Kapitel VII

Clusteranalyse

B. 7 1. (*Gegenstand der Clusteranalyse*)
Ziel der Clusteranalyse (*Cluster* bedeutet: *Büschel, Häufung*) ist bei gleichzeitiger Betrachtung von mehr als zwei Variablen die einzelnen Objekte (Merkmalsträger) so zu Gruppen zusammenzufassen, dass die Ähnlichkeit der Objekte in den Gruppen möglichst groß, die Ähnlichkeit zwischen den Gruppen möglichst gering ist.

B. 7. 2. (*Anwendungsbeispiele*)

- *Werbung*
  Lassen sich die verschiedenen Konsumenten in bestimmte Persönlichkeitstypen segmentieren?

- *Marktforschung*
  Können auf dem relevanten Markt vorhandene Produkte bzw. auch Konsumenten in möglichst homogene Marktsegmente eingeteilt werden? Welche Segmente erscheinen unterbesetzt?

- *Strategische Unternehmensplanung*
  Können die auf dem Markt agierenden Unternehmen aufgrund ihrer Kennziffern in bestimmte Gruppen eingeteilt werden?

B. 7. 3. (*Clusteranalyseverfahren*)
In *SPSS* sind folgende Verfahren implementiert:

- *Hierarchische Clusteranalyse*
- *Clusterzentrenanalyse*
- *Twostep-Cluster*

B. 7. 4. (*Grundidee der hierarchischen Clusteranalyse*)
Bei der hierarchischen Clusteranalyse wird mithilfe verschiedener Algorithmen die gegebene Objektmenge fortlaufend fusioniert, sodass am Ende alle Cluster ein Objekt bilden. Der Anwender muss entscheiden, ab welcher Clusterzahl die beste Anzahl der Cluster vorliegt.

Die Clusterzentralanalyse ist ein iteratives Verfahren, bei dem eine Startpartition in Form von vorläufigen Gruppenmittelwerten vorgegeben wird. Anschließend werden die Objekte durch Umgruppierung so zu diesem zuvor bestimmten Clusterzentren zugeordnet, bis es zu keiner weiteren Abnahme der Streuung durch die Umgruppierung kommt.

Algorithmus (*Clusteralanalyse*)

Schritt 1: Auswahl des Distanz- oder Ähnlichkeitsmaßes.
Schritt 2: Auswahl des Clusteralgorithmus
Schritt 3: Festlegung der Clusterzahl
Schritt 4: Benennung der Cluster

(Es wird oft empfohlen, anstelle der Datenmatrix der Originaldaten eine standardisierte Datenmatrix mit den standardisierten Merkmalsausprägungen zu verwenden, um eine Verzerrung durch die unterschiedlichen Dimensionen zu vermeiden.)

Schritt 1: **Auswahl des Distanz und Ähnlichkeitsmaße**s

Je nach Art des Merkmals wird zwischen *Distanzmaßen* (metrisch- und ordinal skalierte Merkmale) und *Ähnlichkeitsmaßen* (nominal skalierte Merkmale) unterschieden.

Distanzmaße für metrisch skalierte Merkmale

Distanzmaß	**Formel**
Euklidische Distanz	$d_{ik} := \sqrt{\sum_{j=1}^{p} \left(x_{ij} - x_{kj} \right)^2}$
Quadrierte Euklidische Distanz	$d_{ik} = \sum_{j=1}^{p} \left(x_{ij} - x_{kj} \right)^2$
City-Block-Distanz	$d_{ik} := \sum_{j=1}^{p} \left\| x_{ij} - x_{kj} \right\|$
Maximum-Metrik	$d_{ik} := \max_{j} \left\{ \left\| x_{ij} - x_{kj} \right\| \right\}$

(All diese Distanzmaße gehören zu der sog. Minkowski-Metrik, mit der sich durch Gewichtung des Faktors μ neben den oben gezeigten Distanzmaßen auch weitere Distanzmaße bilden lassen

$$d_{ik} := \left(\sum_{j=1}^{p} \left| x_{ij} - x_{kj} \right|^{\mu} \right)^{\frac{1}{\mu}},$$

