Volume 6, Issue 7, July 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Big Data Mining: Analysis of Genetic K- Means Algorithm for Big Data Clustering

Pooja Bisht[1], Kulvinder Singh[2]
M.tech, Department of Computer Science, Uttarakhand Technical University, India[1]
Assistant Professor, Department of Computer Science, Doon Institute of Engineering and Technology, India[2]

Abstract— The amount of data in world is growing day by day because the use of internet, smart phones and social networks. Big data is a collection of data sets which is very large in size as well as complex. Traditional database systems are not able to capture, store and analyze this large amount of data. In this paper, we propose an algorithm for the clustering problem of big data using a combination of the genetic algorithm with the K-Means algorithm. The main idea behind this algorithm is to combines the advantage of Genetic algorithm and K-means to process large amount of data. Experimental results shows that our new genetic k means algorithm take less memory and time to process big data than the simple k means algorithm. This algorithm combines the advantage of Genetic algorithm and K-means.

Keywords— Big Data, Data Mining, Clustering, Genetic Algorithm, K Means

I. INTRODUCTION

In today's Big Data era the data is becoming more and more available due to advances in information and communication technologies, enterprises are gaining meaningful information, relevant knowledge and vision from this huge data based on decision making. Big data mining can be defined as the ability of extracting valuable information from huge and complex set of data or data streams i.e. Big Data. One of the essential data mining techniques for analysing big data is clustering. There are complications for applying clustering techniques to big data duo to large amount of data rising every day. One of the most commonly used clustering technique for data mining is K-means algorithm which is used to extract information from a dataset. But k-means cannot process the data if Data is too large and when we have less storage capacity. As Big Data is refer to terabytes and petabytes of data, the clustering algorithms come with high computational costs, therefore to cope with this difficulty and to deploy clustering approaches to big data to get the outcomes in a reasonable time we propose a genetic k means algorithm. This algorithm is combining the feature of k means and genetic algorithm to process the large data i.e. big data.

II. BIG DATA

Big Data is used to describe massive amounts of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies. Due to the advent of new technologies, devices, and communication like social networking sites, the quantity of data produced by people is growing rapidly each year.

Big Data concern with huge volume, complex, fastest growing data sets with autonomous and multiple sources. There are six key characteristics that define big data. These characteristics are also known as Six V's of big data.

- Volume: It refers to the vast amount of data generated every second.
- Variety: Variety Refers to the type of data that is being stored.
- Velocity: It refers to the speed at which new data is generated and the speed at which data moves around.
- Veracity: It refers to the level of reliability associated with certain types of data.
- Value: Identifying valuable data and then transforming and extracting that data for analysis.
- Variability: It refers to the messiness, inconsistency or trustworthiness of the data.

III. DATA MINING

Data Mining is a technology to extract the knowledge from the data sets. It is used to explore and analyse such data. The data to be mined varies from a small data set to a large data set i.e. big data. The data Mining environment produces a huge volume of the data. The information retrieved in the data Mining step is transformed into the structure that is easily understood by users. Data Mining involves various methods such as genetic algorithm, support vector machines, decision tree, neural network and cluster analysis, to disclose the hidden patterns inside the huge amounts of data set.

Various algorithms are necessary for handling such large amount of data set. These algorithms define various structures and approaches implemented to handle Big Data, and also various tools that were developed for analysing them.

Big Data are the large amount of data being processed by the Data Mining environment. In other words, it is the

collection of data sets which is huge and complex that it becomes difficult to process using traditional data processing applications; hence data mining tools were used. Big Data are mostly unstructured, invaluable, incomplete and complex data into usable information.

IV. CLUSTERING

Clustering means to identify similar types of objects. We can also then identify density and sparse areas in object space and can determine overall distribution patterns and relationships among data attributes. It is a technique of exploring the data, a technique of discovering patterns in the dataset. It is a form of unsupervised learning that means we don't know in advance how data should be group the data objects (similar types) together.

Clustering is one of the most significant research fields in the area of data mining. It deals with discover a structure in a group of unlabeled data. Clustering means creating collections of objects based on their types in a method that the objects belonging to the similar groups are same and those belonging to dissimilar groups are different. The main benefit of clustering is that stimulating patterns and structures can be establish directly from very huge set of data with tiny or not any of the background knowledge. These algorithms can be applied in many domains. Partitioning is one popular approach of clustering. Partitioning approaches transfer objects by moving them from one cluster to another cluster starting from a certain point. The amount of clusters for this technique should be pre-set for this technique. The algorithms uses in this approach are k means, k-medoid algorithm, k-nearest neighbour algorithm etc.

