Clustering

Oliver Brdiczka

PRIMA Research Group

03/05/2004

Overview

- Clustering
  - Explore ”spatial“ structure of data
- Statistical (temporal) Pattern Discovery
  - Explore temporal structure of data

Un-supervised Learning

- Classification
  - Connect data and human interpretation

Supervised Learning

Clustering

- Input: large amounts of high dimensional (sensory) data
- Objective: reduce data in dimension and/or amount
- Results are used for:
  - Visualization
  - Data Compression
  - Representation for statistical pattern discovery and modeling

Clustering types

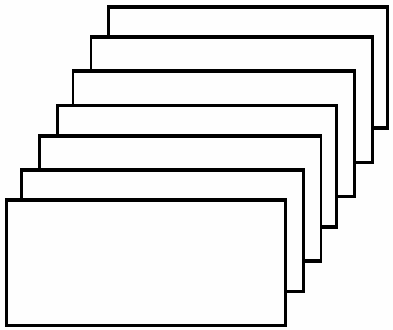
- Projection methods (PCA …)
- Partitioning methods (K-Means...)
- Hierarchical methods
- Density-Based methods
- Grid-based methods

Partitioning vs. Projection

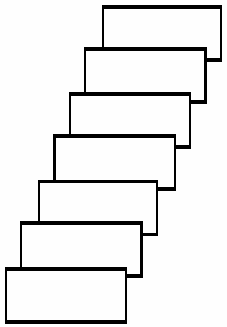
Projection

Partioning

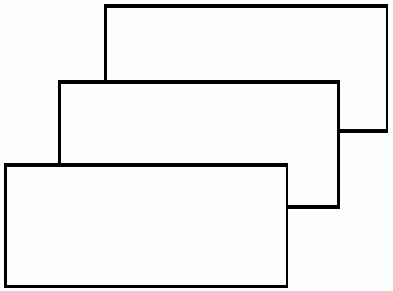
Many (N) high dimensional (d) data



Many (N) low dimensional (q) data



Few (M) high dimensional (d) data



Clustering types

- Projection methods (PCA …)
- Partitioning methods (K-Means...)
- Hierarchical methods
- Density-Based methods
- Grid-based methods

Projection methods

- Principal Component Analysis (PCA)

Minimize reconstruction error

- MDS (Multi Dimensional Scaling):
  - Sammon Mapping
- IsoMap

Preserve distances

- Probabilistic PCA

Maximum likelihood of model

Clustering types

- Projection methods (PCA …)
- Partitioning methods (K-Means...)
- Hierarchical methods
- Density-Based methods
- Grid-based methods

Partitioning methods

- Kmeans, Kmedians

Distance to cluster center

- Competitive Learning ("neural" Kmeans)
  - Self-Organizing Maps (SOM, Kohonen 1982)

- Expectation Maximization (EM)

