CHAPTER-21

A categorization of Major clustering Methods

21 A Categorization of Major Clustering Methods

21.1 Introduction

There exit a large number of clustering algorithms in the literature .The choice of clustering algorithm depends both on the type of data available and on the particular purpose and application. If cluster analysis is used as a descriptive or exploratory tool,it is possible to try several algorithms on the same data to see what the data may disclose.

In general, major clustering methods can be classified into the following categories.

21.2 Partitioning methods:

Given a database of n objects or data tuples,a partition in method constructs k partitions of the data, where each partition represents cluster and K<=n. That is ,it classifies the data into k groups, which together satisfy the following requirements:

(1) each group must contain at least on e object,and

(2) each object must belong to exactly one group-Notice that the second requirement can be relaxed in some fuzzy partitioning technique.

Given K, the number of partitions to construct , a partitining method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another .The general criterion of a good partitioning is that objects in the same clusters are "close" or related to each other,whereas objects of different clusters are "far apart"or very different. there are various kinds of other criteria for judging the quality of partitions.

To achieve global optimality in partitioning-based clustering would require the exhaustive enumeration of all of the possible partitions. Instead, most applications adopt one of two popular heuristic methods;

1. the k-means algorithm,where each cluster is represented by the mean value of the objects in the cluster,and

2. the k-medoids algorithm,where each cluster is represented by one of the objects located near the center of the cluster.These heuristic clustering methods work well for finding spherical-shaped clusters in small to medium -sized databases.To find clusters with complex shapes and for clustering very large data sets, partitioning-based methods need to be extended.Partitioning-based clustering methods are studied in depth later.

21.3 Hierarchical methods:

A hierarchical method creates a hierarchical decomposition of the given set of data objects,A hierarchical method can be classified as being either agglomerative or divisive ,based on how the hierarchical decomposition is formed. The agglomerative approach,also called the bottom -up aproach ,starts with each object forming a separate group, It successively merges the objects or groups close to one another, until all of the groups are merged into one(the topmost level of the hierarchy), or until a trmination condition holds. The divisive approach, also called the top-down approach, starts with all the objects in the same cluster,until eventually each object is in one cluster, or until a termination condition holds,

Hierarchical methods suffer form the fact that once a step(merge o9r split) is done,it can never be undone. This rigidity is useful in that it leads to smaller computation costs by not worrying about a combinatorial number of different choices.However, a major problem of such techniques is that they cannot correct erroneous decisions.There are two approaches to improving the quality of hierarchical partitioning, such as in CURE and Chameleon, or (2) integate hierarchical agglomeration and iterative relocation by first using a hierarchical agglomerative algorithm and then refining the result using iterative relocation by first using a hierarchical aggomerative algorithm and then refining the result using iterative relocation , as in BIRCH.

21.4 Density- based methods:

most partitioning methods cluster objects based on the distance between objects.Such methods can find only spherical-shaped clusters and encounter

difficulty at discovering clusters of arbitrary shapes. Other clustering methods have been developed based on the notion of density.Their general idea is to continue growing the given cluster as long as the density. Their general idea is to continue growing the given cluster as long as the density(number of objects or data points)in the "neighborhood"

exceeds some threshold; that is , for each data point within a given cluster,the neighborhood of a given radius has to contain at least a minimum number of points .Such a method can be used to filter out noise(outliers) and discover clusters of arbitrary shape.

DBSCAN is a typical density-based method that grows clusters according to a density threshold,OPTICS is a density-based method that computes an augmented clustering ordering for automatic and interactive cluster analysis.

Grid -based method:Grid -based methods quantize the object space into a finite number of cells that form a grid structure .All of the clustering operations are performed on the grid structure(i.e., on the quantized space).The main avantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space.

STING is a typical example of a grid-based method.CLIQUE and wave-cluster are two clustering algorithms that are both grid-based and density-based.

model-based methods: Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model. A model- based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model .A model-baed algorithm may locatre clusters by constructing a density function that reflects the spatial distribution of the data points.It also leads to a way of automatical determining the number of clusters based on standard statistics, taking "noise"or outliers into account and thus yielding robust clustering methods .Model-based clustering methods are studied below.

