UNIT-IV

CLUSTERING AND APPLICATIONS

Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.

What is Clustering?

Clustering is the process of making a group of abstract objects into classes of similar objects.

Points to Remember

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

Requirements of Clustering in Data Mining

The following points throw light on why clustering is required in data mining −

- **Scalability** − We need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kinds of attributes** − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- **Discovery of clusters with attribute shape** − The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- **High dimensionality** − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- **Ability to deal with noisy data** − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- **Interpretability** − The clustering results should be interpretable, comprehensible, and usable.

Types Of Data Used In Cluster Analysis - Data Mining

Types of data structures in cluster analysis are

- **Data Matrix** (or object by variable structure)
- **Dissimilarity Matrix** (or object by object structure)

Data Matrix

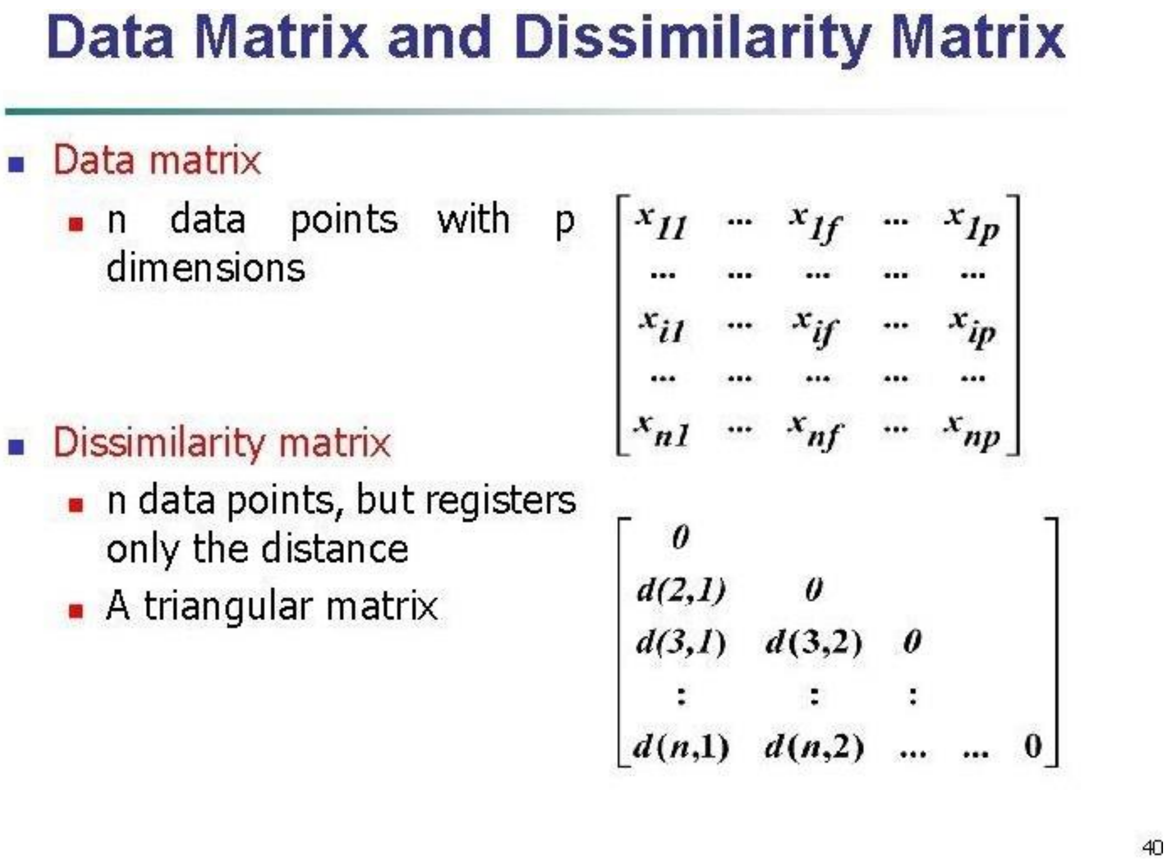
This represents n objects, such as persons, with p variables (also called measurements or attributes), such as age, height, weight, gender, race and so on. The structure is in the form of a relational table, or n-by-p matrix (n objects x p variables)

The Data Matrix is often called a two-mode matrix since the rows and columns of this represent the different entities.

Dissimilarity Matrix

This stores a collection of proximities that are available for all pairs of n objects. It is often represented by a n – by – n table, where d(i,j) is the measured difference or dissimilarity between objects i and j. In general, d(i,j) is a non-negative number that is close to 0 when objects i and j are higher similar or "near" each other and becomes larger the more they differ. Since d(i,j) = d(j,i) and d(i,i) =0, we have the matrix in figure.