Maximum likelihood

Kmeans

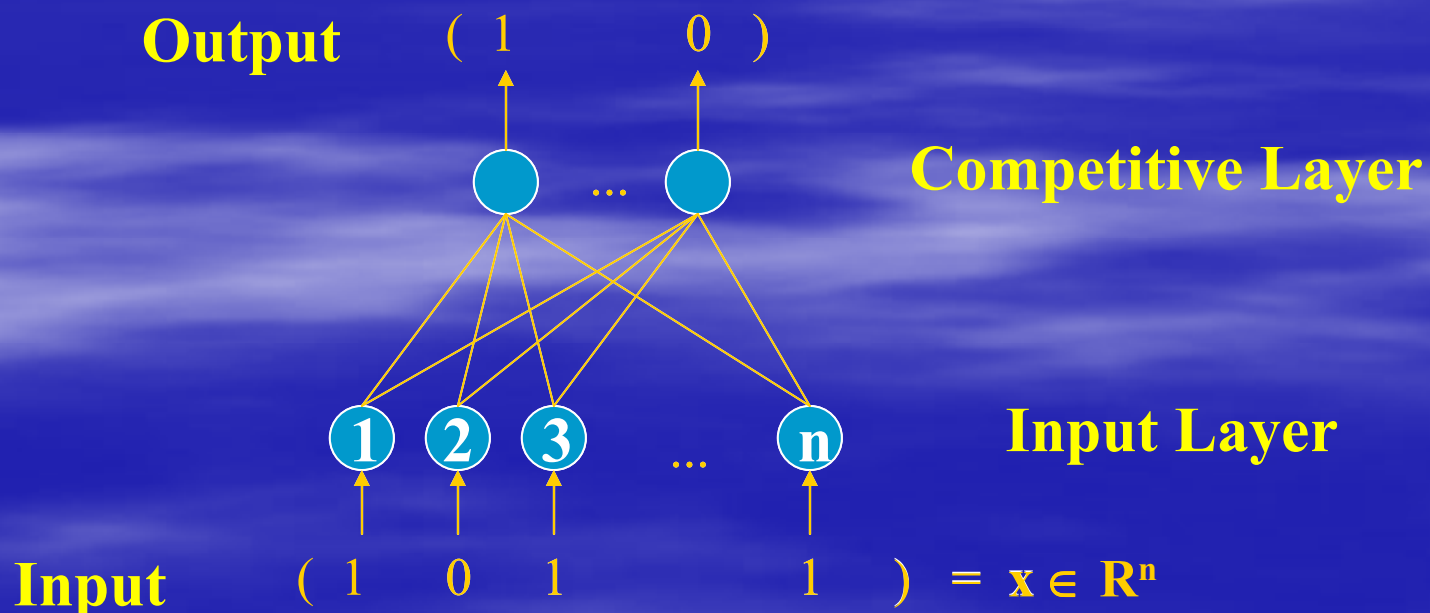
- Cluster center = mean of the objects in the cluster
- Algorithm:
  - (Arbitrarily) choose k centers as initial solution
  - Do until no changes:
    - Compute membership of each object to centers
    - Update cluster centers (=means) according to new memberships
- Kmeans with pruning:
  - Additional steps:
    - Elimination of empty (or weak) clusters
    - Merging of near clusters

Kmedians

- Cluster center = most centrally located object in the cluster
- Algorithm:
  - (Arbitrarily) choose k centers as initial solution
  - Repeat:
    - Randomly pick one of the k centers
    - Replace it with another randomly chosen object from the other (n-k) objects
  - Each object is assigned to the cluster with the closest representative

Competitive Learning

- Idea:
  - competition between neurons
  - One neuron in the competitive layer forms one cluster



Competitive Learning

- The neuron with the highest value for a data item is the winner
- New calculation of the weights of the winner neuron

- Self-Organizing Maps:
  - Additional treatment of “neighborhood”

Expectation Maximization (EM)

- EM = statistical model based on the finite Gaussian mixture model
- Cluster = Gaussian with mean, stddev. and sampling probability
- Basic algorithm:
  - Guess initial values for cluster parameters
  - Repeat until convergence:
    - Estimate the cluster probability for each instance (Expectation)
    - Re-estimate the parameters of the model using the probability score (Maximization)
  - Convergence criteria: e.g. likelihood of the model

Clustering types

- Projection methods (PCA …)
- Partitioning methods (K-Means...)
- Hierarchical methods
- Density-Based methods
- Grid-based methods

Hierarchical Clustering

- Decompose the data into several levels of clusters
- Dendrogram: a tree that splits the data recursively into smaller subsets
- Bottom-up approach (agglomerative)
- Top-down approach (divisive)

- Examples: BIRCH, CHAMELEON

Clustering types

- Projection methods (PCA …)
- Partitioning methods (K-Means...)
- Hierarchical methods
- Density-Based methods
- Grid-based methods

Density-based methods

- Epsilon-neighborhood: *within a radius epsilon of a given object*
- Core object: *epsilon-neighborhood of an object contains at least MinPts*
- Density-reachable: *p is within the epsilon-neighborhood of q,*

  q is core object

Density-based methods

- Steps:
  - Find core objects as new clusters
  - Iteratively split or merge density-reachable clusters
- Use R*-tree (multidim. balanced tree) for good performance
- Examples: DBScan, DENCLUE

Clustering types

- Projection methods (PCA …)
- Partitioning methods (K-Means...)
- Hierarchical methods
- Density-Based methods
- Grid-based methods

Grid-based methods

- Quantize the space into a finite number of cells
- Perform clustering on the grid structure

- Examples: WaveCluster, CLIQUE

SUMMARY

- Partitioning methods:
  - Specify k, number of clusters (→ v-fold cross-validation)
  - No arbitrarily shaped clusters
- Hierarchical clustering:
  - results depends on the ordering of the data (divisive)
  - o frequently used in biology/sociology
- Density-based clustering:
  - \+ Discover clusters of arbitrary shape
  - \+ Separate noise from data
  - \- Many parameters to be adjusted by supervisor
  - \- High complexity