Comparative Study of Different Clustering Algorithms for Association Rule Mining

Ms. Pooja Gupta,
Ms. Monika Jena,
Computer science and Engineering Department,
Amity University,
Noida, India
pooja.srms08@gmail.com
mjena@amity.edu

Ms. Manisha Chowdhary,
Ms. Shilpi Singh,
CSE Department,
United group of Institutions,
Greater Noida, India
manisha.chowdhary5@gmail.com
4.shilpi@gmail.com

Abstract— In data mining, association rule mining is an important research area in today's scenario. Various association rule mining can find interesting associations and correlation relationship among a large set of data items[1]. To find association rules for single dimensional database Apriori algorithm is appropriate. For large databases lots of candidate sets are generated. Thus Apriori algorithm is not efficient for large databases. We need some extension in the existing Apriori algorithm so that it can also work for large multidimensional database or quantitative database. For this purpose to work with apriori in large multidimensional database, data is divided into multiple data sets called as clusters. In order to divide large data bases into clusters we need various clustering algorithms which can be based on Statistical methods, Hierarchical methods, Density Based method or Grid based method. Once clusters are created by these clustering algorithms, the apriori algorithm can be easily applied on clusters of our interest for mining association rules. Since overall process of finding association rules highly depends on clustering algorithms so we have to use best suited clustering algorithm according to given data base ,thus overall execution time will be reduced. In this paper we have compared various clustering algorithms according to size of data set and type of data set.

Keywords- Apriori algorithm, Clustering, Statistical methods, Hierarchical methods, Density Based method, Grid based method, CHAMELEON, BIRCH, DBSCAN, CLARANS.

I. INTRODUCTION

In association is a pattern that states when X occurs, Y occurs with certain probability. In data mining association rule mining can find interesting associations and correlation relationship among a large set of data Items which is represented by A→B where A and B two item sets with property A∩B=ø, A ≠ ø, B ≠ ø. The association is set to be interesting if it satisfy both a minimum support and minimum confidence [8].

Clustering is a process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. The quality of cluster may be represented by its diameter, the maximum distance between any two objects in the cluster. Division of large data into smaller data sets or clusters is done by various clustering algorithms which are based on Statistical methods, Hierarchical methods, Density Based method or Grid based method. After finding clusters by these clustering algorithms, the apriori algorithm can be easily applied on clusters of our interest for mining association rules.

There are some clustering algorithms such as CHAMELEON, BIRCH, DBSCAN, and CLARANS which are used for dividing the large data set into clusters. CHAMELEON and BIRCH are based on hierarchical clustering method. DBSCAN is based on density based method and CLARANS is grid based clustering method. Our main focus is to study the behavior of these algorithms and then compare their relative advantages and disadvantages.

II. CLUSTERING ALGORITHM

Clustering is a task of grouping a set of objects in such a way that objects in the same group(called cluster) are more similar to each other than to those in other groups. Clustering is an unsupervised learning that means there are no pre defined classes.

Chameleon is a hierarchical clustering algorithm that uses dynamic modeling to determine the similarity between pairs of clusters. In chameleon, cluster similarity is assessed based on how well connected objects are

within a cluster and the proximity of clusters. In this algorithm two clusters are merged if their interconnectivity is high and they are close together. Chameleon uses a k- nearest neighbor graph approach to construct a sparse graph where each vertex of the graph represent data object and an edge between two vertices exist if one object is among the k most similar objects to the other.

Chameleon uses a graph partitioning algorithm to partition graph k-nearest graph into a large number of relatively small sub clusters. Thus a cluster C is partition into sub clusters C_i and C_j.then it uses an agglomerative hierarchical clustering algorithm that iteratively merges sub clusters based on their similarity. chameleon determine similarity between each pain of clusters C_i and C_j according to their relative connectivity, $RI(C_i, C_j)$ and their relative closeness $RC(C_i, C_j)$[8].

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is designed for clustering a large amount of data bases. It is recognized as the first clustering algorithm proposed in the database area to handle noise effectively. BIRCH incrementally constructs a CF (Clustering feature) Tree. It is a multiphase clustering technique in which phase 1 scans database to build an initial in- memory CF tree and phase 2 uses an arbitrary clustering algorithm to cluster the leaf nodes of the CF tree.