This is also called as one mode matrix since the rows and columns of this represent the same entity.



Data Matrix and Dissimilarity Matrix

- Data matrix
  - n data points with p dimensions

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix
  - n data points, but registers only the distance
  - A triangular matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

40

Types Of Data In Cluster Analysis Are:

Interval-Scaled Variables

Interval-scaled variables are continuous measurements of a roughly linear scale.

Typical examples include weight and height, latitude and longitude coordinates (e.g., when clustering houses), and weather temperature

Binary Variables

A binary variable is a variable that can take only 2 values.

For example, generally, gender variables can take 2 variables male and female.

Contingency Table For Binary Data

Let us consider binary values 0 and 1
Let p=a+b+c+d

Simple matching coefficient (invariant, if the binary variable is symmetric):
Jaccard coefficient (noninvariant if the binary variable is asymmetric):

Nominal or Categorical Variables

A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green.

Method 1: Simple matching

The dissimilarity between two objects i and j can be computed based on

the simple matching.

m: Let m be no of matches (i.e., the number of variables for which i and j are in the same state).

p: Let p be total no of variables.

Method 2: use a large number of binary variables

Creating a new binary variable for each of the M nominal states.

ordinal Variables

An ordinal variable can be discrete or continuous.

In this order is important, e.g., rank.

It can be treated like interval-scaled

By replacing xif by their rank,

By mapping the range of each variable onto [0, 1] by replacing the i-th object in the f-th variable by,

Then compute the dissimilarity using methods for interval-scaled variables.

Ratio-Scaled Intervals

Ratio-scaled variable: It is a positive measurement on a nonlinear scale, approximately at an exponential scale, such as Ae^Bt or A^e-Bt

Major Clustering Methods

Clustering methods can be classified into the following categories –

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Partitioning Method

Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements −

- Each group contains at least one object.
- Each object must belong to exactly one group.

Points to remember −

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here −

- Agglomerative Approach
- Divisive Approach

Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

Divisive Approach

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

Approaches to Improve Quality of Hierarchical Clustering

Here are the two approaches that are used to improve the quality of hierarchical clustering −

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantages

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

Two classes of clustering tasks

Constraint-Based Method:

The constraint-based clustering method is performed by the incorporation of application or user-oriented constraints. A constraint refers to the user expectation or the properties of the desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. The user or the application requirement can specify constraints.

Clustering high Dimensionality data:

It is a particularly important task in clustering analysis because many applications require the analysis of objects containing a large number of features or dimensions clustering high dimensional data is challenging issue due to curse of dimensional.Many dimensions may not be relevant. As, the number of dimensions increases,the data become increasingly sparse.

PARTIONING METHODS

Given a database of n objects or data tuples, a partitioning method constructs k partitions of the data, where each partition represents a cluster and k <= n.

That is, it classifies the data into k groups, which together satisfy the following requirements

Each group must contain at least one object,

Each object must belong to exactly one group.

Given k, the number of partitions to construct, a partitioning method creates an initial partition.

It uses 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 cluster 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.

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \left| x - m_i \right|^2$$

Here, E is the sum of the square error for all objects in the data set.

x is the point in space representing a given object, and mi is the mean of cluster Ci (both x and mi are multidimensional). In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed.
This criterion tries to make the resulting k clusters as compact and as separate as possible.

Suppose that there is a set of objects located in space as depicted in the rectangle.

Let k =3; i.e. the user would like to cluster the object into three clusters.

According to the algorithm, we arbitrarily choose three objects as the three initial cluster centers, were cluster centers are marked by a “+”.

Each object is distributed to a cluster based on the cluster center to which it is the nearest.

Such distribution forms circled by dotted curves.

Advantages Of K-Means

Relatively efficient: O(tkn), where n is # objects, k is # clusters, and t is # iterations. Normally, k, t << n.

Each object is distributed to a cluster based on the cluster center to which it is the nearest.

Disadvantages Of K-Means

- Applicable only when mean is defined, then what about categorical data
- Need to specify k, the number of clusters, in advance.
- Unable to handle noisy data and outliers.
- Not suitable to discover clusters with non-convex shapes.

K-Medoids Clustering

A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the given dataset.

K-Medoids: Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster.

