Clusteranalyse

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

eingereicht am 7. Dezember 2016
Fachgebiet für Quantitative Methoden der Wirtschaftswissenschaften
bei Prof. Dr. Udo Bankhofer

von

Marco Hanisch	Markus Postler
54829	53992
Bergrat-Mahr-Straße 12	Max-Planck-Ring 4
98693 Ilmenau	98693 Ilmenau

Ilmenau, den 7. Dezember 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

Nach Bankhofer [Ban08] können erfasste Merkmale in quantitative und qualitative Merkmale unterschieden werden. Quantitative Merkmal besitzen einen hohen Informationsgehalt und ihre Ausprägungen werden mit Zahlen benannt. Qualitative Merkmale besitzen einen niedriegeren Informationsgehalt als quantitative Merkmale und werden durch Begriffe beschrieben. Qualitative Merkmale können weiterhin in nominale und in ordinale Merkmale unterschieden werden.

1. Quanitative Merkmale: Diese Merkmale werden auch kardinale oder metrische Merkmale genannt. Sie besitzen den höchstmöglichen Informationsgehalt. Die Merkmalsausprägungen sind Zahlen, welche eine Ordnung besitzen. Dabei können alle möglichen Merkmalsausprägungen auf Skalen geordnet werden. Dadurch kann man den Abstand, sowie das Verhältnis zwischen zei Merkmalsausprägungen bestimmten.
2. Ordinale Merkmale: Diese Merkmale gehören zu den qualitativen Merkmalen. Die Merkmalsausprägungen werden durch Begriffe dargestellt. Dabei können alle Ausprägungen vollständig geordnet werden. Durch diese Ordnung ist ein Vergleich zweier Merkmalsausprägungen möglich, wodurch man die Merkmalsausprägungen in Reihenfolgen bringen kann.
3. Nominale Merkmale: Diese Merkmale gehören zu den qualitativen Merkmalen. Die Merkmalsausprägungen werden durch Begriffe dargestellt. Die Ausprägungen besitzen dabei keine Ordnung. Die Ausprägungen können lediglich auf Gleichheit oder Ungleichheit überprüft werden. Diese Merkmale besitzen den niedrigsten Informationsgehalt.

2.2 explorative und konfirmatorische Lösungen

Clusterverfahren können nach Bacher [Bac10] in explorative und konfirmatorische Clusterverfahren unterschieden werden. Bei explorativen Verfahren steht die Anzahl der Cluster sowie die kennzeichnende Merkmalsausprägung nicht schon im Vorhinein fest. Bei konfirmatorischen Verfahren steht die Anzahl der Cluster und die Charakteristik der Cluster bereits vor dem anwenden der Verfahren fest.

2.3 Disjunkte und nicht-disjunkte Zuordnung

Nach Bankhofer [Ban08] können Clusterlösungen in disjunkte und nicht disjunkte Lösungen unterteilt werden. Diese Klassizifierung der Cluster beschreibt ob Objekte mehreren Klassen

zugeordnet werden können oder nur Bestandteil einer Klasse sind.

1. Disjunkte Lösungen: Ein ist Objekt genau einer Klasse zugeordnet. Es kommtg also nicht zu Überschneidungen von mehreren Klassen.
2. Nicht-Disjunkte Lösungen: Ein Objekt kann mehreren Klassen zugeordnet werden. Es können also Klassen existieren, welche gemeinsame Objekte besitzen. Diese Klassen können auch als überlappende oder überdeckende Klassen bezeichnet werden. Dabei ist zu beachten, dass Teilmengenbeziehungen zwischen Klassen ausgeschlossen sind.

2.4 Exhaustive und nicht-exhaustive Zuordnung

Nach Bankhofer [Ban08] können Cluster unterteilt werden, ob alle Objekte der Datenmenge bei der Klassifizierung einbezogen werden oder nicht. Aufgrundlage dieses Klassifikationstyps kann man Cluster in 2 Gruppen unterteilen:

1. Exhaustive Zuordnung: Alle Objekte der verarbeiteten Datenmenge werden klassifiziert.
2. Nicht Exhaustive Zuordnung: Nur ein Teil der verarbeiteten Datenmenge wird klassifiziert. Die nicht berücksichtigten Objekte werden vernachlässigt.

