Dr. Katharina Best
Sommersemester 2011
12. Mai 2011

ulm university universität ulm

# Reader – Teil 5: Clusteranalyse

## WiMa-Praktikum

Bei der Clusteranalyse wollen wir Gruppen in Daten auffinden. Die Aufgabe ist, in vorhandenen Daten *Klassen* resp. *Cluster* so zu bestimmen, dass sich die Elemente einer Klasse ähneln währed diejenigen aus unterschiedlichen Klassen möglichst unterschiedlich sind.

Betrachten Sie dazu die Abbildung 1. Wie ließe sich da gruppieren?

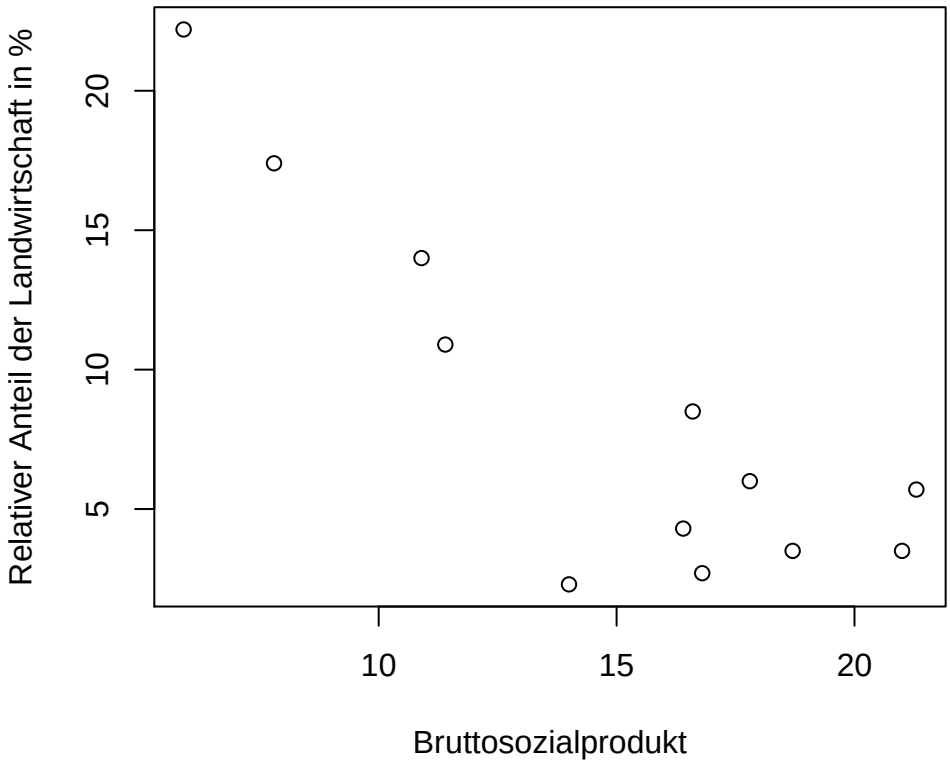


Abbildung 1: Die Daten sind Erhebungen der EU aus dem Jahr 1994. Die waagrechte Achse zeigt das Bruttosozialprodukt, die senkrechte den relativen Anteil der Landwirtschaft am Bruttosozialprodukt. Die Punkte stellen die Länder da, sind jedoch unbeschriftet.

# 1 Vorgehen der Clusteranalyse

Die Verfahren der Clusteranalyse unterscheiden sich im Aufbau in ihrem Distanz- bzw. Ähnlichkeitsbegriff und ihrer algorithmischen Ausrichtung.

**Partitionierende Verfahren** Eine Zielfunktion soll optimiert werden. Zu diesem Zweck wird eine Partition der Punkte festgelegt und diese werden mittels Permutationen und anderen Austauschfunktionen den einzelnen Klassen zugeordnet. Hierbei ist die Anzahl der Klassen von vornherein feststehend.

**Hierarchische Verfahren** Es kann der *top-down* oder der *bottom-up* Ansatz gewählt werden. Dabei wird von der feinsten (jedes Element in einer eigenen Menge) resp. gröbsten Partition (eine Menge) ausgegangen. Anschließend werden diese Partitionen zusammengefasst resp. aufgespalten. Augenscheinlich ist, dass hier eine Abbruchbedingung von Nöten ist, denn sonst endet man mit der gröbsten resp. feinsten Partition.

# 2 Distanz- und Ähnlichkeitsmaße

## 2.1 Definition des Distanzmaßes

Ein Distanzmaß auf dem Raum $S$ ist eine Abbildung $\delta : S \times S \to [0, \infty)$ mit $\delta(x, x) = 0$ und $\delta(x, y) = \delta(y, x)$ für $x, y \in S$. Natürlich ist eine Metrik auch ein Distanzsmaß.

Der Abstand zweier Beobachtungen wird durch den Abstand ihrer Merkmalsvektoren angegeben, gemessen mit $\delta$.

