Fuzzy-Clusteranalyse: Methoden zur Exploration von Daten mit fehlenden Werten sowie klassifizierten Daten

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von Diplominformatiker Heiko Timm,
geboren am 7. Mai 1970 in Göttingen

Gutachter: Prof. Dr. Rudolf Kruse
Prof. Dr. Frank Klawonn
Prof. Dr. Dietrich Behr

Magdeburg, den 21.6.2002

Inhaltsverzeichnis

Abstract

Finding clusters of homogenous data points is an important task in data analysis. The aim of cluster analysis is to divide a given dataset into clusters of homogenous data. One of the main problems is that sometimes clusters are not well separated. That is, there are data points lying between them, which can be seen as belonging (partially) to different clusters. Fuzzy cluster analysis is a method to handle such data points. It is based on the idea to introduce membership degrees between 0 and 1 which are meant to describe how well a data point belongs to a cluster.

Following a brief introduction to fuzzy cluster analysis which reviews the basic ideas and the most important algorithms, I focus on three aspects of fuzzy clustering, which are very important for successful data analysis:

In the first place, I propose an extension of possibilistic fuzzy clustering. This extension is based on cluster repulsion and considerably improves the clustering results in cases in which the clusters are not well separated.

Secondly, I study how missing values can be handled in fuzzy clustering. Since discarding data with missing values throws away valuable information, I concentrate on approaches based on iterative imputation, available case estimation of the cluster parameters and the introduction of a class specific probability for missing values.

Thirdly, I examine how to handle class information in fuzzy cluster analysis, where a class can consist of several clusters. The main problem is to cleanly separate the classes, which I try to solve by introducing a penalty for clusters comprising several classes and a class repulsion term.

Zusammenfassung

Gruppen/Cluster von homogenen Datenpunkten zu finden, ist eine wichtige Aufgabe der Datenanalyse. Das Ziel der Clusteranalyse ist, einen Datensatz in Gruppen von homogenen Daten zu unterteilen. Doch häufig sind die in den Datensätzen vorliegenden Cluster nicht gut voneinander getrennt. D.h., zwischen ihnen liegen Datenpunkte, die man mehreren Clustern zuordnen kann. Die Fuzzy-Clusteranalyse ist eine Möglichkeit, mit solchen Datenpunkten umzugehen, indem sie den Clustern mit einem Zugehörigkeitsgrad zwischen 0 und 1 zugeordnet werden. Der Zugehörigkeitsgrad beschreibt, wie typisch ein Datum für einen Cluster ist.

Aufbauend auf einer kurzen Einführung in die Fuzzy-Clusteranalyse, die die grundlegenden Ideen und die wichtigsten Verfahren vorstellt, werden drei für eine erfolgreiche Datenanalyse wichtige Gebiete untersucht.

Erstens wird eine Erweiterung der possibilistischen Fuzzy-Clusteranalyse vorgestellt. Die Erweiterung basiert auf der Modellierung einer Abstoßung zwischen Clustern und führt zu einer wesentlichen Verbesserung des Klassifikationsergebnisses, wenn die Cluster nicht gut separiert sind.

Zweitens wird betrachtet, wie man Daten mit fehlenden Werten bei der Fuzzy-Clusteranalyse behandeln kann. Das Entfernen von Daten mit fehlenden Werten vor der Fuzzy-Clusteranalyse führt zu einem größeren Informationsverlust. Daher untersuche ich Ansätze basierend auf einer iterierten Schätzung, der „available case“-Berechnung der Clusterparameter und der Verwendung einer clusterspezifischen Wahrscheinlichkeit für fehlende Werte.

Drittens untersuche ich Möglichkeiten, Klasseninformation bei der Fuzzy-Clusteranalyse zu verwenden, wobei eine Klasse aus mehreren Clustern bestehen kann. Das Problem ist, die Klassen (sauber) zu trennen. Hierfür führe ich einen Strafterm für Cluster, die mehrere Klassen umfassen, und eine klassenabhängige Abstoßung ein.

Kapitel 1

Einleitung

1.1 Einführung in „Knowledge Discovery in Databases (KDD)“

Heutzutage ist es möglich, mit geringem Aufwand sehr große Mengen von Daten zu erfassen, zu sammeln und zu speichern. Dies führt dazu, daß eine wachsende Zahl von Unternehmen bzw. wissenschaftlichen und staatlichen Einrichtungen umfassende Datenbestände aufbaut. Diese Datenbestände werden z.B. verwendet, um

- Betrugsfälle zu erkennen (AT&T),
- Kundengruppen zu erkennen und gezielt ansprechen zu können (Amazon),
- Verkaufsdaten auszuwerten (Wal-Mart),
- Fehlerdaten auszuwerten (DaimlerChrysler),
- Projekte zu prognostizieren (HochTief) oder
- Rating-Systeme zu entwickeln und zu beurteilen (Finanzbranche).

Die Auswertung und Nutzung dieser Datenbestände ist jedoch eine schwierige und anspruchsvolle Aufgabe.

Im Gegensatz zu dem Überfluß an Daten fehlt es oft an Werkzeugen und Verfahren, um aus den Datenbeständen sinnvolle Informationen und neues Wissen zu gewinnen. Obwohl die Anwender oft ein grobes Verständnis von den Daten haben, mit dem sie Vermutungen und Hypothesen aufstellen, wissen sie jedoch meistens nicht, wo und wie sie in den Daten die interessanten

bzw. relevanten Informationen finden können, ob diese Informationen ihre Modelle und Hypothesen stützen und ob vielleicht auch weitere interessante Informationen in den Daten enthalten sind.

Mit diesen Fragestellungen beschäftigt sich das Forschungsgebiet des *„Knowledge Discovery in Databases (KDD)“* (Wissensentdeckung in Datenbanken). Eine gängige Beschreibung ist [47]:

> Knowledge discovery in databases (KDD) is a research area that considers the analysis of large databases in order to identify valid, useful, meaningful, unknown, and unexpected relationships.

Für die Formulierung von Modellen des KDD-Prozesses gibt es verschiedene Vorschläge. Ein interessanter Vorschlag, der von mehreren großen Firmen wie NCR, SPSS, DaimlerChrysler und OHRA unterstützt wird, ist das CRISP-DM-Modell (CRoss Industry Standard Process for Data-Mining) [32]. Die Struktur dieses Modells zeigt Abb. 1.1. Der Kreis deutet an, daß es sich um einen mehrstufigen Prozeß handelt, bei dem die Bewertung der Ergebnisse eine erneute Datenauswahl und -aufbereitung und Modellbildung zur Folge haben kann. Der KDD-Prozeß wird in die Phasen

- des Anwendungsverstehens (business understanding),
- des Datenverstehens (data understanding),
- der Datenaufbereitung (data preparation),
- der Modellierung (modelling),
- der Bewertung (evaluation) und
- der Anwendung (deployment)

gegliedert.

In den Phasen des *Anwendungsverstehens* und des *Datenverstehens* werden die Ziele des Projektes definiert, der potentielle Nutzen abgeschätzt und die benötigten bzw. verfügbaren Daten identifiziert und zusammengeführt. Zusätzlich sammelt man Hintergrundwissen über die Daten. Die Daten werden danach im Rahmen der *Datenaufbereitung* in ein passendes Format überführt, ggf. skaliert und von Fehlern und Ausreißern bereinigt.

In der *Modellierungsphase* wendet man Modellierungs- und Entdeckungstechniken auf die vorverarbeiteten Daten an. Dies wird oft auch als *Data-Mining* bezeichnet. Data-Mining ist ein interdisziplinäres Gebiet, das Verfahren aus der Statistik, dem Soft-Computing, der künstlichen Intelligenz und dem maschinellen Lernen umfaßt [96]. Es handelt sich z.B. um Verfahren zur Segmentierung, zur Klassifikation, zur Beschreibung von Konzepten,

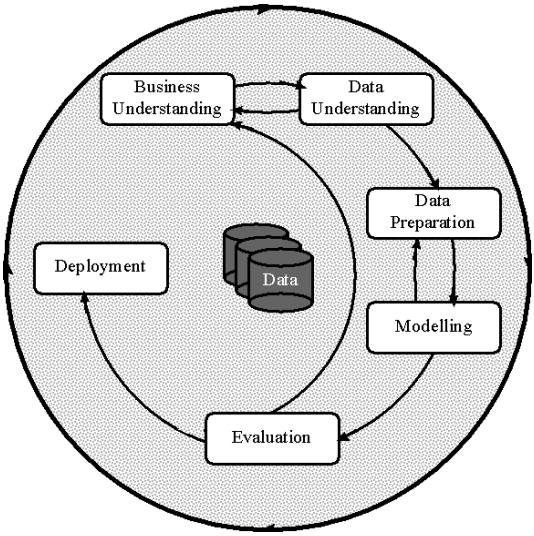

Abbildung 1.1: Das CRISP-DM-Modell

zur Prognose oder zur Abhängigkeitsanalyse. Data-Mining kann als explorative Datenanalyse unter besonderer Berücksichtigung großer Datenbestände angesehen werden.

Bekannte Data-Mining-Verfahren sind z.B. Entscheidungsbäume [30, 102, 103, 24], Schlußfolgerungsnetze [99, 84, 85, 52, 23, 28, 61, 66], Clusteranalyse [22, 112, 111, 95], Neuronale Netze [2, 104, 97] und evolutionäre bzw. genetische Algorithmen [91]. Daneben werden auch klassische statistische Verfahren, wie z.B. die Diskriminanzanalyse, Regressionsanalyse oder Hauptkomponentenanalyse [107], und Verfahren des maschinellen Lernens, wie z.B. induktive logische Programmierung oder fallbasiertes Schließen, dazugezählt [12]. Einen interessanten Überblick über verschiedene kommerzielle Data-Mining-Werkzeuge gibt z.B. [53].

Die Ergebnisse der Data-Mining-Verfahren werden in der *Bewertungsphase* getestet und hinsichtlich ihrer Qualität beurteilt. Gegebenenfalls werden einzelne Phasen des KDD-Prozesses erneut durchlaufen. Abschließend werden die Ergebnisse in der *Anwendungsphase* aufbereitet und verwendet.

Der KDD-Prozeß ist ein interaktiver Prozeß. Der Benutzer prüft und bewertet Ergebnisse und nimmt gegebenenfalls Änderungen und Anpassungen vor.

1.2 Einführung in die Clusteranalyse

Die Clusteranalyse [22, 112, 111, 95, 54] ist eines der Datenanalyseverfahren, die im KDD-Prozeß eingesetzt werden. Sie ist z.B. im Marketing von Interesse, um homogene Kundengruppen zu identifizieren und gezielt ansprechen zu können. Bei größeren Unternehmen ist hierfür der Einsatz von Clusteringverfahren unvermeidbar. Die Kunden sind nicht mehr persönlich bekannt, sondern sie sind für das Unternehmen nur durch die Informationen beschrieben, die es im Rahmen seiner Geschäftsprozesse gesammelt hat. Basierend auf diesen Informationen werden die Kundendaten durch Clusteringverfahren segmentiert. Die Kunden eines Segmentes sind hinsichtlich der betrachteten Informationen homogen und stellen daher eine einheitliche Zielgruppe dar, die entsprechend angesprochen werden kann.

Das Ziel der Clusteranalyse ist, eine Menge von Objekten in homogene Gruppen bzw. Klassen oder Cluster zu unterteilen. Dabei versteht man unter einer Einteilung in Cluster, daß

- die Objekte einer Gruppe untereinander möglichst ähnlich sind. Es wird *Homogenität* innerhalb eines Clusters gefordert.
- die Objekte verschiedener Cluster möglichst unterschiedlich sind. Es wird *Heterogenität* zwischen den Clustern gefordert.

Anschaulich kann man einen Cluster als eine Punktwolke interpretieren. Abb. 1.2 zeigt z.B. einen Datensatz, in dem drei Cluster erkennbar sind. Das Ziel der Clusteranalyse ist, diese Cluster automatisch zu erkennen. Es gibt eine Vielzahl verschiedener Clusteranalyseverfahren, z.B. Mittelwertverfahren, Repräsentantenverfahren, hierarchische Verfahren oder K-Means-Verfahren. Die verschiedenen Clusteranalyseverfahren lassen sich in Abhängigkeit von der Art der Zuordnung der Objekte bzw. der Daten zu den Klassen in disjunkte und nichtdisjunkte Clusteranalyseverfahren unterteilen [22]. Bei den *disjunkten Clusteranalyseverfahren* wird jedes Datum genau einem Cluster zugeordnet. Bei den *nichtdisjunkten Verfahren* können Daten auch mehreren Clustern zugeordnet werden. Die Zuordnung der Daten zu den Clustern kann deterministisch oder probabilistisch sein [4]. Bei den deterministischen Verfahren werden die Daten mit einer Wahrscheinlichkeit

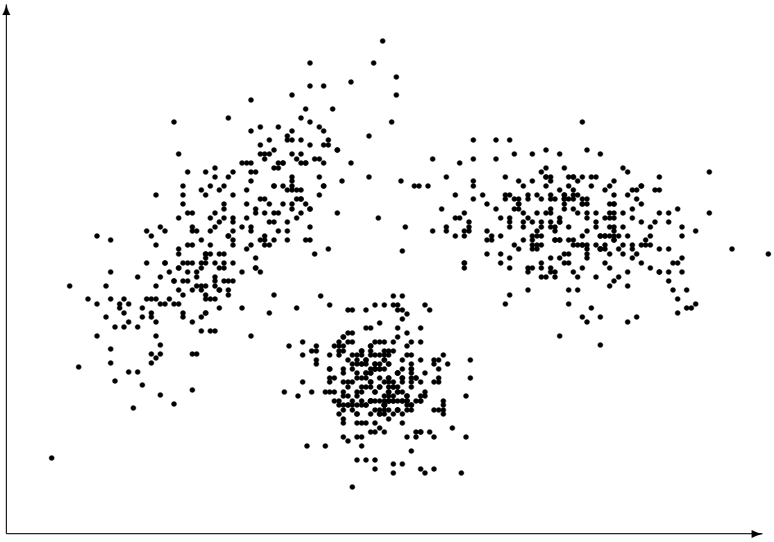

Abbildung 1.2: Ein Datensatz mit drei Clustern.

von 1 einem oder mehreren Clustern zugeordnet. Bei probabilistischen Clusteranalyseverfahren werden die Daten mit einer zwischen 0 und 1 liegenden Wahrscheinlichkeit den Clustern zugeordnet. Ein guter Überblick über die Clusteranalyse wird z.B. in [22, 112, 111, 4, 94] gegeben.

Die Verfahren der *Fuzzy-Clusteranalyse* können im weiteren Sinne zu den probabilistischen Clusteringverfahren gezählt werden, da bei ihnen ebenfalls die Klassifikationsobjekte den Clustern mit einem Zugehörigkeitsgrad zwischen 1 und 0 zugeordnet werden.[1] Dieser Zugehörigkeitsgrad ist jedoch nicht als Wahrscheinlichkeit zu interpretieren. Ein Zugehörigkeitsgrad von 0.7 besagt *nicht*, daß das Datum dem betreffenden Cluster mit einer Wahrscheinlichkeit von 70% zugeordnet wird. Stattdessen sind die Zugehörigkeitsgrade im Sinne der Fuzzy-Logik zu interpretieren.

Abb. 1.3 zeigt die Klassifikation des in Abb. 1.2 dargestellten Datensatzes mit einem Fuzzy-Clusteringverfahren.[2] Nach der Clusteranalyse wurden

[1] Im Gegensatz zu den probabilistischen Clusteringverfahren müssen die Zugehörigkeitsgrade der Daten zu den Clustern sich jedoch nicht auf 1 aufsummieren.

[2] Die Clusteranalyse wurde mit dem in der Arbeitsgruppe von Prof. Dr. R. Kruse, Prof. Dr. F. Klawonn und dem Autor entwickelten Plug-In „Advanced Cluster Analysis“ [25, 121, 124, 119] für das Datenanalysetool DataEngine [98, 136] durchgeführt.

die Daten der Klasse zugeordnet, zu der sie den höchsten Zugehörigkeitsgrad haben.

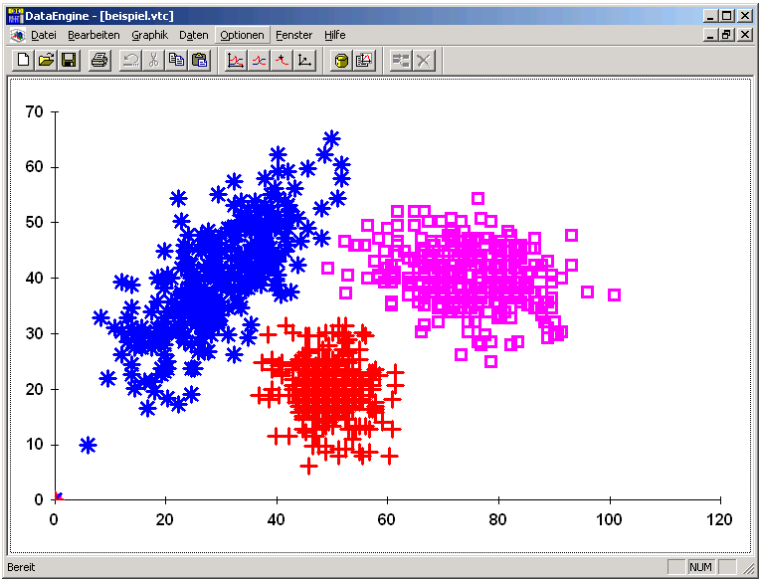

Abbildung 1.3: Klassifikation des Datensatzes aus Abb. 1.2 mit einem Fuzzy-Clusteringverfahren. (Das für die Fuzzy-Clusteranalyse verwendete Verfahren, FMLE, wird in Abschnitt 2.4 vorgestellt.) Der Datensatz ist in drei Cluster (blau, rot und lila) aufgeteilt. Die Daten sind dem Cluster zugeordnet, zu dem sie den größten Zugehörigkeitsgrad besitzen.

1.3 Einführung in die Fuzzy-Logik

Die von Lotfi Zadeh begründete Fuzzy-Mengentheorie bzw. *Fuzzy-Logik* gilt als beherrschender Ansatz zum Umgang mit Vagheit [132, 133, 71, 80, 81, 26, 135, 7, 8, 79]. Bei diesem Ansatz wird versucht, die Problematik des Umgangs mit vagen Begriffen dadurch zu lösen, daß man die Begriffe der Zugehörigkeit zu einer Menge bzw. des Wahrheitswertes fuzzifiziert. Die Idee ist, daß man neben den Begriffen wahr und falsch Zwischenwerte einführt,

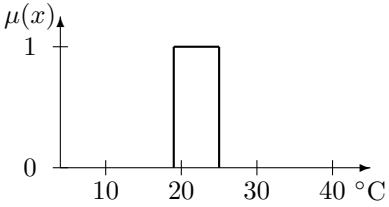

Abbildung 1.4: Eine scharfe Beschreibung des Begriffs „angenehme Temperatur“.

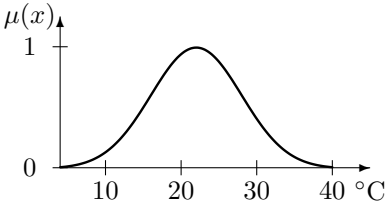

Abbildung 1.5: Beschreibung des Konzepts einer angenehmen Temperatur mit einer Fuzzy-Menge.

die man als Wahrheitsgrad oder Zugehörigkeitsgrad interpretieren kann. Der Zugehörigkeitsgrad sollte um so größer sein, je mehr der Wert unserer Vorstellung von Zugehörigkeit entspricht.

Das Problem läßt sich z.B. gut anhand des linguistischen Ausdrucks der angenehmen Raumtemperatur erläutern. Jeder Mensch hat eine Vorstellung davon, was eine angenehme Temperatur ist. Es ist jedoch problematisch, sie exakt zu definieren, z.B. „Eine Temperatur ist dann angenehm, wenn sie zwischen 19°C und 25°C liegt.“ (vgl. Abb. 1.4). Denn es stellt sich natürlich die Frage, wieso eine Temperatur von 19°C angenehm ist, eine von $18,9$°C jedoch nicht.

Ein gradueller Übergang von einer unangenehmen Temperatur zu einer angenehmen ist zur Modellierung dieses linguistischen Ausdrucks wesentlich sinnvoller. Eine solche Möglichkeit bieten Fuzzy-Mengen. Abb. 1.5 zeigt eine Modellierung des linguistischen Ausdrucks der „angenehmen Temperatur“. Diese Modellierung ermöglicht es z.B. auszudrücken, daß eine Temperatur von 0°C auf keinen Fall angenehm ist ($\mu(0°\text{C}) = 0$), eine von 15°C angenehmer, jedoch nicht so angenehm wie eine von 22°C ist ($\mu(0°\text{C}) < \mu(15°\text{C}) < \mu(22°\text{C})$).

Formal stellt eine Fuzzy-Menge über einer Menge U eine Funktion $\mu : U \rightarrow [0,1]$ dar. Ein Wert $\mu(u_1) = 1, u_1 \in U$, bedeutet, daß das Element u_1 dem durch die Fuzzy-Menge beschriebenen Konzept voll entspricht, und ein Wert $\mu(u_2) = 0, u_2 \in U$, daß das Element u_2 nicht dem durch die Fuzzy-Menge beschriebenen Konzept entspricht.

Wie bei der Wahrscheinlichkeitstheorie stellt sich auch bei der Fuzzy-Logik das Problem der Interpretation. Wie ist ein Zugehörigkeitsgrad von z.B. 0.7 zu deuten? Die Frage der Semantik ist ein fundamentales Pro-

blem der Fuzzy-Logik und wird in den meisten Büchern leider nicht hinreichend beantwortet. Es gibt zwar mit der *Possibilitätstheorie* [134, 42, 43] und ihren verschiedenen Interpretationen Ansätze, diese Frage zu beantworten. So kann z.B. eine Fuzzy-Menge die Unsicherheit über einen scharfen Wert beschreiben, der nicht hinreichend genau beobachtbar bzw. meßbar ist. Eine ausführliche Darstellung dieser Ansätze würde jedoch den Umfang dieses einleitenden Abschnittes sprengen. Daher wird der Begriff des Zugehörigkeitsgrads in dem folgenden Kapitel eingeschränkt auf den Bereich der Fuzzy-Clusteranalyse näher betrachtet.

Neben anderen Gebieten, wie z.B. in der Regelungstechnik, der Qualitätskontrolle oder der Bildverarbeitung, wird die Fuzzy-Logik auch bei der Datenanalyse verwendet [138, 114, 130, 33, 65, 136, 137, 82, 83, 113]. So ist es z.B. bei der Clusteranalyse häufig sinnvoll, graduelle Zugehörigkeitsgrade zu verwenden. Dies ermöglicht es, bei der Klassifikation z.B. zwischen typischen und untypischen Daten für einen Cluster zu unterscheiden und Übergänge zwischen verschiedenen Clustern geeignet zu modellieren. Ein Beispiel hierfür ist die Kundensegmentierung. Kundendaten werden in Cluster unterteilt, die z.B. Zielgruppen beschreiben. Neben typischen Vertretern der einzelnen Kundengruppen gibt es auch Kunden, die verschiedenen Gruppen zuzuordnen sind.

Die Idee der Zugehörigkeitsgrade wird bei der Fuzzy-Clusteranalyse verwendet. Daten können verschiedenen Clustern mit unterschiedlichen Zugehörigkeitsgraden zugeordnet werden. Aus diesen Zugehörigkeitsgraden kann eine Beschreibung der ermittelten Cluster mit Fuzzy-Mengen abgeleitet werden [69].

1.4 Überblick über die Arbeit

In Kapitel 2 dieser Arbeit erfolgt eine Einführung in die Datenanalyse mit Fuzzy-Clusteringverfahren. Die Ideen und grundlegenden Konzepte werden motiviert und erläutert. Die für die Datenanalyse wichtigsten Fuzzy-Clusteringverfahren sind der Fuzzy-C-Means-Algorithmus, der Gustafson–Kessel-Algorithmus und der FMLE-Algorithmus von Gath und Geva. Sie unterteilen einen Datensatz unter Verwendung unterschiedlicher Homogenitätskriterien in wolkenförmige Cluster. Durch die unterschiedlichen Homogenitätskriterien haben diese Verfahren unterschiedliche Eigenschaften und unterscheiden sich hinsichtlich ihrer Leistungsfähigkeit und Flexibilität. Die Verfahren werden näher betrachtet und ihre Möglichkeiten aufgezeigt. Danach werden zwei für die Datenanalyse wichtige Bereiche näher betrachtet:

der Umgang mit verrauschten Daten und die Bewertung einer Klassifikation. Abschließend erfolgt ein kurzer Überblick über weitere Verfahren zur Fuzzy-Clusteranalyse.

Aufgrund des Interesses an Fuzzy-Clusteringverfahren in der Datenanalyse wurden der Fuzzy-C-Means-Algorithmus, der Gustafson–Kessel-Algorithmus und der FMLE als Plug-In „Advanced Cluster Analysis“ [25, 121, 124, 119] für das kommerzielle Datenanalysetool DataEngine [98, 136] von Prof. Dr. R. Kruse, Prof. Dr. F. Klawonn und H. Timm implementiert. Hierdurch ist ein einfacher Einsatz dieser Verfahren in einer professionellen Datenanalyseumgebung und ihre Kombination mit anderen Datenanalyseverfahren möglich. Durch Reaktionen auf dieses Tool und eigene Erfahrungen im Umgang mit der Fuzzy-Clusteranalyse wurden die in dieser Dissertation betrachteten Themen motiviert.

In dieser Arbeit werden drei Problemstellungen der Fuzzy-Clusteranalyse betrachtet, zu denen Lösungsmöglichkeiten entwickelt und bewertet werden:

- *Erweiterung der possibilistischen Fuzzy-Clusteranalyse,*
- *Fuzzy-Clusteranalyse von Daten mit fehlenden Werten und*
- *Fuzzy-Clusteranalyse klassifizierter Daten.*

Hierfür wurden die Verfahren in einem Kommandozeilenprogramm in C implementiert. Die Verfahren können allein oder eingebunden in ein am Lehrstuhl entwickeltes Datenanalyseprogramm „OttoMiner“ verwendet werden. Die Einbindung in „OttoMiner“ ermöglicht die einfache Kombination mit weiteren Datenanalyseverfahren.

Bei der Fuzzy-Clusteranalyse werden meistens probabilistische Zugehörigkeitsgrade verwendet. Hierbei hat jedes Datum das gleiche Gewicht. Diese Verfahren sind robust, ihr Nachteil ist jedoch, daß die Zugehörigkeitsgrade nicht angeben, wie typisch ein Datum für einen Cluster ist. Bei einer größeren Überschneidung von zwei Clustern – es gibt viele Daten, die beiden Clustern zuzuordnen sind – wird die Form der Cluster nicht richtig erkannt, da Daten, die zu beiden Clustern gehören, jeweils nur einen Zugehörigkeitsgrad von 0.5 besitzen.

Eine Alternative ist die Verwendung possibilistischer Zugehörigkeitsgrade. Der Nachteil dieser Verfahren ist jedoch, daß dicht benachbarte Cluster häufig als ein Cluster erkannt werden. Diese Problematik wird in Kapitel 3 betrachtet. Es werden neue Ansätze für die zielfunktionsbasierte Fuzzy-Clusteranalyse und das Alternating Cluster Estimation entwickelt, die possibilistische Zugehörigkeitsgrade besitzen und die Problematik vermeiden,

daß identische Cluster gefunden werden. *Diese neuen Ansätze ermöglichen es, auch bei dicht benachbarten bzw. sich stark überschneidenden Clustern possibilistische Zugehörigkeitsgrade zu verwenden und damit die Form der Cluster gut zu erkennen.*

Neben verrauschten Daten sind Daten mit fehlenden Werten ein häufig auftretendes Problem bei der Datenanalyse. Ein Datum hat fehlende Werte, wenn ein bzw. mehrere Attributwerte nicht beobachtet wurden. Während es für den Umgang mit verrauschten Daten hierfür besonders geeignete Fuzzy-Clusteringverfahren gibt, wurde *der Umgang mit Daten mit fehlenden Werten bisher noch nicht umfassend betrachtet.* Für den Umgang mit Daten mit fehlenden Werten gibt es prinzipiell drei verschiedene Möglichkeiten: Daten mit fehlenden Werten können aus dem Datensatz entfernt werden, fehlende Werte können im Rahmen der Datenvorverarbeitung mit statistischen Verfahren geschätzt werden oder sie können (ggf. nach Modifikation der Verfahren) in den Datenanalyseverfahren berücksichtigt werden. Da für Datenanalyseverfahren der Umgang mit Daten mit fehlenden Werten von großer Bedeutung ist, wird in Kapitel 4 die Fuzzy-Clusteranalyse mit Daten mit fehlenden Werten systematisch betrachtet. Es werden verschiedene Verfahren entwickelt und bewertet, die eine Integration von Daten mit fehlenden Werten in die Fuzzy-Clusteranalyse ermöglichen.

Die Fuzzy-Clusteranalyse zählt zu den nichtüberwachten Klassifikationsverfahren. Manchmal ist jedoch für einige Daten bekannt, zu welcher Klasse sie gehören. Der Begriff der Klasse ist von dem des Clusters zu unterscheiden. Ein Cluster ist eine Menge von homogenen Daten, während eine Klasse aus mehreren Clustern bestehen kann. Für die Berücksichtigung der Information, zu welchem *Cluster* ein Datum gehört, gibt es im Rahmen der teilüberwachten Fuzzy-Clusteranalyse Ansätze. *Für den allgemeineren Fall der Berücksichtigung einer Klasseninformation gibt es jedoch noch keine Untersuchungen.* Da es bei der Datenanalyse sinnvoll ist, *alle* zur Verfügung stehenden Informationen zu verwenden, wird in Kapitel 5 die Fuzzy-Clusteranalyse klassifizierter Daten (Daten mit einer Klasseninformation) betrachtet. Aufbauend auf der teilüberwachten Fuzzy-Clusteranalyse werden neue Ansätze für die zielfunktionsbasierte Fuzzy-Clusteranalyse und das Alternating Cluster Estimation entwickelt, die die Berücksichtigung einer Klasseninformation ermöglichen.

Abschließend wird in Kapitel 6 eine kurze Zusammenfassung und ein Überblick über zukünftig geplante Projekte gegeben.

Kapitel 2

Fuzzy-Clusteranalyse

2.1 Motivation

Das Ziel der Clusteranalyse ist, die Daten eines nicht klassifizierten Datensatzes in Klassen bzw. Cluster einzuteilen. Daten, die zu dem gleichen Cluster gehören, sollen möglichst ähnlich und Daten, die verschiedenen Clustern zugeordnet sind, möglichst verschieden sein. Bei vielen Anwendungen ist eine eindeutige Zuordnung der Daten zu den Clustern jedoch nicht sinnvoll. Falls die Cluster sich z.B. überlappen, wird eine eindeutige Zuordnung zu den Clustern der Struktur der Daten nicht gerecht und kann zu einem Informationsverlust führen.

Das Problem läßt sich anhand des in Abb. 2.1 dargestellten Datensatzes verdeutlichen. Ein Mensch sieht in diesem Datensatz zwei Cluster. Das Datum in der Mitte ist aber weder dem linken noch dem rechten Cluster eindeutig zuzuordnen. Falls man dieses Datum einem Cluster eindeutig zuweist, geht die Information, daß die beiden Cluster spiegelsymmetrisch sind, verloren. Ein anderer Nachteil einer eindeutigen Zuordnung ist, daß man den Daten nach einer Klassifikation nicht mehr ansieht, wie typisch sie für den betreffenden Cluster sind, obwohl diese Information manchmal von Interesse ist.

Eine Möglichkeit, fließende Übergänge zwischen Clustern zu modellieren und bei der Clusteranalyse zu berücksichtigen, ist die Verwendung von graduellen Zugehörigkeiten. Jedem Datum $\vec{x}_j$ wird für jeden Cluster $\vec{\beta}_i$ ein Zugehörigkeitsgrad $u_{i,j} \in [0, 1]$ zugeordnet. Ein Zugehörigkeitsgrad von 1 zeigt an, daß das Datum dem betreffenden Cluster sicher zuzuordnen ist.

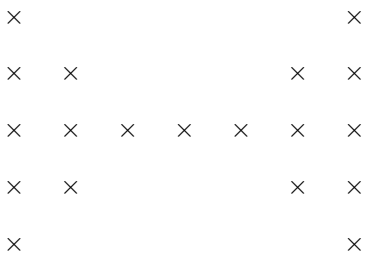

Abbildung 2.1: Ein Datensatz mit zwei Clustern.

Ein Zugehörigkeitsgrad von 0 zeigt dagegen an, daß das Datum dem betreffenden Cluster auf keinen Fall angehört. Da das Datum in der Mitte des in Abb. 2.1 gezeigten Datensatzes beiden Clustern im gleichen Maße zugeordnet werden kann, sollte der Zugehörigkeitsgrad zu beiden Clustern gleich groß sein.

Bei der Fuzzy-Clusteranalyse wird dieses Konzept der Zugehörigkeitsgrade verwendet. Abb. 2.2 zeigt eine Klassifikation dieses Datensatzes mit einem der bekanntesten Fuzzy-Clusteringverfahren, dem Fuzzy-C-Means-Algorithmus (vgl. Abschnitt 2.2). Die Grauwerte geben die Zugehörigkeit an. Die Zentren der Cluster werden durch Quadrate angezeigt. Es ist zu erkennen, daß das Verfahren den Datensatz in zwei spiegelsymmetrische Cluster unterteilt. Das Datum in der Mitte weist zu beiden Clustern den gleichen Zugehörigkeitsgrad auf, der deutlich kleiner als 1 ist. Die so erzeugte Klassifikation entspricht der eines Menschen.

Das Beispiel zeigt, daß Fuzzy-Clusteringverfahren sowohl hinsichtlich der Beschreibung der Cluster als auch bezüglich der Zugehörigkeit der Daten zu den Clustern zu Ergebnissen führen können, die der menschlichen Intuition entsprechen. Typische Vertreter der Cluster und Daten, die als Mischform verschiedener Cluster interpretierbar sind, können nach der Clusteranalyse leicht identifiziert werden. Damit ist die Fuzzy-Clusteranalyse für die Datenanalyse von großem Interesse.

Im folgenden wird der Aufbau und die Vorgehensweise eines Fuzzy-Clusteringverfahrens anhand des weitverbreiteten Fuzzy-C-Means-

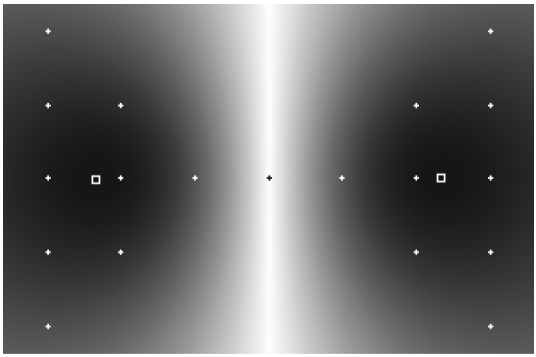

Abbildung 2.2: Klassifikation mit dem Fuzzy-C-Means-Algorithmus. Die Clusterzentren sind durch Quadrate dargestellt.

Algorithmus vorgestellt. Danach werden weitere Fuzzy-Clusteringverfahren, die für die Datenanalyse von Bedeutung sind, einführend betrachtet. Anschließend wird ein kurzer Überblick über weitere Fuzzy-Clusteringverfahren gegeben, die von geringerer Relevanz sind.

Nach der Vorstellung der verschiedenen Fuzzy-Clusteringverfahren wird die Semantik von Zugehörigkeitsgraden näher erläutert. Neben den sehr häufig verwendeten probabilistischen Zugehörigkeitsgraden werden auch possibilistische Zugehörigkeitsgrade betrachtet. Possibilistische Zugehörigkeitsgrade sind eine Möglichkeit, mit verrauschten Daten umzugehen. Da verrauschte Datensätze bei der Datenanalyse ein häufig auftretendes Problem sind, werden danach weitere Verfahren zum Umgang mit verrauschten Datensätzen vorgestellt.

Ein weiterer wichtiger Aspekt bei der Fuzzy-Clusteranalyse ist die Bewertung einer Klassifikation. Die Vorgehensweise und die verwendeten Gütekriterien werden erläutert.

Eine wesentlich weitergehende Vorstellung der Thematik findet sich z.B. in [15, 19, 18, 65, 46, 108, 93].

2.2 Allgemeiner Aufbau eines Fuzzy-Clusteringverfahrens am Beispiel des Fuzzy-C-Means-Algorithmus

Die Fuzzy-Clusteranalyse gehört zu den zielfunktionsbasierten Klassifikationsverfahren. Bei diesen Verfahren wird das Klassifikationsproblem durch eine Zielfunktion beschrieben, die unter Berücksichtigung von Restriktionen zu optimieren ist.

Das Ziel der Clusteranalyse ist, eine Menge von Daten in homogene Gruppen bzw. Klassen oder Cluster zu unterteilen. Wenn man jeden Cluster durch ein typisches Datum beschreibt, kann man die Forderung, daß Daten, die zu einem Cluster gehören, homogen sein sollen, so interpretieren, daß diese Daten dem typischen Datum möglichst ähnlich sein sollten. Wenn man als Homogenitäts- bzw. Ähnlichkeitskriterium den Abstand verwendet, bedeutet dies, daß der Abstand zwischen den Daten eines Clusters und dem typischen Datum möglichst klein sein sollte. Dies ist die Grundidee der bei den Fuzzy-Clusteringverfahren verwendeten Zielfunktion. *Die Daten sind den Clustern so zuzuordnen, daß die Summe der Abstände zwischen den Clustern und den ihnen zugeordneten Daten minimal wird.* Dabei sollten alle Daten das gleiche Gewicht haben.

Der *Fuzzy-C-Means-Algorithmus (FCM)* [15] ist der bekannteste Fuzzy-Clusteringalgorithmus. Das Verfahren ist eng mit dem K-Means-Algorithmus [22] verwandt und kann als seine unscharfe bzw. „fuzzy“ Variante verstanden werden. Das Verfahren versucht, einen Datensatz in c bzw. k Cluster einzuteilen, die durch ihren Mittelwert (Mean) beschrieben werden. Der Unterschied zwischen den beiden Verfahren ist die Zuordnung der Daten zu den Clustern. Während bei dem K-Means Verfahren die Daten den Clustern eindeutig zugeordnet werden, werden bei dem Fuzzy-C-Means-Algorithmus Zugehörigkeitsgrade zwischen 0 und 1 verwendet. Dies ermöglicht es, die Übergänge zwischen den verschiedenen Clustern geeignet zu modellieren (vgl. Abb. 2.2).

Die Zielfunktion des Fuzzy-C-Means-Algorithmus ist:

$$J(\mathbf{X}, \mathbf{U}, \mathbf{B}) = \sum_{j=1}^{n} \sum_{i=1}^{c} u_{i,j}^{m} \cdot d^2(\vec{\beta}_i, \vec{x}_j). \tag{2.1}$$

$\mathbf{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ ist der zu klassifizierende Datensatz,
$\mathbf{B} = \{\vec{\beta}_1, \vec{\beta}_2, \ldots, \vec{\beta}_c\}$ der Vektor der Cluster, die durch ihre jeweiligen Prototypen $\vec{\beta}_i$ beschrieben werden, $\mathbf{U} = \{u_{1,1}, u_{1,2}, \ldots, u_{c,n}\}$ der Vektor der

Zugehörigkeitsgrade $u_{i,j}$ eines Datums $\vec{x}_j$ zu einem Cluster $\vec{\beta}_i$, n die Anzahl der Daten, c die Anzahl der Cluster und $m \in (1, \infty)$ ein Parameter. Als Ähnlichkeitskriterium wird das Quadrat des euklidischen Abstands verwendet.[1]

Die Zielfunktion $J(\mathbf{X}, \mathbf{U}, \mathbf{B})$ ist unter Berücksichtigung der Restriktionen

$$\sum_{i=1}^{c} u_{i,j} = 1 \quad \text{für alle } j \in \{1, \ldots, n\} \tag{2.2}$$

$$\sum_{j=1}^{n} u_{i,j} > 0 \quad \text{für alle } i \in \{1, \ldots, c\} \tag{2.3}$$

zu minimieren.

(2.2) besagt, daß alle Daten das gleiche Gewicht haben.[2] (2.3) bedeutet, daß jedem Cluster Daten zuzuordnen sind.

Die Daten werden den Clustern mit dem Wert $u_{i,j}$ zugeordnet. Der Parameter m wird als Fuzzifier bezeichnet. Durch die Wahl des Fuzzifiers $m \in (1, \infty)$ kann man beeinflussen, ob das Verfahren eher zu einer eindeutigen bzw. harten Zuordnung oder zu einer unscharfen Zuordnung tendiert.[3] Je größer m ist, desto eher wird eine optimale Klassifikation zu Zugehörigkeitsgraden von $\frac{1}{c}$ tendieren. Üblicherweise wird $m = 2$ gewählt.

Restriktion (2.2) verhindert die triviale Lösung des Minimierungsproblems ($u_{i,j} = 0$ für alle Daten $\vec{x}_j$ und alle Cluster $\vec{\beta}_i$).

Da die Zugehörigkeitsgrade durch Restriktion (2.2) stark an eine Wahrscheinlichkeitsverteilung erinnern, wird ein Fuzzy-Clusteringverfahren mit den Restriktionen (2.2) und (2.3) auch als *probabilistisches Fuzzy-Clustering-verfahren* bezeichnet.

Eine direkte Lösung des Optimierungsproblems ist nicht möglich. Daher wird die Zielfunktion durch alternierendes Optimieren minimiert [15]. Die Zielfunktion wird abwechselnd hinsichtlich der Zugehörigkeitsgrade $u_{i,j}$ und der Clusterprototypen $\vec{\beta}_i$ optimiert.

---

[1] Sofern A-priori-Wissen über die den Daten zugrundeliegende Ähnlichkeitsstruktur vorliegt, kann jedoch auch jedes andere Abstandsmaß, wie z.B. der Mahalanobis-Abstand verwendet werden. Im Gegensatz zu anderen Fuzzy-Clusteringverfahren, wie z.B. dem Gustafson–Kessel-Algorithmus (Abschnitt 2.3), wird der Abstand jedoch nicht während des Verfahrens modifiziert.

[2] In der zu minimierenden Zielfunktion und somit auch bei der Berechnung der Clusterprototypen wird $u_{i,j}^m, m \in (1, \infty)$ und nicht $u_{i,j}$ verwendet. Dies bewirkt, daß Daten mit einer „schärferen" Zuordnung ein etwas größeres Gewicht haben.

[3] Manchmal wird auch $m \in [1, \infty)$ erlaubt [15]. Der Fall $m = 1$ erfordert jedoch eine gesonderte Betrachtung bei der Berechnung der Zugehörigkeitsgrade und wird daher meist nicht weiter berücksichtigt.

Die Minimierung der Zielfunktion (2.1) unter Berücksichtigung der Restriktionen (2.2) und (2.3) führt zu folgender Berechnung der Zugehörigkeitsgrade [15, 65]:

$$u_{i,j} = \begin{cases} \dfrac{1}{\sum_{k=1}^{c} \left(\dfrac{d^2(\vec{x}_j, \vec{\beta}_i)}{d^2(\vec{x}_j, \vec{\beta}_k)} \right)^{\frac{1}{m-1}}}, & \text{falls } I_j = \emptyset, \\ 0, & \text{falls } I_j \neq \emptyset \text{ and } i \notin I_j, \\ x, x \in [0,1], \text{ so daß } \sum_{i \in I_j} u_{i,j} = 1 \text{ gilt}, & \text{falls } I_j \neq \emptyset \text{ und } i \in I_j, \end{cases} \tag{2.4}$$

wobei $I_j = \left\{ i | 1 \leq i \leq C, d^2(\vec{x}_j, \vec{\beta}_i) = 0 \right\}$.

(2.4) zeigt, daß die Berechnung der Zugehörigkeitsgrade nur auf den Abständen der Daten zu den Clustern beruht. Die Beschreibung des Clusters, z.B. hinsichtlich Form und Größe, wird nur indirekt mittels der Abstände berücksichtigt. Die Ableitung der Zielfunktion führt daher auch bei anderen probabilistischen Fuzzy-Clusteringverfahren zu dem gleichen Ausdruck, so daß bei *allen* probabilistischen Fuzzy-Clusteringverfahren die Zugehörigkeitsgrade durch (2.4) bestimmt werden.

Bei dem Fuzzy-C-Means-Algorithmus werden die Cluster nur durch ihr Zentrum $\vec{z}_i$, beschrieben. Dieses Zentrum kann als für den Cluster typisches Datum interpretiert werden. Die Ableitung der Zielfunktion (2.1) nach den Clusterprototypen führt zu [15, 65]:

$$\vec{z}_i = \frac{\sum_{j=1}^{n} u_{i,j}^m \vec{x}_j}{\sum_{j=1}^{n} u_{i,j}^m}. \tag{2.5}$$

Die Berechnung des Zentrums durch den Mittelwert der dem Cluster zugeordneten gewichteten Daten entspricht der Intuition.[4]

Ausgehend von einer zufälligen Verteilung der Cluster im Datenraum bzw. einer zufälligen Zuordnung der Daten zu den Clustern werden abwechselnd die Zugehörigkeitsgrade und die Clusterprototypen neu bestimmt. Diese Iteration wird beendet, wenn das Verfahren konvergiert oder wenn die Anzahl der Iterationen eine vorher festgelegte Schranke überschreitet. Algorithmus 2.1 zeigt schematisch den Aufbau des Verfahrens.

[4] Bei dem Fuzzy-C-Median-Algorithmus [67] wird der Mittelwert durch den Median der Daten ersetzt.

Algorithmus 2.1 (Probabilistische Fuzzy-Clusteranalyse)

- *Gegeben sei ein Datensatz* $\mathbf{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$.
- *Wähle die Anzahl der Cluster* c, $2 \leq c < n$. *Jeder Cluster wird durch seinen Prototypen* $\vec{\beta}_i$ *beschrieben. Setze die Anzahl der Iterationen auf 0.*
- *Wähle* $m \in \mathbb{R}_{>1}$
- *Wähle eine Abbruchgenauigkeit und eine maximale Anzahl von Iterationen.*
- *Initialisiere die Clusterprototypen bzw. die Zugehörigkeitsgrade.*
- *REPEAT*
  - *Erhöhe die Anzahl der Iterationen um 1.*
  - *Berechne die Clusterprototypen* $\vec{\beta}_i$, $i \in \{1, \ldots, c\}$.
  - *Berechne die Zugehörigkeitsgrade* $\mathbf{U} = \{u_{1,1}, u_{1,2}, \ldots, u_{c,n}\}$ *nach (2.4).*
- *UNTIL Änderung der Clusterprototypen bzw. Änderung der Zugehörigkeitsgrade kleiner als die Abbruchgenauigkeit oder Überschreitung der maximalen Anzahl der Iterationen.*

Da das Verfahren jedoch nicht immer das Optimum der Zielfunktion findet, ist es sinnvoll, den Algorithmus mit unterschiedlichen Initialisierungen auszuführen und zu bewerten. Die Initialisierung eines Fuzzy-Clusteringverfahrens sowie die Bewertung einer Klassifikation wird in den folgenden Abschnitten näher behandelt.

Der Fuzzy-C-Means-Algorithmus ist ein stabiles und robustes Klassifikationsverfahren, das in verschiedenen kommerziellen Programmen zur Datenanalyse, wie z.B. in einer Toolbox zu Matlab oder DataEngine, enthalten ist. Bei der Anwendung ist jedoch zu berücksichtigen, daß die Daten unter der impliziten Annahme klassifiziert werden, daß alle Cluster ungefähr die gleiche Form und die gleiche Größe besitzen. Üblicherweise wird der euklidische Abstand verwendet, so daß das Verfahren nach kreis- bzw. kugelförmigen Clustern sucht. Es können bei der Verwendung anderer Abstandsmaße jedoch auch andere Clusterformen erkannt werden.

Die in diesem Kapitel vorgestellten Fuzzy-Clusteringverfahren gehen von einer zufälligen Initialisierung aus. Dabei werden zuerst die Zugehörigkeitsgrade der Daten zu den Clustern zufällig bestimmt. Aus diesen Zugehörigkeitsgraden werden dann die Prototypen der Cluster wie üblich bestimmt. Alternativ kann man auch die Cluster zufällig im Datenraum verteilen und anschließend die Zugehörigkeitsgrade der Daten berechnen. Um eine gute Abdeckung des Datenraums zu erreichen, wurde hierfür das Latin-Hypercube-Verfahren verwendet [90]. Häufig werden komplexere Fuzzy-Clusteringverfahren, die eine größere Anzahl von Freiheitsgraden besitzen, mit einfacheren Verfahren initialisiert. Dies verringert die Wahrscheinlichkeit, daß nur eine suboptimale Lösung gefunden wird.

