Clusteranalyse

Seminararbeit zum Proseminar
“Ausgewählte Methoden der Datenanalyse und des Data Mining”

eingereicht am 9. November 2016
Fachgebiet für Quantitative Methoden der Wirtschaftswissenschaften
bei Prof. Dr. Udo Bankhofer

von

Marco Hanisch
54829
Bergrat-Mahr-Straße 12
98693 Ilmenau

Markus Postler
53992
Max-Planck-Ring 4
98693 Ilmenau

Ilmenau, den 9. November 2016

Inhaltsverzeichnis

1 Motivation

Eine Clusteranalyse oder auch Klassifizierung dient dazu, Objekte und/oder Merkmale zu klassifizieren. Dabei sollen Merkmale und/oder Objekte in möglichst homogenen Klassen, die untereinander möglichst heterogen sind, zusammengefasst werden.

Für die Klassifizierung wurden zahlreiche Verfahren entworfen, die sich in Bezug auf die Vorgehensweise und ihren Anforderungen an die Datenstruktur unterscheiden.

Ziel dieser wissenschaftlichen Arbeit soll es sein, einen Überblick über die verschiedenen Verfahren zu gewinnen und Anwendungsempfehlungen bezüglich der Vorgehensweise bei unterschiedlichen Datenstrukturen und Anforderungsszenarien an die Clusteranalyse zu geben. Anhand von Bewertungskriterien soll ein festes Regelwerk in Form eines Entscheidungsbaumes aufgestellt werden, das die Entscheidung für ein geeignetes Vorgehen ermöglichen soll.

2 Eigenschaften von Clusterlösungen

2.1 Klassifizierung von Objekten und/oder Merkmalen

2.2 Disjunkte und nicht-disjunkte Zuordnung

2.3 Exhaustive und nicht-exhaustive Zuordnung

2.4 Homogenität innerhalb der Cluster

2.5 Heterogenität zwischen den Clustern

3 Ein Überblick über Clusteranalyse-Verfahren

4 Distanzmatrizen

5 Bewertungskriterien (erstmal leer lassen

Für die Bewertung der verschiedenen Verfahren ist es notwendig, diese nach verschiedenen Kriterien zu untersuchen. Dabei stellt einer der größten Einflussfaktoren die externen Anforderungen an die Klassifizierung dar. Hier soll untersucht werden, was der Anwender mithilfe der Klassifizierung erreichen will und welche Eigenschaften verschiedene Klassifizierungen beinhalten können.
Einen weiteren Einflussfaktor stellt die zu untersuchende Merkmalsstruktur dar. Die ursprünglichen Merkmale können unterschiedliche Skalenniveaus beinhalten, wovon jede Skalierung eines Merkmals sich anders auf Verfahren der Klassifizierung auswirken kann. Problematisch sind hier vor allem auch Datenmatrizen, die Merkmale mit unterschiedlichen Skalenniveaus beinhalten.
Ein weiteres Kriterium, das betrachtet werden sollte, sind hier besondere Ausprägungen der zu untersuchenden Merkmale. Fehlende Werte, Ausreißer und irrelevante Variablen stellen eine potentielle Fehlerquelle bei der Klassifizierung dar.

5.1 Fehlende Werten

5.2 Ausreißer

5.3 Irrelevante Variablen

6 Hierarchisch-agglomerative Verfahren

In diesem Kapitel sollen einzelne Verfahren hinsichtlich der zuvor genannten Bewertungskriterien untersucht werden. Zudem beschränken wir uns auf die gebräuchlichsten hierarchisch-agglomerativen Verfahren, da sich Austausch-oder divisive Verfahren in der Praxis nicht bewährt haben. Anhand eines Beispiel-Datensatzes sollen die Reaktionen der Verfahren auf die Bewertungskriterien anschaulich aufgezeigt werden.

6.1 Single linkage-Verfahren

6.2 Complete linkage-Verfahren

6.3 Average linkage between groups-Verfahren

6.4 Average linkage within groups-Verfahren

6.5 Centroid Clustering-Verfahren

6.6 Median Clustering-Verfahren

6.7 WARD-Verfahren

7 Regelwerk

Aufbauend auf den vorherigen zwei Kapiteln soll hier ein festes Regelwerk in Form eines Entscheidungsbaumes entstehen, das die Vorgehensweise bei unterschiedlichen Datenstrukturen und Anforderungsszenarien aufzeigt und dem Anwender die Auswahl für ein geeignetes Verfahren erleichtert.

Literaturverzeichnis

[Bac02] Johann Bacher. *Clusteranalyse: Anwendungsorientierte Einführung*. Oldenbourg, München, 2., erg. aufl., [nachdr.] edition, 2002.

[BPW10] Johann Bacher, Andreas Pöge, and Knut Wenzig. *Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren*. Oldenbourg, München, 3., erg., vollst. überarb. und neu gestaltete aufl. edition, 2010.

[BV08] Udo Bankhofer and Jürgen Vogel. *Datenanalyse und Statistik: Eine Einführung für Ökonomen im Bachelor ; [Bachelor geeignet!]*. Lehrbuch. Gabler, Wiesbaden, 1. aufl. edition, 2008.