Clusteranalyse

Ziel:

Auffinden von Gruppen („Cluster“) ähnlicher Objekte (bezogen auf die ausgewählten Variablen). Objekte im selben Cluster haben ähnliche Eigenschaften, Objekte in verschiedenen Clustern haben verschiedene Eigenschaften. (Standardmäßig werden bei der Clusteranalyse Objekte gruppiert. Prinzipiell können auch Variablen in Cluster zusammengefasst werden. Technisch ist dazu die Datenmatrix zu transponieren.)

Mathematische Symbole

n	Anzahl der Objekte, Versuchspersonen
m	Anzahl der Variablen, Indikatoren
k	Anzahl der Cluster, Gruppen
(x_{ij})	eine Datenmatrix **X** (mit i=1,2,...,n und j=1,2,...,m) metrischer Variablen
x_j	ein Objekt

Distanz- bzw. Ähnlichkeitsmaße

Um die Ähnlichkeit bzw. Verschiedenheit von Objekten zu beurteilen, muss ein Maß festgelegt werden: beim Clustern von Objekten in der Regel ein Distanzmaß (distance, dissimilarity), beim Clustern von Variablen meistens ein Ähnlichkeitsmaß (correlation, similarity).

Distanzmaße

Alle Variablen müssen dasselbe Skalenniveau haben, das zunächst mal als metrisch angenommen wird. Distanzmaße sind immer von der Skalierung der Variablen abhängig. Dies wird in der Regel durch eine vorherige Standardisierung der Variablen kompensiert. Darüber hinaus hat die Abhängigkeitsstruktur der Variablen einen Einfluss auf das Ergebnis. (Sind z.B. mehrere Variablen hoch korreliert, so wird die Information dieser Variablen bei der Clusterung entsprechend häufig berücksichtigt.) Das kann durch Verwendung des Mahalanobis-Abstand kompensiert werden. Er stellt eine Verallgemeinerung des euklidischen Abstands dar, in dem Sinne, dass die Korrelationen der Variablen berücksichtigt werden.

Als Alternative, u.a. auch für ordinale Variablen, hat sich die *Manhattan-Distanz* (*City-Block-Metrik, taxicab metric*) etabliert. Für dichotome Variablen ist der *Matching Coefficient* üblich. Im folgenden wird der Abstand für 2 Objekte x_{i_1}, x_{i_2} angegeben:

Euklidischer Abstand $$d(x_{i_1}, x_{i_2}) = \sqrt{\sum_j^m (x_{i_1 j} - x_{i_2 j})^2}$$

Mahalanobis-Abstand $$d(x_{i_1}, x_{i_2}) = \sqrt{(x_{i_1 j} - x_{i_2 j})' S^{-1} (x_{i_1 j} - x_{i_2 j})}$$

Manhattan-Abstand $$d(x_{i_1}, x_{i_2}) = \sum_j^m \left| x_{i_1 j} - x_{i_2 j} \right|$$

Matching Coefficient $$d(x_{i_1}, x_{i_2}) = \frac{Anzahl\{x_{i_1 j} \neq x_{i_2 j}\}}{m}$$

Minkowski Abstand $$d(x_{i_1}, x_{i_2}) = \sqrt[r]{\sum_j^m \left| x_{i_1 j} - x_{i_2 j} \right|^r}$$

Tschebyscheff-Abstand $d(x_{i_1}, x_{i_2}) = max\left|x_{i_1 j} - x_{i_2 j}\right|$

Der *Matching Coefficient* lässt sich auch über den Manhattan-Abstand errechnen (abgesehen von der Normierung).

Ähnlichkeitsmaße

Ein triviales und daher auch häufig eingesetztes Ähnlichkeitsmaß ist der Produkt-Moment- oder auch ein anderer Korrelationskoeffizient.

