Clustering in KDD

Javier Béjar

LSI - MIA

AMLT - 2013/2014
Other topics in clustering

- Unsupervised Learning includes other specific topics:
  - Consensus Clustering/Cluster Ensembles
  - Biclustering/Coclustering
  - Subspace Clustering/Projected Clustering/Correlation clustering
Consensus Clustering/Cluster ensembles

- Inspired by the ensembles algorithms from supervised learning
- Same issues than supervised learning:
  - Clustering algorithms are usually unstable
  - Results have large variance for different samples/parameters
  - Local optima and initialization dependence
- Solution: Combine several clusterings in a consensuated solutions
Clustering diversity: Diversity in the solutions impact in the resulting consensus

Consensus information: How to determine from the clusterings the relations among instances
- Coocurrence matrices (association imputation)
- Relative indices

Consensus function: How to combine the consensus information in a unique clustering
- Clustering the coocurrence matrix
- Finding the median partition

Determining the number of clusters
Subspace Clustering/Correlation Clustering

- Clustering algorithms assume that all attributes are relevant for all clusters.
- Problem: Find a set of clusters and a subset of dimensions for each cluster.
- Formulations:
  - Find clusters in axis parallel subspaces (subspace clustering): related to space partitioning clustering, grid clustering and unsupervised feature selection.
  - Find clusters in non axis parallel subspaces (correlation clustering): apply dimensionality reduction techniques to find projections that represent the clusters.
Biclustering/Co-clustering

- Clustering attributes and instances at the same time
- Goal: To reorganize the data matrix to discover rectangular groups
- Applications:
  - BioInformatics (gene expressions)
  - Document clustering (Words-topics-documents)