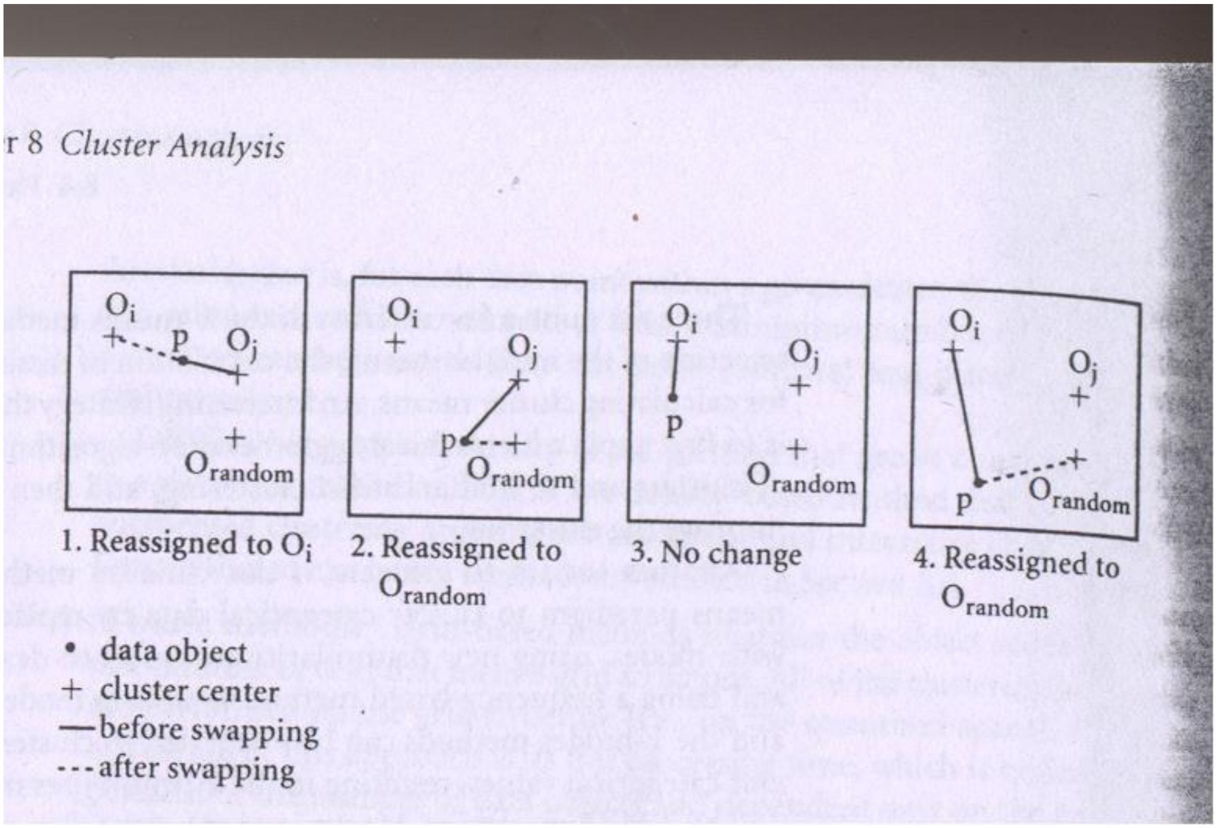
The basic strategy of k-medoids clustering algorithms is to find k clusters in n objects by first arbitrarily finding a representative object (the medoid) for each cluster.

Each remaining object is clustered with the medoid to which it is most similar.

The strategy then iteratively replaces one of the medoids by one of the non-medoids as long as the quality of the resulting clustering is improved.

This quality is estimated by using a cost function that measures the average dissimilarity between an object and the medoid of its cluster.

To determine whether a non-medoid object is "Oi" random is a good replacement for a current medoid "Oj", the following four cases are examined for each of the non-medoid objects "P".



r 8 *Cluster Analysis*

1. Reassigned to O_i 2. Reassigned to O_{random} 3. No change 4. Reassigned to O_{random}

Case 1: "P" currently belongs to medoid "Oj", If "Oj" is replaced by "Orandom", as a medoid and "P" is closest to one of "Oi", it do not belong "j", then "P" is assigned to "Oi".

Case 2: "P" currently belongs to medoid "Oj". If "Oj" is replaced by "Orandom" as medoid and "P" is closest to "Orandom", then "P" is reassigned to "Orandom".

Case 3: "P" currently belongs to medoid "Oi", it does not belong "j". If "Oj" is replaced by "Orandom" as a medoid and "P" is still closest to "Oi", then the assignment does not change.

Case 4: "P" currently belongs to medoid "Oi", it does not belong to "j". If "Oj" is replaced by "Orandom" as a medoid and "P" is closest to "Orandom", then "P" is reassigned to "Orandom".

Which Is More Robust -- K-Means or K-Medoids

The k-medoids method is more robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean.

However, its processing is more costly than the k-means method. Both methods require the user to specify k, the number of clusters.

Aside from using the mean or the medoid as a measure of cluster center, other alternative measures are also commonly used in partitioning clustering methods.

