# Visualisierung des Hierarchischen Clusterings mit Auswahl der Clustering-Methode und des Distanzmaßes

Studienleistung Data Mining mit R

Zoe Willett

WiSe 19/20

# 1 Hierarchisches Clustering

Bei dem hierarchischen Clustering geht es darum, ähnliche Objekte in benachbarten Teilbäumen zu organisieren, um am Ende eine Baumstruktur zu erhalten. In diesem Programm wird das agglomerative Clustering verwendet. Das bedeutet, zu Beginn bildet jedes Objekt ein eigenes Cluster, die dann sukzessive zusammengefasst werden. Es handelt sich also um ein bottom-up Clustering. Um die einzelnen Cluster zusammenzufassen, können verschiedene Methoden verwendetet werden. Welche Methode verwendet werden soll, kann im Programm ausgewählt werden.
Das Distanzmaß kann ebenfalls ausgewählt werden. Es dient dazu die Ähnlichkeit der Records für das Clustering zu bestimmen. Je größer die Distanz ist, desto kleiner ist die Ähnlichkeit.

## 1.1 Clustering Methoden

**Complete Linkage Clustering** Um die Distanz zwischen zwei CLustern zu bestimmen, wird die Distanz zwischen den am weitesten entfernt liegenden Objekten gewählt.

$$d(A,B) = max_{\substack{\alpha=1,\dots,n_A \\ \beta=1,\dots,n_B}} d(a_\alpha, b_\beta)$$

Das Complete Linkage Clustering führt zu kleinen kompakten Clustern

**Single Linkage Clustering** Hier wird die Distanz zwischen zwei Clustern bestimmt, in dem man die Distanz zwischen den am nächsten liegenden Objekten verwendet.

$$d(A,B) = min_{\substack{\alpha=1,\dots,n_A \\ \beta=1,\dots,n_B}} d(a_\alpha, b_\beta)$$

Diese Methode führt häufig zu Chaining.

**Average Linkage Clustering** Diese Methode dient als Kompromiss zwischen den beiden vorher genannten. Es wird die durchschnittliche Distanz zwischen allen Objektpaaren verwendet.

$$d(A,B) = \frac{1}{n_A * n_B} \sum_{\alpha=1}^{n_A} \sum_{\beta=1}^{n_B} d(a_\alpha, b_\beta)$$

## 1.2 Distanzmaße

**Euklidische Distanz**

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Damit die Bedingungen für eine Metrik erfüllt werden, wird bei der euklidischen Distanz quadriert und dann die Wurzel gezogen.

**Manhattan-Distanz**

$$d(X,Y) = \sum_{i=1}^{n} |x_i - y_i|$$

Bei der Manhattan-Distanz werden einfach Betragsstriche verwendet.

### 1.3 Golub-Daten

Die verwendeten Daten stammen aus einem Microarray-Experiment, das mit Daten von Patienten durchgeführt wurde, die an Leukämie leiden. Die Daten wurden in zwei Gruppen unterteilt: ALL (Acute Lymphatic Leukemia) und AML (Acute Myeloid Leukemia). Um mit den Daten arbeiten zu können, wird das *golubEsets* Package von *Bioconductor* verwendet.

## 2 Erklärung R-Skript

Die App wurde mit Shiny erstellt. Das Programm ist unterteilt in das *UI* und die *server logic*.
Im UI-Abschnitt wird festgelegt, wie der Name schon sagt, wie das User Interface aussehen soll.
Im server-logic-Abschnitt werden die eigentlichen Berechnungen durchgeführt.

### 2.1 verwendete Funktionen

**library(GolubEsets)** Hiermit wird das Package mit den verwendeten Daten geladen

**data("Golub_Train")** Hiermit werden die spezifischen benötigten Daten geladen

**x=exprs(Golub_Train)** Die verwendeten Expressionsdaten werden in x gespeichert.

**tx = t(x)** Die Matrix wird transponiert

**x[is.na(x)] = 0** Hier werden die Nullwerte entfernt

**plot(hclust(dist(*Werte*, *Distanzmaß*), *Clusteringmethode*)** Hier wird der eigentliche Plot, also das hierarchische Clustering, erstellt. Der Funktion werden die Daten, das ausgewählte Distanzmaß und die ausgewählte Clusteringmethode übergeben.

### 2.2 Installationsanleitung

Das ganze Programm besteht aus nur einem Skript (app.R). Um das Skript auszuführen muss nur das Shiny und das golubEsets Paket heruntergeladen werden.
Programmiert wurde mit der R-Version 3.4.4 und getestet wurde unter Ubuntu 18.04

# 3 Interpretation

Das hierarchische Clustering zeigt das Clustering der Leukämie Patienten. Da die Patientendaten in zwei Gruppen unterteilt wurden (AML und ALL) erwartet man, dass man eine Trennung zwischen den beiden Gruppen erkennen kann. Dies ist aber nicht der Fall. Die Expression der Gene hängt also von mehr Faktoren ab, als nur der verschiedenen Ausprägung der Krankheit. Man kann aber sehr gut erkennen, dass die Wahl des Distanzmaßes und der Clustering-Methode eindeutige Auswirkungen auf das entstehende Clustering hat.