V. GENETIC ALGORITHMS

Genetic algorithms (GAs) are general-purpose search and optimization techniques based on principles inspired from the genetic and evolution mechanisms observed in natural systems, natural genetics and populations of human beings. Their basic principle is the maintenance of a population of solutions to a problem (genotypes) as encoded information individuals that evolve in time. By simulating natural evolution, a genetic algorithm can effectively search the problem domain and easily solve complex problems. It perform search in complex, large and provide near optimal solutions for objective or fitness function of an optimization problem. The parameters in the search space are represented in the form of strings (chromosome), encoded by a combination of cluster centroids. A set of such chromosomes is called a population. Firstly, a random population is created, which represents different solutions in the search space. Based on the principle of survival of the fittest, a few of chromosomes are selected and each is assigned into the next generation. Chromosomes are binary or continuous encoded strings, representing potential solutions to the optimization problem. Each member becomes evaluated on the fitness function (objective function), giving a measure of the solution quality called the fitness value. Upon candidate solution selection, recombination (crossover and mutation) is being performed, ending in a new candidate solution population. The basic steps of genetic algorithm for data clustering include individual representation and population initialization, fitness computation, selection, crossover and mutation.

Following are the steps of Genetic Algorithm for clustering of data [6]:

Input:

k: the no of clusters
d: the data set containing n objects p: population size
Tmax: Maximum no. of iterations

Output:

A set of K clusters

1) Initialize every chromosome to have k random centroids selected from the set of data.
2) For T=1 to Tmax
   (i) For every chromosome i
      a. Allocate the object data to the cluster with the closest centroid.
      b. Recomputed k cluster centroids of chromosome i as the mean of their data objects.
      c. Compute the chromosome i fitness.
   (ii) Generate the new group of chromosomes using GA selection, crossover and mutation.

VI. K-MEANS ALGORITHM

K-means is very simple and popular unsupervised learning algorithm that solves problems of clustering. It is a partitioning approach for clustering. K-Means clustering approach creates a group of same type of data according to their closeness to each other based on the Euclidean distance. It takes ky as input constraint and partition a set of n number of objects from ky clusters. The object's mean value is taken as similar parameter to build clusters. The cluster mean or center of cluster (centroid) is created by randomly selection of ky object; other objects are assigned to that cluster by comparing the most similarity between them.

The algorithm of simple k mean algorithm is as follows [2]:

Input:

ky - the no. of clusters
dy- data set which contains n number of objects

Output:

1) Input the value of ky and set of data.
2) If ky= = 1 then Exit
3) Else
4) Select k no. of objects from d randomly as the initial centers of cluster.
5) For each data point in the cluster j reprint and state each object into the cluster of similar types objects, based on the mean value of object in the cluster.
6) Update cluster means values and after that for every cluster compute the mean value of objects.
7) Repeat from step iv until no data point was allocated, stop otherwise.

The sufficient criteria can be either no. of iteration or centroid's change of position in consecutive iterations.

Cons of k-means algorithm:

- It is difficult task to find k-value.
- It is data dependent.
- It is not effective when used with large amount of data (Big Data) or universal cluster.
- If dissimilar initial partitions has been selected than it may differ the clusters result.
- If the clusters are of dissimilar size and or dense then it cannot handled by the algorithm.

VII. GENETIC K-MEANS ALGORITHM

This algorithm combines the advantage of Genetic algorithm and K-means. Genetic Algorithm based clustering algorithm is expected to provide an optimal clustering, more superior to that of K-Means approach, but with a little more time complexity.

Following are the steps of the algorithm of GK-means:

1) Set the population.
2) compute fitness of every individual by following equation.
   Fitness (i) =2. (pi - 1)/Q-1 i=individual, p=position, Q=total individuals
3) If satisfied with the fitness condition, then assign solution, Else
4) Calculate sub population and migrate
5) Counting the i_{th} individual depends on the rate si, which is relative to its level of fitness that is
   Si = fitness (i) / summation (fitness (i));
6) Translate population and assets individual wellness.
7) Perform crossover and mutation on each sub population
8) If termination condition satisfies, stop; else go to step 5.

