Intermediate Progress Report HiSC algorithm / Enzymes dataset

Data Mining Group 9:

RAPHAEL BEDNARSKY, MAXIMILIAN FAISSNER,
PETER HUNYADI, LAURA JAHN, NIKOLA VINKO

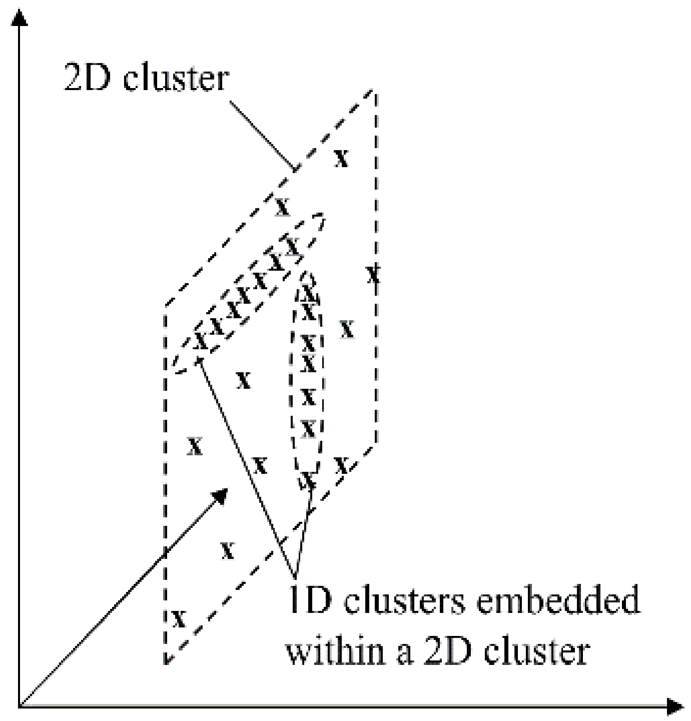
Presentation: Maximilian Faissner

Overview

- HiSC algorithm
  - *Embedded hierarchical structures*
  - *Comparison with OPTICS (related clustering algorithm)*
  - *HiSC Algorithm overview*
  - *Applying HiSC on test inputs / Visualization*

- Exploratory Data analysis of the Enzymes dataset
  - *Presentation by Nikola Vinko*

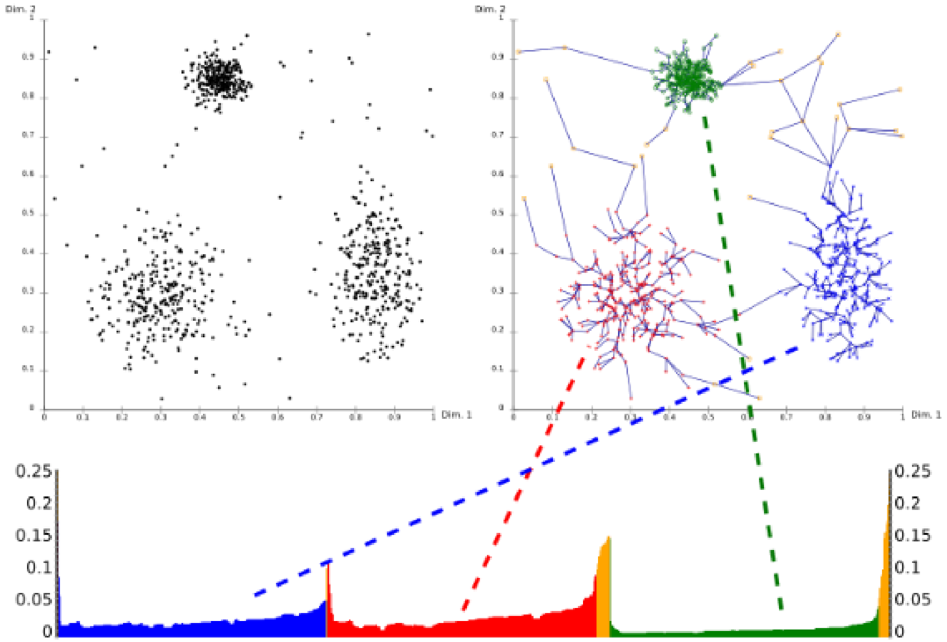
Finding nested subspace clusters



k-dimensional subspace cluster, embedded into l-dimensional subspace cluster (k < l)