$\mu = 2$: Euklidische Distanz
$\mu = 1$: City-Block-Distanz
$\mu = \infty$: Maximum-Metrik

Distanzmaße für ordinal skalierte Merkmale
Für die Berechnung der Distanz zwischen nominal skalierten Merkmalen verwendet man Ähnlichkeitsmaße (*Similarity*). Es gilt dabei für die Überführung eines Ähnlichkeitsmaßes in ein Distanzmaß:

$$d_{ik} = 1 - s_{ik} .$$

Für die Bestimmung der Ähnlichkeit zwischen nominal skalierten Merkmalen kann eine 4-Felder-Tafel zur Verdeutlichung herangezogen werden:

Ausprägung bei Objekt i	Ausprägung bei Objekt k	
	1	0
1	a	b
0	c	d

Je nach der Berücksichtigung der Übereinstimmung bzw. Nichtübereinstimmung können die folgenden Ähnlichkeitsmaße stellvertretend für die 26 verschiedenen Ähnlichkeitsmaße genannt werden, die in SPSS implementiert sind. Die folgende Tabelle zeigt nur eine kleine Auswahl dieser Ähnlichkeitsmaße:

Ähnlichkeitsmaße für nominal skalierte Variable

Distanz- bzw. Ähnlichkeitsmaß	**Formel**
Euklidische Distanz	$d_{ik} := b + c$
Koeffizient von Tanimoto	$s_{ik} := \frac{a+d}{a+b+c+d}$
Koeffizient von Dice	$s_{ik} := \frac{2a}{2a+(b+c)}$
Koeffizient von Sokal & Sneath	$s_{ik} := \frac{2a}{a+(b+c)}$
Koeffizient von Russel & Rao	$s_{ik} := \frac{a}{a+b+c+d}$

Die Ähnlichkeitsmaße, die hier gezeigt sind, beziehen sich auf den Fall, dass die Daten bereits dichotom vorliegen bzw. binär verschlüsselt sind. Der Unterschied bei den einzelnen Ähnlichkeitsmaßen ergibt sich durch die entsprechende Gewichtung der einzelnen möglichen Ausprägungen für das Vorhandensein der Eigenschaft bei keinem, einem bzw. beiden Objekte.

Schritt 2: **Auswahl des Clusteralgorithmus**

Die hierarchischen Verfahren der Clusteranalyse unterscheiden sich in der Art, wie die Cluster A_l und A_m fusioniert werden. Diese Fusionierung kann aufgrund der geringsten Distanz d_{im} zum nächsten geschehen (*Single Linkage, Nearest Neighbour*), oder aufgrund der größten Distanz zum nächsten Objekt (*Complete Linkage, Furthest Neighbour*).
Daneben gibt es eine Reihe weiterer Verfahren.

Die Charakteristik und Leistungsfähigkeit der einzelnen Verfahren ist bereits ausführlich untersucht worden. Zusammenfassend seien die folgenden Punkte hier genannt:

Dilatierende Algorithmen neigen dazu, die Objekte zu sehr in viele einzelne gleich große Gruppen zusammenzufassen.

Kontrahierende Algorithmen tendieren dazu, zunächst wenig große Gruppen zu bilden, denen viele kleine Gruppen gegenüberstehen. Kontrahierende Verfahren sind dadurch besonders geeignet, Ausreißer zu identifizieren.

Weist ein Algorithmus weder die eine oder die andere Eigenschaft auf, so wird er mir *konservativ* bezeichnet

Unter *Kettenbildung* versteht man die Eigenschaft, Brücken zwischen eng aneinanderlegenden Objekten zu bilden, und diese dann zu einer Gruppe zusammenzufassen, obwohl sich auch zwei Gruppen daraus hätten bilden lassen können.

