Computational Time Analysis of K-mean Clustering Algorithm

[1]Praveen Kumari, [2]Hakam Singh, [3]Pratibha Sharma
[1]Student Mtech, CSE 4th SEM, [2]Assistant professor CSE, [3]Assistant professor CSE
Career Point University Hamirpur

***Abstract*-Knowledge discovery in databases is the non-trivial process of identifying valid, novel potentially useful and ultimately understandable patterns from data. In the area of software, data mining technology has been considered as useful means for identifying patterns and trends of large volume of data. This approach is basically used to extract the unknown pattern from the large set of data for business as well as real time applications. It is a computational intelligence discipline which has emerged as a valuable tool for data analysis, new knowledge discovery and autonomous decision making. The raw, unlabeled data from the large volume of dataset can be classified initially in an unsupervised fashion by using cluster analysis i.e. clustering the assignment of a set of observations into clusters so that observations in the same cluster may be in some sense be treated as similar. The outcome of the clustering process and efficiency of its domain application are generally determined through algorithms. The aim this research is to analyze the computation time of k-mean clustering by varying the sample rate using stopwatch for time measurement.**

***Keywords:* data mining, clustering, and k-mean clustering.**

I. INTRODUCTION

Knowledge discovery in databases is the non-trivial process of identifying valid, novel potentially useful and ultimately understandable patterns from data. The process predicts the future trends and behaviors from accumulated large volumes of data to make proactive, knowledge driven decisions. At an abstract level, the core of knowledge discovery process is to map low-level data which is too voluminous in nature to compact, predictive model for estimating the values of future cases. The steps involved in the KDD (knowledge discovery in database) process is diagrammatically represented below.

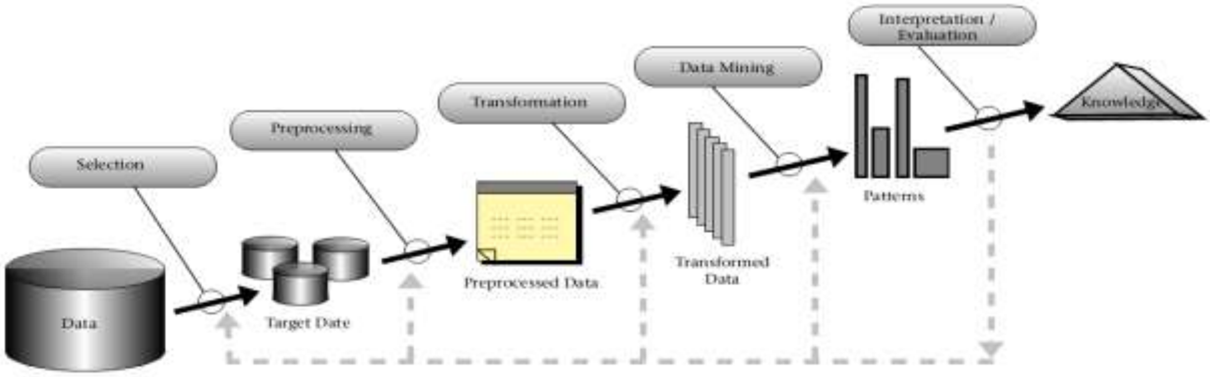


Fig 1: The Overview of the steps in the KDD process

The knowledge discovery process starts with the understanding of the application domain and identifying the goal of the process from the user's perspective. A target data which is relevant to the analysis task is selected on which the discovery process is to be applied. The selected data may be considered dirty which may contain noisy values
or missing values. The preprocessing step cleans the dirty data to a more consistent form. Next step in the process transforms the data into a more consolidated form appropriate for mining by performing operations like normalization, aggregation etc. Data reduction helps to find useful features which can represent the large dataset eliminating the redundant factors.

The next step , which is the actual data mining process does the exploratory analysis on the data and build some model based on known techniques of statistics, neural networks , machine learning and pattern recognition. Data mining[1] is the step in KDD process that consists of applying data analysis and discovery algorithms on preprocessed subsamples, and transformed data. Data mining is the application of specific algorithms for extracting patterns from data.

In general, Data mining has four major relationships. They are:
(a) Classes
(b) Clusters
(c) Associations

(d) Sequential patterns.

(a) Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

(b) Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

(c) Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

(d) Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes[6].

II. CLUSTERING

Clustering can be considered the most important unsupervised learning problem; so, as with every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. The process of organizing objects into groups whose members are similar in some way is a cluster. A collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters[7].

