RUB

Beispiele für Clusteranalyse

11.01.2011

LS Methodenlehre & Sozialstatistik

Christian Dudel

Beispiel 1

Allgöwer, A. et al. (2000): Wie gesund leben Studierende? in: Sonntag, U. et al. (Hrsg.): Gesundheitsfördernde Hochschulen. Konzepte, Strategien und Praxisbeispiele. Weinheim: Juventa

RUB

Beispiel 1: Daten

- 650 Erstsemester (WS 1995/1996) der Universität Bielefeld
- Diverse Fragen zu „Gesundheit“ allgemein
- Beispiel:
  Frage zum Gesundheitsbewusstsein „Wie stark achten Sie auf Ihre Gesundheit?“
  Antwortmöglichkeiten: Sehr, eher mehr, eher weniger, gar nicht
- Zudem Fragen zum Ernährungsbewusstsein, Ernährungsverhalten, Bewegungsbewusstsein, Bewegungsverhalten, Alkoholkonsum, Zigarettenkonsum, Haschischkonsum, Zahnvorsorge, Impfungen

Beispiel 1: Auswertung

- Variablen z-standardisiert
- Hierarchische Clusteranalyse nach Ward
- Kriterium für Clusterwahl: Fehlerquadratsumme (max. 6 Cluster gesetzt)
- k-means, um Partitionierung zu erhalten

Beispiel 1: Ergebnisse

5 Cluster

- „Die moderat Gesundheitsbewussten“
- „Die Sportler“
- „Die Präventionsscheuen“
- „Die gesundheitlich Desinteressierten“
- „Die Drogenkonsumenten“

Beispiel 2

Rüb, F., Werner, D. (2007): Typisierung von SGB II-Trägern. IAB Forschungsbericht 1/2007

Beispiel 2: Fragestellung

„Controlling“ von SGB II-Trägern: Bewertung des „Arbeitsmarkterfolges“ (Integrationsquote etc.)

Dabei Problem: Starke regionale Differenzen am Arbeitsmarkt, die Vermittlungserfolge beeinflussen

Idee: Bewertung nach Typisierung der Bedingungen eines Trägers

Beispiel 2: Daten & Auswertung

- Aus dem Jahr 2006
- 442 SGB II-Träger
- Variablen: Arbeitslosenquote, Bevölkerungsdichte, Ausländeranteil SGB II, Saisondynamik, BIP pro Kopf, SGB II-Kundenquote, Umgebungsvariable
- z-standardisiert
- Hierarchische Clusteranalyse nach Ward
- k-means

Beispiel 2: Ergebnisse

Insgesamt 12 Typen, beispielsweise:

- Gelsenkirchen, Bochum, Essen, Dortmund, Duisburg, Oberhausen, Wuppertal, ...: „Städte in Westdeutschland mit unterdurchschnittlicher Arbeitsmarktlage und sehr hohem Anteil an Langzeitarbeitslosen“
- „Ländliche Gebiete in Westdeutschland mit guter Arbeitsmarktlage und hoher saisonaler Dynamik“
- ...

Beispiel 3

Fiori, K. et al. (2008): Profiles of social relations among older adults: a cross-cultural approach. in: Ageing & Society 28, S. 203-231

Beispiel 3: Fragestellung

1. Was für „Profile“ sozialer Beziehungen finden sich bei älteren Menschen (in Japan/USA)?
2. Unterscheiden sich diese Profile zwischen den USA und Japan (kulturelle Unterschiede)?
3. Haben Profile Einfluss auf körperliche und geistige Gesundheit?

Profile sozialer Beziehungen $\approx$ Umfang, Struktur und Funktionalität sozialer Netzwerke

Beispiel 3: Fragestellung

In anderen Studien zumeist vier Netzwerktypen:

1. diverse network type
2. socially-isolated network type
3. friend-focused network type
4. family-focused network type

Beispiel 3: Daten

- *Social Relations and Mental Health Over the Life Course*
- Stichprobenziehung in Detroit und Yokohama, 1991 bis 1993
- Bei Auswertung nur Personen im Alter von 60+ berücksichtigt
- 514 Personen in USA, 491 Personen in Japan
- 2005 Erhebung, ob Befragte noch leben (nur USA)

Beispiel 3: Daten

Umfang soziale Netzwerke erfasst über:

- Familienstand (verheiratet/nicht-verheiratet)
- Netzwerkgröße (Personen auflisten mit Beziehung)
- Räumliche Nähe der Netzwerkmitglieder (Anteil an Personen die unter einer Stunde Fahrtzeit erreicht werden können)
- Durchschnittliche Kontakthäufigkeit zu Verwandten (0=nie bis 5=jeden Tag)
- Durchschnittliche Kontakthäufigkeit zu Freunden (0=nie bis 5=jeden Tag)

Beispiel 3: Daten

Funktionalität soziale Netzwerke erfasst über:

- Anteil an Personen im „ersten Kreis“
- Indikator für instrumentelle Unterstützung
- Indikator für emotionale Unterstützung
- Indikator für „negative quality“

Beispiel 3: Vorgehen

1. Variablen standardisiert
2. Auswertungen getrennt für USA und Japan
3. Hierarchische Klassifikation nach Ward (SAS)
4. Clusterwahl über die von SAS ausgegebenen Kriterien
5. k-means, um Ergebnis zu erhalten

Beispiel 3: Vorgehen

Wieviele Cluster?

1. In Literatur 4 Typen („robustes Ergebnis“): Clusterzahl sollte nicht weit davon entfernt sein
2. 3 Kennwerte aus SAS (Pseudo-F Statistik, Pseudo-t^2-Statistik, Sarles *cubic clustering criterion*) für hierarchische Klassifikation
3. Interpretation (k-means)

Beispiel 3: Ergebnis

Für USA 6 Netzwerktypen, für Japan 5 Netzwerktypen gefunden

USA: „Diverse/extensive“, „Friend-focused/supported“, „Friend-focused/unsopported“, „Family-focused/negative“, „Structurally restricted“, „Funtionally Restricted“

Japan: „Diverse/extensive“, „Friend-focused“, „Family-focused/close“, „Married and distal“, „Restricted/unsupported“

Beispiel 3: Interpretation

Beispiel: USA (nur 2 Typen)

	Diverse	Structurally restricted
Prop. married	0.99	0.00
Tot. net. size	10.96	8.3
Prop. proximity	0.72	0.63
Contact family	3.83	3.71
Contact friends	1.17	0.07

Beispiel 3: Weitere Ergebnisse

Einfluss Netzwerktypus auf körperliche/geistige Gesundheit

1. CES-Depression Scale: Min. 0, Max. 60 (20 Fragen nach Symptomen in der letzten Woche, Antwortmöglichkeiten meistens 0=nie bis 3=immer)
2. Subjektive Gesundheitseinschätzung (1=hervorragend bis 5=sehr schlecht)
3. Zahl der chronischen Krankheiten

Beispiel 3: Weitere Ergebnisse

- MANCOVA
- Kontrollvariablen: Alter, Geschlecht, Bildung, „race“
- Netzwerktypus als erklärende Variable (Dummy-Kodierung)
- USA: Signifikanter Einfluss Netzwerktyp auf körperliche/geistige Gesundheit („Functionally restricted“ jeweils am schlechtesten)
- Japan: kein signifikanter Zusammenhang

Beispiel 3: Fazit

- 4 typische Netzwerkarten
- Aber auch kulturelle Differenzen (länderspezifische Typen)
- Kulturell vermittelter Effekt von Netzwerken auf Gesundheit

Beispiel 3: Fazit

Mögliche Probleme

- Response Bias Japan: Vermeidung von Extremwerten
- Stichprobe eingeschränkt (lediglich zwei Städte): bspw. keine Landbevölkerung
- Problem der Kausalität: sind Menschen mit kleinen Netzwerken eher depressiv oder haben depressive Personen eher kleine Netzwerke?
- Indikatoren für Funktionalität der Netzwerke nicht differenziert analysiert (bspw. von wem Hilfe gegeben wird)

Beispiel 3: Weitere Probleme

- Subjektive Gesundheitseinschätzungen *extrem* ungenau
- Depression Scale-Ansätze ebenfalls sehr problematisch (Artefakte)
- Stabilität der Clusterlösung? (bspw. k-medoids anstelle von k-means, anderer Fusionierungsalgorithmus bei hierarchischer Klassifikation)
- „Härte“ der Klassifikation? (vielleicht eher fuzzy?)
- MANCOVA: Wenig Kontrollvariablen (Einkommen/ökonomische Situation, Erwerbsstatus, konkrete Wohnsituation, ...)