planung &analyse

Zeitschrift für Marktforschung und Marketing www.planung-analyse.de 5/2015 D11700F

Eine Marke der dfv Mediengruppe

Schwerpunkt

Kunden-zufriedenheit

© etherlore / Flickr

Special

Mobilitäts-forschung

Branche

Frontrunner gesucht

Kongress

Marke in Bewegung

Statistik KOMPAKT

Latent-Class-Clusteranalyse

Klassische Verfahren der Clusteranalyse weisen ein Objekt eindeutig einem Cluster zu. Ergebnis einer Latent-Class-Clusteranalyse sind dagegen Wahrscheinlichkeiten, mit denen Objekte den einzelnen Clustern zugeordnet werden. Es wird davon ausgegangen, dass latente (nicht beobachtbare) Klassen für Unterschiede in den Daten mit verantwortlich sind.

Einführungsbeispiel

In einer Befragung zur Ermittlung der Präferenzen beim Reisen wurde unter anderem gefragt
1. wie Reisen gebucht werden: (a) überwiegend Online / (b) überwiegend Reisebüro / (c) Online & Reisebüro,
2. welche Arten von Reisen in Frage kommen: (a) Bildungsreise / (b) Fernreise / (c) Strandurlaub.

Grundsätzlich sind kategoriale Variablen (auch gemeinsam mit Variablen anderer Skalenniveaus) in einer Latent-Class-Clusteranalyse bei Verwendung geeigneter Software einfach zu handhaben. Frage (1) ist eine Einfachnennung und kann mit den Codes 1, 2, und 3 direkt in die Latent-Class-Clusteranalyse übernommen werden. Frage (2) ist eine Mehrfachnennung und somit mit drei dichotomen Variablen zu erfassen. Auf Basis von 1222 Befragten zeigten sich drei Cluster als beste Lösung. In Abbildung 1 sind beispielhaft die individuellen Zuordnungswahrscheinlichkeiten zu den drei Clustern für drei Personen dargestellt, die sich gemäß ihrer Antwortmuster ergeben. Während Person 1 nahezu eindeutig zu Cluster 1 zählt, ist die Zuordnung bei den beiden anderen Befragten nicht ganz so deutlich.

i	y^i				$p(g=1\|y^i)$	$p(g=2\|y^i)$	$p(g=3\|y^i)$
Person	Frage 1	Frage 2a	Frage 2b	Frage 2c	Cluster 1	Cluster 2	Cluster 3
1	1	0	1	1	0.86	0.04	0.10
2	1	1	0	1	0.29	0.14	0.57
3	2	1	0	1	0.49	0.07	0.44

Abbildung 1: Zuordnungswahrscheinlichkeiten

Zur Beschreibung der drei Cluster wird die mit den individuellen Zuordnungswahrscheinlichkeiten gewichtete Häufigkeitsverteilung der Variablen je Cluster ermittelt (siehe Abbildung 2). Cluster 1 ist beispielsweise charakterisiert durch Strandurlauber, die teilweise auch Fernreisen unternehmen und im Vergleich zu den anderen Gruppen am häufigsten im Reisebüro buchen.

Person	Gesamt	Cluster 1	Cluster 2	Cluster 3
Online	0.50	0.45	0.47	0.59
Reisebüro	0.13	0.19	0.06	0.11
Online & Reisebüro	0.37	0.36	0.47	0.30
Bildungsreise	0.56	0.19	0.97	0.66
Fernreise	0.65	0.60	0.95	0.41
Strandurlaub	0.80	0.99	0.92	0.41

Abbildung 2: Profile der Cluster

Grundidee des Verfahrens

Eine Latent-Class-Clusteranalyse liefert genau genommen bedingte Wahrscheinlichkeiten: unter der Bedingung der von einer Person gegebenen Antworten auf die für die Clusteranalyse relevanten Fragen (formal: die in dem Vektor y^i gesammelten Werte der entsprechenden Variablen für die Person i) zählt sie mit der Wahrscheinlichkeit $p(g = 1 \mid y^i)$ zu Cluster 1, mit der Wahrscheinlichkeit $p(g=2 \mid y^i)$ zu Cluster 2, etc. „Umgekehrt" gibt $p(y^i|g)$ die bedingte Wahrscheinlichkeit an, dass eine Person bestimmte Antworten gibt, unter der Bedingung, dass sie Cluster g angehört. Mit dieser lässt sich die Wahrscheinlichkeit $p(y^i)$ berechnen, die Antworten y^i zu beobachten. Ist p(g) der Anteil eines Clusters g an der Stichprobe, dann gilt nach dem Satz der totalen Wahrscheinlichkeit

$$p(y^i) = \sum_g p(g) \cdot p(y^i \mid g)$$

Die bedingte Wahrscheinlichkeit $p(y^i|g)$ ist bestimmt durch Wahrscheinlichkeitsverteilungen, die für die einzelnen Variablen in Abhängigkeit vom Skalenniveau angenommen werden, und die Parameter, die diese Verteilungen definieren. Die Wahrscheinlichkeit, genau die Antworten aller n Personen einer Stichprobe zu beobachten, ist dann gegeben durch die Likelihood-Funktion

$$L = p(y^1) \cdot p(y^2) \cdot \cdots \cdot p(y^n)$$

Die Parameter der Verteilungen sowie die Anteile p(g) werden so geschätzt, dass die Likelihood-Funktion maximal wird. Über das Bayes-Theorem lässt sich dann die gesuchte Wahrscheinlichkeit berechnen, dass eine Person i einem Cluster, zum Beispiel g=1 angehört:

$$p(g = 1 \mid y^i) = \frac{p(y^i \mid g = 1) \cdot p(g = 1)}{p(y^i)}$$

Die Schätzung der Likelihood-Funktion wird für verschiedene Clusteranzahlen wiederholt. Anhand von Informationskriterien wie dem Akaike Information Criterion (AIC) oder dem Bayesian Information Criterion (BIC) wird die Auswahl der Clusterlösung getroffen. Je näher der Wert der Likelihood-Funktion an 1 bzw. der Wert der logarithmierten Likelihood-Funktion an 0 liegt, desto besser ist die Schätzung. Da die Informationskriterien auf dem maximalen Wert der logarithmierten Likelihood-Funktion basieren, wird die Clusteranzahl gewählt, für die die Informationskriterien 0 am nächsten kommen.

Johannes Lüken und **Prof. Dr. Heiko Schimmelpfennig**, Experten für Multivariate Analysen bei IfaD, Institut für angewandte Datenanalyse.

In Ausgabe 6/2015: Diskriminanzanalyse

► Literatur

Bacher, J., Vermunt, J.K.: Analyse latenter Klassen. In: Wolf, C., Best, H.: Handbuch der sozialwissenschaftlichen Datenanalyse, Wiesbaden, 2010, S. 553-574.

Lüken, J.; Schimmelpfennig, H.: Maximum-Likelihood-Schätzung. In: planung & analyse, Jg. 41/2014, Nr. 1, S. 42.