

Clusteranalyse

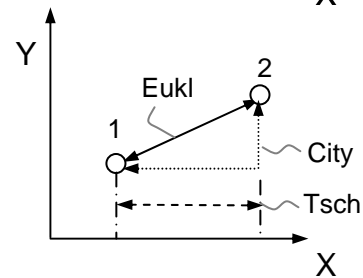
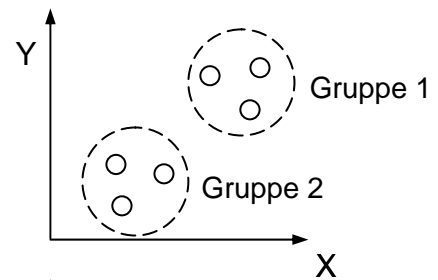
Unter Clusteranalyse versteht man im Wesentlichen eine Gruppierung von ungeordneten Daten (z.B. Messwerte, Bildpunkte, usw.). Die Gruppierung erfolgt durch festzulegende Ähnlichkeitsmerkmale. Das sind in der Regel Abstandsdaten, wie das dargestellte Bild verdeutlicht. In diesem Fall besteht eine hohe Ähnlichkeit, wenn die Datenpunkte einen möglichst geringen Abstand zueinander haben.

Maße für die Beurteilung der Abstände zwischen den Objekten (d =Heterogenitätsmaß)

Euklidische Distanz : $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

City-Block Distanz : $d = |x_2 - x_1| + |y_2 - y_1|$

Tschebyscheff Distanz : $d = \max(|x_2 - x_1|; |y_2 - y_1|)$



Ähnlichkeiten können auch in Form einer Korrelationsmatrix vorliegen. Je höher die Korrelation ist, desto ähnlicher sind die „Objekte“ zueinander. Hier ist also ein größerer Wert maßgebend. Zudem liegen die Ausgangsdaten nicht in Form von Koordinaten vor, sondern als Matrix in der jedes Objekt mit jedem anderen in Beziehung steht (ausgedrückt durch den Korrelationskoeffizient r). Die Objektdistanz ist in diesem Fall mit $d = \text{ArcCos}(r)$ auszudrücken (siehe grafische Interpretation Faktoranalyse). Vereinfacht kann auch mit $d = 1 - r$ gerechnet werden.



Sind die Ähnlichkeiten nicht auf die Zeilen, sondern auf die Titel der Datenspalten zu beziehen, so ist vor der Clusteranalyse eine Korrelationsmatrix zu erstellen.

Die Zielsetzung der Cluster ist:

- Eine vereinfachte übersichtlichere Struktur schaffen
- Datenreduktion
- Erkennen von Zusammenhängen

Es gibt eine Vielzahl von Verfahren zur Bildung von Clustern. Man unterscheidet u.a.

- Partitionierende Methoden, z.B. K-Means-Verfahren
- Hierarchische Methoden, z.B. agglomerative Verfahren

Das sehr häufig zu findende **K-Means**-Verfahren geht folgendermaßen vor.

1. Vorgabe der Anzahl von Clustern
2. Zufällige Vergabe von Clusterzentren
3. Zuweisung der Objekte über kleinsten Abstand zum Clusterzentrum
4. Bestimmung der Summe aller Abstandsquadrate
5. Wiederholung ab Punkt 2. Abbruch, wenn Summe der Abstandsquadrate nicht mehr kleiner werden

Der Vorteil dieses Verfahrens ist:

- Leichte Implementierung möglich

Die Nachteile sind jedoch:

- Algorithmus findet nur lokales Optimum, abhängig von Startcluster
- Jeder Neustart kann deshalb andere Cluster finden
- Anzahl Cluster muss vorher bekannt sein
- Nicht für große Datenmengen geeignet

Als Alternative ist u.a. das **hierarchisch agglomerative Verfahren** zu nennen.

Die Vorteile sind:

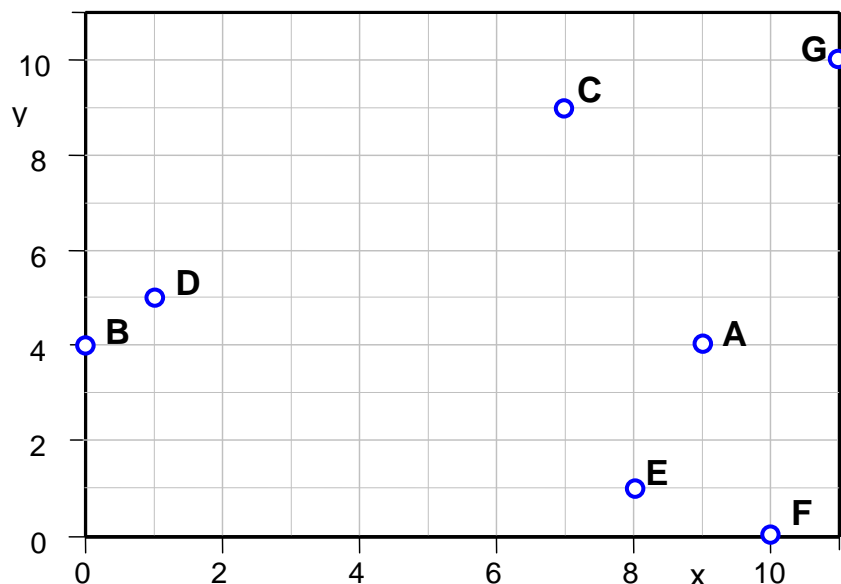
- Keine Festlegung auf Anzahl Cluster notwendig
Nachträgliche Reduzierung der Cluster durch „Grenzdistanz“ möglich
- Jeder Rechendurchlauf ergibt das gleiche Ergebnis
- Leicht zu implementierende effizienter Algorithmus
- Grafische Darstellungsmöglichkeit der Cluster als Baumstruktur

