International Journal of Advanced Computer Research (ISSN (print):2249-7277 ISSN (online):2277-7970)
Volume-3 Number-2 Issue-10 June-2013

Review of various Techniques in Clustering

Megha Gupta[1], Vishal Shrivastava[2]

Abstract

This paper presents the review of various techniques which are used for clustering in data mining.Clustering is the process of dividing data into group of similar objects.Clustering involves various techniques so that the data can be partitioned into groups of same data like k-means algorithm, k-medoid, BIRCH algorithm, chameleon, CLIQUE algorithm, DBSCAN algorithm. In this paper, the overview of these entire algorithms has been discussed.

Keywords

Clustering, Hierarchical clustering, Data Mining.

1. Introduction

The amount of data being generated and stored is growing exponentially, due in large part to the continuing advances in computer technology.This presents tremendous opportunities for those who can unlock the information embedded within this data [10].Such collection of data is referred to as data mining.

Data mining is a new technology, developing with database and artificial intelligence. It is a processing procedure of extracting credible, novel, effective and understandable patterns from database. Cluster analysis is an important data mining technique used to find data segmentation and pattern information. By clustering the data, people can obtain the data distribution, observe the character of each cluster, and make further study on particular clusters [7]. Data mining has its roots in machine learning, artificial intelligence, computer Science and statistics.

There are varieties of different data mining techniques and approaches such as clustering, classification, association rule mining. Data mining is an exploratory process, but can be used for confirmatory investigation.

Megha Gupta, M.Tech scholar, CS, RTU, Jaipur, Rajasthan, India.

Vishal Shrivastava, Associate Professor (CS), RTU, Jaipur, Rajasthan, India.

It is different from other searching and analysis techniques, where other analyses are typically problem-driven and confirmatory.

Cluster analysis or clustering is the process of partitioning a set of data objects into subsets. The task of clustering is the task of organizing data into group such that data objects similar to each other are put in same cluster. Clustering is the tool that analysis, which solves classification problems.

Some of the algorithms have been chosen for investigate, study and analyze. The algorithms that are chosen are: K-means algorithm, k-medoid, BIRCH algorithm, chameleon algorithm, CLIQUE algorithm.

2. Literature Survey

Hierarchical clustering is the process of cluster analysis which seeks to build the hierarchy of clusters. Strategies of clustering fall into two divisions:

1. Agglomerative: This is a bottom-up approach. Each observation starts in its own cluster and pairs of clusters are merged as one moves up to the hierarchy.
2. Divisive: This is a top down approach. All observations starts in its own clusters and splits are performed recursively as one move down the hierarchy [4].

The various methods can be used for clustering in which different algorithms can be used. Clustering is the process of organizing data into group such that data objects which are similar to each other are kept in same cluster. Some of the methods that are involved in clustering are as: Partitioning method, Hierarchical methods etc.

In these methods different algorithms are used. In partitioning methods, the algorithms like k-means algorithm, k-medoid algorithm, and chameleon algorithm are being described. Further, in hierarchical clustering, the algorithms like clique algorithm, BIRCH algorithm are being described.

K-means Algorithm:K-meansalgorithm is a well-known algorithm for partitioning methods. The k-

means algorithm works only for datasets that consist of numerical attributes. It takes number of desired clusters, take data points as input and produce k-clusters as output [2].

Following are the details that characterize the algorithms as follows:

a. arbitrarily chose k objects from D (data set) as the initial cluster centers
b. repeat
c. (re)assign each object to the cluster to the object to which the object is the most similar
d. based on the mean value of the objects in the cluster
e. update the cluster means, that is, calculate the mean value of the objects for each cluster
f. until no change

Advantages

1. If variables are huge, then k-means most of times computationally faster if the value of k is small.
2. K-means produce tighter clusters if the clusters are globular.

Disadvantages

1. Difficult to predict k-value.
2. With global clusters, it does not work well.
3. It does not work with clusters of different size and different density.

K-medoid Algorithm: K-medoid algorithm is also a part of partitioning algorithm.The partitioning aroundmedoid (PAM) is a popular realization of k-medoid clustering. It tackles the problem in an iterative, greedy way.

