Performance analysis of K-means and CLARANS clustering algorithm

Anuradha Rani[1], Sushila Ratre[2] , Pranav More[3]
Department of Computer Science and Engineering,
Amity School of Engineering and Technology,
Amity University Mumbai
[1]arani@mum.amity.edu
[2]suratre@mum.amity.edu
[3]prmore@mum.amity.edu

Abstract - Big data is defined by four V's i.e. volume, variety, velocity and veracity. The biggest challenge in handling Big data is to manage noise and outlier present in dataset efficiently which makes it difficult to get right information. Most of the organizations are spend their time in cleaning and preparing of data. Applying various data mining clustering technique itself became a challenge in terms of execution time as well as cost. In this paper, we have applied k-mean & CLARANS clustering technique on Big data sets. We found that k-means works on big data but it is sensitive with respect to noise and outliers, and the same time if we apply CLARANS clustering algorithm for the same data set, it provides better result in terms of execution time as well as non-sensitive towards outliers.

Keywords: K-means, CLARANS, Big data

I. Introduction

Data Mining is the process of retrieving innovative and useful information or hidden pattern from raw data. Mining is the way to bring useful things out from superfluous and unproductive data[1].We are getting huge amount of data from various online sources day by day which is nothing else but big data and after applying different mining techniques like extraction, transformation, we get data which can be mined for getting useful patterns out of it[8]. This is where big data is related to data mining. Clustering is an unsupervised learning process. It is a way to group similar things together and apart dissimilar things in different groups[3]. There are various clustering algorithms available in data mining as follows.

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Partitioning Method

Suppose, we have n objects in the database, and k partitions are constructed on that database for the data[2]. All the n objects must belong to exactly one partition out of k partition. Main motive of partitioning method is to divide all the n data points into k clusters based upon some objective function for example Euclidean Distance [7]. There are mainly four categories of partitioning methods:

- K-mean
- K-medoid
- CLARA
- CLARANS

In this paper, we are working on comparative analysis of k-mean and CLARANS algorithm for big data. We will mainly focus on outliers of the data from the raw data.

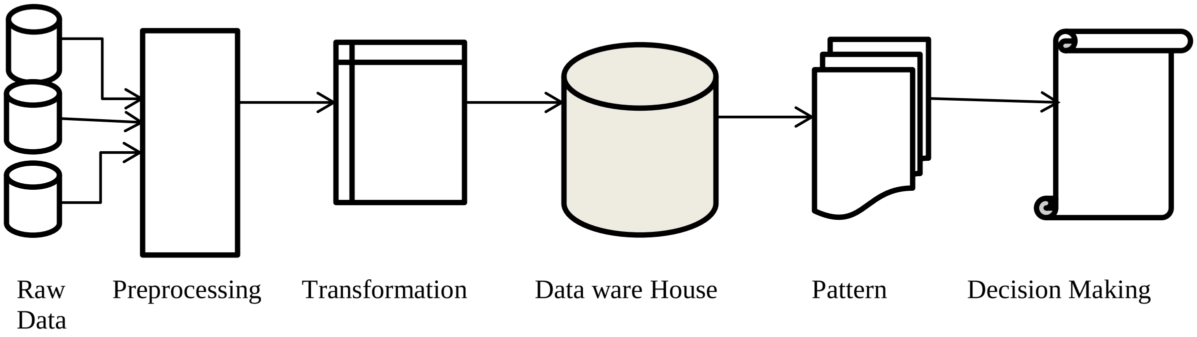

Figure1: Knowledge Discovery in Decision Making

II. Literature Review

Knowledge discovery from data is defined as drawing out of relevant patterns or knowledge from vast amount of data, where data mining becomes the heart of knowledge discovery[3].

For finding the knowledge from data[6], we use the process shown in figure above, first, we select raw data from data ware house, than we preprocess the data, where we clean the data and select attributes from the data, basically we will focus to remove missing values, unwanted values in cleaning process. Than we apply transformation, Dimensionality reduction is done by the transformation process.

K-Mean Algorithm: James Macqueen is developed k-mean algorithm in 1967. Center point or centroid is created for the clusters, i.e. basically the mean value of a one cluster[4]. We use objective function for creating clusters.

Suppose we have n data points and k clusters now we have to distribute n data points into k clusters depending upon their similarity which is based upon objective function [5].The algorithm is divided in two different steps, they are assignment step and update step.

Assignment step: In the Assignment step, Assume any k_x value from n data points as the mean value for k cluster and then assign each observation to any one cluster based upon nearest mean value which is calculated by using Euclidean distance formula between the mean of the cluster and data points.

$$\sum_{i=1}^{x}(n\mathrm{i} - mi)^2$$

Euclidean distance (m, n) = $\sum_{i=1}^{x}(\mathrm{ni}-\mathrm{mi})2$

Where x is the dimension of any data point in data set.

Update step: In the update step, we compute the new means to be the centers of the data points in the new cluster. mean$_\mathrm{i}$ = $\frac{1}{|\ C_i\ |}\ \sum_{x_p \in C_i} x_p$............equation 2

Update the cluster center from equation 2 and reassign each data to the cluster from equation 1 until no change in cluster center[11].

CLARANS stands for Clustering large Application Based on Randomized search. It is a partitioning method for the clustering of large database[9]. Ng and Han has discovered CLARANS in 1994 to overcome the limitations of K-Medoid and K-mean. It is based on medoids.