Die Komplexität einer Iteration der probabilistischen Fuzzy-Clusteranalyse ist $O(n \cdot c)$. (Die Komplexität der Berechnung der Clusterprototypen ist $O(n \cdot c)$, da für jeden Cluster jedes Datum mit seinem Zugehörigkeitsgrad betrachtet wird. Auch die Komplexität für die Berechnung der Abstände zwischen den Daten und den Clustern ist $O(n \cdot c)$ Die Komplexität der Berechnung der Zugehörigkeitsgrade ist ebenfalls $O(n \cdot c)$, da der Ausdruck für die Berechnung der Zugehörigkeitsgrade (2.4) zu $u_{i,j} = 1/\left(c \cdot d^{\frac{2}{m-1}}(\vec{x}_j, \vec{\beta}_i) \cdot \text{dist}_j\right)$ umgeformt und $\text{dist}_j = \sum_{k=1}^{c} d^{\frac{2}{m-1}}(\vec{x}_j, \vec{\beta}_k)$ vorher berechnet werden kann.)

Leider existiert keine allgemeine Konvergenzbetrachtung für alle probabilistischen Fuzzy-Clusteringverfahren. Bezdek hat für den Fuzzy-C-Means-Algorithmus jedoch gezeigt, daß entweder die Iterationsfolge selbst oder jede konvergente Teilfolge davon in einem Sattelpunkt oder Minimum, jedoch nicht in einem Maximum der Bewertungsfunktion konvergiert[14, 17].

Historische Anmerkungen

Die Ursprünge des K-Means-Algorithmus können auf Arbeiten von Gauss zurückgeführt werden [19]. 1802 schrieb Gauss über die Methode der kleinsten Fehlerquadrate für die Parameterschätzung [10]. Als erster expliziter Anwender des K-Means Verfahrens wird von Duda und Hart [44] Thorndike [117] genannt. 1969 stellte Ruspini ein Clusteringverfahren mit probabilistischen Zugehörigkeitsgraden vor [106]. Der Fuzzy-C-Means-Algorithmus wurde 1973 mit $m = 2$ von Dunn vorgestellt [45] und von Bezdek auf den Fall $m > 1$ verallgemeinert [13, 15].

Der ISODATA-Algorithmus [6] ähnelt dem K-Means Verfahren. Der Unterschied zwischen den beiden Verfahren sind Heuristiken für die Bestimmung der Anzahl der Cluster bei dem ISODATA-Verfahren. Es können Clu-

ster vereinigt, geteilt und entfernt werden. In frühen Veröffentlichungen wurde der Fuzzy-C-Means-Algorithmus manchmal auch als Fuzzy-ISODATA bezeichnet. Später setzte sich jedoch der Begriff Fuzzy-C-Means-Algorithmus durch, da die Heuristiken des ISODATA-Verfahrens nicht verwendet werden [19].

2.3 Der Gustafson–Kessel-Algorithmus

Der Fuzzy-C-Means-Algorithmus klassifiziert einen Datensatz unter der Annahme, daß alle Cluster ungefähr die gleiche Form und Größe besitzen. Die Form der Cluster ist durch das verwendete Abstandsmaß festgelegt. Meist ist jedoch bei der Datenanalyse die Form der Cluster nicht bekannt. Auch die Forderung, daß alle Cluster die gleiche Form und Größe aufweisen, entspricht nur selten der Realität. Dies kann bei der Verwendung des Fuzzy-C-Means-Algorithmus zu suboptimalen Ergebnissen führen, da das Verfahren versucht, den Datensatz in kugelförmige Cluster einzuteilen, auch wenn diese den Daten nicht gerecht werden.

Der *Gustafson–Kessel-Algorithmus (GK)* bietet eine Möglichkeit, die in dem Datensatz vorliegende Struktur der Daten bei der Fuzzy-Clusteranalyse besser zu berücksichtigen [56]. Um Cluster mit verschiedenen Formen erkennen zu können, muß das Ähnlicheitskriterium clusterspezifisch sein. Für jeden Cluster wird daher ein eigenes Ähnlichkeitsmaß benutzt. Bei der Fuzzy-Clusteranalyse wird als Ähnlichkeitskriterium der Abstand verwendet. Als clusterspezifisches Abstandsmaß bietet sich der Mahalanobis-Abstand an.

Der Mahalanobis-Abstand zwischen einem Datum $\vec{x}_j$ und einem Clusterzentrum $\vec{z}_i$ ist definiert als [22, 87, 88]:

$$d(\vec{x}_j, \vec{z}_i)^2 = (\vec{x}_j - \vec{z}_i)^\top \mathbf{Cov}_i^{-1} (\vec{x}_j - \vec{z}_i). \tag{2.6}$$

$\mathbf{Cov}_i$ ist die Kovarianzmatrix des Clusters $\vec{\beta}_i$ mit dem Zentrum $\vec{z}_i$. Durch die Verwendung der Kovarianzmatrix wird die Form des Clusters bei der Abstandsberechnung berücksichtigt.[5]

Da bei dem Gustafson–Kessel-Algorithmus der Mahalanobis-Abstand verwendet wird, werden die Cluster durch Prototypen $\vec{\beta}_i = \{\vec{z}_i, \mathbf{Cov}_i\}$ beschrieben. Das Zentrum $\vec{z}_i$ beschreibt die Position des Clusters im Raum und die Kovarianzmatrix $\mathbf{Cov}_i$ die Form des Clusters. Analog zu dem Fuzzy-C-Means-Algorithmus wird angenommen, daß alle Cluster ungefähr die gleiche

[5] Den euklidischen Abstand erhält man, wenn man als Kovarianzmatrix die Einheitsmatrix verwendet.

Größe haben. D.h., der Wert der Determinante der Kovarianzmatrix ist für alle Cluster ungefähr gleich. Üblicherweise wird $\det(\mathbf{Cov}_i) = 1$ gewählt. Bei dem Gustafson–Kessel-Algorithmus wird daher der Abstand eines Datums $\vec{x}_j$ zu einem Cluster $\vec{\beta}_i$ durch

$$d^2(\vec{x}_j, \vec{\beta}_i) = \det(\mathbf{Cov}_i)^{1/p}(\vec{x}_j - \vec{z}_i)^\top \mathbf{Cov}_i^{-1}(\vec{x}_j - \vec{z}_i) \tag{2.7}$$

berechnet. p ist die Anzahl der Attribute des Datums $\vec{x}_j$.

Die Zentren und Kovarianzmatrizen der Cluster werden durch:

$$\vec{z}_i = \frac{\sum_{j=1}^{n} u_{i,j}^m \vec{x}_j}{\sum_{j=1}^{n} u_{i,j}^m} \quad \text{und} \tag{2.8}$$

$$\mathbf{Cov}_i = \frac{\sum_{j=1}^{n} u_{i,j}^m (\vec{x}_j - \vec{z}_i)(\vec{x}_j - \vec{z}_i)^\top}{\sum_{j=1}^{n} u_{i,j}^m} \tag{2.9}$$

berechnet.

Häufig wird anstelle der Kovarianzmatrix $\mathbf{Cov}_i$ eine Normmatrix $\mathbf{A}_i$

$$\mathbf{A}_i = \det(\mathbf{Cov}_i)^{1/p}\mathbf{Cov}_i^{-1} \tag{2.10}$$

verwendet. Mit dieser Normmatrix ist der Abstand eines Datums $\vec{x}_j$ zu einem Cluster $\vec{\beta}_i$

$$d^2(\vec{x}_j, \vec{\beta}_i) = (\vec{x}_j - \vec{z}_i)^\top \mathbf{A}_i(\vec{x}_j - \vec{z}_i). \tag{2.11}$$

Üblicherweise wird der Gustafson–Kessel-Algorithmus mit den Ergebnissen des Fuzzy-C-Means-Algorithmus nach einigen Iterationen initialisiert.

Verglichen mit dem Fuzzy-C-Means-Algorithmus ist bei dem Gustafson–Kessel-Algorithmus der Rechenaufwand wesentlich größer, da die Kovarianzmatrix jedes Clusters invertiert werden muß. Wenn man die Clusterform auf achsenparallele Cluster beschränkt, kann die Inverse der Kovarianzmatrix direkt berechnet werden [68, 65]. Die achsenparalle Variante des Gustafson–Kessel-Algorithmus erfordert daher einen geringeren Rechenaufwand als der Gustafson–Kessel-Algorithmus, dafür ist sie jedoch in der Flexibilität hinsichtlich der Clusterform stark eingeschränkt. Die achsenparalle Variante wird z.B. für die Erzeugung von Fuzzy-Regelsystemen mit Fuzzy-Clusteringverfahren verwendet.

2.4 Fuzzy-Maximum-Likelihood-Estimation-Algorithmus

Im Gegensatz zu dem Fuzzy-C-Means-Algorithmus oder dem Gustafson–Kessel-Algorithmus basiert die *Fuzzy-Maximum-Likelihood-Estimation-Algorithmus (FMLE)* auf einem wahrscheinlichkeitstheoretischen Konzept [51]. Der Datensatz wird klassifiziert unter der Annahme, daß die Daten die Repräsentation von c p-dimensionalen Wahrscheinlichkeitsverteilungen sind. Ausgehend von einer Zuordnung der Daten zu den Clustern mit Zugehörigkeitsgraden $u_{i,j}$ werden diese Verteilungen geschätzt. Eine p-dimensionale Wahrscheinlichkeitsverteilung ist gegeben durch den Mittelwert der Verteilung $\vec{z}_i$, die Kovarianzmatrix der Verteilung $\mathbf{Cov}_i$ und die A-priori-Wahrscheinlichkeit p_i, daß die Daten durch die betreffende Verteilung erzeugt wurden. Der Prototyp eines Clusters ist damit $\vec{\beta}_i = (\vec{z}_i, \mathbf{Cov}_i, p_i)$.

Da das Modell auf einem wahrscheinlichkeitstheoretischen Konzept basiert, ist das Ähnlichkeitsmaß wahrscheinlichkeitstheoretisch motiviert. Der Abstand $d^2(\vec{x}_j, \vec{\beta}_i)$ ist umgekehrt proportional zu der A-posteriori-Wahrscheinlichkeit, daß das Datum von der dem betreffenden Cluster zugrundeliegenden Wahrscheinlichkeitsverteilung erzeugt wurde. Diese Wahrscheinlichkeit ist:

$$\frac{p_i}{\sqrt{\det(\mathbf{Cov}_i)(2\pi)^p}} e^{-\frac{1}{2}(\vec{x}_j - \vec{z}_i)^\top \mathbf{Cov}_i^{-1}(\vec{x}_j - \vec{z}_i)}. \tag{2.12}$$

Bei der Fuzzy-Modifikation des Maximum-Likelihood-Estimation-Algorithmus wird daher der Abstand durch

$$d(\vec{x}_i, \vec{\beta}_j)^2 = \frac{1}{p_i} \sqrt{\det(\mathbf{Cov}_i)} e^{\frac{1}{2}(\vec{x}_j - \vec{z}_i)^\top \mathbf{Cov}_i^{-1}(\vec{x}_j - \vec{z}_i)} \tag{2.13}$$

berechnet.

Die Berechnung des Abstands nach (2.13) führt im Vergleich zu der Verwendung des Mahalanobis-Abstands zu einer „schärferen“ bzw. „härteren“ Zuordnung der Daten zu den Clustern. Der FMLE tendiert daher stärker als die vorhergenannten Verfahren dazu, nur ein lokales Optimum zu finden. Folglich ist die Initialisierung von großer Bedeutung. Häufig werden einige Iterationen des FCM und des GK ausgeführt und der FMLE mit den so ermittelten Clustern initialisiert.

Die Prototypen der Cluster werden durch (2.8), (2.9) und

$$p_i = \frac{\sum_{j=1}^{n} u_{i,j}^m}{\sum_{k=1}^{c} \sum_{i=1}^{n} u_{k,j}^m} \tag{2.14}$$

berechnet. Der Unterschied zu der Berechnung Gaußscher Normalverteilungen mit dem EM-Algorithmus ist die Verwendung des Parameters m bei der Berechnung der Clusterprototypen mit $u_{i,j}^m$ [129, 40].

Der FMLE ist in der Lage, Cluster verschiedener Form und Größe zu erkennen. Es ist damit möglich, die Struktur der Cluster bei der Clusteranalyse genauer zu erkennen und zu berücksichtigen. Der FMLE ist jedoch stärker als der Gustafson–Kessel-Algorithmus von der Initialisierung abhängig.

Ebenso wie bei dem Gustafson–Kessel-Algorithmus kann bei dem FMLE der Rechenbedarf reduziert werden, indem man das Verfahren auf die Erkennung achsenparalleler Cluster beschränkt. Bei der achsenparallelen Variante kann die Inverse der Kovarianzmatrix direkt berechnet werden [68, 65]. Der achsenparallele FLME ist durch die Restriktion jedoch nicht so flexibel wie der FMLE bei der Erkennung von Clustern. Die achsenparalle Variante wird z.B. für die Erzeugung von Fuzzy-Regelsystemen mit Fuzzy-Clusteringverfahren verwendet.

2.5 Lineare Mannigfaltigkeiten als Prototypen

Andere Fuzzy-Clusteringverfahren, die einen Datensatz in nicht kugelförmige Cluster aufteilen, verwenden z.B. lineare Mannigfaltigkeiten als Prototypen. Die Idee ist, die Entfernung in einigen Richtungen nicht für die Abstandsberechnung zu verwenden.

Eines dieser Verfahren ist der *Fuzzy-C-Varieties-Algorithmus (FCV)*, der sich zur Erkennung von Linien, Ebenen und Hyperebenen eignet [19, 65]. Jeder Cluster wird bei diesem Verfahren als r-dimensionale lineare Mannigfaltigkeit dargestellt. $r \in \{0, \ldots, p-1\}$, p ist die Dimension des Vektorraums, hier $\mathbb{R}^p$. Der Prototyp eines Clusters $(\vec{z}_i, \{e_{i_1}, e_{i_2}, \ldots, e_{i_r}\})$ besteht also aus einem Punkt $\vec{z}_i$ des Clusters sowie r linear unabhängigen Vektoren $\{\vec{e}_{i_1}, \vec{e}_{i_2}, \ldots, \vec{e}_{i_r}\} \subset \mathbb{R}^p$, für die $||\vec{e}_{i_k}|| = 1, k \in \{1, \ldots, r\}$ gilt.[6] Der Abstand $d^2(\vec{x}_j, \vec{\beta}_i)$ ist definiert als:

$$d^2(\vec{x}_j, \vec{\beta}_i) = \|\vec{x}_j - \vec{z}_i\|^2 - \sum_{k=1}^{r} ((\vec{x}_j - \vec{z}_i)^\top \vec{e}_{i_k})^2. \tag{2.15}$$

Das Distanzmaß (2.15) kann so interpretiert werden, daß der Cluster in den durch die Vektoren $\vec{e}_{i_k}$ spezifizierten Richtungen eine unendliche Ausdeh-

[6] Für $r = 0$ entspricht der Fuzzy-C-Varieties-Algorithmus dem Fuzzy-C-Means-Algorithmus.

nung hat. Der Vektor $\vec{e}_{i_k}$ wird bei dem FCV als der k-te Eigenvektor der Kovarianzmatrix des Clusters $\vec{\beta}_i$ berechnet. Die Eigenvektoren sind nach ihren zugehörigen Eigenwerten in absteigender Reihenfolge sortiert.

Eine Kombination des FCV mit dem Fuzzy-C-Means-Algorithmus führt zu dem *Fuzzy-C-Elliptotypes-Algorithmus* [16, 19, 65]. Die Idee dieses Verfahrens ist, den euklidischen Abstand des Fuzzy-C-Means-Algorithmus, der zu kugelförmigen Clustern führt, und den Abstand des FCV zu kombinieren. Der Abstand ist definiert als

$$\begin{aligned} d^2(\vec{x}_j, \beta_i) &= \alpha \left(\|\vec{x}_j - \vec{z}_i\|^2 - \sum_{k=1}^{r} ((\vec{x}_j - \vec{z}_i)^\top \vec{e}_{i_k})^2 \right) \\ &\quad +(1-\alpha)\|\vec{x}_j - \vec{z}_i\|^2 \\ &= \|\vec{x}_j - \vec{z}_i\|^2 - \alpha \sum_{k=1}^{r} ((\vec{x}_j - \vec{z}_i)^\top \vec{e}_{i_k})^2. \end{aligned} \tag{2.16}$$

$\alpha \in [0,1]$ gewichtet den Abstand des FCV zu dem des Fuzzy-C-Means-Algorithmus und beeinflußt damit die Form der Cluster.

Eine clusterspezifische Wahl von α in (2.16), die es ermöglicht, Cluster unterschiedlicher Form zu erkennen, führt zu dem *Adaptive-Fuzzy-Clustering-Algorithmus* [34, 19, 65, 115, 116]. Ein Vorschlag zur Wahl von α_i ist z.B. [34]:

$$\alpha_i = 1 - \frac{\lambda_{i,min}}{\lambda_{i,max}}. \tag{2.17}$$

Dabei ist $\lambda_{i,max}$ der größte und $\lambda_{i,min}$ der kleinste Eigenwert der Fuzzy-Kovarianzmatrix des Clusters $\vec{\beta}_i$. [7] Eine andere Möglichkeit ist, α_i so zu wählen, daß α_i im Schnitt das Verhältnis der ersten r Eigenwerte zu den übrigen $p-r$ Eigenwerten angibt[115]:

$$\alpha_i = 1 - \frac{r \sum_{j=r+1}^{p} \lambda_{i,j}}{(p-r) \sum_{j=1}^{r} \lambda_{i,j}} \tag{2.18}$$

$\lambda_{i,j}$ ist der j-te Eigenwert des i-ten Clusters. Auch für die clusterspezifische Berechnung der Anzahl der betrachteten Eigenvektoren $\vec{e}_{i_k}$ gibt es Heuristiken [19, 55]

Die Verwendung der in diesem Abschnitt vorgestellten Verfahren bietet sich an, wenn man vermutet, daß die in dem Datensatz vorliegenden Cluster die Form linearer Mannigfaltigkeiten haben.

[7] Die Fuzzy-Kovarianzmatrix ist definiert durch (2.9).

2.6 Fuzzy-Shell-Clusteringverfahren

Wenn man das Homogenitätskriterium etwas weiter interpretiert, läßt sich auch die Hülle einer Kugel, eines Kreises oder einer Ellipse als Cluster auffassen. Das Homogenitätskriterium ist hier die Eigenschaft der Daten, daß sie auf dieser Hülle liegen. Fuzzy-Clusteringverfahren, die einen Datensatz in solche Cluster unterteilen, werden auch als Fuzzy-Shell-Clusteringverfahren bezeichnet [65, 118, 69, 120]. Anwendungsgebiete dieser Verfahren liegen z.B. in der Bildverarbeitung. Die Idee der Fuzzy-Shell-Clusteringverfahren ist, die Form der Cluster mathematisch zu beschreiben und jeweils den Abstand der Daten zu der geometrischen Struktur zu bestimmen.

Da diese Verfahren für die Datenanalyse nur eine relativ geringe Bedeutung besitzen, wird im folgenden exemplarisch nur der Fuzzy-C-Quadric-Shells-Algorithmus vorgestellt. Dieser-Algorithmus ist in der Lage, bei zweidimensionalen Datensätzen Cluster mit der Form von Geraden, Kugeln, Ellipsen, Hyperbeln und Parabeln und bei höherdimensionalen Datensätzen deren entsprechende höherdimensionale Formen zu erkennen.

Der Fuzzy-C-Quadric-Shells-Algorithmus sucht nach Clustern, die sich mittels einer Gleichung zweiten Grades bzw. durch die Oberfläche einer Hyperquadrik darstellen lassen [74]. Die allgemeine Form der Kontur einer Hyperquadrik ist $\vec{r}_i^{\top}\,\vec{q} = 0$ mit

$$\begin{aligned}
\vec{r}_i^{\top} &= (r_{i(1)}, r_{i(2)}, \ldots, r_{i(p)}, r_{i(p+1)}, \ldots, r_{i(r)}, r_{i(r+1)}, \ldots, r_{i(r+p)}, r_{i(s)}),\\
\vec{q}^{\top} &= (x_{(1)}^2, x_{(2)}^2, \ldots, x_{(p)}^2, x_{(1)}x_{(2)}, \ldots, x_{(p-1)}x_{(p)}, x_{(1)}, x_{(2)}, \ldots, x_{(p)}, 1),\\
s &= \tfrac{p(p+1)}{2} + p + 1 = r + p + 1.
\end{aligned}$$

p ist die Dimension des Datenraums, $x_{(k)}$, $1 \le k \le p$, das k-te Attribut des Vektors $\vec{x}$, $r_i \in \mathbb{R}^s$ und $r = \frac{p(p+1)}{2}$. Da der Cluster die Form einer Hyperquadrik hat, ist der Clusterprototyp bei diesem Verfahren $\vec{\beta}_i = \{\vec{r}_i\}$.

Es wird die algebraische Distanz verwendet. Der Abstand $d^2(\vec{x}_j, \vec{\beta}_i)$ ist gegeben durch:

$$d^2(\vec{x}_j, \vec{\beta}_i) = \vec{r}_i^{\top}\,\vec{q}_j\vec{q}_j^{\top}\,\vec{r}_i. \tag{2.19}$$

Basierend auf diesen Abständen können die Zugehörigkeitsgrade der Daten zu den Clustern wie üblich bestimmt werden.

Die Parameter der Clusterprototypen $\vec{\beta}_i = \{\vec{r}_i\}$ kann man aus $\vec{a}_i$ und $\vec{b}_i$ ableiten. $\vec{a}_i$ und $\vec{b}_i$ sind definiert als

$$a_{i(k)} = \begin{cases} r_{i(k)} & 1 \le k \le p \\ \frac{r_{i(k)}}{\sqrt{2}} & p+1 \le k \le r \end{cases} \tag{2.20}$$

$$b_{i(k)} = \quad r_{i(r+k)} \quad 1 \leq k \leq s-r \ . \tag{2.21}$$

$\vec{a}_i$ ist der dem kleinsten Eigenwert zugeordnete Eigenvektor von $(\mathbf{F}_i - \mathbf{G}_i^\top \mathbf{H}_i^{-1} \mathbf{G}_i)$ und $\vec{b}_i$ wird berechnet durch $\vec{b}_i = -\mathbf{H}_i^{-1} \mathbf{G}_i \vec{a}_i$. Dabei sind:
$\mathbf{F}_i = \sum_{j=1}^{n} u_{i,j}^m \mathbf{R}_j$, $\mathbf{G}_i = \sum_{j=1}^{n} u_{i,j}^m \mathbf{S}_j$, $\mathbf{H}_i = \sum_{j=1}^{n} u_{i,j}^m \mathbf{T}_j$,
$\mathbf{R}_j = \vec{s}_j \vec{s}_j^\top$, $\mathbf{S}_j = \vec{s}_j \vec{t}_j^\top$, $\mathbf{T}_j = \vec{t}_j \vec{t}_j^\top$,
$\vec{s}_j^\top = [x_{j(1)}^2, x_{j(2)}^2, \ldots, x_{j(n)}^2, \sqrt{2}x_{j(1)}x_{j(2)}, \ldots, \sqrt{2}x_{j(k)}x_{j(l)}, \ldots,$
$\sqrt{2}x_{j(n-1)}x_{j(n)}]$,
$\vec{t}_j^\top = [x_{j(1)}, x_{j(2)}, \ldots, x_{j(n)}, 1]$.

Neben dem Fuzzy-C-Quadric-Shells-Algorithmus gibt es weitere Fuzzy-Shell-Clusteringverfahren, die Cluster mit der Form von Hypersphären [35, 73], Ellipsen [36, 49], Kegelschnitten und deren höherdimensionale Formen [74] oder (Hyper)-Rechtecken[64] erkennen. Eine ausführliche Darstellung dieser Verfahren findet sich z.B. in [65, 19].

2.7 Possibilistische Clusteranalyse

Bei der Fuzzy-Clusteranalyse werden meistens probabilistische Zugehörigkeitsgrade verwendet. Für jedes Datum ist die Summe der Zugehörigkeitsgrade zu den Clustern gleich 1. Dies bewirkt, daß jedes Datum ungefähr das gleiche Gewicht hat.[8]

Der Nachteil probabilistischer Fuzzy-Clusteringverfahren ist, daß die Zugehörigkeitsgrade schwer interpretierbar sind. Das Problem wird anhand der in den Abbildungen 2.3 und 2.4 gezeigten Beispieldatensätze gezeigt.

Abb. 2.3 zeigt einen Datensatz mit zwei Clustern. Bei der probabilistischen Fuzzy-Clusteranalyse ist der Zugehörigkeitsgrad der beiden Daten $\vec{x}_1$ und $\vec{x}_2$ zu den beiden Clustern $\vec{\beta}_1$ und $\vec{\beta}_2$ jeweils 0.5. Es wird also nicht unterschieden, daß das Datum $\vec{x}_1$ eher beiden Clustern angehört, während das Datum $\vec{x}_2$ eher ein Stördatum ist, das keinem der beiden Cluster zugeordnet werden sollte. Eine Interpretation der Zugehörigkeitsgrade $u_{i,j}$ kann daher leicht zu ungenauen bzw. fehlerbehafteten Aussagen führen, wenn nicht weitere Informationen, z.B. über die Lage der Cluster, mit berücksichtigt werden. Da die Clusterprototypen basierend auf den (teilweise nichtintuitiven) Zugehörigkeitsgraden berechnet werden, kann ein größerer Anteil von

[8] In der zu minimierenden Zielfunktion und somit auch bei der Berechnung der Clusterprototypen wird $u_{i,j}^m, m \in (1, \infty)$, und nicht $u_{i,j}$ verwendet. Dies bewirkt, daß Daten mit einer „schärferen" Zuordnung ein etwas größeres Gewicht haben. Das entspricht der menschlichen Intuition, daß Daten mit einer klaren Zuordnung stärker berücksichtigt werden sollten als Daten mit unscharfer Zuordnung.

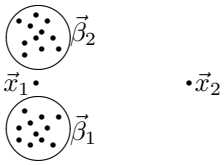

Abbildung 2.3: Beispiel für nichtintuitive Zugehörigkeitsgrade: bei einer probabilistischen Clustereinteilung sind die Zugehörigkeitsgrade der Daten $\vec{x}_1$ und $\vec{x}_2$ gleich.

Stördaten zu einer Berechnung der Clusterprototypen führen, die der Intuition widerspricht.

Auch wenn in den Datensätzen nahezu keine Stördaten enthalten sind, kann die Verwendung probabilistischer Zugehörigkeitsgrade zu schwer interpretierbaren Zugehörigkeitsgraden führen. Abb. 2.4 zeigt einen Datensatz, der aus zwei gut separierten Clustern besteht. Obwohl die Daten $\vec{x}_2$ und $\vec{x}_3$ zu dem Clusterzentrum des linken Clusters, das bei dem Datum $\vec{x}_1$ liegt, den gleichen Abstand haben, haben sie einen unterschiedlichen Zugehörigkeitsgrad zu diesem Cluster. Abb. 2.5 zeigt die Klassifikation des Datensatzes aus Abb. 2.4 mit dem *probabilistischen* Fuzzy-C-Means-Algorithmus. Die Farbsättigung zeigt die Stärke des Zugehörigkeitsgrads zu dem betreffenden Cluster an. Der Zugehörigkeitsgrad steigt mit zunehmender Farbsättigung.

Eine Alternative zu probabilistischen Zugehörigkeitsgraden ist die Verwendung *possibilistischer* Zugehörigkeitsgrade. Im Gegensatz zu probabilistischen Zugehörigkeitsgraden ist bei possibilistischen Zugehörigkeitsgraden der Zugehörigkeitsgrad eines Datums $\vec{x}_j$ zu einem Cluster $\vec{\beta}_i$ *nur* vom Abstand zu dem betreffenden Cluster abhängig. Der Abstand zu den anderen Clustern wird nicht berücksichtigt. Die Zugehörigkeitsgrade spiegeln die Möglichkeit im Sinne der Possibilitätstheorie [42] wieder, daß ein Datum zu dem entsprechenden Cluster gehört.

Possibilitätsverteilungen dienen der Modellierung von Unsicherheit über einen wahren, aber unbekannten Zustand der Wirklichkeit. Der Zugehörigkeitsgrad gibt den Möglichkeitsgrad für diesen Zustand an. Bei der Fuzzy-Clusteranalyse kann der Zugehörigkeitsgrad $u_{i,j}$ als Möglichkeit, daß das Datum $\vec{x}_j$ dem Cluster $\vec{\beta}_i$ zuzuordnen ist, interpretiert werden. $u_{i,j} = 0$ bedeutet, daß es unmöglich ist, das Datum $\vec{x}_j$ dem Cluster $\vec{\beta}_i$ zuzuordnen, während $u_{i,j} = 1$ bedeutet, daß es in keiner Weise eingeschränkt werden kann, das Datum $\vec{x}_j$ dem Cluster $\vec{\beta}_i$ zuzuordnen. Der Möglichkeitsgrad wird bei der possibilistischen Clusteranalyse aus der Relation des Abstands eines Datums $\vec{x}_j$ zu einem Cluster $\vec{\beta}_i$ zu der (vermuteten) Größe des Clusters bestimmt. Die Ähnlichkeit des Datums $\vec{x}_j$ zu dem Cluster $\vec{\beta}_i$ wird in Beziehung zu der Ähnlichkeit „innerhalb“ des Clusters $\vec{\beta}_i$ gesetzt.

Fuzzy-Clusteringverfahren mit possibilistischen Zugehörigkeitsgraden werden auch als *possibilistische Fuzzy-Clusteringverfahren* bezeichnet [75, 38, 9, 76, 78]. Abbildung 2.6 zeigt eine Klassifikation des in Abbildung 2.4 dargestellten Datensatzes mit dem possibilistischen Fuzzy-C-Means-Algorithmus. Der Unterschied zwischen der Verwendung probabilistischer Zugehörigkeitsgrade und der Verwendung possibilistischer Zugehörigkeitsgrade ist deutlich erkennbar. Die possibilistischen Zugehörigkeitsgrade spiegeln im Gegensatz zu den probabilistischen Zugehörigkeitsgraden die Form des Clusters wider. Bei der possibilistischen Fuzzy-Clusteranalyse ist der Zugehörigkeitsgrad ein Maßstab, wie typisch ein Datum für einen Cluster ist, während bei der probabilistischen Fuzzy-Clusteranalyse der Zugehörigkeitsgrad eher die relative Zuordnung eines Datums zu einem Cluster angibt.

Die possibilistische Fuzzy-Clusteranalyse unterscheidet sich von der probabilistischen Fuzzy-Clusteranalyse durch den Verzicht auf Restriktion (2.2) bei der mathematischen Beschreibung des Klassifikationsproblems. Dadurch kann die Summe der Zugehörigkeitsgrade zu den verschiedenen Clustern bei Daten, die zu mehreren Clustern gehören, größer als 1 und bei Stördaten kleiner als 1 sein.

Um die triviale Lösung des Optimierungsproblems, d.h. $u_{i,j} = 0$ für alle $i \in \{1, \ldots, c\}, j \in \{1, \ldots, n\}$, zu vermeiden, wird die Zielfunktion modifiziert, so daß bei einem possibilistischen Clusteringverfahren

$$J(\mathbf{X}, \mathbf{U}, \mathbf{B}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{i,j}^m d^2(\vec{\beta}_i, \vec{x}_j) + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} (1 - u_{i,j})^m \tag{2.22}$$

unter Berücksichtigung der Restriktion $\sum_{j=1}^{n} u_{i,j} > 0, i \in \{1, \ldots, c\}$, minimiert wird. $\eta_i \in \mathbb{R}_{>0}$ gibt den Abstand an, bei dem der Zugehörigkeitsgrad zu dem Cluster β_i $\frac{1}{2}$ betragen soll (vgl. (2.23)).

Die Zugehörigkeitsgrade $u_{i,j}$ werden durch

$$u_{i,j} = \frac{1}{1 + \left(\frac{d^2(\vec{x}_j, \beta_i)}{\eta_i} \right)^{\frac{1}{m-1}}} \tag{2.23}$$

berechnet.[9]

[9] Analog zu der probabilistischen Fuzzy-Clusteranalyse wird der Ausdruck zur Berechnung der possibilistischen Zugehörigkeitsgrade durch Ableitung der Zielfunktion (2.22) nach den Zugehörigkeitsgraden bestimmt.

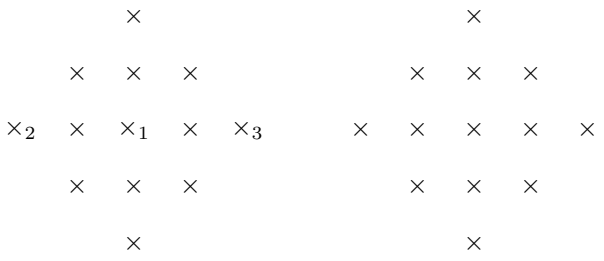

Abbildung 2.4: Datensatz mit 2 Clustern.

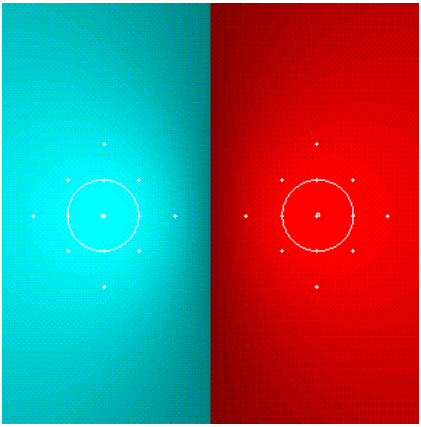

Abbildung 2.5: Clusteranalyse des Datensatzes aus Abb. 2.4 mit dem *probabilistischen* Fuzzy-C-Means-Algorithmus.

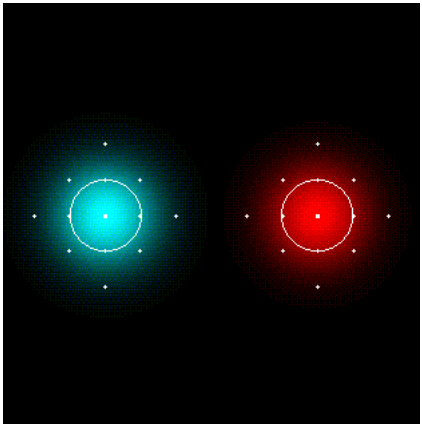

Abbildung 2.6: Clusteranalyse des Datensatzes aus Abb. 2.4 mit dem *possibilistischen* Fuzzy-C-Means-Algorithmus.

Bei dieser Funktion (2.22) bewirkt der erste Summand, daß die Abstände der Daten zu den ihnen zugeordneten Clustern minimiert werden. Der zweite Summand bewirkt, daß die Zugehörigkeitsgrade möglichst groß bestimmt werden. Diese beiden differierenden Ziele werden mittels des Parameters η_i gewichtet.

In [75] wird vorgeschlagen, η_i durch

$$\eta_i = K \frac{\sum_{j=1}^{n} u_{i,j}^m d^2(\vec{\beta}_i, \vec{x}_j)}{\sum_{j=1}^{n} u_{i,j}} \tag{2.24}$$

zu schätzen. K ist ein Parameter, der üblicherweise auf 1 gesetzt wird. Diese Schätzung basiert auf einer vorher durchgeführten probabilistischen Fuzzy-Clusteranalyse.

Die Eigenschaft, daß bei Daten, die für mehrere Cluster typisch sind, die Summe der Zugehörigkeitsgrade zu den verschiedenen Clustern größer als 1 sein kann und bei Daten, die für keinen Cluster typisch sind, die Summe der Zugehörigkeitsgrade zu den verschiedenen Clustern kleiner als 1 sein kann, hat jedoch den Nachteil, daß nicht mehr alle Daten das gleiche Gewicht aufweisen. Das Verfahren gewichtet die Daten also selbständig.

Die in den vorhergehenden Abschnitten vorgestellten Fuzzy-Clusteringverfahren können neben ihrer vorgestellten probabilistischen Variante auch in einer possibilistischen Variante ausgeführt werden. Anstelle probabilistischer Zugehörigkeitsgrade werden possibilistische Zugehörigkeitsgrade verwendet. Die Formeln für die Berechnung der Prototypen der Cluster sind bei beiden Verfahren gleich.

Um ein possibilistisches Fuzzy-Clusteringverfahren zu initialisieren und die Parameter η_i schätzen zu können, wird normalerweise die entsprechende probabilistische Variante zuvor ausgeführt. Häufig wird ein Fuzzifier $m = 1.5$ gewählt.

Da bei der probabilistischen Fuzzy-Clusteranalyse die Zugehörigkeitsgrade nicht ausdrücken, wie typisch ein Datum ist, können probabilistische Zugehörigkeitsgrade zu anderen Clusterprototypen führen als possibilistische Zugehörigkeitsgrade. Bei überlappenden Clustern haben Daten, die beiden Clustern angehören, ein geringeres Gewicht für jeden der Cluster, da die Summe der Zugehörigkeitsgrade zu allen Clustern 1 ist. Die Konsequenz ist, daß bei der probabilistischen Fuzzy-Clusteranalyse die Cluster tendenziell stärker separiert werden als bei der possibilistischen Clusteranalyse, bei der der Zugehörigkeitsgrad aller Daten *nur* davon abhängt, wie typisch sie für den betreffenden Cluster sind. Bei der possibilistischen Fuzzy-Clusteranalyse wird stärker die Form der Daten berücksichtigt, während bei

der probabilistischen Fuzzy-Clusteranalyse die partitionierende Eigenschaft das Ergebnis beeinflußt.

Im Gegensatz zu den probabilistischen Fuzzy-Clusteringverfahren, die einen Datensatz partitionierend aufteilen, können bei den possibilistischen Fuzzy-Clusteringverfahren auch Cluster identisch sein. Der Grund hierfür ist, daß die possibilistischen Verfahren im Gegensatz zu den probabilistischen Verfahren nicht berücksichtigen, ob und inwieweit Daten schon Clustern zugeordnet wurden. Dies wird in Kapitel 3 näher betrachtet.

2.8 Umgang mit Stördaten

Ein allgemeines Problem bei der Datenanalyse ist der Umgang mit Stördaten und Ausreißern. Hierfür gibt es bei der Fuzzy-Clusteranalyse verschiedene Ansätze. Diese werden im folgenden kurz vorgestellt.

Eine Möglichkeit ist die Verwendung possibilistischer Zugehörigkeitsgrade (vgl. Abschnitt 2.7). Die Zugehörigkeitsgrade geben an, wie typisch ein Datum für einen Cluster ist. Sie basieren auf der Relation des Abstands der Daten zu den Clustern zu dem (geschätzten) Abstand, bei dem ein Zugehörigkeitsgrad von $\frac{1}{2}$ vorliegt. Verrauschte Daten und Stördaten erhalten deshalb einen geringeren Zugehörigkeitsgrad als andere Daten und werden daher bei der Berechnung der Clusterprototypen nicht so stark gewichtet.

Ein anderer Ansatz ist die Verwendung eines sogenannten *Noiseclusters*, d.h. eines zusätzlichen Clusters für Stördaten [37]. Da bei der probabilistischen Clusteranalyse die Summe der Zugehörigkeitsgrade eines Datums zu allen Clustern gleich eins ist, reduziert die Zuordnung der Daten zu dem Noisecluster die Zugehörigkeitsgrade zu anderen Clustern. Verrauschte Daten und Stördaten werden dadurch bei der Berechnung der Clusterprototypen nicht mehr so stark gewichtet.

Das Noisecluster ist nicht als Cluster im Sinne einer Punktwolke zu verstehen. Es handelt sich vielmehr um einen fiktiven Cluster, zu dem alle Daten per Definition den gleichen Abstand δ besitzen. Da bei der Fuzzy-Clusteranalyse die Summe der Abstände der Daten zu den Clustern, denen sie zugeordnet sind, minimiert wird, werden Daten dem Noisecluster zugewiesen, wenn der Abstand zu den regulären Clustern größer als δ ist. Die Zielfunktion dieses Ansatzes ist:

$$J(\mathbf{X}, \mathbf{U}, \mathbf{B}) = \sum_{j=1}^{n} \sum_{i=1}^{c} u_{i,j}^{m} \cdot d^2(\vec{\beta}_i, \vec{x}_j) + \sum_{j=1}^{n} \delta^2 \left(1 - \sum_{i=1}^{c} u_{i,j}\right)^m . \quad (2.25)$$

Da die Daten neben den c regulären Clustern auch dem Noisecluster zugeordnet werden können, wird die Restriktion (2.2) abgeschwächt zu

$$\sum_{i=1}^{c} u_{i,j} < 1. \tag{2.26}$$

Die Zugehörigkeitsgrade werden bei diesem Ansatz durch

$$u_{i,j} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d^2(\vec{x}_j, \beta_i)}{d^2(\vec{x}_j, \beta_k)} \right)^{\frac{1}{m-1}} + \left(\frac{d^2(\vec{x}_j, \beta_i)}{\delta} \right)^{\frac{1}{m-1}}} \tag{2.27}$$

berechnet.

Der entscheidende Punkt bei diesem Verfahren ist die Bestimmung von δ. Falls δ zu groß gewählt wird, hat dieser Ansatz nahezu keinen Effekt, da die Stördaten weiterhin den regulären Clustern zugeordnet werden. Falls jedoch δ zu klein gewählt wird, werden zuviele Daten als Stördaten identifiziert. Dies hat den Nachteil, daß unnötig viele Daten bei der Clusteranalyse nicht berücksichtigt werden. Es werden Informationen „verschenkt“.

In [37] wird vorgeschlagen, δ in Abhängigkeit von dem mittleren Abstand der Daten zu den regulären Clustern zu schätzen.

$$\delta^2 = \lambda \frac{\sum_{j=1}^{n} \sum_{i=1}^{c} d^2(\vec{\beta}_i, \vec{x}_j)}{n \cdot c}. \tag{2.28}$$

Da die Wahl von δ auf die Wirksamkeit des Ansatzes einen entscheidenden Einfluß hat, sollte δ nach einigen Iterationen neu berechnet werden.

Weitergehende Betrachtungen finden sich z.B. in [38, 39, 19].

2.9 Bewertung einer Klassifikation — Bestimmung der Clusteranzahl

Die Fuzzy-Clusteranalyse ist ein Verfahren zur unüberwachten Klassifikation. Die Anzahl der Cluster ist vorzugeben. Da diese jedoch oft nicht bekannt ist, ist die Bewertung der Ergebnisse einer Fuzzy-Clusteranalyse von sehr großer Bedeutung. Für die Bewertung werden Gütekriterien verwendet. Man unterscheidet zwischen *globalen Gütekriterien*, die eine Klassifikation als Ganzes bewerten, und *lokalen Gütekriterien*, bei denen jeder Cluster separat bewertet wird.

Um eine Clusteranalyse bei einer unbekannten Clusteranzahl durchzuführen, gibt es unterschiedliche Ansätze:

- Der gebräuchlichste Ansatz ist, das Fuzzy-Clusteringverfahren mit einer unterschiedlichen Anzahl von Clustern auszuführen. Das Ergebnis wird nach jedem Durchlauf bewertet. Das beste Ergebnis wird anschließend ausgegeben.

  Der Begriff des „besten“ Ergebnisses ist jedoch problematisch. Die meisten Gütekriterien tendieren dazu, bei einer steigenden Clusterzahl zu einer besseren Bewertung zu kommen. Sofern man die Clusterzahl und die Anzahl der Daten je Cluster als beliebig ansieht, ist der Zustand als ideal anzusehen, bei dem jedem Cluster genau ein Datum zugeordnet wird, das diesen Cluster beschreibt. Dies entspricht jedoch nicht dem intuitiven Verständnis einer guten Klassifikation. Anstelle des Optimums der Gütefunktion wird daher nach einer signifikanten Änderung der Ableitung der Funktion der Gütewerte — aufgetragen nach der Anzahl der Cluster — gesucht.

- Eine andere Vorgehensweise ist, mit einer größeren Anzahl von Clustern die Clusteranalyse durchzuführen und anschließend ähnliche Cluster zu vereinigen. Ein bekanntes Verfahren, das auf dieser Idee beruht, ist z.B. der „Compatible-Cluster-Merging-Algorithmus“ (CCM).

- Eine weitere Idee ist, nach der Clusteranalyse nach schlechten Clustern zu suchen und diese zu entfernen. Auch bei diesem Ansatz wird die Clusteranalyse mit einer größeren Anzahl von Clustern durchgeführt.

Im folgenden werden verschiedene Gütekriterien exemplarisch vorgestellt.

2.9.1 Globale Gütemaße

Globale Gütemaße bewerten eine Clustereinteilung als Ganzes. Das einfachste globale Gütekriterium ist die zu minimierende *Zielfunktion* $J(\mathbf{X}, \mathbf{B}, \mathbf{U})$. Da die zu minimierende Zielfunktion die Klassifikationsaufgabe beschreibt, kann man aus einer niedrigeren Bewertung auf ein besseres Klassifikationsergebnis schließen.

Die Clusteranzahl kann mit globalen Gütekriterien bestimmt werden, indem das Clusteringverfahren mit einer zunehmenden Anzahl von Clustern durchgeführt und bewertet wird. Da viele Gütekriterien hinsichtlich der Clusteranzahl monoton fallend bzw. steigend sind, wird häufig anstelle der Gütefunktion die Ableitung der Gütefunktion verwendet, um eine „optimale“ Klassifikation zu erkennen.

Im folgenden werden mehrere globale Gütekriterien vorgestellt. Einige dieser Kriterien bewerten „nur“ die Zuordnung der Daten zu den Clustern. Andere Kriterien bewerten die Form der Cluster oder wie gut die Cluster separiert sind. Da die Gütekriterien unterschiedliche Aspekte einer guten Klassifikation betrachten, ist es problematisch, von *dem* besten Gütekriterium zu sprechen. Stattdessen sollten bei der Bewertung einer Clustereinteilung mehrere Gütekriterien betrachtet werden.

Partitionskoeffizient

Der *Partitionskoeffizient (partition coefficient)* ist ein sehr einfaches Gütekriterium, das auf der Idee basiert, daß bei einer guten Klassifikation die Daten eindeutig den Clustern zugeordnet werden können [15]. Die Zugehörigkeitsgrade sollten möglichst nahe bei 1 bzw. nahe bei 0 sein. Eine unscharfe Zuordnung ist ein Indiz für eine schlechte Klassifikation.

Der Partitionskoeffizient ist definiert als

$$\mathrm{PC}(\mathbf{U}) \;=\; \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{i,j}^2. \tag{2.29}$$

Da bei diesem Gütekriterium eine möglichst eindeutige Zuordnung der Daten zu den Clustern angestrebt wird, deutet ein höherer Wert auf eine bessere Klassifikation hin.

Partitionsentropie

Ebenso wie bei dem Partitionskoeffizienten wird auch bei dem Gütekriterium der Partitionsentropie das Ergebnis des Fuzzy-Clusteringverfahrens nur unter Verwendung der Zugehörigkeitsgrade der Daten zu den Clustern beurteilt. Die *Partitionsentropie (partition entropy)* basiert auf der aus der Informationstheorie bekannten Shannon-Entropie [15, 110]. Sie ist definiert als

$$\mathrm{PE}(\mathbf{U}) \;=\; -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{i,j}\ln(u_{i,j}). \tag{2.30}$$

Die Zugehörigkeitsgrade der Daten zu den Clustern werden dabei als Informationsgehalt gedeutet. Wie bei dem Partitionskoeffizient wird auch bei der Partitionsentropie eine möglichst eindeutige Zuordnung der Daten zu den Clustern angestrebt. Hierfür ist die Partitionsentropie zu minimieren.

Normierung des Partitionskoeffizienten

Nachteilig ist bei dem Partitionskoeffizienten und der Partitionsentropie, daß bei beiden Gütekriterien der Wert nicht hinsichtlich der Anzahl der Cluster normiert ist. Ein weiteres Gütemaß, das ebenfalls auf der Eindeutigkeit der Zuordnung der Daten zu den Clustern basiert, jedoch die Anzahl der Cluster berücksichtigt, ist [5, 115]

$$\mathrm{G}(\mathbf{U}) \;=\; \frac{1}{c-1} \sum_{i=1}^{c-1} \sum_{k=i+1}^{c} \frac{1}{n} \sum_{j=1}^{n} u_{i,j} u_{k,j}. \tag{2.31}$$

Das Gütekriterium ist auf den Bereich $[0, 1]$ normiert. Es hat ähnliche Eigenschaften wie der Partitionskoeffizient, da man durch Umformung

$$\mathrm{G}(\mathbf{U}) \;=\; 1 - \frac{c}{c-1}(1 - \mathrm{PC}(\mathbf{U})) \tag{2.32}$$

erhält.

