International Journal of Advanced Scientific Technologies ,Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1,March.2017

Comparative Analysis of Clustering Techniques in Data mining

Dr.E.Kesavulu Reddy
Assistant Professor
Department of Computer Science
S.V.U.College of CM&CS
Tirupati
Email;ekreddysvu2008@gmai.com

Abstract::Data mining refers to extracting useful information from vast amounts of data. It is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Data mining has two types of tasks: Predictive and the descriptive. : The Clustering technique is used to place data elements into related groups without advance knowledge of the group description. The clustering technique belongs to an unsupervised learning and it is used to discover a new set of categories. There are different types of clustering algorithms such as hierarchical, partitioning, grid, density based, model based, and constraint based algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centre based clustering; the value of k-mean is set. Density based clusters are defined as area of higher density then the remaining of the data set. Grid based clustering is the fastest processing time that typically depends on the size of the grid instead of the data.. This paper represents the performance of three clustering algorithms such as Hierarchical clustering, Density based clustering and K Means clustering algorithm.

Keywords: Data mining, Clustering, Hierarchical clustering, Density based clustering, K Means clustering.

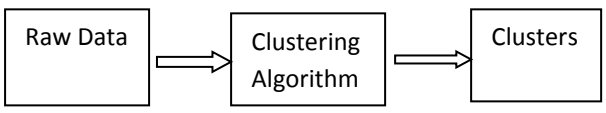
I. INTRODUCTION

Data mining is very computationally intensive process involving very large data sets. Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. Cluster Analysis, an automatic process to find similar objects from a database. It is a fundamental operation in data mining. Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction. A good clustering algorithm is able to identity clusters irrespective of their shapes. There are various clustering algorithms such as partition clustering, hierarchical clustering, Density basedclustering, and fuzzy clustering, etc. These clustering algorithms are classified according to the creation ofclusters of objects [1]. Determining good clusters is one of the main goals in clustering algorithms. The clustering isone of the major problems in variety of domains such as the statistical data analysis, medical image processing,data mining and knowledge discovery, bioinformatics and data classification and compression [2]. Classidentification in spatial databases is the main attractive task in clusteringalgorithms [3]. In supervised learning, the hierarchical clustering is one of the most frequently used methods and it is typically more effective in detectingthe true clustering structure of a data set than partitioningalgorithms.

II. CLUSTERING

Clustering technique is the group of a collection of patterns into clusters based on similarity. The patterns with in valid clusters are more parallel to each other than they are to a pattern belonging to a different cluster. In

clustering, the crisis is to group a given collection of unlabeled patterns into meaningful clusters Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures. Clustering analyses the data objects without consulting a known class label. This is because class labels are not known in the first place, and clustering is used to find those labels.. The superiority also depends on the algorithm's ability to find out some or all of the hidden patterns [2]. The different ways in which clustering methods can be compared are partitioning criteria, separation of clusters, similarity measures and clustering space [3].



III. CLUSTERING ALGORITHMS

Clustering technique is used for finding hiddenpatterns in datamining. The different clustering techniques are stated as follows:
A.Hierarchical clustering
B, Partitioning clustering
C. Density-based clustering
D. Grid-based clustering

A. Hierarchical clustering

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. It is the connectivity based clustering algorithms. The hierarchical algorithms build clusters gradually. Hierarchical clustering generally fall into two types:

❖ *Agglomerative Hierarchical clustering*

It is also known as AGNES. An agglomerative cluster starts with singleton clusters and recursively merges two or more most appropriate clusters. The algorithm forms clusters in a bottom-up manner, as follows [5]:
a) Initially, put each article in its own cluster.
b) Among all current clusters, pick the two clusters with the smallest distance.
c) Replace these two clusters with a new cluster, formed by the smallest distance.
d) Repeat the above two steps until there is only one remaining cluster in the pool.

❖ *Divisive Hierarchical clustering*

This is a "top down" approach. All observations start inone cluster and splits are performed recursively as onemoves down the hierarchy It is also known as DIANA. A divisive cluster starts with one cluster of all data points and recursively splits the most appropriate cluster. This process continues until a stopping criterion (usually the number of requested clusters k) is reached [3]. It uses a top-down strategy.

The algorithm for divisive clustering is as follows [5]:

- Put all objects in one cluster.
- Repeat until all clusters are singletons

i) Choose a cluster to split

ii) Replace the chosen cluster with the sub-cluster.

Advantages of hierarchical clustering

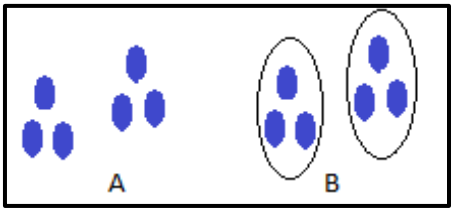
1. Embedded flexibility regarding the level of granularity.
2. Ease of handling ny forms of similarity or distance.
3. Applicability to any attributes type.