VIII. RESULT ANALYSIS

The implementation results of the Genetic k-means algorithm shown in following figures.

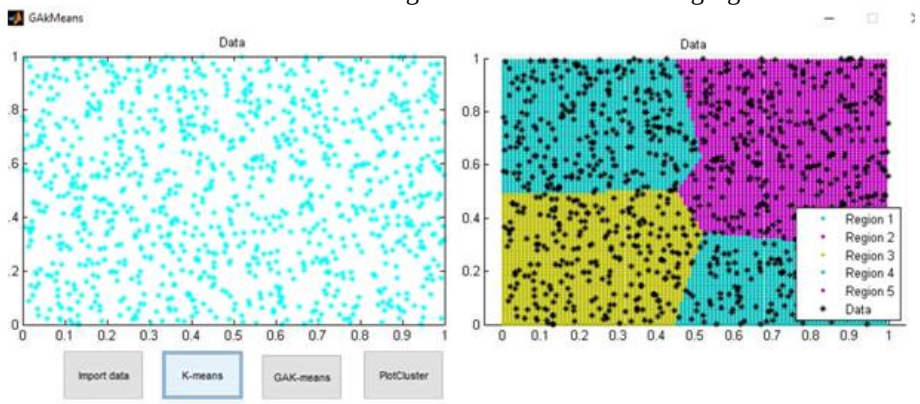

Fig.1 Actual data in clusters and partitions of data using k means

The actual data has shown in Fig.1 in form of clusters. Using simple k means we have partitioned the cluster of data. We have created 5 partitions for our actual data which is around 2000 number of data. But it shows only four partitions in Fig 1.

This is because distances among these data are very less as shown in fig 1. So this is the drawback of k means that it can't process the large amount of data. If we have minimum amount of data then k mean is easy to process but for large amount of data it has a drawback. So in figure 1 it shows only 4 partitioned instead five because of the lack of memory showing that data overlaps into each other.

Fig. 2 shows data processing of the same data using GAK means algorithm that shows which area has the best value. This data processing generating things such random number, parent value, point value, chromosomes etc. based on the hierarchy of data shows in fig. 2 which shows the updating according to hierarchy of data by up down updating bars shown in fig 2.

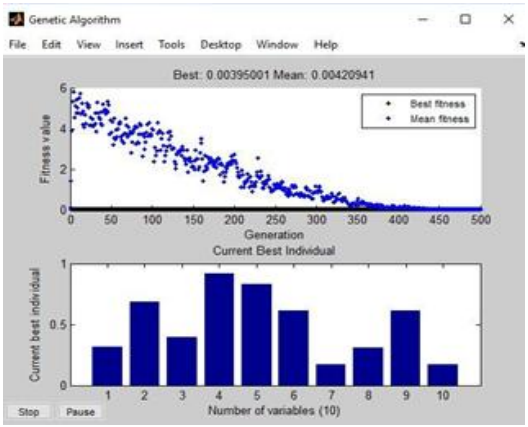

Fig.2 Hierarchy of data using genetic k-means algorithm

The final updating of data based on hierarchy of data and final chart of best individuals after is shown in figure 2 and the final result shows in fig 3.

The final result of GAK- Mean (Genetic k- means algorithm) shows in fig 3 which show the 5 partitions of our actual data. The simple k means will take more processing memory and much time to process large amount of data i.e. big data but our GK-means will take less memory and time to process big data. This is the benefit of GK- Means algorithm over simple k-means algorithm.

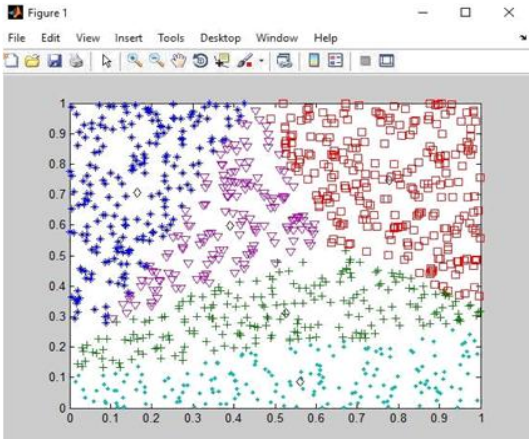

Fig. 3 Partition of data using GAK means algorithm

In this implementation, we have used 1000 by 2 matrixes of data and programmed for 5 partitions. In case of K-Means, the discrimination among the data is not as clear as GK-means algorithm. Except that, in this method, we can also calculate the best value for a certain region. If we use our other dataset of 54500 by 2 matrix, the k-means algorithm in declines processing as it requires 36GB memory to process (Fig 4).