Verhältnis-Repräsentant

Der *Verhältnis-Repräsentant (proportion exponent)* kann als Logarithmus des Maßes für die Anzahl der Einteilungen interpretiert werden, bei denen alle Daten besser als bei der vorliegenden Einteilung klassifiziert werden. Er ist definiert als [65]

$$\mathrm{PX}(\mathbf{U}) \;=\; -\ln \left(\prod_{\vec{x}_j \in \mathbf{X}} \left(\sum_{k=1}^{\lceil \mu_{\vec{x}_j}^{-1} \rceil} (-1)^{k+1} \binom{c}{k} (1 - k\mu_{\vec{x}_j})^{c-1} \right) \right). \tag{2.33}$$

Dabei ist $\mu_{\vec{x}_j} = \max_{1 \le i \le c} u_{i,j}$ und $\mathbf{X}$ die Menge der Daten.

Es ist zu berücksichtigen, daß bei diesem Gütekriterium kein Datum einen Zugehörigkeitsgrad von 1 zu einem Cluster haben darf. Für einen gegen 1 strebenden Zugehörigkeitsgrad strebt der Wert des Gütekriteriums unabhängig von den anderen Zugehörigkeitsgraden gegen ∞. Diese Eigenschaft entspricht der Semantik des Kriteriums, daß nur bessere Einteilungen *aller* Daten berücksichtigt werden.

Trennungsgrad

Der von Xie und Beni eingeführte *Trennungsgrad (separation)* [131] bewertet das Ergebnis einer Clusteranalyse unter dem Gesichtspunkt der Separation

der Cluster. Bei diesem Kriterium wird der Abstand der Daten zu den Clustern, denen sie zugeordnet wurden, in Relation zu dem Abstand der Cluster gesetzt. Er ist definiert als [131]

$$\mathrm{S}(\mathbf{U}) \;=\; \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{i,j}^2 d^2(\vec{x}_j, \vec{\beta}_i)}{n \min\left\{ d^2(\vec{\beta}_j, \vec{\beta}_i) | i, j \in \{1, \ldots c\}, i \neq j \right\}}. \tag{2.34}$$

Dieses Gütekriterium greift die beiden Aspekte des gewünschten Ergebnisses einer Clusteranalyse auf. Der Ausdruck im Zähler kann als Bewertung der Homogenität der Daten innerhalb eines Clusters angesehen werden, da als Homogenitätskriterium der Abstand der Daten zu dem Cluster verwendet wird. Der Zähler sollte daher möglichst klein sein. Im Nenner des Gütekriteriums wird der Aspekt der Heterogenität der Daten aus verschiedenen Clustern aufgegriffen. Da die Daten durch die Cluster, denen sie zugeordnet wurden, repräsentiert werden, wird hier der Abstand der Cluster verwendet. Durch die Verwendung des Minimums wird die Heterogenität tendenziell unterschätzt. Der Ausdruck im Nenner sollte möglichst groß sein. Bei diesem Gütekriterium weist ein kleiner Wert auf eine gute Klassifikation hin.

Fuzzy-Hypervolumen

Das *Fuzzy-Hypervolumen (fuzzy hypervolume)* ist ein Gütekriterium, das die Kompaktheit der Cluster bewertet. Die Idee ist, daß bei einem guten Klassifikationsergebnis die Cluster eine möglichst kleine Ausdehnung haben. Als Maß für die Kompaktheit eines Clusters $\vec{\beta}_i$ kann die Determinante der Kovarianzmatrix des Clusters $\mathbf{Cov}_i$ verwendet werden.

Das Fuzzy-Hypervolumen ist definiert als [51]

$$\mathrm{FHV}(\mathbf{U}) \;=\; \sum_{i=1}^{c} \sqrt{\det(\mathbf{Cov}_i)}. \tag{2.35}$$

Das Fuzzy-Hypervolumen ist als globales Gütekriterium definiert. Es kann jedoch genauso als lokales Gütekriterium zur Bewertung eines Clusters $\vec{\beta}_i$ verwendet werden, indem für diesen Cluster der Wert $\sqrt{\det(\mathbf{Cov}_i)}$ bestimmt wird.

Partitionsdichte

Bei dem Fuzzy-Hypervolumen wird nur das Volumen der Cluster betrachtet. Unabhängig von der Anzahl der ihnen zugeordneten Daten werden große

Cluster schlecht bewertet. Ein Cluster mit einem großen Volumen kann jedoch durchaus gut sein, wenn ihm ein große Zahl von Daten zugeordnet ist. Das Gütekriterium der *Partitionsdichte (partition density)* betrachtet daher die Anzahl der den Clustern zugeordneten Daten in Relation zu dem Volumen der Cluster. Es werden nur die Daten betrachtet, die den Clustern „gut“ zugeordnet sind.

Die Partitionsdichte ist definiert als [51]

$$\mathrm{PD}(\mathbf{U}) \;=\; \frac{\sum_{i=1}^{c} S_i}{\sum_{i=1}^{c} \sqrt{\det(\mathbf{Cov}_i)}}. \tag{2.36}$$

S_i ist ein Maß für die dem Cluster $\vec{\beta}_i$ gut zugeordneten Daten.
$S_i = \sum_{j \in Y_i} u_{i,j}$ mit $Y_i = \{j \in \{1, \ldots, n\} \mid (\vec{x}_j - \vec{z}_i)^\top \mathbf{Cov}_i^{-1} (\vec{x}_j - \vec{z}_i) < 1\}$.

Mittlere Partitionsdichte

Das Gütekriterium der *mittleren Partitionsdichte (average partition density)* bewertet wie die Partitionsdichte die Anzahl der den Clustern gut zugeordneten Daten in Relation zu dem Volumen der Cluster [51]. Im Gegensatz zu der Partitionsdichte wird jedoch bei der mittleren Relationsdichte die Relation für jeden Cluster separat bestimmt.

Die mittlere Partitionsdichte ist definiert als [51]

$$\mathrm{APD}(\mathbf{U}) \;=\; \frac{1}{c} \sum_{i=1}^{c} \frac{S_i}{\sqrt{\det(\mathbf{Cov}_i)}}. \tag{2.37}$$

S_i ist wie bei der Partitionsdichte definiert.

Wie das Fuzzy-Hypervolumen kann auch die mittlere Partitionsdichte als lokales Gütekriteriumverwendet werden, indem $S_i / \sqrt{\det(\mathbf{Cov}_i)}$ für einen Cluster $\vec{\beta}_i$ berechnet wird.

2.9.2 Lokale Gütekriterien

Die Problematik globaler Gütekriterien ist, daß die Clustereinteilung nur als Ganzes betrachtet wird. Oft ist man jedoch auch an einer Bewertung einzelner Cluster interessiert. So kann z.B. die Clusteranzahl bestimmt werden, indem nach der Clusteranalyse schlechte Cluster entfernt werden. Die Clusteranalyse wird dann mit der reduzierten Klassenanzahl erneut durchgeführt.

Einige der im vorhergehenden Abschnitt vorgestellten globalen Gütekriterien können auch als lokale Gütekriterien verwendet werden. So können

z.B. die mit dem Gustafson–Kessel-Algorithmus oder dem FMLE berechneten Cluster mit dem Fuzzy-Hypervolumen oder der Partitionsdichte auch separat bewertet werden, indem bei den Kriterien jeweils nur *ein* Cluster betrachtet wird. Ggf. kann zusätzlich überprüft werden, ob die Verteilung der Daten in dem Cluster der erwarteten bzw. vermuteten entspricht. So wird z.B. bei dem FMLE eine Normalverteilung unterstellt. Die Abweichung der vorliegenden Verteilung von der erwarteten Verteilung gibt die „Güte“ des Clusters an.

Für Fuzzy-Clusteringverfahren mit komplexeren Clusterprototypen, wie z.B. Fuzzy-Shell-Clusteringverfahren, gibt es weitere Kriterien. So wird z.B. bei der *Konturdichte* die Güte des jeweiligen Clusters anhand der Anzahl der Punkte in Relation zu der Länge der Kontur bzw. der Oberfläche der Kontur des Clusters berechnet [77].

2.9.3 Competitive-Agglomeration

Die Bestimmung der Clusteranzahl mit globalen Gütekriterien hat den Nachteil, daß das Clusteringverfahren mehrmals mit einer unterschiedlichen Clusteranzahl ausgeführt wird. Diese Vorgehensweise ist relativ zeitaufwendig.

Einen alternativen Ansatz bietet das *Competitive-Agglomeration-Clustering* [50, 115]. Die Idee des Verfahrens ist, mit einer zu hohen Clusteranzahl die Clusteranalyse zu starten und nach jedem Durchlauf die Clusteranzahl zu reduzieren. Da bei dem jeweils nachfolgenden Durchlauf das Verfahren mit den Ergebnissen des vorhergehenden Durchlaufs initialisiert wird, ist mit einer relativ schnellen Konvergenz der Clusteranalyse für die jeweiligen Clusteranzahlen zu rechnen. Der Nachteil dieser Vorgehensweise ist, daß die Gefahr besteht, die Anzahl der Cluster bei dem Beginn des Verfahrens stark zu überschätzen. Daher kann der Rechenaufwand bei dem ersten Durchlauf verhältnismäßig groß sein.

Bei dem Competitive-Agglomeration-Clustering wird eine modifizierte Zielfunktion verwendet, um die Bildung großer Cluster zu unterstützen. Die Zielfunktion ist:

$$J(\mathbf{X},\mathbf{U},\mathbf{B}) = \sum_{j=1}^{n}\sum_{i=1}^{c} u_{i,j}^2 \cdot d^2(\vec{x}_j,\vec{\beta}_i) - \alpha \sum_{i=1}^{c}\left(\sum_{j=1}^{n} u_{i,j}\right)^2 . \tag{2.38}$$

Die Subtraktion des Ausdrucks $\alpha \sum_{i=1}^{c}\left(\sum_{j=1}^{n} u_{i,j}\right)^2$ bewirkt die Bevorzugung großer Cluster, d.h. von Clustern, für die die Summe $\sum_{j=1}^{n} u_{i,j}$

möglichst groß ist. Der Parameter α bestimmt, wie stark die Clustergröße gegenüber dem auf den Abständen basierenden Optimierungskriterium gewichtet wird.

Die Zugehörigkeitsgrade werden bei diesem Verfahren durch

$$u_{i,j} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d^2(\vec{x}_j, \vec{\beta}_i)}{d^2(\vec{x}_j, \vec{\beta}_k)} \right)} + \frac{\alpha}{d^2(\vec{x}_j, \vec{\beta}_i)} \cdot \left(\sum_{j=1}^{n} u_{i,j} - \frac{\sum_{i=1}^{c} \frac{1}{d^2(\vec{x}_j, \vec{\beta}_i)} \sum_{j=1}^{n} u_{i,j}}{\sum_{i=1}^{c} \frac{1}{d^2(\vec{x}_j, \vec{\beta}_i)}} \right) \tag{2.39}$$

berechnet.

Die Zielfunktion und die Formel für die Berechnung der Zugehörigkeitsgrade zeigen die enge Verwandtschaft des Competitive-Agglomeration-Clustering mit der probabilistischen Clusteranalyse. Der erste Summand der Zielfunktion und der Formel für die Berechnung der Zugehörigkeitsgrade entspricht dem jeweiligen Ausdruck bei der probabilistischen Fuzzy-Clusteranalyse für $m = 2$. Die Möglichkeit, über den Fuzzifier m die Unschärfe der Zugehörigkeitsgrade zu beeinflussen, wurde also beim Competitive-Agglomeration-Clustering gegen die Fähigkeit zur Bevorzugung größerer Cluster eingetauscht. Die verschiedenen Fuzzy-Clusteringverfahren, wie z.B. der Fuzzy-C-Means-Algorithmus oder der Gustafson–Kessel-Algorithmus, können auch als Competitive-Agglomeration-Clustering-Varianten verwendet werden. Hierfür sind die Prototypen der Cluster wie bei der probabilistischen Version zu berechnen.

Beim Competitive-Agglomeration-Clustering wird nach jedem Durchlauf die Größe jedes Clusters mittels

$$\text{card}_i = \sum_{j=1}^{n} u_{i,j} \tag{2.40}$$

bestimmt. Cluster mit einer zu geringen Kardinalität, d.h. $\text{card}_i < \epsilon$, werden entfernt. Anschließend wird die Clusteranalyse mit der reduzierten Anzahl von Clustern erneut durchgeführt.

Für die Leistungsfähigkeit des Verfahrens ist die Wahl des Parameters α von großer Bedeutung. In [50] wird vorgeschlagen, ihn mittels

$$\alpha = \eta(t) \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} u_{i,j}^2 d^2(\vec{x}_j, \vec{\beta}_k)}{\sum_{i=1}^{c} \left(\sum_{j=1}^{n} u_{i,j} \right)^2} \tag{2.41}$$

zu bestimmen. $\eta(t)$ wird dabei durch

$$\eta(t) = \eta e^{\left(\frac{-t}{\tau}\right)} \tag{2.42}$$

definiert. t ist die Nummer des Iterierungsschritts des Competitive-Agglomeration-Clustering.

Die Idee ist, α als das Verhältnis zwischen dem ersten und dem zweiten Term der Zielfunktion (2.38) zu definieren und im Laufe des Clusteringverfahrens exponentiell zu verkleinern. η und τ sind Parameter, die festlegen, wie schnell α verkleinert wird. Durch den exponentiellen Abfall von α entspricht das Competitive-Agglomeration-Clustering nach einigen Iterationen einem probabilistischen Fuzzy-Clusteringverfahren mit dem Fuzzifier $m = 2$.

Bei der Anwendung dieses Verfahrens ist zu berücksichtigen, daß während der ersten Iterationen der Zugehörigkeitsgrad der Daten zu den Clustern nicht probabilistisch zu interpretieren ist.

2.9.4 Compatible-Cluster-Merging

Die Idee des *Compatible-Cluster-Merging (CCM)* ist, ausgehend von einer zu großen Anzahl von Clustern, die in dem Datensatz vorliegende Clusteranzahl zu bestimmen, indem nach jedem Durchlauf ähnliche Cluster vereinigt werden [72]. Wie bei dem im vorhergehenden Abschnitt vorgestellten Competitive-Agglomeration-Clustering werden auch beim Compatible-Cluster-Merging die nachfolgenden Iterationen des Clusteringverfahrens mit den Ergebnissen der vorhergehenden Iteration initialisiert. Wie bei dem Competitive-Agglomeration-Clustering ist daher bei der ersten Ausführung des Fuzzy-Clusteringverfahrens mit einem relativ großen Rechenaufwand zu rechnen, während die nachfolgenden Durchläufe aufgrund der Initialisierung relativ schnell konvergieren.

Das Compatible-Cluster-Merging basiert auf dem Gustafson–Kessel-Algorithmus. Nach jeder Iteration wird die Ähnlichkeit der Cluster hinsichtlich der Lage und der Orientierung der Hyperellipsen bestimmt. Zwei Cluster werden als vereinbar angesehen, wenn die Kompatibilitätsrelation $\doteq_\gamma$ erfüllt ist:

Definition 2.1 (Kompatibilitätsrelation für den CCM)

$$\begin{aligned} \forall x, y \in \mathit{Cluster} : x \doteq_\gamma y \quad \Leftrightarrow \quad & |\vec{e}_x \vec{e}_y^{\top}| \geq \gamma_1 \\ \wedge \quad & \left| \frac{\vec{e}_x + \vec{e}_y}{||\vec{e}_x + \vec{e}_y||} \left(\vec{z}_x - \vec{z}_y\right)^{\top} \right| \leq \gamma_2 \\ \wedge \quad & ||\vec{z}_x - \vec{z}_y|| \leq \gamma_3 \left(\sqrt{\lambda_x} + \sqrt{\lambda_y}\right). \end{aligned}$$

$\vec{z}_x$ bzw. $\vec{z}_y$ sind das Clusterzentrum des Clusters x bzw. y, λ_x bzw. λ_y der größte Eigenwert der Kovarianzmatrix des Clusters x bzw. y und $\vec{e}_x$ bzw. $\vec{e}_y$ der normierte Eigenvektor, der dem kleinsten Eigenwert der Kovarianzmatrix des Clusters x bzw. y zugeordnet ist.

Das Compatible-Cluster-Merging ist nur für den Gustafson–Kessel-Algorithmus definiert. Eine Übertragung der Idee auf andere Fuzzy-Clusteringverfahren erfordert entsprechend angepaßte Ähnlichkeitskriterien zum Vereinigen der Cluster.

2.9.5 Similar-Cluster-Merging

Die Vorgehensweise des „Similar-Cluster-Merging“ [115] kann als Erweiterung bzw. Verallgemeinerung des Compatible-Cluster-Merging-Algorithmus angesehen werden. Wie bei dem CCM wird die Clusteranzahl ermittelt, indem, ausgehend von einer Clusteranalye mit einer zu großen Anzahl von Clustern, nach jedem Durchlauf des Fuzzy-Clusteringverfahrens ähnliche Cluster vereinigt werden.

Im Gegensatz zu dem Compatible-Cluster-Merging ist das Similar-Cluster-Merging nicht nur für den Gustafson–Kessel-Algorithmus sondern auch für andere Fuzzy-Clusteringverfahren geeignet. Dies wird erreicht, indem das Vereinigungskriterium nur auf den Zugehörigkeitsgraden $u_{i,j}$ und den Abständen der Daten zu den anderen Clustern basiert.

Das Ähnlichkeitskriterium a_{ik} zwischen zwei Clustern $\vec{\beta}_i$ und $\vec{\beta}_k$ ist definiert durch [115]

$$a_{ik} = \frac{\sum_{j=1}^{n}(u_{i,j} + u_{j,k}) \min\{d^2(\vec{x}_j, \vec{\beta}_i), d^2(\vec{x}_j, \vec{\beta}_k)\}}{\sum_{j=1}^{n}(u_{i,j} + u_{j,k}) \max\{d^2(\vec{x}_j, \vec{\beta}_i), d^2(\vec{x}_j, \vec{\beta}_k)\}}. \tag{2.43}$$

Es kann als Überlappungsgrad der beiden Cluster interpretiert werden. Die Form der Cluster wird nur indirekt mittels der Zugehörigkeitsgrade der Daten zu den Clustern $u_{i,j}$ und der Abstände der Daten zu den Clustern berücksichtigt.

Bei dem Similar-Cluster-Merging wird nach einer Clusteranalyse mit k Clustern die Ähnlichkeit zwischen allen Clustern bestimmt. Die beiden Cluster $\vec{\beta}_i$ und $\vec{\beta}_k$ mit der größten Ähnlichkeit a_{ik} werden vereinigt, wenn

$$a_{ij} > \gamma \tag{2.44}$$

gilt. Der Parameter $\gamma \in [0, 1]$ bestimmt, wie leicht Cluster vereinigt werden können. In [115] wird vorgeschlagen, γ durch

$$\gamma = f \cdot a^* \tag{2.45}$$

zu bestimmen. a^* ist die höchste Ähnlichkeit im ersten Durchlauf und $f \in (0,1)$ kann durch $\frac{c_{\text{opt}}}{c_{\max}}$ abgeschätzt werden. c_{opt} ist die vermutete Clusteranzahl und $c_{\max}$ die Clusteranzahl, mit der das Verfahren begonnen wurde.

Die beiden Cluster $\vec{\beta}_i$ und $\vec{\beta}_k$ werden vereinigt, indem die Zugehörigkeitsgrade $u_{i^*,j}$ des neuen Clusters durch

$$u_{i^*,j} = \min\{u_{i,j} + u_{k,j}, 1\}, \quad j = 1 \ldots n, \tag{2.46}$$

bestimmt werden. Nach der Vereinigung der beiden Cluster wird das Verfahren mit der reduzierten Clusteranzahl erneut durchgeführt. Dies wird solange wiederholt, bis sich die Clusteranzahl nicht mehr ändert.

2.10 Weitere Verfahren

2.10.1 Überblick

Der gebräuchliche Ansatz bei der Fuzzy-Clusteranalyse ist, das Klassifikationsproblem als Zielfunktion zu formulieren und durch alternierende Optimierung zu lösen. Neben der Vorgehensweise der alternierenden Optimierung gibt es weitere Verfahren zur Fuzzy-Clusteranalyse [3, 70, 57, 65, 105]. So kann die das Klassifikationsproblem beschreibende Zielfunktion z.B. mittels genetischer Algorithmen optimiert werden [3, 70]. Ein anderer Ansatz ist das *Alternating Cluster Estimation (ACE)* [65, 105]. Im Gegensatz zur zielfunktionsbasierten Fuzzy-Clusteranalyse müssen die Ausdrücke zur Berechnung der Zugehörigkeitsgrade und Clusterprototypen kein notwendige Kriterien für die Minimierung der das Klassifikationsproblem beschreibenden Zielfunktion sein. Im folgenden werden die Fuzzy-Clusteranalyse mit evolutionären Algorithmen und das Alternating Cluster Estimation kurz vorgestellt. Für eine ausführliche Betrachtung wird auf die Literatur verwiesen.

2.10.2 Fuzzy-Clusteranalyse mit evolutionären Algorithmen

Bei der zielfunktionsbasierten Fuzzy-Clusteranalyse wird i.a. vorausgesetzt, daß die Zielfunktion differenzierbar ist und zu einem partiell analytisch lösbaren Gleichungssystem führt. Dies bedeutet jedoch eine Einschränkung bei der Formulierung der Zielfunktion. Beispielsweise können Rechteckkonturen nur mit einer nicht differenzierbaren Zielfunktion charakterisiert wer-

den. Die Verwendung genetischer Algorithmen ermöglicht es, auch solche Zielfunktionen zu optimieren.

Evolutionäre Algorithmen sind Optimierungsverfahren, deren Idee auf der biologischen Evolution basiert [91]. Die gesuchte Lösung wird bei diesen Verfahren so kodiert, daß sie in Form eines Parameter-Vektors darstellbar ist. Dieser Vektor wird auch als Chromosom bezeichnet. Ausgehend von einer Startpopulation von Chromosomen werden die folgenden Populationen ermittelt.

Bei der Ermittlung einer neuen Population können einzelne Parameter geändert und Teile der Lösungen von zwei Chromosomen ausgetauscht werden. Diese Operationen sind durch die Mutation und das Crossover in der Biologie motiviert. Anschließend wird aus den Chromosomen mit den besten Lösungen die nächste Population ermittelt. Dieses Verfahren wird solange fortgeführt, bis entweder eine gute Lösung gefunden oder ein anderes Abbruchskriterium erreicht wurde.

Um ein Fuzzy-Clusteringverfahren mit genetischen Algorithmen zu lösen, kann man die Lösung entweder mittels der Zugehörigkeitsgrade der Daten zu den Klassen $\mathbf{U} = \{u_{1,1}, u_{1,2}, \ldots, u_{c,n}\}$ oder mittels der Beschreibung der Cluster $\mathbf{B} = \{\vec{\beta}_1, \vec{\beta}_2, \ldots, \vec{\beta}_c\}$ codieren. Die Verwendung der Clusterprototypen ist der Verwendung der Zugehörigkeitsgrade vorzuziehen, da sie eine kompaktere Codierung ermöglicht.

Eine ausführliche Behandlung dieser Thematik findet sich z.B. in [3, 70].

2.10.3 Alternating Cluster Estimation

Bei der Fuzzy-Clusteranalyse ist die gebräuchliche Vorgehensweise, die das Klassifikationsproblem beschreibende Zielfunktion durch alternierende Optimierung zu minimieren. Die dabei verwendeten Formeln zur Berechnung der Clusterprototypen und der Zugehörigkeitsgrade sind notwendige Kriterien zur Optimierung der Zielfunktion. Sie ergeben sich durch Nullsetzen der partiellen Ableitungen der Zielfunktion unter Berücksichtigung der Restriktionen.

Die Idee des „*Alternating Cluster Estimation (ACE)*“ ist, die Klassifikation analog der alternierenden Optimierung durch das abwechselnde Berechnen der Zugehörigkeitsgrade und der Clusterprototypen zu bestimmen [65]. Im Gegensatz zu der alternierenden Optimierung müssen die dafür verwendeten Ausdrücke jedoch *nicht* mehr notwendige Kriterien zur Minimierung der Zielfunktion sein. Die Klassifikationsaufgabe wird daher nicht mehr durch die Zielfunktion, sondern direkt durch die Formeln für die Berechnung der Zugehörigkeitsgrade bzw. der Clusterprototypen beschrieben.

Dies ermöglicht es, bei der Wahl der Zugehörigkeitsfunktion und der Berechnungsweise der Cluster flexibler zu sein.

Ein Vergleich mit dem *Expectation-Maximization-Algorithmus (EM-Algorithmus)* [40] zeigt die nahe Verwandtschaft der beiden Verfahren. Der EM-Algorithmus kann in zwei Teilschritte zerlegt werden, die iteriert werden:

- Berechnung der Wahrscheinlichkeit $p(\vec{x}_j|\Omega_i)$, daß ein Datum $\vec{x}_j$ der Klasse Ω_i zuzuordnen ist. (*Expectation*)
- Berechnung der Klassifikationsparameter der Klassen Ω_i durch Maximum-Likelihood-Schätzer. (*Maximization*)

Die bei der Ausführung des EM-Algorithmus unterstellten Verteilungen und Modellannahmen werden zuvor festgelegt.

Berechnung der Zugehörigkeitsgrade

Aus den Zugehörigkeitsgraden der Daten zu den Clustern lassen sich Zugehörigkeitsfunktionen ableiten. So können die probabilistischen Zugehörigkeitsgrade der Daten zu einem Cluster $\vec{\beta}_i$ als Zugehörigkeitsfunktion $\mu_{\text{prob},i}$

$$\mu_{\text{prob},i}(\vec{x}) = \begin{cases} 1 \,/ \sum_{j=1}^{c} \left(\frac{\|\vec{x}-\vec{z}_i\|_A}{\|\vec{x}-\vec{z}_j\|_A} \right)^{\frac{2}{m-1}} & \vec{x} \in \mathbb{R}^p \setminus \mathbf{Z}, \mathbf{Z} = \{\vec{z}_1, \ldots, \vec{z}_c\}, \\ 1 & \vec{x} = \vec{z}_i, \\ 0 & \vec{x} \in \mathbf{Z} \setminus \{\vec{z}_i\}, \mathbf{Z} = \{\vec{z}_1, \ldots, \vec{z}_c\} \end{cases} \tag{2.47}$$

und die possibilistischen Zugehörigkeitsgrade der Daten zu einem Cluster $\vec{\beta}_i$ als Zugehörigkeitsfunktion $\mu_{\text{poss},i}$

$$\mu_{\text{poss},i}(\vec{x}) = \frac{1}{1 + \left(\frac{\|\vec{x}-\vec{z}_i\|_A}{\sqrt{\eta_i}} \right)^{\frac{2}{m-1}}}, \quad \eta_i > 0, i = 1, \ldots, c \tag{2.48}$$

interpretiert werden. Der Verzicht auf die Forderung, daß die Zugehörigkeitsgrade die Zielfunktion optimieren müssen, ermöglicht es, anstelle dieser Zugehörigkeitsfunktion $\mu_{\text{prob},i}$ bzw. $\mu_{\text{poss},i}$ auch andere Zugehörigkeitsfunktionen, wie z.B. die Dreiecksfunktion

$$\mu_{\text{Dreieck},i}(\vec{x}) = \begin{cases} 1 - \left(\frac{\|\vec{x}-\vec{z}_i\|}{r_i} \right)^{\alpha} & \|\vec{x} - \vec{z}_i\| \leq r_i,\ \alpha \in \mathbb{R}_{>0} \\ 0 & \text{sonst} \end{cases} \tag{2.49}$$

oder die Exponentialfunktion

$$\mu_{\text{Exp},i}(\vec{x}) = e^{-\left(\frac{\|\vec{x}-\vec{z}_i\|}{\sigma_i}\right)^{\alpha}} \tag{2.50}$$

zu verwenden [65].

Berechnung der Clusterprototypen

Ebenso wie bei der Berechnung der Zugehörigkeitsgrade hat der Anwender auch bei der Berechnung der Clusterprototypen die Möglichkeit, von den Formeln des alternierenden Optimierens abzuweichen. Alternativen zu (2.5) für die Berechnung des Clusterzentrums sind z.B. [65]:

- semilineare Berechnung:

$$\vec{z}_i = \frac{\sum_{j=1}^{n} t_{\text{SLIDE}}(u_{i,j}, \alpha_1, \alpha_2)\vec{x}_j}{\sum_{j=1}^{n} t_{\text{SLIDE}}(u_{i,j}, \alpha_1, \alpha_2)} \tag{2.51}$$

$$t_{\text{SLIDE}}(u_{i,j}, \alpha_1, \alpha_2) = \begin{cases} u_{i,k} & \text{if } u_{i,k} \geq \alpha_1 \\ (1-\alpha_2)\, u_{i,k} & \text{if } u_{i,k} < \alpha_1 \end{cases}$$

  $\alpha_1, \alpha_2 \in [0,1]$

- oder modifizierte semilineare Berechnung:

$$\vec{z}_i = \alpha \frac{\sum_{u_{i,j}=\hat{u}_i} \vec{x}_j}{\sum_{u_{i,j}=\hat{u}_i} 1} + (1-\alpha)\frac{\sum_{j=1}^{n} u_{i,j}\vec{x}_j}{\sum_{j=1}^{n} u_{i,j}} \tag{2.52}$$

  $\alpha \in [0,1]$, $\hat{u}_i = \max\{u_{i,1}, u_{i,2}, \ldots, u_{i,n}\}$. Der Ausdruck $\frac{\sum_{u_{i,j}=\hat{u}_i} \vec{x}_j}{\sum_{u_{i,j}=\hat{u}_i} 1}$ kann als „Mean of Maxima“ interpretiert werden [65].

Kapitel 3

Erweiterung der possibilistischen Fuzzy-Clusteranalyse

3.1 Problematik der possibilistischen Fuzzy-Clusteranalyse

Bei der Fuzzy-Clusteranalyse werden meistens probabilistische Zugehörigkeitsgrade verwendet. Hierbei hat jedes Datum das gleiche Gewicht. Diese Verfahren sind robust, ihr Nachteil ist jedoch, daß die Zugehörigkeitsgrade nicht angeben, wie typisch ein Datum für einen Cluster ist. Die Interpretierbarkeit der Zugehörigkeitsgrade ist damit eingeschränkt. Bei Daten, die typisch für zwei Cluster sind, ist der probabilistische Zugehörigkeitsgrad zu jedem der beiden Cluster 0.5. Bei einer größeren Überschneidung von zwei Clustern — es gibt viele Daten, die beiden Clustern angehören — führt dies dazu, daß die Form der Cluster nicht richtig erkannt wird. Es wird eine stärke Separation der Cluster angezeigt als sie in den Daten vorliegt. Eine Alternative ist die Verwendung possibilistischer Zugehörigkeitsgrade (Abschnitt 2.7).

Bei der possibilistischen Clusteranalyse wird die Zielfunktion $J(\mathbf{X}, \mathbf{U}, \mathbf{B})$

$$J(\mathbf{X}, \mathbf{U}, \mathbf{B}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{i,j}^{m} d^2(\vec{\beta}_i, \vec{x}_j) + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} (1 - u_{i,j})^m \tag{3.1}$$

minimiert unter Berücksichtigung der Restriktion, daß jedem Cluster Daten zugeordnet werden. Der erste Term der Zielfunktion bewertet die Summe der mit den Zugehörigkeitsgraden gewichteten Abstände. Der zweite Term der Zielfunktion verhindert die triviale Lösung, d.h. er verhindert, daß alle Zugehörigkeitsgrade null sind. $\eta_i \in \mathbb{R}_{>0}$ ist ein clusterspezifischer Parameter, der die beiden divergierenden Ziele der Terme zueinander gewichtet.

Auf den ersten Blick sieht dieser Ansatz vielversprechend aus. Wenn man ihn jedoch näher betrachtet, sieht man, daß in dem globalen Optimum der Zielfunktion *j alle* Cluster identisch sind. *Eine optimale Lösung der Zielfunktion eines possibilistischen Fuzzy-Clusteringverfahrens liegt dann vor, wenn in dem Datensatz genau* ein *Cluster erkannt wird.* Die Heterogenitätsforderung der Clusteranalyse — Daten, die zu verschiedenen Clustern gehören, sollen möglichst heterogen sein — wird nicht erfüllt.

Die Ursache dieser unerwünschten Eigenschaft ist, daß bei der Minimierung der Zielfunktion lediglich der Abstand der Daten zu den Clustern und nicht auch die Lage der anderen Cluster berücksichtigt wird. Die Zielfunktion modelliert durch die Minimierung der Abstände zwischen den Clustern und den ihnen zugeordneten Daten *nur* die Forderung der Clusteranalyse, daß Daten, die zu dem gleichen Cluster gehören, möglichst homogen sein sollen. Da kein partitionierender Effekt wie bei der probabilistischen Fuzzy-Clusteranalyse auftritt, gibt es daher „ein“ „optimales“ Clusterzentrum. Bei einer „optimalen“ Lösung der Zielfunktion liegen alle Clusterzentren auf diesem „optimalen“ Clusterzentrum. Nur in sehr seltenen Fällen sehr hoher Symmetrie liegen mehrere optimale Clusterzentren vor.

Die Eigenschaft der possibilistischen Fuzzy-Clusteranalyse, daß bei einer optimalen Lösung alle Cluster identisch sind, wird im folgenden erläutert. Hierfür nehmen wir an, daß ein Datensatz in zwei Cluster $\vec{\beta}_1$ und $\vec{\beta}_2$ zu unterteilen ist. Die beiden Cluster seien nicht identisch. sum_i ist der Wert, den der Cluster β_i zu der Zielfunktion $J(\mathbf{X}, \mathbf{U}, \mathbf{B})$ beiträgt.

$$\text{sum}_i = \sum_{j=1}^{n} u_{i,j}^m d^2(\vec{\beta}_i, \vec{x}_j) + \eta_i \sum_{j=1}^{n} (1 - u_{i,j})^m, \quad i = 1, 2. \tag{3.2}$$

Mit Ausnahme weniger sehr seltener Fälle einer hohen Symmetrie der Daten gilt entweder $\text{sum}_1 > \text{sum}_2$ oder $\text{sum}_2 > \text{sum}_1$. O.B.d.A. sei $\text{sum}_2 > \text{sum}_1$.

Es gilt: $J(\mathbf{X}, \mathbf{U}, \mathbf{B}) = \text{sum}_1 + \text{sum}_2$.

Da $\text{sum}_2 > \text{sum}_1$, kann $J(\mathbf{X}, \mathbf{U}, \mathbf{B})$ auf $\text{sum}_1 + \text{sum}_1 < \text{sum}_1 + \text{sum}_2$ verkleinert werden, indem $\vec{\beta}_2$ auf $\vec{\beta}_1$ gesetzt wird. Im Optimum der Zielfunktion gilt also: $\vec{\beta}_1 = \vec{\beta}_2$. Diese Argumentation kann analog auf mehrere Cluster übertragen werden.

Obwohl im Optimum der Zielfunktion eines possibilistischen Clusteringverfahrens alle Cluster identisch sind[1], führt die possibilistische Fuzzy-Clusteranalyse normalerweise zu guten Ergebnissen. (Zur Initialisierung wird die probabilistische Fuzzy-Clusteranalyse verwendet.) Es werden verschiedene Cluster erkannt. Das Erkennen *verschiedener* Cluster bedeutet jedoch, daß nur ein lokales und kein globales Optimum gefunden wurde. *Es wird also bei einem zielfunktionsbasierten Verfahren eine suboptimale Lösung gesucht.* Dies ist aus theoretischer Sicht extrem unbefriedigend! Es werden daher in diesem Kapitel ein neues zielfunktionsbasiertes possibilistisches Fuzzy-Clusteringverfahren und ein neues possibilistisches Fuzzy-Clusteringverfahren auf der Grundlage des ACE vorgestellt, die possibilistische Zugehörigkeitsgrade verwenden und die Erkennung identischer Cluster verhindern.

3.2 Ein possibilistisches Fuzzy-Clusteringverfahren basierend auf Clusterabstoßung

Die Idee des in diesem Abschnitt vorgestellten possibilistischen Fuzzy-Clusteringverfahrens ist, die Anziehung der Cluster durch die Daten mit einer Abstoßung zwischen verschiedenen Clustern zu kombinieren. Dies wird nicht indirekt wie bei der probabilistischen Clusteranalyse durch eine Restriktion — für jedes Datum ist die Summe der Zugehörigkeitsgrade zu allen Clustern gleich 1 — sondern direkt durch eine Modifikation der Zielfunktion modelliert [127, 128]. Die Abstoßung zwischen verschiedenen Clustern kann als Heterogenitätsforderung interpretiert werden. Die die jeweilige Klasse repräsentierenden Daten sollen möglichst unähnlich sein. Die Zielfunktion beinhaltet bei diesem Ansatz daher sowohl die Homogenitätsforderung für Daten, die zu dem gleichen Cluster gehören, (Minimierung der Abstände zwischen den Clustern bzw. ihren typischen Daten und den ihnen zugeordneten Daten) und die Heterogenitätsforderung zwischen Daten, die zu verschiedenen Clustern gehören (Maximierung der Abstände zwischen den Clustern bzw. zwischen den typischen Daten der Cluster).

Eine geeignete Zielfunktion sollte die folgenden Kriterien erfüllen:

- Der Abstand zwischen den Clustern und den ihnen zugeordneten Daten soll minimiert werden.

[1] Mit Ausnahme einiger seltener Fälle hoher Symmetrie in den Datensätzen.

- Der Abstand zwischen den Clustern soll maximiert werden.
- Es soll keine leeren Cluster geben, d.h. jedem Cluster sind Daten zuzuordnen.
- Die Zugehörigkeitsgrade der Daten sollen möglichst eindeutig, d.h. nahe 1 bzw. nahe 0 sein. Die triviale Lösung, alle Zugehörigkeitsgrade $u_{i,j}$ sind 0, ist zu verhindern.

Die Forderungen sind denen der in Abschnitt 2.7 vorgestellten possibilistischen Fuzzy-Clusteranalyse sehr ähnlich. Sie werden wie folgt modelliert:

- Die Anziehung zwischen den Daten und den Clustern ist durch den Ausdruck $\sum_{i=1}^{c}\sum_{j=1}^{n} u_{i,j}^{m} d^2(\vec{\beta}_i, \vec{x}_j)$ modelliert.
- Um die triviale Lösung zu vermeiden, wird wie in (2.22) der Term $\sum_{i=1}^{c} \eta_i \sum_{j=1}^{n}(1-u_{i,j})^m$ verwendet.
- Die Forderung, daß jedem Cluster Daten zuzuordnen sind, wird durch die Restriktion $\sum_{j=1}^{n} u_{i,j} > 0 \; \forall i \in \{0, \ldots, c\}$ berücksichtigt.
- Die Abstoßung zwischen den Clustern wird durch einen Term analog zu der Anziehung zwischen den Daten und den Clustern modelliert. Dieser Term hat einen kleineren Wert, wenn die Abstände zwischen den Clustern größer werden.

Die einfachste Möglichkeit, die Abstoßung zwischen den Clustern zu modellieren, ist, die Summe des quadrierten Abstands zwischen den Clustern in der Zielfunktion zu subtrahieren. Dieser naheliegende Ansatz ist jedoch nicht geeignet, da diese Summe mit zunehmendem Abstand stärker wächst als die Summe der Abstände zwischen den Clustern und den ihnen zugeordneten Daten. D.h. die Abstände zwischen den Clustern sind das dominierende Klassifikationskriterium. Eine „optimale“ Lösung ist bei dieser Modellierung, alle Daten *einem* Cluster zuzuordnen und alle Cluster möglichst weit voneinander zu entfernen.

Um diese unerwünschte „Explosion“ der Cluster zu vermeiden, sollte der Einfluß der Abstoßung zwischen den Clustern mit zunehmendem Abstand zwischen den Clustern abnehmen, so daß ab einem bestimmten Abstand die Anziehung der Cluster durch die Daten größer ist als die Abstoßung durch die anderen Cluster. Wenn die Cluster gut separiert sind, sollte die Klassifikation (nahezu ausschließlich) auf den Abständen der Daten zu den Clustern basieren. Daneben ist sicherzustellen, daß bei einem sehr geringen Abstand zwischen zwei Clustern die Abstoßung zwischen den Clustern

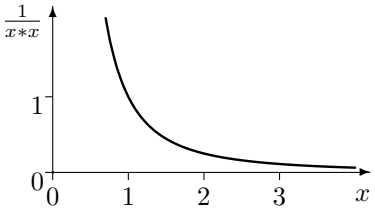

Abbildung 3.1: Die Funktion $f(x) = \frac{1}{x*x}$.

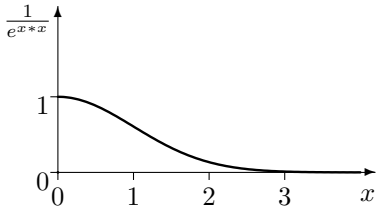

Abbildung 3.2: Die Funktion $f(x) = \frac{1}{e^{x*x}}$.

das dominierende Kriterium bei der Berechnung der Clusterprototypen ist. Die Abstoßung zwischen den Clustern kann z.B. durch $\sum_{k=1,k\neq i}^{c} \frac{1}{d^2(\vec{\beta}_i,\vec{\beta}_j)}$ oder durch $\sum_{k=1,k\neq i}^{c} e^{-d^2(\vec{\beta}_i,\vec{\beta}_k)}$ modelliert werden, vgl. Abb. 3.1 und 3.2. Bei beiden Funktionen ist ab einem Abstand von ungefähr 2 die Abstoßung zwischen benachbarten Clustern vernachlässigbar. Bei der Modellierung der Abstoßung durch den Ausdruck $\frac{1}{d^2(\vec{\beta}_i,\vec{\beta}_j)}$ ist bei dicht benachbarten Clustern die Abstoßung sehr groß. Demgegenüber ist bei der Modellierung mittels $e^{-d^2(\vec{\beta}_i,\vec{\beta}_k)}$ durch die Wahl eines Parameters γ_i, sicherzustellen, daß die Abstoßung nah benachbarter Cluster hinreichend groß ist.

Bei der Berechnung der Strafterme für die Abstoßung zwischen zwei benachbarten Clustern ist ggf. der Abstand zwischen zwei Clustern $d^2(\vec{\beta}_i, \vec{\beta}_k)$ durch die Verwendung von $\zeta \cdot d^2(\vec{\beta}_i, \vec{\beta}_k), \zeta \in \mathbb{R}_+$ zu skalieren, um die Abstoßung den in dem Datensatz vorliegenden Abständen anzupassen.

Im folgenden Abschnitt werden possibilistische Varianten des Fuzzy-C-Means-Algorithmus und des Gustafson–Kessel-Algorithmus betrachtet, bei denen die Abstoßung zwischen den Clustern durch die Ausdrücke $\gamma_i \sum_{k=1,k\neq i}^{c} \frac{1}{\varsigma d^2(\vec{\beta}_i,\vec{\beta}_j)}$ bzw. $\gamma_i \sum_{k=1,k\neq i}^{c} e^{\zeta - d^2(\vec{\beta}_i,\vec{\beta}_k)}$ modelliert wird. Bei der Verwendung anderer Ausdrücke für die Abstoßung zwischen benachbarten Clustern ist die Berechnung der Clusterprototypen entsprechend zu modifizieren.

Die Zielfunktion des Klassifikationsproblems ist

$$J_1(\mathbf{X},\mathbf{U},\mathbf{B}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{i,j}^m d^2(\vec{\beta}_i,\vec{x}_j) + \sum_{i=1}^{c}\eta_i\sum_{j=1}^{n}(1-u_{i,j})^m + \sum_{i=1}^{c}\gamma_i \sum_{k=1,k\neq i}^{c} \frac{1}{\zeta d^2(\vec{\beta}_i,\vec{\beta}_k)} \tag{3.3}$$

bzw.

$$J_2(\mathbf{X},\mathbf{U},\mathbf{B}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{i,j}^m d^2(\vec{\beta}_i,\vec{x}_j) + \sum_{i=1}^{c}\eta_i\sum_{j=1}^{n}(1-u_{i,j})^m + \sum_{i=1}^{c}\gamma_i \sum_{k=1,k\neq i}^{c} e^{-\zeta d^2(\vec{\beta}_i,\vec{\beta}_k)}. \tag{3.4}$$

Diese Funktion ist unter Berücksichtigung der Restriktion $\sum_{j=1}^{n} u_{i,j} > 0$ für alle $i \in \{1,\ldots,c\}$ zu minimieren. γ_i ist ein clusterspezifischer Parameter, der für den Cluster $\vec{\beta}_i$ die Abstoßung durch die anderen Cluster in Relation zu der Anziehung durch die dem Cluster zugeordneten Daten setzt.

Die Minimierung von (3.3) bzw. (3.4) nach den Zugehhörigkeitsgraden führt zu possibilistischen Zugehörigkeitsgraden, vgl. (2.23).

3.3 Berechnung der Clusterprototypen

Die Berechnung der Clusterprototypen hängt von dem verwendeten Fuzzy-Clusteringverfahren ab. Im folgenden werden der Fuzzy-C-Means-Algorithmus und der Gustafson–Kessel-Algorithmus betrachtet.[2]

3.3.1 Variante des Fuzzy-C-Means-Algorithmus

Bei dem Fuzzy-C-Means-Algorithmus werden die Cluster nur durch ihre Zentren $\vec{z}_i$ beschrieben. Es wird der euklidische Abstand verwendet.

[2] Der FMLE wird nicht betrachtet, da die Verbindung der possibilistischen Zugehörigkeitsgrade mit der exponentiellen Abstandsfunktion sehr häufig zu schlechten Klassifikationsergebnissen führt.

Die Ableitung der Zielfunktion (3.3) nach den Clusterzentren unter Berücksichtigung der Restriktion (2.3) führt zu

$$\sum_{j=1}^{n} u_{i,j}^{m}(\vec{x}_j - \vec{z}_i) - \frac{\gamma_i}{\zeta} \sum_{k=1,k\neq i}^{c} \frac{1}{d^2(\vec{z}_k,\vec{z}_i)\cdot d^2(\vec{z}_k,\vec{z}_i)}(\vec{z}_k - \vec{z}_i) = 0. \qquad (3.5)$$

Dies wird im folgenden gezeigt.

