Analytical study of Clustering Techniques

Payal Pahwa[1], Rashmi Chhabra[2]

[1]*Bhagwan Parshuram Institute of Technology, I.P University, Delhi*

[2]*Research Scholar, CSE Department, NIMS University, Jaipur (Rajasthan)*

Abstract

Fast retrieval of the relevant information from the databases has always been a significant issue. Different techniques have been developed for this purpose, one of them is Data Clustering. Clustering is a collection of data objects that are similar to one another within the same cluster and dissimilar to the objects in the other clusters. In this paper Data Clustering is discussed along with its approaches. This paper discusses about clustering, clustering techniques and comparisons of different clustering techniques.

Keyword – Clustering, portioning, Density based, Grid based, Hierarchical

1. Introduction

Data mining is a process of analyzing data from different perspectives and summarizing it into useful information. It is a process that allows users to understand the substance of relationships between data. It reveals patterns and trends that are hidden among the data [1]. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining systems can be classified according to the kinds of databases and mined. Three important components of data mining systems are databases, data mining engine, and pattern evaluation modules. Data mining engine ideally consists of a set of functional modules for tasks such as characterization, association, classification, cluster analysis and evolution. Clustering is one of the first steps in data mining analysis. In our paper we discuss clustering and different clustering methods.

2. Data Clustering

Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects. Cluster analysis is finding similarities between data according to the characteristics found in the data and grouping similar objects into clusters. The criterion for checking the similarity is implementation dependent. It is the basic step in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. Cluster analysis which is to group the data points into clusters is an important task of data mining recently. Unlike classification which analyzes the labeled data, cluster analysis deals with data points without consulting a known label previously. It is a method of unsupervised learning and a common technique for statistical data analysis used in many fields including data mining, pattern recognition, information retrieval and many others. It is based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity.

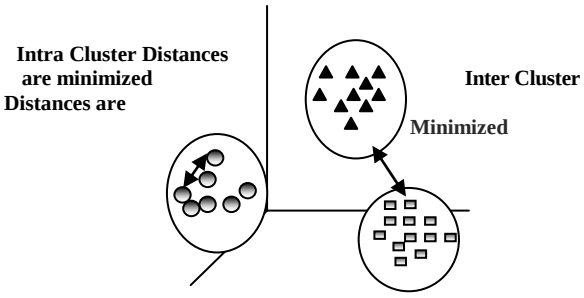

Figure 1 Grouping of different objects

Clustering is challenging field in which its potential applications pose their own requirements. Typical

requirement of clustering in data mining are explained below.

Table 1.Clustering Requirement

Clustering Requirement	
Scalability	Highly scalable clustering algorithms are needed.
Ability to deal with different types of attributes	Algorithm works on different data types.
clusters with arbitrary shape	Cluster could be of any shape
dealing noise and outliers	Algorithms must be able to work on poor quality data
Handle dynamic data	Must be able to handle changeable data
High dimensionality	Algorithm must work on several attributes
Insensitive to order of input records	Input does not affect the Output .

3. Clustering Techniques

There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of dataset. In general, clustering algorithms can be classified into four Categories: partitioning-based, hierarchical-based, density based, and grid-based methods.

3.1. Partitioning Method

The partitioning methods generally result in a set of 'k' clusters, each object belonging to one cluster. The method classifies the data into k groups with condition that each group must contain at least one object and each object must belong to exactly one group. Each cluster may be represented by a centroid or a cluster representative. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset. There are basically two partitioning algorithm one is k-mean[2] and another is k-medoids[3]. In k-.mean method the mean of the cluster is used as the representative object per cluster whereas in k-medoids method instead of taking the mean value of the objects in a cluster as a reference point we can pick actual object to represent the cluster. Both k-means and k-medoid algorithms represent a cluster using a single point and the user provide the parameter, k- the number of clusters, and perform iterative membership relocation until the membership is no longer changed or the change is within a tolerable range. The method always find cluster of spherical shape. The basic algorithm of this method is [4]:

1. Make the first object the centroid for the first cluster.
2. For the next object, calculate the similarity, S, with each existing cluster centroid, using some similarity coefficient.
3. If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re determine the centroid; otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

3.2. Hierarchical Method

A hierarchical clustering method works by grouping data objects into a tree like structure. Hierarchical-based clustering algorithms use a hierarchical tree to represent the closeness of the data objects. The tree is constructed in either bottom-up or top-down. The bottom-up approach starts with each object forming a cluster and recursively merges the clusters based on their closeness measure. This method is known as agglomerative method (AGNES) . On the other hand, the top-down approach starts with all the objects in a single cluster and recursively splits the objects into smaller groups. This method named as divisive method(DIANA). The representative hierarchical based clustering algorithms are BIRCH [5], CURE [6],and CHAMELEON [7].The basic algorithm of AGNES and DIANA are:

AGNES

1. Start with 1 point
2. Recursively add two or more appropriate clusters
3. Stop when k number of clusters is achieved

DIANA

1. Start with a big cluster
2. Recursively divide into smaller clusters
3. Stop when k number of clusters is achieved

3.3. Density Based Method

Density-based clustering algorithms consider clusters as dense regions of objects in the data space and clusters are separated by regions of low density. The main idea of density-based approach is to find regions of high-density

and low density, with high-density regions being separated from low-density regions .These algorithms associate each object with a density value defined by

the number of its neighbor objects within a given radius. An object whose density is greater than a specified threshold is defined as a dense object and initially is formed a cluster itself. Two clusters are merged if they share a common neighbor that is also dense. The representative density-based clustering algorithms are DBSCAN [10], OPTICS [8], HOP [9], and DENCLUE [11]. These methods can separate the noise (outliers) and find arbitrary shape clusters.

