planung &analyse

Zeitschrift für Marktforschung und Marketing www.planung-analyse.de 2/2015 D11700F

Eine Marke der dfv Mediengruppe

Interview

Dr. Stefan Stumpp
CEO TNS Infratest

Special

Im Moment der Kaufentscheidung

Kongresse

Veranstaltungen
PUMa, GOR
MAFO 2015

Schwerpunkt

Segmentierung

Statistik KOMPAKT

Hierarchische Clusteranalyse

Neben den partitionierenden zählen die hierarchischen Verfahren zu den bedeutendsten Methoden der Clusteranalyse. Sie fassen die zu gruppierenden Elemente schrittweise zu immer größeren Clustern zusammen. Dagegen gehen partitionierende Verfahren von einer gegebenen Klassifikation der Elemente aus und versuchen diese durch Umgruppierungen zu verbessern.

Verfahren der Hierarchischen Clusteranalyse

Hierarchische Verfahren werden unterteilt in agglomerative und divisive Algorithmen. Praktische Relevanz besitzt jedoch nur die agglomerative Vorgehensweise. Im Fall einer deterministischen Clusteranalyse mit Objekten beinhaltet sie folgende Schritte:

1. In der Ausgangslösung ist jedes Objekt ein eigenständiges Cluster.

2. Die zwei ähnlichsten Cluster werden sukzessive zu einem neuen Cluster zusammengefasst, bis sich alle Objekte in einem Cluster befinden.

Eine Folge dieses Vorgehens ist, dass einmal zu einem Cluster zusammengefasste Elemente im weiteren Fusionsprozess nicht mehr voneinander getrennt werden können. Durch die Veränderung der Cluster aufgrund der Hinzunahme weiterer Elemente kann es passieren, dass ein Objekt letztlich nicht mehr dem Cluster zugeordnet ist, zu dessen anderen Elementen es am ähnlichsten ist.

Es existiert eine Reihe verschiedener agglomerativer Algorithmen, die sich im Hinblick auf die Bestimmung der Ähnlichkeit zwischen zwei Clustern unterscheiden. Abbildung 1 zeigt ein kleines Datenbeispiel mit vier Objekten, die anhand von zwei Merkmalen charakterisiert sind. Beispielsweise könnte die Ähnlichkeit der beiden Cluster {O1,O2} und {O3,O4} in Abbildung 1b durch die beiden Objekte aus den zwei Clustern bestimmt sein, die sich

- am nächsten sind, das heißt: O1 und O3 (Single-Linkage Verfahren), oder
- am weitesten entfernt voneinander sind, das heißt: O1 und O4 (Complete-Linkage Verfahren).

Ein weiteres, nämlich das Ward-Verfahren gilt unter üblichen Rahmenbedingungen als die beste Methode.

Ward-Verfahren

Zentral für das Ward-Verfahren ist die Streuungsquadratsumme (SQS) einer Klassifikation. Formal ist diese bestimmt durch die quadrierten Abweichungen der Objekte eines Clusters zum Clusterzentrum (= Mittelwerte der Objekte eines Clusters) summiert über alle Cluster. Grafisch bedeutet die SQS beispielsweise für die Klassifikation in Abbildung 1b: die (euklidischen) Distanzen vom Clusterzentrum ● zum Objekt O3 und zum Objekt O4 sowie die (euklidischen) Distanzen vom Clusterzentrum ● zum Objekt O1 und zum Objekt O2 sind zu quadrieren und anschließend alle vier Werte zu addieren. Somit ist die Vorschrift des Ward-Verfahrens: Fasse jeweils die beiden Cluster zu einem neuen Cluster zusammen, durch deren Verschmelzung die SQS am wenigsten erhöht wird.