Disadvantages of hierarchical clustering

1. Vagueness of termination criteria.
2. Most hierarchal algorithm do not revisit once constructed clusters with the purpose of improvement.

B. *Partitioning clustering*

Partitioning algorithms divide data into several subsets. The reason of dividing the data into several subsets is that checking all possible subset systems is computationally not feasible; there are certain greedy heuristics schemes are used in the form of iterative optimization. Specifically, this means different relocation schemes that iteratively reassign points between the k clusters. Relocation algorithms gradually improve clusters.



There are many methods of partitioning clustering

- K Means Method,
- Medoids Method,
- PAM (Partitioning Around Medoids),
- CLARA (Clustering Larger Applications)

❖ *K- Means*

The k-means algorithm is the most popular clustering tool that is used in scientific and industrial applications. It is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. The basic algorithm is very simple

1. Select K points as initial centroids.
2. Repeat.
3. Form K clusters by assigning each point to its closest centroid.
4. Re compute the centroid of each cluster until centroid does not change.

Advantages of k-means

1. It is efficient in processing large data sets.
2. It often terminates at a local optimum.
3. It works only on numeric values.
4. The clusters have convex shapes.

Disadvantages of k-means

1. Usually terminates at the local optimum, and not the global optimum.
2. Can only be used when the mean is defined and hence requires specifying k, the number of clusters, in advance.

❖ *K-Medoids*

In this algorithm we use the actual object to represent the cluster, using one representative object per cluster. Clusters are generated by points which are close to respective methods. The partitioning is done based on minimizing the sum of the dissimilarities between each object and its cluster representative [3].

❖ *PAM*

PAM works in an iterative, greedy way. The initial repress entative objects are chosen randomly, and it is considered whether replacing the representative objects by non-representative objects would improve the quality of clustering. This replacing of representative objects with other objects continues until the quality cannot be improved further. PAM searches for the best k-Medoids among a given data set.

Algorithm

a) Arbitrarily choose k objects in D as the initial representative objects or seeds.

b) Repeat

i) Assign each remaining object to the cluster with the nearest representative object
ii) Randomly select a non-representative object, orandom
iii) Compute the total cost, S, of swapping representative object oj with orandom
iv) If S<0 then swap oj with o random to form the new set of k representative objects.

c) Until no change

❖ *CLARA*

CLARA uses 5 samples, each with 40+2k points, each of which are then subjected to PAM, which computes the best Medoids from the sample [5]. A large sample usually works well when all the objects have equal probability of getting selected. CLARA cannot find a good clustering if any of the best sampled Medoids is far from the best k-Medoids.

➢ *Divisive clustering*

It is also known as DIANA. A divisive cluster starts with one cluster of all data points and recursively splits the most appropriate cluster. This process continues until a stopping criterion (usually the number of requested clusters k) is reached [3]. It uses a top-down strategy. The algorithm for divisive clustering is as follows [5]

a) Put all objects in one cluster.

b) Repeat until all clusters are singletons
i) Choose a cluster to split
ii) Replace the chosen cluster with the sub-cluster.

3.2.4.2. Density Based Clustering

Density based algorithms find the cluster according to the regions which grow with high density. They are one-scan algorithms. There are two major approaches for density-based methods. The first approach called the density-based connectivity clustering pins density to a training data point. Representative algorithms include DBSCAN, GDBSCAN, OPTICS, and DBCLASD. The second approach pins density to a point in the attribute space and is called Density Functions. It includes the algorithm DENCLUE.

➢ DBSCAN (Density-Based Spatial Clustering of Application with Noise)

This algorithm is based on the user defined parameters, and on the same database with different parameters, it can create multiple clusters. The number of clusters is not required initially, because it produces the clusters only on the density basis. The data points in DBSCAN fall into three categories: (i) Core points i.e. points that are at the interior of a cluster, (ii) Boundary points i.e. non-core points inside a boundary and (iii) Outliers i.e. points that are neither core nor boundary points [6]. DBSCAN cannot handle clusters of different densities.The basic idea of DBSCAN algorithm is that a neighborhood around a point of a given radius must contain at least minimum number of points.

C. .DENCLUE (Density-based Clustering)

Denclueis a clustering method that depends upon density distribution function. DENCLUE uses a gradient hill-climbing technique for finding local maxima of density functions [6]. These local maxima are called density attractors, and only those local maxima whose kernel density approximation is greater than the noise threshold are considered in the cluster. [3].

