CLIQUE: CLustering in QUEst

Agrawal et al, SIGMOD 1998

France ROSE, Jan 2017

Clustering approaches

- “Clustering aims at dividing datasets into subsets (clusters), where objects in the same subset are similar to each other with respect to a given similarity measure, whereas objects in different clusters are dissimilar.”
- Clustering can be used:
  - To better understand the data: data mining, pattern recognition, information retrieval, machine learning
  - As a first step for different purposes: indexing, data compression

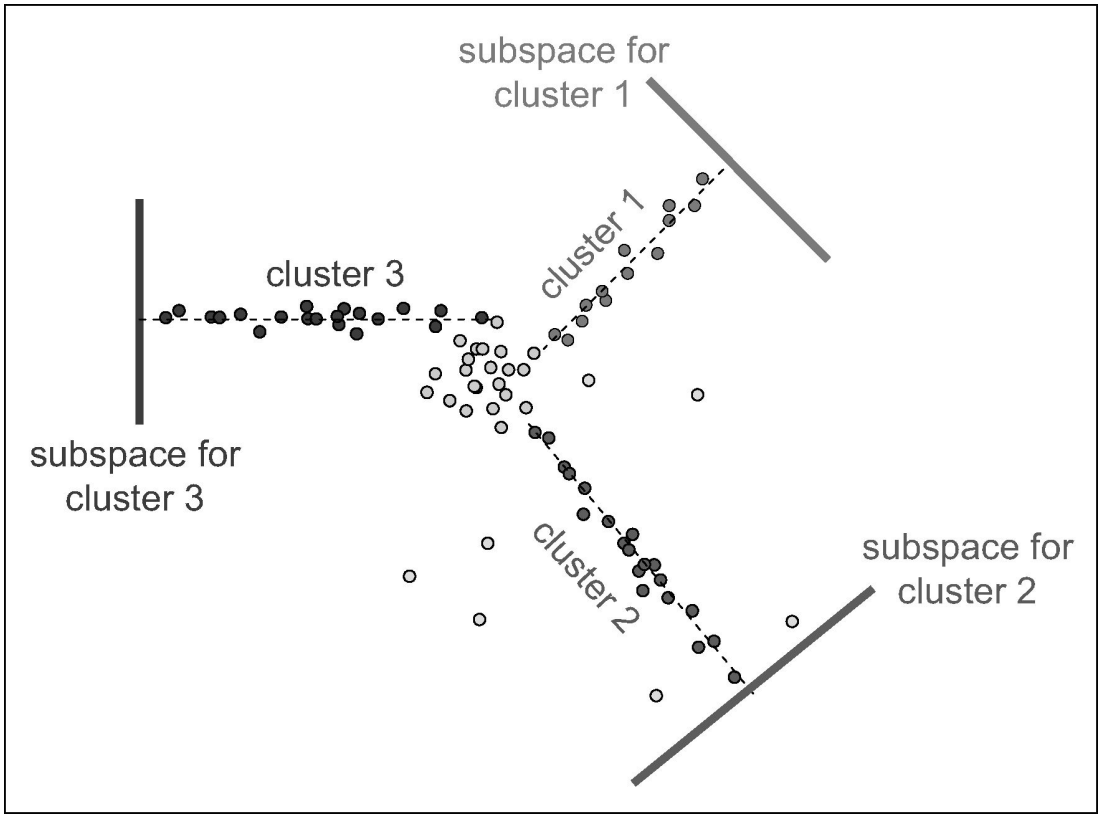
Kriegel et al, 2009

Context and concepts

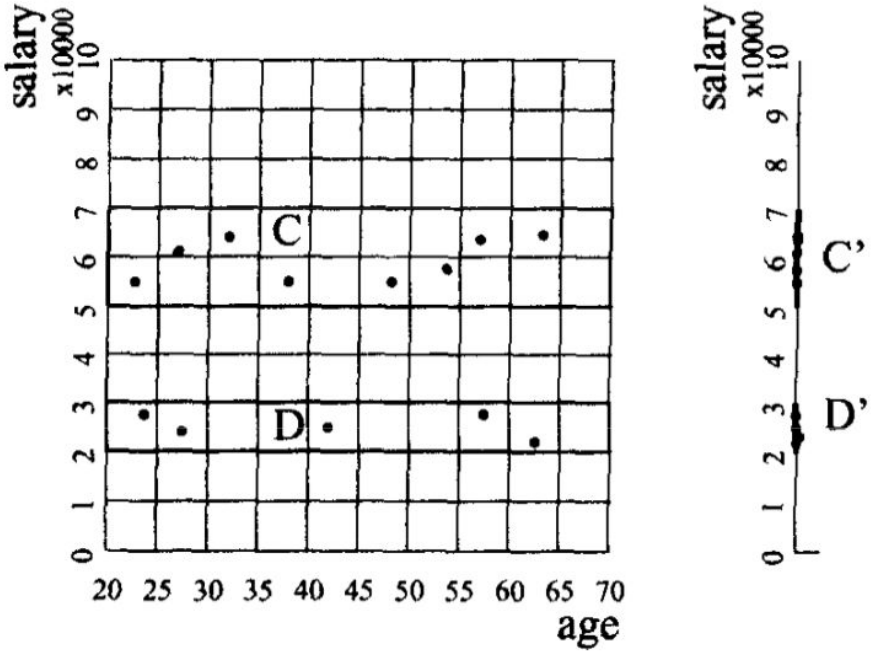
- Clustering techniques: partitional (single level) or hierarchical
- Distance based (k-means) or connectivity based (graph-based or grid-based)
- Special case of high-dimensional data:
  - Irrelevance of distances;
  - Sparsity of the data;
  - Local feature relevance: *different features or a different correlation of features may be relevant for varying clusters*