## 2.2 Beispiele von Distanzmaßen

Die bekanntesten Distanzmaße sind durch $p$-Normen erzeugte Metriken, genannt *Minkowski-Metriken*. Dabei ist $\delta_p(x, y) = \|x - y\|_p$ und

$$\|x\|_p := \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \tag{1}$$

die bekanntesten Distanzmaße sind

- $p = 1$: Manhattan-Metrik,
- $p = 2$: Euklidische Abstand,
- $p = \infty$: Maximumsabstand.

Ein weiteres wichtiges Distanzmaß ist die *Mahalanobis-Distanz*, gegeben durch

$$\delta_S := \left((y_i - y_j)^T \boldsymbol{S_{y,n}}^{-1} (y_i - y_j)\right)^{1/2}, \tag{2}$$

wobei

$$\boldsymbol{S_{y,n}} := \frac{1}{n-1} \sum_{i=1}^{n} (y_i - y_j)(y_i - y_j)^T \tag{3}$$

die empirische Kovarianzmatrix der Werte $y_1, \ldots y_n \in \mathbb{R}^r$ ist.

## 2.3 Eigenschaften von Distanzmaßen

**Skaleninvarianz** Ein Distanzmaß $\delta$ ist *skaleninvariant* auf der Menge $\{y_1, \ldots, y_n\}$, falls $\delta(y_i, y_j) = \delta(\boldsymbol{\alpha} y_i, \boldsymbol{\alpha} y_j)$ gilt für alle $i, j \in \{1, \ldots, n\}$ und alle Diagonalmatrizen $\boldsymbol{\alpha} = \text{diag}(\alpha_1, \ldots, \alpha_r)$.

**Translationsinvarianz** Ein Distanzmaß $\delta$ ist *translationsinvariant* auf der Menge $\{y_1, \ldots, y_n\}$, falls $\delta(y_i, y_j) = \delta(y_i + z, y_j + z)$ gilt für alle $i, j \in \{1, \ldots, n\}$ und alle $z \in \mathbb{R}^r$.

**Invarianz unter orthogonalen Transformationen** Ein Distanzmaß $\delta$ ist *invariant unter orthogonalen Transformationen* auf der Menge $\{y_1, \ldots, y_n\}$, falls $\delta(y_i, y_j) = \delta(\boldsymbol{\alpha} y_i, \boldsymbol{\alpha} y_j)$ gilt für alle $i, j \in \{1, \ldots, n\}$ und alle orthogonalen Matrizen $\boldsymbol{\alpha}$, d. h. $\boldsymbol{\alpha}^T \boldsymbol{\alpha} = I_r$, wobei $I_r$ die Einheitsmatrix des $\mathbb{R}^r$ ist.

Die Minkovski-Metriken sind zwar translationsinvariant, jedoch nicht skaleninvariant. Der euklidische Abstand ist darüber hinaus noch invariant unter orthogonalen Transformationen. Die Mahalanobis-Distanz ist so beliebt, da sie sowohl skalen- als auch translationsinvariant ist, desweiteren auch invariant unter orthogonalen Transformationen.

## 2.4 Definition des Ähnlichkeitsmaßes

Ein Ähnlichkeitsmaß auf dem Raum $S$ ist eine Abbildung $\rho : S \times S \to [0, 1]$ mit $\rho(x, x) = 1$ und $\rho(x, y) = \rho(y, x)$ für $x, y \in S$. Natürlich sind Distanz- und Ähnlichkeitsmaße verwandt, allein schon dadurch, dass zwei Punkte, die bezüglich eines Distanzmaßes einen kleinen Abstand haben, ähnlich sind.

Überlegen Sie sich, wie diese ineinander übergeführt werden können.

Insbesondere bei kategorialen Daten werden lieber (aus historischen Gründen) ausgewiesene Ähnlichkeitsmaße verwendet.

## 2.5 Beispiele von Ähnlichkeitsmaßen

Die bekanntesten Ähnlichkeitsmaße bei Dummy-kodierten Merkmalsvektoren sind

- *Jacard-Koeffizient*, gegeben durch

$$\rho_J(i,j) := \frac{y_i^T y_j}{r-(1-y_i)^T(1-y_j)} \mathbb{1}_{\{(1-y_i)^T(1-y_j)<r\}} + \mathbb{1}_{\{(1-y_i)^T(1-y_j)=r\}},$$

- *Czekanowsky-Koeffizient*, gegeben durch

$$\rho_X(i,j) := \frac{2y_i^T y_j}{y_i^T y_j + r-(1-y_i)^T(1-y_j)} \mathbb{1}_{\{(1-y_i)^T(1-y_j)<r\}} + \mathbb{1}_{\{(1-y_i)^T(1-y_j)=r\}},$$

- *M-Koeffizient*, gegeben durch

$$\rho_M(i,j) := \frac{y_i^T y_j + (1-y_i)^T(1-y_j)}{r}.$$

# 3 Hierarchische Verfahren

## 3.1 Single-Linkage-Verfahren

Hierbei wird der Abstand zweier Mengen $A, B$ als der minimale Abstand der Elemente,

$$\delta(A,B) = \min_{i\in A, j\in B} \delta(i,j)$$

definiert. Die Methode wird als *Single-Linkage-* oder *Nearest-Neighbour-Verfahren* bezeichnet.