The median can be used, resulting in the k-median method, where the median or "middle value" is taken for each ordered attribute. Alternatively, in the k-modes method, the most frequent value for each attribute is used.

Hierarchical Clustering

A hierarchical clustering method works by grouping data objects into a tree of clusters.

Hierarchical clustering methods can be further classified into agglomerative and divisive hierarchical clustering, depending on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion.

(AGNES)Agglomerative Hierarchical Clustering:

This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters until all of the objects are in a single cluster or until certain termination conditions are satisfied.

(DIANA)Divisive Hierarchical Clustering:

This top-down strategy does the reverse of agglomerative 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.

Data set of five objects a, b, c, d. Initially, AGNES places each object into a cluster of its own.

The clusters are then merged step-by-step according to some criterion.

For example, clusters C1 and C2 may be merged if an object in C1 and an object in C2 form the minimum Euclidean distance between any two objects from different clusters.

This is a single-linkage approach in that each cluster is represented by all of the objects in the cluster, and the similarity between two clusters is measured by the similarity of the closest pair of data points belonging to different clusters.

The cluster merging process repeats until all of the objects are eventually merged to form one cluster.

In DIANA, all of the objects are used to form one initial cluster.

The cluster is split according to some principle, such as the maximum Euclidean distance between the closest neighboring objects in the cluster.

The cluster splitting process repeats until, eventually, each new cluster contains only a single object.

In either agglomerative or divisive hierarchical clustering, the user can specify the desired number of clusters as a termination condition.

A tree structure called a dendrogram is commonly used to represent the process of hierarchical clustering.

Decompose data objects into several levels of nested partitioning (tree of clusters), called a dendrogram.

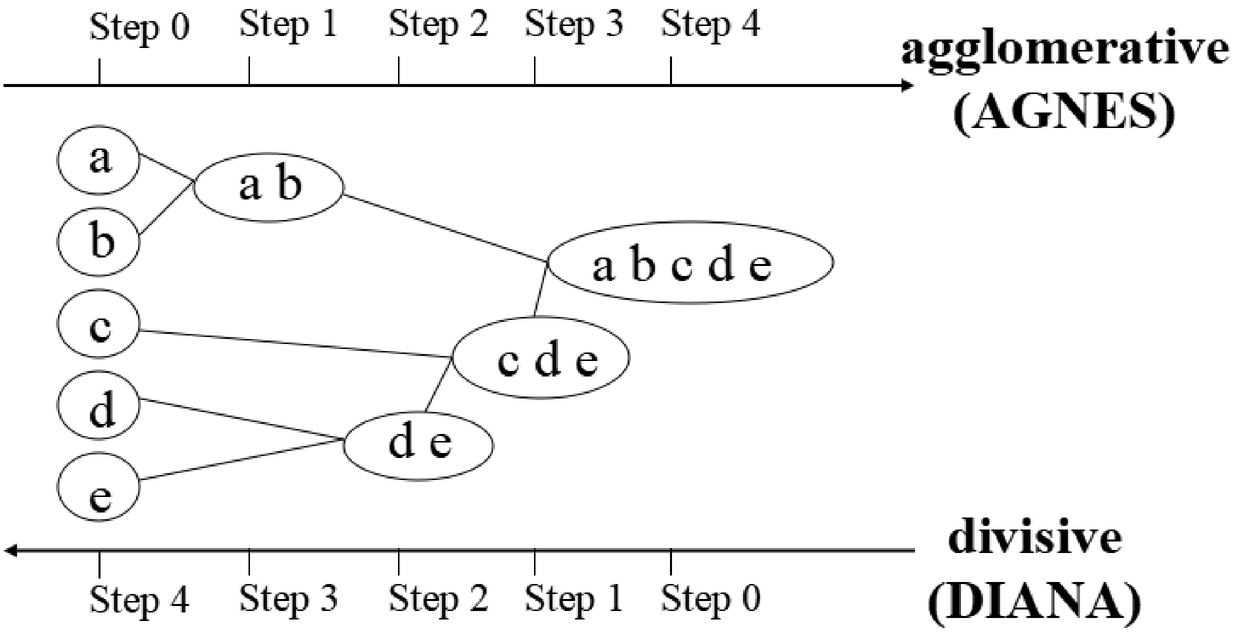
A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.

It shows how objects are grouped together step by step.

The figure shows a dendrogram for the five objects presented in previous Fig, where l = 0 shows the five objects as singleton clusters at level 0.

At l = 1, objects a and b are grouped together to form the first cluster, and they stay together at all subsequent levels.

We can also use a vertical axis to show the similarity scale between clusters. For example, when the similarity of two groups of objects, {a, b} and {c, d, e} is roughly 0.16, they are merged together to form a single cluster.