Types of Clustering Algorithms

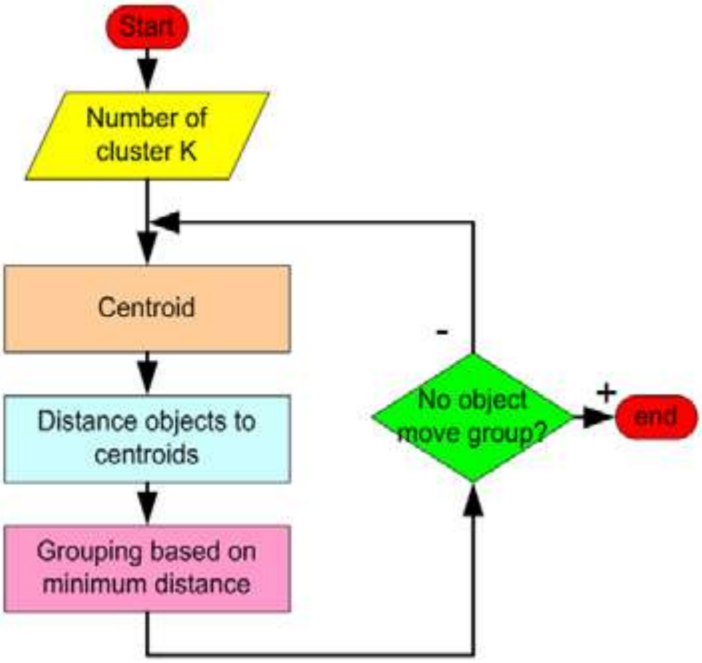
A. Hierarchical Clustering Algorithm: The hierarchical clustering algorithm[8] is a group of data objects forming a tree shaped structure. It can be broadly classified into agglomerative hierarchical clustering and divisive hierarchical clustering. In the agglomerative approach, which is also called as the bottom up approach, each data point is considered to be a separate cluster, and on each iteration the clusters are merged, based on a criterion.

B. Spectral Clustering Algorithm: Spectral clustering refers to a class of techniques, which relies on the Eigen structure of a similarity matrix. Clusters are formed by partitioning data points using the similarity matrix. Any spectral clustering algorithm will have three main stages. They are preprocessing, spectral mapping and post mapping. Preprocessing deals with the construction of the similarity matrix. Spectral Mapping deals with the construction of Eigen vectors for the similarity matrix. Post Processing deals with the grouping of data points.

C. Density based Clustering Algorithm: The density based algorithm allows the given cluster to continue to grow as long as the density in the neighborhood exceeds a certain threshold. This algorithm is suitable for handling noise in the dataset. The following points are enumerated as the features of this algorithm: it handles clusters of arbitrary shape, Handles noise, needs only one scan of the input dataset, and the density parameters to be initialized. DBSCAN, DENCLUE and OPTICS [4] are examples of this algorithm [7].

III. K-mean Clustering Algorithm

K-means clustering is a partition-based cluster analysis method. According to this algorithm we firstly select k data value as initial cluster centers, then calculate the distance between each data value and each cluster center and assign it to the closest cluster, update the averages of all clusters, repeat this process until the criterion is not match. K-means clustering[11] aims to partition data into k clusters in which each data value belongs to the cluster with the nearest mean.


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

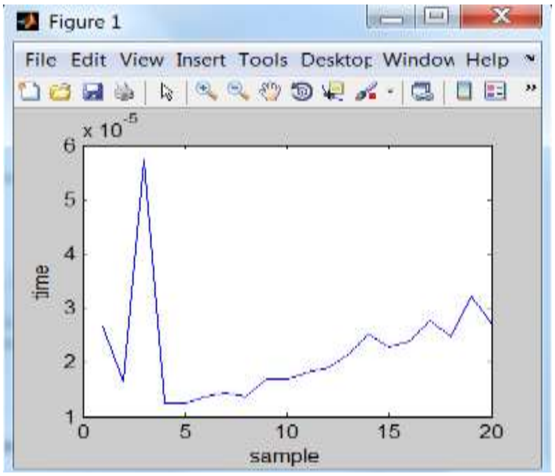
Flow Chart of K-mean Algorithm[19]

