Comparative Analysis of various algorithms for large Dataset

Preeti Arora
Department of Computer Science and Engineering
Bhagwan Parshuram Insititute of Technology, New Delhi, India

Garvit Magoo
Department of Computer Science and Engineering
Bhagwan Parshuram Insititute of Technology, New Delhi, India

Mansi Bansal
Department of Computer Science and Engineering
Bhagwan Parshuram Insititute of Technology, New Delhi, India

Abstract- Clustering is a data mining technique where the dataset is grouped into various clusters in such a way that the data points of the same cluster are more similar to each other than to those lying in the other cluster. These clusters are built using different clustering algorithms. A comparative study of some of these clustering algorithms is performed here. The analysis is done using WEKA, which is a data mining tool. Using the Amazon dataset, K means and Farthest first algorithm are analysed. The clustering algorithms are analysed on the basis of time taken, number of seeds and number of clusters formed. The dataset used for the analysis is Amazon product co-purchasing network March 02, 2003.

Keywords – Cluster, K-means algorithm, Farthest First algorithm, Big Data.

I. Introduction

Clustering is grouping of data into groups of similar objects. These clusters (or groups) are built in such a way that the data points belonging to the same cluster are similar to each other and the data points belonging to different clusters are dissimilar to each other in one or the other way [1]. WEKA is used to analyze the algorithms over a dataset. WEKA provides a good interface for the comparison to the users, as compared to other data mining tools[2]. K-means and Farthest First clustering algorithm has been compared.

We have done the cluster analysis on big data because nowadays most of the data that is available to us, is BIG DATA. Big data is a term for datasets so large and complex that traditional data processing techniques are inadequate to be performed. The clustering analysis on such data helps in understanding the factors on the basis of which clusters have been made. We have picked up the Amazon dataset to analyze the three different clustering algorithms. The clustering algorithms are analyzed on the basis of time taken, number of seeds and number of clusters formed.

A. K-means Clustering Algorithm

K-means clustering algorithm is the simplest clustering algorithm. It is a partitioning clustering algorithm, which was first proposed by Macqueen in 1967. It classifies the given dataset into k different clusters It does so through an iterative method. And this iterative method tends to converge to a local minimum. The algorithm consists of two phases. In the first phase, the user specifies the number of clusters (k). The algorithm selects the k centers randomly. K means clustering algorithm calculates the distance between each data object and the cluster centers to assign each data point to a cluster. This distance is calculated by a distance function, called Euclidean distance.

The first step is completed when all the data objects are assigned to one or the other clusters. Now, the second phase of the algorithm starts. The second phase recalculates the average of the early formed clusters. This iterative process continues until the criterion function (Euclidean distance) becomes the minimum [3].

B. Farthest First Algorithm

Farthest first is a modified K-Means algorithm. This algorithm places each cluster center in turn at the point further most from the existing cluster center. This point must lies within the data area. This greatly increases the clustering speed in most of the cases since less reassignment and modification is needed [4].
Implements the "Farthest First Traversal Algorithm" by Hochbaum and Shmoys 1985: A best possible heuristic for the k-center problem, Mathematics of Operations Research, 10(2):180-184, as cited by Sanjoy Dasgupta "performance guarantees for hierarchical clustering"[6], colt 2002, Sydney works as a fast simple approximate clustered [6] modeled after Simple Means, might be a useful initialize for it Valid options are:
N -Specify the number of clusters to generate.
S -Specify random number seed

The paper is divided into V sections. The introduction and the aim of implementing the algorithms has been explained in the section I. The section II gives a brief introduction to the used dataset. The pseudo codes of the two algorithms are explained in section III. The IV and the V section explain the Experimental results and the conclusion respectively.

II. DATABASE USED

The dataset that has been used for the analysis of the three clustering algorithms is the Amazon product co-purchasing network March 02, 2003. The dataset is an undirected graph where there is an edge from a node i to node j , if the product i was bought along with product j , for the month March,2003. The dataset has two attributes , FromNodeId and ToNodeId and contains 1048573 rows.