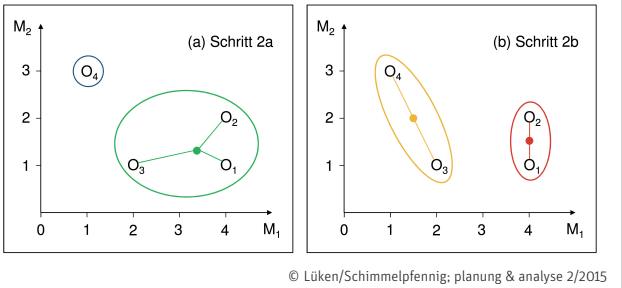

Abbildung 1: Beispiel mit zwei Merkmalen

Veranschaulichung des Algorithmus anhand des Beispiels aus Abbildung 1:

Schritt 0: Ausgangspunkt sind vier Cluster, die jeweils ein Objekt enthalten.

Schritt 1: Abbildung 2 zeigt die sechs verschiedenen Möglichkeiten, zwei Cluster zu einem neuen zusammen zu fassen. Werden die Cluster {O1} und {O2} zu einem fusioniert, so erhöht sich die SQS am wenigsten: die quadrierte Distanz von O1 zum Clusterzentrum beträgt $(4-4)^2 + (1-1{,}5)^2 = 0{,}25$, die quadrierte Distanz von O2 zum Clusterzentrum $(4-4)^2 + (2-1{,}5)^2 = 0{,}25$. Insofern ergibt sich eine Streuungsquadratsumme von 0,5, da in den übrigen beiden Clustern nur jeweils ein Objekt enthalten ist.

Schritt 1: Fusion von		Zentrum des neues Clusters		
Cluster	Cluster	M_1	M_2	SQS
{O1}	**{O2}**	**4,0**	**1,5**	**0,5**
{O1}	{O3}	3,0	1,0	2,0
{O1}	{O4}	2,5	2,0	6,5
{O2}	{O3}	3,0	1,5	2,5
{O2}	{O4}	2,5	2,5	5,0
{O3}	{O4}	1,5	2,0	2,5

Schritt 2: Fusion von		Zentrum des neues Clusters		
Cluster	Cluster	M_1	M_2	SQS
{O1,O2}	{O3}	3,3	1,3	3,3
{O1,O2}	{O4}	3,0	2,0	8,0
{O3}	**{O4}**	**1,5**	**2,0**	**3,0**

Abbildung 2: Streuungsquadratsumme für verschiedene Klassifikationen

Schritt 2: Neben den Möglichkeiten zu dem im Schritt zuvor gebildeten Cluster {O1,O2} entweder das Cluster {O3} (siehe Abbildung 1a) oder das Cluster {O4} (ohne eigene Abbildung) hinzuzufügen, könnten auch die Cluster {O3} und {O4} zu einem Cluster vereinigt werden (siehe Abbildung 1b). Letztere führt zur geringsten SQS, so dass sich nach dem zweiten Schritt 2 Cluster mit jeweils 2 Objekten ergeben.

Schritt 3: Die beiden Cluster werden abschließend zu einem fusioniert, das dann alle Objekte enthält.

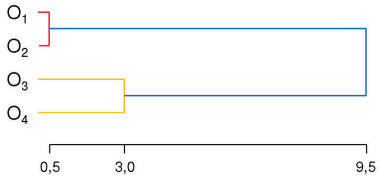

Abbildung 3: Dendrogramm zum Ward-Verfahren

Abbildung 3 veranschaulicht den Fusionsprozess in einem Dendrogramm. Mögliche Klassifikationen sind somit neben der anfänglichen, in der alle Objekte eigenständige Cluster sind, bzw. der letzten, in der alle Objekte zu einem Cluster gehören, {O1,O2}, {O3} und {O4} nach Schritt 1 und {O1,O2} und {O3,O4} nach Schritt 2. Kriterien, die zur Auswahl einer Klassifikation verwendet werden können, werden in einem kommenden Beitrag dieser Reihe vorgestellt.

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

In Ausgabe 3/2015: Partitionierende Clusteranalyse

► Literatur

Bacher, J.; Pöge, A.; Wenzig, K.: Clusteranalyse. München, 2010, S. 285-297

Backhaus, K.; Erichson, B.; Plinke, W., Weiber, R.: Clusteranalyse. In: Multivariate Analysemethoden. Berlin, Heidelberg, 2011, S. 395-436