Unsupervised Data Mining: From Batch to Stream Mining Algorithms

Prof. Dr. Stefan Kramer

Johannes Gutenberg-Universität
Mainz

Outline

- Selected topics in clustering

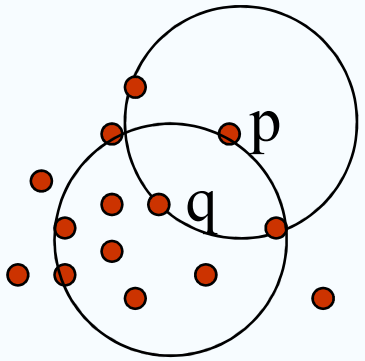
Acknowledgements

- Hans-Peter Kriegel
- Pauli Miettinen

Selected Topics in Clustering

Density-Based Clustering: Basic Concepts

- Clustering based on density (*local* cluster criterion), such as density-connected points: *clusters of arbitrary shape*
- Epsilon neighborhood $N_{Eps}(p) = \{q \text{ belongs to } D \mid dist(p,q) \leq Eps\}$, where Eps is maximum radius of neighborhood
- *Directly density-reachable*: A point p is directly *density-reachable* from a point q w.r.t. Eps, MinPts if p belongs to $N_{Eps}(q)$ and $|N_{Eps}(q)| \geq MinPts$ (core point condition)

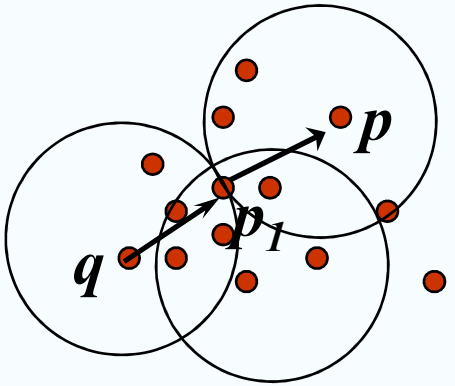


MinPts = 5

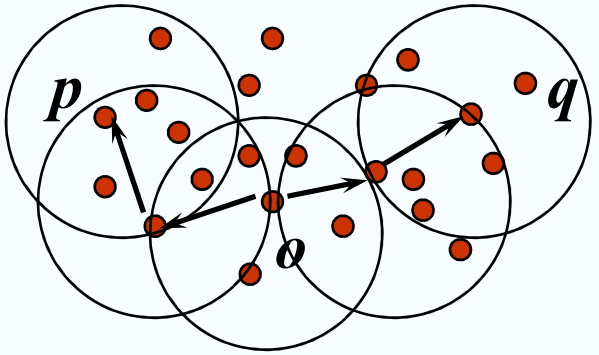
Eps = 1 cm

Density-Reachable and Density-Connected

- A point p is *density-reachable* from a point q w.r.t. Eps, MinPts if there is a chain of points p_1, …, p_n, p_1 = q, p_n = p such that p_{i+1} is *directly density-reachable* from p_i



- A point p is *density-connected* to a point q w.r.t. Eps, MinPts if there is a point o such that both, p and q are density-reachable from o w.r.t. Eps and MinPts

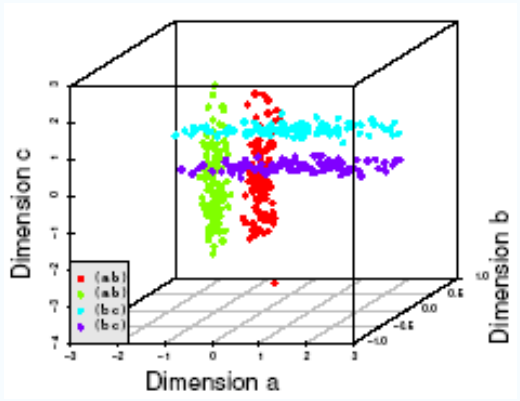


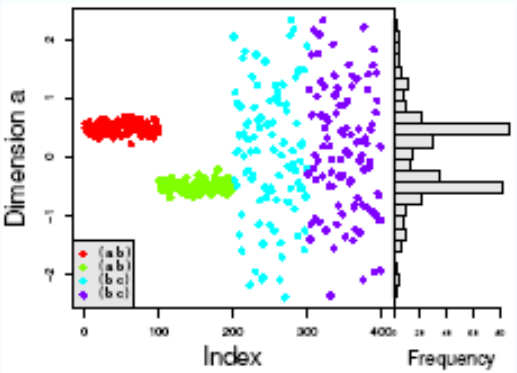
- A cluster is defined as a maximal set of density-connected points
- DBScan, OPTICS and others