Disadvantages Of Hierarchical Clustering

The hierarchical clustering method, though simple, often encounters difficulties regarding the selection of merge or split points.

Such a decision is critical because once a group of objects is merged or split, the process at the next step will operate on the newly generated clusters.

It will neither undo what was done previously nor perform object swapping between clusters.

Thus merge or split decisions, if not well chosen at some step, may lead to low-quality clusters.

Moreover, the method does not scale well, because each decision to merge or split requires the examination and evaluate a good number of objects or clusters.

More On Hierarchical Methods Are

BIRCH (1996): Uses CF-tree and incrementally adjusts the quality of sub-clusters.

CURE (1998): Selects well-scattered points from the cluster and then shrinks them towards the center of the cluster by a specified fraction.

Density-Based Clustering method is one of the clustering methods based on density (local cluster criterion), such as density-connected points.

The basic ideas of density-based clustering involve a number of new definitions. We intuitively present these definitions and then follow up with an example.

The neighborhood within a radius ε of a given object is called the ε-neighborhood of the object.

If the ε-neighborhood of an object contains at least a minimum number, MinPts, of objects, then the object is called a core object.

Density-reachable:

A point p is density-reachable from a point q wrt. Eps, MinPts if there is a chain of points p1, …, pn, p1 = q, pn = p such that pi+1 is directly density-reachable from pi

Density-connected

A point p is density-connected to a point q wrt. Eps, MinPts if there is a point o such that both, p and q are density-reachable from o wrt. Eps and MinPts.

Working Of Density-Based Clustering

Given a set of objects, D' we say that an object p is directly density-reachable from object q if p is within the ε-neighborhood of q, and q is a core object.

An object p is density-reachable from object q with respect to ε and MinPts in a set of objects, D' if there is a chain of objects p1,.,.,.pn, where p1 = q and pn = p such that pi+1 is directly density-reachable from pi with respect to e and MinPts, for 1/n, pi € D.

An object p is density-connected to object q with respect to ε and MinPts in a set of objects, D', if there is an object o, belongs D such that both p and q are density-reachable from o with respect to ε and MinPts.

Density-Based Clustering - Background

Two parameters:

Eps: Maximum radius of the neighborhood.

MinPts: Minimum number of points in an Eps-neighbourhood of that point.

NEps(p): {q belongs to D | dist(p,q) <= Eps}
Directly density-reachable: A point p is directly density-reachable from a point q wrt. Eps, MinPts if p belongs to NEps(q)

core point condition:|NEps (q)| >= MinPts

Major features:

- It is used to discover clusters of arbitrary shape.
- It is also used to handle noise in the data clusters.
- It is a one scan method.
- It needs density parameters as a termination condition.

Density-Based Methods

DBSCAN: Ester, et al. (KDD'96)

OPTICS: Ankerst, et al (SIGMOD'99).

DENCLUE: Hinneburg & D. Keim (KDD'98)

CLIQUE: Agrawal, et al. (SIGMOD'98)

DBSCAN(Density-Based Spatial Clustering of Applications with Noise)

It relies on a density-based notion of cluster: A cluster is defined as a maximal set of density-connected points.

It discovers clusters of arbitrary shape in spatial databases with noise.

DBSCAN Algorithm

Arbitrary select a point p.

Retrieve all points density-reachable from p wrt Eps and MinPts.

If p is a core point, a cluster is formed.

If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.

Continue the process until all of the points have been processed.

say, let MinPts = 3.

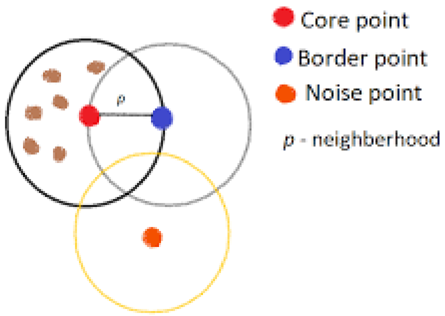
Of the labeled points, m, p, o, and r are core objects because each is in an ε-neighborhood containing at least three points.

q is directly density-reachable from m. m is directly density-reachable from p and vice versa.

q is (indirectly) density-reachable from p because q is directly density-reachable from m and m is directly density-reachable from p.

However, p is not density-reachable from q because q is not a core object.

Similarly, r and s are density-reachable from o, and o is density-reachable from o, and o is density-reachable from R.



OPTICS - A Cluster-Ordering Method

OPTICS: Ordering Points To Identify the Clustering Structure.

It produces a special order of the database with respect to its density-based clustering structure.

This cluster-ordering contains info equivalent to the density-based clusterings corresponding to a broad range of parameter settings.

It is good for both automatic and interactive cluster analysis, including finding an intrinsic clustering structure.