2.5 Homogenität innerhalb der Cluster

Die Homogenität innerhalb eines Clusters beschreibt wie ähnlich sich die Objekte des Clusters sind. Die Objekte innerhalb eines Clusters sollten maximal homogen, also maximal ähnlich zu einander sein (vgl. Bacher et al [Bac10], S.16). Nach Bankhofer [Ban08] kann die Homogenität durch die Innerklassenverschiedenheit abgebildet werden. Dies wird anhand der Maximaldistanz zwischen Objekten des Clusters dargestellt. Bei einelementigen Clustern ist dabei die Homogenität maximal.

Formeln hinzufügen ???

2.6 Heterogenität zwischen den Clustern

Die Heterogenität zwischen den Clustern beschreibt, wie verschieden die Objekte von verschiedenen Clustern sind. Die Objekte von verschiedenen Clustern sollen minimal homogen sein. Die Objekte sollen heterogen, also verschieden sein (vgl. Bacher et al [Bac10], S.16). Nach Bankhofer [Ban08] kann die Heterogenität durch die Zwischenklassenverschiedenheit beschrieben werden. Diese Verschiedenheit wird anhand von Distanzen dargestellt. Diese Distanzen können auf Grundlage von verschiedenen Methoden berechnet werden:

1. Single Linkage: Die minimale Distanz zwischen zwei Objekten der betrachteten Cluster wird zur Darstellung genutzt.
2. Average Linkage: Die mittlere Distanz zwischen den Objekten der betrachteten Cluster wird zur Darstellung genutzt.

3. Complete Linkage: Die maximale Distanz zwischen zwei Objekten der betrachteten Cluster wird zur Darstellung genutzt.

2.7 Fusionierungseigenschaften

Clusterverfahren lassen sich nach Backhaus [Bac16] S.488/489 anhand ihrer Fusionierungseigenschaften in drei Gruppen unterteilen:

1. Dilatierende Verfahren: Bei diesen Verfahren werden die Objekte in einzelne etwa gleich große Gruppen zusammengefasst.
2. Kontrahierende Verfahren: Bei diesen Verfahren stehen viele kleine Gruppen wenigen großen Gruppen gegenüber.
3. Konservative Verfahren: Diese Verfahren weisen weder Konstrahierende noch Dilatierende Merkmale auf.

3 Ein Überblick über Clusteranalyseverfahren

Nach Backhaus et al [Bac16] S. 476 lassen sich vier übergeordnete Gruppierungen von Clusteranalyseverfahren darstellen:

1. Partitionierende Verfahren: Diese Algorithmen benötigen eine vorgegebene Clusteranzahl, in die sie die Objekte einzuordnen versuchen. Unterschiede zwischen den einzelnen Verfahren entstehen hierbei vor allem durch die unterschiedliche Messung der Verbesserung der Clusterbildung und in der Regelung des Austauschs der Objekte zwischen den Clustern.
2. Hierarchische Verfahren: Im Gegensatz zu partitionierenden Verfahren benötigen diese Algorithmen keine vorgegebene Clusteranzahl, sondern iterieren alle möglichen Clusteranzahlen durch. Hierarchisch divisive Verfahren gehen dabei von der größtmöglichen Partition aus, die sie Schritt für Schritt in die kleinstmöglichen Partitionen zerlegen (ein Objekt in einer Partition). Hierarchisch agglomerative Verfahren dagegen fassen die feinsten Partitionen zu immer größeren Gruppen zusammen, bis schließlich die größtmögliche Partition erreicht ist, die alle Objekte enthält.
3. Graphentheoretische Verfahren
4. Optimierungsverfahren

[Bac10] S. 18 unterscheidet andere Clusteranalyseverfahren auf Grundlage der Zuordnung der Klassifikationsobjekte zu den Clustern:

1. Unvollständige Clusteranalyseverfahren: auch geometrische Methoden, Repräsentations- oder Projektverfahren, führen nur zu einer räumlichen Darstellung, nur bis dreidimensionalen Raum möglich
2. Deterministische Clusteranalyseverfahren: Klassifikationsobjekte werden mit Wahrscheinlichkeit 1 einem oder mehreren Clustern zugeordnet,
3. Probabilistische Clusteranalyseverfahren: Fuzzy, Klassifikationsobjekte werden verschiedenen Clustern mit einer Wahrscheinlichkeit zwischen 0 und 1 zugeordnet, Verallgemeinerung der Deterministischen Clusteranalyseverfahren (Annahme, dass w = 0 oder 1 wird fallen gelassen)

S.20/21 [Bac10] Unterscheidung auch in heuristische und modellbasierte Verfahren, Grafik zeigt Vergleich zwischen diesen und dem Bacher.2010-Schema

Gedanke: Graphentheoretische Verfahren/Optimierungsverfahren in Bacher.2010-Schema einordnen?

Die partitionierenden Verfahren lassen sich wiederum in Austauschverfahren und iterierte Mi-

nimaldistanzverfahren unterscheiden.

Xu [Xu99] erwähnt weiterhin noch Single Scan Clustering, den BIRCH-Algorithmus, den STING-Algorithmus und Grid Clustering. Diese speziellen Algorithmen dienen der Klassifizierung bei räumlichen Datenbanken (Spatial Databases).

+ Abbildung in [Bac16] S.476??? Überblick normal über Clusterverfahren + Abbildung in [Xu99] S. 21??? Unterschied hierarchisch/agglomerativ sehr gut + Ref auf Chapter hierarchisch-agglomerative Verfahren

4 Proximitätsmaße

Um die Einteilung in möglichst homogene Gruppen vornehmen zu können, müssen die zu untersuchenden Objekte bezüglich ihrer zu beobachtenden Eigenschaften auf Ähnlichkeit untersucht werden.

Nach Backhaus et al [Bac16] lassen sich zwei Arten von Proximitätsmaßen unterscheiden:

1. *Ähnlichkeitsmaße*: Diese Maße spiegeln die Ähnlichkeit zweier Objekte wider. Je höher der zugewiesene Wert für zwei Objekte, desto höher ist auch ihre Ähnlichkeit.
2. *Distanzmaße* (auch *Unähnlichkeitsmaße*): Diese drücken die Unähnlichkeit zweier Objekte aus. Je größer die angegebene Distanz, desto unähnlicher sind die Objekte, wobei eine Distanz von Null ausdrückt, dass die zwei Objekte hinsichtlich ihrer Klassifikationsmerkmale vollkommen identisch sind. Die Distanzmaße lassen sich als entgegengesetzter Pol der Ähnlichkeitsmaße auffassen, wobei diese Eigenschaft die Überführung beider Maße ineinander ermöglicht. (Eckey et al. [Eck02], S. 205).

[Eck02] wenn mathematische Eigenschaften noch mit aufgenommen werden sollen

Durch unterschiedliche Skalenniveaus der betrachteten Merkmale lassen sich eine Vielzahl von unterschiedlichen Proximitätsmaßen bestimmen.

Die Wahl eines Proximitätsmaß für eine Clusteranalyse hängt laut maßgeblich davon ab, ob die Clusteranalyse Objekte oder Variablen zu klassifizieren versucht (Vgl. Bacher et al. [Bac10], S. 196) Welche Eigenschaft? -¿ Welches Maß

Übersicht generieren:

Dichotome Variablen	City-Block-Metrik Quadrierte Euklidische Distanz Produkt-Moment-Korrelationskoeffizient
Nominale Variablen	Dummybildung sonst wie Dichotome Variablen
Ordinale Variablen	City-Block Metrik Canberra-Metrik JACCARD-II-Koeffizient verallgemeinerter Simple-Matching-Koeffizient für ordinale Variablen Übereinstimmungskoeffizient für ordinale Variablen
Quantitative Variablen	Produkt-Moment-Korellation City-Block-Metrik euklidische Distanz quadrierte euklidische Distanz Chebychev-Distanz

[Bac10] S. 200 Alle Ähnlichkeitsmaße a lassen sich durch

$$u_{ij} = 1 - a_{ij} \quad bzw. \quad u_{g,g*} = 1 - a_{g,g*} \tag{4.1}$$

in Distanzmaße u umwandeln.