Begriffe

Zentroid:
komponentenweises arithmetisches Mittel aller Objekte eines Clusters

Medoid:
das Objekt eines Clusters, das den minimalen Durchschnitt der Distanzen zu allen anderen Objekten desselben Clusters hat

Abstand von Clustern

Ein Distanz- bzw. Ähnlichkeitsmaße legt nur den Abstand zwischen zwei Objekten fest, damit ist noch nicht erklärt, wie sich der Abstand eines Objekts zu einem Cluster oder der Abstand zwischen zwei Clustern errechnet. Dabei ist erster Fall als Spezialfall des zweiten anzusehen, nämlich wenn ein Cluster aus genau einem Objekt besteht. Hierfür gibt es nun wiederum eine Reihe von Definitionen bzw. Methoden. Hierbei sei angenommen, es liegen zwei Cluster vor: C_1 mit n_1 Objekten und C_2 mit n_2 Objekten.

Single linkage (nearest neighbour)
Es werden alle Objekte aus C_1 mit allen Objekten aus C_2 in Beziehung gesetzt und n_1*n_2 Distanzmaße bestimmt. Der Abstand zwischen C_1 und C_2 wird als Minimum der Distanzen festgelegt.

Complete linkage (furthest neighbour)
Es werden alle Objekte aus C_1 mit allen Objekten aus C_2 in Beziehung gesetzt und n_1*n_2 Distanzmaße bestimmt. Der Abstand zwischen C_1 und C_2 wird als Maximum der Distanzen festgelegt.

Average linkage (linkage zwischen den Gruppen)
Es werden alle Objekte aus C_1 mit allen Objekten aus C_2 in Beziehung gesetzt und n_1*n_2 Distanzmaße bestimmt. Der Abstand zwischen C_1 und C_2 wird als arithmetisches Mittel der Distanzen festgelegt.

linkage innerhalb der Gruppen
Es werden nicht nur alle Objekte aus C_1 mit allen Objekten aus C_2 in Beziehung gesetzt sondern auch alle Paare innerhalb der Cluster C_1 und C_2 . Der Abstand zwischen C_1 und C_2 wird als arithmetisches Mittel der Distanzen festgelegt.

Centroid linkage
Es werden die Zentroide beider Cluster ermittelt. Der Abstand zwischen C_1 und C_2 wird als Distanz der Zentroide festgelegt.

Ward minimum variance linkage
Der Abstand zwischen C_1 und C_2 wird als Zuwachs der Zwischen-Cluster-Streuung festgelegt, wenn C_1 und C_2 zu einem Cluster fusionieren. Rechnerisch lässt er sich aus dem Centroid linkage d_c ableiten: $\frac{n_1 n_2}{n_1 + n_2} d_c$
Single linkage führt zu „Ketten", d.h. langgestreckten Clustern, da ein Objekt nur ein ähnliches in einem Cluster benötigt. Complete linkage führt zu vielen kleinen Clustern, da ein Objekt zu allen Objekten eines Clusters ähnlich sein muss. Bevorzugt werden Average linkage und Ward linkage.

Arten/Methoden von Clusteranalysen:

***Partitionierende Verfahren*:**

Es wird eine Clusterzahl vorgegeben. Jedes Objekt wird genau einem Cluster zugeordnet.

Das bekannteste Verfahren ist die K-Means-Clusteranalyse. Beginnend von einer Zufallspartition werden iterativ die Objekte genau dem Cluster zugeordet, zu dem es den kleinsten Abstand hat, wobei die Centroid linkage angewandt wird. Da sich dadurch die Cluster ändern, muss dieser Prozess so lange wiederholt werden, bis sich keine Verbesserung durch die Umordnung eines Objekts erreichen lässt. Die K-Means-Clusteranalyse basiert auf der euklidischen Distanz. (S-Plus-Kommando: `kmeans` , SPSS: `Clusterzentrenanalyse`)