DBSCAN (Density based spatial clustering of application with noise) is a density based clustering algorithm, providing generation of a number of clusters starting from the estimated density distribution of corresponding notes DBSCAN is based on two main concepts :density reachablity and density connectiblity. Both of these concepts depend on two input parameters: size of epsilon neighborhood e and the minimum points in a cluster m. The number of point parameter impacts detection of outliers. Points are declare to be outlier if there are few other points in the e –Euclidean neighborhood parameter controls the size of neighborhood, as well as size of clusters. if e is big enough, then it would be one big cluster and no outliers in the figure[1].

CLARANS (Clustering Algorithm based on Randomized Search) is a clustering process which can be presented as searching a graph where every node is a potential solution. It draws sample of neighbors dynamically. It is more efficient and scalable. It's main aim is to identify spatial structure that may be present the data. It can handle not only point objects, but also polygon objects efficiently [9].

III. COMPARISON METHODOLOGY

Chameleon, BIRCH, DBSCAN and CLARANS are compared according to the following factors:

- The size of data sets
  - i) Based on large data set
  - ii) Based on small data set
- Type of data set

For each factor, four tests are made one for each algorithm. For example for size of data each algorithm is executed twice one for small size data set and another for large size of the dataset. The Table 1 explains how the four algorithms are compared.

TABLE I. COMPARISON METHODOLOGY

Name of Algorithm	Size of Data Set	Type of the Data set
Chameleon	Large data set and small data set	Average data set
BIRCH	Large data set and small data set	Average data set
DBSCAN	Large data set and small data set	Average data set
CLARANS	Large data set and small data set	Average data set

IV. RESULT ANALYSIS

We have analyzed the performances of clustering algorithms on the basis of the size of data sets in both situations, that is, small size data sets and large size data sets. Fig.1 shows the relationship between data set (small size) and performance of four algorithms: Chameleon, BIRCH, DBSCAN, and CLARANS.

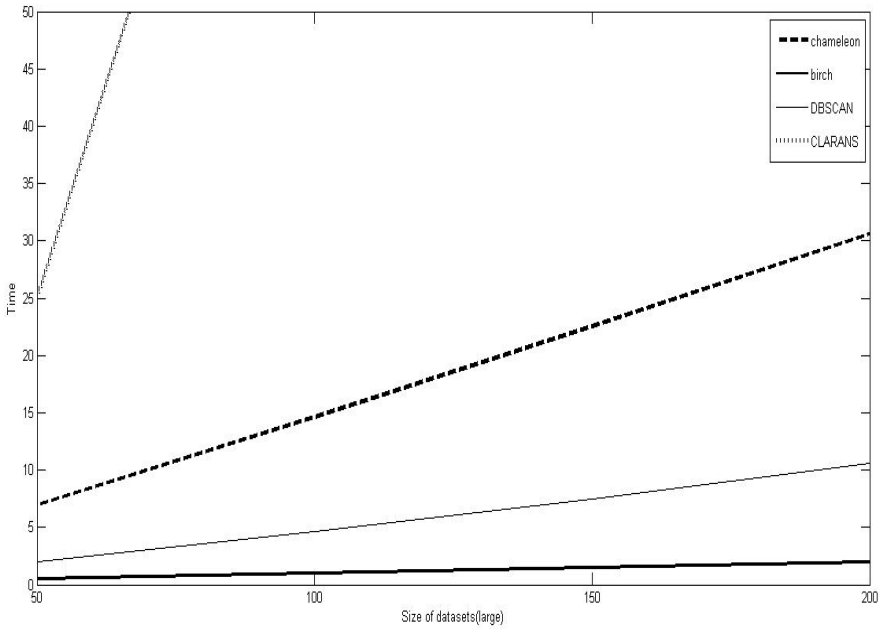

Figure 1. Relationship between size of dataset (small) and performance (time)

Fig.2 shows the relationship between the size of dataset (large) and performance of all four algorithms: Chameleon, BIRCH, DBSCAN, and CLARANS.