HiSC as OPTICS extension

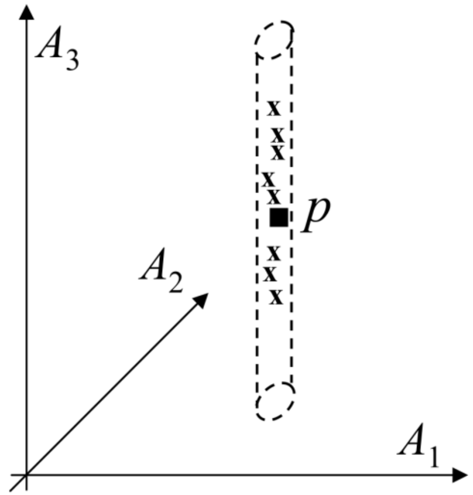
- OPTICS:
  - *DBSCAN-based approach*
  - *Outputs datapoints in a computed ordering with help of reachability distances (kNN)*
  - *Perform a Walk, deterministic walk succession by reachability distance*
  - *Reachability plot:*
    - valleys correspond to clusters (labels are not computed)
- HiSC extension:
  - *Top-Down based (axis-parallel) subspace extension*
  - *Each datapoint is assigned to a subspace dimension & weighted reachability distance*



OPTICS reachability plot (from https://de.wikipedia.org/wiki/OPTICS)

HiSC algorithm overview

- Pre-processing step:
  - *Assign subspace preference vectors to every datapoint*
  - *Based on nearest neighbour (kNN, input parameter k is required)*

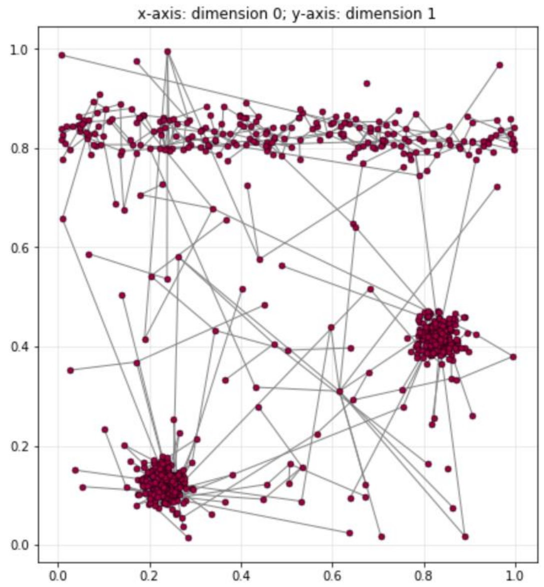


Subspace dimensionality of a single point

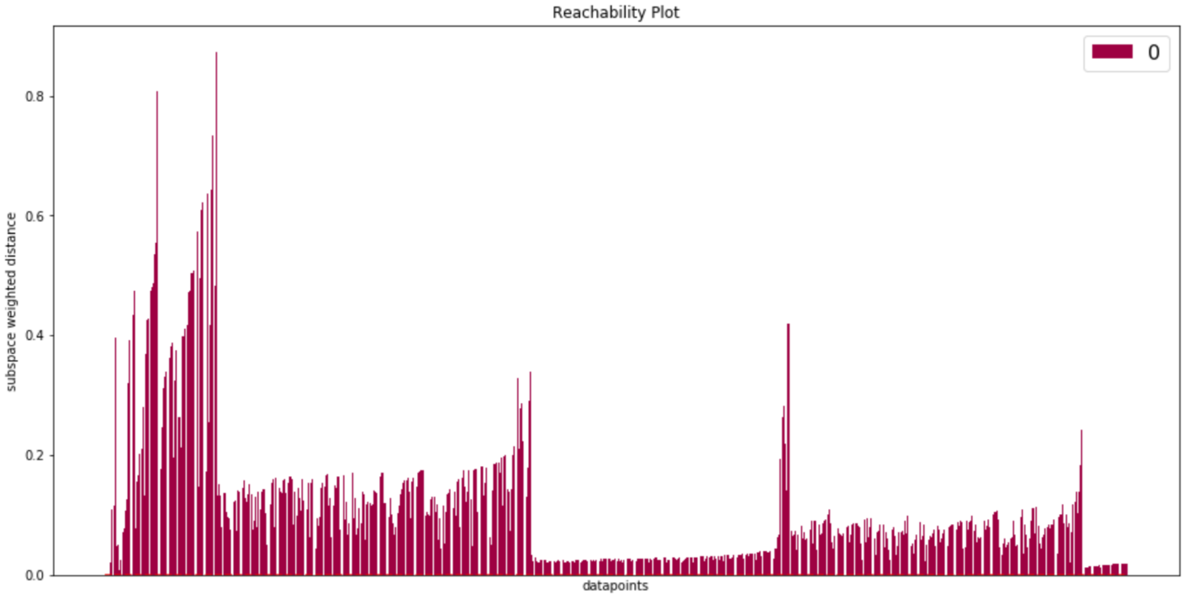
- Perform walk through the dataset like OPTICS:
  - *Next point has the smallest subspace distance to last point*
  - *Initialize priority queue with the first data point randomly chosen*
- Walk succession: Calculate metrics to all remaining datapoints:
  - *Primary sorting by d_1: integer value based on subspace preference vector between 2 points p and q. Requires 2nd input parameter α.*
  - *Secondary sorting by d_2: Subspace-weighted Euclidean distance (based on combined subspace weighting vector)*