Ähnliches funktioniert das Medoid-Verfahren. Einziger Unterschied: anstatt die Abstände der Objekte zu den Zentroiden zu berechnen, werden hier die Abstände zu einem Repräsentanten der Cluster („Medoid") ermittelt. Auch hier beginnt der Algorithmus mit einer Zufallspartition. Dann wird iterativ geprüft, ob durch Zuordnung von Objekten zu anderen Clustern die durchschnittlichen Distanzen innerhalb der Cluster reduziert werden können. (S-Plus-Kommandos: `pam`, `clara`)

***Agglomerative hierarchische Verfahren*:**

Im 1. Schritt bildet jedes Objekt ein eigenes Cluster, so dass zu Beginn *n* Cluster - *n* sei die Anzahl der Objekte - existieren. Anschließend werden in (*n-1*) Schritten jeweils die beiden Cluster zu einem neuen vereinigt, die den kleinsten Abstand haben, bis schließlich im letzten Schritt nur noch ein „entartetes" Cluster aller Objekte besteht. Das heißt, es gibt Lösungen für jede beliebige Clusterzahl. Allerdings sind diese, insbesondere bei größerer Fallzahl, suboptimal hinsichtlich der erzielbaren Homogenität der Cluster. Daher wird häufig eine Mischung aus agglomerativ hierarchischen und dem K-Means-Verfahren gewählt, zumal dadurch der Iterationsprozess des letzteren deutlich abgekürzt werden kann. (S-Plus-Kommandos: `agnes` und `hclust`, SPSS: `hierarchische Cluster`)

Durch die verschiedenen Definitionen für den Abstand von Clustern existieren dementsprechend viele agglomerative hierarchische Verfahren.

***Divisive hierarchische Verfahren*:**

Diese sind praktisch die Umkehrung der letzten Klasse von Verfahren. D.h. die Gesamtheit aller n Objekte wird zunächst geteilt, dann zukzessive eines der bereits bestehenden Cluster. Kriterium ist der größtmögliche Abstand, der beiden zu bildenden Cluster. (S-Plus-Kommando: `diana`)

Monothetische hierarchische Verfahren:

Diese haben eine Ähnlichkeit mit den divisiven hierarchischen Verfahren. Allerdings müssen hierbei die Variablen binär (dichotom) sein. Im ersten Schritt wird diejenige der m Variablen gesucht, die den Datensatz so teilt, dass der Abstand zwischen den beiden entstandenen Clustern maximal wird. Iterativ werden die Cluster dann anhand der übrigen Variablen weitergeteilt. Dabei gibt es hier zwei Varianten: zum einen wird die Variable gesucht, die alle bestehenden Cluster am besten teilt, zum anderen kann die teilende Variable für jedes Cluster eine andere sein. (S-Plus-Kommando: `mona`)