Following are the details that characterize the algorithms as follows:

a. arbitrarily chose k objects in D (data set) as the initial representative objects or seeds.
b. repeat
c. assign each remaining object to the cluster with the nearest representative object;
d. randomly select a no representative object, $o_{random;}$
e. compute total cost, S, of swapping representative object, o_j, with o_{random};
f. if S<0 then swap $o_j$with o_{random} to form the new set of k representative objects;
g. until no change;

Advantages

It is better than k-means algorithm as it tackles the problem in iterative,greedy way. It divides the algorithm in multiple phases so that clustering is done in multiple phases and it is quite easy to understand.

Disadvantages

1. Its performance is less effectively over very large databases and did not consider the case where a dataset was too large to fit in main memory.

BIRCH Algorithm: Balanced Iterative Reducing and Clustering using Hierarchies is designed for clustering a large amount of numeric data by integrating hierarchical clustering and other clustering methods such as iterative partitioning. It operates in two phase. During the first phase, it forms a hierarchical clustering obtained at the first end phase is reasonable sufficient. However, BIRCH recommends a second phase for further refinement in which the leaf cluster obtained at end are reclusters.

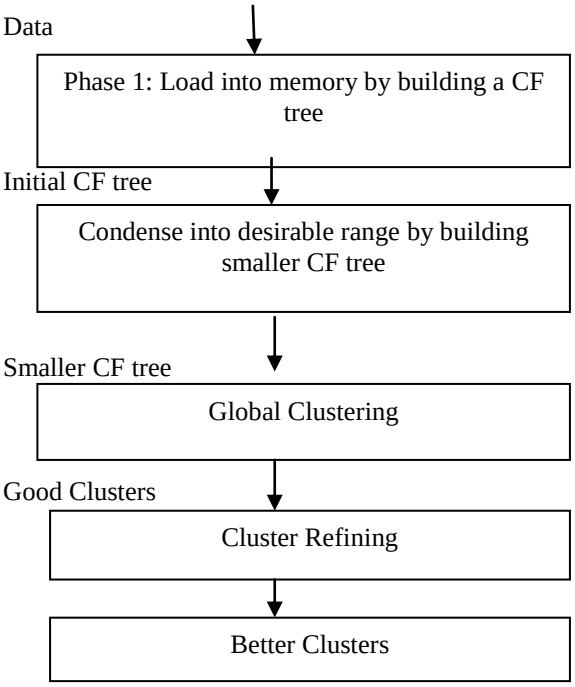

Figure 1: Basic phases involved in BIRCH algorithm

The above figure represents the basic phases that are involved in BIRCH algorithm [3]. It consists of four phases: 1. loading 2.Optional Condensing 3. Global Clustering 4. Optional Refining. In first phase, all

data is scanned and build an initial in-memory-CF-tree using the given amount of memory and recycling space on disk. This CF-tree tries to reflect clustering information of data set to the memory space.After phase 1 subsequent phase will be fast because:

a. no I/O operations are needed
b. the problem of clustering the original data is reduced to smaller problem of clustering the subclusters in the leaf entries.

It is further accurate because:
a. outliers can be eliminated.
b. The remaining data is described at the finest granularity that can be achieved in the available memory.

The BIRCH algorithm is very scalable with respect to the number of records in a datasets. The complexity of phase 1 is linear with respect to the dataset size. Following are the advantages of using BIRCH algorithm:

1. Embedded flexibility regarding a level of granularity.
2. Ease of handling of any forms of similarity or distance.
3. Consequently applicability to any attributes types.
4. Hierarchical clustering algorithms are more versatile.

Advantages
1. It is made without scanning all data points and currently existing clusters.
2. It makes full use of available memory to derive the finest possible sub-clusters while minimizing I/O costs.

Chameleon Algorithm: Chameleon is a hierarchical algorithm that uses dynamic modeling to determine the similarity between pairs of clusters. Cluster similarity is based on

a. How well connected objects are within a cluster.
b. The proximity of clusters.

That is, two clusters are merged if their interconnectivity is high and they are close together. This does not depend on a static, user-supplied model and can automatically adapt to the internal characteristics of the clusters being merged. The merge process facilitates the discovery of natural and homogeneous clusters and applies to all data types as long as a similarity function can be specified.