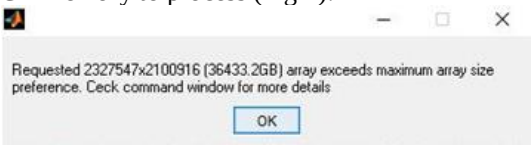

Fig. 4 K-means decline to process large data

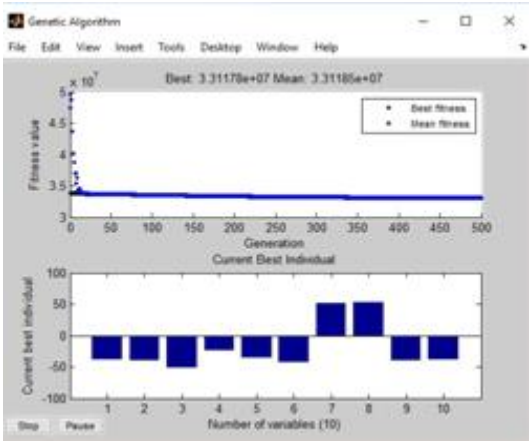

Fig. 5 processing chart of GK- Means

While our approach still works fine with that data (Fig. 5) and it shows the partition of large amount of data shows in Fig 6 which k-mean algorithm cannot process. Therefore it is proved that the Genetic k-means algorithm converges to the global optimum. This Genetic K-mean algorithm is faster than some of the other approaches of clustering.

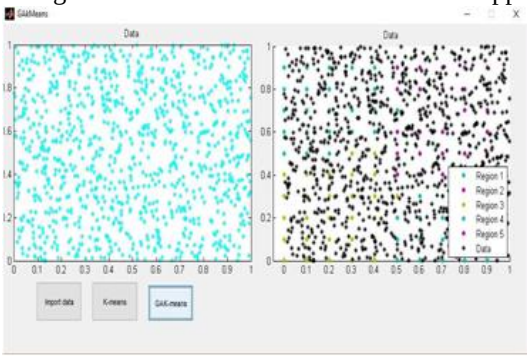

Fig. 6 Partition of large data using GK-Means Algorithm

IX. CONCLUSIONS

This paper introduces a combination of a genetic algorithm and the k-means for the big data clustering. K-Means algorithm is most common and simple clustering algorithm for partition of large data. But it has some drawback to process huge amount of data and when we have less memory space to process. Genetic algorithm is search and optimization technique to process big data. But it can't the partition of the data so the new algorithm is used by combining Genetic and k-means algorithm known as GKA or Genetic K-means algorithm. The future scope of this algorithm is we can use Hadoop's MapReduce with k-means algorithm or modified versions of k-means to make it more suitable to process big data or mining the data.