3.4. Grid-based methods

Grid-based clustering algorithms first cover the problem space domain with a uniform grid mesh[15]. Statistical attributes are collected for all the data objects located in each individual mesh cell and clustering is, then, performed on the grid, instead of data objects themselves. These algorithms typically have a fast processing time, since they go through the data set once to compute the statistical values for the grids and the performance of clustering depends only on the size of the grids which is usually much less than the data objects. The representative grid-based clustering algorithms are STING [12], WaveCluster [14], and CLIQUE [13]. All these methods employ a uniform grid mesh to cover the whole problem. For the problems with highly irregular data distributions, the resolution of the grid mesh must be fine enough to obtain a good clustering quality.

4. Comparison of clustering techniques

As we know there are variety of clustering techniques available for data mining. Here we draw a comparison chart of all the techniques on the basis of input parameters, technique used for clustering, shape of the cluster and type of database on which clustering is applied, complexity and methods of each technique.

Table 2.Comparison Chart

Clustering Methods	Input Parameters	Technique Used	Cluster Shape	Database Size	Complexity	Methods	Limitation
Partitioning Method	'k' No. of clusters Dataset containing 'n' objects	Iterative relocation Technique	Spherical shaped	Small-medium size database	O(n2)	K-Means K-Medoids	Unable to find clusters of complex shape
Hierarchical Methods	--------	Merging or divisive approach	Arbitrary Shaped	Large data set	O(n2)	BIRCH CURE CHAMELEON	Once merge or split is done, it can never be undone
Density Based Method	Cluster Radius, Minimum number of Objects	Density based connectivity analysis	Arbitrary Shaped	Spatial database	O(n2) in DBSCAN	DBSCAN OPTICS HOP DENCLUE	Density is more difficult to define for high dimensional data.
Grid-based methods	Number of objects in a cell	Multi resolution clustering technique	Arbitrary Shaped	Large data sets	O(n) of WAVE CLUSTER where n – number of objects O(K) of STING Where K – Number of Cells at bottom Layer	STING WaveCluster CLIQUE	The quality of STING clustering depends on the granularity of the lowest level of the grid structure.

5. Conclusion

Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval , and bioinformatics. There are different clustering techniques of cluster analysis, but they are broadly classified into hierarchical and nonhierarchical techniques. In the hierarchical procedures, we construct a hierarchy to find the relationship among objects, whereas in the non-hierarchical method a position in the measurement is taken as central place and distance is measured from such central point. In our paper we discuss all these methods and compare them. We concluded that Grid based algorithm has the minimum complexity and can handle large data set efficiently. Also it discovers the clusters of arbitrary shapes.

REFERENCES

[1] I. K. Ravichandra Rao, Data Mining and Clustering Techniques,DRTC Workshop on Semantic Web , Bangalore, December, 2003.

[2] J. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In the 5th Berleley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297, 1967.

[3] L. Kaufman and P. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons Inc.,New York, 1990.

[4] Raza Ali , Usman Ghani ,Aasim Saeed ,Data Clustering and Its Applications

[5] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an Efficient Data Clustering Method for Very Large Databases.In International Conference Management of Data (SIGMOD'96), pages 103–114, Jun. 1996.

International Conference Management of Data (SIGMOD'99), pages 49–60, Jun. 1999

[6] S. Guha, R. Rastogi, and K. Shim. CURE: An Efficient Clustering Algorithm for Large Databases", Databases. In International Conference Management of Data (SIGMOD'98), pages 73–84, Jun. 1998.

[7] G. Karypis, E. Han, and V. Kumar. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling.Computer, 32(8):68–75, 1999.

[8] M. Ankerst, M. Breunig, H.P. Kriegel, and J. Sander. Optics: Ordering Points to Identify the Clustering Structure. In International Conference Management of Data (SIGMOD'99), pages 49–60, Jun. 1999

[9] D. Eisenstein and P. Hut. Hop: A New Group Finding Algorithm for N-body Simulations. Astrophysics Journal, 498(1):137–142, 1998.

[10] M. Ester, H. Kriegel, J. Sander, , and X. Xu. A Density based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pages 226–231, 1996.

[11] A. Hinneburg and D. Keim. An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In International Conference on Knowledge Discovery and Data Mining (KDD'98), pages 58–65, Aug. 1998.

[12] W. Wang, J. Yang, and R. R. Muntz. STING: A Statistical Information Grid Approach to Spatial Data Mining. In the 23rd International Conference on Very Large Data Bases (VLDB'97), pages 186–195, 1997

[13] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan.Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. In International Conference Management of Data (SIGMOD'98), pages 94–105, Jun. 1998.

[14] G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster:A Multi-Resolution Clustering Approach for Very Large Spatial Databases. In the 24th International Conference on Very Large Data Bases (VLDB'98), pages 428–439, Aug. 1998.

[15] Wei-keng LiaoA, Grid-based Clustering Algorithm using Adaptive Mesh Refinement