ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Research Article*
www.ijptonline.com

EFFECTIVE CLUSTERING INBIG DATA FOR EFFICIENT KNOWLEDGE DISCOVERY USING PARALLEL K-MEANS AND ENHANCED K-MEANS ALGORITHM

E.Vijayan[1], N.C.Senthil Kumar[2], SiddhantAgnihotry[3], MehaboobSubuhani[4]
School of Information Technology and Engineering, VIT University.
Email: vijayvsi81@gmail.com

Received on 06-08-2016 *Accepted on 10-09-2016*

Abstract

Now days the environment produces large amount of data which is needed to be analysed, stored and secured in an efficient way. If we use BIG DATA efficiently, it can solve tremendous amount of problems. For analysis of big data the concept of Data Miningalong with the concept of machine learning is used. In this paper we are going to discuss on how to use parallel k-means algorithm and the adaptive k-means algorithm which reduces the iterations to cluster in big data, but time complexity of classic k-means algorithm is always high in large data sets. We have implemented an algorithm which is the fusion of parallel k-means algorithm with reducing the number of iterations. For our implementation we took prone accidental zones data set because life is more important than everything.We will also discuss about the applications of this algorithm that are used in real time environment. Last but not least this thesis will also discuss the various issues and future challenges in BIG DATA environment.

General Keywords: Time complexity, algorithm, cluster.

Keywords: BIG DATA, classic k-means, parallel k-means, adaptive k-means.

Introduction: BIG DATA will be useful if we analyse and make use of the data inan efficient way. Till date handling BIG DATA is a big challenge. If we fuse the concept of clustering inBIG DATA we can overcome certain disadvantages of the big data. For our study we took this application data sets which is also a big data by considering the data sets of all the prone zones for accidents in one location and placing the ambulance in a correct prone zone so that any accident happens, the ambulance can go on time to the spot to save the life. If the ambulance is not in the correct prone zone then the consequences are really high. Hence keeping this in mind we have developed an algorithm which works under less time complexity to place the ambulance in the correct locations. Clustering is a technique for pigeonholing the elements which are similar to the cluster but dissimilar to other clusters. Various algorithms are developed using parallel k-means and reducing the iterations but in this paper we will fuse both the

algorithms parallel k-means and algorithm to reduce the iterations. This concept is implemented so that we can handle the BIG DATAefficiently and reduce the time complexity to analyse the data.

Literature Survey:

1. According to**V.Ramesh , K.Ramar , S.Babu**in their paper[3] they have discussed about the parallel k-means algorithm where the master divide the work to all the slaves and compute the mean of all the slaves.

2. According to **MugdhaJain**and **ChakradharVerma**[7] they have enhanced the classic k-means algorithm by reducing the number of iterations of the k-means algorithm.

3. According to ChanchalYadav, Shuliang Wang and Manoj Kumar in their paper they have discussed about the various algorithms to handle BIG DATA.

4. T.Rathika and J.SenthilMurugan they conveyed about the various technologies and tools to handleBIG DATA.

Description about Big Data:

BIG DATA is the key for success for all the organizations. Big data can be classified as **4V'sVolume,Velocity, Variety, Value.**

Our view on BIG DATA:

1. Walking through this technology called BIG DATAwe can define this jargon as - "Big data is the collection of structured and unstructured data that comes from different sources which can't be handled by the traditional database systems, so we need to use data mining techniques to analyse these large data sets and get the knowledge out of these chunks of data, which is really large."

This proves that BIG DATA is totally different from the data that is stored in the normal databases and warehouses. Hence to handle this big data different technologies is been introduced and handling big data is really very complex in reality. We need to have entirely different frameworks in order to handle big data.

Right now the challenges of big data are -

1. The system we developed should handle very large amount of data to handle the BIG DATA.

2. Providing security to the BIG DATA is complex

3. Mining the efficient data from the large chunks of data is complex.

Clustering in BIG DATA:

Clustering in BIG DATA is useful in order to measure the similarity of the data in the large chunks of data hence once we have groups of similar data then it will be so easy to analyse required information with less time complexity.

There are many clustering algorithms which can be used according to the requirement.

These algorithms can be classified according to the following characteristics:-

1. Time complexity.

2. Depending on data order.

3. Finding the clusters of irregular shape.

4. Working in high dimensional data.

A good clustering algorithm proposesarbitrary shape clusters. Many clustering algorithms has only convex shaped clusters.

K-means algorithm:

The classic k-means algorithm has a the time complexity of O(n^2) and it resembles like a template for all the clustering concepts and it was the key algorithm as well. This k-means algorithm provides the better result and provides as good results however the k-means algorithm is very expensive when we go with large data sets.