Clustering High-Dimensional Data

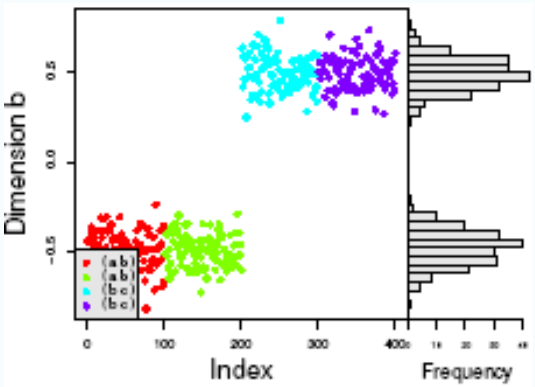
- Clustering high-dimensional data, e.g., in text or microarray data
  - many irrelevant dimensions may *mask clusters*
  - *curse of dimensionality*: distance measure becomes meaningless — instances equidistant and *very far apart*
  - *clusters* may *exist* only in some *subspaces*
- Methods
  - feature transformation (e.g., PCA): only effective if most dimensions are relevant
  - feature selection: wrapper or filter approaches
  - subspace-clustering: find clusters in *all* the possible subspaces
- Finding *random clusters* in subspaces has to be avoided

Subspace Clustering (Parsons *et al.*, 2004)

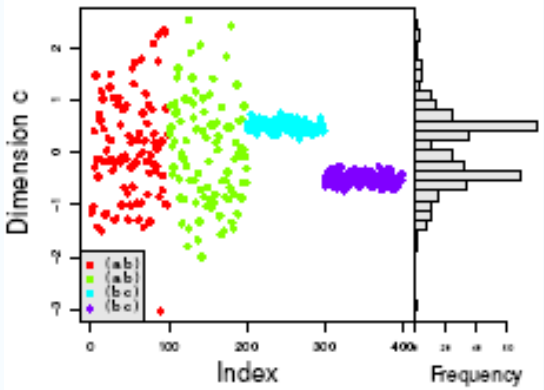




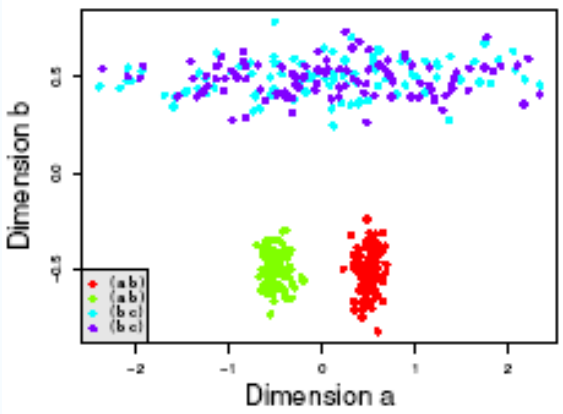
(a) Dimension a



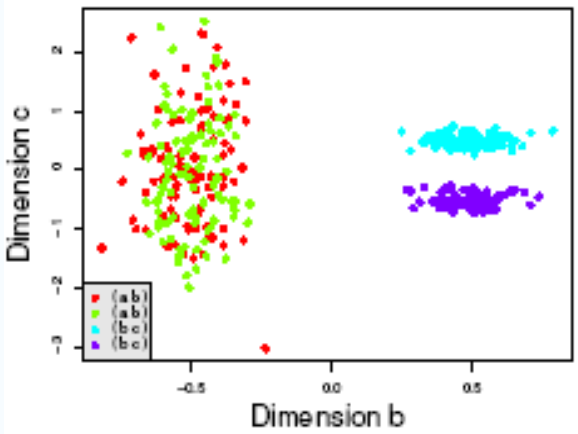
(b) Dimension b



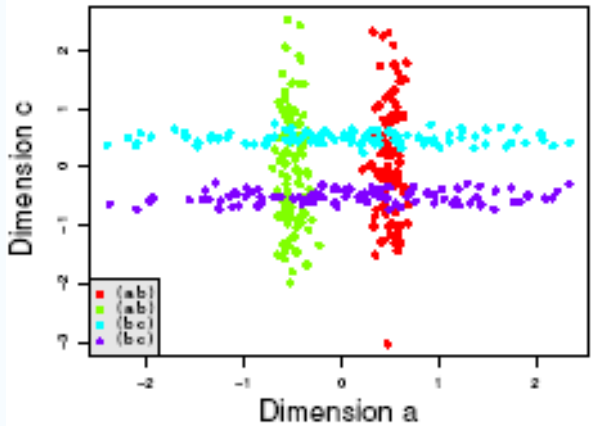
(c) Dimension c



(a) Dims a & b



(b) Dims b & c



(c) Dims a & c

Constraint-Based Clustering

- Desirable to have *user-guided* (i.e., *constrained*) cluster analysis: *users know their applications best*
- Different constraints in cluster analysis:
  - constraints on individual objects (do selection first)
    - cluster on houses worth over $300K
  - constraints on distance or similarity functions
    - weighted functions, obstacles (e.g., rivers, lakes)
  - constraints on the selection of clustering parameters
    - # of clusters, MinPts, etc.
  - user-specified constraints
    - contain at least 500 valued customers and 5000 ordinary ones
  - constraints on objects having to/not allowed to belong to the same cluster
  - *semi-supervised*: giving small training sets as "constraints" or hints
- *The art of **pushing constraints** into search*

Summary

- Cluster analysis groups objects based on their similarity: wide range of applications
- Different *types of clustering algorithms*: partitioning, hierarchical, model-based, graph-based, density-based, … methods
- Current clustering techniques *do not address* all the requirements *adequately*, still an active area of research