

```
private static int findK(Sample s, double minSD)
{
    // Idea 2: split sample while standard deviation decreases
    double tolerance = 0.001, thisSD;

    s.sort();

    if (minSD < 0) minSD = s.getStandardDeviation();

    Sample workingSample1 = getHalf(1, s);
    Sample workingSample2 = getHalf(2, s);

    thisSD = workingSample1.getStandardDeviation();
    if (thisSD+tolerance < minSD)
        return 1 + findK(workingSample1, thisSD);

    thisSD = workingSample2.getStandardDeviation();
    if (thisSD+tolerance < minSD)
        return 1 + findK(workingSample2, thisSD);

    return 0;
}
```

7. DATA SETS AND RESULTS

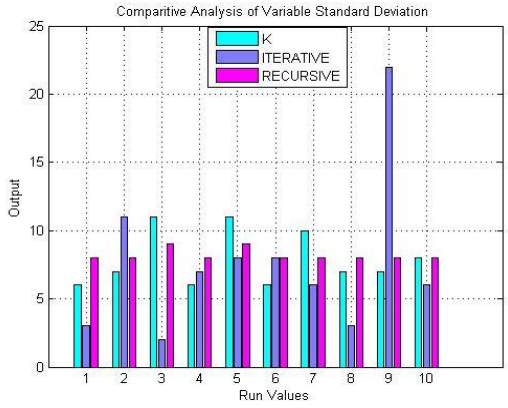
Presented here in tabular form are the results of 20 experimental runs. More were done, of course, but due to space requirements, only 20 are presented. In both sets, the number of samples was uniformly random between 5 and 12, the size of each sample varied from 50 to 100, and the mean and standard deviation of each sample spanned 0-100 and 0-5, respectively. In the first set of ten runs, the standard deviations of the initial samples were allowed to vary. The second set used a fixed standard deviation of two.

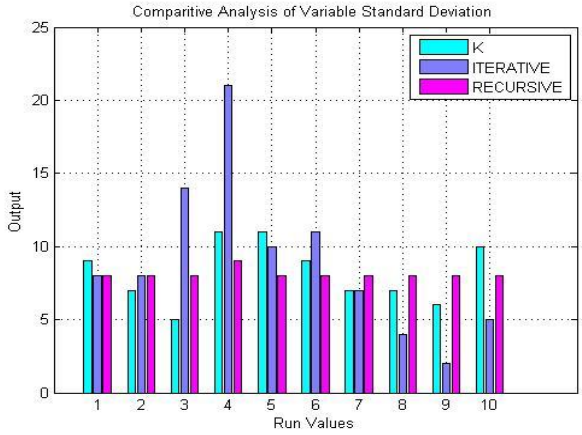
Set 1: Variable Standard Deviation										
Run:	1	2	3	4	5	6	7	8	9	10
k:	6	7	11	6	11	6	10	7	7	8
Iterative:	3	11	2	7	8	8	6	3	22	6
Recursive:	8	8	9	8	9	8	8	8	8	8

As the data show, neither algorithm was particularly successful at correctly finding k. The recursive algorithm is a lot less erratic, and seems to be closer to finding k than its iterative counterpart. It should be noted, however, that in several of the cases where the recursive algorithm found too few clusters, particularly when it was

one or two off, looking at the individual generated clusters reveals that actually, it was very good - just that two or three of the means were incredibly close, and the algorithm could not differentiate between all the sets. For example, in the first run of the fixed standard deviation, one of the means was 87.68 with another at 89.78. In this case, it is unlikely even a human could have differentiated between the two. Similarly, in the sixth run of the same set, two of the generated means were less than a standard deviation apart. However, a glaring issue with the recursive trials is that it appears to always hover around eight or nine indicating that perhaps the tolerance was so low, it encouraged splitting the sample all the time. However, even varying the side of the equation in which the tolerance was applied (or subtracting it instead of adding it) did not change the variety of results obtained.

Set 2: Fixed Standard Deviation = 2										
Run:	1	2	3	4	5	6	7	8	9	10
k:	9	7	5	11	11	9	7	7	6	10
Iterative:	8	8	14	21	10	11	7	4	2	5
Recursive:	8	8	8	9	8	8	8	8	8	8





8. CONCLUSIONS

A big data can be handled effectively by using clustering approach and concentrating on attributes of big data such as Heterogeneity and evolves complex relationship. Here the use of iterative k means defines the exact clustering method which helps to distribute data. A classification of big data carried out in such a way that objects in the one cluster are having similarity but objects in other clusters are dissimilar to same. Whenever user want to retrieve a data by considering several attributes can possible with the help of defined approach.