Die folgende Tabelle zeigt in einem Überblich weit verbreitete Verfahren der Clusteranalyse und deren wesentlichen Eigenschaften auf das Gruppierungsergebnis:

Algorithmen zur Clusteranalyse

Algorithmus	**Berechnung**	**Eigenschaft**
Single-Linkage	$d_{lm} := \min\{d_{ik} \mid i \in A_l,\ k \in A_m\}$	Kontrahierend, neigt zu Kettenbildung
Complete-Linkage	$d_{lm} := \max\{d_{ik} \mid i \in A_l,\ k \in A_m\}$	Dilatierend, neigt zur Bildung kleiner Gruppen
Average-Linkage	$d_{lm} := \frac{1}{2}(d_s + d_c)$	Konservativ
Centroid-Verfahren	$d_{lm} := d\left(\bar{x}_l, \bar{x}_m\right)$	Konservativ
Median-Verfahren	$d_{lm} := d\left(x_{0.5l}, x_{0.5m}\right)$	konservativ
Verfahren von Ward	$d_{lm} := \frac{n_l \cdot n_m}{n_l + n_m} \sum_{j=1}^{p}\left(\bar{x}_{lj} - \bar{x}_{mj}\right)^2$	Neigt zur Bildung gleich großer Gruppen

Die Größen bedeuten:

d_s : minimale Distanz nach dem Single-Linkage-Algorithmus

d_c : maximale Distanz nach dem Complete-Linkage-Algorithmus

d_{lm} : Distanz zwischen zwei Elementen der Cluster A_l und A_m

Schritt 3: **Festlegung der Clusterzahl**

Die Festlegung der „richtigen“ Anzahl von Clustern bleibt bei dem hierarchischen Verfahren dem Anwender überlassen. Oft ist dies erst nach mehrmaligem Ausführen der Prozedur möglich.

Hilfreich zur Entscheidung sind neben sachlichen Überlegungen auch Graphiken:

Eine spezielle Darstellungsform, die den Ablauf des Fusionierungsprozesses visualisiert, ist das *Dendrogramm*. Das Dendrogramm zeigt die Abnahme der Streuung in Form eines Homogenitätsmaßes durch die zunehmende Fusion der einzelnen Cluster.

Im Extremfall bildet jedes Objekt einen eigenen Cluster (Homogenität = 0) bzw. alle Objekte nur noch einen Cluster (Homogenität = Max). Der Anwender muss selbst entscheiden, ab welcher Stufe der Abnahme der Varianz die „richtige“ Anzahl an Clustern festzulegen ist.

Eine andere Darstellungsart des Fusionsprozesses sind die sogenannten *Eiszapfenplots*. Diese zeigen ebenfalls horizontal bzw. vertikal die Abnahme der Streuung durch die fortlaufende Fusion der einzelnen Objekte zu Clustern. Bei dem Eiszapfplot lässt sich ablesen, bei welcher Stufe der Fusionierung welche Objekte zusammengefasst werden. Dies gibt wesentliche Hinweise darauf, wie die einzelnen Cluster zusammengesetzt sind. Dadurch lässt sich entscheiden, welche Grupperung sich als in den Augen des Anwenders als gelungene Grupperung erweist

Eine weitere Möglichkeit bietet sich durch ein sogenanntes *Struktogramm* an. In diesem Diagramm wird an der Ordinate die Abnahme der Streuung (Homogenitätsmaß) und an der Abszisse die Anzahl der Objekte eingezeichnet. Grundsätzlich ließe sich anstelle des Homogenitätsmaßes auch die Distanz aufgetragen, ab welcher es zu einer Fusionierung zwischen zwei Objekten gekommen ist. Werden die Punkte durch eine Linie verbunden, so könnte man an der Stelle, an der ein optisch signifikanter *Knick*, auch *Elbow* genannt, entsteht, die „richtige“ Anzahl an Clustern festlegen. Dieses Verfahren ist jedoch in *SPSS* nicht vorgesehen, lässt sich aber mit Hilfe eines Liniendiagramms nachträglich erzeugen.

Schritt 4: **Benennung der Cluster**
Diese werden als Ergebnis der Clusteranalyse erscheinen.

(Letzte Aktualisierung: 28.03.11)