Die Clustereinteilung minimiere die Zielfunktion $J_1(\mathbf{X},\mathbf{U},\mathbf{B})$. Dann sind alle Richtungsableitungen von J_1 nach $\vec{z}_i$ gleich 0. Daher gilt für alle $\vec{\xi} \in \mathbb{R}^p$ mit $t \in \mathbb{R}$:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \vec{z}_i} J_1(\mathbf{X},\mathbf{U},\mathbf{B}) \\
&= \frac{\partial}{\partial \vec{z}_i}\left(\sum_{i=1}^{c}\sum_{j=1}^{n} u_{i,j}^{m} d^2(\vec{\beta}_i,\vec{x}_j) + \sum_{i=1}^{c}\eta_i \sum_{j=1}^{n}(1-u_{i,j})^m \right. \\
&\quad \left. + \sum_{i=1}^{c}\gamma_i \sum_{k=1,k\neq i}^{c} \frac{1}{\zeta d^2(\vec{\beta}_i,\vec{\beta}_k)}\right) \\
&= \sum_{j=1}^{n} u_{i,j}^{m} \frac{\partial}{\partial \vec{z}_i}||\vec{x}_j - \vec{z}_i||^2 + \frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c}\frac{\partial}{\partial \vec{z}_i}\left(\frac{1}{||\vec{z}_k - \vec{z}_i||^2}\right) \\
&= \sum_{j=1}^{n} u_{i,j}^{m} \lim_{t\to 0}\frac{||\vec{x}_j - (\vec{z}_i + t\vec{\xi})||^2 - ||\vec{x}_j - \vec{z}_i||^2}{t} \\
&\quad + \frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c}\lim_{t\to 0}\left(\frac{1}{t}\left(\frac{1}{||\vec{z}_k - (\vec{z}_i + t\vec{\xi})||^2} - \frac{1}{||\vec{z}_k - \vec{z}_i||^2}\right)\right) \\
&= \sum_{j=1}^{n} u_{i,j}^{m} \lim_{t\to 0}\frac{||(\vec{x}_j - \vec{z}_i) - t\vec{\xi}||^2 - ||\vec{x}_j - \vec{z}_i||^2}{t} \\
&\quad + \frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c}\lim_{t\to 0}\left(\frac{1}{t}\left(\frac{1}{||(\vec{z}_k - \vec{z}_i) - t\vec{\xi}||^2} - \frac{1}{||\vec{z}_k - \vec{z}_i||^2}\right)\right) \\
&= \sum_{j=1}^{n} u_{i,j}^{m} \lim_{t\to 0}\frac{-2t(\vec{x}_j - \vec{z}_i)^\top\vec{\xi} + t^2\vec{\xi}^\top\vec{\xi}}{t} \\
&\quad + \frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c}\lim_{t\to 0}\left(\frac{1}{t}\frac{2t(\vec{z}_k - \vec{z}_i)^\top\vec{\xi} - t^2\vec{\xi}^\top\vec{\xi}}{||(\vec{z}_k - \vec{z}_i) - t\vec{\xi}||^2\cdot||\vec{z}_k - \vec{z}_i||^2}\right)
\end{aligned}$$

$$= -2\sum_{j=1}^{n} u_{i,j}^{m}(\vec{x}_j - \vec{z}_i)^\top \vec{\xi} + \frac{\gamma_i}{\zeta} \sum_{k=1,k\neq i}^{c} \frac{2(\vec{z}_k - \vec{z}_i)^\top \vec{\xi}}{||\vec{z}_k - \vec{z}_i||^2 \cdot ||\vec{z}_k - \vec{z}_i||^2}$$

$$\Rightarrow$$

$$0 = \sum_{j=1}^{n} u_{i,j}^{m}(\vec{x}_j - \vec{z}_i) - \frac{\gamma_i}{\zeta} \sum_{k=1,k\neq i}^{c} \frac{1}{d^2(\vec{z}_k, \vec{z}_i) \cdot d^2(\vec{z}_k, \vec{z}_i)}(\vec{z}_k - \vec{z}_i)$$

Analytisch ist das Gleichungssystem (3.5) nicht allgemein für beliebige $p \in \mathbb{N}_{>0}$ lösbar. Alternativ bietet sich daher eine iterative Berechnung der Clusterzentren durch

$$\vec{z}_i = \frac{\sum_{j=1}^{n} u_{i,j}^{m}\vec{x}_j - \frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c} \frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\vec{z}_k}{\sum_{j=1}^{n} u_{i,j}^{m} - \frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c} \frac{1}{d^4(\vec{z}_k,\vec{z}_i)}} \tag{3.6}$$

an. (3.6) erhält man durch Auflösung von (3.5) nach $\vec{z}_i$. Die Abstände zwischen den Clustern $d^2(\vec{z}_k, \vec{z}_i)$ werden dabei nicht nach $\vec{z}_i$ aufgelöst. Die Berechnung von $\vec{z}_i$ wird solange wiederholt, bis $||\vec{z}_i^{(\text{new})} - \vec{z}_i^{(\text{old})}|| < \epsilon$ gilt.

Wenn man sich vorübergehend von der dynamischen Betrachtung löst und in jeder Iteration die Clusterzentren wie Datenpunkte, d.h. fixe Punkte, betrachtet, kann in (3.6) der Ausdruck $\frac{1}{d^4(\vec{z}_k,\vec{z}_i)}$ als Zugehörigkeitsgrad bzw. Einflußgrad des Clusters $\vec{\beta}_k$ zu dem Cluster $\vec{\beta}_i$ analog zu dem Zugehörigkeitsgrad $u_{i,j}^m$ des Datums $\vec{x}_j$ zu dem Cluster $\vec{\beta}_i$ gedeutet werden. D.h. das neue Zentrum wird berechnet als das mit Zugehörigkeitsgraden gewichtete Mittel der Daten und der mit den (negativen) Einflußgraden gewichteten anderen Clusterzentren. Bei dieser Betrachtung entfällt die Problematik der iterativen Berechnung der Clusterzentren innerhalb der Iterationen der alternierenden Optimierung des Fuzzy-Clusteringverfahrens.

Das Ziel der Berechnung der Clusterprototypen ist, zu verhindern, daß identische Cluster gefunden werden. Die exakte Stärke der Abstoßung ist in den ersten Iterationen des Fuzzy-Clusteringverfahrens demgegenüber noch nicht so wichtig. In den späteren Iterationen verändern die Clusterzentren ihre Lage nur noch geringfügig. Es gilt $\vec{z}_i^{(neu)} \approx \vec{z}_i^{(alt)}$. Die nach (3.6) berechnete Lösung des Fuzzy-Clusteringverfahrens erfüllt damit das notwendige Kriterium für eine optimale Lösung der Zielfunktion.

Analog kann gezeigt werden, daß die Ableitung der Zielfunktion (3.4)

nach den Clusterzentren unter Berücksichtigung der Restriktion (2.3) zu

$$\sum_{j=1}^{n} u_{i,j}^{m}(\vec{x}_j - \vec{z}_i) - \gamma_i\zeta \sum_{k=1,k\neq i}^{c} (\vec{z}_k - \vec{z}_i)e^{-\zeta d^2(\vec{z}_k,\vec{z}_i)} = 0 \tag{3.7}$$

führt.

Ebenso wie (3.5) ist auch das Gleichungssystem (3.4) nicht allgemein analytisch lösbar. Die Clusterzentren werden daher durch

$$\vec{z}_i = \frac{\sum_{j=1}^{n} u_{i,j}^{m}\vec{x}_j - \gamma_i\zeta \sum_{k=1,k\neq i}^{c} \vec{z}_k e^{-\zeta d^2(\vec{z}_k,\vec{z}_i)}}{\sum_{j=1}^{n} u_{i,j}^{m} - \gamma_i\zeta \sum_{k=1,k\neq i}^{c} e^{-\zeta d^2(\vec{z}_k,\vec{z}_i)}} \tag{3.8}$$

berechnet. Analog zu (3.6) werden auch bei (3.8) auf der rechten Seite die Werte der vorhergehenden Iteration verwendet.

Sowohl (3.6) als auch (3.8) verdeutlichen das Charakteristikum dieses Ansatzes. Die Clusterzentren werden von den ihnen zugeordneten Daten angezogen und von den anderen Clustern abgestoßen. Die Ausdrücke zur Berechnung der Clusterzentren zeigen, daß die Abstoßung zwischen den Clustern bei der Berechnung der Clusterprototypen durch das „Entfernen“ bzw. „Nichtberücksichtigen“ von Daten aus der Gegend eines benachbarten Clusterzentrums modelliert wird. Das mit einem „Einflußgrad“ gewichtete Clusterzentrum wird bei der Berechnung des Zentrums subtrahiert. Diese Vorgehensweise ist jedoch nur sinnvoll, wenn eine Anziehung vorliegt. Damit der Ausdruck sinnvoll bleibt, muß daher

$$\sum_{j=1}^{n} u_{i,j}^{m} - \frac{\gamma_i}{\zeta} \sum_{k=1,k\neq i}^{c} \frac{1}{d^4(\vec{z}_k,\vec{z}_i)} > 0$$

bzw.

$$\sum_{j=1}^{n} u_{i,j}^{m} - \gamma_i\zeta \sum_{k=1,k\neq i}^{c} e^{-\zeta d^2(\vec{z}_k,\vec{z}_i)} > 0$$

gelten. Die Problematik wird im folgenden kurz erläutert.

Falls zwei Cluster $\vec{\beta}_i$ und $\vec{\beta}_l$ nahezu identisch sind, ist die abstoßende Kraft zwischen den beiden Clustern wesentlich größer als die anziehende Kraft durch die Daten oder die abstoßende Kraft durch andere Cluster. Dies führt bei der Berechnung der Clusterzentren nach (3.6) zu

$$\vec{z}_i = \frac{\sum_{j=1}^{n} u_{i,j}^{m}\vec{x}_j - \frac{\gamma_i}{\zeta} \sum_{k=1,k\neq i}^{c} \frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\vec{z}_k}{\sum_{j=1}^{n} u_{i,j}^{m} - \frac{\gamma_i}{\zeta} \sum_{k=1,k\neq i}^{c} \frac{1}{d^4(\vec{z}_k,\vec{z}_i)}}$$

$$
\begin{aligned}
&\approx \frac{-\frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c}\frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\vec{z}_k}{-\frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c}\frac{1}{d^4(\vec{z}_k,\vec{z}_i)}} \\
&\approx \frac{-\frac{\gamma_i}{\zeta}\frac{1}{d^4(\vec{z}_l,\vec{z}_i)}\vec{z}_l}{-\frac{\gamma_i}{\zeta}\frac{1}{d^4(\vec{z}_l,\vec{z}_i)}} \\
&\approx \vec{z}_l.
\end{aligned}
$$

Analog gilt für die Berechnung des Zentrums $\vec{z}_l$: $\vec{z}_l \approx \vec{z}_i$. Die Cluster tauschen ihre Lage.

Der Fall, daß bei einem Cluster $\vec{\beta}_i$ das Gewicht der Abstoßung durch die Cluster stärker als das Gewicht der Anziehung durch die Daten ist, kann so interpretiert werden, daß für die Beschreibung der in diesem „Gebiet“ des Datensatzes vorliegenden Cluster der Cluster $\vec{\beta}_i$ nicht benötigt wird. Es bietet sich daher an, diesen Cluster in ein anderes „Gebiet“ des Datensatzes zu „legen“, in dem weniger Cluster sind. Dies kann z.B. erfolgen, indem man das Clusterzentrum dieses Clusters auf das Datum mit dem geringsten maximalen Zugehörigkeitsgrad zu einem Cluster ($\min_{1\leq j\leq n}\{\max_{1\leq i\leq c}\{u_{i,j}\}\}$) setzt. Hierbei sollte man sicherstellen, daß es sich nicht um ein Stördatum handelt. D.h. es sollten in der Umgebung des Datums weitere Daten liegen.

3.3.2 Variante des Gustafson–Kessel-Algorithmus

Bei dem Gustafson–Kessel-Algorithmus werden die Cluster durch ihre Zentren $\vec{z}_i$ und ihre Kovarianzmatrizen $\mathbf{Cov}_i$ bzw. die aus $\mathbf{Cov}_i$ abgeleitete Normmatrix $\mathbf{A}_i$ beschrieben. Es wird der Mahalanobis-Abstand verwendet. Der Abstand eines Datums zu einem Cluster ist von den Parametern des Clusters abhängig. Im Gegensatz zu dem Fuzzy-C-Means-Algorithmus ist der Abstand zwischen zwei Clustern bei dem Gustafson–Kessel-Algorithmus nicht eindeutig definiert. Es kann z.B. der euklidische Abstand zwischen den Clusterzentren oder der Mahalanobis-Abstand zwischen den Clustern mit Verwendung der Normmatrizen *beider* Cluster verwendet werden. Die Verwendung des Mahalanobis-Abstands bietet gegenüber dem euklidischen Abstand den Vorteil, daß die Form der Cluster mit berücksichtigt wird. Daher wird in diesem Abschnitt der Abstand zwischen Clustern durch den Mahalanobis-Abstand berechnet.

Der Abstand zwischen zwei Clustern $\vec{\beta}_i$ und $\vec{\beta}_k$ wird in diesem Abschnitt als

$$
d^2(\vec{\beta}_i,\vec{\beta}_k) = \frac{1}{2}\left((\vec{z}_i-\vec{z}_k)^\top\mathbf{A}_i(\vec{z}_i-\vec{z}_k) + (\vec{z}_i-\vec{z}_k)^\top\mathbf{A}_k(\vec{z}_i-\vec{z}_k)\right) \qquad (3.9)
$$

definiert.

Alternativ könnte auch der Abstand

$$d^2(\vec{\beta}_i, \vec{\beta}_k) = (\vec{z}_i - \vec{z}_k)^\top \mathbf{A}_i (\vec{z}_i - \vec{z}_k)$$

verwendet werden. Der Abstand eines Clusters $\vec{\beta}_i$ zu einem anderen Cluster $\vec{\beta}_k$ wird in diesem Fall *nur* unter Berücksichtigung der Form des Clusters $\vec{\beta}_i$ bestimmt. Dies hat jedoch den Nachteil, daß $d^2(\vec{\beta}_i, \vec{\beta}_k) = d^2(\vec{\beta}_k, \vec{\beta}_i)$ *nur* gilt, wenn beide Cluster die gleiche Form haben. Ansonsten ist $d^2(\vec{\beta}_i, \vec{\beta}_k) \neq d^2(\vec{\beta}_k, \vec{\beta}_i)$. Dies bedeutet, daß die abstoßende Wirkung, die der Cluster $\vec{\beta}_k$ auf den Cluster $\vec{\beta}_i$ ausübt, ungleich der ist, die der Cluster $\vec{\beta}_i$ auf den Cluster $\vec{\beta}_k$ ausübt. Dies widerspricht jedoch der Intuition!

Die Verwendung des Abstands (3.9) führt bei der Variante des Algorithmus von Gustafson–Kessel für die Zielfunktion $J_1(\mathbf{X}, \mathbf{U}, \mathbf{B})$ zu folgendem Ausdruck für die Berechnung der Clusterzentren:

$$\sum_{j=1}^{n} u_{i,j}^m (\vec{x}_j - \vec{z}_i) - \frac{\gamma_i}{\zeta} \sum_{k=1, k\neq i}^{c} \frac{1}{d^4(\vec{z}_k, \vec{z}_i)} \frac{1}{2} \left((\vec{z}_k - \vec{z}_i)^\top \mathbf{A}_i + (\vec{z}_k - \vec{z}_i)^\top \mathbf{A}_k \right) = 0. \tag{3.10}$$

Dies wird im folgenden gezeigt.

Die Clustereinteilung minimiere die Zielfunktion $J_1(\mathbf{X}, \mathbf{U}, \mathbf{B})$. Dann sind alle Richtungsableitungen von J_1 nach $\vec{z}_i$ gleich 0. Mit

$$(\vec{x}_j - \vec{z}_i)^\top \mathbf{A}_i (\vec{x}_j - \vec{z}_i) = ||\vec{x}_j - \vec{z}_i||^2_{\mathbf{A}_i}$$

gilt für alle $\vec{\xi} \in \mathbb{R}^p$ mit $t \in \mathbb{R}$:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \vec{z}_i} J_1(\mathbf{X}, \mathbf{U}, \mathbf{B}) \\
&= \frac{\partial}{\partial \vec{z}_i} \left(\sum_{i=1}^{c} \sum_{j=1}^{n} u_{i,j}^m d^2(\vec{\beta}_i, \vec{x}_j) + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} (1 - u_{i,j})^m \right. \\
&\quad \left. + \sum_{i=1}^{c} \gamma_i \sum_{k=1, k\neq i}^{c} \frac{1}{\zeta d^2(\vec{\beta}_i, \vec{\beta}_k)} \right) \\
&= \sum_{j=1}^{n} u_{i,j}^m \frac{\partial}{\partial \vec{z}_i} ||\vec{x}_j - \vec{z}_i||^2_{\mathbf{A}_i} \\
&\quad + \frac{\gamma_i}{\zeta} \sum_{k=1, k\neq i}^{c} \frac{\partial}{\partial \vec{z}_i} \left(\frac{1}{\frac{1}{2} \left(||\vec{z}_i - \vec{z}_k||^2_{\mathbf{A}_i} + ||\vec{z}_i - \vec{z}_k||^2_{\mathbf{A}_k} \right)} \right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{n} u_{i,j}^{m} \lim_{t\to 0} \frac{||\vec{x}_j - (\vec{z}_i + t\vec{\xi})||_{\mathbf{A}_i}^2 - ||\vec{x}_j - \vec{z}_i||_{\mathbf{A}_i}^2}{t} \\
&\quad + \frac{\gamma_i}{\zeta} \sum_{k=1,k\neq i}^{c} \lim_{t\to 0} \left(\frac{1}{t} \left(\frac{1}{\frac{1}{2}\left(||(\vec{z}_i + t\vec{\xi}) - \vec{z}_k||_{\mathbf{A}_i}^2 + ||(\vec{z}_i + t\vec{\xi}) - \vec{z}_k||_{\mathbf{A}_k}^2 \right)} \right. \right. \\
&\quad \left. \left. - \frac{1}{\frac{1}{2}\left(||\vec{z}_i - \vec{z}_k||_{\mathbf{A}_i}^2 + ||\vec{z}_i - \vec{z}_k||_{\mathbf{A}_k}^2 \right)} \right) \right) \\
&= \sum_{j=1}^{n} u_{i,j}^{m} \lim_{t\to 0} \frac{||(\vec{x}_j - \vec{z}_i) - t\vec{\xi})||_{\mathbf{A}_i}^2 - ||\vec{x}_j - \vec{z}_i||_{\mathbf{A}_i}^2}{t} \\
&\quad + \frac{\gamma_i}{\zeta} \sum_{k=1,k\neq i}^{c} \lim_{t\to 0} \left(\frac{1}{t} \left(\frac{1}{\frac{1}{2}\left(||(\vec{z}_i - \vec{z}_k) + t\vec{\xi}||_{\mathbf{A}_i}^2 + ||(\vec{z}_i - \vec{z}_k) + t\vec{\xi})||_{\mathbf{A}_k}^2 \right)} \right. \right. \\
&\quad \left. \left. - \frac{1}{\frac{1}{2}\left(||\vec{z}_i - \vec{z}_k||_{\mathbf{A}_i}^2 + ||\vec{z}_i - \vec{z}_k||_{\mathbf{A}_k}^2 \right)} \right) \right) \\
&= \sum_{j=1}^{n} u_{i,j}^{m} \lim_{t\to 0} \frac{-2t(\vec{x}_j - \vec{z}_i)^\top \mathbf{A}_i \vec{\xi} + t^2 \vec{\xi}^\top \mathbf{A}_i \vec{\xi}}{t} \\
&\quad + \frac{\gamma_i}{\zeta} \sum_{k=1,k\neq i}^{c} \lim_{t\to 0} \\
&\quad \left(\frac{1}{t} \frac{-2t(\vec{z}_i - \vec{z}_k)^\top \mathbf{A}_i \vec{\xi} - t^2 \vec{\xi}^\top \mathbf{A}_i \vec{\xi} - 2t(\vec{z}_i - \vec{z}_k)^\top \mathbf{A}_k \vec{\xi} - t^2 \vec{\xi}^\top \mathbf{A}_k \vec{\xi}}{\frac{1}{2}(||(\vec{z}_i - \vec{z}_k) + t\vec{\xi}||_{\mathbf{A}_i}^2 + ||(\vec{z}_i - \vec{z}_k) + t\vec{\xi})||_{\mathbf{A}_k}^2)} \right. \\
&\quad \left. \cdot \frac{1}{(||\vec{z}_i - \vec{z}_k||_{\mathbf{A}_i}^2 + ||\vec{z}_i - \vec{z}_k||_{\mathbf{A}_k}^2)} \right) \\
&= -2 \sum_{j=1}^{n} u_{i,j}^{m} (\vec{x}_j - \vec{z}_i)^\top \vec{\xi} \\
&\quad + \frac{\gamma_i}{\zeta} \frac{2(\vec{z}_k - \vec{z}_i)^\top \mathbf{A}_i \vec{\xi} + 2(\vec{z}_k - \vec{z}_i)^\top \mathbf{A}_k \vec{\xi}}{\frac{1}{2}\left(||\vec{z}_i - \vec{z}_k||_{\mathbf{A}_i}^2 + ||\vec{z}_i - \vec{z}_k||_{\mathbf{A}_k}^2 \right) \cdot \left(||\vec{z}_i - \vec{z}_k||_{\mathbf{A}_i}^2 + ||\vec{z}_i - \vec{z}_k||_{\mathbf{A}_k}^2 \right)} \\
&\Rightarrow \\
0 &= \sum_{j=1}^{n} u_{i,j}^{m} (\vec{x}_j - \vec{z}_i)^\top \mathbb{1}
\end{aligned}
$$

$$-\frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c}\frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\frac{1}{2}\left((\vec{z}_k-\vec{z}_i)^\top\mathbf{A}_i+(\vec{z}_k-\vec{z}_i)^\top\mathbf{A}_k\right)$$

Analytisch ist das Gleichungssystem (3.10) nicht allgemein für beliebige $p \in \mathbb{N}_{>0}$ lösbar. Alternativ bietet sich eine iterative Berechnung der Clusterzentren durch

$$\begin{aligned}\vec{z}_i &= \left(\sum_{j=1}^{n}u_{i,j}^m\,\mathbb{I}-\frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c}\frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\frac{1}{2}\left(\mathbf{A}_i^\top+\mathbf{A}_k^\top\right)\right)^{-1}\\ &\quad\cdot\left(\sum_{j=1}^{n}u_{i,j}^m\vec{x}_j^\top-\frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c}\frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\frac{1}{2}\left(\vec{z}_k^\top\mathbf{A}_i+\vec{z}_k^\top\mathbf{A}_k\right)\right)^\top \quad (3.11)\end{aligned}$$

an. Die Berechnung wird solange wiederholt, bis $||\vec{z}_i^{(\text{new})}-\vec{z}_i^{(\text{old})}||<\epsilon$ gilt.

Analog zu den Betrachtungen bei der Variante des Fuzzy-C-Means-Algorithmus kann in (3.11) der Ausdruck $\frac{1}{d^4(\vec{z}_k,\vec{z}_i)}$ als Zugehörigkeitsgrad bzw. Einflußgrad des Clusters $\vec{\beta}_k$ zu dem Cluster $\vec{\beta}_i$ analog zu dem Zugehörigkeitsgrad $u_{i,j}^m$ gedeutet werden. Das neue Zentrum wird als das mit den Zugehörigkeitsgraden der Daten und der (negativen) Einflußgrade der Cluster gewichtete Mittel der Daten und der Clusterzentren berechnet. Eine iterative Berechnung ist bei dieser Berechnung nicht erforderlich. Nach Konvergenz des Verfahrens gilt $\vec{z}_i^{(neu)}\approx\vec{z}_i^{(alt)}$. Das notwendige Kriterium für eine optimale Lösung der Zielfunktion ist erfüllt.

(3.11) erhält man durch Umformung von (3.10):

$$\begin{aligned}&\sum_{j=1}^{n}u_{i,j}^m\vec{z}_i^\top\mathbb{I}-\frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c}\frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\frac{1}{2}\left(\vec{z}_i^\top\mathbf{A}_i+\vec{z}_i^\top\mathbf{A}_k\right)\\ =\;&\sum_{j=1}^{n}u_{i,j}^m\vec{x}_j^\top\mathbb{I}-\frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c}\frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\frac{1}{2}\left(\vec{z}_k^\top\mathbf{A}_i+\vec{z}_k^\top\mathbf{A}_k\right)\\ \Rightarrow\quad&\left(\left(\sum_{j=1}^{n}u_{i,j}^m\mathbb{I}-\gamma_i\sum_{k=1,k\neq i}^{c}\frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\frac{1}{2}\left(\mathbf{A}_i^\top+\mathbf{A}_k^\top\right)\right)\vec{z}_i\right)^\top\\ =\;&\sum_{j=1}^{n}u_{i,j}^m\vec{x}_j^\top\mathbb{I}-\gamma_i\sum_{k=1,k\neq i}^{c}\frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\frac{1}{2}\left(\vec{z}_k^\top\mathbf{A}_i+\vec{z}_k^\top\mathbf{A}_k\right)\end{aligned}$$

$$
\begin{aligned}
&\Rightarrow \\
&\left(\sum_{j=1}^{n} u_{i,j}^{m}\,\mathbb{I} - \gamma_i \sum_{k=1,k\neq i}^{c} \frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\frac{1}{2}\left(\mathbf{A}_i^{\top}+\mathbf{A}_k^{\top}\right)\right)\vec{z}_i \\
&= \left(\sum_{j=1}^{n} u_{i,j}^{m}\vec{x}_j^{\top}\,\mathbb{I} - \frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c} \frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\frac{1}{2}\left(\vec{z}_k^{\top}\mathbf{A}_i+\vec{z}_k^{\top}\mathbf{A}_k\right)\right)^{\top} \\
&\Rightarrow \\
\vec{z}_i &= \left(\sum_{j=1}^{n} u_{i,j}^{m}\,\mathbb{I} - \frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c} \frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\frac{1}{2}\left(\mathbf{A}_i^{\top}+\mathbf{A}_k^{\top}\right)\right)^{-1} \\
&\quad\cdot\left(\sum_{j=1}^{n} u_{i,j}^{m}\vec{x}_j^{\top}\,\mathbb{I} - \frac{\gamma_i}{\zeta}\sum_{k=1,k\neq i}^{c} \frac{1}{d^4(\vec{z}_k,\vec{z}_i)}\frac{1}{2}\left(\vec{z}_k^{\top}\mathbf{A}_i+\vec{z}_k^{\top}\mathbf{A}_k\right)\right)^{\top}
\end{aligned}
$$

Die Ableitung der Zielfunktion (3.3) nach der Normmatrix $\mathbf{A}_i$ unter Berücksichtigung der Restriktion (2.3) führt zu :

$$\mathbf{A}_i = \sqrt[p]{\det(\mathbf{S}_i)}\,\mathbf{S}_i^{-1}. \tag{3.12}$$

Dabei ist

$$\mathbf{S}_i = \sum_{j=1}^{n} u_{i,j}^{m}(\vec{x}_j-\vec{z}_i)(\vec{x}_j-\vec{z}_i)^{\top} - \frac{\gamma_i}{\zeta}\sum_{k=1}^{c}(\vec{z}_k-\vec{z}_i)(\vec{z}_k-\vec{z}_i)^{\top}\frac{1}{2d^4(\vec{\beta}_i,\vec{\beta}_k)}. \tag{3.13}$$

Ebenso wie der Ausdruck zur Berechnung des Clusterzentrums (3.11) stellen auch die Ausdrücke zur Berechnung von $\mathbf{A}_i$ (3.12) und $\mathbf{S}_i$ (3.13) keine analytische Lösung dar, da der Abstand zwischen den Clustern $d^2(\vec{\beta}_i,\vec{\beta}_k)$ auf $\mathbf{A}_i$ basiert. Im folgenden werden (3.12) und (3.13) hergeleitet.

Bei dem Gustafson–Kessel-Algorithmus besitzt jeder Cluster die gleiche Größe. Hierfür wird die Determinante der Normmatrix $\mathbf{A}_i$ auf 1 normiert. Unter Verwendung von c Lagrangeschen Multiplikatoren λ_i wird daher die Zielfunktion modifiziert zu

$$
\begin{aligned}
J_1(\mathbf{X},\mathbf{U},\mathbf{B}) &= \sum_{i=1}^{c}\sum_{j=1}^{n} u_{i,j}^{m} d^2(\vec{\beta}_i,\vec{x}_j) + \sum_{i=1}^{c}\eta_i\sum_{j=1}^{n}(1-u_{i,j})^m \\
&\quad + \sum_{i=1}^{c}\gamma_i\sum_{k=1,k\neq i}^{c}\frac{1}{\zeta d^2(\vec{\beta}_i,\vec{\beta}_k)} - \sum_{i=1}^{c}\lambda_i(\det(\mathbf{A})_i - 1)
\end{aligned} \tag{3.14}
$$

Es gilt $\nabla \vec{x}_j^\top \mathbf{A}_i \vec{x}_j = \vec{x}_j \vec{x}_j^\top$ und $\nabla \det(\mathbf{A}_i) = \det(\mathbf{A}_i)\mathbf{A}_i^{-1}$ [65]. Aus der Minimalität folgt das Verschwinden des Gradienten. Es gilt:

$$
\begin{aligned}
0 &= \nabla J_1(\mathbf{X}, \mathbf{U}, \mathbf{B}) \\
&= \nabla \left(\sum_{i=1}^{c} \sum_{j=1}^{n} u_{i,j}^m (\vec{x}_j - \vec{z}_i)^\top \mathbf{A}_i (\vec{x}_j - \vec{z}_i) + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} (1 - u_{i,j})^m \right. \\
&\quad + \sum_{i=1}^{c} \frac{\gamma_i}{\zeta} \sum_{k=1, k\neq i}^{c} \frac{1}{\frac{1}{2}\left((\vec{z}_i - \vec{z_k})^\top \mathbf{A}_i (\vec{z}_i - \vec{z}_k) + (\vec{z}_i - \vec{z_k})^\top \mathbf{A}_k (\vec{z}_i - \vec{z}_k)\right)} \\
&\quad \left. - \sum_{i=1}^{c} \lambda_i (\det(\mathbf{A})_i - 1) \right) \\
&= \sum_{j=1}^{n} u_{i,j}^m (\vec{x}_j - \vec{z}_i)(\vec{x}_j - \vec{z}_i)^\top \\
&\quad + \frac{\gamma_i}{\zeta} \sum_{k=1}^{c} (-1) \frac{1}{2} (\vec{z}_i - \vec{z}_k)(\vec{z}_i - \vec{z}_k)^\top \\
&\quad \cdot \frac{1}{\left(\frac{1}{2}\left((\vec{z}_i - \vec{z_k})^\top \mathbf{A}_i (\vec{z}_i - \vec{z}_k) + (\vec{z}_i - \vec{z_k})^\top \mathbf{A}_k (\vec{z}_i - \vec{z}_k)\right)\right)^2} \\
&\quad - \lambda_i \det(\mathbf{A}_i) \mathbf{A}_i^{-1} \\
&= \sum_{j=1}^{n} u_{i,j}^m (\vec{x}_j - \vec{z}_i)(\vec{x}_j - \vec{z}_i)^\top \\
&\quad + \frac{\gamma_i}{\zeta} \sum_{k=1}^{c} (-1) \frac{1}{2} (\vec{z}_i - \vec{z}_k)(\vec{z}_i - \vec{z}_k)^\top \frac{1}{d^2(\vec{\beta}_i, \vec{\beta_k}) d^2(\vec{\beta}_i, \vec{\beta_k})} \\
&\quad - \lambda_i \det(\mathbf{A}_i) \mathbf{A}_i^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&\Rightarrow \\
\lambda_i \det(\mathbf{A}_i) \mathbf{A}_i^{-1} &= \sum_{j=1}^{n} u_{i,j}^m (\vec{x}_j - \vec{z}_i)(\vec{x}_j - \vec{z}_i)^\top \\
&\quad + \frac{\gamma_i}{\zeta} \sum_{k=1}^{c} (-1) \frac{1}{2} (\vec{z}_i - \vec{z}_k)(\vec{z}_i - \vec{z}_k)^\top \\
&\quad \cdot \frac{1}{d^2(\vec{\beta}_i, \vec{\beta_k}) d^2(\vec{\beta}_i, \vec{\beta_k})}
\end{aligned}
$$

Da $\det(\mathbf{A}_i) = 1$, gilt mit (3.13)

$$\begin{aligned} & \mathbf{S}_i = \lambda_i \mathbf{A}_i^{-1} \\ \Leftrightarrow \quad & \mathbf{S}_i \mathbf{A}_i = \lambda_i \, \mathbb{I} \\ \Rightarrow \quad & \det(\mathbf{S}_i \mathbf{A}_i) = \lambda_i^p \\ \Leftrightarrow \quad & \lambda_i = \sqrt[p]{\det(\mathbf{S}_i)\det(\mathbf{A}_i)} = \sqrt[p]{\det(\mathbf{S}_i)} \\ \Rightarrow \quad & \mathbf{A}_i = \sqrt[p]{\det(\mathbf{S}_i)}\mathbf{S}_i^{-1} \end{aligned}$$

Analog kann man zeigen, daß die Ableitung der Zielfunktion (3.4) nach den Clusterzentren $\vec{z}_i$ zu

$$\sum_{j=1}^{n} u_{i,j}^m (\vec{x}_j - \vec{z}_i) - \gamma_i \zeta \sum_{k=1, k\neq i}^{c} e^{-\zeta d^2(\vec{\beta}_i, \vec{\beta}_k)} \frac{1}{2} \left((\vec{z}_k - \vec{z}_i)^\top \mathbf{A}_i + (\vec{z}_k - \vec{z}_i)^\top \mathbf{A}_k \right) = 0$$

führt.

Durch Umformung erhält man:

$$\begin{aligned} \vec{z}_i \quad = \quad & \left(\sum_{j=1}^{n} u_{i,j}^m \, \mathbb{I} - \gamma_i \zeta \sum_{k=1, k\neq i}^{c} e^{-d^2(\vec{\beta}_i, \vec{\beta}_k)} \frac{1}{2} \left(\mathbf{A}_i^\top + \mathbf{A}_k^\top \right) \right)^{-1} \\ & \cdot \left(\sum_{j=1}^{n} u_{i,j}^m \vec{x}_j - \gamma_i \zeta \sum_{k=1, k\neq i}^{c} e^{-d^2(\vec{\beta}_i, \vec{\beta}_k)} \frac{1}{2} \left(\vec{z}_k^\top \mathbf{A}_i + \vec{z}_k^\top \mathbf{A}_k \right) \right)^\top \end{aligned} \tag{3.15}$$

Die Ableitung der Zielfunktion (3.4) nach den Normmatrizen $\mathbf{A}_i$ führt zu

$$\mathbf{A}_i = \sqrt[p]{\det(\mathbf{S}_i)}\mathbf{S}_i^{-1}. \tag{3.16}$$

Hierbei ist

$$\mathbf{S}_i = \sum_{j=1}^{n} u_{i,j}^m (\vec{x}_j - \vec{z}_i)(\vec{x}_j - \vec{z}_i)^\top - \gamma_i \zeta \sum_{k=1}^{c} (\vec{z}_k - \vec{z}_i)(\vec{z}_k - \vec{z}_i)^\top e^{-\zeta d^2(\vec{\beta}_i, \vec{\beta}_k)}. \tag{3.17}$$

Die Ausdrücke zur Berechnung der Clusterzentren und der Normmatrizen veranschaulichen den Effekt der Idee des Verfahrens. Wenn zwei Cluster nah benachbart sind, stoßen sie sich durch die Berechnung der Zentren gegenseitig ab. Gleichzeitig wird die Form der Cluster etwas modifiziert. Bei

weit entfernten Clustern werden demgegenüber die Clusterzentren und die Normmatrix ohne (wesentliche) Beeinflussung durch andere Cluster berechnet.

Falls bei einem Cluster $\vec{\beta}_i$ die abstoßende Wirkung der anderen Cluster die anziehende Wirkung durch die Daten übersteigt, bietet es sich an, den Cluster in ein anderes „Gebiet“ des Datensatzes zu „legen“, in dem weniger Cluster sind (vgl. die diesbezüglichen Betrachtungen für die Variante des Fuzzy-C-Means-Algorithmus). Als neue Form des Clusters sollte eine Hyperkugel gewählt werden. D.h. die Normmatrix $\mathbf{A}_i$ ist die Einheitsmatrix.

3.3.3 Bestimmung des Parameters γ_i

Der Parameter $\gamma_i \in \mathbb{R}_{>0}$ setzt die Abstoßung des Clusters $\vec{\beta}_i$ durch benachbarte Cluster in Relation zu der Anziehung durch Daten. Bei $\gamma_i = 0$, $i = 1 \ldots c$ liegt das in Kapitel 2.7 vorgestellte possibilistische Fuzzy-Clusteringverfahren vor. Je größer γ_i gewählt wird, desto stärker wird die Abstoßung durch benachbarte Cluster gewichtet. Durch die Wahl der Abstoßungsfunktion ist jedoch sichergestellt, daß eine Abstoßung nur durch *benachbarte* Cluster und nicht durch alle Cluster stattfindet. Mit zunehmendem Abstand der Cluster von einander verliert der Parameter γ_i durch die Modellierung der Abstoßung bzw. des Einflußgrads benachbarter Cluster stark an Bedeutung. Der Abstoßungseffekt geht mit zunehmender Entfernung der Cluster voneinander gegen 0.

Neben der Gewichtung der Anziehung durch die Daten zu der Abstoßung durch andere Cluster kann der Parameter γ_i als Korrekturfaktor für die Anzahl der Daten bzw. der Summe ihrer Zugehörigkeitsgrade dienen. Wenn man z.B. den Ausdruck zur Berechnung der Cluster

$$\vec{z}_i = \frac{\sum_{j=1}^{n} u_{i,j}^m \vec{x}_j - \gamma\zeta \sum_{k=1,k\neq i}^{c} \vec{z}_k e^{-\zeta d^2(\vec{z}_k,\vec{z}_i)}}{\sum_{j=1}^{n} u_{i,j}^m - \gamma\zeta \sum_{k=1,k\neq i}^{c} e^{-\zeta d^2(\vec{z}_k,\vec{z}_i)}}$$

mit $\gamma_i = \gamma$ betrachtet, sieht man, daß bei „großen“ Clustern die Abstoßung geringer ist als bei „kleinen“ Clustern. Durch eine clusterspezifische Abstoßung $\gamma_i = \gamma \sum_{j=1}^{n} u_{i,j}^m$ kann diese Problematik vermieden werden.

Bei der Wahl von γ_i ist sicherzustellen, daß die Ausdrücke zur Berechnung der Clusterprototypen sinnvoll bleiben. Die Abstoßung durch einen Cluster kann als Aufhebung der Anziehung durch Daten aus dem Gebiet dieses Clusters interpretiert werden. Die Abstoßung durch Cluster sollte daher kleiner sein als die Anziehung durch Daten. Falls dies für einen Cluster

nicht erfüllt ist, bietet es sich an, diesen Cluster neu zu initialisieren. Anschaulich kann der Fall, daß die Abstoßung durch andere Cluster größer als die Anziehung durch Daten ist, so interpretiert werden, daß die betreffende Region im Datenraum durch Cluster „gut abgedeckt ist“.

Im folgenden wird der Parameter γ_i anhand der possibilistischen Variante des Fuzzy-C-Means-Algorithmus betrachtet. Die Abstoßung zwischen benachbarten Clustern wird durch

$$\gamma_i \sum_{k=1,k\neq i}^{c} \frac{1}{\zeta d^2(\vec{\beta}_i,\vec{\beta}_j)} \quad \text{bzw.} \quad \gamma_i \sum_{k=1,k\neq i}^{c} e^{\zeta - d^2(\vec{\beta}_i,\vec{\beta}_k)}$$

modelliert. Die Clusterzentren werden durch (3.6) bzw. (3.8) berechnet. Der Parameter ζ für die Normierung des Abstands wird auf 1 gesetzt. Entsprechend den vorhergehenden Betrachtungen wird γ_i durch $\gamma_i = \gamma \sum_{j=1}^{n} u_{i,j}^m$ berechnet. Der Parameter η sowie die Zugehörigkeitsgrade $u_{i,j}$ für die Berechnung von γ_i werden mit den Ergebnissen des vorher durchgeführten probabilistischen Fuzzy-C-Means-Algorithmus berechnet. Als Testdatensatz dienen die Irisdaten [48]. Zur Klassifikation wurden nur die Attribute „petal length“ und „petal width“ verwendet, da diese die meiste klassifikationsrelevante Information enthalten. Dies ermöglicht eine bessere Visualisierung der Wirkung von γ_i.

Abb. 3.3 zeigt die Klassifikation des Datensatzes durch den *probabilistischen* Fuzzy-C-Means-Algorithmus. Die Cluster haben die Farben blau, grün und rot. Mit einem abnehmenden Zugehörigkeitsgrad wird die Farbe dunkler. Die partitionierende Eigenschaft der probabilistischen Fuzzy-Clusteranalyse ist gut erkennbar.

Abb. 3.4 zeigt die Klassifikation mit dem *possibilistischen* Fuzzy-C-Means-Algorithmus. Da das Verfahren mit dem probabilistischen Fuzzy-C-Means-Algorithmus initialisiert wurde, werden zwei Cluster erkannt. Der grüne und der blaue Cluster der probabilistischen Fuzzy -Clusteranalyse werden durch *einen* Cluster beschrieben.

Für $\gamma = 0$ ist die in diesem Kapitel vorgestellte Variante der possibilistischen Fuzzy-Clusteranalyse mit der „üblichen“ possibilistischen Fuzzy-Clusteranalyse identisch. Eine abstoßende Wirkung zwischen Clustern tritt bei $\gamma > 0$ auf. Die Abbildungen 3.5, 3.6, 3.7, 3.8, 3.9, 3.10, 3.11, 3.12, 3.13 und 3.14 verdeutlichen die Abstoßung zwischen den Clustern in Abhängigkeit von γ für die Berechnung der Clusterzentren durch (3.6) bzw. (3.8). Mit einem größeren Wert γ werden der blaue und der grüne Cluster zunehmend separiert, während der rote Cluster unverändert bleibt. Die Ursache ist,

daß der Abstand zwischen den Zentren des roten Clusters und des grünen bzw. des blauen Clusters hinreichend groß ist, so daß der Abstoßungsgrad $\frac{1}{d^4(\vec{z}_k,\vec{z}_i)}$ bzw. $e^{-\zeta d^2(\vec{z}_k,\vec{z}_i)}$ sehr klein ist. Eine Multiplikation dieses Wertes mit γ_i führt zu keiner relevanten Veränderung bei der Berechnung der Zugehörigkeitsgrade. Demgegenüber ist der Abstand zwischen den Zentren des blauen und des grünen Clusters gering. Die Multiplikation des Abstoßungsgrads mit γ_i spiegelt sich daher bei der Berechnung der Clusterprototypen wider.

Sofern der Abstand bei der Bestimmung des Abstoßungsgrads so normiert wird, daß $\frac{1}{d^4(\vec{z}_k,\vec{z}_i)}$ bzw. $e^{-\zeta d^2(\vec{z}_k,\vec{z}_i)}$ nur bei fast identischen Clustern relevante Werte annimmt, kann das Verfahren hinsichtlich der Wahl von γ als ziemlich robust angesehen werden.

3.4 Ein weiterer Ansatz, basierend auf dem Alternating Cluster Estimation

In den vorhergehenden Abschnitten dieses Kapitels 3 wurde die zielfunktionsbasierte possibilistische Fuzzy-Clusteranalyse betrachtet. Eine Erweiterung der zielfunktionsbasierten Fuzzy-Clusteranalyse ist das „Alternating Cluster Estimation“. Bei diesem Ansatz werden die Ausdrücke zur Berechnung der Clusterprototypen bzw. der Zugehörigkeitsgrade „geschätzt“. Sie müssen keine notwendigen Kriterien zur Optimierung der Zielfunktion sein. Bei der possibilistischen Version des „Alternating Cluster Estimation“ werden possibilistische Zugehörigkeitsgrade verwendet. Die Berechnung der Clusterprototypen erfolgt durch entsprechend motivierte Ausdrücke. Eine formale Herleitung ist nicht erforderlich.

Bei dem in den vorhergehenden Abschnitten vorgestellten Ansatz werden identische Cluster verhindert, indem bei der Zielfunktion die Abstände zwischen den Clustern berücksichtigt werden. Bei einer Modellierung durch $\sum_{k=1,k\neq i}^{c} \frac{1}{\zeta d^2(\vec{\beta}_i,\vec{\beta}_j)}$ führt dies bei dem Fuzzy-C-Means-Algorithmus für die Berechnung der Clusterzentren zu

$$\vec{z}_i = \frac{\sum_{j=1}^{n} u_{i,j}^m \vec{x}_j - \frac{\gamma_i}{\zeta} \sum_{k=1,k\neq i}^{c} \frac{1}{d^4(\vec{z}_k,\vec{z}_i)} \vec{z}_k}{\sum_{j=1}^{n} u_{i,j}^m - \frac{\gamma_i}{\zeta} \sum_{k=1,k\neq i}^{c} \frac{1}{d^4(\vec{z}_k,\vec{z}_i)}}.$$

Dieser Ausdruck kann so interpretiert werden, daß ein Teil der Anziehung durch die Daten durch die Abstoßung der Clusterzentren benachbarter Cluster aufgehoben wird. Dieser Ausdruck kann auch bei der possibilistischen

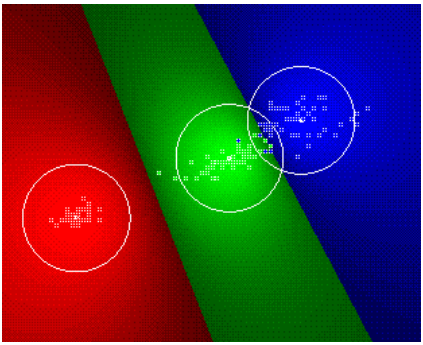

Abbildung 3.3: Clusteranalyse des Irisdatensatzes mit dem *probabilistischen* Fuzzy-C-Means-Algorithmus. Attribute „petal length“ und „petal width“.

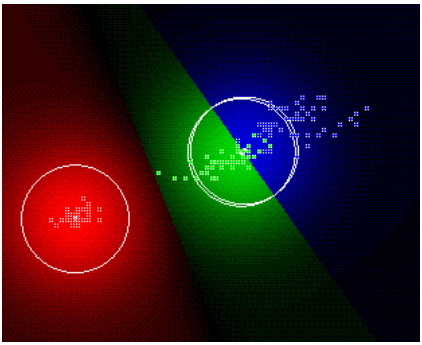

Abbildung 3.4: Clusteranalyse des Irisdatensatzes mit dem *possibilistischen* Fuzzy-C-Means-Algorithmus. Attribute „petal length“ und „petal width“.

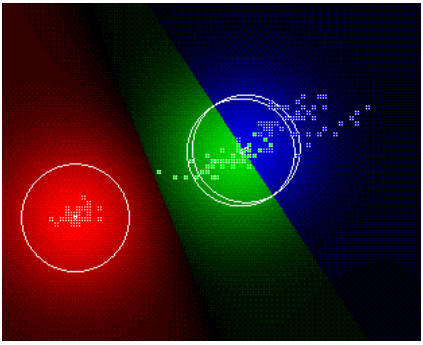

Abbildung 3.5: Clusteranalyse des Irisdatensatzes mit dem in diesem Abschnitt vorgestellten Ansatz basierend auf der *Zielfunktion (3.3)*. $\gamma = 0.00001$, Attribute „petal length“ und „petal width“.

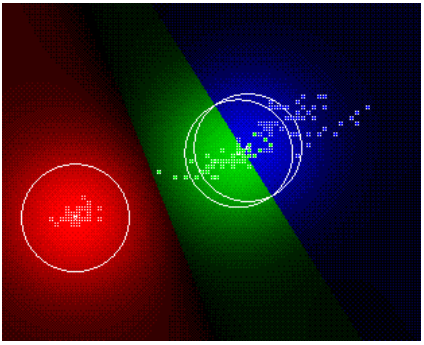

Abbildung 3.6: Clusteranalyse des Irisdatensatzes mit dem in diesem Abschnitt vorgestellten Ansatz basierend auf der *Zielfunktion (3.3)*. $\gamma = 0.0001$, Attribute „petal length“ und „petal width“.

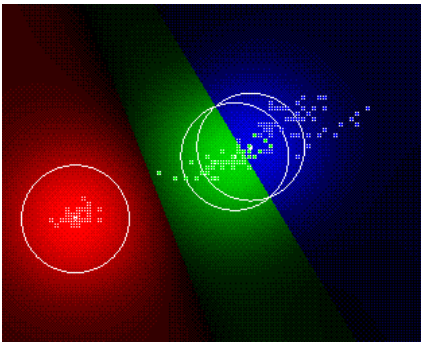

Abbildung 3.7: Clusteranalyse des Irisdatensatzes mit dem in diesem Abschnitt vorgestellten Ansatz basierend auf der *Zielfunktion (3.3)*. $\gamma = 0.001$, Attribute „petal length“ und „petal width“.

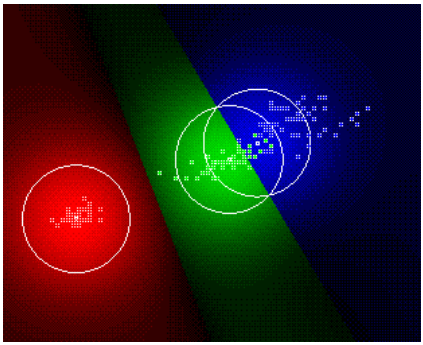

Abbildung 3.8: Clusteranalyse des Irisdatensatzes mit dem in diesem Abschnitt vorgestellten Ansatz basierend auf der *Zielfunktion (3.3)*. $\gamma = 0.01$, Attribute „petal length“ und „petal width“.

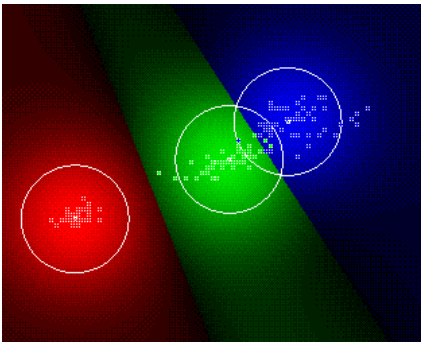

Abbildung 3.9: Clusteranalyse des Irisdatensatzes mit dem in diesem Abschnitt vorgestellten Ansatz basierend auf der *Zielfunktion (3.3)*. $\gamma = 0.1$, Attribute „petal length“ und „petal width“.