ACKNOWLEDGEMENT

I take this opportunity to express my hearty thanks to all those who helped me in the completion of this paper. I express my deep sense of gratitude to my dissertation guide and Head of Department Dr. S. V. Gumaste, Computer Engineering, Sharadchandra Pawar College of Engineering, Dumberwadi, Otur for his guidance and continuous motivation.

REFERENCES

[1]. Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE *Data Mining with Big Data* IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO. 1, JANUARY 2014.

[2]. Yang Song, Gabriel Alatorre, Nagapramod Mandagere, and Aameek Singh *Storage Mining: Where IT Management Meets Big Data Analytics* IEEE International conference on Big Data 2013.

[3]. Abdul-Aziz Rashid Al-Azmi *Data, Text, and Web Mining for Business Intelligence: A Survey* International Journal of Data Mining and Knowledge Management Process 2013.

[4]. Li Liu,Murat Kantarcioglu and Bhavani Thuraisingham *Privacy Preserving Decision Tree Mining from Perturbed Data* International Conference on System Sciences 2009.

[5]. Shanqing Li, Lirong Song, Hui Zhao *A Discriminant Framework for Detecting Similar Scientific Research Projects Based on Big Data Mining* 2014 IEEE International Congress on Big Data.

[6]. Leila Ismail, sMohammad M. Masud ,Latifur Khan *FSBD: A Framework for Scheduling of Big Data Mining in Cloud Computing* 2014 IEEE International Congress on Big Data.

[7].Katsuya Suto, Hiroki Nishiyama, Nei Kato, Kimihiro Mizutani, Osamu Akashi, And Atsushi Takahara An *Overlay-Based Data Mining Architecture Tolerant to Physical Network Disruptions* IEEE VOLUME 2, No. 3, September 2014.

[8].Jemal H. Abawajy, Andrei Kelarev, and Morshed Chowdhury *Large Iterative Multitier Ensemble Classifiers for Security of Big Data* IEEE Volume 2, NO. 3, September 2014.

[9] Scott W. Cunningham, Wil A. H. Thissen *Three Business and Societal Cases for Big Data: Which of the Three Is True?* IEEE Vol. 42, No. 3, Third Quarter, September 2014.

[10] Xiaochun Cao, Hua Zhang, Xiaojie Guo, Si Liu, and Dan Meng, *SLED: Semantic Label Embedding Dictionary Representation for Multilabel Image Annotation* IEEE Transactions On Image Processing, Vol. 24, No. 9, September 2015.

[11] Shifeng Fang, Li Da Xu, Yunqiang Zhu, Jiaerheng Ahati, Huan Pei, Jianwu Yan, and Zhihui *Liu An Integrated System for Regional Environmental Monitoring and Management Based on Internet of Things* IEEE Transactions On Industrial Informatics, Vol. 10, No. 2, May 2014.

[12].J. Gerard Wolff, *Big Data and the SP Theory of Intelligence* IEEE Volume 2, 2014.

[13]. Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, And Yong Ren *Information Security in Big Data: Privacy and Data Mining* IEEE Volume 2, 2014.

[14] Rajiv Ranjan *Streaming Big Data Processing in Data center Clouds* IEEE Cloud Computing 2014.

[15] Abdur Rahim Mohammad Forkan, Ibrahim Khalil, Ayman Ibaida, and Zahir Tari *BDCaM: Big Data for Context-aware Monitoring - A Personalized Knowledge Discovery Framework for Assisted Healthcare* IEEE Transactions On Cloud Computing , Vol. X, No. X, February 2015.

[16] SHI Wenhua, ZHANG Xiaohang, GONG Xue, LV Tingjie *Identifying Fake and Potential Corporate Members in Telecommunications Operators* China Communications August 2013

[17]R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.

[18] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches toCrawling Important Pages Early," Knowledge and InformationSystems, vol. 33, no. 3, pp 707-734, Dec. 2012.

[19] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.

[20] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.