Table 1. Description of Dataset

S.No	Attribute	Types	Range
1	**FromNodeId**	**Numeric**	**0-220105**
2	**ToNodeId**	**Numeric**	**1-206998**

II. COMPARISON AND ANALYSIS

A. K means algorithm

K means algorithm consists of two phases and follows the following steps to partition the dataset into clusters. It minimises an objective function called as squared error function[7]. This function is given by:
J(S)=$\sum_{k=1}^{m}\sum_{l=1}^{mi}(g_k - h_l)^2$
Where
(g_k – h_l) Is the euclidean distance between $g_k$and h_l.
m is the number of cluster centres.
mi is the number of data points in ith cluster.

1. Let G = {g_1,g_2,g_3,……..,g_n} be the set of data points and H = {h_1,h_2,…….,h_c} be the set of centers.
2. Select *'m'* cluster centers, randomly.
3. Repeat
4. Distance between each data point and cluster centers is calculated next.
5. The data point is assigned to that cluster center, whose distance from that data point is minimum.
6. the new cluster center is recalculated using :

   h_k = (1/mi) $\sum_{l=1}^{mi} g_k$
7. The steps are repeated until no new clusters are found.

B. Farthest First algorithm

Farthest first is a variant of K Means. In this algorithm the cluster center is placed at a point furthest from the existing cluster [4]. This point must lie within the data area. The points that are farther are clustered together first.

This feature of farthest first clustering algorithm speeds up the clustering process in many situations like less reassignment and adjustment is needed
The farthest-first traversal of a finite point set may be computed by a greedy algorithm that maintains the distance of each point from the previously-selected points, performing the following steps:

1. Initialize the sequence of selected points to the empty sequence, and the distances of each point to the selected points to infinity.
2. While not all points have been selected, repeat the following steps.
3. Scan the list of not-yet-selected points to find a point *p* that has the maximum distance from the selected points.
4. Remove *p* from the not-yet-selected points and add it to the end of the sequence of selected points.
5. For each remaining not-yet-selected point *q*, replace the distance stored for *q* by the minimum of its old value and the distance from *p* to *q*.

For a set of *n* points, this algorithm takes $O\ (n^2)$ steps and $O\ (n^2)$ distance computations

III. FACTORS FOR COMPARISON

Performance of K-means, Farthest First clustering algorithm has been analyzed based on the following factors:

1. Time taken by algorithms when numbers of clusters are same.
2. Time taken by algorithms when numbers of clusters are different.
3. Time taken by algorithms when numbers of seeds are different.

A dataset is applied to WEKA and the results for the number of clusters and time taken are noted. Table 2 describes the time taken by the algorithms when number of clusters is 2.

Table 2. Time taken (sec) to form clusters

Algorithm	Number of Clusters	Time Taken
K-means	2	23.07
Farthest First	2	2.39

The results after applying the dataset are noted. Figure 1 shows the graphical representation of the results

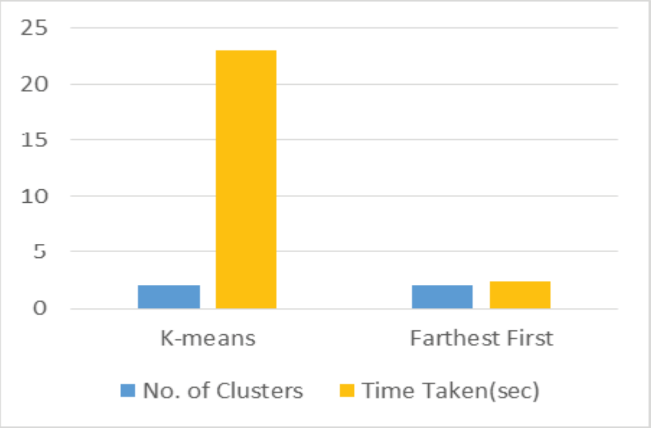

Figure 1. Time taken when number of clusters is same

Table 3 shows the time taken by K-means and farthest first algorithm as the number of clusters increase