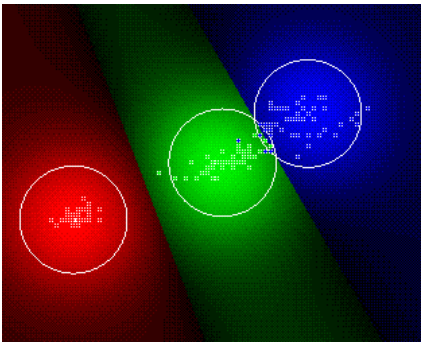

Abbildung 3.10: Clusteranalyse des Irisdatensatzes mit dem in diesem Abschnitt vorgestellten Ansatz basierend auf der *Zielfunktion (3.3)*. $\gamma = 0.5$, Attribute „petal length“ und „petal width“.

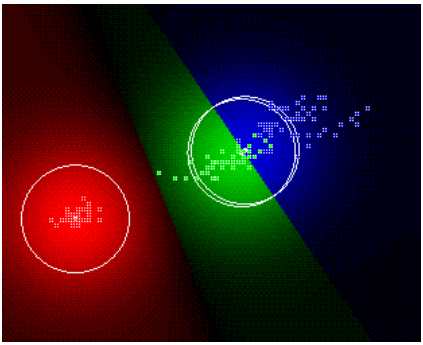

Abbildung 3.11: Clusteranalyse des Irisdatensatzes mit dem in diesem Abschnitt vorgestellten Ansatz basierend auf der *Zielfunktion (3.4)*. $\gamma = 0.01$, Attribute „petal length“ und „petal width“.

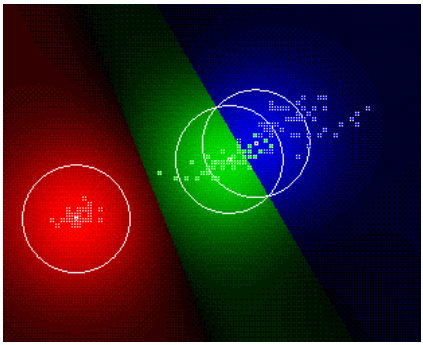

Abbildung 3.12: Clusteranalyse des Irisdatensatzes mit dem in diesem Abschnitt vorgestellten Ansatz basierend auf der *Zielfunktion (3.4)*. $\gamma = 0.1$, Attribute „petal length“ und „petal width“.

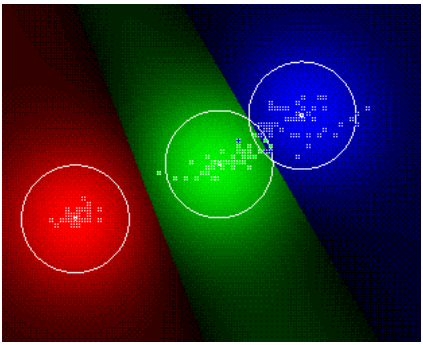

Abbildung 3.13: Clusteranalyse des Irisdatensatzes mit dem in diesem Abschnitt vorgestellten Ansatz basierend auf der *Zielfunktion (3.4)*. $\gamma = 1$, Attribute „petal length“ und „petal width“.

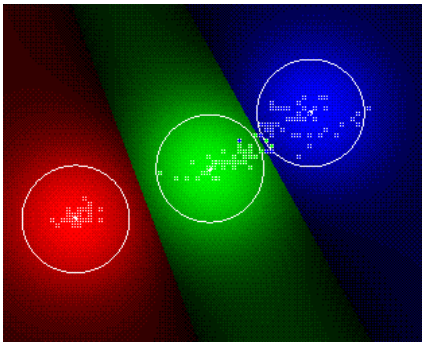

Abbildung 3.14: Clusteranalyse des Irisdatensatzes mit dem in diesem Abschnitt vorgestellten Ansatz basierend auf der *Zielfunktion (3.4)*. $\gamma = 4$, Attribute „petal length“ und „petal width“.

Version des „Alternating Cluster Estimation“ verwendet werden. Etwas anschaulicher als die Abstoßung durch „Aufhebung der Anziehung durch Daten“ bei diesem Ausdruck ist jedoch die Abstoßung durch Anziehung aus der entgegengesetzten Richtung. Diese Idee wird im folgenden näher betrachtet.

Die Idee ist, die Abstoßung durch ein Clusterzentrum durch die Anziehung durch einen fiktiven Punkt zu ersetzen. Die Lage des Punktes erhält man durch Spiegelung des abstoßenden Clusterzentrums $\vec{z}_k$ an dem Clusterzentrum $\vec{z}_i$, auf das die abstoßende Kraft wirkt. Um die Stärke der Abstoßung und die Richtung der abstoßenden Kraft getrennt zu modellieren, soll der fiktive Punkt von dem Clusterzentrum $\vec{z}_i$ einen Abstand von 1 haben. Der fiktive Punkt wird berechnet durch $\left(\vec{z}_i - \frac{\vec{z}_k - \vec{z}_i}{||\vec{z}_k - \vec{z}_i||}\right)$. Dieser Punkt wird mit dem Abstoßungsgrad bzw. Einflußgrad des Clusters $\vec{\beta}_k$ auf den Cluster $\vec{\beta}_i$ gewichtet. Mit dieser Modellierung werden die Clusterzentren durch

$$\vec{z}_i = \frac{\sum_{j=1}^{n} u_{i,j}^{m} \vec{x}_j + \gamma_i \sum_{k=1,k\neq i}^{c} u_{i,k}^{(\vec{\beta}_i)} \left(\vec{z}_i - \frac{\vec{z}_k - \vec{z}_i}{||\vec{z}_k - \vec{z}_i||}\right)}{\sum_{j=1}^{n} u_{i,j}^{m} + \gamma_i \sum_{k=1,k\neq i}^{c} u_{i,k}^{(\vec{\beta}_i)}}$$

berechnet. γ_i gewichtet die Abstoßung durch die anderen Cluster zu der Anziehung durch die dem Cluster zugeordneten Daten. Die Abstoßung durch die Cluster wird damit durch

$$\gamma_i \sum_{k=1,k\neq i}^{c} u_{i,k}^{(\vec{\beta}_i)} \left(\vec{z}_i - \frac{\vec{z}_k - \vec{z}_i}{||\vec{z}_k - \vec{z}_i||}\right)$$

modelliert.

Der Abstoßungsgrad $u_{i,k}^{(\vec{\beta}_i)}$ des Clusters $\vec{\beta}_k$ auf den Cluster $\vec{\beta}_i$ kann analog zu den Betrachtungen der vorhergehenden Abschnitte z.B. als

$$u_{i,k}^{(\vec{\beta}_i)} = \frac{1}{\zeta d^2(\vec{z}_k, \vec{z}_i)}, \tag{3.18}$$

oder

$$u_{i,k}^{(\vec{\beta}_i)} = e^{-\zeta d^2(\vec{z}_k, \vec{z}_i)} \tag{3.19}$$

definiert werden.[3] Der Parameter γ_i ist in Abhängigkeit von der Modellierung des Einflußgrades $u_{i,k}^{(\vec{\beta}_i)}$ und ggf. der Summe der Zugehörigkeitsgrade der dem Cluster zugeordneten Daten zu wählen.

[3] Die Modellierung der Abstoßung durch $\frac{1}{d^2(\vec{\beta}_i, \vec{\beta}_j)}$ in der Zielfunktion führt bei der Berechnung der Zugehörigkeitsgrade zu einem Abstoßungsgrad von $\frac{1}{d^4(\vec{z}_k, \vec{z}_i)}$, vgl. 3.6. Die abstoßende Wirkung der Clusterzentren nimmt daher mit zunehmendem Abstand stärker ab.

Für die Berechnung weiterer Parameter der Clusterprototypen bei anderen Fuzzy-Clusteringverfahren, wie z.B. der Kovarianzmatrix bzw. Normmatrix bei dem Gustafson–Kessel-Algorithmus, sind analoge Überlegungen hinsichtlich einer „korrekten“ bzw. „richtigen“ Berechnung der Parameter erforderlich.[4] Die Kovarianzmatrix kann z.B. berechnet werden, indem sie in Richtung nah benachbarter Cluster gestaucht wird, so daß die Cluster sich weniger stark überlappen. Der Nachteil ist jedoch, daß hierdurch die Form der Cluster nicht mehr allein von den dem Cluster zugeordneten Daten abhängt. Bei dicht benachbarten Clustern hebt die Abstoßung durch die Cluster den Einfluß der Daten aus dem Gebiet dieses Clusters zumindest teilweise auf. In dem überlappenden Bereich der beiden Cluster kann dies als eine Partitionierung der Daten aufgefaßt werden.

Alternativ kann die Berechnung der Kovarianzmatrix auch ohne Berücksichtigung benachbarter Cluster erfolgen. Die Form der Cluster basiert nur auf den dem Cluster zugeordneten Daten. Hierdurch werden auch bei nah benachbarten Clustern die Cluster möglichst präzise beschrieben. Durch die Abstoßung benachbarter Clusterzentren wird dabei sichergestellt, daß die Cluster nicht identisch sind. Diese Vorgehensweise wird bei den Beispielen im folgenden Abschnitt verwendet.

3.5 Beispiele

Die in diesem Kapitel vorgestellten possibilistischen Varianten des Fuzzy-C-Means-Algorithmus und des Gustafson–Kessel-Algorithmus werden im folgenden anhand des Irisdatensatzes [48] und des Weindatensatzes [1] näher betrachtet. Der Irisdatensatz wird mit dem Fuzzy-C-Means-Algorithmus und der Weindatensatz mit dem Gustafson–Kessel-Algorithmus klassifiziert.

Der Irisdatensatz [48] besteht aus den drei Klassen Iris Setosa, Iris Versicolour und Iris Virginica mit je 50 Daten. Die Daten enthalten vier Attribute, „sepal length“, „sepal width“„petal length“ und „petal width“. Für die Clusteranalyse wurden alle Attribute verwendet.

Bei der Clusteranalyse mit dem probabilistischen Fuzzy-C-Means-Algorithmus werden 15 Daten falsch klassifiziert. Abb. 3.15 zeigt die Projektion der Cluster auf die Attribute „petal length“ und „petal width“. Die Stärke des Zugehörigkeitsgrads zu den Clustern wird durch die Färbung angedeutet. Eine dunklere Färbung deutet auf einen geringeren Zugehörig-

[4] Bei dem Alternating Cluster Estimation sind die Ausdrücke zur Berechnung der Clusterprototypen und Zugehörigkeitsgrade nicht zwingend notwendige Kriterien zur Optimierung der Zielfunktion.

keitsgrad hin. Neben den Daten werden die Clusterzentren und die kreisförmige Form der Cluster angezeigt. Die partitionierende Eigenschaft der probabilistischen Fuzzy-Clusteranalyse ist gut erkennbar. Abb. 3.17 zeigt die Zugehörigkeitsgrade der Daten zu den Clustern an. Daten, die der gleichen Klasse angehören, sind benachbart. Eine scharfe Klassifikation ohne fehlklassifizierte Daten würde drei Rechtecke aufweisen. Bei der Clusteranalyse mit dem possibilistischen Fuzzy-C-Means-Algorithmus werden nur zwei verschiedene Cluster erkannt. Zwei Cluster sind identisch. Es wurden zwei Cluster gefunden, da die probabilistische Fuzzy-Clusteranalyse mit den Ergebnissen der probabilistischen Fuzzy-Clusteranalyse initialisiert wird. (Die gefundene Lösung ist also nur ein lokales Optimum.) Abb. 3.18 zeigt die Zugehörigkeitsgrade der Daten zu den Clustern und Abb. 3.16 zeigt die Projektion der Cluster auf die Attribute „petal length“ und „petal width“. Der Nachteil der possibilistischen Fuzzy-Clusteranalyse, bei der Berechnung der Cluster die anderen Cluster weder direkt noch indirekt zu berücksichtigen, ist offensichtlich. Die Abbildungen 3.19, 3.20, 3.23, 3.24, 3.21, 3.22, 3.25 und 3.26 zeigen die Ergebnisse der Clusteranalyse mit den verschiedenen in diesem Kapitel vorgestellten possibilistischen Varianten. Der Parameter ζ wurde auf 1 gesetzt. Die Anzahl der fehlklassifizierten Daten ist bei jeder Variante 11. Die Klassifikationsgüte hinsichtlich der Anzahl der fehlklassifizierten Daten ist zumindest mit der des probabilistischen Ansatzes vergleichbar. Die Zugehörigkeitsgrade der Daten zu den Clustern verdeutlichen, daß die Cluster klar erkannt werden. Im Gegensatz zu der probabilistischen Clusteranalyse geben die Zugehörigkeitsgrade an, wie typisch ein Datum für einen Cluster ist. Die Projektionen der Cluster auf die Attribute „petal length“ und „petal width“ verdeutlichen den Unterschied. Während bei der probabilistischen Fuzzy-Clusteranalyse der gesamte Datenraum zwischen den Clustern partitioniert ist, entspricht bei der possibilistischen Fuzzy-Clusteranalyse der Zugehörigkeitsgrad der Form der Cluster.

Der Weindatensatz [1] besteht aus drei Klassen mit 59, 71 und 48 Daten. Die Daten sind das Resultat einer chemischen Analyse von Weinen aus der gleichen Region. Bei der Analyse wurden 13 Bestandteile der drei verschiedenen Weintypen untersucht. Für die Fuzzy-Clusteranalyse wurden von den 13 Attributen die Attribute 7, 10 und 13 verwendet. Die Abbildungen 3.27, 3.28 und 3.29 zeigen den Datensatz hinsichtlich der drei betrachteten Attribute. Der Datensatz wurde hinsichtlich jedes Attributs auf den Wertebereich $[0, 10]$ skaliert.

Bei der probabilistischen Fuzzy-Clusteranalyse wird der Datensatz mit 7 Fehlern klassifiziert. Abb. 3.30 zeigt die Projektion des Ergebnisses auf die Attribute 7 und 10. Bei der possibilistischen Fuzzy-Clusteranalyse des

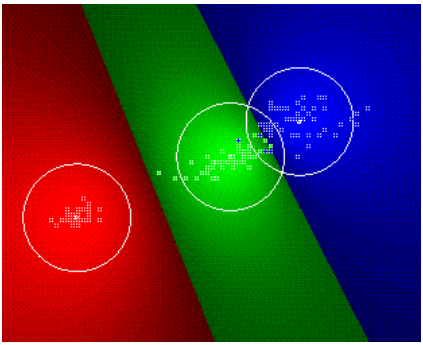

Abbildung 3.15: Clusteranalyse des Irisdatensatzes mit dem *probabilistischen* Fuzzy-C-Means-Algorithmus. Attribute „petal length“ und „petal width“.

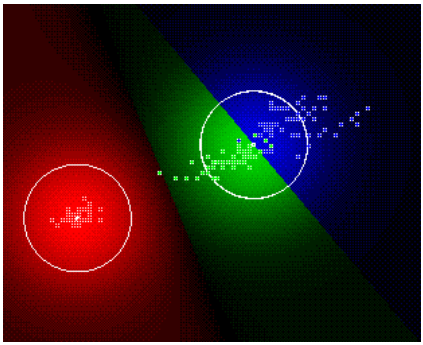

Abbildung 3.16: Clusteranalyse des Irisdatensatzes mit dem *possibilistischen* Fuzzy-C-Means-Algorithmus. Attribute „petal length“ und „petal width“.

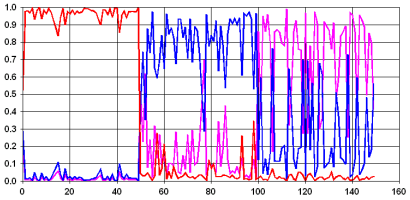

Abbildung 3.17: Zugehörigkeitsgrade der Daten zu den Clustern bei der Clusteranalyse des Irisdatensatzes mit dem *probabilistischen* Fuzzy-C-Means-Algorithmus. Daten, die zu dem gleichen Cluster gehören, sind benachbart.

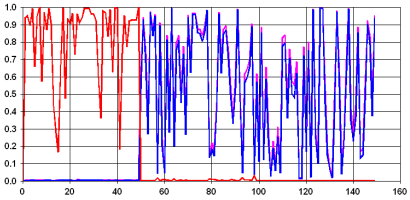

Abbildung 3.18: Zugehörigkeitsgrade der Daten zu den Clustern bei der Clusteranalyse des Irisdatensatzes mit dem *possibilistischen* Fuzzy-C-Means-Algorithmus. Daten, die zu dem gleichen Cluster gehören, sind benachbart.

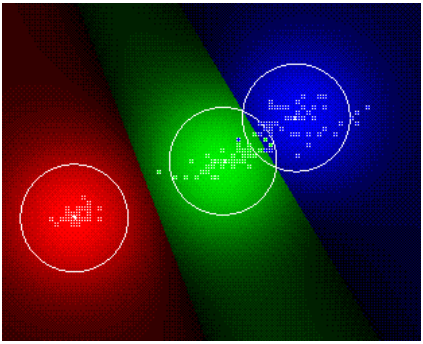

Abbildung 3.19: Clusteranalyse des Irisdatensatzes mit dem possibilistischen Ansatz basierend auf der *Zielfunktion (3.4)*. $\gamma = 1$, Attribute „petal length“ und „petal width“.

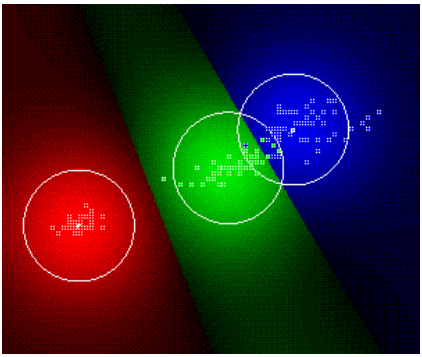

Abbildung 3.20: Clusteranalyse des Irisdatensatzes mit dem possibilistischen Ansatz basierend auf der *Zielfunktion (3.3)*. $\gamma = 1$, Attribute „petal length“ und „petal width“.

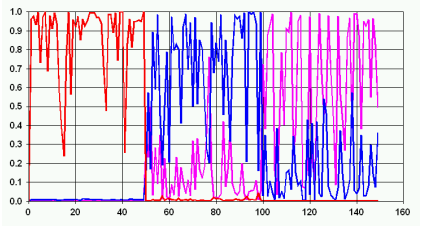

Abbildung 3.21: Zugehörigkeitsgrade der Daten zu den Clustern bei der Clusteranalyse des Irisdatensatzes mit dem possibilistischen Fuzzy-C-Means-Algorithmus basierend auf der *Zielfunktion (3.4)*($\gamma = 1$). Daten, die zu dem gleichen Cluster gehören, sind benachbart.

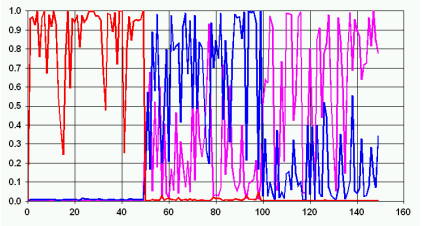

Abbildung 3.22: Zugehörigkeitsgrade der Daten zu den Clustern bei der Clusteranalyse des Irisdatensatzes mit dem possibilistischen Fuzzy-C-Means-Algorithmus basierend auf der *Zielfunktion (3.3)*($\gamma = 1$). Daten, die zu dem gleichen Cluster gehören, sind benachbart.

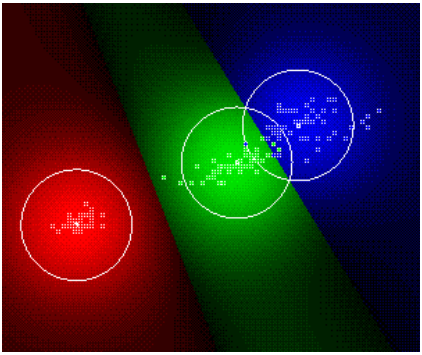

Abbildung 3.23: Clusteranalyse des Irisdatensatzes mit dem possibilistischen ACE und der *Modellierung der Abstoßung durch (3.18)*. $\gamma = 0.3$, Attribute „petal length“ und „petal width“.

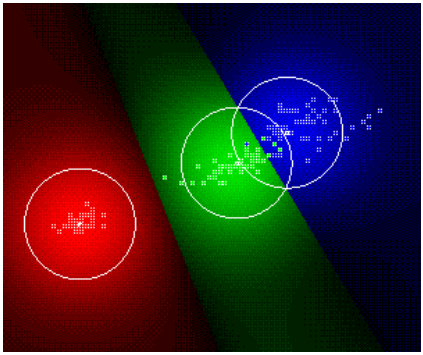

Abbildung 3.24: Clusteranalyse des Irisdatensatzes mit dem possibilistischen ACE und der *Modellierung der Abstoßung durch (3.19)*. $\gamma = 0.5$, Attribute „petal length“ und „petal width“.

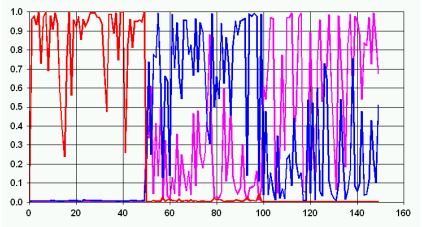

Abbildung 3.25: Zugehörigkeitsgrade der Daten zu den Clustern bei der Clusteranalyse des Irisdatensatzes mit dem possibilistischen ACE und der *Modellierung der Abstoßung durch (3.18)*($\gamma = 0.3$). Daten, die zu dem gleichen Cluster gehören, sind benachbart.

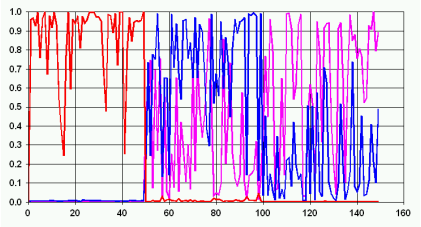

Abbildung 3.26: Zugehörigkeitsgrade der Daten zu den Clustern bei der Clusteranalyse des Irisdatensatzes mit dem possibilistischen ACE und der *Modellierung der Abstoßung durch (3.19)*($\gamma = 0.5$). Daten, die zu dem gleichen Cluster gehören, sind benachbart.

Weindatensatzes mit dem Gustafson–Kessel-Algorithmus sind trotz der Initialisierung mit dem probabilistischen Verfahren alle Cluster identisch, vgl. Abb. 3.31. Das Verfahren hat das (unerwünschte) globale Optimum der Zielfunktion gefunden. Durch die in diesem Kapitel 3 vorgestellte Idee der Berücksichtigung einer Abstoßung zwischen benachbarten Clustern kann dieses Zusammenfallen der Cluster vermieden werden. Die Abbildungen 3.32, 3.33, 3.34, 3.35, 3.36 und 3.37 zeigen die Ergebnisse der verschiedenen possibilistischen Varianten des Gustafson–Kessel-Algorithmus. Um die Einteilung des Datenraums in Regionen besser erkennen und die Ergebnisse besser mit der probabilistischen Fuzzy-Clusteranalyse vergleichen zu können, wurde der Datenraum leicht in der Farbe des Cluster, zu dem der größte Zugehörigkeitsgrad vorliegt, eingefärbt.

Mit $\gamma = 1$ werden bei dem Ansatz, basierend auf der Zielfunktion (3.3), 28 Daten und bei dem Ansatz, basierend auf der Zielfunktion (3.4), 35 Daten fehlklassifiziert. Bei dem Ansatz, basierend auf dem ACE und der Modellierung der Abstoßung durch (3.18), werden für $\gamma = 1$ 25 Daten fehlklassifiziert. Bei dem possibilistischen Ansatz, basierend auf dem ACE und der Modellierung der Abstoßung durch (3.19), muß γ auf 10 erhöht werden, um eine vergleichbare Klassifizierung zu erhalten. Es werden dann 28 Daten fehlklassifiziert. Die Ursache für die im Vergleich zum probabilistischen Gustafson–Kessel-Algorithmus aufgetretenen Abweichungen ist anhand der Abbildungen 3.32, 3.33, 3.34 und 3.35 erkennbar. Der blaue Cluster wird durch die Daten des roten und des grünen Clusters angezogen. Im Gegensatz zu dem probabilistischen Fuzzy-Clusteringverfahren handelt es sich bei dem possibilistischen um *kein* partitionierendes Verfahren. Eine große Punktwolke in der Mitte übt daher auf alle Cluster eine starke Anziehungskraft aus. Die Form des blauen Clusters ist daher ein Kompromiß zwischen den blauen Datenpunkten am Rand und den Datenpunkten in der Mitte.

Durch eine Erhöhung des Abstoßungsfaktors γ auf 6 kann bei dem Ansatz, basierend auf der Zielfunktion (3.3), und bei dem Ansatz, basierend auf dem ACE und der Modellierung der Abstoßung durch (3.18), die Form des blauen Clusters besser beschrieben werden, vgl. Abb. 3.36 und 3.37. Bei dem zielfunktionsbasierten Ansatz werden 35 Daten und bei dem „Alternating Cluster Estimation“ nur 6 Daten falsch klassifiziert. Bei der Modellierung der Abstoßung zwischen den Clustern durch die e-Funktion konnte durch eine Veränderung des Parameters γ der blaue Cluster nicht besser beschrieben werden. Die Ursache hierfür ist, daß bei der Normierung der Daten auf den Bereich $[0, 10]$ der Abstand des blauen Clusters zu den anderen Clustern so groß ist, daß bei der Modellierung des Abstoßungsgrads mit der e-Funktion der Abstoßungsgrad so klein ist, daß die Multiplikation

mit dem Parameter γ nicht ausreicht. Bei der Berechnung der Abstoßung durch $e^{-d^2(\vec{\beta}_i,\vec{\beta}_k)}$ ist ab einem Abstand von 2 zwischen benachbarten Clustern die Abstoßung gering und ab einem Abstand von 3 irrelevant, vgl. Abb. 3.2. Demgegenüber ist bei der Verwendung des Ausdrucks $\frac{1}{d^2(\vec{\beta}_i,\vec{\beta}_k)}$ auch bei einem etwas größeren Abstand zwischen benachbarten Clustern noch eine geringfügige Abstoßung vorhanden, die durch eine Erhöhung von γ ausgenutzt werden kann, vgl. Abb. 3.1.

Der Unterschied in der Form der berechneten Cluster zwischen den zielfunktionsbasierten Ansätzen und den Ansätzen, basierend auf dem ACE, beruht auf der unterschiedlichen Berechnung der Kovarianzmatrix der Cluster. Bei den zielfunktionsbasierten Ansätzen wird die Kovarianzmatrix durch benachbarte Cluster beeinflußt. Bei den Ansätzen, basierend auf dem ACE, ist man demgegenüber bei der Berechnungsvorschrift frei. Um die Form der Cluster möglichst exakt zu beschreiben, wird daher die Kovarianzmatrix allein aus den dem Cluster zugeordneten Daten berechnet.[5]

3.6 Bewertung

Im Gegensatz zu der probabilistischen Fuzzy-Clusteranalyse können bei der possibilistischen Fuzzy-Clusteranalyse die Zugehörigkeitsgrade im Sinne der Possibilitätstheorie interpretiert werden. Aus dem Ergebnis einer possibilistischen Fuzzy-Clusteranalyse kann direkt abgelesen werden, wie typisch ein Datum für einen Cluster ist. Dies bietet Vorteile bei der Generierung von Fuzzy-Mengen aus den Zugehörigkeitsgraden, z.B. der Erzeugung von Fuzzy-Regelsystemen mit Fuzzy-Clusteranalyse [75, 68].

Auch bei Datensätzen, bei denen sich Cluster stark überschneiden, ist die Verwendung possibilistischer Zugehörigkeitsgrade sinnvoll. Im Gegensatz zu der probabilistischen Fuzzy-Clusteranalyse können Daten, die für mehrere Cluster typisch sind, zu jedem dieser Cluster einen hohen Zugehörigkeitsgrad aufweisen. Dies führt dazu, daß bei einem großen Anteil von Daten, die mehreren Clustern zuzuordnen sind, die possibilistische Fuzzy-Clusteranalyse die Struktur der Daten besser beschreibt, während die probabilistische Fuzzy-Clusteranalyse die Cluster stärker separiert.

Für die Berechnung possibilistischer Zugehörigkeitsgrade wird für jeden Cluster der Abstand η_i benötigt, bei dem ein Zugehörigkeitsgrad von 0.5 vorliegen soll. Da eine Bestimmung dieses Abstands vor der Fuzzy-

[5] Eine Berechnung der Kovarianzmatrix unter Berücksichtigung benachbarter Cluster ist natürlich auch möglich.

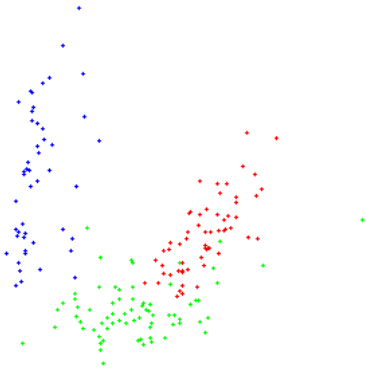

Abbildung 3.27: Projektion des Weindatensatzes auf die Attribute 7 und 10.

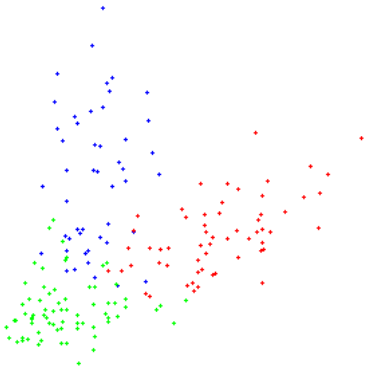

Abbildung 3.28: Projektion des Weindatensatzes auf die Attribute 10 und 13.

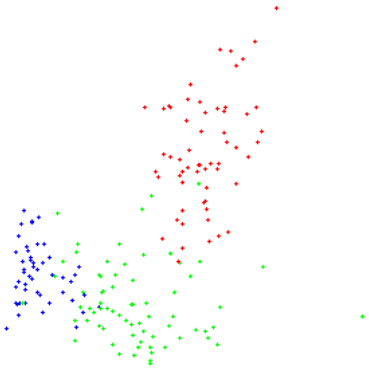

Abbildung 3.29: Projektion des Weindatensatzes auf die Attribute 7 und 13.

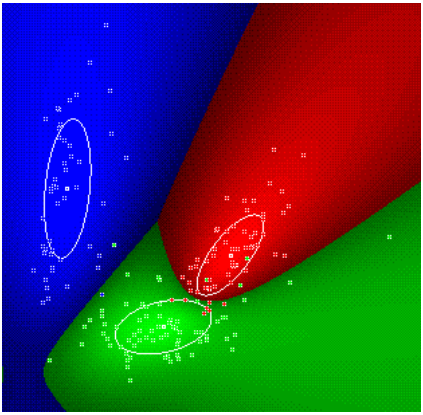

Abbildung 3.30: Clusteranalyse des Weindatensatzes mit dem *probabilistischen* Gustafson–Kessel-Algorithmus. Attribute 7 und 10.

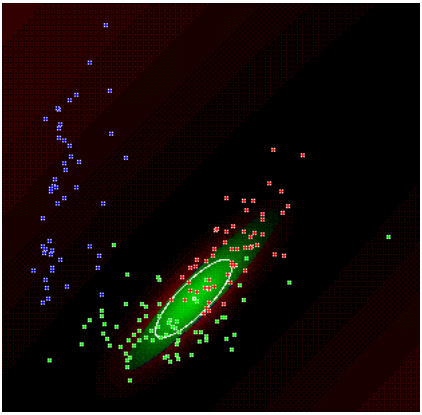

Abbildung 3.31: Clusteranalyse des Weindatensatzes mit dem *possibilistischen* Gustafson–Kessel-Algorithmus. Attribute 7 und 10.

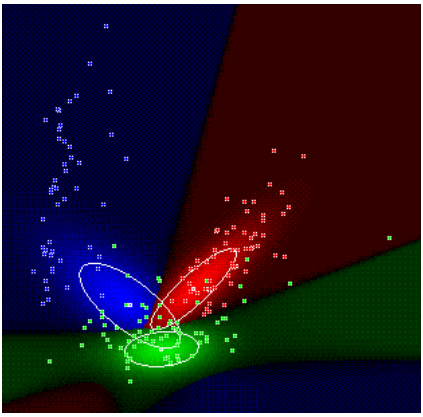

Abbildung 3.32: Clusteranalyse des Weindatensatzes mit dem possibilistischen Ansatz basierend auf der *Zielfunktion (3.3)*. $\gamma = 1$, Attribute 7 und 10.

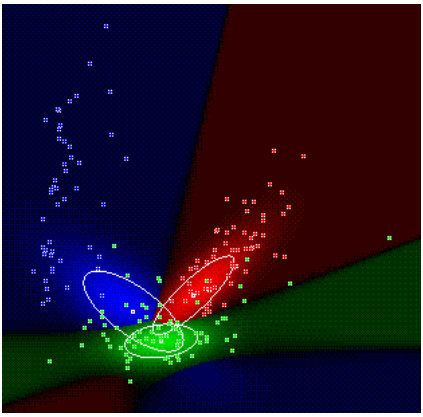

Abbildung 3.33: Clusteranalyse des Weindatensatzes mit dem possibilistischen Ansatz basierend auf der *Zielfunktion (3.4)*. $\gamma = 1$, Attribute 7 und 10.

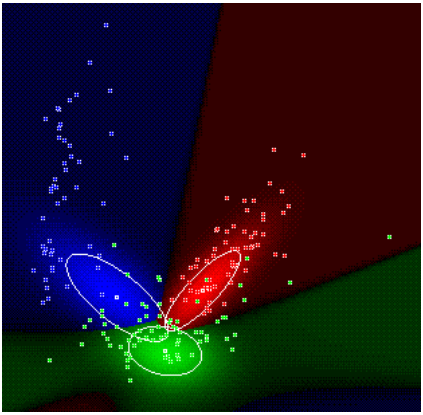

Abbildung 3.34: Clusteranalyse des Weindatensatzes mit dem possibilistischen ACE und der *Modellierung der Abstoßung durch (3.18)*. $\gamma = 1.0$, Attribute 7, 10.

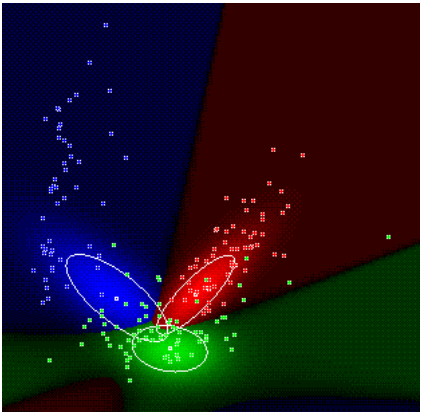

Abbildung 3.35: Clusteranalyse des Weindatensatzes mit dem possibilistischen ACE und der *Modellierung der Abstoßung durch (3.19)*. $\gamma = 10.0$, Attribute 7, 10.

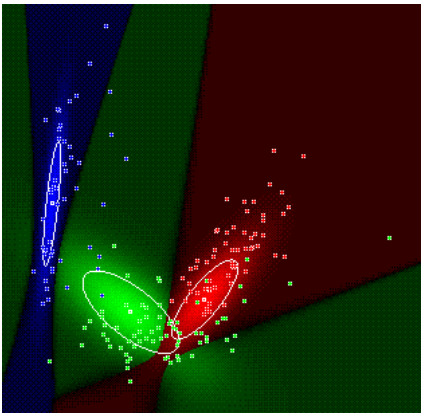

Abbildung 3.36: Clusteranalyse des Weindatensatzes mit dem possibilistischen Ansatz basierend auf der *Zielfunktion (3.3)*. $\gamma = 6.0$, Attribute 7 und 10.

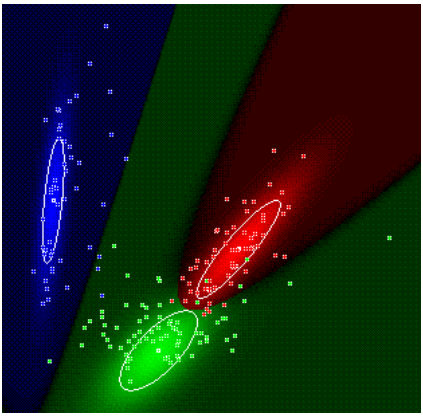

Abbildung 3.37: Clusteranalyse des Weindatensatzes mit dem possibilistischen ACE und der *Modellierung der Abstoßung durch (3.18)*. $\gamma = 6.0$, Attribute 7, 10.

Clusteranalyse sehr problematisch ist, wird die Größe der Cluster und damit auch η_i durch eine probabilistische Fuzzy-Clusteranalyse geschätzt. Die Initialisierung des possibilistischen Fuzzy-Clusteringverfahrens mit den Ergebnissen des probabilistischen Fuzzy-Clusteringverfahrens führt meistens zu guten Ergebnissen. Diese Ergebnisse stellen jedoch lokale Optima dar. Bei einem globalen Optimum sind alle Cluster identisch. Dieses unerwünschte globale Optimum tritt leicht auf, wenn die Cluster nicht gut separiert sind. Um auch bei nicht gut separierten Clustern eine possibilistische Fuzzy-Clusteranalyse durchführen zu können, ist eine Modifikation des Verfahrens erforderlich, die identische Cluster verhindert.

Identische Cluster können vermieden werden, indem neben der Anziehung der Cluster durch die Daten eine Abstoßung zwischen den Clustern modelliert wird. Hierfür wurde die Zielfunktion, die das Klassifikationsproblem beschreibt, entsprechend modifiziert. Die Abstoßung zwischen den Clustern wurde hierbei durch die Ausdrücke $\sum_{k=1,k\neq i}^{c} \frac{1}{\zeta d^2(\vec{\beta}_i,\vec{\beta}_j)}$ bzw. $\sum_{k=1,k\neq i}^{c} e^{\zeta - d^2(\vec{\beta}_i,\vec{\beta}_k)}$ modelliert. Neben den zielfunktionsbasierten Verfahren wurden auch die auf dem Alternating Cluster Estimation basierenden Verfahren betrachtet. Hierbei wurde die Abstoßung zwischen benachbarten Clustern direkt bei der Berechnung der Clusterprototypen betrachtet. Die Abstoßung wird clusterspezifisch in Relation zu der Anziehung durch die Daten durch den Parameter $\gamma_i = \frac{\sum_{j=1}^{n} u_{i,j}^m}{\gamma}$ gewichtet. Der Parameter ζ wird zur Skalierung des Abstands bei der Berechnung des Abstoßungsgrads verwendet.

Die Modellierung der Abstoßung durch $\frac{1}{\zeta d^2(\vec{\beta}_i,\vec{\beta}_j)}$ bietet gegenüber $e^{\zeta - d^2(\vec{\beta}_i,\vec{\beta}_k)}$ den Vorteil, daß durch eine Vergrößerung des Parameters γ auch eine abstoßende Wirkung auf nicht sehr dicht benachbarte Cluster ausgeübt werden kann. Die Skalierung des Abstands bzw. die Wahl des Parameters ζ ist daher nicht so kritisch. Die Ursache dieser Problematik ist die Verwendung des absoluten Abstands bei der Modellierung der Abstoßung. Nachteilig bei dem Ausdruck $\frac{1}{\zeta d^2(\vec{\beta}_i,\vec{\beta}_j)}$ ist die extrem große Abstoßung bei dicht benachbarten Clustern. Da die possibilistische Fuzzy-Clusteranalyse jedoch mit den Ergebnissen der probabilistischen Fuzzy-Clusteranalyse initialisiert wird, ist lediglich das „Zusammenfallen“ von Clustern zu verhindern. Ausgehend von unterschiedlichen Clustern verhindert die Abstoßung den Fall nahezu identischer Cluster.

Bei den zielfunktionsbasierten possibilistischen Ansätzen ist der Fall zu berücksichtigen, daß die Abstoßung durch benachbarte Cluster größer als die Anziehung durch die Daten ist. Dieser Fall kann so interpretiert werden, daß

der Cluster in dem betreffenden Gebiet des Datenraums „überflüssig“ ist. In diesem Fall kann der Cluster in einem noch nicht hinreichend abgedeckten Gebiet des Datenraums neu initialisiert werden. Bei den auf dem ACE basierenden possibilistischen Verfahren kann diese Problematik vermieden werden, indem die Abstoßung durch eine Anziehung aus der entgegengesetzten Richtung modelliert wird.

Die Berücksichtigung der anderen Cluster bei der Berechnung der Clusterprototypen erhöht bei allen vorgestellten Ansätzen die Komplexität *einer* Iteration des Verfahren von $O(n \cdot c)$ auf $O(n \cdot c + c^2)$. Die Laufzeit einer Iteration ändert sich hierdurch jedoch nur sehr geringfügig, da die Anzahl der Cluster c wesentlich kleiner als die Anzahl der Daten n ist und keine aufwendigen Rechenoperationen durchgeführt werden.

Anhand der Beispieldatensätze wurde die Wirkung der Abstoßung zwischen den Clustern aufgezeigt. Auch bei nicht gut separierten Clustern können die Cluster durch possibilistische Fuzzy-Clusteringverfahren erkannt werden. Die in diesem Kapitel vorgestellten Ansätze führen daher bei nicht gut separierten Clustern zu besseren Ergebnissen als die „normale“ possibilistische Fuzzy-Clusteranalyse. Bei gut separierten Clustern verhalten die beiden Verfahren sich gleich, da die abstoßende Wirkung bei hinreichend voneinander entfernten Clustern vernachlässigbar klein ist. Die in diesem Kapitel 3 vorgestellten Ansätze stellen damit eine Erweiterung der possibilistischen Fuzzy-Clusteranalyse dar.

Kapitel 4

Fuzzy-Clusteranalyse von Daten mit fehlenden Werten

4.1 Motivation

Ein häufig auftretendes Problem bei der Datenanalyse ist die Qualität der Daten. Oft sind Daten verrauscht, fehlerhaft oder es fehlen einzelne Attributwerte. Das Fehlen eines Attributwertes bei einem Datum wird auch als *„missing value“* bezeichnet. Die Probleme, die aus der Datenerhebung bzw. Datenerfassung resultieren, sind häufig unvermeidbar und im Nachhinein sehr schwer bzw. überhaupt nicht mehr korrigierbar. Daher ist bei der Datenanalyse die Fähigkeit, mit solchen Problemen umzugehen, von sehr großer Relevanz. Üblicherweise werden hierfür Verfahren aus dem Bereich der Datenvorverarbeitung verwendet. Daneben bieten viele Verfahren die Möglichkeit an, mit diesen Problemen direkt umzugehen.

Bei der Fuzzy-Clusteranalyse gibt es für den Umgang mit verrauschten oder fehlerhaften Daten mehrere Ansätze. Die wichtigsten dieser Verfahren wurden in Kapitel 2.8 kurz vorgestellt.

Im Gegensatz zu dem Umgang mit verrauschten oder fehlerhaften Daten ist bei der Fuzzy-Clusteranalyse die Berücksichtigung von Daten mit fehlenden Werten bisher noch nicht vertiefend betrachtet worden. Daher wird in diesem Abschnitt untersucht, wie Daten mit fehlenden Werten be-

rücksichtigt werden können.

Für den Umgang mit Daten mit fehlenden Werten („missing values“) gibt es prinzipiell drei verschiedene Ansätze:

- *Daten mit fehlenden Werten bzw. Attribute, in denen Daten fehlende Werte aufweisen, werden bei der Datenanalyse nicht berücksichtigt.* Diese Vorgehensweise ist häufig als Default-Ansatz implementiert. Sie ist sinnvoll, wenn fehlende Werte nur in wenigen Attributen konzentriert auftreten oder nur ein geringer Anteil der Daten fehlende Werte besitzt. Sofern jedoch der Anteil der fehlenden Werte größer ist, besteht die Problematik, daß zu viele Daten aus dem Datensatz entfernt werden und für die Datenanalyse nicht mehr zur Verfügung stehen. Ein weiteres Problem tritt auf, wenn bei einigen Clustern das Auftreten von Daten mit fehlenden Werten besonders ausgeprägt ist. Das Entfernen dieser Daten kann u.U. dazu führen, daß diese Cluster nicht mehr korrekt erkannt werden können.

  Die Analyse eines Datensatzes, bei dem Daten mit fehlenden Werten entfernt worden sind, wird auch als „complete-case analysis“ bezeichnet [86].

- *Fehlende Werte werden mittels statistischer Verfahren im Rahmen der Datenvorverarbeitung geschätzt („imputation“).* Hierfür gibt es zahlreiche Verfahren, wie z.B. die Verwendung des Means, Regressionsmethoden, „Expectation-Maximization“ Verfahren oder „Maximum-Likelihood“ Verfahren. Diese Verfahren werden z.B. in [86, 109] näher betrachtet.

  Der Nachteil dieser Vorgehensweise ist, daß bei den in den nachfolgenden Datenanalyseschritten eingesetzten Verfahren nicht mehr zwischen den geschätzten Werten und den beobachteten Werten unterschieden werden kann. Die Qualität der verwendeten Schätzverfahren beeinflußt daher u.U. die Ergebnisse der nachfolgenden Datenanalyseverfahren erheblich.

- *Die Verfahren zur Datenanalyse werden entsprechend adaptiert, so daß diese Verfahren mit Daten mit fehlenden Werten umgehen können.* Diese Methode ermöglicht es, bei der Datenanalyse Daten mit „missing values“ entsprechend zu berücksichtigen. Es besteht die Möglichkeit, die Nachteile der beiden o.g. Vorgehensweisen zu vermeiden bzw. zu umgehen. Die zur Verfügung stehenden Daten können in ihrem vollen Umfang bei der Analyse berücksichtigt werden. Es besteht die

Möglichkeit, bei der Datenanalyse zwischen Originalwerten und eventuell geschätzten „missing values“ zu unterscheiden.

In Rahmen dieses Kapitel 4 wird untersucht, ob, inwieweit und in welcher Form es möglich ist, die Verfahren der Fuzzy-Clusteranalyse so zu adaptieren, daß sie mit Daten mit fehlenden Werten umgehen können.

4.2 Arten von fehlenden Werten

4.2.1 Motivation

Die Ursachen für das Vorliegen von fehlenden Werten („missing values“) sind vielfältig. Fehlende Werte können z.B. durch Probleme bei der Datenerfassung verursacht werden, wie dem Ausfall von Sensoren bei dem Überwachen von Prozessen, oder bei Umfragen lassen Befragte bei Fragebögen einzelne Fragen unbeantwortet, weil sie Fragen übersehen oder die Beantwortung dieser Fragen ablehnen. Wenn man die ungenaue Beantwortung von Fragebögen näher betrachtet, erkennt man, daß es sinnvoll ist, bei dem Umgang mit Daten mit fehlenden Werten diese in verschiedene Kategorien einzuteilen. So kann man aus fehlenden Werten, die dadurch entstehen, daß bei dem Fragebogen Fragen übersehen werden, vielleicht Rückschlüsse auf die Gestaltung des Fragebogens ziehen. Eventuell lassen sich bei diesen Fragen die Probanden auch in die Gruppe der aufmerksamen und der eher unaufmerksamen Befragten einteilen. Weitergehende Rückschlüsse, die auf den Inhalt der Frage abzielen, sind jedoch problematisch. Demgegenüber kann das bewußte Nichtbeantworten einzelner Fragen durchaus gruppenspezifische Informationen beinhalten. So kann die Bereitschaft, Fragen nach dem Einkommen, dem Familienstand oder dem sozialen Status zu beantworten, bei verschiedenen Personengruppen unterschiedlich hoch sein. Die Information über ein Vorliegen von fehlenden Werten kann also bei der Kundensegmentierung von großer Relevanz sein.

Auch bei der Zuordnung von Daten zu Klassen können Informationen über eine eventuelle klassenspezifische Häufigkeit von fehlenden Werten vorteilhaft sein. In der Abbildung 4.1 werden zwei Kreise mit den Zentren $(2, 2)$ und $(6, 2)$ gezeigt. Wenn wir ein Datum betrachten, daß in dem ersten Attribut einen fehlenden Wert besitzt, z.B. $(?, 2)$, und wir wissen, daß fehlende Werte zufällig auftreten bzw. keine Informationen über die Ursache des fehlenden Wertes vorliegen, ist die Zuordnung dieses Datums zu einer der beiden Klassen willkürlich. Eine naheliegende Klassifikation ist daher, das

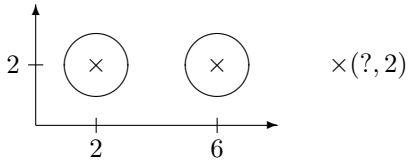

Abbildung 4.1: Ein Datensatz mit zwei kugelförmigen Klassen. Die Zentren sind durch × markiert. Falls Informationen über eine clusterspezifische Häufigkeit von fehlenden Werten vorliegen, können sie bei der Klassifikation des Datums $(?, 2)$ verwendet werden. Ansonsten sollte der Zugehörigkeitsgrad des Datums zu beiden Klassen gleich sein.

