Hierarchische Clusteranalyse

Probleme der hierarchischen Clusteranalyse bei qualitativen Merkmalen

Abdurahman Maarouf[1]

[1]Universität Bonn

14. Juli 2017

Inhaltsverzeichnis

Inhaltsverzeichnis

Problemstellung

Ähnlichkeiten erkennen

- Untersuchung von Ähnlichkeiten unter den Untersuchungsobjekten
  Untersuchungsobjekte: Personen, Unternehmen, Produkte, etc.
- Sinnvolle Gruppierung anhand der Merkmalsvariablen
  Merkmalsvariablen: Geschlecht, Alter, Größe, Wohnort, etc.
- Anwendung in vielen Bereichen (z.B. Medizin, Soziologie, Biologie, Wirtschaftswissenschaften, Alltag)

Ziel und Funktion einer Clusteranalyse

Clusteranalyse als exploratives Verfahren der multivariaten Datenanalyse

- Gruppen sind im Ausgangspunkt unbekannt
- Gruppen werden erst durch das Clusterverfahren herbeigeführt
- Hierbei werden alle Merkmalsvariablen der Untersuchungsobjekte gleichzeitig berücksichtigt
- Strukturierung und Gruppierung der Untersuchungsobjekte im Hinblick auf folgende Kriterien:

Kriterium 1: Homogenität innerhalb der Klassen

Objekte in einer Klasse sollen möglichst ähnlich sein

Kriterium 2: Heterogenität zwischen den Klassen

Verschiedene Klassen sollen möglichst unterschiedliche Objekte enthalten

Aufbau einer Clusteranalyse

Partitionierendes vs. Hierarchisches Clustering

Partitionierendes Clustering:

- Annahme einer gegebenen Gruppierung und einer festen Anzahl an Gruppen
- Objekte werden den Gruppen so zugeordnet, dass eine gegebene Zielfunktion ihr Optimum erreicht

Hierarchisches Clustering:

- Agglomerative Algorithmen: Zusammenfassung der Objekte in Gruppen
- Divisive Algorithmen: Aufteilung der Gesamtheit in Gruppen

Aufbau des hierarchischen Clustering

1. Bestimmung der Ähnlichkeit

2 Objekte werden auf Ähnlichkeiten/Unterschiede untersucht.
Ähnlichkeiten/Unterschiede werden mit einem Proximitätsmaß gemessen.

2. Auswahl des Fusionierungsalgorithmus

Objekte werden anhand ihrer Proximitätsmaße zu Gruppen zusammengefasst (agglomerativ).
Dabei werden Kriterien 1 und 2 so weit wie möglich erfüllt.

3. Bestimmung der Clusteranzahl

Welche Anzahl an Clustern ist die „beste“ Lösung?
Handhabbarkeit vs Homogenitätsanforderung

Inhaltsverzeichnis

Definition

Ähnlichkeits- und Distanzmaße

Ähnlichkeitsmaße:

- Je größer der Wert desto ähnlicher sind die zwei Objekte

Distanzsmaße:

- Je größer der Wert desto unähnlicher sind die zwei Objekte
- Falls Objekte vollkommen identisch sind liegt der Wert bei 0

Wahl des Proximitätsmaßes hängt von dem Skalenniveau des Merkmals ab. Hier werden Variablen mit binären (Ähnlichkeitsmaß) und nominalen (Distanzmaß) Skalenniveaus betrachtet.

Ähnlichkeitsfunktionen bei binären Merkmalen

Von Rohdatenmatrizen zu Ähnlichkeitsmatrizen

Rohdatenmatrix:

$$\begin{array}{c} \\ Objekt_1 \\ Objekt_2 \\ Objekt_3 \\ Objekt_4 \end{array} \begin{array}{c} \begin{array}{ccc} Merkmal_1 & Merkmal_2 & Merkmal_3 \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{pmatrix} \end{array}$$

wird transformiert zu:

Ähnlichkeitsmatrix:

$$\begin{array}{c} \\ Objekt_1 \\ Objekt_2 \\ Objekt_3 \\ Objekt_4 \end{array} \begin{array}{c} \begin{array}{cccc} Objekt_1 & Objekt_2 & Objekt_3 & Objekt_4 \end{array} \\ \begin{pmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{pmatrix} \end{array}$$

wobei S_{ij} den Ähnlichkeitswert der Objekte i und j darstellt.