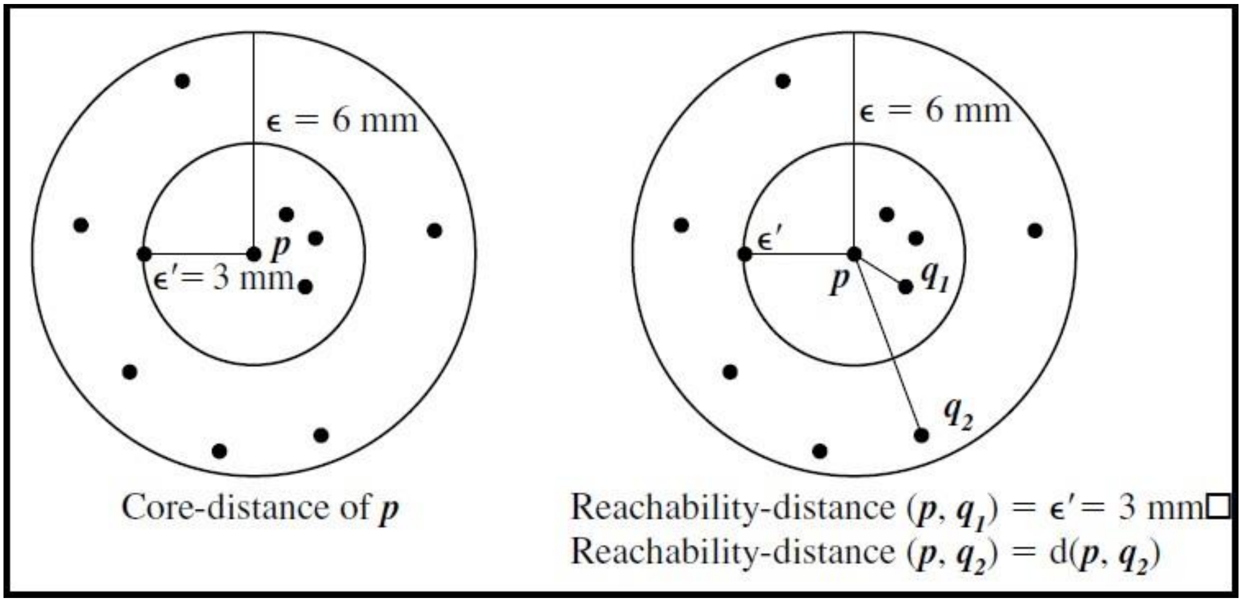
It can be represented graphically or using visualization techniques.

Core-distance and reachability-distance: The figure illustrates the concepts of core-distance and reachability-distance.

Suppose that e=6 mm and MinPts=5.
The core distance of p is the distance, e0, between p and the fourth closest data object.

The reachability-distance of q1 with respect to p is the core-distance of p (i.e., e0 =3 mm) because this is greater than the Euclidean distance from p to q1.

The reachability distance of q2 with respect to p is the Euclidean distance from p to q2 because this is greater than the core-distance of p.


ε = 6 mm
p
ε′= 3 mm
Core-distance of p
ε = 6 mm
ε′
p
q1
q2
Reachability-distance (p, q1) = ε′= 3 mm□
Reachability-distance (p, q2) = d(p, q2)

DENCLUE - Using Density Functions

DENsity-based CLUstEring by Hinneburg & Keim (KDD’98)

Major Features

It ha got a solid mathematical foundation.

It is definitely good for data sets with large amounts of noise.

It allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets.

It is significantly faster than the existing algorithm (faster than DBSCAN by a factor of up to 45).

But it needs a large number of parameters.

DENCLUE - Technical Essence

It uses grid cells but only keeps information about grid cells that do actually contain data points and manages these cells in a tree-based access structure.

Influence function: This describes the impact of a data point within its neighborhood.

The Overall density of the data space can be calculated as the sum of the influence function of all data points.

The Clusters can be determined mathematically by identifying density attractors.

The Density attractors are local maxima of the overall density function.

Summary

Density-Based Clustering -> It is one of the clustering methods based on density (local cluster criterion), such as density-connected points.

Grid-Based Clustering

Grid-Based Clustering method uses a multi-resolution grid data structure.

STING (a STatistical INformation Grid approach) by Wang, Yang, and Muntz (1997)

WaveCluster by Sheikholeslami, Chatterjee, and Zhang (VLDB'98) - A multi-resolution clustering approach using wavelet method

CLIQUE - Agrawal, et al. (SIGMOD'98)

STING - A Statistical Information Grid Approach

STING was proposed by Wang, Yang, and Muntz (VLDB'97).

In this method, the spatial area is divided into rectangular cells.

There are several levels of cells corresponding to different levels of resolution.