Datum beiden Klassen zu dem gleichen Grad zuzuweisen. Falls jedoch bekannt ist, daß bei Daten des Clusters mit dem Zentrum $(2, 2)$ eine geringere Wahrscheinlichkeit für das Fehlen von Attributen vorliegt als bei Daten des anderen Clusters, ist es naheliegend, das Datum $(?, 2)$ eher dem Cluster mit dem Zentrum $(6, 2)$ zuzuordnen. Der Zugehörigkeitsgrad des Datums zu dieser Klasse sollte in diesem Fall daher größer als der zu der anderen Klasse sein.

Die o.g. Beispiele verdeutlichen, daß es sinnvoll ist, zwischen verschiedenen Arten von fehlenden Werten zu unterscheiden. Daher werden im folgenden Abschnitt die verschiedenen Arten von „missing values“ näher erläutert.

4.2.2 Formale Betrachtung

Für den Umgang mit fehlenden Werten ist es sinnvoll, zuerst die Modellierung fehlender Werte näher zu betrachten. Die Einteilung der verschiedenen Arten von fehlenden Werten orientiert sich an [86, 109].

Das Modell für fehlende Werte basiert auf folgenden Annahmen:

- Der die Daten erzeugende Prozeß läßt sich durch die Angabe von zwei Parametersätzen θ und ξ vollständig beschreiben:
  - Der Parametersatz θ bestimmt, eventuell zusammen mit dem Parametersatz ξ, die Wahrscheinlichkeitsverteilungen der Zufallsvariablen, deren Realisierungen die (wahren) Daten sind. θ hat keinen Einfluß darauf, welche Daten beobachtbar sind.
  - Der Parametersatz ξ bestimmt, welche Realisierungen beobachtet werden können. Daneben kann er eventuell mit θ die Wahr-

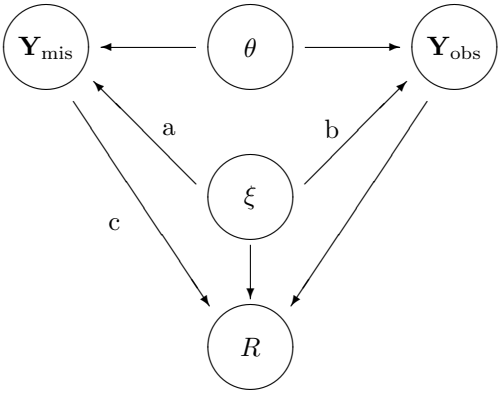

Abbildung 4.2: Ein allgemeines Modell für „missing values“. (Bedingter Unabhängigkeitsgraph)

scheinlichkeitsverteilungen der Zufallsvariablen, deren Realisierungen die (wahren) Daten sind, beeinflussen.

- Die Parametersätze θ und ξ sind, aufgefaßt als Zufallsvariablen, marginal unabhängig.

- Gegeben θ und ξ sind die Zufallsvariablen, deren Realisierungen die wahren Werte sind, bedingt unabhängig.

Wenn man die wahren Daten $\mathbf{Y}$ in die beobachteten Daten $\mathbf{Y}_{\text{obs}}$ und in die unbeobachteten (fehlenden) Daten $\mathbf{Y}_{\text{mis}}$ aufteilt und eine Zufallsvariable bzw. Indikatormatrix R verwendet, die angibt, ob ein Wert von $\mathbf{Y}$ beobachtet werden kann (er ist in $\mathbf{Y}_{\text{obs}}$) oder nicht beobachtbar ist (er ist in $\mathbf{Y}_{\text{mis}}$), kann ein allgemeines Modell fehlender Werte durch den bedingten Unabhängigkeitsgraph in Abb. 4.2 dargestellt werden.

Eine erste Einteilung von fehlenden Werten erfolgt in die Klassen *„ignorable“* und *„non-ignorable“*. Daten werden als „ignorable“ bezeichnet, wenn für die Schätzung der Parametersätze θ und ξ die wahren fehlenden Werte $\mathbf{Y}_{\text{mis}}$ ignoriert werden können. Andernfalls werden sie als „non-ignorable“

bezeichnet. [1]

Eine notwendige und hinreichende Bedingung für den „ignorable“-Fall ist das Fehlen der drei Kanten a, b, c in Abb. 4.2. Denn damit θ und ξ unabhängig von den wahren Werten $\mathbf{Y}_{\text{mis}}$ geschätzt werden können, müssen R und $\mathbf{Y}_{\text{mis}}$ bedingt unabhängig gegeben $\mathbf{Y}_{\text{obs}}$ sein. Mit der Kante a gibt es jedoch den aktiven Pfad R — ξ — $\mathbf{Y}_{\text{mis}}$, mit der Kante b den aktiven Pfad R — ξ — $\mathbf{Y}_{\text{obs}}$ — θ — $\mathbf{Y}_{\text{mis}}$ und mit der Kante c den aktiven Pfad R — $\mathbf{Y}_{\text{mis}}$.

Das Fehlen der Kanten a und b wird als *„distinctness“* der Parametersätze θ und ξ bezeichnet. Die Parametersätze θ und ξ (aufgefaßt als Zufallsvariablen) sind marginal unabhängig [86, 109].

Bei fehlenden Werten, die „non-ignorable“ sind, können die Parametersätze θ und ξ ohne $\mathbf{Y}_{\text{mis}}$ nicht geschätzt werden. Da $\mathbf{Y}_{\text{mis}}$ aber nicht bekannt ist, ist eine Schätzung von θ ohne weiteres Wissen über das Verfahren, das fehlende Werte verursacht, und dessen Parametersatz ξ nicht möglich. Im folgenden werden daher nur noch Daten, die “ignorable“ sind, betrachtet.

Das Modell fehlender Daten, die als „ignorable“ bezeichnet werden, sieht daher wie in dem in Abb. 4.3 gezeigten bedingten Unabhängigkeitsgraph aus.

Wenn die Wahrscheinlichkeit, daß ein Datum nicht beobachtet werden kann, von den beobachteten Daten $\mathbf{Y}_{\text{obs}}$, nicht jedoch von den fehlenden Daten $\mathbf{Y}_{\text{mis}}$ abhängt, werden die fehlenden Daten als *„missing at random“ (MAR)* bezeichnet. Bei Daten „missing at random“ gilt $P(R|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \xi) = P(R|\mathbf{Y}_{\text{obs}}, \xi)$ [86, 109]. In dem in Abb. 4.3 gezeigten bedingten Unabhängigkeitsgraphen wird dies durch die Kante d modelliert.

Daten „missing at random“ liegen z.B. vor, wenn bei einer Studie über die Leistungsfähigkeit von Studenten in einem Semester unter anderem die Ergebnisse verschiedener Tests als Attributwerte verwendet werden. Unter der Annahme, daß leistungsschwächere Studenten häufiger die Vorlesung nicht bis zum Ende besuchen als leistungsstärkere Studenten und somit auch an den letzten Tests nicht teilnehmen, kann das Fehlen der den letzten Tests zugeordneten Attributwerte aus den vorliegenden abgeleitet werden.

Bei Daten „missing at random“ kann die Wahrscheinlichkeit für das Fehlen von Daten aus den beobachteten Daten $\mathbf{Y}_{\text{obs}}$ abgeleitet werden. Wenn

[1] Eine gebräuchliche Definition für den Begriff „ignorable“ ist die Forderung, daß die fehlenden Werte „missing at random“ sind und die Parametersätze θ und ξ „distinct“ sind [86, 109]. Diese Definition entspricht der anschaulichen Darstellung in diesem Abschnitt. Die Begriffe „missing at random“ und „distinct“ werden im weiteren Verlauf dieses Abschnitts vorgestellt.

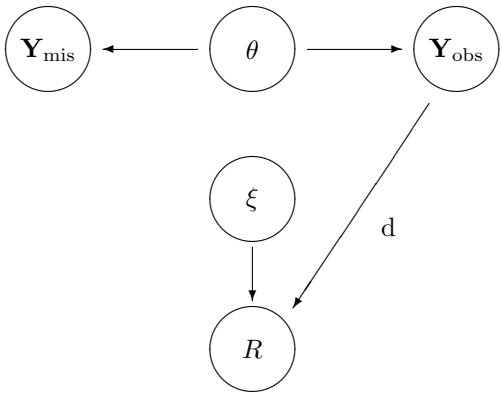

Abbildung 4.3: Ein Modell für „missing values“, die “ignorable“ sind. (Bedingter Unabhängigkeitsgraph)

dies nicht möglich ist (die Wahrscheinlichkeit, daß ein Datum fehlt, kann nicht aus den Daten $\mathbf{Y}_{\text{obs}}$ oder $\mathbf{Y}_{\text{mis}}$ abgeleitet werden), werden die fehlenden Daten als *„missing completely at random“ (MCAR)* bezeichnet. Es gilt $P(R|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \xi) = P(R|\xi)$ [86, 109]. In dem in Abb. 4.3 gezeigten bedingten Unabhängigkeitsgraphen bedeutet dies, daß bei Daten „missing completely at random“ die Kante d nicht vorliegt.

Ein anschauliches Beispiel für Daten mit „missing values missing completely at random“ ist [62]: Gegeben sei ein vollständiger Datensatz. Ein Mitarbeiter mischt diesen Datensatz und entfernt willkürlich einzelne Attributwerte. Diese Werte fehlen „missing completely at random“.

4.3 Fehlende Werte „missing completely at random“

Sofern nur ein kleiner Teil der Daten fehlende Werte „missing completely at random“ aufweist, ist die übliche Vorgehensweise, diese Daten bei der Fuzzy-Clusteranalyse zu vernachlässigen. Diese Vorgehensweise wird auch als „complete-case analysis“ bezeichnet. Sofern man auch an einer Klassi-

fikation der Daten mit fehlenden Werten interessiert ist, können diese den Clustern nach der Fuzzy-Clusteranalyse zugeordnet werden, indem der Abstand zwischen den Daten und den Clustern hinsichtlich der beobachteten Attribute bestimmt wird. Das Entfernen von Daten mit fehlenden Werten „missing completely at random“ ist aus statistischer Sicht zulässig, da dies als eine zufällige Reduktion des Datensatzes interpretiert werden kann. Sofern jedoch ein größerer Teil der Daten fehlende Werte besitzt, ist das Entfernen problematisch, da dies zu einer zu starken Verkleinerung des Datensatzes führt. In diesem Fall bietet sich die Verwendung einer der in den folgenden Abschnitten vorgestellten Methoden an.

4.3.1 Ein naheliegender Ad-Hoc Ansatz — Schätzen während der Fuzzy-Clusteranalyse

Bei der Fuzzy-Clusteranalyse wird die Klassifikationsaufgabe durch eine zu optimierende Zielfunktion beschrieben. Die Zielfunktion wird minimiert, indem abwechselnd die Zugehörigkeitsgrade $u_{i,j}$ der Daten zu den Clustern und die Clusterprototypen $\vec{\beta}_i$ berechnet werden. Bei den Solid-Clusteringverfahren können die dabei berechneten Zentren $\vec{z}_i$ der Cluster als für den jeweiligen Cluster typisches Datum aufgefaßt werden. Es ist daher naheliegend, den entsprechenden Attributwert des Clusterzentrums $\vec{z}_i$ als Schätzwert für den fehlenden Wert eines Datums, das diesem Cluster zugeordnet ist, zu verwenden.

Dieser Ansatz entspricht im Prinzip dem Schätzen von fehlenden Werten im Rahmen der Datenvorverarbeitung. Der Unterschied ist, daß der fehlende Wert nicht einmalig vor, sondern mehrmals während der Fuzzy Clusteranalyse, z.B. bei jeder n-ten Iteration, geschätzt wird. Die Schätzung während der Fuzzy-Clusteranalyse bietet im Vergleich zu der Schätzung im Rahmen der Datenvorverarbeitung den Vorteil, daß die inzwischen berechnete Clustereinteilung für die Schätzung der fehlenden Werte verwendet werden kann.

Der Aufbau eines Fuzzy-Clusteringverfahrens basierend auf diesem Ansatz ist:

Algorithmus 4.1 (Schätzung während der Fuzzy-Clusteranalyse)

- *Schätzung der fehlenden Werte (1).*
- *Initialisierung des Fuzzy-Clusteringverfahrens.*
- *REPEAT*
  - *Berechnung der Clusterprototypen.*
  - *Berechnung der Abstände der Daten zu den Clustern.*
  - *Berechnung der Zugehörigkeitsgrade.*
  - *Schätzung der fehlenden Werte (2).*
- *UNTIL Konvergenz des Verfahrens oder Überschreitung der maximalen Anzahl von Iterationen.*

Die Initialisierung des Fuzzy-Clusteringverfahrens sowie die Berechnung der Clusterprototypen und der Abstände der Daten zu den Clustern der Zugehörigkeitsgrade erfolgt wie in Kapitel 2 vorgestellt. Die Schätzung der fehlenden Werte unter (1) erfolgt im Rahmen der Datenvorverarbeitung. Bei einer geringen Anzahl von fehlenden Werten kann auch eine zufällige „Schätzung“ erfolgen. Bei der Schätzung der fehlenden Werte unter (2) sollte die bis dahin verwendete Klassifikation (die Zentren der Cluster und die Zugehörigkeitsgrade der Daten zu den Clustern) verwendet werden. [2] Im folgenden werden einige Ansätze für die Schätzung fehlender Werte näher betrachtet.

Möglichkeiten, fehlende Werte zu schätzen

Bei Solid-Clusteringverfahren kann das Clusterzentrum $\vec{z}_i$ als typisches Datum für den betreffenden Cluster interpretiert werden. Somit ist es naheliegend, fehlende Werte durch die entsprechenden Attributwerte des Clusterzentrums des Clusters zu schätzen, zu dem das betreffende Datum den höchsten Zugehörigkeitsgrad aufweist [122, 123, 86, 20, 59, 92]. Falls ein Datum $\vec{x}_j$ bei dem k-ten Attributwert einen fehlenden Wert hat und $\vec{\beta}_i$ der Cluster mit dem höchsten Zugehörigkeitsgrad ist, gilt $x_{j,k} = z_{i,k}$.

[2] Falls eine zuverlässige Schätzung ohne diese Zwischenergebnisse erfolgen kann, ist die Schätzung in jeder Iteration gleich. Das Verfahren entspricht damit einer „normalen“ Fuzzy-Clusteranalyse, bei der die fehlenden Daten im Rahmen der Datenvorverarbeitung geschätzt werden.

Diese Vorgehensweise bietet sich insbesondere bei dem Fuzzy-C-Means-Algorithmus an, da durch die Schätzung der Clusterprototyp (bei gleichbleibenden Zugehörigkeitsgraden) nicht beeinflußt wird. Bei der Anwendung bei anderen Fuzzy-Clusteringverfahren, wie z.B. dem Gustafson–Kessel-Algorithmus oder dem FMLE, ist eine weitere Modifikation erforderlich, um die Form der Cluster korrekt zu erkennen. Das Problem ist, daß bei der Schätzung fehlender Werte durch den entsprechenden Attributwert des Clusterzentrums zwar der Mittelwert (das Clusterzentrum) der dem Cluster zugeordneten Daten unverändert bleibt, die Varianzen und Kovarianzen jedoch unterschätzt werden. Dies führt dazu, daß die Form der Cluster bei einer größeren Anzahl von Daten mit fehlenden Werten schlecht erkannt wird.

In einem Datensatz mit n Daten sei bei $n^{(j)}$ Daten das j-te Attribut und bei $n^{(jk)}$ Daten das j-te Attribut *und* k-te Attribut beobachtet worden. Dann ist nach Schätzung der fehlenden Attributwerte durch den Mittelwert die Varianz aus den beobachteten und den geschätzten Attributwerten $\frac{(n^{(j)}-1)}{(n-1)} s_{jj}^{(j)}$. Dabei ist $s_{jj}^{(j)}$ die Varianz für das j-te Attribut, die aus den beobachteten Attributwerten berechnet wurde. Analog ist die Kovarianz zwischen dem j-ten Attribut und dem k-ten Attribut nach Schätzung der fehlenden Werte $\frac{(n^{(jk)}-1)}{(n-1)} \tilde{s}_{jk}^{(jk)}$. Dabei ist $\tilde{s}_{jk}^{(jk)}$ die Kovarianz für das j-te Attribut und das k-te Attribut, die aus den beobachteten Attributwerten berechnet wurde.[3] Indem man die Varianzen und die Kovarianzen mit $\frac{(n-1)}{(n^{(j)}-1)}$ bzw. $\frac{(n-1)}{(n^{(jk)}-1)}$ multipliziert, kann diese Deformation der Kovarianzmatrix vermieden werden [86].

Dies kann auf die Fuzzy-Clusteranalyse übertragen werden, indem man statt der Anzahl der Daten die Summe der Zugehörigkeitsgrade $u_{i,j}^m$ betrachtet. Nach der Schätzung der fehlenden Werte sollte die Varianz von Cluster $\vec{\beta}_i$ für das j-te Attribut mit $\frac{\sum_{l=1}^{n} u_{i,l}^m}{\sum_{\vec{x}_h \in \mathbf{X}_{obs}^{(j)}} u_{i,h}^m}$ multipliziert werden. $\mathbf{X}_{obs}^{(j)}$ ist dabei die Menge der Daten $\vec{x}_h$, bei denen das j-te Attribut beobachtet wurde. Analog sind die Kovarianzen von Cluster $\vec{\beta}_i$ zwischen dem j-ten Attribut und dem j-ten Attribut mit $\frac{\sum_{l=1}^{n} u_{i,l}^m}{\sum_{\vec{x}_h \in \mathbf{X}_{obs}^{(jk)}} u_{i,h}^m}$ zu multiplizieren. $\mathbf{X}_{obs}^{(jk)}$ ist dabei die Menge der Daten $\vec{x}_h$, bei denen das j-te Attribut *und* das k-te Attribut beobachtet wurden.

Ein Problem dieses Schätzverfahrens ist, daß bei allen Fuzzy-Clustering-

[3] Die Berechnung der Varianzen und der Kovarianzen aus den beobachteten Attributwerten wird in Abschnitt 4.3.3 näher betrachtet.

verfahren der Abstand des betreffenden Datums zu diesem Cluster unterschätzt wird. Der Abstand zu den anderen Clustern kann sich sowohl vergrößern als auch verkleinern. Da die Ähnlichkeit auf den Abständen der Daten zu den Clustern basiert (vgl. (2.4) und (2.23)), wird tendenziell die Ähnlichkeit und damit der Zugehörigkeitsgrad zu dem Cluster, zu dem das Datum den größten Zugehörigkeitsgrad aufweist, überschätzt. Dies bedeutet, daß Daten mit fehlenden Werten, verglichen mit Daten, bei denen alle Attributwerte beobachtet wurden, tendenziell ein größeres Gewicht bei der Berechnung der Clusterprototypen des Clusters aufweisen, zu dem sie den größten Zugehörigkeitsgrad besitzen. Dieses höhere Gewicht widerspricht jedoch der menschlichen Intuition, sich eher auf vollständige als auf unvollständige Informationen zu verlassen. Der tendenziell höhere Zugehörigkeitsgrad und damit das tendenziell höhere Gewicht bei des Berechnung der Clusterprototypen ist kritisch zu sehen, da die mit den Schätzwerten berechneten Clusterzentren wieder zu der Berechnung der fehlenden Werte verwendet werden.

Eine Möglichkeit, dieses Problem zu vermeiden, ist, die fehlenden Werte z.B. nach der Methode von Buck [31, 86] zu schätzen. Jedoch auch dieser Ansatz führt zu einer Unterschätzung der Varianzen und Kovarianzen, die jedoch geringer ausfällt im Vergleich zu der Schätzung durch den Mittelwert [86]. Eine andere Möglichkeit ist, das Gewicht von Daten mit fehlenden Werten bei der Fuzzy-Clusteranalyse zu verringern (vgl. Abschnitt 20).

Bei der Schätzung eines fehlenden Wertes durch den entsprechenden Attributwert des Clusterzentrums, zu dem dieses Datum den größten Zugehörigkeitsgrad aufweist, wird nicht berücksichtigt, wie typisch ein Datum für einen Cluster ist. So wird ein fehlender Wert bei einem Datum, das ein typischer Vertreter eines Clusters ist, genauso geschätzt, wie bei einem Datum, das eher mehreren Clustern zuzuordnen ist. Eine Möglichkeit, die Zugehörigkeitsgrade bei der Schätzung zu berücksichtigen, ist die Schätzung von fehlenden Werten durch das mit den Zugehörigkeitsgraden zu den Clustern gewichtete Mittel der entsprechenden Attributwerte aller Clusterprototypen [122, 123]. Hierbei können die Zugehörigkeitsgrade $u_{i,j}$ mit dem Fuzzifier m potenziert werden, um Cluster, zu denen das Datum einen höheren Zugehörigkeitsgrad besitzt, gegenüber Clustern, zu denen der Zugehörigkeitsgrad geringer ist, stärker zu gewichten. Falls ein Datum $\vec{x}_j$ bei dem k-ten Attributwert einen fehlenden Wert besitzt, gilt $x_{j,k} = \frac{\sum_{i=1}^{c} u_{i,j}^m z_{i,k}}{\sum_{i=1}^{c} u_{i,j}^m}$.

Diese Vorgehensweise bietet gegenüber der Schätzung durch den Attributwert nur eines Clusters den Vorteil, daß die geschätzten Attributwerte

während der Clusteranalyse nicht von einem Wert zu einem anderen „springen“, sondern ihren Wert allmählich ändern. Darüberhinaus können Daten, die einem Cluster nicht eindeutig zuzuordnen sind, sondern eher dem Grenzbereich bzw. Überlappungsbereich mehrerer Cluster zuzuordnen sind, besser berücksichtigt werden.

Problematik dieses Ansatzes

Ebenso wie bei dem Schätzen fehlender Werte im Rahmen der Datenvorverarbeitung wird auch bei diesem Ansatz während der Clusteranalyse nicht zwischen geschätzten Werten und beobachteten Werten unterschieden. Dadurch hat die Schätzung der Werte bei einer größeren Anzahl fehlender Werte einen großen Einfluß auf das Ergebnis der Fuzzy-Clusteranalyse. Ausgehend von einer eventuell unzuverlässigen oder schlechten Schätzung werden die Zugehörigkeitsgrade und die Clusterprototypen berechnet. Dabei sind sowohl die Zugehörigeitsgrade von Daten mit fehlenden Werten als auch die Clusterprototypen von einer eventuell schlechten Schätzung betroffen. Basierend auf diesen hierdurch möglicherweise unzuverlässigen Werten werden die fehlenden Werte erneut geschätzt. Dies kann zu einer Reduktion der Zuverlässigkeit der Schätzung und damit des Klassifikationsergebnisses führen.

Daneben sind ab der zweiten Iteration auch die Zugehörigkeitsgrade von Daten ohne fehlende Werte zu den Clustern von dieser Unsicherheit bzw. Unzuverlässigkeit betroffen, da die hierfür berechneten Abstände der Daten zu den Clustern durch die eventuell unzuverlässigen Prototypen der Cluster auch unzuverlässig bzw. unsicher sind. Mit zunehmender Iterationsanzahl wächst bei einer größeren Anzahl von Daten mit fehlenden Werten die Unsicherheit des Klassifikationsergebnisses.

Die Verwendung von Daten mit fehlenden Werten führt somit u.U. zu einer Reduktion der Zuverlässigkeit des Ergebnisses der Fuzzy-Clusteranalyse, verglichen mit einer Fuzzy-Clusteranalyse unter Auslassung von Daten mit fehlenden Werten.

4.3.2 Bestimmung fehlender Attributwerte als Optimierungsproblem

Der Umgang mit fehlenden Attributwerten läßt sich auch als Optimierungsproblem ansehen [20, 58, 59]. Bei dieser Betrachtung will man neben Zugehörigkeitsgraden der Daten zu den Clustern und den Clusterprototypen

auch die fehlenden Werte so bestimmen, daß die Zielfunktion des Verfahrens optimiert wird.

Der Ausdruck für die Bestimmung der fehlenden Werte ergibt sich durch Ableitung der Zielfunktion. Bei diesem Ansatz werden die Zugehörigkeitsgrade der Daten zu den Clustern, die Clusterprototypen und die fehlenden Werte abwechselnd berechnet, bis das Verfahren konvergiert. Dies wird als „tri-level alternating optimization“ bezeichnet. Das Verfahren führt zu einem Minimum bzw. einem Sattelpunkt der Zielfunktion [20, 59, 60]. Es wird als „optimal completion strategy“ bezeichnet.

Bei der Variante des Fuzzy-C-Means-Algorithmus werden die fehlenden Werte durch

$$x_{j,k} = \frac{\sum_{i=1}^{c} u_{i,j}^{m} z_{i,k}}{\sum_{i=1}^{c} u_{i,j}^{m}} \tag{4.1}$$

berechnet [20, 59, 60].

Für die Fuzzy-Clusteranalyse mit dem Fuzzy-C-Means-Algorithmus entspricht damit dieser Ansatz dem in Abschnitt 4.3.1 vorgestellten Ansatz mit Schätzung fehlender Attributwerte durch das gewichtete Mittel der entsprechenden Attributwerte aller Clusterprototypen. Der Unterschied zu dem vorhergehenden Abschnitt ist, daß sich der Ausdruck zur Berechnung fehlender Werte während der Fuzzy-Clusteranalyse aus der Zielfunktion motiviert.

Bei dem Gustafson–Kessel-Algorithmus führt die Minimierung der Zielfunktion nach einem Datum x_j ebenfalls zu

$$x_{j,k} = \frac{\sum_{i=1}^{c} u_{i,j}^{m} z_{i,k}}{\sum_{i=1}^{c} u_{i,j}^{m}}. \tag{4.2}$$

Dies kann analog zu der Herleitung des Ausdrucks für die Berechnung der Clusterzentren gezeigt werden. Die Herleitung wird daher im folgenden nur kurz skizziert.

$$\begin{aligned}
0 &= \frac{\partial}{\partial \vec{x}_j} \left(\sum_{i=1}^{c} \sum_{j=1}^{n} u_{i,j}^{m} d^2(\vec{\beta}_i, \vec{x}_j) \right) \\
&= \sum_{i=1}^{c} u_{i,j}^{m} \frac{\partial}{\partial \vec{x}_j} ||\vec{x}_j - \vec{z}_i||_{\mathbf{A}_i}^2 \\
&\Rightarrow \\
0 &= \sum_{i=1}^{c} u_{i,j}^{m} (\vec{x}_j - \vec{z}_i) \\
&\Rightarrow
\end{aligned}$$

$$\vec{x}_j = \frac{\sum_{i=1}^{c} u_{i,j}^m \vec{z}_i}{\sum_{i=1}^{c} u_{i,j}^m}$$

Die Kovarianzmatrizen der Cluster werden also bei der zielfunktionsbasierten Schätzung fehlender Werte nur indirekt über die Zugehörigkeitsgrade berücksichtigt.

4.3.3 Fuzzy-Clusteranalyse nach der „available case“-Methode

Berechnung der Clusterprototypen

Die Problematik der in Abschnitt 4.3.1 vorgestellten Vorgehensweise ist, daß die fehlenden Werte aus den entsprechenden Attributwerten der Clusterzentren bestimmt werden *und* die Prototypen der Cluster aus *allen* dem Cluster zugeordneten Daten, d.h. aus den beobachteten und den geschätzten Attributwerten berechnet werden. Diese Problematik läßt sich vermeiden, indem man die Clusterprototypen *nur* aus den beobachteten Attributwerten berechnet. Dies ist zulässig, da fehlende Werte „missing completely at random“ als eine zufällige Verkleinerung des Datensatzes aufgefaßt werden können.

Bei der „available case“-Methode [86] wird der Mittelwert $\bar{x}_j^{(j)}$ für das j-te Attribut bei einem Datensatz mit n Daten, bei dem bei $n^{(j)}$ Daten das j-te Attribut beobachtet wurde, durch

$$\bar{x}_j^{(j)} = \frac{1}{n^{(j)}} \sum_{\vec{x}_h \in \mathbf{X}_{obs}^{(j)}} x_{h,j}$$

berechnet. $\mathbf{X}_{obs}^{(j)}$ ist dabei die Menge der Daten $\vec{x}_h$, bei denen das j-te Attribut $x_{h,j}$ beobachtet wurde. Die Varianz für das j-te Attribut $s_{jj}^{(j)}$ wird analog durch

$$s_{jj}^{(j)} = \sum_{\vec{x}_h \in \mathbf{X}_{obs}^{(j)}} (x_{h,j} - \bar{x}_j^{(j)})^2$$

berechnet. Analog wird auch die Kovarianz für das j-te Attribut und das k-te Attribut $s_{jk}^{(jk)}$ berechnet. Hierbei werden jedoch nur Daten verwendet, bei denen das j-te Attribut *und* das k-te Attribut beobachtet wurden. Diese Daten werden als $\mathbf{X}_{obs}^{(jk)}$ bezeichnet. Die Anzahl der Daten aus $\mathbf{X}_{obs}^{(jk)}$ sei n_{jk}.

Die Kovarianz $s_{jk}^{(jk)}$ wird berechnet durch

$$s_{jk}^{(jk)} = \frac{1}{n_{jk}} \sum_{\vec{x}_h \in \mathbf{X}_{obs}^{(jk)}} (x_{h,j} - \bar{x}_j^{(jk)})(x_{h,k} - \bar{x}_k^{(jk)}).$$

Alternativ kann man anstelle der Mittelwerte $\bar{x}_j^{(jk)}$ und $\bar{x}_k^{(jk)}$, die aus $\mathbf{X}_{obs}^{(jk)}$ berechnet wurden, auch die Mittelwerte $\bar{x}_j^{(j)}$ bzw. $\bar{x}_k^{(k)}$ verwenden. Dies bietet den Vorteil, daß die Mittelwerte aus einer größeren Anzahl von Werten berechnet werden. Die Kovarianz ist in diesem Fall gegeben durch

$$\tilde{s}_{jk}^{(jk)} = \frac{1}{n_{jk}} \sum_{\vec{x}_h \in \mathbf{X}_{obs}^{(jk)}} (x_{h,j} - \bar{x}_j^{(j)})(x_{h,k} - \bar{x}_k^{(k)}).$$

Die Berechnung der Mittelwerte, der Varianzen und der Kovarianzen bei der „available case“-Methode kann auf die Fuzzy-Clusteranalyse übertragen werden. Dies wird für den Fuzzy-C-Means-Algorithmus, den Gustafson–Kessel-Algorithmus und den FMLE[4] anhand der Berechnung der Clusterzentren und der Kovarianzmatrizen aufgezeigt [122, 123, 20, 59].

Die Clusterzentren werden durch

$$z_{i,k} = \frac{\sum_{j=1}^{n} u_{i,j}^m i_{j,k} x_{j,k}}{\sum_{j=1}^{n} u_{i,j}^m i_{j,k}} \tag{4.3}$$

berechnet. Dabei ist $z_{i,k}$ der k-te Attributwert des Zentrums $\vec{z}_i$, $x_{j,k}$ der k-te Attributwert des Datums $\vec{x}_j$ und $i_{j,k}$ der k-te Attributwert des Indexvektors $\vec{i}_j$. $i_{j,k}$ gibt an, ob der k-te Attributwert des Datums $\vec{x}_j$ beobachtet wurde, d.h. $x_{j,k} = 1$, falls das k-te Attribut des Datums $\vec{x}_j$ vorliegt, und $x_{j,k} = 0$, falls das k-te Attribut des Datums $\vec{x}_j$ nicht beobachtet wurde.

Analog wird die Kovarianzmatrix $\mathbf{Cov}_i$ berechnet durch

$$\mathbf{Cov}_{i(k,l)} = \frac{\sum_{j=1}^{n} (u_{i,j})^m i_{j,k} i_{j,l} (x_{j,k} - z_{i,k})(x_{j,l} - z_{i,l})^\top}{\sum_{j=1}^{n} u_{i,j}^m i_{j,k} i_{j,l}}. \tag{4.4}$$

Berechnung der Abstände und der auf ihnen basierenden Zugehörigkeitsgrade

Bei der Fuzzy-Clusteranalyse werden die Zugehörigkeitsgrade der Daten zu den Clustern basierend auf den Abständen der Daten zu den Clustern bestimmt. Als Abstandsmaß wird bei dem Fuzzy-C-Means-Algorithmus der

[4] Die Berechnung der A-priori-Wahrscheinlichkeiten bei dem FMLE kann durch (2.14) erfolgen.

euklidische Abstand, bei dem Gustafson–Kessel-Algorithmus der Mahalanobisabstand und bei dem FMLE ein wahrscheinlichkeitsbasiertes Abstandsmaß verwendet (vgl. Kapitel 2). Der Abstand und damit auch der Zugehörigkeitsgrad kann bei Daten mit fehlenden Werten daher *nicht* berechnet werden. Eine Schätzung des Abstands bzw. des Zugehörigkeitsgrads ist erforderlich.

Schätzung des Abstands

Eine gute Vorgehensweise, um den Abstand bei Daten mit fehlenden Werten zu schätzen, ist, den Abstand hinsichtlich der beobachteten Attribute zu berechnen und ihn anschließend durch Multiplikation mit dem Ausdruck (Anzahl der Attribute/Anzahl der beobachteten Attribute) zu skalieren [41, 20]. So wird z.B. der Abstand bei dem Fuzzy-C-Means-Algorithmus durch

$$d(\vec{x}_j, \vec{\beta}_i) = \frac{p}{\sum_{k=1}^{p} i_{j,k}} \sum_{k=1}^{p} i_{j,k}(x_{j,k} - z_{i,k})^2 \tag{4.5}$$

berechnet. $i_{j,k}$ ist dabei ein Index, der angibt, ob das k-te Attribut des Datums $\vec{x}_j$ beobachtet wurde. Wenn das k-te Attribut fehlt, ist $i_{j,k} = 0$, sonst ist $i_{j,k} = 1$. p ist die Anzahl der Attribute. Mit diesen Abständen werden die Zugehörigkeitsgrade wie üblich berechnet.

Probabilistische Zugehörigkeitsgrade

Bei der probabilistischen Fuzzy-Clusteranalyse werden die Zugehörigkeitsgrade durch

$$u_{i,j} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d^2(\vec{x}_j, \vec{\beta}_i)}{d^2(\vec{x}_j, \vec{\beta}_k)} \right)^{\frac{1}{m-1}}}$$

bestimmt. Der Zugehörigkeitsgrad $u_{i,j}$ eines Datums $\vec{x}_j$ zu einem Cluster $\vec{\beta}_i$ basiert auf dem Verhältnis zwischen dem Abstand zu dem Cluster $\vec{\beta}_i$ und den Abständen zu den anderen Clustern $\vec{\beta}_k$ und nicht auf einer isolierten Betrachtung des Abstands $d^2(\vec{x}_j, \vec{\beta}_i)$. Da man bei der probabilistischen Fuzzy-Clusteranalyse primär an den Zugehörigkeitsgraden der Daten zu den Clustern und nicht so sehr an den exakten Abständen zwischen den Daten und den Clustern interessiert ist, bietet es sich an, diese Relation unter Berücksichtigung der beobachteten Werte direkt zu schätzen, indem man die Zugehörigkeitsgrade basierend auf den Abständen hinsichtlich der beobachteten Daten berechnet [122].

Diese Überlegungen führen zu den gleichen Zugehörigkeitsgraden wie die Schätzung des Abstands durch (4.5), da bei der Berechnung probabi-

listischer Zugehörigkeitsgrade der Skalierungsfaktor $\frac{p}{\sum_{k=1}^{p} i_{j,k}}$ weggekürzt werden kann.

Die in diesem Abschnitt vorgestellte Vorgehensweise der Schätzung der Zugehörigkeitsgrade unter Verwendung der Relation der Abstände hinsichtlich der beobachteten Attributwerte kann als optimistische Schätzung der Abstände interpretiert werde. Die Nichtberücksichtigung fehlender Attribute kann als das Schätzen durch den entsprechenden Attributwert des gerade betrachteten Clusterzentrums verstanden werden. Bei dieser Interpretation wird daher bei jedem Cluster ein „anderes“ Datum betrachtet anstelle eines gemeinsamen Datums.

Die Berechnung der Zugehörigkeitsgrade unter Verwendung der Abstände lediglich hinsichtlich der *beobachteten* Attributwerte führt bei dem Fehlen von klassifikationsrelevanten Attributen zu unschärferen Zugehörigkeitsgraden. Bei dem Fehlen von Attributen, die nicht klassifikationsrelevant sind, werden die Zugehörigkeitsgrade eher schärfer.

Possibilistische Zugehörigkeitsgrade
Bei possibilistischen Zugehörigkeitsgraden wird der Zugehörigkeitsgrad $u_{i,j}$ eines Datums $\vec{x}_j$ zu einem Cluster $\vec{\beta}_i$ basierend auf dem Abstand $d^2(\vec{x}_j, \vec{\beta}_i)$ in Relation zu dem Abstand η_i berechnet, bei dem der Zugehörigkeitsgrad $0,5$ beträgt. Bei Daten mit fehlenden Werten ist daher entweder der (nicht beobachtete) Abstand eines Datums mit fehlenden Werten zu schätzen oder der Abstand η_i ist entsprechend der Anzahl der beobachteten Attributwerte geeignet zu skalieren bzw. zu berechnen.

Reduktion des Gewichtes von Daten mit fehlenden Werten

Sowohl bei der probabilistischen als auch bei der possibilistischen Clusteranalyse ist eine Schätzung bei der Berechnung der Zugehörigkeitsgrade von Daten mit fehlenden Werten zu Clustern unvermeidbar. Der Zugehörigkeitsgrad bei Daten mit fehlenden Werten hat daher eine geringere Zuverlässigkeit als der bei Daten ohne fehlende Werte.

Diese geringere Zuverlässigkeit kann modelliert und somit bei der Fuzzy-Clusteranalyse berücksichtigt werden, indem bei Daten mit fehlenden Werten die Zugehörigkeitsgrade erniedrigt werden. Diese Verkleinerung der Zugehörigkeitsgrade sollte von der Anzahl der fehlenden Werte sowie, falls vorhanden, von der Relevanz der fehlenden Attribute abhängen, damit die Reduktion des Zugehörigkeitsgrads dem Verlust an Zuverlässigkeit entspricht.[5]

[5] Wenn keine Informationen über die Relevanz der Attribute vorliegen, sind die Attri-

Eine einfache Möglichkeit hierfür ist z.B.:

$$u_{i,j}^{(\text{neu})} = \left(\frac{\text{Anzahl der beobachteten Attribute von } \vec{x}_j}{\text{Anzahl aller Attribute von } \vec{x}_j} \right)^2 \cdot u_{i,j(\text{alt})}. \quad (4.6)$$

Dabei ist $u_{i,j}^{(\text{alt})}$ der Zugehörigkeitsgrad des Datums $\vec{x}_j$ zu dem Cluster $\vec{\beta}_i$ vor der Reduktion der Zugehörigkeitsgrade und $u_{i,j}^{(\text{neu})}$ der Zugehörigkeitsgrad danach.

Durch die Reduktion des Zugehörigkeitsgrads wird die Restriktion bei probabilistischen Fuzzy-Clusteringverfahren, daß jedes Datum das gleiche Gewicht haben soll, verletzt. Es handelt sich daher bei der Verwendung dieser Zugehörigkeitsgrade um kein probabilistisches Fuzzy-Clusteringverfahren im engeren Sinne mehr.

Es ist anzumerken, daß durch die Reduktion des Zugehörigkeitsgrads kein zielfunktionsbasiertes Fuzzy-Clusteringverfahren im engeren Sinne mehr vorliegt. Die Reduktion der Zugehörigkeitsgrade kann nicht aus der Zielstellung der Optimierung der Zielfunktion unter Berücksichtigung der Restriktionen abgeleitet werden. Das Verfahren entspricht daher eher dem Schema des Alternating Cluster Estimation (vgl. Abschnitt 2.10.3).

4.3.4 Testergebnisse

Die verschiedenen Ansätze zum Umgang mit fehlenden Attributwerten „missing completely at random“ wurden anhand des Brustkrebsdatensatzes [89] und des Weindatensatzes [1] aus dem „UCI Machine Learning Repository“ [21] hinsichtlich ihrer Leistungsfähigkeit näher untersucht.

Der Brustkrebsdatensatz (Wisconsin Breast Cancer Database) [89] besteht aus 699 Daten mit 9 Attributen. Die Daten unterteilen sich in gutartig (458 Daten) und bösartig (241 Daten). 16 Attributwerte wurden nicht beobachtet. Der Datensatz wurde mit dem probabilistischen Fuzzy-C-Means-Algorithmus klassifiziert. Der Datensatz wurde jeweils in zwei bzw. drei Cluster unterteilt.

Bei den im Datensatz enthaltenen 16 fehlenden Attributwerten unterscheidet sich die Clusteranalyse unter Auslassung von Daten mit fehlenden Attributwerten [6] nicht von den Verfahren, bei denen Daten mit fehlenden

bute als gleich relevant anzunehmen.

[6] Daten mit fehlenden Attributwerten werden nach der Fuzzy-Clusteranalyse dem Cluster zugeordnet, zu dem sie unter Berücksichtigung der beobachteten Attributwerte die größte Ähnlichkeit haben.

Werten in das Fuzzy-Clusteringverfahren integriert werden. Daher wurde die Anzahl fehlender Werte MCAR künstlich erhöht. Es wurden jeweils zehnmal fehlende Werte „missing completely at random“ mit einer Wahrscheinlichkeit von 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% bzw. 50% generiert.

Die Abbildungen 4.4 und 4.5 zeigen die gemittelten Ergebnisse der verschiedenen Verfahren in Abhängigkeit von der Wahrscheinlichkeit, daß Attributwerte fehlen.[7] Die rote Kurve zeigt die Anzahl der fehlklassifizierten Daten bei einer Fuzzy-Clusteranalyse unter Auslassung von Daten mit fehlenden Werten und anschließender Zuordnung der Daten mit fehlenden Werten zu dem ähnlichsten Cluster. Die grüne Kurve und die lila Kurve zeigen die Ergebnisse bei Schätzung der fehlenden Werte während der Fuzzy-Clusteranalyse (vgl. Abschnitte 4.3.1 und 4.3.2). Bei der grünen Kurve werden fehlende Werte durch den entsprechenden Attributwert des Clusterzentrums, zu dem sie den größten Zugehörigkeitsgrad besitzen, geschätzt. Bei der lila Kurve werden fehlende Werte durch das mit den Zugehörigkeitsgraden gewichtete Mittel der entsprechenden Attributwerte aller Clusterzentren berechnet. Die blaue und die hellblaue Kurve zeigen die Ergebnisse bei der „available case“-Methode (vgl. Abschnitt 4.3.3). Nur der Abstand bzw. der Zugehörigkeitsgrad wird hier geschätzt. Bei der hellblauen Kurve wurden Daten mit fehlenden Werten geringer gewichtet(vgl. Abschnitt 20).

Die Ergebnisse zeigen, daß es sinnvoll ist, Daten mit fehlenden Werten bei der Fuzzy-Clusteranalyse zu berücksichtigen. Während das Entfernen von Daten schon ab einer Wahrscheinlichkeit von ungefähr 10% für fehlende Werte „missing completely at random“ (bei 9 beobachteten Attributen) zu einem starken Anstieg der fehlklassifizierten Daten bei der Fuzzy-Clusteranalyse führt, sind die anderen Verfahren wesentlich leistungsfähiger. Der Vergleich der Ergebnisse der Fuzzy-Clusteranalyse mit 2 und mit 3 Clustern zeigt dabei, daß die Anzahl der Cluster für die Klassifikationsgüte der Verfahren, bei denen die fehlenden Werte geschätzt werden, von großer Relevanz ist. So führt bei der Fuzzy-Clusteranalyse mit 2 Clustern das Schätzen fehlender Werte durch den entsprechenden Attributwert des Clusterzentrums, zu dem das Datum den größten Zugehörigkeitsgrad besitzt, zu den besten Ergebnissen, während die Verwendung des gewichteten Mittels der Attributwerte der Clusterzentren wesentlich schlechter ist. Bei drei Clustern hingegen führen alle Ansätze zu ungefähr gleichen Ergebnissen. Die Ursache hierfür ist, daß die Schätzung auf den Clusterparametern basiert.

Auch bei einer Erhöhung der Wahrscheinlichkeit für fehlende Werte auf 50% zeigte sich die Leistungsfähigkeit der vorgestellten Ansätze. Bei dieser

[7] In Anhang B ist die Anzahl der gemittelten Fehler tabellarisch dargestellt.

Wahrscheinlichkeit für fehlende Werte war eine Fuzzy-Clusteranalyse basierend auf Daten ohne fehlende Werte bei zwei als auch bei drei Clustern nicht mehr möglich. Demgegenüber führten sowohl der Ansatz basierend auf der Schätzung fehlender Werte durch das gewichtete Mittel der Clusterzentren als auch der „available case“-Ansatz zu einer verhältnismäßig geringen Anzahl von Fehlklassifikationen. Bei der Einteilung in zwei Cluster wurden bei dem Ansatz basierend auf dem Schätzen fehlender Werte im Mittel 76 Daten und bei den „available case“-Ansätzen im Mittel 45 Daten fehlklassifiziert.[8] Bei der Einteilung in drei Cluster wurden bei dem Ansatz basierend auf dem Schätzen fehlender Werte im Mittel 39 Daten und bei den „available case“-Ansätzen im Mittel 35 Daten fehlklassifiziert.[9] Ein Vergleich mit der Anzahl fehlklassifizierter Daten bei einer Wahrscheinlichkeit für fehlende Werte von 40% zeigt keinen Einbruch der Klassifiaktionsgüte. Die Verfahren können die Redundanz der Informationen in dem neun-dimensionalen Datensatz gut nutzen. Lediglich bei dem Ansatz basierend auf einer Schätzung fehlender Attribute durch den entsprechenden Attributwert des Clusterzentrums mit dem höchsten Zugehörigkeitsgrad war eine Clusteranalyse in den meisten Fällen nicht mehr möglich.

Der Weindatensatz (Wine Recognition Data) [1] wurde mit dem probabilistischen Gustafson–Kessel-Algorithmus klassifiziert. Er besteht aus drei Klassen mit 59, 71 und 48 Daten. Die Daten sind das Resultat einer chemischen Analyse von Weinen aus der gleichen Region. Bei der Analyse wurden 13 Bestandteile der drei verschiedenen Weintypen untersucht. Für die Fuzzy-Clusteranalyse wurden von den 13 Attributen die Attribute 7, 10 und 13 verwendet. [10] Der Datensatz wurde hinsichtlich jedes Attributs auf den Wertebereich $[0, 10]$ skaliert. Es wurden jeweils zehnmal fehlende Werte „missing completely at random“ mit einer Wahrscheinlichkeit von 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% bzw. 50% generiert. Die Abbildung 4.6 zeigt die gemittelten Ergebnisse der verschiedenen Verfahren in Abhängigkeit von der Wahrscheinlichkeit, daß Attributwerte fehlen.[11]

Die rote Kurve zeigt die Anzahl der fehlklassifizierten Daten bei einer Fuzzy-Clusteranalyse unter Auslassung von Daten mit fehlenden Werten und anschließender Zuordnung der Daten mit fehlenden Werten zu dem ähn-

[8] 3 Datensätze waren mit keinem Verfahren mehr klassifizierbar. Sie wurden bei der Bestimmung des Mittels nicht berücksichtigt.

[9] 2 bzw. 3 Datensätze waren nicht klassifizierbar. Sie wurden bei der Bestimmung des Mittels nicht berücksichtigt.

[10] Die Abbildungen 3.27, 3.28 und 3.29 zeigen den Datensatz hinsichtlich der drei betrachteten Attribute.

[11] In Anhang B ist die Anzahl der gemittelten Fehler tabellarisch dargestellt.