Berechnung des Ähnlichkeitswerts zweier Objekte

Allgemeines Ähnlichkeitsmaß:

$$S_{ij} = \frac{a + \delta d}{a + \delta d + \gamma(b + c)} \tag{1}$$

wobei:

- S_{ij}: Ähnlichkeitswert der Objekte i und j
- δ/γ : mögliche Gewichtungsfaktoren (abhängig vom Ähnlichkeitskoeffizienten)
- a: Anzahl der Merkmale, die bei beiden Objekten vorhanden sind
- b: Anzahl der Merkmale, die nur bei Objekt j vorhanden sind
- c: Anzahl der Merkmale, die nur bei Objekt i vorhanden sind
- d: Anzahl der Merkmale, die bei beiden Objekten nicht vorhanden sind

gilt.

Ähnlichkeitskoeffizienten bei binären Merkmalen

1. Jaccard-Koeffizient: $\delta = 0$; $\gamma = 1$

$$S_{ij} = \frac{a}{a + b + c} \tag{2}$$

2. M-Koeffizient: $\delta = 1$; $\gamma = 1$

$$S_{ij} = \frac{a + d}{M}; M = a + b + c + d \tag{3}$$

3. Russel und Rao-Koeffizient:

$$S_{ij} = \frac{a}{M}; M = a + b + c + d \tag{4}$$

Auswahl des Ähnlichkeitskoeffizienten

- Kein Ähnlichkeitskoeffizient allgemeingültig vorziehbar
- Alle drei Ähnlichkeitskoeffizienten gleich, falls $d = 0$
- Falls $d > 0$ gilt, ist der M-Koeffizient am größten und der RR-Koeffizient am kleinsten
- Bedeutung von „Merkmal nicht vorhanden“ bestimmt die Auswahl des Ähnlichkeitskoeffizienten
- M-Koeffizient: Wenn „Merkmal nicht vorhanden“ die gleiche Aussagekraft hat wie „Merkmal vorhanden“
- Jaccard-Koeffizient: Wenn „Merkmal nicht vorhanden“ nicht die gleiche Aussagekraft hat (Bsp: Deutsch vs. Nicht-Deutsch)

Ähnlichkeits-/ Distanzfunktionen bei nominalen Merkmalen

Variante 1: Transformation in binäre Variable

- Nominale Merkmale können in binäre Hilfsvariablen zerlegt werden
- (Deutsch, Französisch, Spanisch,...) $\Rightarrow$ (Deutsch, Nicht-Deutsch)
- Restliche Vorgehensweise gleich wie bei binären Variablen
- Achtung: Je höher die Anzahl an Merkmalsausprägungen desto verzerrter die binären Proximitätsmaße!
  $\Rightarrow$ Genauere Ergebnisse mit Variante 2: Analyse von Häufigkeiten

Variante 2: Analyse von Häufigkeiten

Rohdatenmatrix:

$$\begin{array}{c} \\ Objekt_1 \\ Objekt_2 \\ Objekt_3 \\ Objekt_4 \end{array} \begin{array}{c} \begin{array}{ccc} Merkmal_1 & Merkmal_2 & Merkmal_3 \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{pmatrix} \end{array}$$

wird transformiert zu:

Distanzmatrix:

$$\begin{array}{c} \\ Objekt_1 \\ Objekt_2 \\ Objekt_3 \\ Objekt_4 \end{array} \begin{array}{c} \begin{array}{cccc} Objekt_1 & Objekt_2 & Objekt_3 & Objekt_4 \end{array} \\ \begin{pmatrix} D_{11} & D_{12} & D_{13} & D_{14} \\ D_{21} & D_{22} & D_{23} & D_{24} \\ D_{31} & D_{32} & D_{33} & D_{34} \\ D_{41} & D_{42} & D_{43} & D_{44} \end{pmatrix} \end{array}$$

wobei D_{ij} den Distanzwert der Objekte i und j darstellt.
Distanzwerte werden mithilfe des χ^2-Maß berechnet.