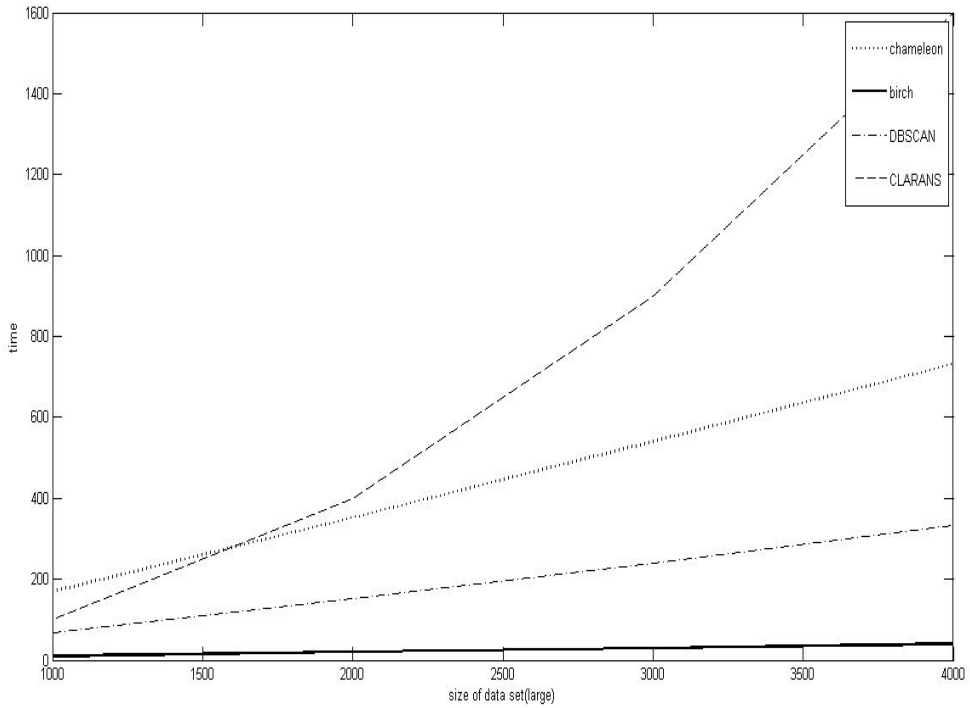

Figure 2. Relationship between size of data set (large) and performance (time)

V. CONCLUSION

After analyzing the result of clustering algorithm using different factors and different situations following results are obtained:

- If clustering performance of BIRCH is compared with other three algorithms then we have found that BIRCH algorithm has best performance in both cases that is, large and small datasets.
- If clustering performance of DBSCAN and Chameleon are compared then we have found that DBSCAN has a superior performance than Chameleon in both cases that is, large and small datasets..
- If clustering performance of CLARANS and Chameleon are compared then we have found that Chameleon has a superior performance than CLARANS in the case of small datasets.
- It has found that clustering performance is worst as compared to other three algorithms in the case of small data sets.
- If clustering performance of Chameleon and CLARANS is compared then we found that initially CLARANS has better performance than Chameleon when the size of data set is less than 1500 but after that Chameleon has superior performance in the case when size of datasets are increased from 1500 .

VI. FUTURE WORK

There are many other clustering algorithms that are worth comparing with each other and useful in various applications. For example it is good to relate the quality performance of multiple clustering algorithms using identical testing datasets. Some algorithms are never compared with each other because they are created or become popular in different times, or belonging to different research fields.

Some enhancement is also possible in this work, if these comparisons are applied on any particular application. This may also applied for document clustering and on basis of performance of clustering algorithm any one of the clustering algorithm can be applied for this purpose, which one is best suited for this.

ACKNOWLEDGEMENT

We would like to thanks to Computer Science and Engineering Department of Amity University and United Group of Institutions, Greater Noida for continuous support and encouragement.

REFERENCES

[1] Saket Agarwal And Leena Singh “Study and analysis of clustering algorithm for association rule mining “ in IJCSSA publications ISSN 0973-7448.
[2] Ravikumar kondadadi,Robert Kozma, “A Modified fuzzy art for soft document clustering” , 0-7803-7278-6,2002,IEEE.
[3] Jie Chen “Comprision of clustering algorithms and its application to Document clustering”,The international Arab journal of Information Technology,5(3),2008.
[4] R.agarwal ,T. Imielinski, A. Swami, “Mining Assocition Rules between sets of items in large Databases”.Proceedings of ACM SIGMOD International conference on management of data,May 1993,pp 207-216.
[5] Gail A.Carpenter,Stephen Grossperg, Natalya Markuzon,John H. Reynolds and David B. Rosen, “Fuzzy ARTMAP: A Neural network Architecture for Incremental Learning of Analog Multidimentional maps” ,1045-9227/92 1992 IEEE.
[6] Nikhil R. Pal ,Kuhu Pal,James M.Keller, and James C. Bezdek,”A Problistic Fuzzy c means clustering Algorithm” ,IEEE Transaction on Fuzzy System,13(4),2005.
[7] Wang Xuping,NI Zijian ,CAO Haiyan,”Research on Assocition Rules Mining Based on ontology in E-Commerce”. 14244-1312-5 2007 IEEE.
[8] Data mining Book “Han and Kamber Data Mining concepts” Definition of clustering algorithms.
[9] www.wikipedia.com
[10] www.google.com