planung &analyse

Zeitschrift für Marktforschung und Marketing www.planung-analyse.de 3/2015 D11700F

Eine Marke der dfv Mediengruppe

Special

Forschung für die Medien

Veranstaltungen

60 Jahre ADM und Designforschungstag

MaFo-Spitzen

von Stephan Grünewald

Schwerpunkt

Marke als Orientierung

Statistik KOMPAKT

Partitionierende Clusteranalyse

Partitionierende Verfahren zählen neben den hierarchischen zu den bedeutendsten Methoden der Clusteranalyse. Sie gehen von einer gegebenen Klassifikation der Elemente aus und versuchen diese durch Umgruppierungen zu verbessern. Beide Verfahrenstypen sind aber nicht sich ausschließende Alternativen, sondern können gemeinsam eingesetzt werden, um die Stärken beider zu nutzen.

K-Means-Algorithmus

Das bekannteste Verfahren der partitionierenden Clusteranalyse basiert auf dem K-Means-Algorithmus, der folgende Schritte umfasst:

1. Für eine vorgegebene Anzahl an Clustern wird eine zufällige Klassifikation ermittelt.
2. Für jedes Cluster ist das Clusterzentrum (= Mittelwerte der Variablen über die Objekte eines Clusters) zu bestimmen.
3. Jedes Objekt wird dem Cluster zugeordnet, zu dessen Clusterzentrum es am nächsten liegt, das heißt formal die geringste quadrierte euklidische Distanz aufweist.
4. Falls es keine Umgruppierung gab, bricht der Algorithmus ab. Ansonsten wird mit Schritt 2 fortgefahren.

Veranschaulichung des Algorithmus anhand des Beispiels aus Abbildung 1:

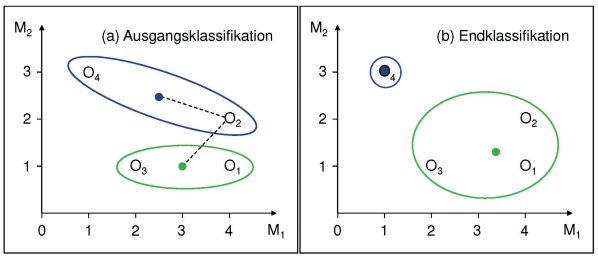

Abbildung 1: Beispiel mit zwei Merkmalen

Abbildung 1a zeigt für ein kleines Datenbeispiel mit vier Objekten, die anhand von zwei Merkmalen beschrieben sind, mit {01,03} und {02,04} eine mögliche, zufällig bestimmte Ausgangsklassifikation.

	Ausgangsklassifikation		Klassifikation nach dem ersten Iterationsschritt (Endklassifikation)	
	$\{O_1,O_3\}$	$\{O_2,O_4\}$	$\{O_1,O_2,O_3\}$	$\{O_4\}$
O_1	1,00	4,50	**0,56**	13,00
O_2	2,00	2,50	**0,89**	10,00
O_3	1,00	2,50	**1,89**	5,00
O_4	8,00	2,50	8,22	**0,00**

Abbildung 2: Quadrierte euklidische Distanzen der Objekte zu den Clusterzentren

Nur für Objekt O2 ist die quadrierte euklidische Distanz zum Zentrum ● des Clusters, zu dem es nicht gehört, kleiner als zu dem Zentrum ● des Clusters, in dem es sich befindet (siehe Abbildung 2). Insofern wird O2 statt dem *blauen* dem *grünen* Cluster zugeordnet, so dass sich die Klassifikation {01,02,03} und {04} ergibt (siehe Abbildung 1b). Da nun kein Objekt eines Clusters mehr zum Zentrum des anderen Clusters eine geringere Distanz aufweist, ist damit auch die abschließende Klassifikation erreicht.

Einfluss der Ausgangsklassifikation

Der K-Means-Algorithmus verfolgt das Ziel, die Streuungsquadratsumme – die quadrierten euklidischen Distanzen der Objekte eines Clusters zum Clusterzentrum summiert über alle Cluster – zu minimieren. Diese beträgt 3,34 (= 0,56 + 0,89 + 1,89 + 0,00) für die Klassifikation {01,02,03} und {04}. Jedoch ist {01,02} und {03,04} eine Klassifikation, deren Streuungsquadratsumme mit 3 noch geringer ist (siehe zur Berechnung den Beitrag dieser Reihe in Ausgabe 2/2015). Insofern garantiert der K-Means-Algorithmus nicht, das globale Minimum zu finden. Welche abschließende Klassifikation gefunden wird, ist abhängig von der Ausgangsklassifikation. Schließlich hätte auch {01,02} und {03,04} bereits die zu Beginn zufällig bestimmte Klassifikation sein können.

In statistischen Softwarepaketen ist die Ausgangsklassifikation zumeist abhängig von der Reihenfolge, in der die Fälle in dem Datensatz vorliegen. Die Ergebnisse eines typischen Datenbeispiels mit 22 auf einer vierstufigen Likert-Skala von 400 Befragten zu beurteilenden Statements mögen die Vielfalt der Ergebnisse illustrieren. Zehn Mal nacheinander wurde jeweils die erste Datenzeile an das Ende verschoben und somit zehn Datensätze leicht unterschiedlicher Reihenfolge erzeugt. Die Anwendung des K-Means-Algorithmus führte zu drei verschiedenen Klassifikationen mit drei Clustern sowie sechs verschiedenen Klassifikationen mit vier Clustern.

Two-Stage Clustering

Neben der mangelnden Eindeutigkeit des Ergebnisses wird die Notwendigkeit der Vorgabe einer Anzahl an Clustern als Hindernis für die Anwendung von partitionierenden Verfahren angeführt. Diesen begegnet das Two-Stage Clustering. Mit Hilfe eines hierarchischen Verfahrens werden in der ersten Stufe die Anzahl an Clustern und eine Ausgangsklassifikation bestimmt, die in der zweiten Stufe mittels partitionierendem Verfahren verbessert wird. Das Two-Stage Clustering ermöglicht somit im Gegensatz zur alleinigen Anwendung eines hierarchischen Verfahrens, Objekte eines Clusters wieder voneinander zu trennen. Da das (hierarchische) Ward-Verfahren in jedem Fusionsschritt ebenfalls auf die Minimierung der Streuungsquadratsumme abzielt, bietet sich seine Kombination mit dem K-Means-Algorithmus an.

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

In Ausgabe 4/2015: Bestimmung der Clusteranzahl

► Literatur

Bacher, J.; Pöge, A.; Wenzig, K.: Clusteranalyse. München 2010, S. 299-305

Punj, G.; Stewart; D.W.: Cluster Analysis in Marketing Research: Review and Suggestions for Application. In: Journal of Marketing Research, Jg. 20/1983, Nr. 2, S. 134-148