Notes 06-1: Introduction to Clustering

Clustering techniques attempt to derive a model of the data such that objects grouped together are most similar to one another, but most dissimilar to objects in other groups.

In the simplest case, each object is associated with one and only one cluster.

DIAGRAM = Clustering.A.1.a

Some approaches may allow an object to belong to more than one cluster.

DIAGRAM = Clustering.A.1.b

Some approaches associate objects with clusters probabilistically so that for every object there is a probability or degree of membership for each cluster.

DIAGRAM = Clustering.A.1.c

More formally, the clustering problem is defined as follows: Given a database $D = \{t_1, t_2, \ldots, t_n\}$ of tuples and an integer k (i.e., the number of clusters to find), the clustering problem is to define a mapping $f : D \rightarrow \{K_1, K_2, \ldots, K_k\}$, where each t_i is assigned to one cluster K_a, $1 \leq a \leq k$. A cluster, K_a, contains precisely those tuples mapped to it; that is, $K_a = \{t_i \mid f(t_i) = K_a\ (1 \leq i \leq n) \wedge t_i \in D\}$.

Two fundamental problems encountered while clustering:

- The (best) number of clusters may not be known in advance (i.e., how many clusters are "actually" described by the data).

- Interpreting the semantic meaning of clustering may be difficult (i.e., it may not be intuitively clear or obvious why a particular item was assigned to a cluster).

The major clustering methods can be classified according to *five* general approaches: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods.

Partitioning methods: Classify objects into *k* groups using an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another until the objects in a cluster are "close" and those in different clusters are "far" according to some quality criteria.

Hierarchical methods: Create a hierarchical decomposition of the objects, where a nested set of clusters is created such that at the highest level, all objects belong to the same cluster, and at the lowest level, each object is in its own cluster.

An *agglomerative approach* (i.e., bottom-up) starts with each object in a separate group, and then merges "close" groups until all the groups are merged into one or until some termination condition is satisfied.

A *divisive approach* (i.e., top-down) starts with all objects in the same group, and then splits groups into smaller groups until each object is in its own cluster or until some termination condition is satisfied.

Density-based methods: Find clusters of arbitrary shapes, where for each data point within a given cluster, the neighborhood of a given radius has to contain at least some given number of points.

Grid-based methods: Divide the object space into a finite number of cells that form a grid structure in, possibly, multiple dimensions.

Model-based methods: Hypothesize a model for each cluster and find the best fit of the data to the model.

Example – General idea behind k-means clustering (a partitioning method).

DIAGRAM = Clustering.A.5

Data Representation

A *data matrix* (a.k.a. an *object-by-variable structure*) represents *n* objects with *p* attributes (a.k.a. measurements or variables). Also called a *two-mode* matrix because the rows and columns represent different entities.

DIAGRAM = Clustering.B.1

A *dissimilarity matrix* (a.k.a. an *object-by-object structure*) represents a collection of proximities that are available for all pairs of *n* objects.

DIAGRAM = Clustering.B.2

Each $d(i, j)$ is the measured difference or dissimilarity between objects i and j, where $d(i, j)$ is a non-negative number, usually close to 0 when i and j are similar and larger when i and j are different.

Also, called a *one-mode* matrix because the rows and columns represent the same entity.

Using a difference or dissimilarity measure, a data matrix can be converted to a dissimilarity matrix.