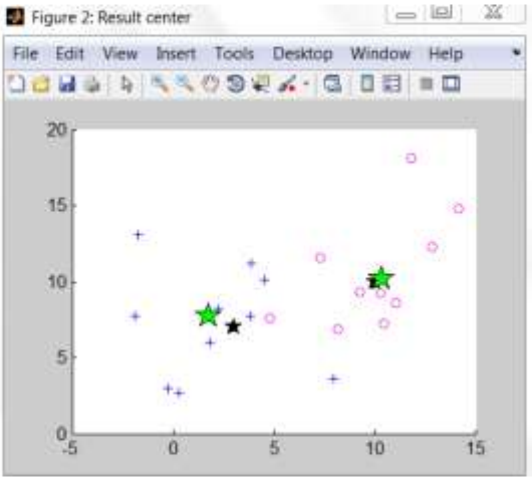
k-mean algorithm

```
1) Arbitrarily choose k data item from D dataset as initial cluster centriod
2) Calculate d(di, cj) = the distance between each data item and all k cluster center. Assign data item to nearest   cluster. Where di : is data item in data set D
          (1<=i<= n)
          Cj : is the cluster center
          (1<=j<= k)
3) For each data object di find the closest center cj assign di to cluster j.
4) Assign the sample rate.
5) For time measurement use stopwatch function.
6) Store the label of cluster center in which data object di is in array cluster[]. Store distance of data object di to the      nearest cluster in array dist[].
7) Set Cluster[i] = j, j is the label of nearest cluster.
8) Set Dist[i] = d(di, cj), d(di, cj) is the nearest Euclidean distance to the closest center.
9) For each cluster j(1<=j<=k) calculate cluster center;
10) Repeat
11) For each data object di compute its distance to the center of present nearest cluster
    a. If this distance<=Dist[i] then
       Data object stay in initial cluster;
    b. else
       for every cluster center cj compute distance d(di, cj) of each data object to all the center assign data object di to
       nearest cluster center cj.
       set cluster[i]=j;
       set Dist[i]=d(di, cj);
12) For each cluster center j recalculate the centers;
13) Until convergence criteria met
14) Output the clustering results;[15]
```

1V. EXPERIMENTAL RESULTS AND SIMULATION:

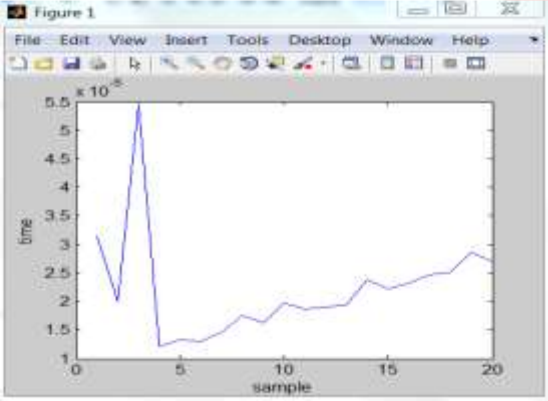
1. When we take the value of sample rate = 10, the output is

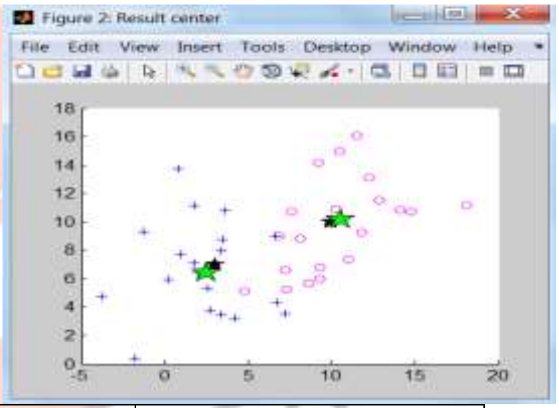




Sr. No.	Sample	Time
1.	0-5	5.8
2.	5-10	1.8
3.	10-15	2.5
4.	15-20	3.2

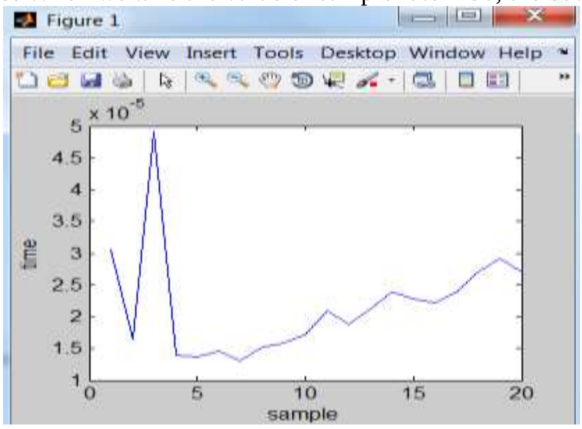
2. When we take the value of sample rate = 20, the output is

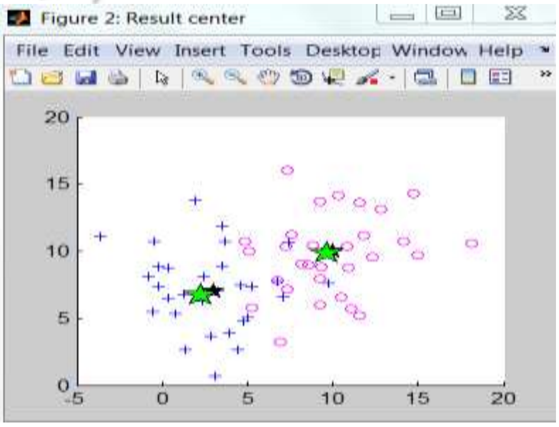




Sr. No.	Sample	Time
1.	0-5	5.5
2.	5-10	2
3.	10-15	2.4
4.	15-20	2.8

3. When we take the value of sample rate = 30, the output is





Sr. No.	Sample	Time
1.	0-5	4.9
2.	5-10	1.8
3.	10-15	2.4
4.	15-20	2.8

CONCLUSION: K-mean is one of the most typical and widely used algorithm for clustering. This research paper analyses the standard algorithm and gives an approach that the increasing sample rate in k-mean clustering algorithm decreases the computation time of cluster formation. So from the above result it has been proved that K-mean clustering algorithm is best for hugged data.