This is an iterative algorithm which works on iterations and the first initial centroids are given to the clusters then the algorithm walks through these steps

1. Once after the centroids are assigned the algorithm computes the Euclidean distance of all the data points where the distance is smaller from the computed clusters.

$$\text{The distance function between two points } a=(x1, y1) \text{ and } b=(x2, y2) \text{ is defined as: } \rho(a, b) = |x2 - x1| + |y2 - y1|$$

2. Once the clusters is associated in step-1 it associates once again or recomputed once again in order to have a good mean between the data point (or) accurate means of the data points.

3. Step 1 and step 2 is repeated until the accurate mean of the clusters is achieved.

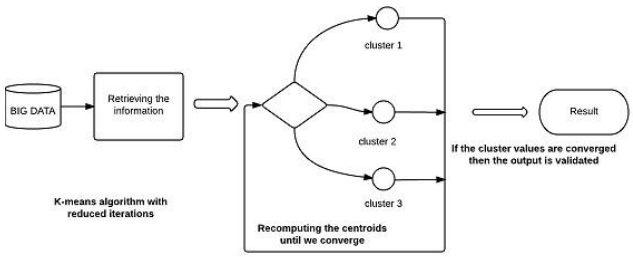


Fig 1: k-means algorithm.

Enhanced K-means by reducing the iterations:

MugdhaJainand **ChakradharVerma**has gone through the classic K-means algorithm and enhanced the classic k-

means algorithm by reducing the complexity of the algorithm. This algorithm says that the number of iterations required is the truly depends on the number of attributes need for clustering.

The main aim of this algorithm is when we reduce the iterations of the classic k-means algorithm by not initializing the centroid in prior hence the centroid is randomly generated and we don't need to update the centroid each and every time it automatically does update the cluster.

Parallel k-means algorithm:

In parallel K-means algorithm **V.Ramesh , K.Ramar , S.Babu**has deployed that this algorithm can be used for large data sets which in turn called as Big data this algorithm can scale up to O(K) times in a single machine. They have implemented this algorithm using MPI(Message Passing Interface).

Working Methodology:

1. **This** algorithm use MPI(message passing interface) method where the data set is split up among the machines.
2. Each data set is taken through the machines and they compute the centroids hence forth we compute the centroids in parallel.
3. Hence the mean of the centroids is calculated by the supervisor of all the machines (Controller).
4. If the result is converged the process terminates else the supervisor sends again the data set to all the machines once again to calculate the centroids it goes on.

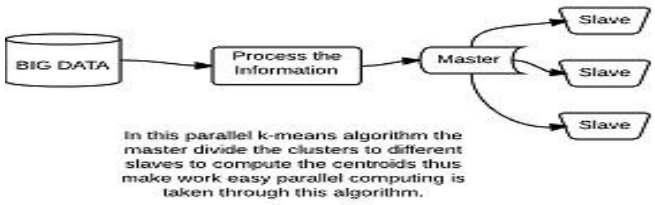


Fig: parallel k-means algorithm.

Proposed algorithm:

We have gone through two algorithms hence we develop our algorithm which do reduce the iterations and our algorithm works parallel as well. As we have seen that parallel K-means algorithm use MPI this methodology use threads we share the work to the kernel threads in the OS. Rather than using MPI this gives a better performance when we compute in a single machine.

Methodology

1. The working methodology is same as we the k-means algorithm were we generate the random centroids in the

first phase.

2. Once the centroids are generated the data set is divided and given to the available threads (Individual threads are given a separate set of datasets).
3. Once the threads compute the centroids the new centroids are updated.
4. If converged then we stop the process and calculate the mean.
5. Once all the clusters are assigned in a correct spot the new mean of all these clusters can be calculated by the Master thread.

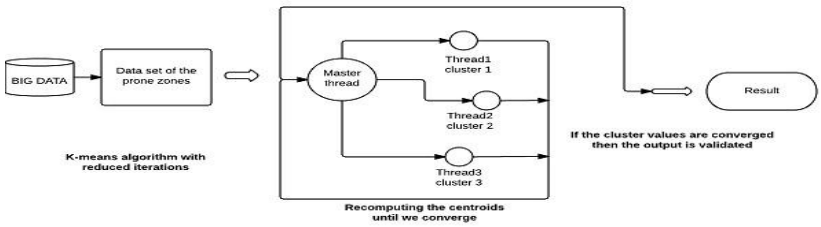


Fig: fusion of parallel k-means& adaptive k-means algorithm.

Benefits:

Reducing the iterations:

Reducing the iterations of the classic K-means algorithm by randomly generating the clusters or centroids.

Parallel processing:

Implanting the k-means algorithm in threads which is simpler to implement and this will achieve that it can handle a big data.