Table 3. Time taken (sec) when number of clusters are different

Algorithm	Time Taken when number of clusters is 2	Time Taken when number of clusters is 3	Time Taken when number of clusters is 4
K-means	23.07	49.01	67.39
ssFarthest First	2.39	3.67	4.32

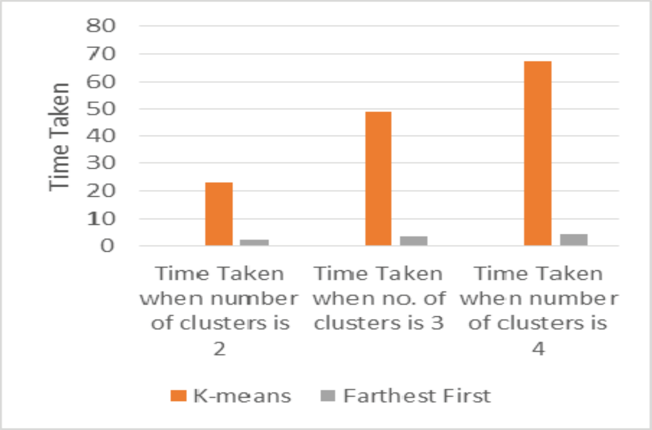

Figure 2. Time taken as number of clusters increase

Table 4 shows the time taken as we increase the number of seeds required initially for the algorithms

Table 4. Time taken (sec) when number of seeds are different

Algorithm	Time Taken when number of seeds is 10	Time Taken when number of seeds is 20	Time Taken when number of seeds is 50
K-Means	23.54	22.17	20.15
Farthest First	1.8	1.7	1.6

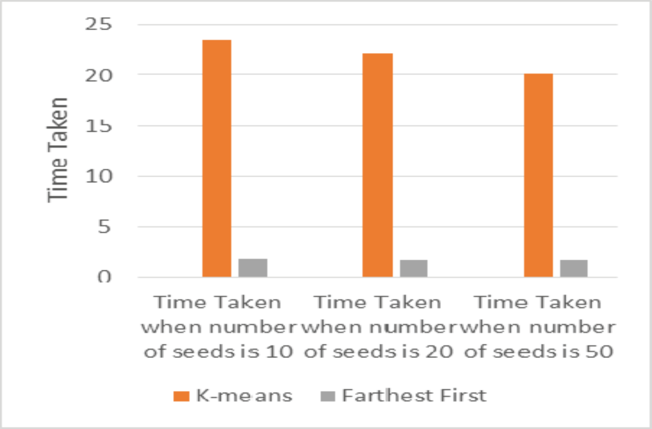

Figure 3. Time Taken as number of seeds increase

IV.CONCLUSION

A comparative study of K-means, farthest first algorithm has been performed. The performance has been measured on the basis of time taken by the algorithm when number of clusters is same, when number of clusters increase and when the number of seeds increases. The results are depicted in the form of graphs. From the results obtained we conclude that Farthest first takes less time than K-means algorithm when number of clusters are same. Also we can conclude time to form clusters when number of clusters increase, while it decreases if we increase the number of seeds.

REFERENCES

[1] Osama Abu Abbas "Comaprison between Data Clustering Algorithms" The International Arab Journal of Information Technology, Volume 5, July 2008.

[2] Garima Sehgal,Kanwal Garg, "Comparison of Various Clustering Algorithms", International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3074-3076

[3] Richa Loohach and Dr. Kanwal Garg " Effect of Distance Functions on Simple K-means Clustering Algorithm" International Journal of Computer Applications, Volume 49, July 2012.

[4] Narendra Sharma , Aman Bajpai and Ratnesh Litoria “Compariso the various Clustering Algorithms of weka tools”, International Journal of Emerging Technology and Advanced Engineering,ISSN:2250-2459,Volume 2, Issue 5, May 2012.

[5] S.Revathi,T Nalini, “Performance Comparison of Various Clustering Algorithm”, “2013 International Journal of Advanced Research in Computer Science and Software Engineering”, pp. 67-72, vol.24, 2013.

[6] E.B Fawlkes and C.L. Mallows. A method for comparing two hierarchical clusterings. Journal of the American Statistical Association, 78:553–584, 1983.

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