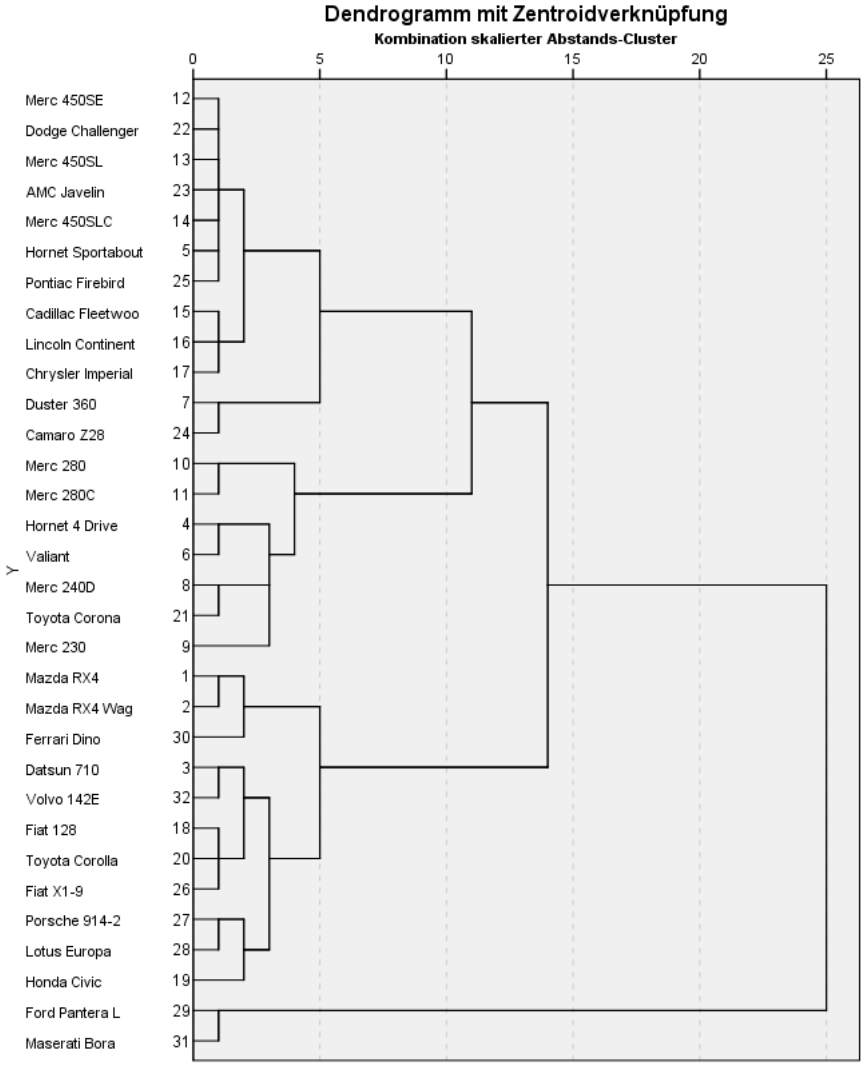
Fuzzy-Clustering, Verfahren mit unscharfen Partitionen:

Bei allen bisher aufgeführten Verfahren wurden die Objekte immer genau einem Cluster zugeordnet. Bei dieser Klasse von Verfahren wird bei der Lösung für jedes Objekt und jedes Cluster die Zugehörigkeitswahrscheinlichkeit errechnet, d.h. die Wahrscheinlichkeit, mit der ein Objekt zu einem bestimmten Cluster gehört. Das bekannteste in dieser Kategorie ist der *Normix-Algorithmus*: Bei *k* angestrebten Clustern werden die Daten als eine Mischung von *k* m-dimensionalen Normalverteilungen angenommen, d.h. die Objekte jedes Cluster werden als eine Stichprobe aus einer multivariaten Normalverteilung (mit verschiedenen Lageparametern) angenommen. Es werden dann die Parameter der *k* Normalverteilungen geschätzt und anschließend die Wahrscheinlichkeiten ausgerechnet. (S-Plus-Kommando: `fanny`)

Graphiken

Für agglomerative oder divisive hierarchische Verfahren gibt es als Standarddarstellung die Baum-Diagramme, die die sukzessive Vereinigung der Objekte und der Cluster miteinander

veranschaulichen.



In S-Plus ermöglichen einige Verfahren, so u.a. `pam` und `clara` (Medoid-Verfahren) sowie `fanny` (Fuzzy-Clustering), die Ausgabe einer Punktewolke der ersten beiden Hauptkomponenten, in der die Clusterzugehörigkeiten durch verschiedene Symbole gekennzeichnet sind und ein Ellipsoid die Konturen der Cluster widerspiegelt. Die entsprechende Funktion ist `clusplot`.

Alternativ können für Partitionen die Clusterzugehörigkeiten der einzelnen Objekte als zusätzliche Variable gespeichert werden. Darüber lassen sich Punktewolken erstellen, worin die Objekte verschiedener Cluster in verschiedenen Farben dargestellt werden.

In SPSS muss dieses in einzelnen Schritten selbst vorgenommen werden: Berechnung der Hauptkomponenten, Durchführung der Clusteranalyse mit Speichern der Clusterzugehörigkeit als neue Variable und schließlich Erstellen der Punktewolke der ersten beiden Hauptkomponenten mit Kennzeichnung der Clusterzugehörigkeit.