5. DeCember 2020

Author: Laura Jahn, Raphael, Max, Peter, Nikola

01426256

WS2020/2021

Report – Data Mining

Exploratory Data Analysis

Inhaltsverzeichnis

[1. Introduction 2](#_Toc58143501)

[2. Cluster analysis 2](#_Toc58143502)

[2.1. What is Clustering? 2](#_Toc58143503)

[2.2. Curse of Dimensionality 3](#_Toc58143504)

[2.3. Subspace Clustering 3](#_Toc58143505)

[3. HiSC 4](#_Toc58143506)

# Introduction

In this intermediate submission we had to choose a dataset and a clustering algorithm which is suited to the dataset. After that we had to implement the algorithm in Python and compare the results with the algorithm implemented in ELKI.

After that we had to validate and analyze the results of our implementation and perform an exploratory data analysis.

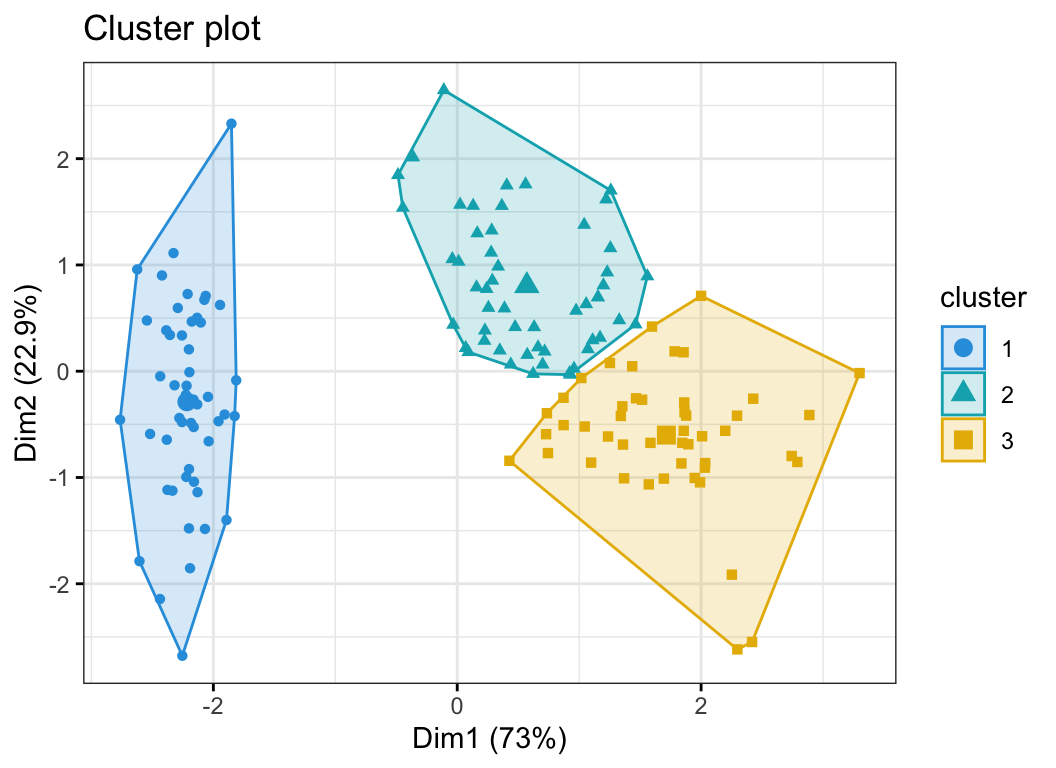
We chose the dataset ENZYMES with the algorithm HiSC. In the following chapters we will describe the usages and the challenges of our implementations.

# Cluster analysis

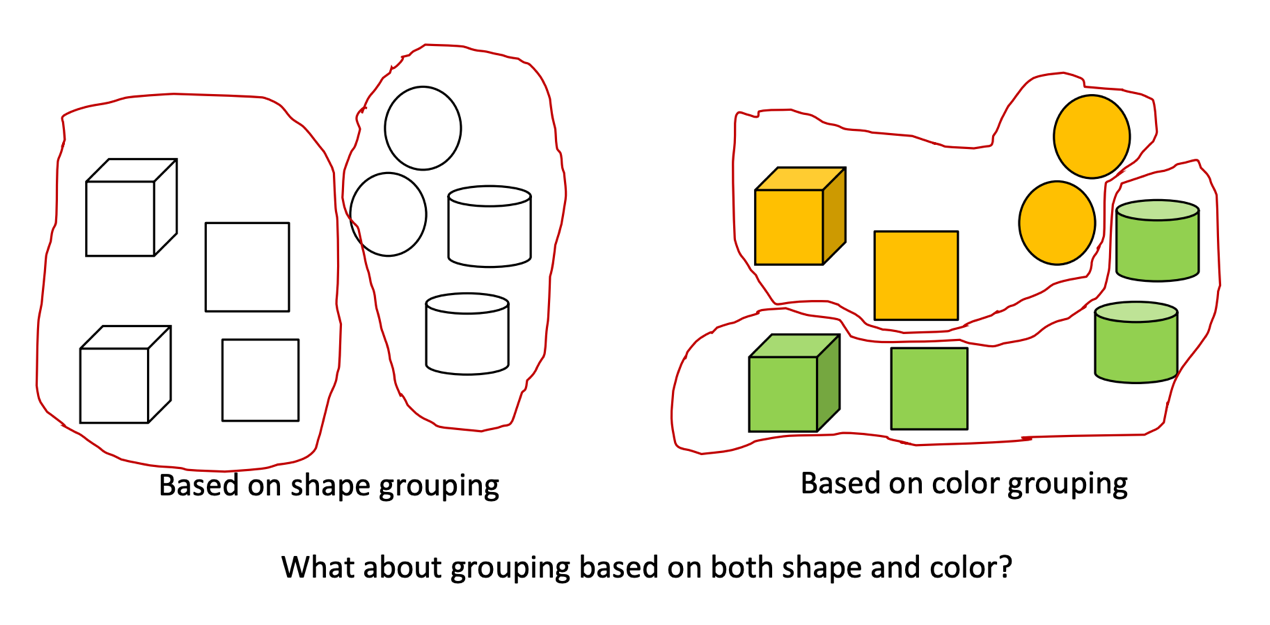
## 2.1. What is Clustering?