## 3.2 Complete-Linkage-Verfahren

Der Abstand zweier Mengen $A, B$ ist der maximale Abstand der Elemente,

$$\delta(A,B) = \max_{i\in A, j\in B} \delta(i,j).$$

Die Methode wird als *Complete-Linkage-* oder *Furthest-Neighbour-Verfahren* bezeichnet.

## 3.3 Average-Linkage-Verfahren

Der Abstand zweier Mengen $A, B$ mit $|A| = n_A, |B| = n_B$ wird hier gemittelt über die Einzelabstände,

$$\delta(A,B) = \frac{1}{n_A n_B} \sum_{i\in A} \sum_{j\in B} \delta(i,j)$$

definiert.

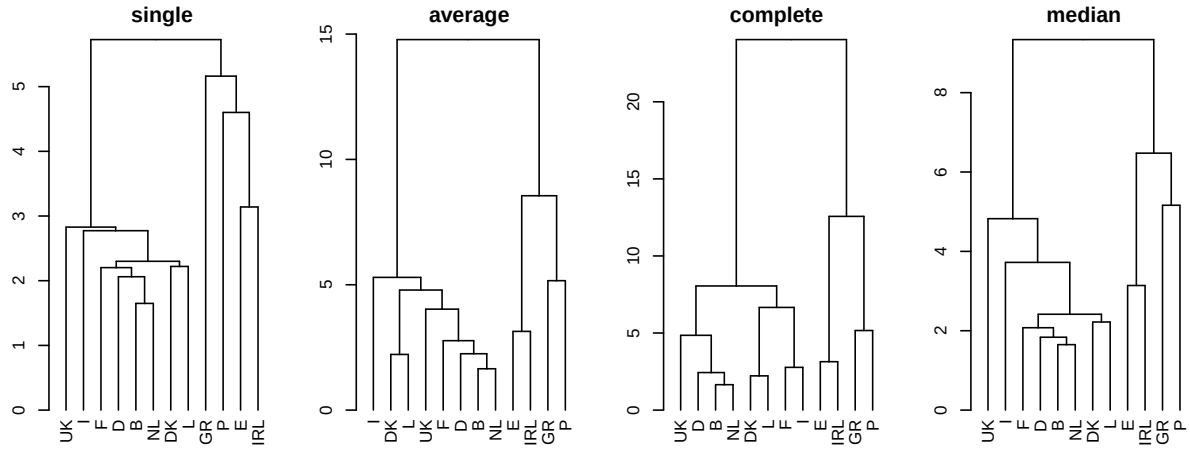


Abbildung 2: Dendrogramme der oberen Landwirtschaftsdaten für verschiedene hierarchische Verfahren, von links nach rechts: single linkage, average linkage, complete linkage, median.

### 3.4 Zentroid- und Median-Verfahren

Bei dieser Abstandsmessung wird jeweils ein Schwerpunkt der Menge gebildet, also $\mathcal{Y}_A := \frac{1}{n_A} \sum_{i \in A} y_i$ und $\mathcal{Y}_B := \frac{1}{n_B} \sum_{j \in B} y_j$.

Das Zentroid-Verfahren gibt dann als Abstand zweier Mengen $\delta(A, B) = \delta_2(\mathcal{Y}_A, \mathcal{Y}_B)^2$, während das Median-Verfahren $\delta(A, B) = \delta_1(\mathcal{Y}_A, \mathcal{Y}_B)$, wählt.

## 4 Abbruch des hierarchischen Prozesses

### 4.1 Darstellung der Clusterbildung

Das Vorgehen des Vereinigens kann mit einem *Dendrogramm* dargestellt werden. Hierbei werden Klassen minimaler Distanz verbunden. Die Distanzen werden durch die Abstände zwischen den Verbindungen kodiert, siehe dazu die Abbildung 2 So kann visuell eine geeignete Distanz, in der das Verfahren abgebrochen wird, bestimmt werden.

Darüber hinaus erhält man eine schöne Visualisierung des Effektes der verschiedenen Abstandsmaße.

### 4.2 Clusteranzahl

Das visuelle Vorgehen von oben kann auch funktional aufbereitet werden. Dazu wird eine Funktion $\varphi$ mittels

$$\varphi(k) := \min_{A,B \in \text{ Partition zur Zeit } k+1} \delta(A, B) \tag{4}$$

definiert, die Abstände der Zukunft abbildet und deren Steigung dann ein Abbruchkriterium liefert.