Agrawal et al., 1998. Kriegel et al, 2009

Data case 1


subspace for
cluster 1
cluster 1
cluster 3
subspace for
cluster 3
cluster 2
subspace for
cluster 2

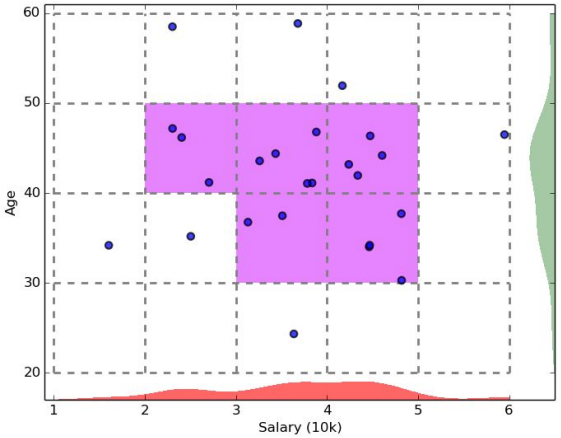
Data case 2

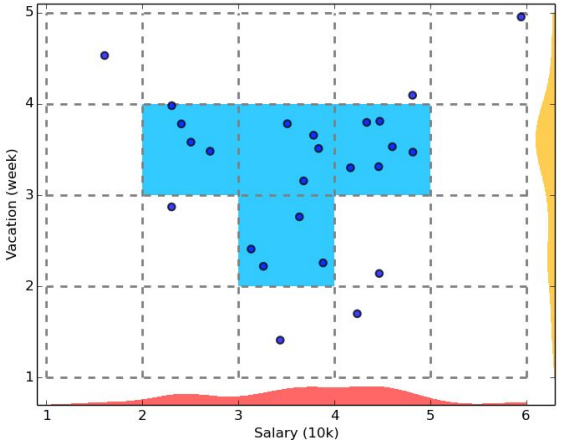

salary
x10000
10
9
8
7
6
5
4
3
2
1
0
C
D
20 25 30 35 40 45 50 55 60 65 70
age
salary
x10000
10
9
8
7
6
5
4
3
2
1
0
C'
D'

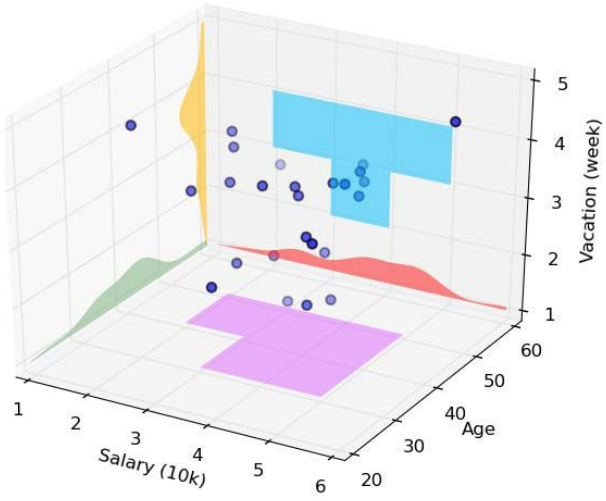
CLIQUE: Grid-Based Subspace Clustering

- CLIQUE is a density-based and grid-based subspace clustering algorithm
  - **Grid-based:** It discretizes the data space through a grid and estimates the density by counting the number of points in a grid cell
  - **Density-based:** A cluster is a maximal set of connected dense units in a subspace
    - A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter
- **Subspace clustering**: A subspace cluster is a set of neighboring dense cells in an arbitrary subspace. It also discovers some minimal descriptions of the clusters
- It **automatically** identifies subspaces of a high dimensional data space that allow better clustering than original space using the Apriori principle

Bottom-up approach







Apriori principle: *If a collection of points S is a cluster in a k-dimensional space, then S is also part of a cluster in any (k-1) dimensional projections of this space*

Major Steps of the CLIQUE Algorithm

- Identify subspaces that contain clusters
  - Partition the data space and find the number of points that lie inside each cell of the partition
  - Identify the subspaces that contain clusters using the Apriori principle
- Identify clusters
  - Determine dense units in all subspaces of interests
  - Determine connected dense units in all subspaces of interests
- Generate minimal descriptions for the clusters
  - Determine maximal regions that cover a cluster of connected dense units for each cluster
  - Determine minimal cover for each cluster

Comments on CLIQUE

- **Strengths**
  - Automatically finds subspaces of the highest dimensionality as long as high density clusters exist in those subspaces
  - Insensitive to the order of records in input and does not presume some canonical data distribution
  - Scales linearly with the size of input and has good scalability as the number of dimensions in the data increases $O(C^k + mk)$
  - Simple method and interpretability of results
- **Weaknesses**
  - As in all grid-based clustering approaches, the quality of the results crucially depends on the appropriate choice of the number and width of the partitions and grid cells

References

- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. SIGMOD'98
- Charu Aggarwal. An Introduction to Clustering Analysis. in Aggarwal and Reddy(eds.). Data Clustering: Algorithms and Applications (Chapter 1). CRC Press, 2014
- Kriegel, H.-P., Kröger, P., & Zimek, A. (2009). Clustering high-dimensional data. ACM Transactions on Knowledge Discovery from Data, 3(1), 1–58.
- Jiawei Han's video on CLIQUE (extract of a coursera/UIUC MOOC) https://www.youtube.com/watch?v=QqkHPJxAXoE
- ELKI framework https://elki-project.github.io/