Als Nachteil ist zu nennen:

- Zuordnung der Objekte zu den Clustern ist fest
- Bei hoher Datenmenge großer Strukturbaum mit Speicherbedarf notwendig
- Rekursive Verfahren wegen „Speicher-Stack“ limitiert

Es gibt jedoch auch nicht rekursive Algorithmen, die einen geringen Datenaufwand erfordern. Das Verfahren soll an einem einfachen Beispiel verdeutlicht werden. Gegeben sind folgende Objekte mit ihren Koordinaten:

| | x | y |
|----------|----|----|
| A | 9 | 4 |
| B | 0 | 4 |
| C | 7 | 9 |
| D | 1 | 5 |
| E | 8 | 1 |
| F | 10 | 0 |
| G | 11 | 10 |

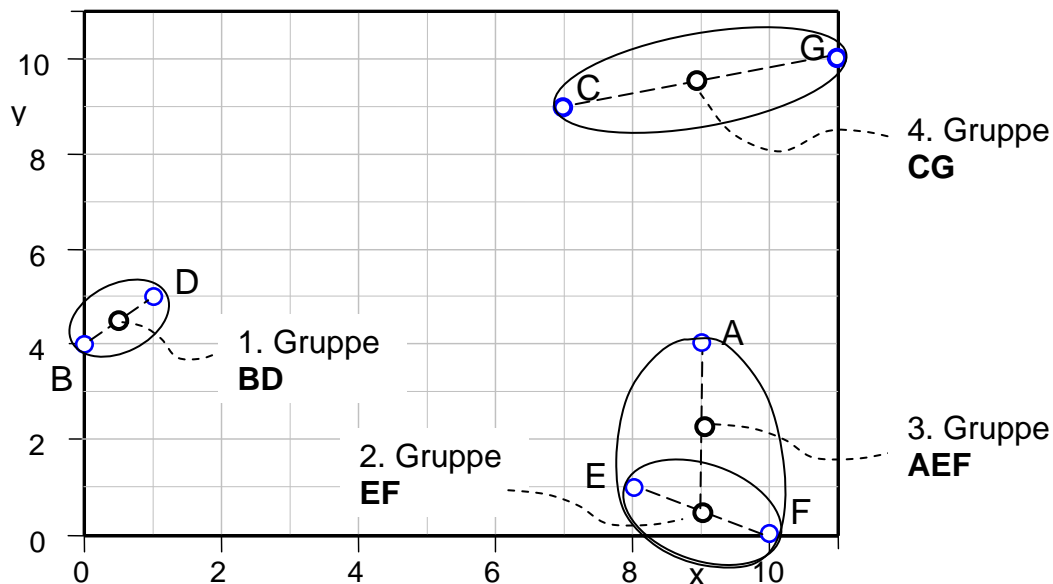


Hier sind nur 2 Koordinaten dargestellt. Denkbar sind n-dimensionale Koordinaten (Spalten), wobei man sich bei 3 Koordinaten die Anordnung als 3D Darstellung noch gut vorstellen kann.

Aus den Koordinaten ergibt sich die Distanzmatrix (Werte=euklidische Distanz):

| | A | B | C | D | E | F | G |
|---|-----|------|-----|------|-----|------|------|
| A | | 9,0 | 5,4 | 8,1 | 3,2 | 4,1 | 6,3 |
| B | 9,0 | | 8,6 | 1,4 | 8,5 | 10,8 | 12,5 |
| C | 5,4 | 8,6 | | 7,2 | 8,1 | 9,5 | 4,1 |
| D | 8,1 | 1,4 | 7,2 | | 8,1 | 10,3 | 11,2 |
| E | 3,2 | 8,5 | 8,1 | 8,1 | | 2,2 | 9,5 |
| F | 4,1 | 10,8 | 9,5 | 10,3 | 2,2 | | 10,0 |
| G | 6,3 | 12,5 | 4,1 | 11,2 | 9,5 | 10,0 | |

Diese Suche nach dem ersten Cluster (Objekt-Paar) erfolgt über die kleinste Distanz. Diese liegt bei 1,4 zwischen B und D. Zwischen diesen Punkten wird ein neuer Mittelpunkt BD gebildet.