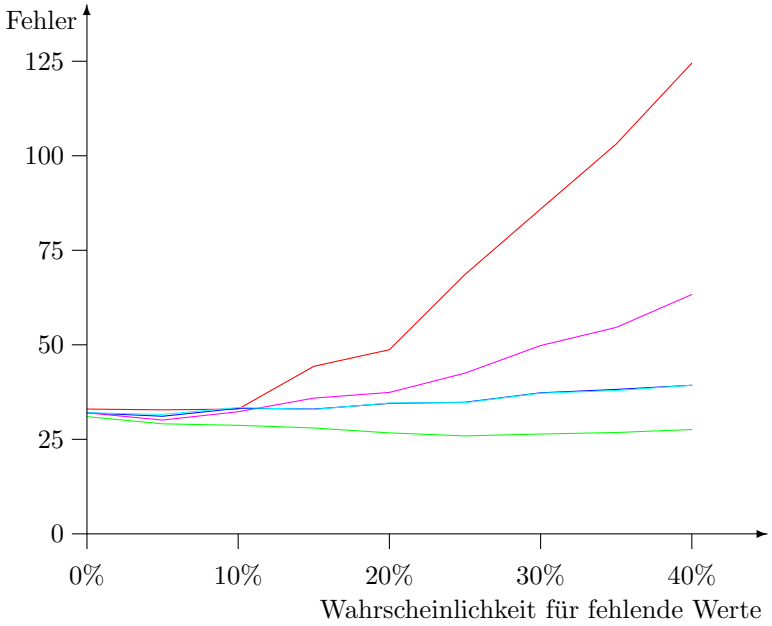

Abbildung 4.4: Anzahl der falsch klassifizierten Daten bei der Analyse des Brustkrebsdatensatzes (Wisconsin Breast Cancer Database) mit dem probabilistischen *Fuzzy-C-Means-Algorithmus und 2 Clustern*. Rot: Fuzzy-Clusteranalyse ohne Daten mit fehlenden Werten, grün: Fuzzy-Clusteranalyse mit Schätzung fehlender Werte durch den Attributwert eines Clusters, lila: Fuzzy-Clusteranalyse mit Schätzung fehlender Werte durch das gewichtete Mittel der Attributwerte aller Cluster, blau: Fuzzy-Clusteranalyse mit allen beobachteten Werten, hellblau: Fuzzy-Clusteranalyse mit allen beobachteten Werten und Reduktion des Gewichtes von Daten mit fehlenden Werten. Die blaue und die hellblaue Kurve fallen in dieser Abbildung nahezu zusammen.

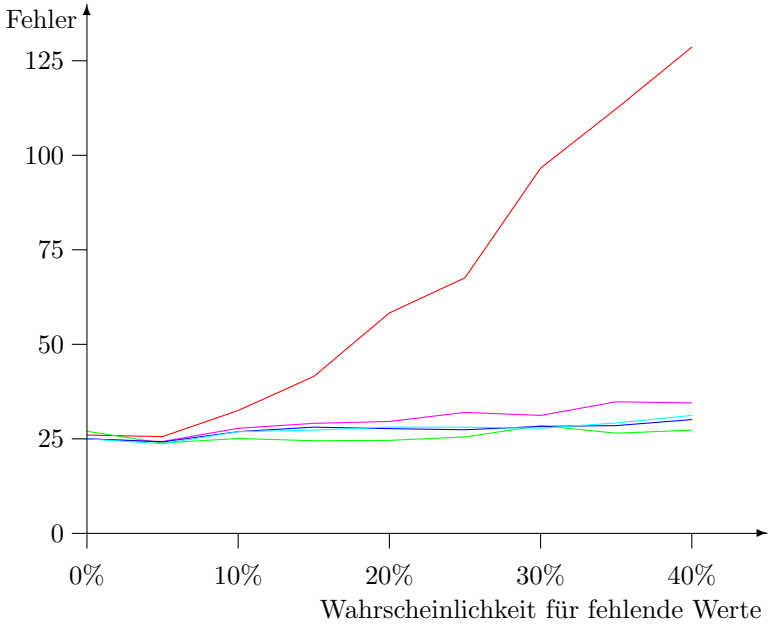

Abbildung 4.5: Anzahl der falsch klassifizierten Daten bei der Analyse des Brustkrebsdatensatzes (Wisconsin Breast Cancer Database) mit dem probabilistischen *Fuzzy-C-Means-Algorithmus und 3 Clustern.* Rot: Fuzzy-Clusteranalyse ohne Daten mit fehlenden Werten, grün: Fuzzy-Clusteranalyse mit Schätzung fehlender Werte durch den Attributwert eines Clusters, lila: Fuzzy-Clusteranalyse mit Schätzung fehlender Werte durch das gewichtete Mittel der Attributwerte aller Cluster, blau: Fuzzy-Clusteranalyse mit allen beobachteten Werten, hellblau: Fuzzy-Clusteranalyse mit allen beobachteten Werten und Reduktion des Gewichtes von Daten mit fehlenden Werten.

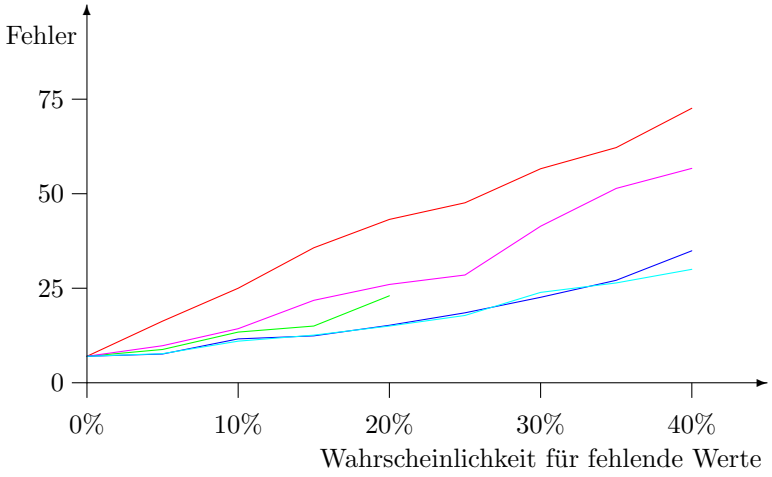

Abbildung 4.6: Anzahl der falsch klassifizierten Daten bei der Analyse des Weindatensatzes (Wine Recognition Data) mit dem probabilistischen *Gustafson–Kessel-Algorithmus und 3 Clustern*. Rot: Fuzzy-Clusteranalyse ohne Daten mit fehlenden Werten, grün: Fuzzy-Clusteranalyse mit Schätzung fehlender Werte durch den Attributwert eines Clusters, lila: Fuzzy-Clusteranalyse mit Schätzung fehlender Werte durch das gewichtete Mittel der Attributwerte aller Cluster, blau: Fuzzy-Clusteranalyse mit allen beobachteten Werten, hellblau: Fuzzy-Clusteranalyse mit allen beobachteten Werten und Reduktion des Gewichtes von Daten mit fehlenden Werten.

lichsten Cluster. Die grüne Kurve und die lila Kurve zeigen die Ergebnisse bei Schätzung der fehlenden Werte während der Fuzzy-Clusteranalyse (vgl. Abschnitte 4.3.1 und 4.3.2). Bei der grünen Kurve werden fehlende Werte durch den entsprechenden Attributwert des Clusterzentrums, zu dem sie den größten Zugehörigkeitsgrad besitzen, geschätzt. Um die Verringerung der Varianzen und Kovarianzen durch die Schätzung fehlender Werte zu vermeiden, wurde die Kovarianzmatrix nach dem in Abschnitt 18 vorgestellten Ansatz berechnet. Bei der lila Kurve werden fehlende Werte durch das mit den Zugehörigkeitsgraden gewichtete Mittel der entsprechenden Attributwerte aller Clusterzentren errechnet. Die blaue und die hellblaue Kurve zeigen die Ergebnisse bei der „available case“-Methode (vgl. Abschnitt 4.3.3). Nur der Abstand bzw. der Zugehörigkeitsgrad wird hier geschätzt. Bei der hellblauen Kurve wurden Daten mit fehlenden Werten geringer gewichtet(vgl. Abschnitt 20).

Bei dem Gustafson–Kessel-Algorithmus zeigt das Schätzen fehlender Werte durch den entsprechenden Attributwert des Clusterzentrums mit dem höchsten Zugehörigkeitsgrad trotz der Korrektur der Kovarianzmatrix bei einer höheren Anzahl von fehlenden Werten „missing completely at random“ Schwächen. Ab einer Wahrscheinlichkeit von 25% für fehlende Werte entsprach das Klassifikationsergebnis nur noch der Voraussage der Mehrheitsklasse. Die Cluster werden hier nicht mehr erkannt. Die Kurve hört daher bei einer Wahrscheinlichkeit von 20% für fehlende Werte auf.

Das Schätzen fehlender Werte durch das mit den Zugehörigkeitsgraden gewichtete Mittel der entsprechenden Attributwerte aller Cluster führte zu relativ schlechten Ergebnissen. Die Schätzung fehlender Werte führt zwar zu besseren Ergebnissen als das Entfernen von Daten mit fehlenden Werten vor der Fuzzy-Clusteranalyse. Auch die Problematik des Schätzens durch den entsprechenden Attributwert des Clusterzentrums, daß eine Fuzzy-Clusteranalyse nur bis zu einer Wahrscheinlichkeit von 20% für fehlende Werte durchgeführt werden konnte, trat nicht auf. Es wurden jedoch wesentlich mehr Daten fehlklassifiziert als bei dem „available case“-Ansatz. Dieser Ansatz zeigte sich auch gegenüber größeren Zahlen von Daten mit fehlenden Werten als sehr robust und führte zu guten Klassifikationsergebnissen.

Bei den Experimenten zeigte es sich, daß die Verfahren, die auf einer Schätzung fehlender Werte basieren, sehr empfindlich gegenüber der Initialisierung des Verfahrens sind.

Sowohl bei dem Brustkrebsdatensatz als auch bei dem Weindatensatz führte die Reduktion des Gewichtes von Daten mit fehlenden Werten zu keiner größeren Änderung des Klassifikationsergebnisses. Eine mögliche Ursache hierfür ist, daß die Berechnung des Abstands hinsichtlich der beobach-

teten Attribute tendenziell zu unschärferen Zugehörigkeitsgraden $u_{i,j}$ führt. Da bei der Berechnung der Clusterprototypen der Ausdruck $u_{i,j}^m$ verwendet wird, werden Daten mit fehlenden Werten automatisch geringer gewichtet, so daß eine zusätzliche Reduktion des Gewichtes sich nicht mehr so stark auswirkt.

Sowohl bei der Schätzung fehlender Werte *vor* der Fuzzy-Clusteranalyse als auch bei dem Entfernen von Daten bzw. Attributen mit fehlenden Werten aus dem Datensatz wird die Fuzzy-Clusteranalyse „normal durchgeführt“. Ein erhöhter Rechenaufwand durch Daten mit fehlenden Werten liegt daher nur im Rahmen der Datenvorverarbeitung vor. Die Höhe des Rechenaufwands ist dabei von dem verwendeten Schätzverfahren abhängig. Die Integration von Daten mit fehlenden Werten durch Schätzen *während* der Fuzzy-Clusteranalyse führt in Abhängigkeit von dem Schätzverfahren zu einem erhöhten Rechenbedarf während der Fuzzy-Clusteranalyse. Bei der Veränderung einfacher Schätzverfahren, wie die Verwendung des entsprechenden Attributwerts des Clusterzentrums, ist der zusätzliche Aufwand jedoch gering, da die Clusterzentren und die Zugehörigkeitsgrade vorher berechnet wurden. Die Komplexität *einer* Iteration bleibt bei $O(n \cdot c)$. Auch die Integration fehlender Werte nach dem „available case“-Ansatz führt nur zu einem geringfügig erhöhten Rechenbedarf. Alle vorgestellten Ansätze sind daher problemlos auch bei größeren Datenmengen einsetzbar.

Bei einer geringen Anzahl von Daten mit fehlenden Attributwerten „missing completely at random“ bietet es sich daher an, Daten, bei denen nicht alle Attributwerte beobachtet wurden, vor der Fuzzy-Clusteranalyse zu entfernen. Bei einem größeren Anteil von Daten mit fehlenden Werten sollte jedoch der „available case“-Ansatz verwendet werden. Die Schätzung fehlender Werte bietet sich an, wenn ein „gutes“ Schätzverfahren verwendet werden kann. Dies kann jedoch in Abhängigkeit von dem verwendeten Schätzverfahren gegebenenfalls zu einem erhöhten Rechenaufwand führen, da die Schätzung bei jeder Iteration durchgeführt wird.

4.4 Daten mit einer clusterspezifischen Wahrscheinlichkeit für fehlende Werte

4.4.1 Allgemeine Betrachtungen

Im Gegensatz zu „missing values missing completely at random“ hängt bei „missing values missing at random“ die Wahrscheinlichkeit, daß ein Datum nicht beobachtet wurde, von den beobachteten Werten ab. Es liegen

also zusätzliche Informationen vor, die bei der Behandlung von Daten mit fehlenden Werten verwendet werden können.

Bei der Clusteranalyse wird ein Datensatz in Cluster von homogenen Daten eingeteilt. Aufgrund der Homogenitätsforderung wird im folgenden angenommen, daß die Wahrscheinlichkeit für ein nicht beobachtetes Datum innerhalb eines Clusters gleich ist.[12] Die Annahme von „missing values missing at random“ ist in diesem Fall als clusterspezifische Wahrscheinlichkeit für fehlende Werte interpretierbar, wenn man annimmt, daß man die Zugehörigkeit der Daten zu den Clustern kennt.[13] Diese clusterspezifische Wahrscheinlichkeit kann während der Clusteranalyse geschätzt und für den Umgang mit Daten mit fehlenden Werten verwendet werden.

4.4.2 Ein wahrscheinlichkeitsbasierter Abstand

Das wahrscheinlichkeitsbasierte Abstandsmaß des FMLE bietet eine Möglichkeit, den Abstand mit der clusterspezifischen Wahrscheinlichkeit für fehlende Werte direkt zu kombinieren. Der FMLE unterteilt einen Datensatz in Cluster unter der Annahme, daß die Daten der Cluster Realisierungen p-dimensionaler Wahrscheinlichkeitsverteilungen sind. Diese Wahrscheinlichkeitsverteilungen werden bei der Ausführung des FMLE bestimmt. Der Abstand der Daten zu den Clustern ist dabei umgekehrt proportional zu der Wahrscheinlichkeit, daß das Datum von der Wahrscheinlichkeitsverteilung erzeugt wurde, die dem betreffenden Cluster zugrundeliegt.

Die Idee des in diesem Abschnitt 4.4.2 vorgestellten Ansatzes ist, die clusterspezifische Wahrscheinlichkeit für das Fehlen von Daten in das Modell des FMLE zu integrieren [125]. Die Daten werden dabei als Realisierung einer p-dimensionalen Wahrscheinlichkeitsverteilung N_i gesehen, bei der das k-te Attribut mit einer Wahrscheinlichkeit $p_{i,k}^{(\mathrm{mv})}$ fehlt. Die Wahrscheinlichkeitsverteilung wird dabei mit einer Wahrscheinlichkeit P_i ausgewählt. Da bei einer Betrachtung der Daten getrennt nach Clustern die fehlenden Werte „missing completely at random sind“ und die Daten durch eine Normalverteilung erzeugt wurden, kann das Modell wie folgt modifiziert werden: Zuerst wird mit einer Wahrscheinlichkeit P_i eine Klasse i ausgewählt. Danach wird mit Wahrscheinlichkeiten $\vec{p}_i^{(\mathrm{mv})}$ entschieden, welche Attribute eines Datums beobachtbar sind. Danach wird das Datum durch die Normalverteilung N_{il} erzeugt. l ist dabei ein Index, der angibt, welche Attribute fehlen. Da die

[12] Nach Clustern getrennt liegen daher fehlende Werte „missing values missing completely at random“ vor.

[13] Formal ist dies nicht ganz korrekt, da man die Clusterzugehörigkeit nur schätzt und nicht beobachtet.

Daten eines Clusters $\vec{\beta}_i$ durch dieselbe Normalverteilung N_i erzeugt werden sollen, sind die Normalverteilungen N_{il} die Marginalverteilungen von N_i. Daher sind die A-posteriori-Wahrscheinlichkeiten für beide Modelle gleich.

Diese Annahme führt zu der folgenden A-posteriori-Wahrscheinlichkeit (Likelihood), daß ein Datum $\vec{x}_j$ mit einem fehlendem Wert in dem k-ten Attribut durch die Normalverteilung N_l erzeugt wurde.

$$\frac{P_i \cdot \left(1 - p_{i,1}^{(\mathrm{mv})}\right) \cdot \ldots \cdot \left(1 - p_{i,k-1}^{(\mathrm{mv})}\right) \cdot p_{i,k}^{(\mathrm{mv})} \cdot \left(1 - p_{i,k+1}^{(\mathrm{mv})}\right) \cdot \ldots \cdot \left(1 - p_{i,p}^{(\mathrm{mv})}\right)}{(2\pi)^{p/2} \sqrt{\det(\mathbf{A}_i)} e^{-\frac{1}{2}(\vec{x}_j - \vec{z}_i)^\top \mathbf{A}_i^{-1}(\vec{x}_j - \vec{z}_i)}} \tag{4.7}$$

Den vorhergehenden Betrachtungen folgend werden bei der Berechnung des Ausdrucks

$$\frac{e^{-\frac{1}{2}(\vec{x}_j - \vec{z}_i)^\top \mathbf{A}_i^{-1}(\vec{x}_j - \vec{z}_i)}}{(2\pi)^{p/2}\sqrt{\det(\mathbf{A}_i)}} \tag{4.8}$$

nicht beobachtete Attributwerte des Datums $\vec{x}_j$ nicht berücksichtigt.

Der Abstand zwischen einem Datum $\vec{x}_j$ und einem Cluster $\vec{\beta}_i$ ist bei dem FMLE umgekehrt proportional zu der Wahrscheinlichkeit, daß das Datum durch die dem Cluster zugrundeliegende Wahrscheinlichkeitsverteilung erzeugt wurde. Diese Idee führt bei Daten mit fehlenden Werten zu dem Abstand

$$d^2\left(\vec{x}_j, \left(\vec{z}_i, \mathbf{A}_i, P_i, \vec{p}_i^{(\mathrm{mv})}\right)\right) = \frac{1}{P_i \cdot \left(1 - p_{i,1}^{(\mathrm{mv})}\right) \cdot \ldots \cdot \left(1 - p_{i,k-1}^{(\mathrm{mv})}\right) \cdot p_{i,k}^{(\mathrm{mv})} \cdot \left(1 - p_{i,k+1}^{(\mathrm{mv})}\right) \cdot \ldots \cdot \left(1 - p_{i,p}^{(\mathrm{mv})}\right)} \cdot \sqrt{\det(\mathbf{A}_i)} e^{\frac{1}{2}(\vec{x}_j - \vec{z}_i)^\top \mathbf{A}_i^{-1}(\vec{x}_j - \vec{z}_i)}. \tag{4.9}$$

Analog zu der Bestimmung der A-posteriori-Wahrscheinlichkeiten werden Attribute, bei denen $\vec{x}_j$ einen fehlenden Wert hat, bei der Berechnung des Ausdrucks

$$\sqrt{\det(\mathbf{A}_i)} e^{\frac{1}{2}(\vec{x}_j - \vec{z}_i)^\top \mathbf{A}_i^{-1}(\vec{x}_j - \vec{z}_i)} \tag{4.10}$$

nicht berücksichtigt. Bei der Bestimmung der Determinante der Matrix $\mathbf{A}_i$ werden die entsprechenden Zeilen und Spalten der Attribute gestrichen, bei denen $\vec{x}_j$ einen fehlenden Wert hat, da die Marginalverteilungen betrachtet werden.

Da alle Daten, die zu dem gleichen Cluster gehören, durch die gleiche Normalverteilung erzeugt werden, unabhängig davon, ob sie fehlende Werte haben oder nicht, werden der Mittelwert $\vec{z}_i$ sowie die Kovarianzmatrix und

damit auch $\mathbf{A}_i$ nach der „available case“-Methode aus den beobachteten Attributwerten berechnet (vgl. Abschnitt 4.3.3).

Die Wahrscheinlichkeiten P_i und $\vec{p}_i^{(\mathrm{mv})}$ werden wie folgt berechnet:

$$P_i = \frac{\sum_{j=1}^{n} u_{i,j}^m}{\sum_{j=1}^{n}\sum_{l=1}^{c} u_{i,j}^m}, \tag{4.11}$$

$$p_{i,k}^{(\mathrm{mv})} = \frac{\sum_{j=1}^{n} u_{i,j} i_{j,k}}{\sum_{j=1}^{n} u_{i,j}}. \tag{4.12}$$

$\vec{i}_j$ ist ein Indexvektor für fehlende Werte bei dem Datum $\vec{x}_j$. $i_{j,k}$ ist 1, wenn das k-te Attribut nicht beobachtet wurde, und 0 sonst. $p_{i,k}^{(\mathrm{mv})}$ ist das k-te Attribut des Vektors $\vec{p}_i^{(\mathrm{mv})}$.

4.5 Experimentelle Ergebnisse

Der vorgestellte Ansatz zur Berücksichtigung clusterspezifischer Wahrscheinlichkeiten wurde anhand des Weindatensatzes [1] näher betrachtet. Er besteht aus drei Klassen mit 59, 71 und 48 Daten. Es wurden die Attribute 7, 10 und 13 verwendet. [14] Der Datensatz wurde hinsichtlich jedes Attributs auf den Wertebereich $[0, 10]$ skaliert. Der Datensatz enthält fehlende Werte MCAR mit einer Wahrscheinlichkeit von 5% und fehlende Werte mit einer clusterspezifischen Wahrscheinlichkeit zwischen 20% und 60%. Die fehlenden Werte mit einer clusterspezifischen Wahrscheinlichkeit treten jeweils in einem unterschiedlichen Attribut auf.

Als Vergleichsverfahren zu dem wahrscheinlichkeitsbasierten Verfahren wurde die in diesem Kapitel vorgestellte Variante des FMLE für den Umgang mit fehlenden Werten MCAR verwendet, bei der Daten mit fehlenden Werten bei der Clusteranalyse berücksichtigt werden, indem alle beobachteten Attribute verwendet werden. Der menschlichen Intuition entsprechend wurde das Gewicht fehlender Daten reduziert, vgl. Abschnitt 20. Der Unterschied zwischen den beiden Verfahren ist somit nur die zusätzliche Berücksichtigung einer clusterspezifischen Wahrscheinlichkeit für fehlende Werte. Beide Varianten des FMLE wurden mit dem Fuzzy-C-Means-Algorithmus für den Umgang mit fehlenden Werten MCAR initialisiert. Durch diese Initialisierung wird die Vergleichbarkeit der Ergebnisse sichergestellt. Abbildung 4.7 zeigt die Ergebnisse der Fuzzy-Clusteranalyse des Weindatensatzes

[14] Die Abbildungen 3.27, 3.28 und 3.29 zeigen den Datensatz hinsichtlich der drei betrachteten Attribute.

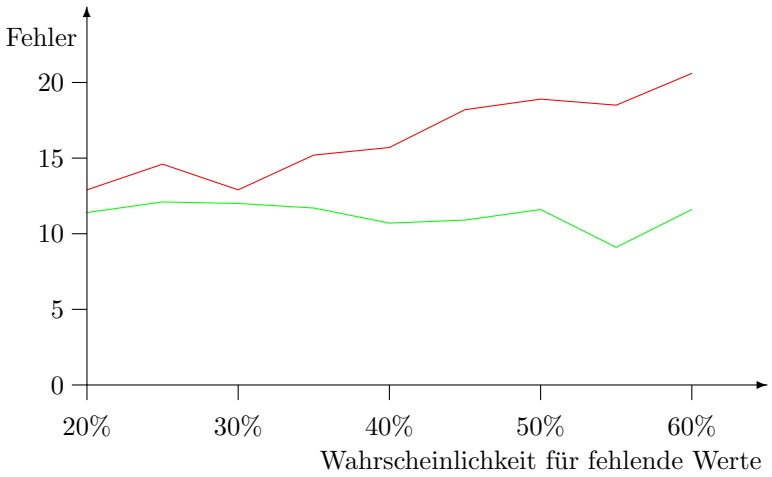

Abbildung 4.7: Anzahl der falsch klassifizierten Daten bei der Analyse des Weindatensatzes (Wine Recognition Data) mit dem probabilistischen FMLE-Algorithmus und 3 Clustern. Rot: Fuzzy-Clusteranalyse mit allen beobachteten Werten und Reduktion des Gewichtes von Daten mit fehlenden Werten, grün: Berücksichtigung der clusterspezifischen Wahrscheinlichkeit für fehlende Werte.

mit den beiden Verfahren mit 3 Clustern. Die rote Kurve zeigt die Anzahl der falsch klassifizierten Daten bei dem FMLE für fehlende Werte MCAR und die grüne Kurve bei dem FMLE mit Berücksichtigung einer clusterspezifischen Wahrscheinlichkeit für fehlende Werte an. Bei einer clusterspezifischen Wahrscheinlichkeit von 20% sind die Ergebnisse beider Verfahren ungefähr vergleichbar. Die Berücksichtigung der clusterspezifischen Wahrscheinlichkeit führt nur zu geringfügig besseren Ergebnissen. Bei einer Zunahme fehlender Werte mit einer clusterspezifischen Wahrscheinlichkeit verschlechtert sich jedoch die Klassifikationsgüte bei der Nichtberücksichtigung dieser Information, während sie bei der Berücksichtigung dieser clusterspezifischen Wahrscheinlichkeit ungefähr gleich bleibt. Der Informationsverlust durch die fehlenden Werte wird ausgeglichen durch die Berücksichtigung der clusterspezifischen Wahrscheinlichkeit.

Die Berücksichtigung einer clusterspezifischen Wahrscheinlichkeit für feh-

lende Werte führt verglichen mit dem „available case“ Ansatz nur zu einem geringfügig größeren Rechenaufwand (die Berechnung einer clusterspezfischen Wahrscheinlichkeit für jeden Cluster und die Berücksichtigung dieser Wahrscheinlichkeit bei der Berechnung des Abstands). Die Komplexität *einer* Iteration des Verfahrens ändert sich nicht. Der Ansatz ist daher auch auf größeren Datensätze anwendbar.

4.6 Bewertung

Daten mit fehlenden Werten können im Rahmen der Datenvorverarbeitung geschätzt werden. Hierfür gibt es verschiedene Ansätze aus der Statistik. Die Problematik ist jedoch, daß in den nachfolgenden Schritten eines Datenanalyseprozesses nicht mehr zwischen geschätzen und beobachteten Werten unterschieden wird. Bei einer größeren Anzahl fehlender Werte hat damit die Güte des Schätzverfahrens einen entscheidenden Einfluß auf die Ergebnisse der nachfolgenden Datenanalyseverfahren.

Eine andere Möglichkeit ist, Daten mit fehlenden Attributen bzw. Attribute, bei denen Daten fehlende Werte aufweisen, vor der Fuzzy-Clusteranalyse aus dem Datensatz zu entfernen. Das Entfernen von Daten mit fehlenden Werten bzw. Attributen, in denen fehlende Werte vorliegen, ist nur bei einem geringen Anteil fehlender Werte sinnvoll. Bei einem zu großen Anteil fehlender Werte besteht die Gefahr, daß der zu untersuchende Datensatz zu stark verkleinert wird und dadurch eventuell vorliegende Strukturen nicht mehr erkannt werden können.

Eine Alternative ist die Integration von Daten mit fehlenden Werten in das Datenanalyseverfahren. Aufgrund der Bedeutung des Umgangs mit fehlenden Werten bei der Fuzzy-Clusteranalyse wurde die Integration fehlender Werte in ein Fuzzy-Clusteringverfahren umfassend betrachtet. Es wurde aufgezeigt, daß bei einer größeren Anzahl von Daten mit fehlenden Attributwerten die Integration der Daten mit fehlenden Werten in die Fuzzy-Clusteranalyse dem Entfernen der Daten mit fehlenden Werten vor der Fuzzy-Clusteranalyse weit überlegen ist.

Die Integration von Daten mit fehlenden Werten in die Fuzzy-Clusteranalyse kann durch Schätzung der fehlenden Werte während der Fuzzy-Clusteranalyse oder durch Berücksichtigung nur der beobachteten Werte erfolgen. Die Schätzung kann z.B. durch den entsprechenden Attributwert des Clusterzentrums, zu dem ein Datum mit einem fehlenden Attributwert den höchsten Zugehörigkeitsgrad hat, oder durch das mit den Zugehörigkeitsgraden gewichtete Mittel der entsprechenden Attributwerte der Clusterzentren

erfolgen. Bei Fuzzy-Clusteringverfahren, bei denen auch die Kovarianzmatrix berechnet wird, ist die Kovarianzmatrix entsprechend zu korrigieren, da sonst die Varianzen und Kovarianzen unterschätzt werden. Die Schätzung durch den entsprechenden Attributwert der Clusterzentren führt nur zu einem geringfügig erhöhten Rechenaufwand. Alternativ können auch andere Schätzverfahren verwendet werden. Hierbei ist jedoch der zusätzliche Rechenbedarf im Auge zu behalten, da die Schätzung während jeder Iteration des Fuzzy-Clusteringverfahrens durchgeführt wird.

Die Methode des Schätzens fehlender Werte hat den Nachteil, daß die Prototypen der Cluster aus den beobachteten Attributwerten und den geschätzten Attributwerten der Cluster berechnet werden. Bei einer größeren Anzahl von fehlenden Werten zeigte sich das Verfahren abhängig von der Initialisierung.

Bei dem „available case“-Ansatz wird die Problematik des Schätzens während der Fuzzy-Clusteranalyse vermieden, indem die Clusterprototypen nur unter Berücksichtigung der beobachteten Werte berechnet werden. Für die Schätzung der Zugehörigkeitsgrade bietet es sich an, die Zugehörigkeitsgrade basierend auf den Abständen hinsichtlich der beobachteten Attribute zu berechnen, da bei der probabilistischen Fuzzy-Clusteranalyse die Zugehörigkeitsgrade auf der Relation der Abstände eines Datums zu den verschiedenen Clustern beruhen. Die Berechnung der Zugehörigkeitsgrade nach dieser Idee führt zu dem gleichen Ergebnis wie das Schätzen des Abstands durch die Berechnung der Abstände hinsichtlich der beobachten Attribute und der anschließenden Skalierung des Abstands auf die Zahl aller Attribute.

Bei der Fuzzy-Clusteranalyse des Brustkrebsdatensatzes mit 2 Clustern zeigte das Schätzen fehlender Werte durch den entsprechenden Attributwert des Clusterzentrums, zu dem das Datum den höchsten Zugehörigkeitsgrad besitzt, die besten Ergebnisse, während das Schätzen durch das mit den Zugehörigkeitsgraden gewichtete Mittel bei 2 Clustern zu wesentlich schlechteren Ergebnisse führte. Die Ergebnisse des „available case“-Ansatzes lagen dazwischen. Bei der Einteilung in drei Cluster führten alle drei Ansätze zu vergleichbar guten Ergebnissen. Auch bei einer Wahrscheinlichkeit für fehlende Werte „missing completely at random“ änderte sich die Anzahl der fehlklassifizierten Daten nicht wesentlich. Die Redundanz der Informationen durch die hohe Anzahl von neun Attributen in dem Datensatz konnte ausgenutzt werden.

Anhand des Weindatensatzes wurde das Verhalten der Verfahren bei einer geringeren Dimensionalität und nicht kreisförmigen Clustern betrachtet. Auch hier zeigte der „available case“-Ansatz ein gutes Verhalten. Dagegen

waren die Ergebnisse basierend auf der Schätzung durch den Mittelwert und durch das gewichtete Mittel der Clusterzentren nicht so gut. Trotz einer Korrektur der Kovarianzmatrix, um den Effekt der Schätzung auszugleichen, konnte der Ansatz basierend auf der Schätzung durch den entsprechenden Attributwert des Clusterzentrums mit dem höchsten Zugehörigkeitsgrad den Datensatz nur bis zu einer Wahrscheinlichkeit für fehlende Werte von 20% klassifizieren. (Die Ergebnisse waren schlechter als bei dem „available case"-Ansatz.) Es bestätigte sich, was sich bei dem Brustkrebsdatensatz mit einer Wahrscheinlichkeit für fehlende Werte von 50% schon angedeutet hatte. Der „available case"-Ansatz und der Ansatz basierenend auf der Schätzung fehlender Werte durch das gewichtete Mittel der Attributwerte der Clusterzentren können mit einer höheren Anzahl von fehlenden Werten umgehen als der Ansatz basierend auf einer Schätzung durch das Clusterzentrum mit dem höchsten Zugehörigkeitsgrad.

Bei dem Weindatensatz zeigte der Ansatz basierend auf der Schätzung fehlender Werte durch das gewichtete Mittel der Attributwerte der Clusterzentren ein schlechtes Klassifikationsverhalten. Ein möglicher Grund hierfür ist die Verwendung der Kovarianzmatrizen bei dem Gustafson–Kessel-Algorithmus. Der Ausdruck für die Schätzung fehlender Werte entspricht bei dem Gustafson–Kessel-Algorithmus dem bei dem Fuzzy-C-Means-Algorithmus. Eine einfache Korrektur der Kovarianzmatrix, um den Effekt des Schätzens auszugleichen, ist nicht bestimmbar, da nicht der Mittelwert (Clusterzentrum) der Daten *eines* Clusters sondern das gewichtete Mittel der Mittelwerte (Clusterzentren) aller Cluster verwendet wird. Bei Daten mit fehlenden Werten werden daher die Varianzen und Kovarianzen entweder unterschätzt (bei einem hohen Zugehörigkeitsgrad zu einem Cluster) oder die Varianzen und Kovarianzen werden durch die anderen Clusterzentren beeinflußt. Die in einem Cluster vorliegenden Varianzen und Kovarianzen werden dem gegenüber bei dem Umgang mit fehlenden Werten nicht berücksichtigt. Dieser Ansatz ist daher für flexiblere Fuzzy-Clusteringverfahren wie den Gustafson–Kessel-Algorithmus oder den FMLE nicht so geeignet.

Aufgrund der vorstehenden Betrachtungen ist bei Daten mit fehlenden Werten die Verwendung des „available case"-Ansatzes empfehlenswert. Bei *guten* Schätzverfahren, die die Eigenschaften des verwendeten Fuzzy-Clusteringverfahrens berücksichtigen (z.B. Verwendung der Kovarianzmatrix) ist auch der Ansatz des iterierten Schätzens von Interesse. Hierbei ist jedoch der zusätzliche Aufwand für das Schätzen, der bei dem „available case"-Ansatz nicht anfällt, zu berücksichtigen. Einfache Ansätze zum Schätzen, wie die Verwendung des Mittelwertes, stellen nur bei dem Fuzzy-

C-Means-Algorithmus eine Alternative dar.

Falls fehlende Werte mit einer clusterspezifischen Wahrscheinlichkeit auftreten, kann diese zusätzliche Information bei der Fuzzy-Clusteranalyse verwendet werden. Als Verfahren bietet sich hierfür der FMLE aufgrund seines wahrscheinlichkeitsbasierten Abstandsmaßes an. Die Parameter der Cluster werden dabei unter Berücksichtigung aller beobachteten Attributwerte berechnet („available case"-Ansatz). Durch eine zusätzliche Berücksichtigung der clusterspezifischen Wahrscheinlichkeit kann eine Verbesserung der Klassifikationsgüte erreicht werden. Da auch der „available case"-Ansatz schon zu guten Ergebnissen bei Daten mit fehlenden Werten führt, ist der Unterschied zwischen den beiden Ansätzen jedoch wesentlich geringer als der zwischen einer Berücksichtigung fehlender Werte und dem Entfernen fehlender Werte vor der Datenanalyse.

Entsprechend der menschlichen Intuition können auch bei der Fuzzy-Clusteranalyse Daten mit fehlenden Werten gegenüber Daten mit allen beobachteten Attributwerten geringer gewichtet werden. Eine signifikante Verbesserung der Klassifikationsergebnisse trat hierdurch jedoch nicht auf. Bei den durchgeführten Experimenten schien die Reduktion des Gewichts von Daten mit fehlenden Werten jedoch zu einem etwas stabileren Verhalten zu führen.

Als robustes Verfahren zum Umgang mit fehlenden Werten sollte bei der Fuzzy-Clusteranalyse der „available case"-Ansatz immer zur Verfügung stehen, um eine gute Möglichkeit zur Berücksichtigung fehlender Werte zu haben.

Kapitel 5

Fuzzy-Clusteranalyse mit klassifizierten Daten

5.1 Motivation

Der Begriff der Klasse ist von dem des Clusters zu unterscheiden. Ein Cluster ist eine Menge von homogenen Daten. Eine Klasse ist eine Menge von Daten mit einem gemeinsamen Attribut oder einer gemeinsamen Eigenschaft. Die Daten einer Klasse müssen nicht homogen sein. Eine Klasse kann daher aus mehreren Clustern bestehen.

Die Fuzzy-Clusteranalyse ist ein Verfahren zur Suche von Clustern in Daten. Üblicherweise werden dabei nicht klassifizierte Datensätze klassifiziert. Daneben kann sie jedoch auch verwendet werden, um nach Teilklassen gegebener Klassen zu suchen [27, 29]. Die ermittelten Teilklassen können z.B. verwendet werden, um die Güte von Klassifikatoren zu verbessern oder um z.B. Teilgruppen bzw. Teilklassen bei Marketingmaßnahmen gezielt ansprechen zu können.

Allgemein läßt sich eine Information über die Klassenangehörigkeit bzw. Klassenzugehörigkeit von Daten so interpretieren, daß Daten, die zu verschiedenen Klassen gehören, nicht demselben Cluster zugeordnet werden sollen. Aus der Klasseninformation der Daten läßt sich daher eine Klasseninformation der Cluster ableiten. Hierfür kann z.B. die mit den Zugehörigkeitsgraden gewichtete Majoritätsklasse der dem Cluster zugeordneten Daten verwendet werden.

In diesem Kapitel 5 wird betrachtet, wie sich Klasseninformationen bei

der Fuzzy-Clusteranalyse verwenden lassen. Dabei werden zuerst die naheliegenden allgemeinbekannten Ansätze betrachtet. Danach werden Ansätze zur Clusteranalyse teilklassifizierter Daten kurz vorgestellt. Anschließend werden drei neue Ansätze für die Integration der Klasseninformation in die Fuzzy-Clusteranalyse entwickelt und analysiert.

5.2 Einfache Möglichkeiten der Berücksichtigung von Klasseninformationen bei der Fuzzy-Clusteranalyse

In diesem Abschnitt 5.2 werden mehrere einfache Ansätze betrachtet, wie eine Information über die Klassenzugehörigkeit der zu betrachtenden Daten bei der Fuzzy-Clusteranalyse verwendet werden kann.

Eine Möglichkeit ist, den Datensatz nach der Klasseninformation in Teildatensätze zu zerlegen. Die Teildatensätze werden getrennt analysiert. Die Ergebnisse der Clusteranalyse werden anschließend kombiniert.

Der Nachteil dieser Vorgehensweise ist, daß bei der Fuzzy-Clusteranalyse Informationen über Cluster, die zu anderen Klassen gehören, nicht berücksichtigt werden.[1] Sofern die Klassen sich teilweise überlagern bzw. nicht gut separiert sind, kann dies zu einem suboptimalen Klassifikationsergebnis führen.

Eine andere Möglichkeit ist, die Information über die Klassenzugehörigkeit der Daten als weiteres Attribut der Daten aufzufassen. Die Anzahl der Attribute der Daten wird in diesem Fall um eins erhöht. Mit dieser vergrößerten Anzahl von Attributen wird die Fuzzy-Clusteranalyse normal durchgeführt. Die Betrachtung der Klasseninformation als zusätzliches Attribut ermöglicht es, auf die Aufteilung des Datensatzes in Klassen zu verzichten. Die Klasseninformation wird durch das verwendete Ähnlichkeits- bzw. Distanzmaß mitberücksichtigt.

Die Berechnung des Abstands eines Datums zu einem Cluster unter Verwendung des Klassenattributs ist jedoch problematisch. Bei der Klasseninformation handelt es sich in den meisten Fällen um ein symbolisches Attribut. Ohne zusätzliche Informationen sind daher die Abstände zwischen verschiedenen Klassen alle gleich. Es wird nur zwischen Gleichheit

[1] Probabilistische Fuzzy-Clusteringverfahren haben eine partitionierende Eigenschaft. Die Ursache ist die Restriktion $\sum_{j=1}^{n} u_{i,j} = 1, i \in \{1, \ldots, c\}$. Dicht benachbarte Cluster beeinflussen sich daher gegenseitig.

und Ungleichheit hinsichtlich dieses Attributs unterschieden. Dieses Attribut bzw. sein „Abstand“ ist in Relation zu den anderen Attributen bzw. deren Abständen zu dem Cluster zu gewichten.

Eine automatische Gewichtung, wie z.B. durch die Berechnung der Kovarianzmatrizen der Cluster und die Verwendung des Mahalanobis-Abstands beim Gustafson–Kessel-Algorithmus, ist problematisch, da das Klassenattribut ein symbolisches Attribut ist. Daneben ist sowohl bei einer automatischen als auch bei einer manuellen Gewichtung sicherzustellen, daß das Klassenattribut nicht zu dominant, aber auch nicht nur eins unter vielen ist.

Wenn das Klassenattribut zu dominant ist (es ist das einzige klassifikationsrelevante Merkmal), ist der Abstand zu Daten mit einer anderen Klasseninformation unabhängig von den weiteren Attributen immer hoch. Die Klassen werden daher jede für sich separat geclustert. Cluster, die zu verschiedenen Klassen gehören, beeinflussen sich gegenseitig nicht. Das Ergebnis entspricht daher dem einer Aufteilung des Datensatzes in Teilklassen.

Falls jedoch andererseits die Klasseninformation ein zu geringes Gewicht hat, kann es sein, daß sie nur eins unter vielen Attributen der Daten ist. In diesem Fall besteht die Gefahr, daß eine große Zahl von Attributen klassifikationsrelevanter ist als die Klassenzugehörigkeit und ggf. eine Heterogenität hinsichtlich des Klassenattributs toleriert wird.[2]

Ein weiterer Ansatz ist, die Klasseninformation als Gütekriterium zu verwenden. Die Idee ist hierbei, die Anzahl der Cluster solange zu erhöhen, bis in jedem Cluster die meisten Daten zu der gleichen Klasse gehören. Der Nachteil dieser Vorgehensweise besteht darin, daß die Clusterinformation zwar bei der Verifikation der Ergebnisse, nicht jedoch bei der Bestimmung der Cluster berücksichtigt wird. Daher tendiert dieser Ansatz dazu, eine größere Anzahl von Clustern mit einer geringen Anzahl von Daten zu erzeugen, wenn die Klassen nicht gut separiert sind.

5.3 Teilüberwachte Fuzzy-Clusteranalyse

Die Thematik der teilüberwachten Fuzzy-Clusteranalyse ist sehr nah mit der in diesem Kapitel 5 behandelten Thematik der Fuzzy-Clusteranalyse klassifizierter Daten verwandt. Bei der *teilüberwachten Fuzzy-Clusteranalyse*

[2]Sofern auch bei einer „guten“ Gewichtung der Klassenzugehörigkeit gegenüber anderen Attributen einige Cluster hinsichtlich der Klassenzugehörigkeit heterogen sind (es gibt Cluster, bei denen nicht nur ein geringer Teil der Daten zu einer anderen Klasse gehört), ist das Klassenattribut kritisch zu hinterfragen.

(semi supervised fuzzy cluster analysis) ist für einige der Daten vor der Fuzzy-Clusteranalyse bekannt, zu welchem Cluster diese Daten gehören sollen. Diese Information kann zur Verbesserung der Ergebnisse der Fuzzy-Clusteranalyse oder auch zum Benennen (Labeln) einzelner Cluster verwendet werden. Der Unterschied zur Fuzzy-Clusteranalyse klassifizierter Daten ist zum einen, daß nur einige Daten klassifiziert sind, und zum anderen, daß die Daten einem Cluster und nicht einer Klasse zugeordnet sind.

Für die teilüberwachte Fuzzy-Clusteranalyse gibt es verschiedene Ansätze [19, 100, 101, 11, 63]. Diese Ansätze werden im folgenden kurz vorgestellt und hinsichtlich ihrer Eignung für die Fuzzy-Clusteranalyse klassifizierter Daten betrachtet.

Den Ansätzen zur teilüberwachten Fuzzy-Clusteranalyse ist gemeinsam, daß die klassifizierten Daten die Fuzzy-Clusteranalyse „leiten“. Dies wird bei dem einen Ansatz [11, 19] erreicht, indem Daten mit einer Klasseninformation dem der betreffenden Klasse zugeordneten Cluster mit einem Zugehörigkeitsgrad von 1 zugeordnet werden.[3] Dieser Zugehörigkeitsgrad wird während der Fuzzy-Clusteranalyse *nicht* verändert. Nur für Daten ohne eine Klasseninformation werden die Zugehörigkeitsgrade wie üblich berechnet. Die Clusterprototypen werden wie üblich unter Verwendung aller Daten berechnet. Dabei werden Daten mit einer Klasseninformation höher gewichtet, wenn die Klasseninformation nur für wenige Daten vorliegt. Die Clusterzentren werden in diesem Fall durch

$$\vec{z}_i = \frac{\sum_{j=1}^{n_c} \left(w_j u_{i,j}^m \vec{x}_j^{(\text{class.})} \right) + \sum_{j=1}^{n_{nc}} \left(u_{i,j}^m \vec{x}_j^{(\text{notclass.})} \right)}{\sum_{j=1}^{n_c} w_j u_{i,j}^m + \sum_{j=1}^{n_{nc}} u_{i,j}^m} \tag{5.1}$$

berechnet. w_j ist das Gewicht des klassifizierten Datums $\vec{x}_j^{(\text{class.})}$. n_c ist die Anzahl der klassifizierten Daten und n_{nc} ist die Anzahl der nicht klassifizierten Daten $\vec{x}_j^{(\text{notclass.})}$.

Die Initialisierung des Verfahrens, d.h. die Berechnung der Clusterzentren in der ersten Iteration, erfolgt *nur* unter Verwendung der klassifizierten Daten.

Der Nachteil dieser Vorgehensweise ist, daß die Klasseninformation als „absolut“ angesehen wird. Auch wenn die Berechnung der Clusterprototypen auf eine andere Klassenzugehörigkeit der klassifizierten Daten hinweist, wird der Zugehörigkeitsgrad der klassifizierten Daten nicht geändert.

Das Verfahren ist geeignet für die Benennung von Clustern anhand von

[3] Es wird von einer clusterspezifischen Zuordnung ausgegangen.

Beispieldatensätzen sowie für die Erkennung von Clustern mit einer stark variierenden Größe.[4]

Das Verfahren ist *nicht* geeignet, wenn für fast alle Daten bekannt ist, zu welchem Cluster sie gehören, da die Zugehörigkeitsgrade während der Fuzzy-Clusteranalyse bei diesen Daten nicht geändert werden.

Ein anderer Ansatz basiert auf der Idee, die Zuordnung eines klassifizierten Datums zu einem „falschen“ Cluster zu bestrafen. Im Gegensatz zu dem vorhergehend vorgestellten Ansatz können klassifizierte Daten jedoch anderen Clustern zugeordnet werden [100].

Der „Strafterm“ für die Zuordnung eines klassifizierten Datums $\vec{x}_j$ zu einem anderen Cluster ist

$$\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{i,j}-b_j f_{i,j})^m d^2(\vec{\beta}_i,\vec{x}_j). \tag{5.2}$$

b_j ist 1, wenn ein Datum klassifiziert ist und 0 sonst. Die Klassenzugehörigkeit ist in der $c \times n$ Matrix $\mathbf{F} = [f_{i,j}]_{c\times n}$ gespeichert. $f_{i,j}$ ist 1, wenn das Datum $\vec{x}_j$ dem Cluster $\vec{\beta}_i$ durch die vorgegebene Klassifikation zugeordnet wurde.

Wenn der Zugehörigkeitsgrad $u_{i,j}$ eines Datums $\vec{x}_j$ zu einem Cluster $\vec{\beta}_i$ der vorgegebenen Klassifikation entspricht, ist $(u_{i,j} - b_j f_{i,j}) = 0$. Eine Abweichung von der vorgegebenen Klassifikation führt zu einer Erhöhung des Ausdrucks.[5]

Die Verwendung des Strafterms (5.2) führt zu der Zielfunktion [100, 19]

$$J(\mathbf{X},\mathbf{U},\mathbf{B}) = \alpha\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{i,j}-b_j f_{i,j})^m d^2(\vec{\beta}_i,\vec{x}_j) + \sum_{i=1}^{c}\sum_{j=1}^{n}u_{i,j}^m d^2(\vec{\beta}_i,\vec{x}_j). \tag{5.3}$$

α ist ein Wichtungsfaktor.

[4] Die Initialisierung der Clusterprototypen erfolgt nur mittels der klassifizierten Daten. Hierdurch werden zu Beginn alle Cluster, denen Daten manuell zugeordnet wurden, erkannt. Bei einer hinreichenden Separierung der Cluster und durch die Gewichtung der klassifizierten Daten werden auch kleine Cluster in den nachfolgenden Iterationen nicht „aufgegeben“. (Das Clusterzentrum „wandert“ nicht woanders hin.)

[5] Bei der probabilistischen Fuzzy-Clusteranalyse gilt meist $m = 2$. Ansonsten ist ggf. hilfsweise $|u_{i,j} - b_j f_{i,j}|^m$ zu verwenden.