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 hierarchisch-agglomerative Verfahren untersucht werden, da sich Austausch- oder divisive Verfahren in der tatsächlichen Anwendung nicht bewährt haben. Der Grund hierfür ist die höhere Rechenzeit für Algorithmen der divisiven Verfahren in Verbindung mit dem Fehlen eines Nachweises, dass divisive Strategien präzisere Clusterlösungen liefern als agglomerative Verfahren. In der Praxis wird daher meist auf hierarchisch-agglomerative Verfahren zurückgegriffen (Vgl. Pedrycz [Ped10] S. ???) Nach Piegorsch ([Pie15], S. 378) kann der Austausch des zugrundeliegenden Verfahrens Fluch oder Segen sein: Die Verwendung eines anderen Verfahrens auf Basis der selben Daten zeigt interessante oder unerwartete Änderungen im Ergebnis der Clusteranalyse. Dies ist eine Beobachtung, die den zugrundeliegenden Prozess der Wissensschöpfung verbessern kann.

6.1 Single Linkage-Verfahren

Bei dem Single Linkage-Verfahren können sowohl Ähnlichkeits- als auch Distanzmaße verwendet werden. Dabei werden jeweils die zwei Cluster zusammengefasst, welche die kleinste Distanz zwischen sich aufweisen. Die Distanz entspricht dabei der kleinstmöglichen Distanz zwischen Objekten der verschiedenen Cluster. Nach dem Bilden eines neuen Clusters müssen die Distanzen jeweils neu ausgerechnet werden und die neue Minimaldistanz identifiziert werden. Nach Backhaus [Bac16] kann die neue Distanz vereinfacht mit folgender Formel errechnet werden:

$$D(R; P + Q) = min\{D(R, P); D(R, Q)\} \tag{6.1}$$

Aufgrund des Vorgehens wird dieses Verfahren auch ”Nearest-Neighbour-Verfahren”genannt. (Vgl. Eckey et al. [Eck02], S. 231) Dieses Verfahren ist ein kontrahierendes Verfahren. Zudem kann es genutzt werden um Ausreißer zu identifizieren. (Vgl. Backhaus [Bac16], S.481-483) Dieses Verfahren erzeugt oft Cluster, die ziemlich diffus, langgestreckt und/oder unförmig sind. (Vgl. [Pie15], S. 377) Hierbei entsteht allerdings auch das Problem des Verkettungseffekts: Cluster werden zusammengefasst, die nur durch eine Brücke verbunden sind, im Raum aber deutlich separiert voneinander liegen. Dies kann zu eher heterogenen Clustern führen. (Vgl. Eckey et al. [Eck02], S. 233) Nach Clarke et al. ([Cla09], S. 416) sind solche ungewöhnlichen Strukturen in der Natur allerdings durchaus üblich. Er zitiert: ”...it’s unclear in general wehether such properties are features or bugs.”Der Einfluss solcher Fusionierungseigenschaften ist noch nicht vollständig untersucht.

6.2 Complete Linkage-Verfahren

Bei dem Complete Linkage-Verfahren können ebenfalls sowohl Ähnlichkeits- als auch Distanzmaße verwendet werden. Es werden jeweils die Cluster mit der geringsten Distanz zusammengefasst. Die Distanz berechnet sich allerdings anders als beim Single Linkage-Verfahren. Die Distanz zwischen Clustern entspricht hier nicht der kleinstmöglichen, sondern der größtmöglichen Distanz zwischen Objekten der verschiedenen Cluster. Auch hier muss nach jedem Zusammenfassen zweier Cluster die Distanzen jeweils neu ausgerechnet werden und die neue Maximaldistanz identifiziert werden. Nach Backhaus [Bac16] kann die neue Distanz vereinfacht mit folgender Formel errechnet werden:

$$D(R; P + Q) = max\{D(R, P); D(R, Q)\} \tag{6.2}$$

Aufgrund des Vorgehens wird dieses Verfahren auch ”Furthest-Neighbour-Verfahren”genannt. Dieses Verfahren ist ein dilatierendes Verfahren. Es eignet sich im Gegensatz zum Single Linkage-Verfahren nicht gut, um Ausreißer zu identifizieren, da es eher kleine Gruppen bildet. (Vgl. Backhaus [Bac16], S.483/484) Ein Problem der Orientierung an den maximal entfernten Objekten zweier Cluster stellt das Ausbleiben einer Fusionierung dar, selbst wenn die mittlere Distanz dieser zweier Objekte keine merkliche Erhöhung der Heterogenität im neu zu bildenden Cluster anzeigen würde. (Vgl. Eckey et al. [Eck02], S.236)

6.3 Average Linkage-Verfahren

6.4 Centroid-Verfahren

6.5 Median-Verfahren

6.6 WARD-Verfahren

+ Abbildung in [Bac16] S.489??? Fusionierungseigenschaften -¿ Wie reagieren Verfahren? Abbildung kann rein, Fusionierungseigenschaften vielleicht unter 2.Clusterlösungen erklären?

Verfahren	Eigenschaft	Monoton?	Proximitätsmaße	Bemerkungen
Single Linkage	kontrahierend	ja	alle	neigt zur Kettenbildung
Complete Linkage	dilatierend	ja	alle	neigt zu kleinen Gruppen
Average Linkage	konservativ	ja	alle	
Centroid	konservativ	nein	Distanzmaße	
Median	konservativ	nein	Distanzmaße	
WARD	konservativ	ja	Distanzmaße	bildet etwa gleich große Gruppen

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.

[Bac10] Andreas andWenzig Knut Bacher, Johann andPöge. *Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren*. Oldenbourg, München, 3., erg., vollst. überarb. und neu gestaltete aufl. edition, 2010.

[Bac16] Bernd andPlinke Wulff andWeiber Rolf Backhaus, Klaus andErichson. *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*. Springer Gabler, Berlin and Heidelberg, 14., überarbeitete und aktualisierte auflage edition, 2016.

[Ban08] Jürgen Bankhofer, Udo andVogel. *Datenanalyse und Statistik: Eine Einführung für Ökonomen im Bachelor ; [Bachelor geeignet!]*. Lehrbuch. Gabler, Wiesbaden, 1. aufl. edition, 2008.

[Cla09] Ernest andZhang Hao Helen Clarke, Bertrand andFokoue. *Principles and Theory for Data Mining and Machine Learning*. Springer Series in Statistics. Springer New York, New York, NY, 1. aufl. edition, 2009.

[Eck02] Reinhold andRengers Martina Eckey, Hans-Friedrich andKosfeld. *Multivariate Statistik: Grundlagen - Methoden - Beispiele*. Gabler Verlag, Wiesbaden, 2002.

[Ped10] Witold Pedrycz. *Knowledge-based clustering: From data to information granules*. Wiley, Hoboken, N.J, 2010.

[Pie15] Walter W. Piegorsch. *Statistical data analytics: Foundations for data mining, informatics, and knowledge discovery*. Wiley, Chichester, West Sussex, 2015.

[Pun83] David W. Punj, Girish andStewart. Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research : JMR*, 20(2):134–148, 1983.

[Wu04] Hui andShekhar Shashi Wu, Weili andXiong, editor. *Clustering and information retrieval*, volume 11 of *Network theory and applications*. Kluwer Acad. Publ, Norwell, Mass., 2004.

[Xu99] Xiaowei Xu. *Efficient clustering for knowledge discovery in spatial databases: Zugl.: München, Univ., Diss., 1998*. Berichte aus der Informatik. Shaker, Aachen, als ms. gedr edition, 1999.