Visualization of sample inputs

- Various multi-dimensional test datasets used as input
- Presented example: subspaces_5d.csv, HiSC parameters $\alpha = 0.02, \mathrm{k} = 4$
  - *source of file: ELKI clustering framework*
- Invoke HiSC, plot predecessor & reachability plot without label considerations:



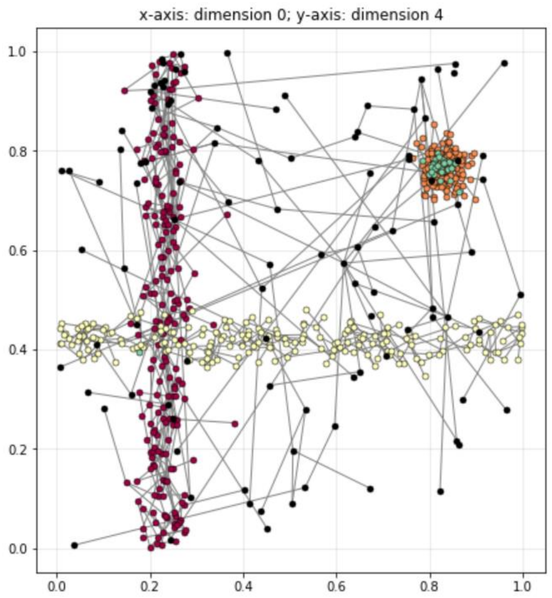
Predecessor plot (2d representation)