D. Grid Based Clustering

Grid-based clustering method maps all the objects in a cluster into a number of square cells, known as grids. Grid based clustering has a fast processing time that typically depends on the size of the grid instead of the data. Grid Density based algorithms [10] require the users to specify a grid size or the density threshold.The grid-based clustering algorithms are STING, Wave Cluster, and CLIQUE.

➢ STING (StatisticalInformation Grid approach)

This approach breaks the available space of objects into cells of rectangular shapes in a hierarchy. It follows the top down approach and the hierarchy of the cells can contain multiple levels corresponding to

multiple resolutions [4]. The statistical information like the mean, maximum and minimum values of the attributes is precomputed and stored as statistical parameters, and is used for query processing and other data analysis tasks. The statistical parameters for higher level cells can be computed from the parameters for lower level cells.

- CLIQUE

Cliqueis a grid-based method that finds density-based clusters in subspaces. CLIQUE performs clustering in two steps. In the first step, CLIQUE partitions each dimension into non-overlapping rectangular units, thereby partitioning the entire space of data objects into cells. At the same time it identifies the dense cells in all the subspaces. When the fraction of total data points contained in the unit exceeds the input model parameter then a unit is dense. In the second step, CLIQUE uses these dense cells to form clusters, which can be arbitrary.

IV. COMPARATIVE ANALYSIS OF CLUSTERING TECHNIQUES

In Hierarchical clustering methods discover to the cluster with arbitrary shape. Partitioning the data sets into several sub-clusters using partitioning algorithms and constructs a fuzzy graph of sub-clusters [6]. DBCURE [7] is used to find clusters with varying densities and suitable for parallelizing the algorithms with map reduce. DBCUREMR [8] finds several clusters together in parallel. It finds clusters efficiently without being sensitive to the clusters with varying densities and scales up with the frame work. Density based k-Mediods clustering algorithm to overcome the drawbacks of DBSCAN and K-MEDIODS clustering algorithms to handle the clusters circularly distributed points slightly overlapped clutters.

V. CONCLUSION

We presentthe overview of the algorithms used in different clustering techniques along with their respective advantages and disadvantages. The different clustering methods are focused i.e. partitioning clustering, hierarchical clustering, density based clustering and grid based clustering. Under partitioning method, a brief description of k-means and k-Medoids algorithms have been studied. In hierarchical clustering, the BIRCH and CHAMELEON algorithms have been described. The DBSCAN [9] and DENCLUE algorithms under the density based methods have been studied. Finally, under grid-based clustering method the STING and CLIQUE algorithms have been described. The comparisons with clustering analysis are mainly that different clustering techniques give substantially different results on the same data.

REFERENCES

[1] 1.Peter Scherer et al., ―Using SVM and Clustering Algorithms in IDSSystems‖ 2011, pp108–119, ISBN 978-80-248-2391-1.
[2] 2. Ming-chuan hung et al., ―An Efficient k-Means ClusteringAlgorithm using
[3] Simple Partitioning‖ journal of information scienceand engineering 21, 1157-
[4] 1177 (2005).
[5] 3. Martin Ester et al., ―A Density-Based Algorithm for DiscoveringClusters in
[6] Large Spatial Databases with Noise
[7] 4. Younghoon Kim, et al., ―DBCURE-MR: An efficient density-basedclustering
[8] Algorithmfor large data using MapReduce‖, InformationSystems 42 (2014)
[9] 15–35.
[10] 5. Christophe ambroise, et al., ―Christophe ambroise, et al.,―Hierarchical
[11] clustering of Self-organizing maps for cloudclassification, Neurocomputing,
[12] Volume 30, Issues 1–4, January2000, Pages 47–52.
[13] 6. Yihong Dong et al., ―A hierarchical clustering algorithm based onfuzzy graph
[14] connectedness‖, Fuzzy Sets and Systems, Volume 157,1 July 2006, Pages 1760–
[15] 1774.
[16] 7. Clark F. Olson., ―Parallel algorithms for hierarchical clustering Parallel
[17] Computing,Volume 21, Issue 8, August 1995, Pages 1313–1325
[18] 8. Younghoon Kim et al., ―DBCURE-MR: An efficient density-basedclustering
[19] algorithm for large data using MapReduce‖, InformationSystems, Volume 42,
[20] June 2014, Pages 15–35.
[21] 9. P. Viswanath et al., Rough-DBSCAN: A fast hybrid based densitybased
[22] clustering method for large datasets‖, Pattern RecognitionLetters 30 (2009)
[23] 1477–1488.
[24] 10. Ashish Ghosh et al.,Aggregation pheromone density based dataclustering
[25] InformationSciences, volume 178, Issue 13, July 2008,Pages 2816-2831.