For each cell, the high level is partitioned into several smaller cells in the next lower level.

The statistical info of each cell is calculated and stored beforehand and is used to answer queries.

The parameters of higher-level cells can be easily calculated from parameters of lower-level cell

Count, mean, s, min, max

Type of distribution—normal, uniform, etc.

Then using a top-down approach we need to answer spatial data queries.

Then start from a pre-selected layer—typically with a small number of cells.

For each cell in the current level compute the confidence interval.

Now remove the irrelevant cells from further consideration.

When finishing examining the current layer, proceed to the next lower level.

Repeat this process until the bottom layer is reached.

Advantages:

It is Query-independent, easy to parallelize, incremental update.

O(K), where K is the number of grid cells at the lowest level.

Disadvantages:

All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected.

WaveCluster

It was proposed by Sheikholeslami, Chatterjee, and Zhang (VLDB'98).

It is a multi-resolution clustering approach which applies wavelet transform to the feature space

A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub-band.

It can be both grid-based and density-based method.

Input parameters:

No of grid cells for each dimension

The wavelet, and the no of applications of wavelet transform.

How to apply the wavelet transform to find clusters

It summaries the data by imposing a multidimensional grid structure onto data space.

These multidimensional spatial data objects are represented in an n-dimensional feature space.

Now apply wavelet transform on feature space to find the dense regions in the feature space.

Then apply wavelet transform multiple times which results in clusters at different scales from fine to coarse.

Why is wavelet transformation useful for clustering

It uses hat-shape filters to emphasize region where points cluster, but simultaneously to suppress weaker information in their boundary.

It is an effective removal method for outliers.

It is of Multi-resolution method.

It is cost-efficiency.

Major features:

The time complexity of this method is O(N).

It detects arbitrary shaped clusters at different scales.

It is not sensitive to noise, not sensitive to input order.

It only applicable to low dimensional data.

CLIQUE - Clustering In QUEst

It was proposed by Agrawal, Gehrke, Gunopulos, Raghavan (SIGMOD'98).

It is based on automatically identifying the subspaces of high dimensional data space that allow better clustering than original space.

CLIQUE can be considered as both density-based and grid-based:

It partitions each dimension into the same number of equal-length intervals.

It partitions an m-dimensional data space into non-overlapping rectangular units.

A unit is dense if the fraction of the total data points contained in the unit exceeds the input model parameter.

A cluster is a maximal set of connected dense units within a subspace.

Partition the data space and find the number of points that lie inside each cell of the partition.

Identify the subspaces that contain clusters using the Apriori principle.

Identify clusters:

Determine dense units in all subspaces of interests.

Determine connected dense units in all subspaces of interests.

Generate minimal description for the clusters:

Determine maximal regions that cover a cluster of connected dense units for each cluster.

Determination of minimal cover for each cluster.

Advantages

It automatically finds subspaces of the highest dimensionality such that high-density clusters exist in those subspaces.

It is insensitive to the order of records in input and does not presume some canonical data distribution.

It scales linearly with the size of input and has good scalability as the number of dimensions in the data increases.

Disadvantages

The accuracy of the clustering result may be degraded at the expense of the simplicity of the method.

Summary

Grid-Based Clustering -> It is one of the methods of cluster analysis which uses a multi-resolution grid data structure.

OUTLIER ANALYSIS

The set of objects are considerably dissimilar from the remainder of the data ○ Example: Sports: Michael Jordon, Wayne Gretzky, ...

Problem: Define and find outliers in large data sets

Applications:

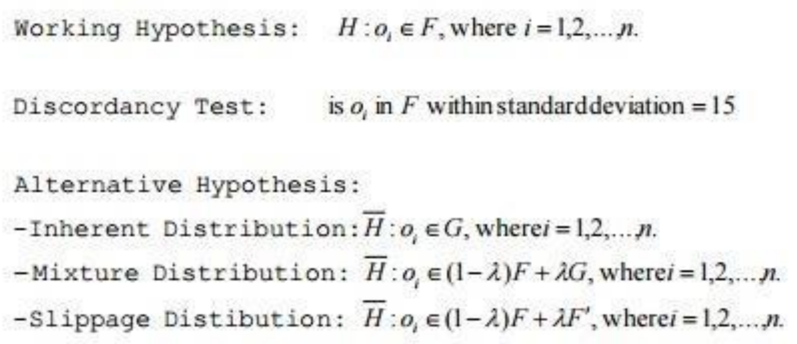
- Credit card fraud detection
- Telecom fraud detection
- Customer segmentation
- Medical analysis

Statistical Distribution-based outlier detection- Identify the outlier with respect to the model using discordancy test