REFERENCES

[1] Ahmed and Saeed. A Survey of Big Data Cloud Computing Security. *International Journal of Computer Science and Software Engineering (IJCSSE)*, Volume 3, Issue 1, December 2014.
[2] Arora, Deepali, Varshney, Analysis of K-Means and K-Medoids Algorithm For Big Data, *International Conference on Information Security & Privacy* (ICISP2015), 2015.
[3] Big Data Strategy — Issues Paper, Commonwealth of Australia 2013.
[4] Carlo Vaccari, Big Data in Official Statistics - PhD Thesis in Computer Science - University of Camerino, JULY 2014.
[5] Cooper, Mell, Tackling Big Data, NIST Information Technology Laboratory Computer Security Division.
[6] Dash and Dash, Comparative Analysis of K-means and Genetic Algorithm Based Data Clustering. *International Journal of Advanced Computer and Mathematical Sciences*ISSN 2230-9624. Vol 3, Issue 2, 2012.
[7] ENISA, Privacy by design in big data, Final 1.0, Public, December 2015.
[8] European data protection supervisor, Meeting the challenges of big data, Opinion 7/2015.
[9] Gaddam, Securing your Big Data Environment, Black Hat USA 2015.
[10] Gandomi and Haider. Beyond the hype: Big data concepts, methods, and analytics, *International Journal of Information Management* (2015).
[11] Garg, Trivedi, Panchal, A Comparative study of Clustering Algorithms using MapReduce in Hadoop, *International Journal of Engineering Research & Technology (IJERT)*, ISSN: 2278-0181, Vol. 2 Issue 10, October – 2013.
[12] Geethakumari, Srivatsava, Big Data Analysis for Implementation of Enterprise Data Security, *International Journal of Computer Science and Information Technology & Security (IJCSITS)*, ISSN: 2249-9555 Vol. 2, No.4, August 2012.
[13] http://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT
[14] http://techspective.net/2016/02/04/stronger-security-requires-advanced-authentication/
[15] http://www.edupristine.com/blog/hadoop-ecosystem-and-components
[16] http://www.fruitionit.co.uk/2016/02/our-5-key-predictions-for-information-security-in- 2016/
[17] http://www.isaca.org/Knowledge/Center/Research/ResearchDeliverables/Pages/Security- As-A-Service.aspx
[18] http://www.sas.com/en_us/insights/big-data/internet-of-things.html
[19] http://www.tutorialspoint.com/hadoop/index.htm
[20] http://www.tutorialspoint.com/map_reduce/
[21] http://www8.hp.com/h20195/V2/GetPDF.aspx/4AA5-0863ENW.pdf
[22] https://datajobs.com/what-is-hadoop-and-nosql
[23] https://en.wikipedia.org/wiki/Internet_of_Things
[24] https://www.ciphercloud.com/blog/cloud-encryption-trends-predictions-for-2016-taking-a-proactive-approach-to-data-protection/
[25] https://www.mssqltips.com/sqlservertip/3132/big-data-basics--part-1--introduction-to-big-data/
[26] Inukollu, Arsi and Ravuri, Security Issues Associated With Big Data in Cloud Computing. *International Journal of Network Security & Its Applications (IJNSA)*, Vol.6, No.3, May 2014.

[27] Jainendra Singh, Big Data Analytic and Mining with Machine Learning Algorithm, *International Journal of Information and Computation Technology* ISSN 0974-2239 Volume 4, Number 1 (2014).

[28] K.U. and David, Issues, Challenges, and Solutions: Big Data Mining. M.E.S College, Marampally, Aluva, Cochin, India.

[29] Kaisler, Armour, Espinosa, Money, Big Data: Issues and Challenges Moving Forward. 2013 46th Hawaii International Conference on System Sciences.

[30] Khan, Yaqoob, Hashem, Inayat, Ali, Alam,Shiraz, Gani, Big Data: Survey, Technologies, Opportunities, and Challenges, Volume 2014, Article ID 712826.

[31] Samuel1, RVP2, Sashidhar3 and Bharathi4, A Survey on Big Data and Its Research Challenges, *ARPN Journal of Engineering and Applied Sciences*, VOL. 10, NO. 8, MAY 2015.

[32] Saraladevia, Pazhanirajaa, Paula, Bashab and Dhavachelvanc,Big Data and Hadoop-A Study in Security Perspective,2nd International Symposium on Big Data and Cloud Computing (ISBCC'15).

[33] Shirudkar, Motwani, Big-Data Security. *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 5, Issue 3, March 2015.

[34] Thakur and Manish Mann, Data Mining for Big Data: A Review, *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 4, Issue 5, May 2014.

[35] Toshniwal, Ghosh, Nath, Big Data Security Issues and Challenges. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, ISSN: 2349-2163 Issue 2, Volume 2 (February 2015).

[36] Ularu, Puivan, Apostu,Velicanu, Perspectives on Big Data and Big Data Analytics, Database Systems Journal vol. III, no. 4/2012.

[37] Ustaomer, Information Security in Big Data: Privacy and Data Mining(IEEE, 2014)

[38] Venkatesh H, Perur, Jalihal, A Study on Use of Big Data in Cloud Computing Environment. *International Journal of Computer Science and Information Technologies*, Vol. 6 (3), 2015, 2076-2078.

[39] Vinit Gopal Savant, Approaches to Solve Big Data Security Issues and Comparative Study of Cryptographic Algorithms for Data Encryption, *International Journal of Engineering Research and General Science* Volume 3, Issue 3, May-June, 2015.

[40] Yadav,Wang and Kumar, Algorithm and approaches to handle large, IJCSN *International Journal of Computer Science and Network*, Vol 2, Issue 3, 2013.