Some clustering algorithms integrate the ideas of several clustering methods,so that it is sometimes difficult to classify a given algorithm as uniquely belonging to only one clusteing method category. Furthermore ,some applications may have clustering creteria that require the integration of seeral clustering techniques.

In the following sections,we examine each of the above five clustering methods in detail. We also introduce algorithms that integrate the ideas of several clustering methods.outlier analysis , which typically involves clustering, is described at the end of this section.

21.5 Partitioning Methods

Given a database of objects and k , the number of clusters to form , a partitioning algorithm organizes the objects into k partitions(k<=n), where each partition represents a cluster.The clusters are formed to optimize an objective-partitioning criterion,often called a similarity function ,such as distance ,so that the objects within a cluste are "similar," whereas the objects of different clusters are "dissimilar"in terms of the database attributes.

21.6 Classical Partitioning Methods: k-means and k-medoids

The most well -known and commonly used partitioning methods are k-means,k-nedoids, and their variations.

Centroid-Based Technique: The K-Means method

The fc-means algorithm takes the input paramete,k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low.cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's center of gravity.

"How does the k-means algorithm work ?" The k-means algorithm proceeds as follows.First, it randomly selects k of the objects, each of which initially represents a cluster mean or center.For each of the remaiining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean.It then computes the new mean for each cluster.This process iterates until the criterion function converges.Typically, the squared-error criterion is used,defined as

E=sigma sigmap=ci|p-mi| square

where E is the sum of square-error for all obects in the database ,p is the point in space representing a given object, and mi, is the mean of cluster ci (both p and mi, are multidimensional).This criterion tries to make the resulting k clusters as compact and as separate as possible.

The algorithm attempts to determine K partitions that minimize the squared-error function. It works when the clusters are compact clouds that are rather well separated from one another.The method is relatively scalable and efficient in processing large data sets because the computational complexity of the algorithm is

O(nkt), where n is the total number of objects,k is the number of clusters , and t is the number of iterations . nNormally,k<<n and t<<n.The method often terminates

at a local optimum.

The k-means method,however,can be applied only when the mean of a cluster is denned-This may not be the case in some applications , such as when data with categorical attributes are involved-the necessity for users to specify k, the number of clusters, in advance can be seen as a desadvantage.the k-means method is not suitable for discovering clusters with nonconvex shapes or clusters of very different size.Moreover,it is sensitive to noise and outlier data points since a small number of such data can substantially influence the mean value.

21.7 Hierarchical Methods

A hierarchical clustering method works by grouping data objects into a tree clusters.Hierarchical clustering methods can be further classified into agglomerative and divisive hierarchical clustering,depending on whether the hierarchy decomposition is formed in a bottom-up or top-down fashion.The quality of a pure hierarchical clustering method suffers from its inalbility to perform adjustment once a merge or split decision has been executed.Tecent studies have emphasized the integration of hierarchical agglomeration with iterative relocation methods.

Agglomerative and Divisive Hierarchical Clustering

In general, there are two types of herearchical clustering methods:

Agglomerative hierarchical clustering: This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic dusters into larger and larger clusters,,until all of the objects are in a single cluster or until certain termination conditions are satisfied .Most herearchical clustering methods belong to this category.They differ only in their definition of inter cluster similarity.

Divisive heerarchical clustering: This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster.It subdivides the cluster into smaller and smaller pieces,until each object forms a cluster on its own or until it satisfies certain termination conditions,such as a desired number of clusters is obtained or the distance between the two closest clusters is above a certain threshold distance.

Four widely used measures for distance between clusters are as follows,where |p-p'| is the distance between two objects or points p and p',m,is the mean for cluster C, and n, is the number of objects of in Ci.

minimum distance :

maximium distance:

mean distance:

Average distance:

21.8 Review questions

1 Explian about Partitioning methods:

2 Explian About Hierarchical methods:

3 Explian about Density- based methods:

4 Explain about Partitioning Methods

5 Expalian about   Classical Partitioning Methods: k-means and k-medoids

21.9 References.

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition