Clusteranalyse

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

eingereicht am 13. 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 13. 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. Die Wahl eines Proximitätsmaß für eine Clusteranalyse hängt maßgeblich davon ab, ob die Clusteranalyse Objekte oder Variablen zu klassifizieren versucht (Vgl. Bacher et al. [Bac10], S. 196).

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).

[Bac10] S. 200 Alle Ähnlichkeitsmaße $\ddot{a}$ lassen sich durch

$$u_{ij} = 1 - \ddot{a}_{ij} \quad \text{bzw.} \quad u_{g,g*} = 1 - \ddot{a}_{g,g*} \tag{4.1}$$

in Distanzmaße u umwandeln.

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

4.1 Proximitätsmaße für dichotome Merkmale

Es kann von dichotomen Merkmalen (auch binären Merkmalen) gesprochen werden, wenn deren Modalitäten mit 0 und 1 kodiert werden; polytome Merkmale hingegen besitzen mehrere Ausprägungen. (Vgl. Eckey et al. [Eck02], S. 218).

Um die Kombinationen der Merkmalsausprägungen bei einem Objekt- bzw. Variablenpaar zu definieren, kommt oft eine Kontingenztabelle (auch *Vierfeldertafel*) zum Einsatz:

		Variable j oder Objekt $g*$		
		0	1	$\sum$
Variable i oder Objekt g	0 (Nichtbesitz)	a	b	a + b
	1 (Besitz)	c	d	c + d
	$\sum$	a + c	b + d	a + b + c + d = m

Die Variablen *a* und *d* geben Auskunft über Übereinstimmungen. Die Variable *a* gibt dabei Auskunft zur Übereinstimmung des Nichtbesitzes, dass bedeutet beide betrachteten Objekte besitzen das untersuchte Merkmal nicht. Die Variable *d* gibt Auskunft zur Übereinstimmung des Besitzes, dass bedeutet beide betrachteten Objekte besitzen das untersuchte Merkmal. Die Variablen *b* und *c* geben Auskunft zur Nichtübereinstimmungen, dass heißt nur eines der Objekte besitzt das untersuchte Merkmal.

Ausgehend von Bacher et al. ([Bac10], S. 199)

$$\ddot{a}_{ij} \quad \text{bzw.} \quad \ddot{a}_{g,g*} = \frac{\alpha \cdot a + \beta \cdot d}{\delta \cdot a + \beta \cdot d + \gamma \cdot (b + c)} \tag{4.2}$$

lassen sich durch die unterschiedlichen Gewichtungsfaktoren α, β, γ und δ verschiedene Ähnlichkeitsmaße herleiten. Dabei gehen je nach Ähnlichkeitsmaß manche Variablen mehr und manche weniger in die Berechnung ein. Die Maße sind nach dem jeweiligen Entwickler benannt. Nachfolgend sind einige Ähnlichkeitsmaße, nach Bacher et al. ([Bac10], S. 200) mitsamt ihrer Berechnung dargestellt.

Ähnlichkeitsmaß	Berechnungsformel	Eigenschaften
Jaccard I	$\frac{0 \cdot a + 1 \cdot d}{1 \cdot a + 1 \cdot d + 1 \cdot (b+c)}$	Gemeinsamer Nichtbesitz geht nicht in Berechnung ein
Dice	$\frac{0 \cdot a + 2 \cdot d}{1 \cdot a + 1 \cdot d + 1 \cdot (b+c)}$	Gemeinsamer Nichtbesitz geht nicht in Berechnung ein, gemeinsamer Besitz doppelt
Sokal & Sneath I	$\frac{0 \cdot a + 1 \cdot d}{1 \cdot a + 1 \cdot d + 2 \cdot (b+c)}$	Gemeinsamer Nichtbesitz geht nicht in Berechnung ein, Nichtübereinstimmung doppelt
Russel & Rao	$\frac{0 \cdot a + 1 \cdot d}{1 \cdot a + 1 \cdot d + 1 \cdot (b+c)}$	Gemeinsamer Nichtbesitz wird nicht als Ähnlichkeit betrachtet, geht aber in den Nenner ein
Simple-Matching	$\frac{1 \cdot a + 1 \cdot d}{1 \cdot a + 1 \cdot d + 1 \cdot (b+c)}$	Alles wird gleich gewichtet
Sokal & Sneath II	$\frac{2 \cdot (a+d)}{2 \cdot (a+d) + 1 \cdot (b+c)}$	Übereinstimmungen werden doppelt gewichtet
Rogers & Tanimoto	$\frac{1 \cdot a + 1 \cdot d}{1 \cdot a + 1 \cdot d + 2 \cdot (b+c)}$	Nichtübereinstimmung wird doppelt gewichtet

4.2 Proximitätsmaße für polytome Merkmale

Polytome Merkmale können auf einem nominalen oder ordinalen Skalenniveau gemessen werden und immer anhand einer Dichotomisierung durch mehrere dichotome Merkmale (sogenannte Dummy-Variablen) ersetzt werden. Ein Merkmal mit *r* Merkmalsausprägungen erfordert somit *r* Dummy-Variablen (Vgl. Bankhofer/Vogel [Ban08], S. 159). Hierbei ist aber vor allem darauf zu achten, dass ein erheblicher Informationsverlust bei der Dichotomisierung ordinal skalierter Merkmale in Kauf genommen werden muss. In diesem Fall würde sich eher eine Distanzmessung über die Rangdistanz anbieten, die vorteilhaft anwendbar ist, sofern die Ränge als Intervallskala interpretiert werden und die meisten Merkmale sich in ihren Ausprägungen unterscheiden (Vgl. Eckey et al. [Eck02], S. 225).

4.3 Proximitätsmaße für hierarchische Merkmale

Bei hierarchischen Merkmalen wird von einzelnen Hierarchieebenen ausgegangen, denen ein Wert stellvertretend für die Aggregation der Hierarchie zugewiesen wird. Ausgehend von einer Baumstruktur der verästelten hierarchischen Merkmale wird untersucht, auf welcher Hierarchieebene zwei Merkmalsausprägungen zusammentreffen. Anhand dieser Beobachtung kann ihnen dann eine spezifische Distanz zugeordnet werden (Vgl. Bankhofer/Vogel [Ban08], S. 160).

4.4 Proximitätsmaße für quantitative Merkmale

Bei einer objektorientierten Clusteranalyse wird vor allem von Distanzmaßen ausgegangen, die sich aus der verallgemeinerten Minkowski-Metrik[1] ableiten (Vgl. Bacher et al. [Bac10], S. 219):

$$d(q,r)_{g,g*} = \left[\sum_i |x_{gi} - x_{g*i}|^r\right]^{\frac{1}{q}} \tag{4.3}$$

Anhand der Parameter lassen sich verschiedene Distanzmaße für einzelne Merkmale unterscheiden:

$r = 1$	und	$q = 1$	für die City-Block-Distanz
$r = 2$	und	$q = 2$	für die euklidische Distanz
$r = 2$	und	$q = 1$	für die quadrierte euklidische Distanz
$r = \infty$	und	$q = \infty$	für die Chebychev-Distanz

[Bac10] S. 219

Der Parameter *r* bestimmt, wie stark größere Unterschiede in wenigen Variablen bzw. wie kleinere Unterschiede in vielen Variablen Einfluss nehmen. Beispielsweise gehen bei der City-Block-Distanz Merkmalswertdifferenzen proportional zu ihrem Ausmaß in die Objektdistanz ein, während Merkmalswertdifferenzen bei der euklidischen Distanz einen überproportional großen Einfluss besitzen (Vgl. Eckey [Eck02], S. 212). Die Verwendung des Namens L_p-Distanzen ist durchaus üblich. Die Unterscheidung erfolgt hierbei durch den Parameter *p* mit *p* = *q* = *r*, wodurch die quadrierte euklidische Distanz für gewöhnlich nicht zu den L_p-Distanzen gezählt wird.

Des weiteren spielt im Falle von hoch korrelierten Merkmalen die Malahanobis-Distanz eine wichtige Rolle, da sie Korrelationen und Varianzen zwischen den Variablen beseitigt:

$$MAHA(x_g, x_{g*}) = (x_g - x_{g*}) \cdot S^{-1} \cdot (x_g - x_{g*}) \tag{4.4}$$

mit x_g und x_{g*} als Merkmalsvektoren der Objekte *g* bzw. *g**. S^{-1} bezeichnet die Inverse der zugrundeliegenden Kovarianzmatrix (Vgl. Bacher et al [Bac10], S. 339 und Bankhofer/Vogel [Ban08], S. 168).

[1] In der Literatur wird oft auch *q* = *r* gesetzt. Hier spricht man von der gewöhnlichen Minkowski-Metrik.

5 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.

5.1 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:

$$D(R; P + Q) = min\{D(R, P); D(R, Q)\} \tag{5.1}$$

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) Klemm S. 19 Grafik 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.

5.2 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:

$$D(R; P+Q) = max\{D(R,P); D(R,Q)\} \tag{5.2}$$

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)

5.3 Average Linkage-Verfahren

Das Average Linkage-Verfahren ist eine Modifikation vom Single und Complete Linkage-Verfahren. Dadurch wird versucht die Vorteile der beiden Verfahren zu kombinieren und die Nachteile zu eliminieren. Die Distanz berechnet sich dabei auf Grundlage einer Durchschnittsbildung von den Distanzen der vereinten Objekten. Nach Bacher et al. [Bac10], S.264 lässt sich der Mittelwert mithilfe von drei verschiedenen Verfahren berechnen:

1. Average-Linkage

   $$u_{(p+q),i} = \frac{u_{p,i} + u_{q,i}}{2} \tag{5.3}$$

   Hier findet eine einfache Durchschnittsbildung statt. Everitt et al. ([Eve11], S. 79) beschreiben dieses Verfahren als relativ robust und dazu geneigt, Cluster mit geringen Varianzen zu fusionieren.

2. Weighted-Average-Linkage

   $$u_{(p+q),i} = \frac{u_{p,i}n_p + u_{q,i}n_q}{n_p + n_q} \tag{5.4}$$

   Hier werden Merkmalsausprägungen in kleinen Clustern höher gewichtet als Merkmalsausprägungen in großen Clustern (Vgl. Everitt et al. [Eve11], S.79)

3. Within-Average-Linkage selbe Formel wie Weighted-Average-Linkage? -¿ ja, weil sie sich nur minima

Dabei bezeichnet $u_{p,i}$ die Unähnlichkeit zwischen den Clustern p und i, $u_{q,i}$ die Unähnlichkeit zwischen den Clustern q und i. Dieses Verfahren ist ein konservatives Verfahren. (Vgl. Bacher [Bac10], S.264-266)

5.4 Centroid-Verfahren

Das Centroid-Verfahren benötigt zur Berechnung quantitative Daten. Bei diesem Verfahren werden Clusterzentren gebildet, welche als Repräsentaten für den jeweiligen Cluster dienen. Dabei werden jewils die Clusterpaare zusammengefasst, deren euklidische Distanz zum Clusterzentrum am geringsten ist. Nachdem die Clusterpaare zusammengefasst wurden muss das Clusterzentrum neu berechnet werden. Die ständige Neuberechnung des Clusterzentrums kann man umgehen, wenn man stattdessen die Unähnlichkeit der Objekte berechnet. Das Centroid-Verfahren ist ein konservatives Verfahren. (Vgl. Bacher et al. [Bac10],S. 285-289) Nach Everitt et al. ([Eve11], S. 79) dominiert nach einer Fusionierung der größere Cluster von den ursprünglichen beiden. Eckey et al. ([Eck02], S. 243) weisen auf das Problem hin, dass die Bildung eher heterogener Cluster nicht auszuschließen ist, sofern die Clusterzentren nur nahe genug beieinander liegen.

Reminder: Hier kommt auch das Problem mit den Umkehrungen/Inversionen für Centroid, Median, Ward

5.5 Median-Verfahren

Das Median-Verfahren benötigt genauso, wie das Centroid-Verfahren quantitative Daten zur Berechnung. Ebenso werden die Clusterpaare zusammengefasst, deren euklidische Distanz zu dem Clusterzentrum minimal ist. Nach dem Zusammenfassen der Clusterpaare wird das neue Clusterzentrum errechnet. Dafür wird der Median aus den zwei alten Clusterzentren gebildet. Dies geschieht nach Bacher [Bac10] nach folgender Formel:

$$x_{(p+q),j} = \frac{x_{pj} + x_{qj}}{2} \tag{5.5}$$

Hier ist ebenso die ständige Neuberechnung der Clusterzentren nicht notwendig, wenn man stattdessen die Unähnlichkeit der Objekte berechnet. Dabei ist das Verschmelzungsschema ähnlich wie bei dem Average Linkage-Verfahren schwer zu interpretieren (Vgl. Bacher et al. [Bac10], S. 285.289). Neu entstehende Cluster liegen in der Mitte der beiden ursprünglichen Cluster. [Eve11] S. 79 Im Gegensatz zum Centroid-Verfahren berücksichtigt das Median-Verfahren die Besetzungszahlen der beiden Cluster bei der Fusion nicht. Bei gleicher Objektanzahl können beide Verfahren allerdings als identisch angesehen werden (Vgl. Eckey et al. [Eck02], S. 242).

5.6 Ward-Verfahren

Das Ward-Verfahren wird oft als sehr guter Clusteralgorithmus gesehen, da es im Vergleich zu anderen Verfahren sehr gute Partitionen findet und Objekte ”richtig”den Clustern zuordnet. Dazu müssen jedoch folgende Bedingungen gegeben sein, da dieses Verfahren sehr sensibel reagiert:

- Verwendung eines geeigneten Distanzmaßes
- alle Variablen müssen metrisches Skalenniveau
- Ausreißer müssen vor Anwendung des Verfahrens eliminiert werden

- alle Variablen müssen unkorreliert sein
- die Elementanzahl sollte in jeder Gruppe als in etwa gleich groß erwartet werden
- die einzelnen Cluster müssen die gleiche Ausdehnung besitzen

Im Verfahren werden jeweils die Objekte zusammengefasst, welche die Streuung innerhalb eines Cluster möglichst wenig erhöhen. Die Cluster werden also so gebildet, dass sie möglichst homogen sind (Vgl. Backhaus et al. [Bac16] S.484).
Ward neigt jedoch dazu, möglichst gleich große Cluster zu bilden und langgestreckte Cluster bzw. kleine Cluster mit wenigen Objekten nicht zu erkennen [Bac16] S. 489, auch Grafik

Abbildung kann rein, Fusionierungseigenschaften vielleicht unter 2.Clusterlösungen erklären?

Idee: Inversionen ja/nein einfügen

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

6 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.

Nach Backhaus [Bac16], S. 510 muss man einige Vorüberlegungen anstellen um das ideale Clusterverfahren auswählen zu können. Die Vorüberlegungen sind folgende:

1. Anzahl der Objekte
2. Problem der Ausreißer
3. Anzahl der betrachteten Merkmale
4. Gewichtung der Merkmale
5. Vergleichbarkeit der Merkmale

Bei dem Single Linkage-Verfahren hat man keine Probleme mit Ausreißern, da man dieses Verfahren zum Teil sogar nutzt um Außreißer zu identifizieren. Bei dem Complete Linkage-Verfahren kann man nur schlecht Ausreißer erkennen, das Verfahren ist aber auch mit Ausreißern in den Daten durchführbar. Bei dem Ward-Verfahren muss man vor der Ausführung die Ausreißer elimienieren um ordentliche Ergebnisse zu erhalten.

Vielleicht lieber andere Kriterien als bei Backhaus?? (z.B. Proximitätsmaß, Ausreißer, Gruppengröße) -¿ Kö

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.

[Boc74] Hans Hermann Bock. *Automatische Klassifikation: Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten ; (Cluster-Analyse)*, volume 24 of *Studia mathematica*. Vandenhoeck & Ruprecht, Göttingen, 1974.

[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.

[Eve11] Brian S. Everitt. *Cluster analysis*. Wiley series in probability and statistics. Wiley, Chichester, 5. ed. edition, 2011.

[Gor99] Allan D. Gordon. *Classification*, volume 82 of *Monographs on statistics and applied probability*. Chapman & Hall/CRC, Boca Raton, Fla., 2. ed. edition, 1999.

[Kle95] Elmar Klemm. *Das Problem der Distanzbindungen in der hierarchischen Clusteranalyse: Zugl.: Berlin, Techn. Univ., Diss., 1995*, volume 271 of *Europäische Hochschulschriften Reihe 22, Soziologie*. Lang, Frankfurt am Main, 1995.

[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 re-*

trieval, 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.