Application:

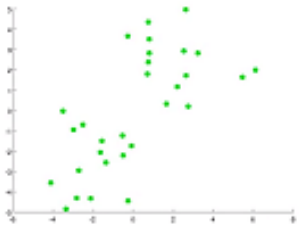
1. We use the data set of the prone zone accidents areas where we can have the ambulance in the correct spot so that it covers the accidental zones of all the prone areas.
2. We consider the ambulances to be the clusters and the data points are the prone zones.
3. Once we place the ambulance in the spot we calculate the mean of all the clusters (ambulance) and place the hospitals in that mean.

Results:

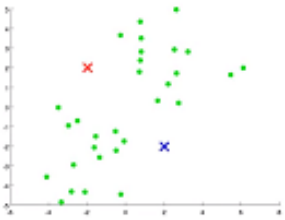
We use Net beans 7.02 and Microsoft Visual Studio and hence we have tested our tool with the student's data sets in a university aswell. We can conclude that the performance of the two algorithms can be fused together in order to give a tremendous performance and reduced time complexity. We use two implementations to differentiate the

normal k-means and k-means implemented through threads using Microsoft visual studio.

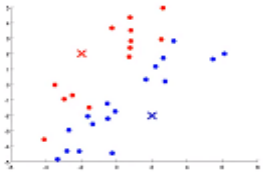
1) Green says the prone areas(data points).



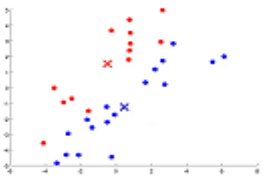
2) Now randomly placing the centroids anywhere in the graph which is depicting the data points. Here we assumed two centroids, but it's just a prototype as data sets increases the number of centroids increases.



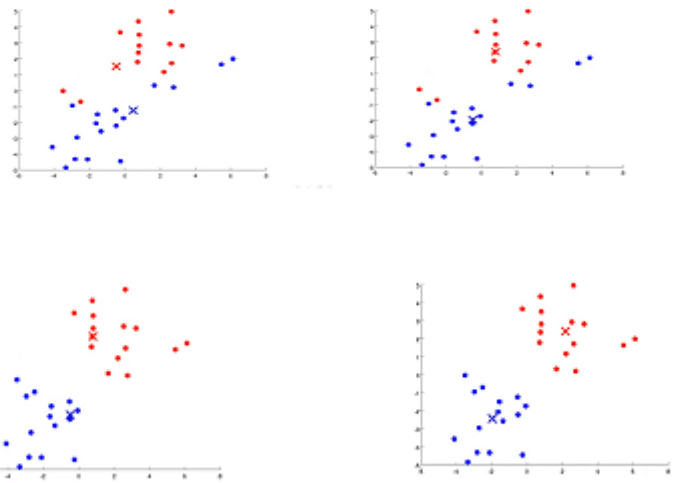
3) After calculating Euclidean distance we are dividing the data points into clusters. Here in this case the data sets is divided into two clusters.



4) Again calculate the Euclidean distance and get the new coordinates for the centroids and place the centroids on the new coordinates.



5) Repeat Step – 4 until our coordinates for centroids is fixed i.e. they are not changing on iterations.



Conclusion:

We conclude by saying that our implementation of this algorithm can handle a large data set where algorithms fail to handle if the data sets arehuge. This algorithm using java cant have the power of parallel as well as reduced iterations fusing both of these algorithms together we get the efficient algorithm as a whole here by we conclude this paper and future work is given below.

Future work:

Since due to the time constraint we can implement this in JAVA language but if this is implemented in the concept of threads according to the methodology we proposed. We are sure that the result comes out of the process is more efficient than k-means and parallel k-means algorithm.

References:

1. 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.
2. 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.
3. Understanding how is the k-means algorithm works http://www.edureka.co/blog/k-means-clustering/
4. Applications of clustering in real life http://www.slideshare.net/EdurekaIN/applications-of-clustering-in-real-life
5. What is K-means algorithm http://www.onmyphd.com/?p=k-means.clustering&ckattempt=2
6. K-means clustering with Map reduce http://codingwiththomas.blogspot.in/2011/05/k-means-clustering-with-mapreduce.html
7. MugdhaJain, Chakradhar Verma Adapting k-means for Clustering in BIG DATA.pdf from International Journal of Computer Applications (0975 – 8887), Volume 101– Number1, September 2014.
8. https://www.gliffy.com/go/html5/launch?app=1b5094b0-6042-11e2-bcfd-0800200c9a66
9. https://www.draw.io/
10. https://www.lucidchart.com/documents/edit/ad26553b-e8d7-4a1c-a6c5-a4bde63908f4

Corresponding Author:

E.Vijayan*,

Email: *vijayvsi81@gmail.com*