



Cluster ing

By

R. Siva Narayana

Index

- Introduction to Clustering
- Applications
- Basic Clustering Methods



Clustering

Clustering is the process of grouping a set of data objects into multiple groups.

Objects within a cluster have high similarity, but very dissimilar objects in the other cluster.

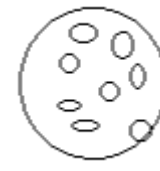
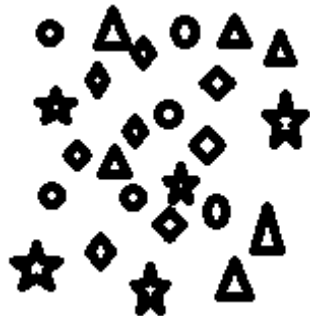
Similarities and Dissimilarities are based on the attribute values describing the objects often involve distance measures.



Cluster Analysis

- Cluster analysis is the process of partitioning a set of data objects into subsets, each subset is a cluster.
- The set of clusters resulting from a cluster analysis can be referred to as a **Clustering**

- Different clustering methods may generate different clusterings on the same data set.



- Clustering also called as
 - **automatic classification**
 - **Data Segmentation**
 - **Outlier Detection**
 - **Unsupervised Learning**

Applications

- Business Intelligence
- Image Pattern Recognition 2 2 2 2 2
- Web Search
- Biology
- Security

Basic Clustering Methods

- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods

Partitioning Method

- Given a set of n objects, we have to construct k partitions of the data, where each partition represents a cluster and $k \leq n$.
- Divides the data into k groups such that each group must contain at least one object.
- Adopts **exclusive cluster separation** and are **distance-based**.
- Given k , the number of partitions to construct, a partitioning 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.

Partitioning Method

- Achieving global optimality is often computationally prohibitive, potentially requiring all possible partitions
- K-means and K-medoids algorithms are greedy approach progressively improves the clustering quality and approaches a local optimum.
- These methods work well for finding **spherical shaped** clusters in a **small to medium size** databases.

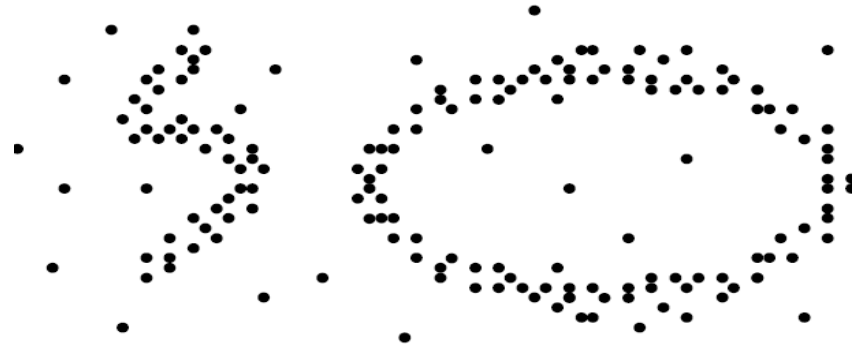
Hierarchical Methods

- A hierarchical method creates a hierarchical decomposition of the given set of data objects.
 - ***agglomerative (Bottom-Up)***
 - ***Divisive (Top-down)***
- The *agglomerative approach*, starts with each object forming a separate group.
- It successively merges the objects or groups close to one another, until all the groups are merged into one (the topmost level of the hierarchy), or a termination condition holds.
- The *divisive approach*, starts with all the objects in the same cluster.
- In each successive iteration, a cluster is split into smaller clusters, until eventually each object is in one cluster, or a termination condition holds.
- Hierarchical clustering methods can be distance-based or density- and continuity based.

Density Based Clustering

- To find clusters of arbitrary shape developed based on the notion of **density**
- General idea is to continue growing a given cluster as long as the density (number of objects or data points) in the “neighborhood” exceeds some threshold
- Ex: for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points.
- Used to filter out noise or outliers and discover clusters of arbitrary shape

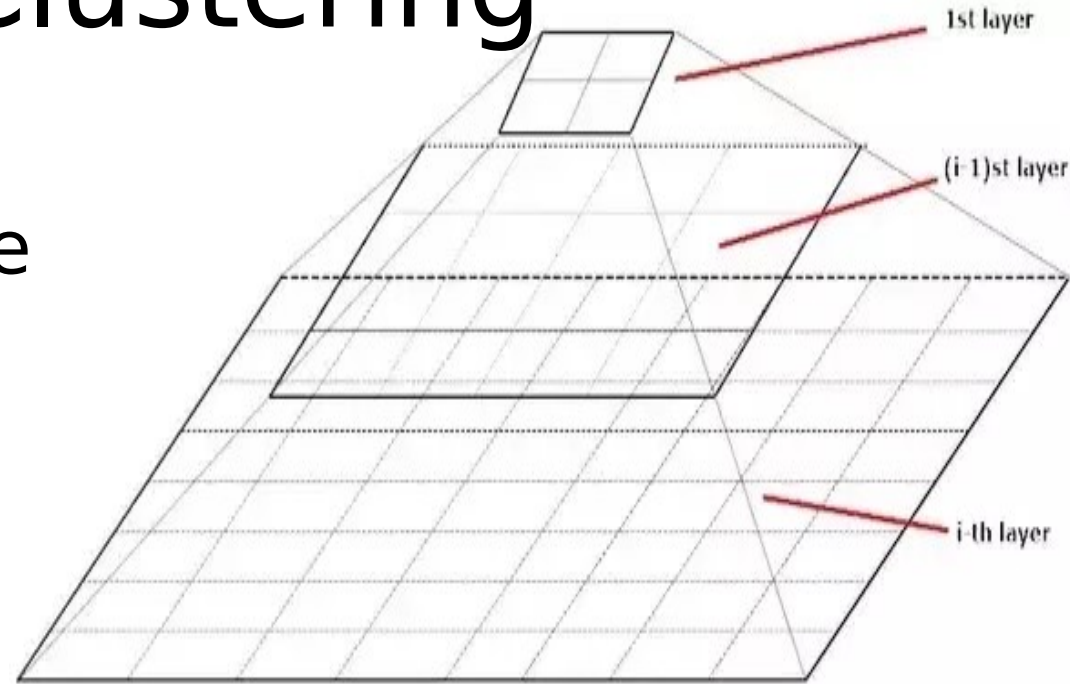
Density Based Clustering



- We can model clusters based on dense regions in the data space, separated by sparse regions. This is the main strategy behind *density-based clustering methods*, which can discover clusters of nonspherical shape.
- Ex: DBSCAN, OPTICS and DENCLUE

Grid-Based Clustering

- Grid-based methods quantize the object space into a finite number of cells that form a grid structure.
- All the clustering operations are performed on the grid structure (i.e., on the quantized space).
- **Advantage:** Fast processing time
- Using grids is often an efficient approach to many spatial data mining problems, including



Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object “linkages”
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its “neighborhood” – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

Thanks