The following steps to be performed:

i.) Take two input parameters, num_local and max_neighbour.
ii.)Select any K object on random basis from the database object D.
iii.) Divide Si and non-selected Si, by Marking these K object as selected Si and rest of all other as non-selected Si.
iv.) Calculate the cost T for selected Si.
v.) We will get two values of T, Positive and Negative. If T is negative update medoid set. Otherwise selected medoid chosen as local optimum.
vi)Restart the selection of another set of medoid and find local optimum again.
vii.) CLARANS stops until returns the best.

According to the authors (ibid) CLARANS uses two parameters – numlocal and max neighbor. Numlocal means number local minima obtained and max neighbor means maximum number of neighbor examined. The higher the value of latter, the closer will be CLARANS to PAM and longer will each search of local minima. This is an advantage because the quality of local minima is higher and less number of local minima are to be found out[10].

Advantages of CLARANS:

- It is easy to handle outliers.
- CLARANS result is more the effective as compare PAM and CLARA.

Disadvantages of CLARANS:

- It does not guarantee to give search to a localized area.
- It uses randomize samples for neighbors.

III. Comparison of k-means and CLARANS

We have taken Sales data set from UCI machine repository. We have executed k means and CLARANS algorithm for the above mentioned data set. After execution, overlapping of clusters using CLARANS is better as compared to k-means as shown below in figure2 and figure3

k-means

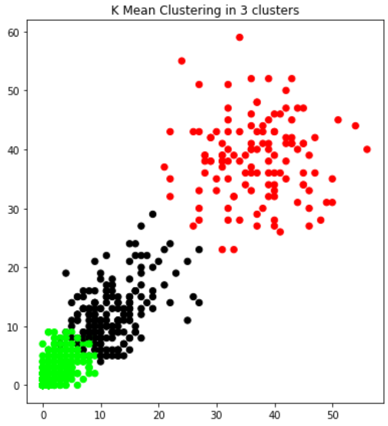

Figure 2

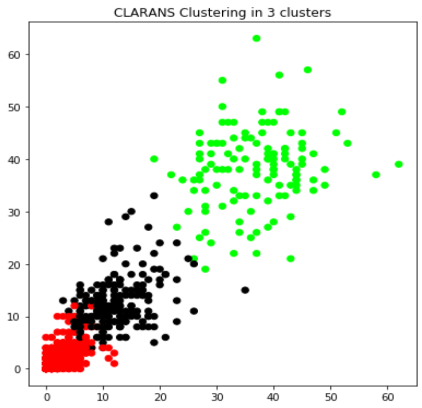

Figure 3

Execution time for k-means and CLARANS

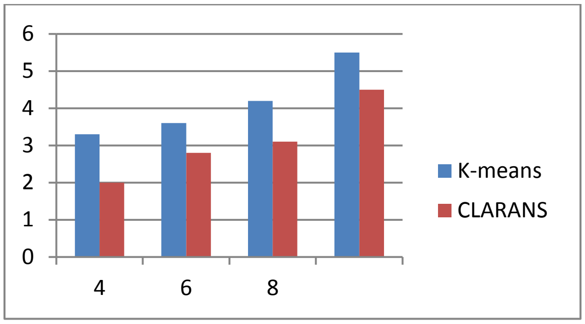

Figure 4

IV. Conclusion

In this paper we compared the performance of both K-mean and CLARANS clustering algorithms with respect to the number of clusters formed and distance metric. The comparison results show that time taken in cluster formation and overlapping of cluster is better in CLARANS rather than K-Means. Also the result of dataset shows that CLARANS is better in all aspects such as less execution time, less sensitive to outliers and noise.

V. Future Scope

We have compared K-Means and CLARANS in this paper, in future different types of partition based techniques can be made hybrid and with different distance metric, these can be compared for getting better result.

References

[1] J. Han and M. Kamber, Data Mining Concepts and Techniques, 2nd Ed. ,US: Moorgan Kaufmann Pub.,2011.

[2]. Swarndeep Saket J, Dr. Sharnil Pandya**,** An Overview of Partitioning Algorithms in Clustering Techniques

[3] Pradeep Rai and Shubha Singh, ―A Survey of Clustering Techniques‖, International Journal of Computer Applications (0975-8887) Vol 7-No. 12, pp. 1-5, October 2010.

[4]Rekha Awathi, Anil K Tiwari,Seema Pathak "Empirical evaluation on k Means clustering with effect of distance function for bank database" in IJITR ,pp2013,233-235 vol -2,2013

[5] T. Velmurugan and T. Santhanam Department of Computer Science, DG Vaishnav College, Chennai, India"Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points"

[6] P. Bradley, U. Fayyad, and C. Reina, "Scaling Clustering Algorithms to Large Databases," Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining, pp. 9–15, 1998.

[7] T. Velmurugan, T. Santhanam, "A Survey of Partition based Clustering Algorithm in Data Mining: An Experimental Approach" Information Technology Journal 10(3): 478-484, 2011 ISSN 1812-5638

[8] Ali Seyed Shirkhorshid, Saeed Aghabozorgi,. Teh Ying Wah, *and* Tutut Herawan" Big Data Clustering: A Review", B. Murgante et al. (Eds.): ICCSA 2014, Part V, LNCS 8583, pp. 707–720, 2014. © Springer International Publishing Switzerland 2014

[9] Raymond T. Ng and Jiawei Han," CLARANS: A Method for Clustering Objects for Spatial Data Mining", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 14, NO. 5, SEPTEMBER/OCTOBER 2002

[10] Garima, Hina Gulati, P.K.Singh, “Clustering Techniques in Data Mining: A Comparison”, IEEE, 2015

[11] Saurabh Shah, Manmohan Singh,” Comparison of A Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid algorithm”,IEEE, 2012