ACKNOWLEDGEMENT: The author express their sincere thanks to Department Head of Computer Science And Engineering Mrs. Pratibha Sharma, Ass. Prof. Mr. Hakam Singh, Career Point University Hamirpur (H.P.) for theis valueable suggestions and guidelines and I would also like to thanks my friend Vinesh Thakur for her help in this research.

REFERENCES

[1] Vitthal Yenkar, Prof. Mahip Bartere, “Review on “Data Mining with Big Data” ”, International Journal of Computer Science and Mobile Computing, Vol.3 Issue. 4, April-2014,pg.97-102.

[2] Nikita Jain, Vishal Srivastava, “DATA MINING TECHNIQUES: A SURVEY PAPER”, International Journal of Research in Engineering and Technology, eISSN: 2319-1163| pISSN: 2321-7308.

[3] Kalyani M Raval, “Data Mining Techniques”, International Journal of advance Research in Computer Science and Software Engineering, Volume 2, Issue 10, October 2012.

[4] Sumit Garg, Arvind K. Sharma, “Comparative Analysis of Data Mining Techniques on Educational Dataset”, International Journal of Computer Applications (0975-8887), Volume 74-No. 5, July 2013.

[5] Neelamadhab Padhy, et.al, “The Survey of Data Mining Applications And Feature Scope”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT),Vol.2, No.3, June 2012.

[6] Dr. Sankar Rajagopal, “Customer Data Clustering Using Data Mining Technique”, International Journal of Database Management Systems (IJDMS) Vol. 3, No. 4, November 2011.

[7] P. IndiraPriya, Dr. D.K. Ghosh, “A Survey on Different Clustering Algorithms in Data Mining Technique”, International Journal of Modern Engineering Research (IJMER) Vol.3, Issue.1, Jan-Feb. 2013 pp-267-274 ISSN:2249-6645.

[8] Amandeep Kaur Mann, Navneet Kaur, “Survey Paper on Clustering Techniques”, International Journal of Science, Engineering and Technology Research Volume 2, Issue 4, April 2013.

[9] Astha Joshi, Rajneet Kaur, “A Review: Comparative Study of various Clustering Techniques in Data Mining”, International Journal of advance Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.

[10] Yujie Zheng, “Clustering Methods in Data Mining with its Applications in High Education”, International Conference on Education Technology and Computer (ICETC2012), IPCSIT vol. 43(2012).

[11] Narendra Sharma, et.al, “Comparison the various clustering algorithms of weka tools”, International Journal of Emerging Technology and Advance Engineering (ISSN 2250-2459, Volume 2, Issue 5, May 2012).

[12] Osama Abu Abbas, “Comparisons Between Data Clustering Algorithms”, The International Arab Journal Of Information Technology, Vol. 5, No. 3, July 2008.

[13] Tapas Kanungo, et.al, “An Efficient k-Means Clustering Algorithm: Analysis and Implementation”, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, July 2002.

[14] Barkha Narag, et.al, “Application based, advantageous K-means Clustering Algorithm in Data Mining- A Review”, International Journal of Latest Trends in Engineering and Technology (IJLTET) ISSN: 2278-621X.

[15] Jyoti Yadav, Monika Sharma, “A Review of K-mean Algorithm”, International Journal of Engineering Trends and Technology (IJETT) – Volume 4, Issue 7 – July 2013.

[16] Ming-Chuan Hung, et.al, “An efficient k-mean Clustering Algorithm Using Simple Partitioning”, Journal of Information Science and Engineering 21,1157-1177 (2005).

[17] Neha Aggraval, Kirti Aggraval, “A Mid – Point based k-mean Clustering Algorithm for Data Mining”, International Journal on Computer Science and Engineering (IJCSE).

[18] Madhu Yedla, et.al, “Enhancing K- means Clustering Algorithm with Improved Initial Center”, International Journal of Computer Science and Information Technologies, Vol. 1(2), 2010, 121-125.

[19] Dr. Aishwarya Batra, “Analysis and Approach: K-Means and K-Medoids Data Mining Algorithms”, batra .ashwarya@gmail.com.