**11.** In conventional cluster analysis, an object is assigned to one cluster exclusively. However, in some applications, there is a need to assign an object to one or more clusters in a fuzzy or probabilistic way. **Fuzzy clustering** and **probabilistic model-based clustering** allow an object to belong to one or more clusters.

A **partition matrix** records the membership degree of objects belonging to clusters.

**Probabilistic model-based clustering** assumes that a cluster is a parameterized distribution.

Using the data to be clustered as the observed samples, we can estimate the parameters of the clusters.

A **mixture model** assumes that a set of observed objects is a mixture of instances from multiple probabilistic clusters. Conceptually, each observed object is generated independently by first choosing a probabilistic cluster according to the probabilities of the clusters, and then choosing a sample according to the probability density function of the chosen cluster.

An **expectation-maximization algorithm** is a framework for approaching maximum likelihood or maximum a posteriori estimates of parameters in statistical models.

Expectation-maximization algorithms can be used to compute fuzzy clustering and probabilistic model-based clustering.

**High-dimensional data** pose several challenges for cluster analysis, including how to model high-dimensional clusters and how to search for such clusters.

There are two major categories of clustering methods for high-dimensional data: subspace clustering methods and dimensionality reduction methods.

**Subspace** **clustering methods** search for clusters in subspaces of the original space. Examples include **subspace search methods**, **correlation-based clustering methods**, and **biclustering methods**. **Dimensionality reduction methods** create a new space of lower dimension and search for clusters there

**Biclustering methods** cluster objects and attributes simultaneously.

Types of biclusters include biclusters with **constant values**, **constant values on rows/columns**, **coherent values**, and **coherent evolutions on rows/columns**.

Two major types of biclustering methods are **optimization-based methods** and **enumeration methods**.

**Spectral clustering** is a **dimensionality reduction method**. The general idea is to construct new dimensions using an affinity matrix.

**Clustering graph and network data** has many applications such as social network analysis. Challenges include how to measure the similarity between objects in a graph, and how to design clustering models and methods for graph and network data.

**Geodesic distance** is the number of edges between two vertices on a graph. It can be used to measure similarity. Alternatively, similarity in graphs, such as social networks, can be measured using structural context and random walk.

**SimRank** is a similarity measure that is based on both structural context and random walk.

Graph clustering can be modeled as computing **graph cuts**. A **sparsest cut** may lead to a good clustering, while **modularity** can be used to measure the clustering quality.

**SCAN** is a graph clustering algorithm that searches graphs to identify well-connected components as clusters.

**Constraints** can be used to express application-specific requirements or background knowledge for cluster analysis. Constraints for clustering can be categorized as constraints on **instances**, on **clusters**, or on **similarity measurement**. Constraints on instances include **must-link** and **cannot-link** constraints. A constraint can be **hard** or **soft**.

**Hard constraints for clustering** can be enforced by strictly respecting the constraints in the cluster assignment process. **Clustering with soft constraints** can be considered an optimization problem. Heuristics can be used to speed up constrained clustering.