On the one hand clusters divide data into groups, so that you can get a better understanding of the similarity of the data. For example if you want to find similar functionality of genes. On the other hand cluster analysis is used for utility like compressing data and find the nearest neighors of a point.

The goal of cluster analysis is to find groups based on the information in the given data where the datapoints are related or similar to the other points in the group. Below you can see an example of how that could look like:[[1]](#footnote-1)



But what if you get higher dimensions in the dataset? How to cluster for example the figure below?[[2]](#footnote-2)[[3]](#footnote-3)



## 2.2. Curse of Dimensionality

If you add more dimensions to a featureset the similarity of two points will probably decrease and therefor the clustering will be more difficult. -> This is called the Curse of Dimensionality.

To avoid this there is a method called Feature Reduction. The idea is to find a low dimensional feature space where redundant and irrelevant features are summarized and so this lower dimensional space represents the original space in a good way but you can cluster it better and easier. But you have tob e careful by reducing dimensions because different attributes are relevant for different clusters.

There are a few ways how to perform the Feature Reduction like Principal Component Analysis (PCA), Singular value decomposition (SVD), Kernel PCA etc.[[4]](#footnote-4)2

## 2.3. Subspace Clustering

The goal of clustering high dimensional data is to search for clusters in subspaces of the original feature space. But there are some challenges you have to face:

#### Find the correct subspace of each cluster

#### Find the correct cluster in each relevant subspace

Because both challenges deppend on each other you have to integrate the subspace search into the clustering process. There are some methods how to do this:

* Bottom-Up Approaches:  
  Find all clusters in all subspaces. (Subspace Clustering)
* Top-Down Approaches:  
  Each point is assigned to one subspace cluster or noise. (Projected Clustering, Correlation Clustering)
* Some novel Approaches:  
  Find the best subspace for all clusters, each point is assigened to one cluster. (Subspace-centered)
* Multiple clusterings:  
  Each point is assigned to multiple clusters in different subspaces. (Non-redundant)[[5]](#footnote-5)4

In the following chapters we will describe the algorithm we implemented which has a Top-Down Approach in a more detailed way:

# HiSC

HiSC is a hierarchical subspace clustering algorithm: Like PreDeCon or PROCLUS, HiSC is a top-down-based approach: Each data point gets an assigned subspace which is a subset of the d-dimensional feature space. This is similar to PreDeCon, which is the most well known subspace clustering algorithm.

In contrast to other subspace clustering algorithms, HiSC allows for overlapping subspace clusters, which allows to detect potential hierarchies, i.e. nested cluster formations.

HiSC was implemented by the author Elke Achtert within the ELKI framework. This implementation was compared with our own implementation with

First step: 2 points a common subspace cluster.

In general, HiSC can be seen as a subspace extension of OPTICS, which is in itself an extension of DBSCAN. The basic approach of DBSCAN will not be explained at this point, however, a quick introduction of OPTICS is necessary since cluster extraction works for HiSC and OPTICS in a similar fashion.

OPTICS attempts to improve upon DBSCAN by sorting neighbours within the ε-neighbourhood with the k-nearest-neighbours algorithm. This tries to alleviate one of DBSCANs weaknesses: clusters of varying densities. A reachability distance is computed, which is set to 0 (or undefined) if the connecting cluster is not sufficiently dense.

OPTICS connects every point with other points with a priority queue: For each unprocessed point, remaining reachability distances are sorted, the closest point is then processed. This generates an ordering, which is presented as output in the form of spanning tree (dimensional plot showing relations) and of a reachability plot (ordering on x-axis and reachability distance on y-axis):

pretty image here

a priority queue, unprocessed points are sorted by the reachability distance for each unprocessed point in the dataset.

Since now (almost) every point

OPTICS. DBscan extension, with same parameters epsilon distance, and minpts. Sorting step of spatially close points to enable clustering of varying density clusters.

p,o: nearest neighbours.

ξ cluster extraction

TODO: Inhaltsverzeichnis aktualisieren, Namen & Matrikelnummern auf Deckenblatt

1. https://www.datanovia.com/en/blog/k-means-clustering-visualization-in-r-step-by-step-guide/ [↑](#footnote-ref-1)
2. https://link.springer.com/chapter/10.1007/978-3-662-08968-2\_16, on 6.December 2020 [↑](#footnote-ref-2)
3. Data Mining Course, p.3: https://moodle.univie.ac.at/pluginfile.php/11513657/mod\_resource/content/2/DM-2-HDData.DimensionalityReduction.pdf [↑](#footnote-ref-3)
4. [↑](#footnote-ref-4)
5. 4 Data Mining Course: https://moodle.univie.ac.at/pluginfile.php/11513692/mod\_resource/content/4/DM-3-HDData.ClusteringHighDimensionalDataPart1.pdf [↑](#footnote-ref-5)