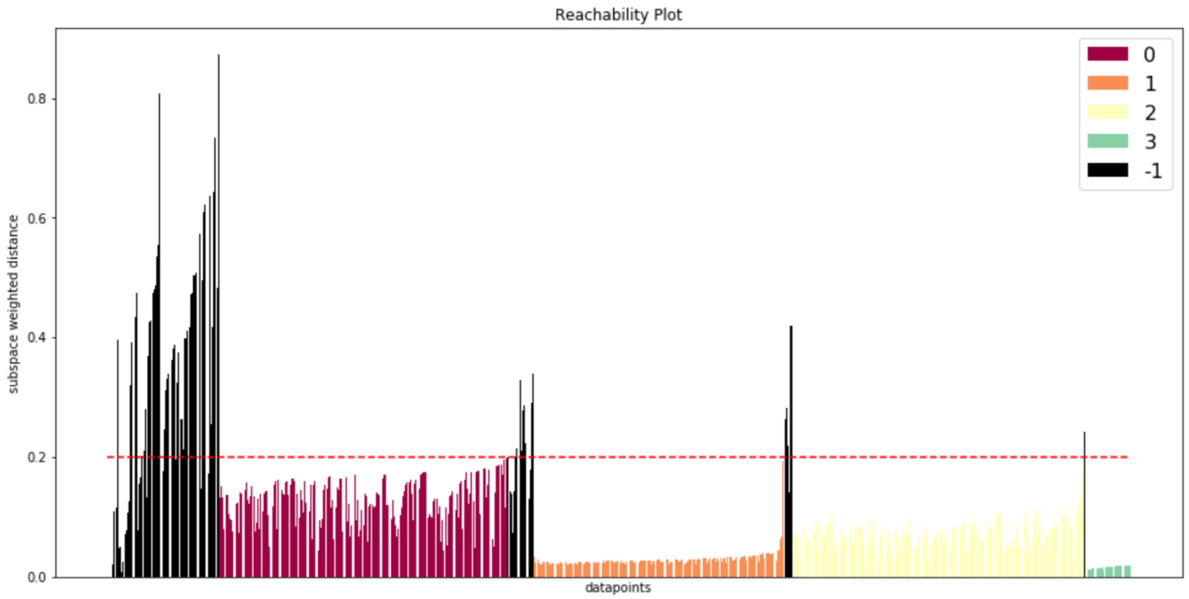
Reachability plot without labels

Cluster Generation

- Set threshold weighted distance k (y-axis of reachability plot)
- Assign all consecutive points to same cluster, if:
  - *Below threshold value k*
  - *A minimum number of points are present in a cluster*

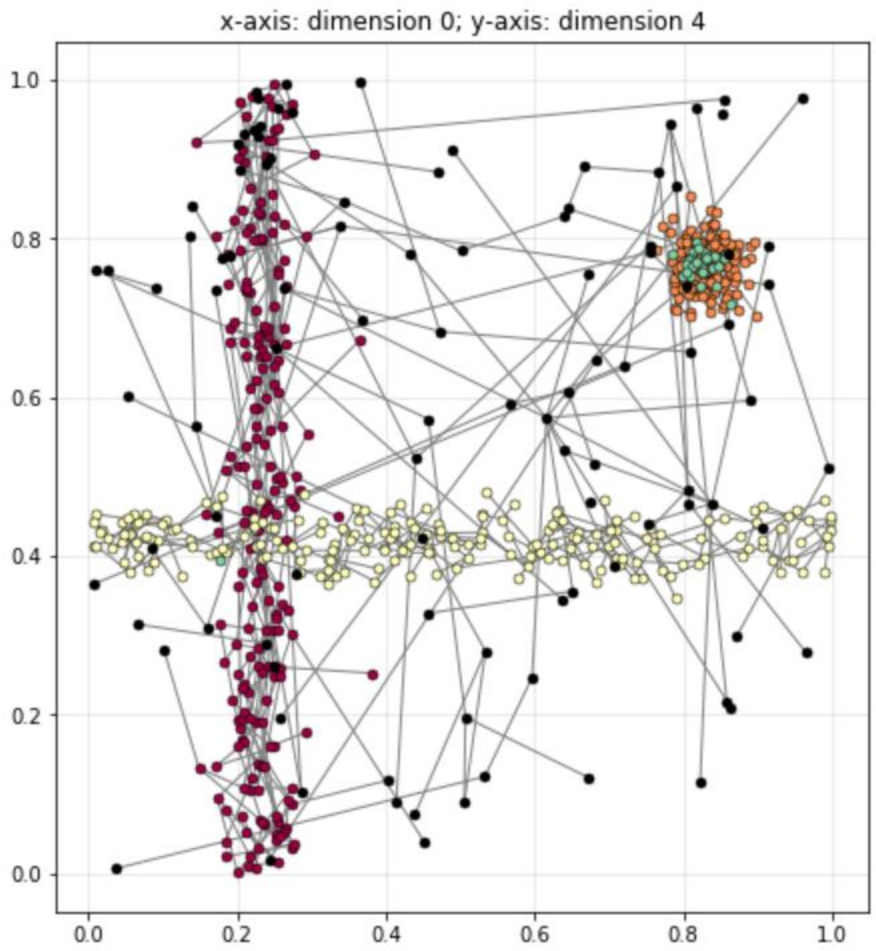


Predecessor plot (2d representation)



Reachability plot with predicted labels

Discussion



- Inconsistent performance on tested input datasets
- HiSC often “finds” embedded structures of clusters with no embeddings
- Selection of parameters α and k is not straight-forward
- Performance on Enzymes dataset: see exploratory data analysis report