Die Koordinaten der neuen Gruppe erfolgt durch $X_{BD} = 1/2 (X_B + X_D)$. $Y_{BD} = 1/2 (Y_B + Y_D)$. Entsprechend gilt für die nächste Gruppe $X_{AEF} = 1/3 (X_A + X_E + X_F)$... Liegt jedoch nur eine Abstandsmatrix vor, so kann der Clustermittelpunkt auch über folgende geometrische Beziehung bestimmt werden:

$$d = \frac{\sqrt{2 \cdot (d_{AE}^2 + d_{AF}^2) - d_{EF}^2}}{2}$$

Die Ergebnisse beider Varianten ergeben jedoch nicht exakt das Selbe, da hier nur näherungsweise die Schwerpunkte getroffen werden.

Der Abstand von E und F beträgt 2,2 und stellt somit die 2. Gruppe dar. Die 3. Gruppe ist bereits eine Zusammenfassung von 3 Punkten AEF. Nach jedem Durchlauf muss die gesamte Tabelle neu aufgebaut werden.

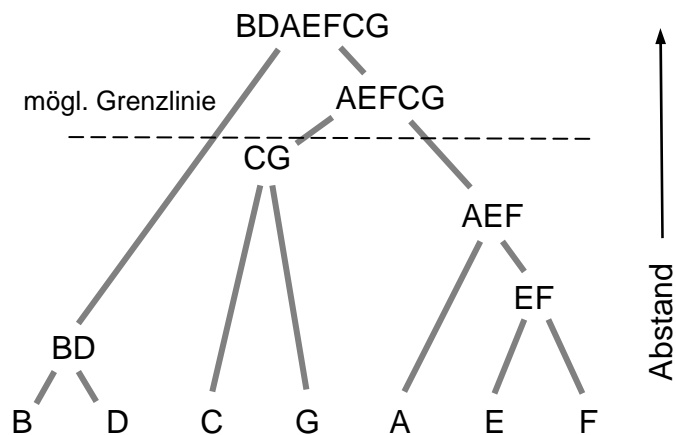
Bei der ersten Zusammenfassung wird der Partner B gelöscht (Werte in Spalte und Zeile). Anstelle von D wird der Titel BD gesetzt und die neuen Abstände mit den verbliebenen Objekten über die angegebene Formel berechnet (fett gedruckte Werte).

Wegen der späteren Auslistung der Cluster ist es vorteilhaft, den Titel des Partners aus der ersten Zeile dem der ersten Spalte voranzustellen. Deshalb hier BD und nicht DB.

| | A | B | C | BD | E | F | G |
|----|-----|---|-----|------|-----|------|------|
| A | | | 5,4 | 8,1 | 3,2 | 4,1 | 6,3 |
| B | | | | | | | |
| C | 5,4 | | | 7,2 | 8,1 | 9,5 | 4,1 |
| BD | 8,1 | | 7,2 | | 8,1 | 10,3 | 11,2 |
| E | 3,2 | | 8,1 | 8,1 | | 2,2 | 9,5 |
| F | 4,1 | | 9,5 | 10,3 | 2,2 | | 10,1 |
| G | 6,3 | | 4,1 | 11,2 | 9,5 | 10,1 | |

Die Tabelle reduziert sich immer weiter, bis nur noch 2 Partner übrig bleiben. Durch dieses Verfahren, kann auf eine rekursive (sich selbst aufrufende Routinen) Vorgehensweise verzichtet werden. Bei einer großen Datenmenge wird allerdings vorausgesetzt, dass die Datenhaltung von mehr als 256 Spalten möglich ist (Excel-Einschränkung!). Die einzelnen Schritte lassen sich als Baumstruktur, auch **Dendrogramm** genannt, verdeutlichen. Von unten nach oben werden die Abstände der Gruppen größer. Führt man das Verfahren immer bis zum Ende, so ist die letzte Gruppe die Zusammenfassung aller.

Man kann bei Festlegung eines Mindestabstandes die Zusammenfassung jedoch abbrechen und erhält eine gewünschte Anzahl Cluster. Im dargestellten Beispiel (mögl. Grenzlinie) gibt es also 4 Cluster. Die letzten beiden Zusammenfassungen werden nicht durchgeführt.



Kategoriale Merkmale lassen sich nicht direkt umsetzen. Denkbar ist hier jedoch eine Überführung in eine numerische Skalierung durch einfaches Durchnummerieren. Die Zuordnung der Merkmale in die Cluster ist dann aber von der Reihenfolge der Ausprägungen abhängig. Besser ist es eigene Spalten (Dimensionen) für jede Ausprägung zu erzeugen. Ist y eine kategoriale Variable, so ist folgende numerische Umwandlung nötig:

| | x | y |
|---|----|---|
| A | 9 | a |
| B | 0 | b |
| C | 7 | c |
| D | 1 | a |
| E | 8 | b |
| F | 10 | c |
| G | 11 | a |

→

| | x | ya | yb | yc |
|---|----|----|----|----|
| A | 9 | 1 | 0 | 0 |
| B | 0 | 0 | 1 | 0 |
| C | 7 | 0 | 0 | 1 |
| D | 1 | 1 | 0 | 0 |
| E | 8 | 0 | 1 | 0 |
| F | 10 | 0 | 0 | 1 |
| G | 11 | 1 | 0 | 0 |