UNIKASSEL
VERSITÄT

Institut für Betriebswirtschaftslehre
Dialog Marketing Competence Center

Masterthesis – Exposé

Bestimmung der optimalen Clusterzahl in der empirischen Marktforschung

Vorgelegt von: Karina Cvetkova (33263005)
Hinter der Brücke 37
34134 Kassel
Tel.: 0152/34511704

Vorgelegt bei: Prof. Dr. Ralf Wagner

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abstrakt

Titel - Bestimmung der optimalen Clusterzahl in der empirischen Marktforschung

Theoretischer Hintergrund – Die Clusteranalyse ist ein statistisches Datenanalyseverfahren, das auf einer Gruppenbildung von ähnlichen Objekten basiert. Die Anzahl der Gruppen ist dabei nur selten offensichtlich und muss mit geeigneten Clusteralgorithmen erst ermittelt werden. Auf Grund der Tatsache dass die Festlegung einer „richtigen" Clusteranzahl eine Voraussetzung für eine gelungene Clusteranalyse darstellt und zugleich eine absolute Herausforderung ist, haben Wissenschaftsforscher verschiedene Lösungsansätze dazu entwickelt.

Methode – Nach der Einführung in die Clusteranalyse und deren unterschiedlichen Clusteralgorithmen, beschäftigt sich die vorliegende Masterthesis mit existierenden Methoden zur Bestimmung der optimalen Clusteranzahl. Von einer Reihe existierender Lösungsansätze zur Festsetzung der Clusteranzahl werden in dieser Arbeit acht Ansätze erläutert, darunter „Dendrogramm", „Elbow-Kriterium", „Gap Statistik", „Ward-Methode", „K-Means-Methode", „simple overage linkage", „Silhoutten-Koeffizient", „Gitter-Methode". Anschließend erfolgt eine Bewertung jedes einzelnen Verfahrens.

Schlüsselwörter - Clusteranalyse, Ähnlichkeits- und Distanzmatrix, hierarchische Clustererfahren, Dendrogramm, partitionierende Clusterverfahren, Clusteranzahl.

1. Einleitung

Hinführung zum Thema

In der heutigen Gesellschaft sind Informationen wichtiger denn je und bedeuten für sowohl Unternehmen als auch für Privatpersonen oft den maßgeblichen Wissensvorsprung und somit gleichzeitig auch den Wettbewerbsvorteil. Für die Unternehmen, die über die Jahre hinweg große Mengen an Daten gesammelt haben, bedeutet es nicht nur eine bloße Datensammlung, um beispielsweise Kontakt zu ihren Kunden aufzunehmen, es ist vielmehr eine Datenbank mit wichtigen Informationen u.a. über das Kaufverhalten der Kunden (Mooi, Sarstedt, 2010). Gleichzeitig führt der technischer Fortschritt zum rasanten Anstieg der Möglichkeiten der elektronischen Erfassung und Speicherung der großen Datenmengen. Das Volumen und die Komplexität der Daten ist demnach in den letzten Jahren enorm gewachsen. Extrahieren des potenziell wichtiges Wissens aus der Menge von den zu verarbeitenden Daten stellt daher eine absolute Herausforderung dar. Vor diesem Hintergrund wird der Verwendung von unterschiedlichen Datenanalyseverfahren bzw. Segmentierungsmodellen eine hohe Bedeutung zugeschrieben. (Sprenger, 2005). Die häufig angewandte Methode bei der Reduktion von großen und unübersichtlichen Datenvolumen sowie deren Analyse ist die Clusteranalyse, die verschiedene Verfahren zur Zuordnung von einzelnen Elementen bzw. Objekten zu einer Gruppe bzw. zu einem sogenannten Cluster umfasst (Gluchowski, Gabriel, Dittmar, 2008). Unter Anwendung der Clusteranalyse werden die Objekte aufgrund bestimmter Merkmale in möglichst homogene Gruppen eingeteilt und demzufolge kann eine überschaubare Strukturierung der Datenmenge erfolgen, die eine Vergleichbarkeit und Analyse der Daten ermöglicht (Hansen, Jaumard, 1997; Fahrmeir, Hamerle, Tutz, 1996).

Oft wird das Clusterverfahren mit der Klassifikation gleichgesetzt, es gibt jedoch einen wesentlichen Unterschied zwischen den beiden Verfahren: während bei der Klassifikation die Daten den bereits bestehenden Klassen zugeordnet werden, ist das Ziel der Clusteranalyse neue Gruppen zu identifizieren (Anderberg, 2014). Trotz des Unterschieds zwischen der Clusteranalyse und der Klassifikation werden beide Verfahren bei den meisten Autoren entweder auf eine Stufe gestellt und als Synonym verwendet oder die Clusteranalyse wird als Teil der Klassifikation gesehen. In diesem Kontext wird Clustering oft als „unsupervised classification“ bezeichnet oder beispielsweise in der bekannten und oft zitierten Literaturquelle „Journal of Marketing Research“ wird der Zusammenhang der beiden Verfahren wie folgt beschrieben: „Cluster analysis is a statistical model for classification“

(Jain, Murty, Flynn, 1999; Punj, Stewart, 1983). Das Verfahren der Zuordnung bzw. der Klassifikation von Objekten ist gewiss einer der wesentlichen Bestandteile der Wissensgenerierung von jeder Wissenschaft (Kaufman, Rousseuw, 2005). So wurde die Wichtigkeit des Verfahrens bereits vor Jahrhunderten von Philosophen als ein zentraler Punkt der Erkenntnistheorie festgestellt:
"All the real knowledge which we possess, depends on methods by which we distinguish the similar from the dissimilar." (Linnaeus; Genera Plantarum, 1737)

Die Clusteranalyse findet in zahlreichen wissenschaftlichen Disziplinen statt, so unter anderem in der Psychologie (z.B. Erstellung von Patientenprofilen), in der Anthropologie (z.B. Aufdecken von homogenen Kulturregionen), in den Sozial- und Politikwissenschaften (z.B. Typologisierung von Ländern oder Individuen) aber vor allem in der Betriebswirtschaft (z.B. Erstellung von Konsumentenprofilen) (Romesburg, 2009; Kaufman, Rousseuw, 2005). Um auf das letztere näher einzugehen, so werden in der Marktforschung beispielsweise Personen in möglichst ähnliche Gruppen zu Käuferschichten zusammengefasst basierend auf Eigenschaften wie Geschlecht, Alter, Bildung, Einkommen, Einstellungen, Einkaufsgewohnheiten, Interessen und weitere. Dabei wird das Ziel der Marktsegmentierung verfolgt, die unter anderem eine gezielte Kundenansprache und -anpassung ermöglicht. Ein weiteres Beispiel, das in der Marktforschung eingesetzt wird ist das Clustern von Regionen oder Städten um somit unterschiedliche Marketingstrategien für komparable Städte auszutesten. (Janssen, Laatz, 1999).

Die zentrale und gleichzeitig schwierige Frage stellt die Bestimmung der optimalen Clusteranzahl dar, denn diese ist a prioro nicht bekannt und muss aus der Datenmenge zunächst konstruiert werden. Das Herausfinden einer „richtigen" Clusteranzahl spielt dabei eine sehr wichtige Rolle, denn diese bedingt die Stabilität und somit auch die Qualität der Ergebnisse. Grundlegend wird das sogenannte „Elbow-Kriterium" angewendet um die optimale Clusteranzahl zu wählen (Tibshirani, Walther, Hastie, 2001; Backhaus, Erichson, Plinke, Weiber, 2015). Neben „Elbow-Kriterium" gibt es aber eine Reihe weiterer Methoden, die im Laufe der Jahre von Wissenschaftsforschern entwickelt worden sind und sich ebenfalls in der Anwendung zur Festlegung der optimalen Clusteranzahl etabliert haben. Einige davon werden im weiteren Verlauf dieses Kapitels kurz erläutert.

Stand der Forschung

Die unten aufgeführte Tabelle verschafft einen Überblick über die relevante Literatur die sich mit dem Thema der Bestimmung der optimalen Clusteranzahl beschäftigt. Dabei ist anzumerken, dass es tatsächlich eine Menge von Literaturquellen zu dem Thema existiert, jedoch werden in der folgenden Tabelle nur elf davon kurz erläutert. Bei der Auswahl der relevanten Literatur wurde nicht hauptsächlich auf die Häufigkeit des Zitierens einer Literaturquelle geachtet, sondern vor allem auf die Ausführung verschiedener Methoden.

Autor	Jahr	Titel	Quelle	Inhalt
Tibshirani, Walther, Hastie	2001	Estimating the number of clusters in a data set via the gap statistic	Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 63(2)	Schätzung der Anzahl der Cluster mit einer sog. „gap statistic“ (Lückenstatistik). Verwendung eines beliebigen Clusteralgorithmus und Vergleich der Änderung der Streuung der Cluster innerhalb mit der Streuung der zu erwartenden geeigneten Nullverteilung. Laut eine Simulationsstudie übertrifft die Lückenstatistik andere Methoden.
Punj, Stewart	1983	Cluster Analysis in Marketing Research: Review and Suggestions for Application	Journal of Marketing Research, Vol.20(2)	Erläuterung der Clusteranalyse und deren unterschiedlichen Verfahren. Schätzung der Clusteranzahl mit der „Ward`s minimum vaiance“ und „simple overage linkage“.
Sugar, James	2003	Finding the Number of Clusters in a Dataset	Journal of the American Statistical Association, Vol. 98(463)	Entwicklung einer neuen, einfachen, dennoch leistungsstarken nicht-parametrischer Methode. Die Auswahl der Clusteranzahl basiert auf sog. Verzerrung („distortion“), einer Größe, die den durchschnittlichen Abstand zwischen jeder Beobachtung und ihrem nächsten Clusterzentrum misst.

Tan, Broach, Floudas	2007	A novel clustering approach and prediction of optimal number of clusters: global optimum search with enhanced positioning	Journal of Global Optimization, Vol.39(3)	Bestimmung der „richtigen“ Clusteranzahl“ mit Hilfe einer neuen Methode des Clusteralgorithmus, der sog. „Global Optimum Search“. Der Ansatz beinhaltet eine Vorgruppierung von Datenpunkten. Schlecht platzierte Datenpunkte werden entfernt, wodurch eine enge Gruppierung zwischen den Datenpunkten gewährleistet wird. Somit wird die Anzahl der Cluster bis zu einem Optimum erhöht.
Fang, Wang	2012	Selection of the number of clusters via the bootstrap method	Computational Statistics & Data Analysis, Vol.56(3)	Betrachtung des Problems des Auswahl der Clusteranzahl in der Clusteranalyse. Entwicklung eines Schätzverfahrens der Clusterinstabilität basierend auf sog. „bootstrap“. Die Anzahl der Cluster wird so gewählt, dass die entsprechende geschätzte Clusterinstabilität minimiert wird.
Chiang, Mirkin	2010	Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads	Journal of Classification, Vol.27(1)	Bestimmung der Clusteranzahl mit Hilfe von der sog. „intelligenten“ K-Means-Methode (engl. ik-means). Finden der optimalen Clusteranzahl durch eine Extraktion von „anomalen Mustern“ („anomalous patterns) aus den Daten eins nach dem anderen. Vergleich mit sieben weiteren Methoden.
Hartigan	1985	Statistical theory in clustering	Journal of Classification, Vol.2(1)	Überprüfung verschiedener hierarchischer Clusteralgorithmen. Auswertung der Algorithmen durch ihre Fähigkeit Regionen mit einer hohen Bevölkerungsdichte zu entdecken.

Milligan, Cooper	1985	An examination of procedures for determining the number of clusters in a data set	Psychometrika, Vol.50(2)	Die Bestimmung der Clusteranzahl erfolgt anhand sog. „Monte Carlo-Experimente“. Diese bewertet vier hierarchische Clusterverfahren, die mit Hilfe von künstlichen Datensätzen durchgeführt werden. Die Simulationsergebnisse offenbarten, dass eine Anwendung der sog. „Stoppregel“ (stopping rules) gute unterstützende Funktion bei der Bestimmung der richtigen Clusteranzahl leisten kann.
Fraley, Raftery	2002	Model-Based Clustering, Discriminant Analysis, and Density Estimation	Journal of the American Statistical Association, Vol.97(458)	Untersuchung der allgemeinen Methodik des modellbasierten Clustering sowie Bereitstellung eines allgemeinen statistischen Ansatzes für des Problem der Bestimmung der richtigen Clusteranzahl. Der Ansatz ist auch auf die Diskriminanzanalyse und multivariate Dichteabschätzung übertragbar.
Rousseeuw	1987	Silhouettes: A graphical aid to the interpretation and validation of cluster analysis	Journal of Computational and Applied Mathematics, Vol.20	Entwicklung einer neuen grafischen Darstellung bei partitioniernenden Clusteralgorithmen. Jeder Cluster wird durch eine sog. „Silhouette“ dargestellt, die auf dem Vergleich ihrer Dichtheit und Trennung beruht. Diese Silhouette zeigt welche Objekte innerhalb ihres Clusters gut liegen und welche nur irgendwo zwischen den Clustern liegen. Das gesamte Clustering wird angezeigt, indem die Silhouetten zu einem einzigen Diagramm zusammengefasst werden. Dies ermöglicht eine Bewertung der

				Qualität der Cluster sowie ermöglich eine Übersicht über die Datenkonfiguration
Chiu	1994	Fuzzy Model Identification Based on Cluster Estimation	Journal of Intelligent and Fuzzy Systems, Vol.2(3)	Präsentation einer einfachen und effektiven Methode zur Schätzung der optimalen Clusteranzahl, der sog. „Mountain-Methode“ , die 1992 von Yager und Filev entwickelt wurde. Das Verfahren basiert auf der Darstellungsform eines Gitters. Die Daten werden in das Gitter übertragen und es wird ein potenzieller Wert für jeden Gitterpunkt auf der Grundlage seiner Abstände zu den tatsächlichen Datenpunkten berechnet. Ein Gitterpunkt mit vielen Datenpunkten in der Nähe hat ein hohen Potenzialwert. Der Gitterpunkt mit dem höchsten Potenzialwert wird als Clusterzentrum gewählt.

Tabelle 1: Übersicht der relevanten Literatur.

Zielsetzung und Aufbau der Arbeit

Vor dem dargestellten Hintergrund beschäftigt sich die vorliegende Arbeit mit der Frage der optimalen Clusteranzahl in der Clusteranalyse. Das Ziel dieser Arbeit ist es verschiedene Methoden zur Bestimmung der Clusterzahl aufzuzeigen und diese zu bewerten.
Die Masterarbeit gliedert sich in drei Teile: Zum besseren Verständnis der vorliegenden Arbeit werden im Rahmen des ersten Teils die theoretischen Grundlagen zu dem Thema erarbeitet. Dabei wird auf die Erläuterung des Begriffs der Clusteranalyse sowie deren Anwendungsgebiete in der Marktforschung eingegangen. Im zweiten Teil werden die gängigen Algorithmen der Clusteranalyse vorgestellt. Der Hauptteil der Arbeit befasst sich mit den Methoden zur Bestimmung der Clusteranzahl sowie deren Beurteilung. Abgerundet wird die Arbeit durch ein Fazit und Ausblick.

2. Theoretische Grundlagen

2.1 Einführung in die Clusteranalyse

Clusteranalyse ist ein klassisches, statistisches Verfahren der multivariaten Datenanalyse, dem eine sehr hohe Bedeutung sowohl in der Wissenschaft als auch in der Praxis zugeschrieben wird (Kohrmann, 2003). Das Verfahren basiert auf der Gruppenbildung von Untersuchungsobjekten in einem Datensatz. Das Ziel dabei ist eine Vielzahl von Objekten mit einer möglichst geringer Streuung bzw. mit einem möglichst ähnlichem Informationsgehalt zu Gruppen zusammenzufassen. Gleichzeitig sollten die Gruppen untereinander jedoch möglichst unähnlich sein (Backhaus, Erichson, Weiber, Plinke, 2015). Kurz gefasst besteht die Funktion einer Clusteranalyse darin, eine heterogene Gruppe von Objekten in homogene Untergruppen aufzuteilen. Die Gruppenzuordnung wird dabei als Clustering bezeichnet und die Gruppen mit ähnlichen Objekten als Cluster (Ammann, 2007). In diesem Kontext kann die Clusteranalyse folgendermaßen definiert werden:

(1) „[…] activity of dividing a set of objects into a smaller number of classes in such a way that objects in the same class are similar to one another and dissimilar to objects in other classes." (Gordon, 1987).

(2) "Cluster analysis is the art of findings groups in the data." (Kaufman, Rousseeauw, 2009)

(3) "Ziel der Clusteranalyse (Cluster = Büschel, Häufung) ist, bei simultaner Betrachtung vom mehreren (mehr als zwei) Variablen die einzelnen Objekte (Merkmalsträger) so zu Gruppen zusammenzufassen, dass die Ähnlichkeit der Objekte in Gruppen möglichst groß ist, die Ähnlichkeit zwischen den Gruppen aber möglichst gering ist." (Martens, 2003)

Die Clusterbildung beruht somit auf folgenden zwei Prinzipien (Bacher, Pöge, Wenzig, 2010):

- *Innere Homogenität*: Innerhalb der Cluster soll Homogenität vorliegen, das heißt, dass die Unterschiede innerhalb der einzelnen homogenen Gruppen minimal sein sollen.
- *Externe Heterogenität*: Zwischen der Cluster soll Heterogenität vorliegen, das heißt, dass die Unterschiede zwischen den einzelnen homogenen Gruppen maximal sein sollen.

Es ist zu betonen, dass die Clusteranalyse trotz ihres Verfahrens der Gruppeneinteilung, grundsätzlich ein Instrument der Aufdeckung von Gruppen ist. Es dient dem Aufdecken von Strukturen, die im Datensatz vorhanden sind, die jedoch anderenfalls nicht sichtbar wären (Rabe-Hesketh, Everitt 2004).

Die unten aufgeführte Abbildung 1 zeigt in einer vereinfachten Darstellung das Prinzip der Einteilung der vielfältigen und unstrukturierten Daten bzw. Elemente in Cluster. (Während auf der linken Seite die Elemente unstrukturiert „in einem Raum“ dargestellt sind, so sind diese Elemente nach einer Clusteranalyse nach ihrer Form zu Gruppen zusammengefasst.) Die gebildeten Cluster werden dann als neue und eigenständige Objekte betrachtet.

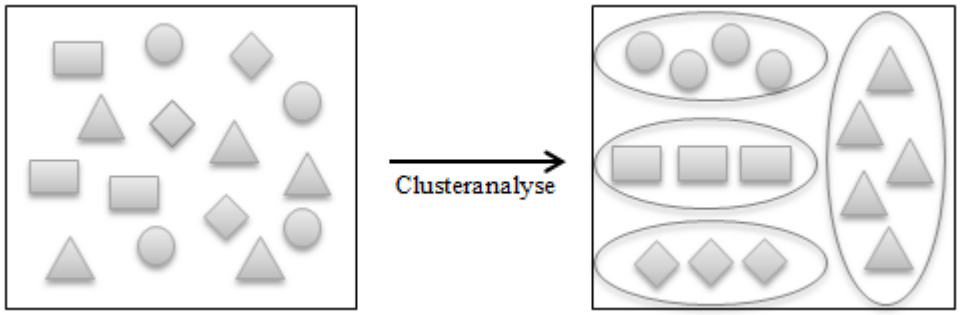

Abbildung 1: Das Prinzip der Einteilung des Datensatzes. Eigene Erstellung.

Die Einsatzfelder der Clusteranalyse sind sehr vielseitig, jedoch wird diese häufig vor allem im Bereich der Marktforschung zur Marktsegmentierung angewandt. Die Objekte können hierbei nicht nur Individuen (z.B. befragte Personen) sein, sondern beispielsweise auch Gegenstände (z.B. verschiedene Produkte) oder auch Aggregate (z.B. Länder oder Organisationen). Handelt es sich um Verbraucher bzw. potenzielle Verbraucher / Abnehmer, so werden diese durch bestimmte Merkmale charakterisiert und dann mittels ihrer Merkmalsausprägungen in Gruppen zusammengefasst. Die Angehörigen innerhalb einer Gruppe sind ähnlich, unterscheiden sich jedoch von Angehörigen anderer Gruppen. (Damit ist das zuvor erläuterte Prinzip der internen Homogenität und der externen Heterogenität beleuchtet.) Anhand der Clusteranalyse können somit einzelne Verbrauchertypen selektiert werden und infolgedessen können Marketingmaßnahmen gezielt auf die unterschiedlichen Kundengruppen angepasst werden. (Hering, Mühleisen, 1987).

Um einen besseren Einblick in die praktische Anwendung der Clusteranalyse sowie in deren vielseitigen Einsatzfelder zu gewinnen, so werden nachfolgend zwei Beispiele aus der Marktforschung geschildert.

Beispiel 1: Im Jahr 2013 beschäftigten sich Vinerean, Cetina, Dumitrescu und Tichindelean mit der Frage „who are the people interacting online and how engaged are they in online activities?“ (S.66). Es wurden 236 Social-Media-Nutzer betrachtet. Mittels einer Clusteranalyse wurde zunächst eine Segmentierung der Nutzer in verschiedene Typen vorgenommen und anschließend wurde untersucht wie sich verschiedene Merkmale der Online-Werbungen in den sozialen Netzwerken auf die Nutzergruppen auswirken. Das Ziel der Studie war es herauszufinden welche Typen von Nutzern überhaupt existieren und wie diese auf verschiedene Eigenschaften der Online-Werbungen reagieren um somit Online-Marketing-Strategien effektiver gestalten zu können und infolgedessen deren Wirkung maximieren zu können.

Beispiel 2: Sehr oft wird die Clusteranalyse zur Marktsegmentierung im Bereich des Tourismus angewandt. Als ein Beispiel hierzu untersuchten Kau und Lim im Jahr 2005 den Tourismus von China nach Singapur. Das Ziel war dabei die chinesischen Touristen nach ihren Beweggründen für die Reise nach Singapur zu segmentieren. Die 240 befragten Besucher im Februar 2003 von China nach Singapur konnten in vier Hauptsegmente gegliedert werden. Basierend auf den Ergebnissen wurden Vorschläge gemacht wie Singapur ihre Marketingstrategien weiterentwickeln könnte. Im Jahr 2003 war China noch das drittgrößte Touristenland für Singapur, jedoch bereits im Jahr 2004 stieg die touristische Ankunft aus China so stark an, dass es die zweitgrößte Touristengruppe nach Indonesien wurde.

2.2 Vorgehensweise einer Clusteranalyse

Die Vorgehensweise einer Clusteranalyse gliedert sich in mehrere Schritte:

1. Auswahl und Aufbereitung der Variablen
2. Bestimmung der Distanzen bzw. Unähnlichkeiten zwischen allen betrachteten Merkmalen.
3. Auswahl und Anwendung eines Fusionierungsalgorithmus zur Gruppeneinteilung.
4. Bestimmung der optimalen Clusterzahl
5. Validitätsprüfung und inhaltliche Interpretation der Ergebnisse.

Im weiteren Verlauf werden die einzelnen Schritte näher erläutert. Dabei werden in diesem Unterkapitel die Schritte 1. und 2. vorgeführt. Die verschiedenen Clusteralgorithmen werden im Kapitel 3 vorgestellt und der Hauptteil dieser Arbeit, Kapitel 4, befasst sich mit den

verschiedenen Methoden zur Bestimmung der optimalen Clusterzahl sowie deren Beurteilung.

Auswahl und Aufbereitung der Variablen

Um ein gutes Clustering-Verfahren durchführen zu können, müssen zunächst bestimmte Voraussetzungen erfüllt sein (Zaïane, Foss, Lee, Wang, 2002; Martens, 2003):

- *Einheitliches Messniveau der Merkmale:* Die Merkmale der einzuteilende Objekte können unterschiedliches Messniveau besitzen (siehe Tabelle 2). Ist es der Fall, so können die Daten nicht in ein (sinnvolles) Skalenniveau übertragen werden. Damit das Problem ausgeschlossen werden kann, so sollen alle Merkmale in einem einheitlichen Messniveau (binär, nominal/ordinal oder metrisch) betrachtet werden. Anderenfalls ist keine Vergleichbarkeit möglich.

Binär	Zwei Ausprägungen 0 und 1 (Beispiel: Geschlecht)
Nominal/Ordinal	Mehr als zwei Ausprägungen (Beispiel: Kundenstatus gemessen an fünf Ausprägungen von 0 bis 4)
Metrisch	Zahl (Beispiel: Einkommen, Alter)

Tabelle 2: Skalenniveau von Merkmalen. Eigene Erstellung.

- *Standardisierung der Variablen:* Im Fall, dass Variablen große Unterschiede im Bezug auf ihren Wertebereich aufweisen, sollten diese mit Hilfe einer z-Transformation standardisiert werden. Die z-Tansformation ermöglicht eine Vergleichbarkeit indem diese gewährleistet, dass der Mittelwert gleich Null ist ($\mu = 0$) und der Streuungswert / Standardabweichung gleich Eins ist ($\sigma=1$). Sind die Wertebereiche gleich, so ist eine z-Transformation überflüssig und daher nicht notwendig (Zöfel, 2000).

- *Anzahl der Merkmale:* Für das Ergebnis der Clusteranalyse spielt die Auswahl der Merkmale, nach denen die Daten gruppiert werden sollen, eine entscheidende Rolle. Möchte man beispielsweise Personen als Objekte betrachten, so kann man diese sehr einfach in zwei Gruppen unterteilen, indem man alle Frauen einem Cluster und alle Männer einem anderen Cluster zuordnet. Somit hätte man zwei in sich homogene und voneinander heterogene Cluster. In diesem Fall erfolgt Clustering ausschließlich über das Geschlecht, d.h. über nur eine Merkmalsausprägung. Oftmals ist es nicht ausreichend bzw. nicht zielführend nur eine Variable zu betrachten, sondern in der Regel ist eine

Kombination von mehreren Variablen bzw. Merkmalen erforderlich. Die Größe der Kombination spielt dabei eine ausschlaggebende Rolle: Berücksichtigung von zu wenigen Merkmalen führt zugleich zu einer geringen Anzahl von Clustern, die sich jedoch eventuell auch weiter differenzieren ließe wenn man zusätzliche Merkmale beachten würde. Berücksichtigung von zu vielen Merkmalen kann zu wenig differenzierten Clustern führen. Damit aussagekräftige und bedeutungsvolle Ergebnisse mittels der Clusteranalyse erreicht werden können, soll die Auswahl der Merkmale somit sehr bedacht ausgewählt werden.

- *Skalierbarkeit:* "The cluster method should be applicable to huge databases and performance should decrease linearly with data size increase".

- *Eliminierung der Ausreißer:* Als Ausreißer werden diejenigen Objekte genannt, deren Werte im Gesamtvergleich stark abweichen. Solche Extremwerte sollten eliminiert werden, da es anderenfalls zur Beeinflussung des Fusionierungsprozesses kommt und infolgedessen die Zusammenhänge nicht richtig erkannt werden können.

- *Eliminierung von hoch korrelierbaren Variablen:* Durch hoch korrelierbare Variablen kann das Ergebnis durch Überbewertungen verzerrt werden. Aus diesem Grund sollten diese besser ausgeschlossen werden.

Ähnlichkeitsmaße und Distanzmaße

Zur Berechnungsgrundlage der Clusteranalyse werden Zahlenwerte (oder auch Proximitätsmaße genannt) herangezogen, die Ähnlichkeiten oder Unähnlichkeiten zwischen den verschiedenen Elementen paarweise quantifizieren. Das heißt, dass bei der Clusteranalyse entweder Ähnlichkeit oder die Distanz von Objekten bestimmt werden kann. (Ammann, 2007). Dazu stehen zwei Arten zur Verfügung (Schendera, 2010):

1. Mit dem Ähnlichkeitsmaß wird die Ähnlichkeit zwischen zwei Objekten beschrieben. Je höher das Ähnlichkeitsmaß, desto ähnlicher sind sich die Objekte.
2. Mit dem Distanzmaß wird die Unähnlichkeit zwischen zwei Objekten beschrieben. Je niedriger das Distanzmaß, desto ähnlicher sind sich zwei Objekte.

Somit können mithilfe von Ähnlichkeits- und Distanzmaßen unterschiedliche Objekte getrennt und ähnliche Objekte zusammengefasst werden. Dabei ist zu beachten, dass die Ähnlichkeit bzw. die Unähnlichkeit anhand derer die Einteilung der Objekte in Gruppen

vorgenommen wird, nur mittels der Merkmale definiert werden kann, die für alle zu sortierende Objekte in derselben Maßeinheit vorliegen. Besitzen die einzuteilende Objekte unterschiedliche Maßeinheit, so liegt keine Vergleichbarkeit vor und dementsprechend können die Objekte nicht sinnvoll zusammengefasst werden. (Bacher, 1989)
Grundsätzlich ist zu sagen, dass beim Vorliegen von binären Variablen (z.B. (t)=true und (f)=false oder codierte Variablen 0 und 1) Ähnlichkeitsmaße Anwendung finden, beim Vorliegen von der metrischen Datenstruktur (nummerische Werte wie z.B. Einkommen und Alter) Distanzmaße.

Die berechneten Ähnlichkeits- bzw. die Distanzmaße aller Paare der zu zuordneten Objekte werden in einer Distanzmatrix *D* zusammengefasst.

	Objekt 1	Objekt 2	…	Objekt K
Objekt 1	P_{11}			
Objekt 2	P_{21}	P_{11}		
…	…	…	..	
Objekt K	P_{k1}	P_{k2}	P_{k3}	P_{kk}

Abbildung 2: Aufbau einer Distanzmatrix. Eigene Erstellung in Anlehnung an Biemann, 2009, S. 195

Oberhalb der Diagonalen können die Zellen leer bleiben, da sie lediglich eine Spiegelung dieser unterhalb der Diagonalen wären. Die P-Koeffizienten in der Datenmatrix geben den Wert der Ausprägung der Objekte.
Die dann im weiteren Verlauf der Clusteranalyse verwendeten Ähnlichkeits- bzw. Distanzmaßen spielen eine ausschlaggebende Rolle auf die Ergebnisse der Clusteranalyse. Bereits im Jahr 1970 stellte Wishart fest, dass „[…resulting cluster depend more on the underlying similarity criterion than on the physical process of cluster formation.]" (S.1)

<u>Proximitätsmaße bei binärem bzw. nominalen Skalenniveau:</u>
Liegen alle zu untersuchenden Merkmale in Form von codierten Variablen (0 und 1) bzw. von zu Dummy-Variablen umgewandelten nominalen Merkmale, so stehen mehrere Ähnlichkeitsdefinitionen zur Auswahl. Eine der weit verbreiteten Ähnlichkeitsdefinitionen ist die Simple-Matching (1) (Fett, 2008).

$$\frac{a+d}{a+b+c+d} = \frac{a+d}{m} \tag{1}$$

Dabei sind:

- a Anzahl der Variablen bzw. Merkmalen, die bei den beiden Objekten zutreffend sind (Bsp.: $X_i = Y_i = 1$)
- d Anzahl der Variablen bzw. Merkmalen, die bei den beiden Objekten nicht zutreffend sind (Bsp.: $X_i = Y_i = 0$)
- b Anzahl der Variablen bzw. Merkmalen, die nur bei dem Objekt X zutreffend sind ($X_i = 1$, $Y_i = 0$)
- c Anzahl der Variablen bzw. Merkmalen, die nur bei dem Objekt Y zutreffend sind ($X_i = 0$, $Y_i = 1$)
- m Anzahl alle betrachteten Merkmale (a+b+c+d)

Proximitätsmaße bei metrischem Skalenniveau:

Es gibt eine Fülle von Abstandsdefinitionen. Eine der weit verbreiteten Abstandsdefinitionen ist die Euklische Distanz (2) (Meullenet, Xiong, Findlay, 2007). Das Maß der Distanz ergibt sich dabei als Quadratwurzel aus der Summe der Werteabstände zweier Objekten, die quadriert werden (Duran, Odell, 1974). Aufgrund der Quadrierung ist die Gewichtung der großen Differenzwerten stärker als die der kleinen.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{K} (x_{ip} - x_{jp})^2} \tag{2}$$

Eine weitere oft gebräuchliche Abstandsdefinition ist die quadrierte Euklische Distanz (3). Im Gegensatz zu der oben genannten Euklischen Distanz bleibt die Rangordnung der ähnlichen Objektpaare bei der quadrierten Euklischen Distanz gleich.

$$d(x_i, x_j) = \sum_{k=1}^{K} (x_{ip} - x_{jp})^2 \tag{3}$$

Eine ebenfalls oft angewandte Abstandsdefinition ist die City-Block-Distanz (4). Das Distanzmaß wird hierbei als die Summe der absoluten Differenzen zweier Merkmalsausprägungen definiert.

$$d(x_i, x_j) = \left| x_{ip} - x_{jp} \right| \tag{4}$$

Dabei sind:

- ($x_{i,}$ x_j) die Objekte bzw. Merkmale zwischen denen der Abstand gemessen wird,
- k die Indikatoren bzw. die Variablen (k = 1,2,…,K),
- x_{ip} und x_{jp} konkrete Ausprägungen der beiden Objekte x_i und x_j auf der p-ten Variable, die sich aus der Datenmatrix ergeben.
- Der Betrag wird betrachtet um Abweichungen sowohl nach unten als auch nach oben berücksichtigen zu können.

3. Algorithmen der Clusteranalyse

Auf der Grundlage der zuvor erläuterten Proximitätsmaße, die den Abstand zweier Objekte vorgeben, werden in diesem Abschnitt Fusionierungsalgorithmen vorgestellt, die zur Bestimmung der Distanz zweier Clustern herangezogen werden. Dabei stellt die Clusteranalyse zahlreiche Algorithmen zur Verfügung, um Objekte in Cluster einzuteilen. Im weiteren Verlauf dieses Kapitels werden jedoch nur einige davon vorgestellt.

Generell lassen sich die Clusterverfahren in zwei grundlegende Kategorien einteilen: hierarchische und partitionierende Verfahren. In der folgenden Abbildung 3 ist eine Übersicht ausgewählter gängiger Clusterverfahren abgebildet.

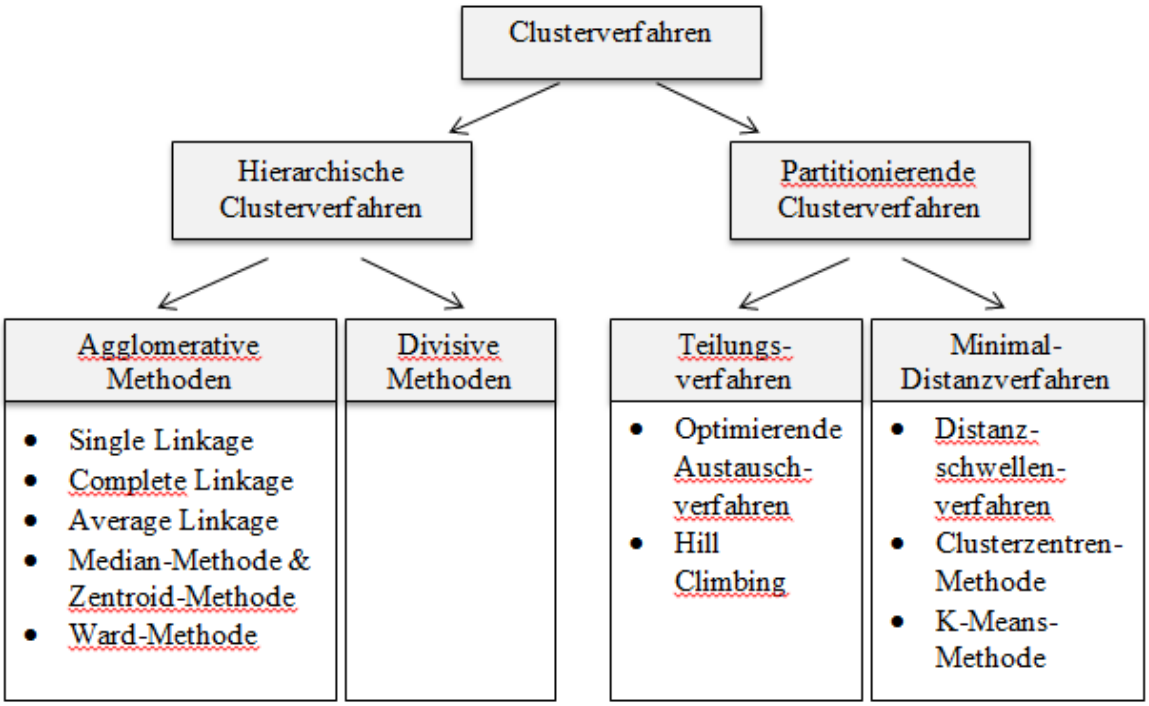

Abbildung 3: Übersicht der Clustererfahren. Eigene Erstellung.

Hierarchische Clusteranalyse

> *„Hierarchical methods represent the solution as a tree structure that explains the history of combining or splitting groups."* (Akkucuk, 2011, S.19).

Die hierarchische Clusteranalyse wird in Form eines Stammbaumes dargestellt. Am Anfang des Verfahrens bildet jedes Objekt ein eigenes Cluster. Mit Hilfe von verschiedenen Algorithmen wird die zu Beginn gegebene Objektmenge im stetigen Fortgang miteinander fusioniert bis nur ein Cluster bleibt. Das heißt, dass es mit der „feinsten" Einteilung beginnt

und mit der „gröbsten“ Einteilung endet, bei der alle Elemente in einem einzigen Cluster versammelt sind. Die „richtige“ Clusteranzahl liegt dabei irgendwo zwischen den beiden Extremen, wobei die Entscheidung allein dem Anwender obliegt.

„Cutting the tree at any point reveals the solution with the desired number of clusters. This is practical when the number of objects studies is very small.” (Akkucuk, 2011, S.19)

4. Bestimmung der optimalen Clusteranzahl

Die Qualität der Lösungen einer Clusteranalyse ist stark von der angenommenen Clustermenge abhängig. Das folgende Kapitel befasst sich mit verschiedenen Ansätzen zur Bestimmung der „richtigen“ Clusteranzahl.

Wird die Clusteranalyse mit einer vorgegeben Clusteranzahl durchgeführt, so besteht das Risiko, dass wichtige Zusammenhänge verloren gehen wenn diese von der optimalen Anzahl der Cluster abweichen (Bostanci, 2011). In der folgenden Abbildung wird der Zusammenhang zwischen einer optimalen und einer nicht optimalen Clusteranzahl verdeutlicht:

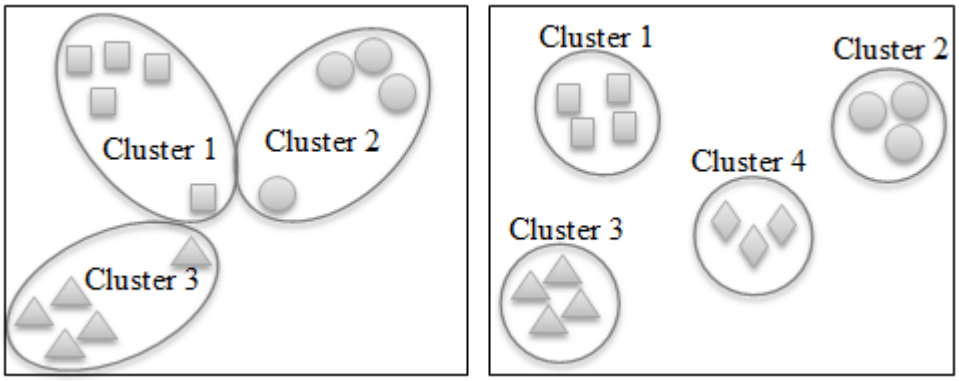

Abbildung 4: Nicht optimale und optimale Clusteranzahl,

5. Inhaltsverzeichnis

6. Zeitplan

Zeit	Aktivität
Februar 2017	-Beendigung der theoretischen Grundlagen -Clusteralgorithmen -Weitere Literaturrecherche
März-April 2017	-Verbesserung der Kapitel 1-3 -Methoden zur Bestimmung der Clusteranzahl und deren Bewertung -Weitere Literaturrecherche -Beendigung des Kapitels 4
Mai 2017	-Fazit und erste Korrektur, Finale Korrektur

Literaturverzeichnis

Akkucuk, U. (2011). A Study on the Competitive Positions of Countries Using Cluster Analysis and Multidimensional Scaling. *European Journal of Economics, Finance and Administrative Sciences, 37, pp. 17-26*

Ammann, P. (2007). Marktsegmentierung: Erfolgsnischen finden und besetzen. 2.Auflage. Symposion Publishing GmbH, Düsseldorf.

Anderberg, M.R. (2014). Cluster Analysis for Applications: Probability and Mathematical Statistics. Acadamic Press, New York.

Bacher, J. (1989). Einführung in die Clusteranalyse mit SPSS-X für Historiker und Sozialwissenschaftler. *Historical Social Research 14, 2, pp. 6-167.*

Bacher, J., Pöge, A., Wenzig, K. (2010). Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren. 3.Auflage. Oldenbourg Verlag München.

Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (2015). Multivariate Analysemethoden: Eine anwendungsorientierte Einführung, 14. Auflage. Springer Gabler, Berlin.

Biemann, T. in: Albers, S., Klapper, D., Konradt, U., Walter, A., Wolf, J. (Hrsg.) (2009). Methodik der empirischen Forschung. 3. Auflage. Springer, Wiesbaden.

Bostanci, H. (2011). Clusterbasierte Datenanalyse auf Grundlage genetischer Algorithmen in SAP-BI: Ein Verfahren zur selbstständigen Ermittlung der optimalen Anzahl Cluster. Diplomatica Verlag, Hamburg.

Chauhan, R., Kaur, H., Alam, M.A. (2010). Data Clustering for Discovering Clusters in Spatial Cancer Databases. *International Journal of Computer Applications, 10 (6), pp. 9-14.*

Chiang, M.M., Mirkin, B. (2010). Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads. *Journal of Classification, 27 (1), pp. 3-40.*

Chiu, S.L. (1994). Fuzzy Model Identification Based on Cluster Estimation. *Journal of Intelligent and Fuzzy Systems, 2 (3), pp. 267-278.*

Duran, B.S., Odell, P.L. (1974). Cluster Analysis: A Survey. Springer-Verlag Berlin, Heidelberg.

Fahrmeir, L., Hamerle, A. and Tutz, G. (1996). Multivariate statistische Verfahren. 2.Auflage. Walter de Gruyter, Berlin.

Fang, Y., Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis, 56 (3), pp. 468-477.*

Fett, K. (2008). Clusteranalyse in CRM, Sales und Marketing: Grundlagen und praktische Anwendung. Books on Demand, Norderstedt.

Fraley, C., Raftery, A.E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association, 97 (452), pp. 611-631*

Gluchowski, P., Gabriel, R., Dittmar, C. (2008). Management Support Systeme – Computergestützte Informationssysteme für Fach- und Führungskräfte. 2. Auflage. Springer-Verlag Berlin Heidelberg.

Gordon, A. D. (1987). A Review of Hierarchical Classification. *Journal of the Royal Statistical Society. 150 (2), pp. 119-137.*

Hansen, P., Jaumard, B. (1997). Cluster analysis and mathematical programming. *Mathematical Programming, 79(1), pp. 191-215.*

Hartigan, J.A. (1985). Statistical theory in clustering. *Journal of Classification, 2 (1), pp. 63-76.*

Hering, E., Mühleisen, U. (1987). Marketing mit dem PC: Basic-Programme für IBM PC und Kompatible. Friedr. Vieweg & Sohn, Braunschweig, Wiesbaden.

Jain, A.K., Murty, M.N., Flynn, P.J. (1999). *Data clustering: a review. Journal ACM Computing Surveys (CSUR), 31 (3), pp.264-323.*

Janssen, J., Laatz, W. (1999). Statistische Datenanalyse mit SPSS für Windows: Eine anwendungsorientierte Einführung in das Basissystem Version 8 und das Modul Exakte Tests. 3. Auflage. Springer-Verlag Berlin Heidelberg.

Kau, A.K., Lim, P.S. (2005). Clustering of Chinese tourists to Singapore: an analysis of their motivations, values and satisfaction. *International Journal of Tourism Research, 7 (4), pp. 231-248.*

Kaufman, L., Rousseuw, P.J. (2005). Finding Groups in Data: An Introduction to Cluster Analysis, 2. edition, JohnWiley & Sons, New Jersey.

Kohrmann, O. (2003). Mehrstufige Marktsegmentierung zur Neukundenakquisition: Am Beispiel der Telekommunikation. 1. Auflage. Deutscher Universitäts-Verlag, Wiesbaden.

Linnaeus, C. (1737). Genera Plantarum.

Martens, J. (2003). Statistische Datenanalyse mit SPSS für Windows. Oldenbourg Wissenschaftsverlag, München.

Meullenet, J.-F., Xiong, R., Findlay, C.J. (2007). Multivariate and Probabilistic Analyses of Sensory Science Problems. Blackwell Publishing Ltd, Oxford.

Milligan, G.W., Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50 (2), pp. 159-179.*

Mooi, E., Sarstedt, M. (2010). A Concise Guide to Market Research. Springer, Berlin Heidelberg.

Punj, G., Stewart, D. O. (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research, 20 (2), pp. 134-148.*

Rabe-Hesketh, S., Everitt, B. (2004). A Handbook of Statistical Analyses using Stata. 3.Auflage, CRC Press, Florida.

Romesburg, C. (2009). Cluster Analysis for Researchers. Lulu Press, North Carolina.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20, pp. 53-65*

Schendera, C. (2010). Clusteranalyse mit SPSS: Mit Faktorenanalyse. Oldenbourg Verlag München.

Schwaiger, M., Meyer, A. (2011). Theorien und Methoden der Betriebswirtschaft: Handbuch für Wissenschaftler und Studierende. Vahlen, München.

Sprenger, T. (2005). Clusteranalyse und Qualitätsmanagement: Visuelle Clusteranalyse. AdNovum Informatik AG, Zürich.

Steinhausen, D., Langer, K. (1977). Clusteranalyse: Einführung in Methoden und Verfahren der automatischen Klassifikation. Walter de Gruyter, Berlin.

Sugar, C.A., James, G.M. (2003). Finding the Number of Clusters in a Dataset. *Journal of the American Statistical Association, 98 (463), pp. 750-763.*

Tan, M.P., Broach, J.R., Floudas, C.A. (2007). A novel clustering approach and prediction of optimal number of clusters: global optimum search with enhanced positioning. *Journal of Global Optimization, 39 (3), pp. 323-346.*

Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63 (2), pp. 411-423.*

Vinerean, S., Cetina, I., Dumitrescu, L., Tichindelean, M. (2013). The Effect of Social Media Marketing on Online Consumer Behavior. *International Journal of Business and Management, 8 (14), pp. 66-79.*

Wishart, D. (1970). The treatment of various similarity criteria in relation to Clustan. University of St. Andrews, Edinburgh.

Zaïane, O.R., Foss, A., Lee C-H., Wang, W. (2002). On Data Clustering Analysis: Scalability, Constraints, and Validation. *Lecture Notes in Computer Science, 2336, pp. 28-39.*

Zöfel, P. (2000). Statistik verstehen: ein Begleitbuch zur computergestützten Anwendung. 1.Auflage. Addison-Wesley Verlag, München.

Eidesstattliche Versicherung

Hiermit versichern wir, dass wir die vorliegende Arbeit selbstständig und ohne unerlaubte Hilfe und Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst und die den benutzten Quellen wörtlich, inhaltlich oder sinngemäß entnommenen Stellen aus veröffentlichten oder unveröffentlichten Schriften als solche kenntlich gemacht haben. Keinen Teil dieser Arbeit haben wir bei einer anderen Stelle zur Erlangung einer Studien- und/oder Prüfungsleistung eingereicht.

Kassel,

Ort, Datum

Karina Cvetkova