Inhaltsverzeichnis

Auswahl des Fusionierungsalgorithmus

Agglomerative hierarchische Algorithmen

- Gewonnene Distanz- bzw. Ähnlichkeitsmatrix bilden den Ausgangspunkt der Clusteralgorithmen
- Breites Spektrum an Clusteralgorithmen
- Schwerpunkt auf agglomerative hierarchische Algorithmen, da sie in der Praxis häufig zur Anwendung kommen
- Cluster werden anhand der Ähnlichkeits- bzw. Distanzwerte erstellt.
- Im Ausgangspunkt stellt jedes Objekt ein Cluster dar

Hierarchische Verfahren

Ablauf der agglomerativen Verfahren

Schritt 1

Ausgangssituation: Jedes Objekt stellt ein Cluster da

Schritt 2

Für alle Objekte werden paarweise Ähnlichkeits- bzw. Distanzwerte bestimmt

Schritt 3

Die beiden Cluster mit dem größten Ähnlichkeitswert bzw. kleinsten Distanzwert werden gesucht

Ablauf der agglomerativen Verfahren

Schritt 4

Die beiden Cluster mit dem größten Ähnlichkeitswert bzw. kleinsten Distanzwert werden zu einem neuen Cluster zusammengefasst

Schritt 5

Neue Ähnlichkeits- bzw. Distanzwerte werden mit den übrigen Gruppen berechnet

$\Rightarrow$ reduzierte Ähnlichkeits- bzw. Distanzmatrix

Schritt 6

Schritt 3 bis Schritt 5 werden solange wiederholt, bis es nur ein Cluster gibt

$\Rightarrow$ Ein-Cluster-Lösung

Zu Schritt 5: Berechnung der neuen Distanzwerte

Schritt 5

Neue Ähnlichkeits- bzw. Distanzwerte werden mit den übrigen Gruppen berechnet

$\Rightarrow$ reduzierte Ähnlichkeits- bzw. Distanzmatrix

- Zusammenfassung von Cluster X und Cluster Y zu Cluster X+Y
- Neuer Distanzwert von Cluster X+Y und Cluster Z:

$$D(Z; X+Y) = A \cdot D(Z; X) + B \cdot D(Z; Y) + E \cdot D(X; Y) + G \cdot |D(Z; X) - D(Z; Y)| \quad (5)$$

Zu Schritt 5: Berechnung der neuen Distanzwerte

$$D(Z; X+Y) = A \cdot D(Z; X) + B \cdot D(Z; Y) + E \cdot D(X; Y) + G \cdot |D(Z; X) - D(Z; Y)|$$

mit

$$D(I; J) = \text{Distanz zwischen den Clustern I und J}$$

- A,B,E und G sind Konstanten
- Sie werden vom agglomerativen Verfahrensalgorithmus bestimmt