Advantages
1. It uses agglomerative method to merge the clusters.
2. It uses graph partitioning method to divide data set into set of individual clusters.

Disadvantages
1. Different domains require different measures for different connectivity and closeness.

CLIQUE Algorithm: CLIQUE is abbreviated as clustering in Quest. It has been designed to handle datasets with a large number of dimensions. A high level pseudo code for the algorithm is as follows [1]:
1. for each dimension d in D (data set).
2. Partition d into equal intervals
3. Identify dense units
4. K=2
5. while (1):
6. for each combination of k dimension d1, d2…dk.
7. For each intersection I of dense units along the k dimension
8. If I is dense
9. Marks I as dense unit
10. If no units marked as dense, break from while (1) loop.

Advantages
1. It can automatically find subspaces of the highest dimensionality such that high-density cluster exist in those subspaces.

DBSCAN Algorithm: It is a density based clustering algorithmbecause it finds the number of clusters starting from estimated density distribution of corresponding nodes. It is one of the most common clustering algorithms. The complexity of this algorithm is O(n. log n).

Advantages
1. It does not require one to specify number of clusters in the data as a priori, as opposed to k-means.
2. It just only requires two parameters.
3. It has a notion of noise.

Disadvantages
1. It cannot cluster data sets well with large differences in densities.
2. The quality of DBSCAN depends on the distance used in the function regionQuery.

3. Analysis

In the algorithms above described, the BIRCH algorithm is considered the best one as it is easy to handle any forms of similarity or distance in between clusters. Further, it is applicable to any attributes types. Moreover, hierarchical clustering is more versatile than any other method that is used for clustering.

4. Conclusion & Future Scope

In this paper, we have discussed various methods for data clustering. We have tried to describe the various algorithms that are involved in clustering. We have analyzed the algorithms as K-means, k-medoid, chameleon, BIRCH, CLIQUE and DBSCAN algorithm. In future, more algorithms can be used to describe the partitioning methods and hierarchical methods with more advantages and disadvantages along with their working in details.

References

[1] Ameet Shah, PritishTijare," Cluster Computing", International Journals of Computer and Organization Trends, Volume-2, Issue-I, 2012.
[2] Jiawei, Han, and Micheline Kamber. "Data mining: concepts and techniques." San Francisco, CA, itd: Morgan Kaufmann 5 (2001).
[3] Zhang, Tian, Raghu Ramakrishnan, and Miron Livny. "BIRCH: A new data clustering algorithm and its applications." Data Mining and Knowledge Discovery 1, no. 2 (1997): 141-182.
[4] Bhupendra Kumar Pandya, Umesh Kumar Singh, Keerti Dixit, Kamal Bunkar," International Journal of Engineering Trends and Technology- Sep to Oct Issue 2011.
[5] Basheer aShaikh, G.Vijayadeep,"K best cluster based neighbor classification using improved DDG approaches", International Journals of Engineering Trends and Technology- Volume -3 Issue-5,2012.
[6] Zheng, Yujie. "Clustering Methods in Data Mining with its Applications in High Education." International Proceedings of Computer Science & Information Technology 43 (2012).
[7] Ali, Raza, Usman Ghani, and Aasim Saeed. "Data clustering and its applications." Available at the web address: http://members. tripod. com/asim_saeed/paper. htm (1998).
[8] Jain, Anoop Kumar, and Satyam Maheswari. "Survey of recent clustering techniques in data mining." Int. J. Comput. Sci. Manage. Res 1 (2012): 72-78.
[9] Gupta, Er Arpit, Er Ankit Gupta, and Er Amit Mishra. "Research Paper On Cluster Techniques Of Data Variations." International Journal of Advance Technology & Engineering Research 1, no. 1 (2011): 39-47.
[10] Gary M.Weiss, Brian D. Davisson, "Data Mining", Handbook of Technology Management, John Wiley and Sons, 2010.

Megha Gupta has received B.Tech (CS) degree in 2009 and presently pursuing M.Tech in Computer Science from Rajasthan Technical University Kota. Her area of interest is Data Mining and Clustering.

Vishal Shrivastava is professor in Arya College of Engineering and Information Technology Jaipur. He has been guide for many M.Tech research Scholars. His Research Area is Networking and Data mining.