How discordancy test work

Data is assumed to be part of a working hypothesis (working hypothesis)-H

Each data object in the dataset is compared to the working hypothesis and is either accepted in the working hypothesis or rejected as discordant into an alternative hypothesis (outliers)- H



Working Hypothesis: $H: o_i \in F$, where $i = 1,2,\ldots,n.$

Discordancy Test: is o_i in F within standard deviation $= 15$

Alternative Hypothesis:

-Inherent Distribution: $\overline{H}: o_i \in G$, where $i = 1,2,\ldots,n.$

-Mixture Distribution: $\overline{H}: o_i \in (1-\lambda)F + \lambda G$, where $i = 1,2,\ldots,n.$

-Slippage Distibution: $\overline{H}: o_i \in (1-\lambda)F + \lambda F'$, where $i = 1,2,\ldots,n.$

Distance-Based outlier detection

Imposed by statistical methods

We need multi-dimensional analysis without knowing data distribution Algorithms for mining distance-based outliers

Index-based algorithm

Indexing Structures such as R-tree (R+-tree), K-D (K-D-B) tree are built for the multi-dimensional database

The index is used to search for neighbors of each object O within radius D around that object.

Once K (K = N(1-p)) neighbors of object O are found, O is not an outlier.

Worst-case computation complexity is O(K*n^2), K is the dimensionality and n is the number of objects in the dataset.

·Pros: scale well with K

·Cons: the index construction process may cost much time

·Nested-loop algorithm

Divides the buffer space into two halves (first and second arrays)

Break data into blocks and then feed two blocks into the arrays.

Directly computes the distance between each pair of objects, inside the array or between arrays

Decide the outlier.

Divide the dataset into cells with length

·K is the dimensionality, D is the distance

·Define Layer-1 neighbors – all the intermediate neighbor cells. The maximum distance between a cell and its neighbor cells is D

·Define Layer-2 neighbors – the cells within 3 cell of a certain cell. The minimum distance between a cell and the cells outside of Layer-2 neighbors is D

Criteria

- Search a cell internally. If there are M objects inside, all the objects in this cell are not outlier
- Search its layer-1 neighbors. If there are M objects inside a cell and its layer-1 neighbors, all the objects in this cell are not outlier

· •Search its layer-2 neighbors. If there are less than M objects inside a cell, its layer-1 neighbor cells, and its layer-2 neighbor cells, all the objects in this cell are outlier

- Otherwise, the objects in this cell could be outlier, and then need to calculate the distance between the objects in this cell and the objects in the cells in the layer-2 neighbor cells to see whether the total points within D distance is more than M or not.

Density-Based Local Outlier Detection

Distance-based outlier detection is based on global distance distribution

It encounters difficulties to identify outliers if data is not uniformly distributed

Ex. C_1 contains 400 loosely distributed points, C_2 has 100 tightly condensed points, 2 outlier points o_1, o_2

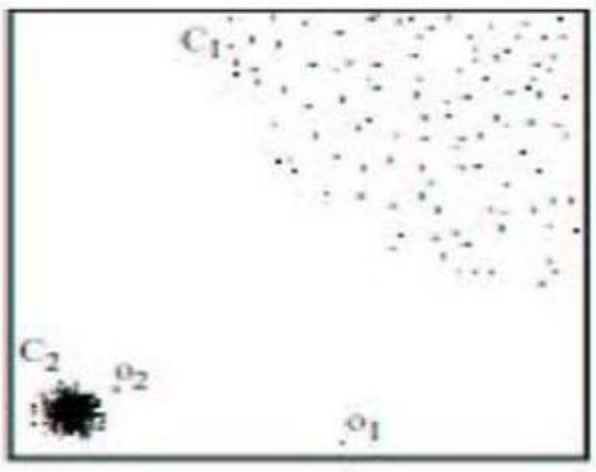
Some outliers can be defined as global outliers, some can be defined as local outliers to a given cluster

O_2 would not normally be considered an outlier with regular distance-based outlier detection, since it looks at the global picture

Each data object is assigned a *local outlier factor (LOF)*

Objects which are closer to dense clusters receive a higher LOF

·LOF varies according to the parameter MinPts



Deviation-Based Outlier detection

Identifies outliers by examining the main characteristics of objects in a group

Objects that ―deviate‖ from this description are considered outliers

Sequential exception technique

simulates the way in which humans can distinguish unusual objects from among a series of supposedly like objects

Dissimilarities are assed between subsets in the sequence the techniques int roduce following key term Exception set, dissimilarity function, cardinality function, smoothing factor

OLAP data cube technique

Deviation detection process is overlapped with cube computation

Recomputed measures indicating data exceptions are needed

A cell value is considered an exception if it is significantly different from the expected value, based on a statistical model

Use visual cues such as background color to reflect the degree of exception