Agglomerative Verfahrensalgorithmen

1. Single-Linkage-Verfahren: $A = 0.5$; $B = 0.5$; $E = 0$; $G = -0.5$

$$D(Z; X+Y) = 0.5 \cdot (D(Z; X) + D(Z; Y) - |D(Z; X) - D(Z; Y)|) \quad (6)$$

2. Complete-Linkage-Verfahren: $A = 0.5$; $B = 0.5$; $E = 0$; $G = 0.5$

$$D(Z; X+Y) = 0.5 \cdot (D(Z; X) + D(Z; Y) + |D(Z; X) - D(Z; Y)|) \quad (7)$$

3. Average-Linkage-Verfahren: $A = 0.5$; $B = 0.5$; $E = 0$; $G = 0$

$$D(Z; X+Y) = 0.5 \cdot (D(Z; X) + D(Z; Y)) \quad (8)$$

Dendrogramm

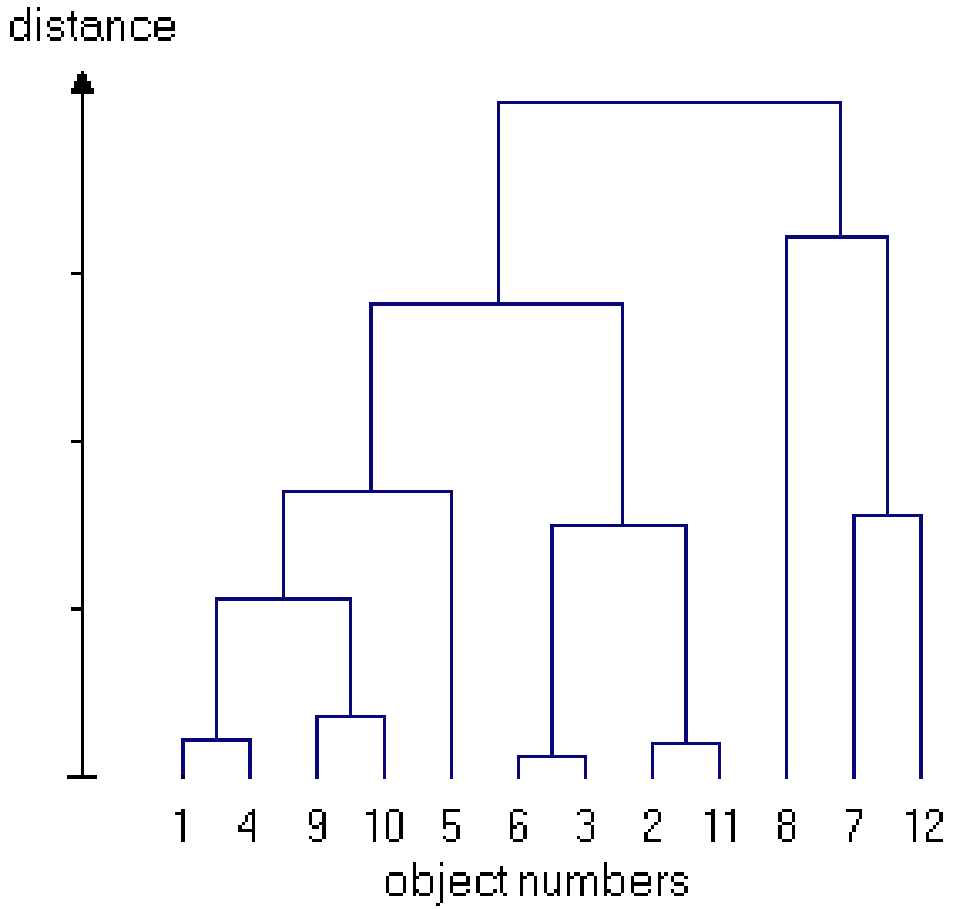

Abbildung 1

Inhaltsverzeichnis

Spezielle Probleme

Probleme der praktischen Durchführung

1. Große Anzahl an Untersuchungsobjekten erschweren die Durchführung einer hierarchischen Clusteranalyse
2. Große Anzahl an Merkmalsvariablen erschwert die Interpretation der Ergebnisse
3. Bei Datenauswertungen fehlen häufig Daten oder sind ungültig

Handhabbarkeit vs. Homogenitätsanforderung

- Nach dem agglomerativen Verfahren muss die „beste“ Anzahl von Clustern bestimmt werden
- Einerseits Erfüllung der Homogenitätsanforderung
- Andererseits Maximierung der Handhabbarkeit
- Beide Anforferung gleichzeitig zu erfüllen fällt schwer
  $\Rightarrow$ Elbow-Kriterium

Bestimmung der Clusteranzahl

Elbow-Kriterium

Optische Identifikation eines „Sprungs"(Elbow) im Dendrogramm oder im Scree-Plot:

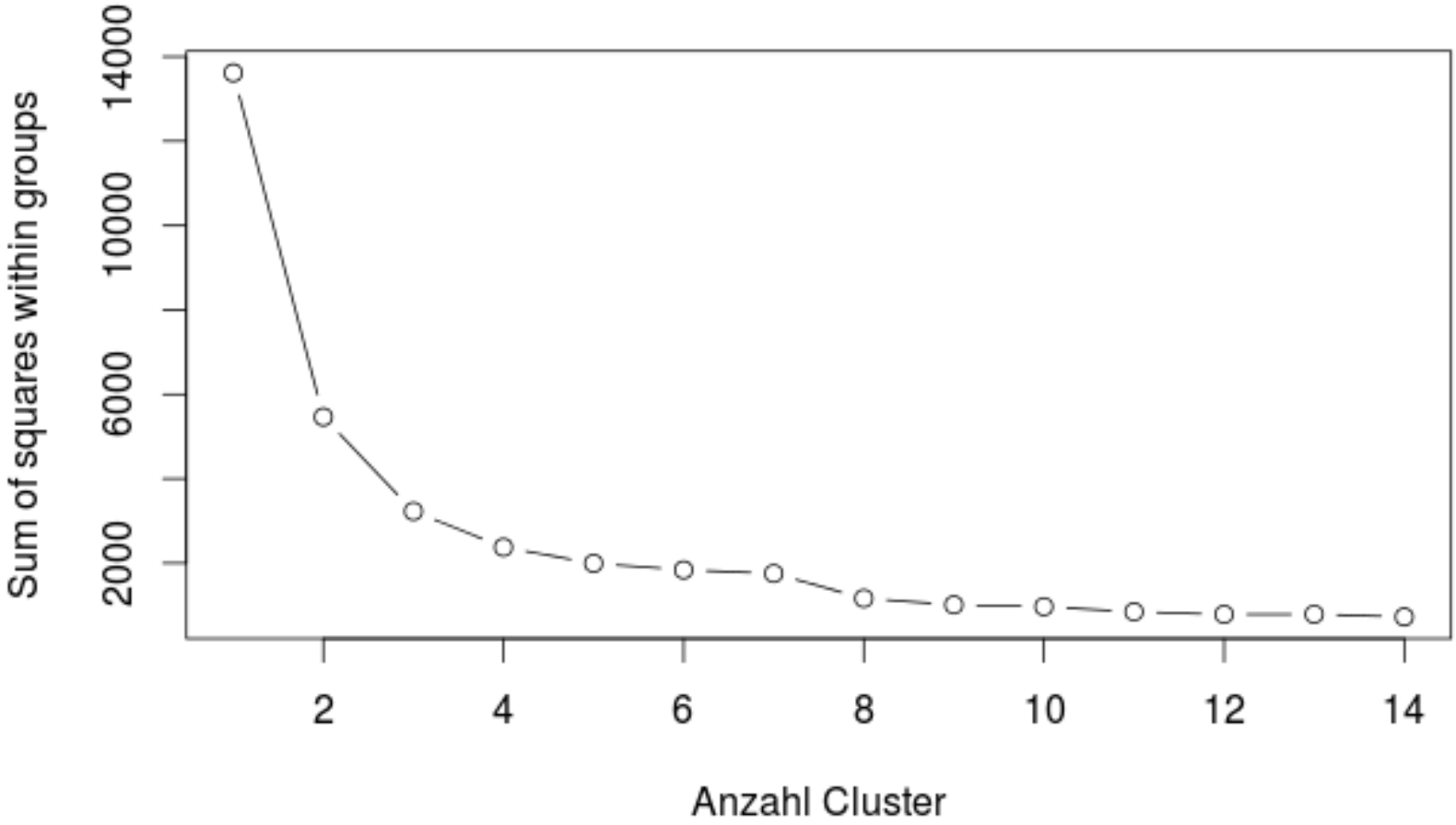

Abbildung 2

Quellen

1. Detlef Steinhausen:„Clusteranalyse, Einführung in Methoden und Verfahren der automatischen Klassifikation", 1977
2. Joachim Hartung:„Multivariate Statistik: Lehr- und Handbuch der angewandten Statistik", 1995
3. Klaus Backhaus:„Multivariate Analysemethoden: eine anwendungsorientierte Einführung", 2016
4. Chris Fraley:„How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis", 1998
5. Michael Wiedenbeck: „Klassifikation mit Clusteranalyse : grundlegende Techniken hierarchischer und K-means-Verfahren", 2001

Abbildungen:

1. Abbildung 1:
   $http://www.statistics4u.info/fundstat_eng/img/hl_dendrogram.png$
2. Abbildung 2:
   https://kamihoeferl.wordpress.com/2014/05/01/clusteranalyse-mit-r-ech

Vielen Dank für Ihre Aufmerksamkeit!