Die Zugehörigkeitsgrade werden berechnet durch [101, 19]

$$u_{i,j} = \frac{1}{1+\alpha^{\frac{1}{m-1}}} \left(\frac{1+\alpha^{\frac{1}{m-1}} \left(1 - b_k \sum_{k=1}^{c} f_{k,j}\right)}{\sum_{k=1}^{c} \left(\frac{d^2(\vec{\beta}_i, \vec{x}_j)}{d^2(\vec{\beta}_k, \vec{x}_j)}\right)^{\frac{2}{m_1}}} + \alpha^{\frac{1}{m-1}} b_j f_{i,j} \right). \quad (5.4)$$

Dieser Ansatz kann von teilklassifizierten Datensätzen auf klassifizierte Datensätze übertragen werden. Der Index b_j ist dabei für alle Daten $\vec{x}_j$ 1. Damit jedoch eine Zuordnung zu *Klassen* und nicht zu *Clustern* bei der Fuzzy-Clusteranalyse berücksichtigt werden kann, ist eine Modifikation des Strafterms erforderlich.

Sofern eine Klasse nur aus einem Cluster besteht oder wenn man weiß, zu welchem Cluster einer Klasse ein Datum gehört, können die o.g. teilüberwachten Verfahren auch für die Analyse teilklassifizierter Verfahren eingesetzt werden. Falls jedoch nur bekannt ist, zu welcher Klasse ein Datum gehört, nicht jedoch zu welchem Cluster, sind die teilüberwachten Verfahren nicht geeignet.

5.4 Ein zielfunktionsbasierter Ansatz

In diesem Abschnitt 5.4 wird ein zielfunktionsbasierter Ansatz für die Fuzzy-Clusteranalyse klassifizierter Daten vorgestellt. Dieser Ansatz ähnelt dem o.g. teilüberwachten Ansatz 5.3.

Das Kennzeichen einer Fehlklassifikation eines Datums ist, daß die Klasseninformation des Datums von der des Clusters, dem es zugeordnet ist, abweicht. Im Idealfall sollte kein Datum einem Cluster mit der „falschen“ Klasseninformation zugeordnet sein. Ein Kennzeichen für die Fehlklassifikation eines Datensatzes ist daher

$$\sum_{i=1}^{c} \sum_{j=1}^{n} (1 - \text{class}(i,j))\, u_{i,j}^{m}. \quad (5.5)$$

class(i,j) ist 1,wenn der Cluster $\vec{\beta}_i$ und das Datum $\vec{x}_j$ der gleichen Klasse angehören. Sonst ist class(i,j) 0.

Die Integration der Forderung, möglichst wenige Daten falsch zu klassifizieren, führt zu der Zielfunktion

$$J(\mathbf{X}, \mathbf{U}, \mathbf{B}) = \sum_{j=1}^{n} \sum_{i=1}^{c} u_{i,j}^{m} \cdot d^2(\vec{\beta}_i, \vec{x}_j) + \alpha \sum_{i=1}^{c} \sum_{j=1}^{n} (1 - \text{class}(i,j))\, u_{i,j}^{m}. \quad (5.6)$$

α gewichtet die Forderung, Fehlklassifikationen zu vermeiden, gegen die Forderung, daß der Abstand zwischen den Clustern und den ihnen zugeordneten Daten möglichst klein sein soll.

Als notwendiges Kriterium für die Optimierung der Zielfunktion sind bei der probabilistischen Fuzzy-Clusteranalyse die Zugehörigkeitsgrade durch

$$u_{i,j} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d^2(\vec{\beta}_i, \vec{x}_j) + \alpha(1 - \text{class}(i,j))}{d^2(\vec{\beta}_k, \vec{x}_j) + \alpha(1 - \text{class}(k,j))} \right)^{\frac{1}{m-1}}} \tag{5.7}$$

zu bestimmen. Dies wird im folgenden gezeigt.

Bei der probabilistischen Fuzzy-Clusteranalyse ist die Zielfunktion (5.6) unter Berücksichtigung der Restriktion $\sum_{i=1}^{c} u_{i,j} = 1$ für alle Daten $\vec{x}_j$ zu minimieren. Unter Verwendung von n Lagrangeschen Multiplikatoren λ_j wird die Zielfunktion modifiziert zu

$$\begin{aligned} J(\mathbf{X}, \mathbf{U}, \mathbf{B}) &= \sum_{j=1}^{n} \sum_{i=1}^{c} u_{i,j}^m \cdot d^2(\vec{\beta}_i, \vec{x}_j) \\ &\quad + \alpha \sum_{i=1}^{c} \sum_{j=1}^{n} (1 - \text{class}(i,j))\, u_{i,j}^m \\ &\quad - \sum_{j=1}^{n} \lambda_j \left(\sum_{i=1}^{c} u_{i,j} - 1 \right). \end{aligned} \tag{5.8}$$

Für die Bestimmung der Zugehörigkeitsgrade muß $\frac{\partial}{\partial u_{i,j}} J(\mathbf{X}, \mathbf{U}, \mathbf{B}) = 0$ gelten.

$$\begin{aligned} 0 &= \frac{\partial}{\partial u_{i,j}} J(\mathbf{X}, \mathbf{U}, \mathbf{B}) \\ &= \frac{\partial}{\partial u_{i,j}} \Big(\sum_{j=1}^{n} \sum_{i=1}^{c} u_{i,j}^m \cdot d^2(\vec{\beta}_i, \vec{x}_j) \\ &\quad + \alpha \sum_{i=1}^{c} \sum_{j=1}^{n} (1 - \text{class}(i,j))\, u_{i,j}^m \\ &\quad - \sum_{j=1}^{n} \lambda_j \left(\sum_{i=1}^{c} u_{i,j} - 1 \right) \Big) \\ &= m \cdot u_{i,j}^{m-1} \cdot d^2(\vec{\beta}_i, \vec{x}_j) + \alpha \cdot m \cdot u_{i,j}^{m-1} (1 - \text{class}(i,j)) - \lambda_j \end{aligned}$$

$$\Rightarrow \quad u_{i,j} = \left(\frac{\lambda_j}{m\left(d^2(\vec{\beta}_i, \vec{x}_j) + \alpha(1 - \text{class}(i,j))\right)} \right)^{\frac{1}{m-1}}$$

$$\begin{aligned} 1 &= \sum_{i=1}^{c} u_{i,j} \\ &= \lambda_j^{\frac{1}{m-1}} \left(\sum_{i=1}^{c} \left(\frac{1}{m\left(d^2(\vec{\beta}_i, \vec{x}_j) + \alpha(1 - \text{class}(i,j))\right)} \right)^{\frac{1}{m-1}} \right) \end{aligned}$$

$$\Rightarrow \quad \lambda_j = \frac{1}{\sum_{i=1}^{c} \left(\frac{1}{m\left(d^2(\vec{\beta}_i, \vec{x}_j) + \alpha(1 - \text{class}(i,j))\right)} \right)^{\frac{1}{m-1}}}$$

$$\Rightarrow \quad u_{i,j} \quad = \quad \frac{1}{\sum_{k=1}^{c} \left(\frac{d^2(\vec{\beta}_i, \vec{x}_j) + \alpha(1 - \text{class}(i,j))}{d^2(\vec{\beta}_k, \vec{x}_j) + \alpha(1 - \text{class}(k,j))} \right)^{\frac{1}{m-1}}}$$

Durch den Strafterm für die Klassenzuordnung wird der Zugehörigkeitsgrad zu Clustern mit einer anderen Klasseninformation reduziert, während der Zugehörigkeitsgrad von Daten mit der „richtigen“ Klasseninformation erhöht wird.

Die Berechnung der Clusterprototypen erfolgt ohne die Berücksichtigung einer Klasseninformation, da die Ableitung des Strafterms für fehlklassifizierte Daten $\sum_{i=1}^{c} \sum_{j=1}^{n} (1 - \text{class}(i,j))\, u_{i,j}^m$ nach den Clusterprototypen 0 ist.

Das in diesem Abschnitt 5.4 vorgestellte zielfunktionsbasierte Verfahren ähnelt dem in Abschnitt 5.3 vorgestellten Verfahren. Der Unterschied ist, daß anstelle der Zugehörigkeit zu einem Cluster die Zugehörigkeit zu einer Klasse, d.h. zu einer Menge von Clustern, verwendet wird. Der Strafterm (5.2) des zielfunktionsbasierten Verfahrens aus Abschnitt 5.3 besagt, daß bei der Zuordnung eines Datums $\vec{x}_j$ zu einem „falschen“ Cluster $\vec{\beta}_i$ mit dem Zugehörigkeitsgrad $u_{i,j}$ der Wert der Zielfunktion um $u_{i,j}^m$ und bei der Zuordnung zu dem „richtigen“ Cluster um $(|u_{i,j} - 1|)^m$ erhöht wird. Demgegenüber wird nur bei der Zuordnung eines Datum $\vec{x}_j$ zu einem Cluster $\vec{\beta}_i$ mit einer anderen Klasseninformation der Zugehörigkeitsgrad um $u_{i,j}^m$ erhöht. Bei einer Zuordnung zu einem Cluster mit der gleichen Klasseninformation bleibt der Wert der Zielfunktion unverändert. Daher hat bei dem in diesem Abschnitt 5.4 vorgestellten Ansatz der Strafterm in Relation zu dem Ausdruck, der die Minimierung der Abstände zwischen den Daten und den ihnen zugeordneten Clustern fordert, ein etwas geringeres Gewicht.

5.5 Zwei intuitive Ansätze basierend auf der Abstoßung fremder Klassen

Bei dem zielfunktionsbasierten Ansatz zur Fuzzy-Clusteranalyse klassifizierter Daten führt die Zuordnung eines Datums zu einem Cluster mit einer anderen Klasseninformation zu einer Reduktion des Zugehörigkeitsgrads des Datums zu diesem Cluster. Eine einfache Reduktion des Zugehörigkeitsgrads, bis hin zu einem Zugehörigkeitsgrad von 0, führt jedoch zu der Problematik, daß das Verfahren dazu tendiert, die Daten nach Klassen ge-

trennt zu klassifizieren, wenn alle Daten klassifiziert sind. Die partitionierende Wirkung der probabilistischen Fuzzy-Clusteranalyse wird abgeschwächt, da Daten, die dicht bei einem Cluster liegen, zu diesem keinen hohen Zugehörigkeitsgrad aufweisen müssen. Sie können anderen Clustern zugeordnet werden und „ziehen“ diese damit an.

Eine Alternative zu einer Reduktion des Zugehörigkeitsgrads ist, die Abstoßung der Cluster durch Daten mit einer anderen Klasseninformation direkt zu modellieren. Hierfür bietet sich das Modell des „Alternating Cluster Estimation“ an. Die in diesem Abschnitt 5.5 vorgestellten beiden Ansätze folgen der menschlichen Intuition, daß Daten mit der gleichen Klasseninformation eine positive Anziehungskraft und Daten mit einer anderen Klasseninformation eine negative Anziehungskraft besitzen sollten [126].

Ein naheliegender Ansatz ist, die Zugehörigkeitsgrade der Daten zu den Clustern von dem Intervall $[0, 1]$ auf das Intervall $[-1, 1]$ zu erweitern, indem man bei Daten mit einer anderen Klasseninformation als der des Clusters den Zugehörigkeitsgrad $u_{i,j}^m$ mit -1 multipliziert. Dies führt bei den Ausdrücken zur Berechnung der Clusterprototypen zu der Verwendung von $|u_{i,j}|^m \cdot \text{sgn}(u_{i,j})$ anstelle von $u_{i,j}^m$. Die Verwendung eines negativen Zugehörigkeitsgrads für ein Datum $\vec{x}_j$ kann als „Reduktion“ der Anziehung in Richtung des Datums $\vec{x}_j$ betrachtet werden. Durch die Reduktion der Anziehung „wandert“ der Cluster weg.

Die Stärke der Abstoßung zwischen Daten und Clustern mit einer unterschiedlichen Klasseninformation kann erhöht werden, indem der Zugehörigkeitsgrad $u_{i,j}$ eines Datums $\vec{x}_j$ zu einem Cluster $\vec{\beta}_i$ mit einem Wichtungsfaktor α multipliziert wird, wenn $\vec{x}_j$ und $\vec{\beta}_i$ eine unterschiedliche Klasseninformation besitzen.

Aus statistischer Sicht ist die Verwendung negativer Zugehörigkeitsgrade für die Berechnung des Mittelwerts bzw. der Kovarianzen nicht korrekt. Die Vorgehensweise kann jedoch als Heuristik interpretiert werden, mittels derer die Anziehung der Cluster aus einer Region des Datenraums reduziert wird. Es ist sicherzustellen, daß bei der Verwendung negativer Zugehörigkeitsgrade „sinnvolle“ Clusterprototypen berechnet werden. Es dürfen nicht zu viele Daten einem Cluster mit negativen Zugehörigkeitsgraden zugeordnet werden. Die Verwendung negativer Zugehörigkeitsgrade beeinflußt sowohl die Berechnung der Clusterzentren als auch die der Kovarianzmatrix.

Die Stärke der Abstoßung zwischen Daten und Clustern mit einer unterschiedlichen Klasseninformation kann erhöht werden, in dem der Zugehörigkeitsgrad $u_{i,j}$ eines Datums $\vec{x}_j$ zu einem Cluster $\vec{\beta}_i$ mit einem Wichtungsfaktor α multipliziert wird, wenn $\vec{x}_j$ und $\vec{\beta}_i$ eine unterschiedliche Klassen-

information besitzen.

Die Problematik, daß die Summe der Zugehörigkeitsgrade echt größer 0 sein muß, läßt sich vermeiden, wenn man die Abstoßung durch ein Datum $\vec{x}_j$ nicht durch negative Zugehörigkeitsgrade, sondern durch eine Anziehung des Clusters $\vec{\beta}_i$ aus der entgegengesetzten Richtung mittels eines fiktiven Datums $\vec{x}'_{i,j}$ modelliert. Das fiktive Datum $\vec{x}'_{i,j}$ kann durch Spiegelung des Datums $\vec{x}_j$ an dem Clusterzentrum $\vec{\beta}_i$ bestimmt werden:

$$\vec{x}'_{i,j} = \vec{z}_i - (\vec{x}_j - \vec{z}_i). \tag{5.9}$$

Da die Stärke der Abstoßung von dem Ausmaß der Fehlklassifikation, d.h. von dem Zugehörigkeitsgrad des Datums $\vec{x}_j$ zu einem Cluster mit einer anderen Klasseninformation abhängt, sollte der Zugehörigkeitsgrad des Datums $\vec{x}'_{i,j}$ zu dem Cluster $\vec{\beta}_i$ dem Zugehörigkeitsgrad des Datums $\vec{x}_j$ zu diesem Cluster entsprechen. Nach diesem Modell werden die Clusterzentren berechnet durch:

$$\vec{z}_i = \frac{1}{\sum_{j=1}^{n}(u_{i,j})^m} \sum_{j=1}^{n}(u_{i,j})^m \vec{x}'_{i,j}. \tag{5.10}$$

$\vec{x'_{i,j}}$ ist definiert durch

$$\vec{x}'_{i,j} = \begin{cases} \vec{x}_j, & \text{falls } \vec{x}_j \text{ und } \vec{\beta}_i \text{ zu der gleichen Klasse gehören.} \\ \vec{z}_i - (\vec{x}_j - \vec{z}_i), & \text{falls } \vec{x}_j \text{ und } \vec{\beta}_i \text{ zu verschiedenen Klassen gehören.} \end{cases} \tag{5.11}$$

Da der Zugehörigkeitsgrad $u_{i,j}$ eines Datums $\vec{x}_j$ zu einem Cluster $\vec{\beta}_i$ bei Daten mit einer anderen Klasseninformation die Stärke der Abstoßung angibt, kann die Abstoßung verstärkt werden, indem bei einer unterschiedlichen Klasseninformation zwischen einem Datum und einem Cluster der Zugehörigkeitsgrad mit einem Wichtungsfaktor α multipliziert wird.

Im Gegensatz zu der Verwendung negativer Zugehörigkeitsgrade führt die Abstoßung durch Anziehung zu keiner Veränderung der Kovarianzmatrizen der Cluster.

5.6 Vergleich und Bewertung der Verfahren

Die Eigenschaften der verschiedenen Verfahren zur Fuzzy-Clusteranalyse klassifizierter Daten werden anhand des in Abb. 5.1 dargestellten Datensatzes näher betrachtet. Der Datensatz besteht aus 150 Daten. Die Daten

gehören zu zwei Klassen mit 100 Daten (blau dargestellt) und 50 Daten (grün dargestellt). Die blau dargestellten Daten wurden durch zwei Normalverteilungen mit je 50 Daten und die grün dargestellten Daten durch eine Normalverteilung erzeugt. Abb. 5.2 zeigt eine Fuzzy-Clusteranalyse dieses Datensatzes mit dem Gustafson–Kessel-Algorithmus ohne Verwendung der Klasseninformation. Die blau dargestellten Daten der einen Klasse werden durch den roten und den blauen Cluster, die grün dargestellten Daten der anderen Klasse werden durch den grünen Cluster beschrieben. Da bei der Clusteranalyse keine Klasseninformation verwendet wird, werden dem grünen Cluster auch Daten aus dem Bereich zugeordnet, in dem die Klassen sich überlappen. Aufgrund der partitionierenden Eigenschaft der probabilistischen Fuzzy-Clusteranalyse werden der rote und der blaue Cluster „zur Seite gedrückt“. Der grüne Cluster „wandert“ leicht nach oben.

Eine einfache Möglichkeit, eine Klasseninformation zu berücksichtigen, ist die Clusteranalyse der Daten getrennt nach Klassen. Abb. 5.3 zeigt die Klassifikation nach dem Zusammenfügen der Cluster. Es wurde nicht erkannt, daß die beiden blau und rot markierten Cluster symmetrisch sind.

Abb. 5.4 zeigt das Ergebnis einer Fuzzy-Clusteranalyse, bei der Daten mit der falschen Klasseninformation bei der Berechnung der Clusterprototypen nicht berücksichtigt werden. Hierdurch wird verhindert, daß durch die Zuordnung von Daten mit einer anderen Klasseninformation ein Cluster „wegwandert“. Der grüne Cluster liegt bei diesem Verfahren tiefer als bei der Nichtberücksichtigung der Klasseninformation, vgl. Abb. 5.2. Gleichzeitig werden der rote und der blaue Cluster nicht „zur Seite gedrückt“. Die blau markierten Daten aus dem Überlappungsbereich werden dem grünen Cluster zwar teilweise zugeordnet, d.h. sie haben hinsichtlich der Berechnung des roten und des blauen Clusters ein geringeres Gewicht. Da sie jedoch bei der Berechnung des grünen Clusters nicht berücksichtigt werden, wird der grüne Cluster nicht breiter.

Eine Fuzzy-Clusteranalyse, bei der Daten mit der falschen Klasseninformation bei der Berechnung der Clusterprototypen nicht berücksichtigt werden, kann auch als Fuzzy-Clusteranalyse mit einer mit 0 gewichteten Abstoßung verstanden werden. Die Abbildungen 5.5, 5.6, 5.7, 5.8, 5.9 und 5.10 zeigen die in Abschnitt 5.5 vorgestellten Ansätze zur Berücksichtigung einer Klasseninformation. Die Abstoßung ist dabei unterschiedlich stark gewichtet. Bei den in den Abbildungen 5.5 und 5.6 dargestellten Klassifikationen wird die Klasseninformation der Daten berücksichtigt, indem Daten mit einer „falschen“ Klasseninformation Cluster mittels eines fiktiven Datums abstoßen. Dies führt zu einer Verschiebung des Clusterzentrums. Da die fiktiven Daten durch Spiegelung am Clusterzentrum erzeugt werden,

werden Daten mit einer „falschen“ Klasseninformation weiterhin bei der Berechnung der Kovarianzmatrix berücksichtigt. Die Abbildungen 5.7 und 5.8 zeigen die Ergebnisse dieses Verfahrens, wenn Daten mit einer „falschen“ Klasseninformation nur bei der Berechnung der Clusterzentren, nicht jedoch bei der Berechnung der Kovarianzmatrix berücksichtigt werden. Eine andere Möglichkeit, eine Abstoßung zwischen Daten und Clustern mit einer unterschiedlichen Klasseninformation zu modellieren, ist die Verwendung negativer Zugehörigkeitsgrade. Die Abbildungen 5.9 und 5.10 zeigen die Ergebnisse dieses Ansatzes. Bei der Gewichtung der Abstoßung ist darauf zu achten, daß die so berechneten Clusterzentren und Kovarianzmatrizen interpretierbar bleiben. Bei einer zu starken Gewichtung ist eine Invertierbarkeit der Kovarianzmatrizen oft nicht mehr möglich. Diese Problematik spricht für die Modellierung der Abstoßung durch Anziehung durch ein fiktives Datum.

Im Gegensatz zu den in Abschnitt 5.5 vorgestellten Ansätzen wird die Abstoßung der Cluster durch Daten mit einer anderen Klasseninformation bei dem in Abschnitt 5.4 vorgestellten Ansatz nicht direkt bei der Berechnung der Prototypen modelliert. Stattdessen wird die Zuordnung von Daten zu Clustern mit einer anderen Klasseninformation durch einen Strafterm in der das Klassifikationsproblem beschreibenden Zielfunktion modelliert. Die Ableitung dieses Ansatzes führt zu einer geänderten Berechnung der Zugehörigkeitsgrade. Wenn die Klassenzugehörigkeit eines Datums mit der eines Clusters nicht übereinstimmt, wird der Abstand zwischen dem Datum und diesem Cluster erhöht. Dies führt zu einer Verringerung des Zugehörigkeitsgrads zu Clustern mit einer anderen Klassenzugehörigkeit und zu einer Erhöhung des Zugehörigkeitsgrads zu Clustern mit der gleichen Klasseninformation. Eine Abstoßung eines Clusters mit einer anderen Klasseninformation erfolgt jedoch nur indirekt durch die Anziehung eines anderen Clusters. Ein Datum übt auch auf Cluster mit einer anderen Klasseninformation stets eine positive Anziehung aus. Der Effekt der Abstoßung ist daher etwas geringer als bei der direkten Modellierung. Die Abbildungen 5.11 und 5.12 zeigen die Ergebnisse dieses Ansatzes.

Für die Fuzzy-Clusteranalyse mit klassifizierten Daten ist sowohl der Ansatz basierend auf Abstoßung mittels eines fiktiven Datums als auch der zielfunktionsbasierte Ansatz geeignet. Für den zielfunktionsbasierten Ansatz spricht seine mathematisch saubere Modellierung, während der Ansatz basierend auf Abstoßung mittels eines fiktiven Datums eine stärkere Berücksichtigung der Klasseninformation ermöglicht.

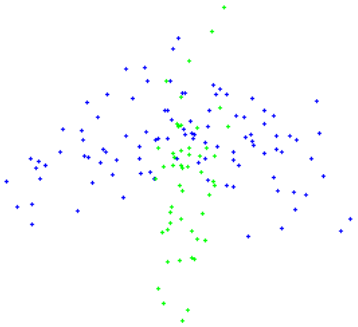

Abbildung 5.1: Datensatz mit zwei Klassen.

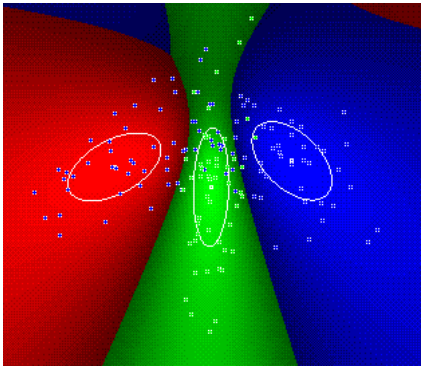

Abbildung 5.2: Fuzzy-Clusteranalyse mit dem Gustafson–Kessel-Algorithmus *ohne Klasseninformation*.

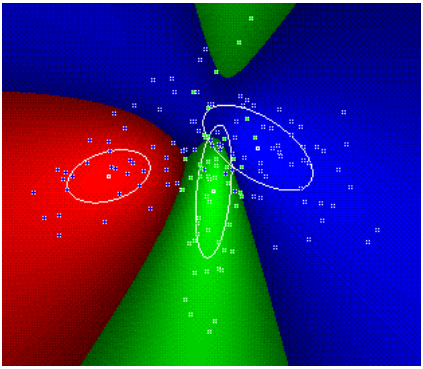

Abbildung 5.3: Fuzzy-Clusteranalyse mit dem Gustafson–Kessel-Algorithmus. Die Clusteranalyse erfolgte *für jede Klasse separat*. Die Ergebnisse wurden kombiniert.

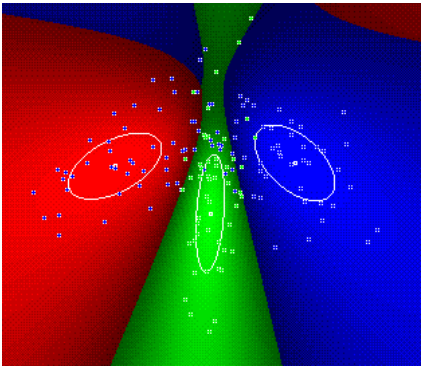

Abbildung 5.4: Fuzzy-Clusteranalyse mit dem Gustafson–Kessel-Algorithmus. *Daten mit falscher Klasseninformation werden nicht berücksichtigt.*

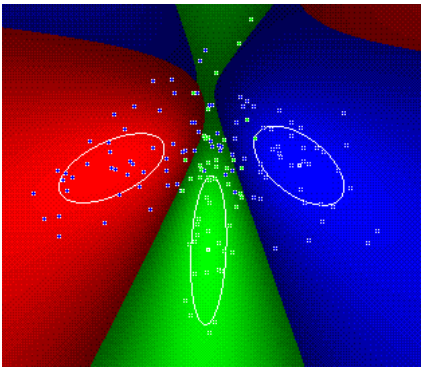

Abbildung 5.5: Fuzzy-Clusteranalyse mit dem Gustafson–Kessel-Algorithmus. Bei falscher Klasseninformation erfolgt eine *Abstoßung durch Verwendung eines fiktiven Datums.* $\alpha = 1$.

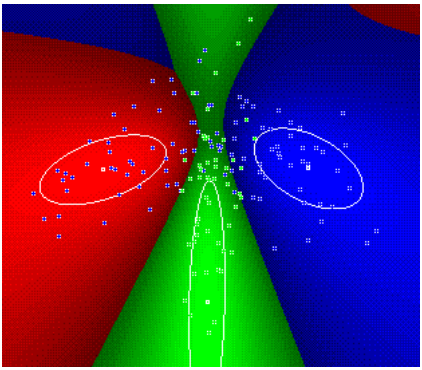

Abbildung 5.6: Fuzzy-Clusteranalyse mit dem Gustafson–Kessel-Algorithmus. Bei falscher Klasseninformation erfolgt eine *Abstoßung durch Verwendung eines fiktiven Datums.* $\alpha = 2$.

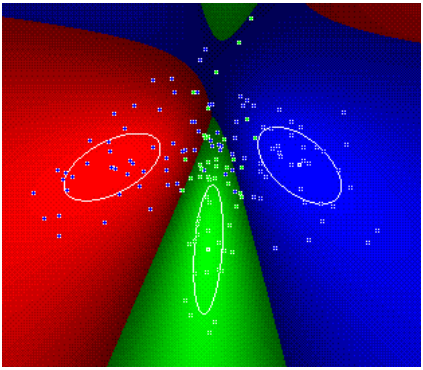

Abbildung 5.7: Fuzzy-Clusteranalyse mit dem Gustafson–Kessel-Algorithmus. Bei falscher Klasseninformation erfolgt eine *Abstoßung durch Verwendung eines fiktiven Datums.* $\alpha = 1$, *Berechnung der Kovarianzmatrix ohne Daten mit falscher Klasseninformation.*

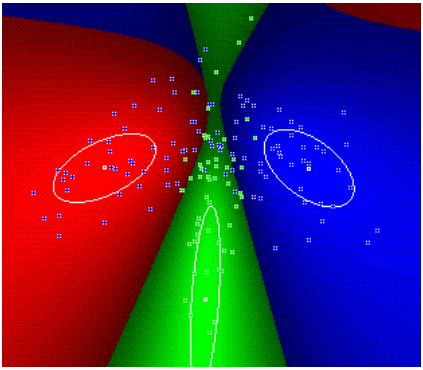

Abbildung 5.8: Fuzzy-Clusteranalyse mit dem Gustafson–Kessel-Algorithmus. Bei falscher Klasseninformation erfolgt eine *Abstoßung durch Verwendung eines fiktiven Datums.* $\alpha = 2$, *Berechnung der Kovarianzmatrix ohne Daten mit falscher Klasseninformation.*

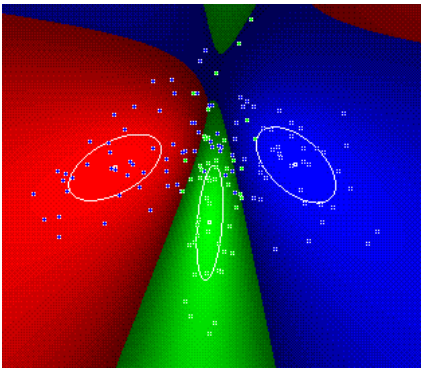

Abbildung 5.9: Fuzzy-Clusteranalyse mit dem Gustafson–Kessel-Algorithmus. Bei falscher Klasseninformation werden *negative Zugehörigkeitsgrade* verwendet. $\alpha = 0.25$.

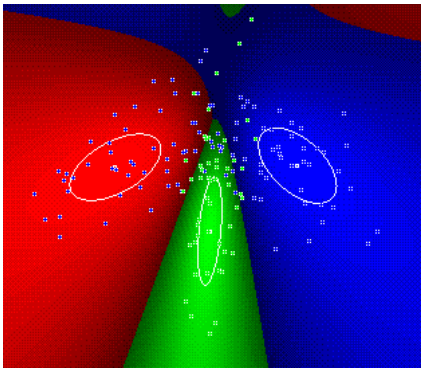

Abbildung 5.10: Fuzzy-Clusteranalyse mit dem Gustafson–Kessel-Algorithmus. Bei falscher Klasseninformation werden *negative Zugehörigkeitsgrade* verwendet. $\alpha = 0.5$.

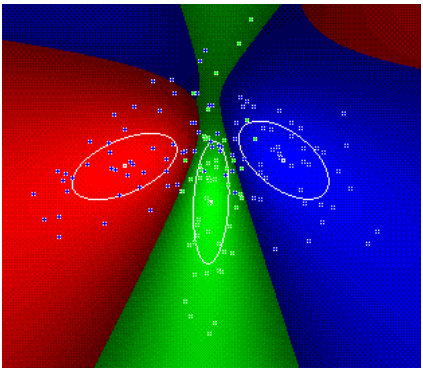

Abbildung 5.11: *Zielfunktionsbasierte* Fuzzy-Clusteranalyse mit dem Gustafson–Kessel-Algorithmus mit Berücksichtigung der Klasseninformation. $\alpha = 4$.

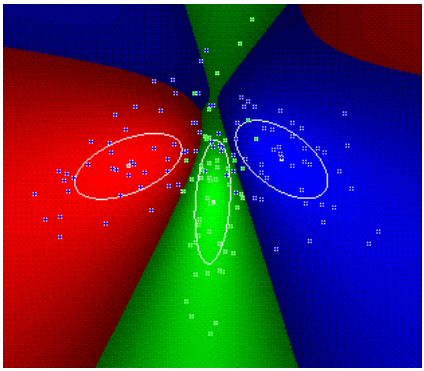

Abbildung 5.12: *Zielfunktionsbasierte* Fuzzy-Clusteranalyse mit dem Gustafson–Kessel-Algorithmus mit Berücksichtigung der Klasseninformation. $\alpha = 8$.

5.7 Verwendung der neuen Ansätze bei der teilüberwachten Fuzzy-Clusteranalyse

Die in diesem Kapitel 5 vorgestellten Verfahren zur Fuzzy-Clusteranalyse klassifizierter Daten können auch für die teilüberwachte Fuzzy-Clusteranalyse verwandt werden. Bei der teilüberwachten Fuzzy-Clusteranalyse ist nur bei einigen Daten die Klasseninformation bekannt. Ein Anwendungsgebiet der teilüberwachten Fuzzy-Clusteranalyse ist z.B. das „Labeling“. Anhand der klassifizierten Daten werden die Cluster identifiziert bzw. benannt.

Die in Abschnitt 5.5 vorgestellten Verfahren basieren auf einer „Bestrafung“ der Zuordnung eines Datums zu einem Cluster mit einer anderen Klasseninformation. Wenn bei Daten ohne Klasseninformation immer angenommen wird, daß die Klasseninformation des Datums mit der des Clusters übereinstimmt, ist die Klasseninformation nur bei klassifizierten Daten von Belang. Falls nur bei einem geringen Teil der Daten eine Klasseninformation zur Verfügung steht, ist eine Erhöhung des Gewichts dieser Daten erforderlich, damit die Klasseninformation bei der Berechnung der Clusterprototypen eine relevante Rolle spielt.

Der Vorteil der in den Abschnitten 5.4 und 5.5 vorgestellten neuen Verfahren gegenüber den in Abschnitt 5.3 vorgestellten Verfahren ist, daß bei der teilüberwachten Fuzzy-Clusteranalyse auch Klassen, die aus mehreren Clustern bestehen, berücksichtigt werden können, ohne daß vorher bekannt sein muß, ob die „gelabelten“ Vertreter derselben Klasse auch demselben Cluster angehören. Damit sind die in diesem Kapitel 5 vorgestellten Verfahren zur Fuzzy-Clusteranalyse klassifizierter Daten eine Erweiterung der in Abschnitt 5.3 vorgestellten teilüberwachten Verfahren.

Kapitel 6

Fazit

Die Zielstellung der Clusteranalyse, Strukturen bzw. Cluster in Daten zu finden, ist bei der Datenanalyse von großem Interesse. Häufig sind die in den Datensätzen vorliegenden Cluster nicht deutlich separiert. Viele Daten sind Mischformen bzw. Hybride verschiedener Cluster. Die Fuzzy-Clusteranalyse bietet eine Möglichkeit, diese Daten bei der Clusteranalyse entsprechend zu berücksichtigen. Hierbei werden die Daten den Clustern mit einem Zugehörigkeitsgrad zwischen 0 und 1 zugeordnet.

Im ersten Teil der Arbeit erfolgte eine *Einführung in die Datenanalyse mit Fuzzy-Clusteringverfahren*. Aufbauend auf einer Vorstellung der Ideen und Konzepte wurden die wichtigsten Verfahren für die Fuzzy-Clusteranalyse vorgestellt. Daneben wurden zwei für die Datenanalyse relevante Bereiche, der Umgang mit verrauschten Daten und die Bewertung einer Klassifikation, näher betrachtet. Danach erfolgte ein kurzer Überblick über weitere Fuzzy-Clusteringverfahren.

Aufgrund des Interesses an Verfahren zur Fuzzy-Clusteranalyse wurden die wichtigsten Fuzzy-Clusteringverfahren als Plug-In „Advanced Cluster Analysis“ für das kommerzielle Datenanalysetool DataEngine implementiert. Die in dieser Dissertation betrachteten *drei Themen — Fuzzy-Clusteranalyse mit possibilistischen Zugehörigkeitsgraden, Behandlung von Daten mit fehlenden Werten bei der Fuzzy-Clusteranalyse und Fuzzy-Clusteranalyse klassifizierter Daten* — wurden durch Reaktionen auf dieses Tool und eigene Erfahrungen im Umgang mit der Fuzzy-Clusteranalyse motiviert.

Bei der Fuzzy-Clusteranalyse können neben den weitverbreiteten probabilistischen Zugehörigkeitsgraden auch *possibilistische Zugehörigkeitsgra-*

de verwendet werden. Die Verwendung possibilistischer Zugehörigkeitsgrade ermöglicht eine bessere Interpretation der Zugehörigeitsgrade der Daten zu den Clustern. Insbesondere, wenn in dem Datensatz ein größerer Anteil von Daten vorliegt, die für mehrere Cluster typisch sind, können die Cluster durch possibilistische Zugehörigkeitsgrade präziser beschrieben werden. Bei der possibilistischen Fuzzy-Clusteranalyse besteht jedoch die Problematik, daß dicht benachbarte Cluster häufig als ein Cluster beschrieben werden. Daher wurden in dieser Arbeit Ansätze für die zielfunktionsbasierte Fuzzy-Clusteranalyse und das Alternating Cluster Estimation entwickelt, die diese Problematik durch die Modellierung einer Abstoßung zwischen den Clustern vermeiden. Wenn die Cluster hinreichend voneinander entfernt sind, verhalten sich die vorgestellten neuen Verfahren wie ein possibilistisches Fuzzy-Clusteringverfahren ohne Abstoßung. Die vorgestellten Ansätze sind daher als Erweiterung bzw. Ergänzung der possibilistischen Fuzzy-Clusteranalyse aufzufassen.

Ein bei der Datenanalyse häufig auftretendes Problem ist der *Umgang mit Daten mit fehlenden Attributwerten*. Für den Umgang mit ihnen gibt es prinzipiell drei verschiedene Möglichkeiten: Daten mit fehlenden Werten bzw. Attribute, in denen fehlende Werte vorliegen, werden aus dem Datensatz entfernt, die fehlenden Werte werden im Rahmen der Datenvorverarbeitung geschätzt oder die Datenanalyseverfahren werden so erweitert, daß auch Daten mit fehlenden Werten analysiert werden können. Da der Umgang mit Daten mit fehlenden Werten bei der Fuzzy-Clusteranalyse bisher noch nicht systematisch untersucht wurde und der Umgang mit Daten mit fehlenden Werten bei der Datenanalyse von sehr großer Bedeutung ist, wurde in Kapitel 4 die Fuzzy-Clusteranalyse mit Daten mit fehlenden Werten systematisch untersucht. Für Daten mit fehlenden Werten „missing completely at random“ wurden zwei verschiedene Ansätze betrachtet: das Schätzen fehlender Werte während der Clusteranalyse z.B. durch den ensprechenden Attributwert des Clusterzentrums und die Berechnung der Clusterprototypen nach der „available case“-Methode. Der für den Umgang mit fehlenden Werten benötigte Aufwand hängt bei der Schätzung der fehlenden Werte von dem Schätzverfahren ab und ist bei der „available case“-Methode vernachlässigbar. Aufgrund der theoretischen Betrachtungen und der in den Tests beobachteten Ergebnisse sowie des vernachlässigbaren zusätzlichen Aufwands ist die „available case“-Methode vorzuziehen. Es wurde gezeigt, daß diese Methode auch bei einem relativ großen Anteil von fehlenden Werten in dem zu betrachtenden Datensatz zu guten Ergebnissen führt. Falls nicht nur fehlende Werte „missing completely at random“ sondern auch fehlende Werte mit einer clusterspezifischen Wahrscheinlichkeit auftreten, ist

es sinnvoll, diese zusätzliche Information bei der Fuzzy-Clusteranalyse zu berücksichtigen. Hierfür wurde ein neues Modell basierend auf dem FMLE entwickelt.

Die Fuzzy-Clusteranalyse ist ein nichtüberwachtes Klassifikationsverfahren. Der Datensatz wird in Cluster von homogenen Daten unterteilt. Manchmal ist jedoch für einige Daten bekannt, zu welcher Klasse sie gehören. (Der Begriff der Klasse ist hier von dem des Clusters zu unterscheiden. Eine Cluster ist eine Menge von homogenen Daten, während eine Klasse aus mehreren Clustern bestehen kann.) Für die Berücksichtigung der Information, zu welchem *Cluster* ein Datum gehört, gibt es im Rahmen der teilüberwachten Fuzzy-Clusteranalyse Ansätze. Der allgemeinere Fall einer Klasseninformation wurde jedoch noch nicht betrachtet. Um diese Information, falls sie vorliegt, bei der Fuzzy-Clusteranalyse zu nutzen, wurden Ansätze für die zielfunktionsbasierte Fuzzy-Clusteranalyse und das Alternating Cluster Estimation entwickelt, die die *Berücksichtigung einer Klasseninformation* ermöglichen. Die Idee dieser neuen Ansätze ist, daß eine Fehlklassifikation (ein Datum wird einem Cluster mit einer anderen Klasseninformation zugeordnet) bestraft wird bzw. daß Daten Cluster mit einer anderen Klasseninformation nicht anziehen sondern abstoßen.

Die im Rahmen dieser Arbeit entwickelten Verfahren wurden als Kommandozeilenprogramm in C implementiert. Durch Einbindung dieses Kommandozeilenprogramms in das am Lehrstuhl für Neuronale Netze und Fuzzy-Systeme, Institut für Wissens- und Sprachverarbeitung der Otto-von-Guericke-Universität Magdeburg, entwickelte Datenanalyseprogramm „OttoMiner“ können die Fuzzy-Clusteringverfahren in einer grafischen Umgebung leicht mit anderen Datenanalyseverfahren kombiniert und in Projekten eingesetzt werden. Aufgrund der Relevanz fehlender Werte für die Datenanalyse ist darüber hinaus geplant, das Plug-In „Advanced Cluster Analysis“für das Datenanalysetool DataEngine in einer zukünftigen Version hinsichtlich des Umgangs mit fehlenden Werten zu erweitern.

Anhang A

Software

Die verschiedenen im Rahmen dieser Arbeit vorgestellten Fuzzy-Clusteringverfahren sind in einem Kommandozeilenprogramm in C implementiert. Neben einem separaten Einsatz von Verfahren der Fuzzy-Clusteranalyse für die Betrachtung von Einzelproblemen ist die Kombination mit anderen Methoden zur Datenanalyse bei Datenanalyseprojekten von großer Bedeutung. Hierfür kann das Programm in das am Lehrstuhl für Neuronale Netze und Fuzzy-Systeme, Institut für Wissens- und Sprachverarbeitung der Otto-von-Guericke-Universität Magdeburg, entwickelte Datenanalyseprogramm „OttoMiner“ integriert werden. „OttoMiner“ stellt eine grafische Oberfläche zur Verfügung, bei der in Form der Darstellung eines Datenflusses verschiedene Datenanalyseverfahren miteinander kombiniert werden können.

Aufgrund des allgemeinen Interesses an Verfahren zur Fuzzy-Clusteranalyse wurden daneben der Fuzzy-C-Means-Algorithmus, der Gustafson–Kessel-Algorithmus und der FMLE als Plug-In für das kommerzielle Datenanalyseprogramm DataEngine von Prof. Dr. R. Kruse, Prof. Dr. F. Klawonn und H. Timm implementiert. Die Einbindung der Verfahren als Plug-In in ein kommerzielles Programm ermöglicht es dem Anwender, ohne detaillierte Kenntnisse der Fuzzy-Clusteranalyse diese Verfahren mit anderen Methoden zur Datenanalyse einfach zu kombinieren und in seinen Projekten einzusetzen. Die Kombination der Verfahren zur Datenanalyse wird bei DataEngine ebenfalls in Form eines Datenflusses dargestellt.

Anhang B

Experimentelle Ergebnisse

Die in Kapitel 4 vorgestellten Verfahren für die *Fuzzy-Clusteranalyse von Daten mit fehlenden Werten* wurden anhand des Brustkrebsdatensatzes [89] und des Weindatensatzes [1] aus dem „UCI Machine Learning Repository“[21] hinsichtlich ihrer Leistungsfähigkeit getestet. Da der Bruskrebsdatensatz nur 16 Daten mit fehlenden Attributwerten besitzt und bei dem Weindatensatz alle Attributwerte beobachtet wurden, wurden zusätzlich fehlende Werte erzeugt.

Bei der Betrachtung der Verfahren für den Umgang mit *fehlenden Werten „missing completely at random“* wurde der Brustkrebsdatensatz mit zwei und mit drei Clustern und der Weindatensatz mit drei Clustern klassifiziert. Der Brustkrebsdatensatz wurde mit allen 9 Attributen ohne Normierung verwendet. Bei dem Weindatensatz wurden von den 13 Attributen die Attribute 7, 10 und 13 verwendet. Der Datensatz wurde hinsichtlich jedes Attributs auf den Wertebereich $[0, 10]$ skaliert. Mit einer Wahrscheinlichkeit p von 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% und 50% wurden fehlende Werte „missing completely at random“ generiert. In den Tabellen B.1, B.2 und B.3 ist die über 10 Datensätze gemittelte Anzahl der fehlklassifizierten Daten für die Fuzzy-Clusteranalyse des Brustkrebsdatensatzes mit dem Fuzzy-C-Means-Algorithmus mit zwei und drei Clustern und des Weindatensatzes mit dem Gustafson–Kessel-Algorithmus mit 3 Clustern angegeben.

Die Datensätze wurden mit folgenden Verfahren klassifiziert:

- *Verfahren I:* Daten mit fehlenden Werten werden vor der Fuzzy-Clusteranalyse aus dem Datensatz entfernt. Nach der Fuzzy-Cluster-

p	I	II	III	IV	V
0	33	31	32	32	32
5	32,8	29,1	30,1	31,1	31,6
10	33	28,7	32,3	33,2	33,4
15	44,3	28	35,9	33	32,9
20	48,7	26,7	37,4	34,5	34,6
25	68,6	25,9	42,5	34,8	34,7
30	85,9	26,4	49,8	37,3	37,2
35	103,1	26,8	54,6	38,2	37,9
40	124,5	27,6	63,3	39,3	39,3
50	-	27,5*	76,125	45,75	45,125

Tabelle B.1: Gemittelte Anzahl der fehlklassifizierten Daten bei der Fuzzy-Clusteranalyse des Brustkrebsdatensatzes mit dem Fuzzy-C-Means-Algorithmus und zwei Clustern. (* nur 2 von 10 Datensätzen konnten klassifiziert werden.) p: Wahrscheinlichkeit für fehlende Werte MCAR.

analyse werden die entfernten Daten klassifiziert unter Berücksichtigung der beobachteten Attribute.

- *Verfahren II:* Fehlende Werte werden während der Fuzzy-Clusteranalyse durch den entsprechenden Attributwert des Clusters geschätzt, zu dem sie den höchsten Zugehörigkeitsgrad besitzen.
- *Verfahren III:* Fehlende Werte werden während der Fuzzy-Clusteranalyse durch das mit den Zugehörigkeitsgraden gewichtete Mittel der Clusterzentren geschätzt.
- *Verfahren IV:* Die Clusterprototypen werden nach der „available case"-Methode berechnet. Die Zugehörigkeitsgrade werden unter Berücksichtigung aller beobachteten Attributwerte geschätzt.
- *Verfahren V:* Die Clusterprototypen werden nach der „available case"-Methode berechnet. Die Zugehörigkeitsgrade werden unter Berücksichtigung aller beobachteten Attributwerte geschätzt. Zusätzlich werden Daten mit fehlenden Werten bei der Fuzzy-Clusteranalyse geringer gewichtet.

Bei der Betrachtung der Verfahren für den Umgang mit *fehlenden Werten mit einer clusterspezifischen Wahrscheinlichkeit* wurde der Weindatensatz mit 3 Clustern klassifiziert. Von den 13 Attributen wurden die Attribute

p	I	II	III	IV	V
0	26	27	25	25	25
5	25,6	23,9	24,3	24,2	23,7
10	32,5	25,1	27,8	26,9	26,9
15	41,5	24,5	29,1	28,1	27,3
20	58,3	24,6	29,6	27,7	28,1
25	67,6	25,5	32	27,4	28,1
30	96,6	28,5	31,2	28,3	27,8
35	112,3	26,5	34,8	28,5	29,2
40	128,6	27,3	34,5	30,1	31,2
50	-	40*	39,714	35,75	35,125

Tabelle B.2: Gemittelte Anzahl der fehlklassifizierten Daten bei der Fuzzy-Clusteranalyse des Brustkrebsdatensatzes mit dem Fuzzy-C-Means-Algorithmus und drei Clustern.(* nur 2 von 10 Datensätzen konnten klassifiziert werden.) p: Wahrscheinlichkeit für fehlende Werte MCAR.

p	I	II	III	IV	V
0	7	7	7	7	7
5	16,3	8,8	9,8	7,6	7,7
10	25	13,4	14,3	11,6	11
15	35,7	15	21,8	12,4	12,6
20	43,2	23	26	15,2	15
25	47,6	-	28,5	18,5	17,8
30	56,6	-	41,4	22,6	23,9
35	62,2	-	51,4	27,1	26,4
40	72,6	-	56,7	34,9	30

Tabelle B.3: Gemittelte Anzahl der fehlklassifizierten Daten bei der Fuzzy-Clusteranalyse des Weindatensatzes (Attribute 7,10,13) mit dem Gustafson–Kessel-Algorithmus und drei Clustern. p: Wahrscheinlichkeit für fehlende Werte MCAR.

p	I	II
20	12,9	11,4
25	14,6	12,1
30	12,9	12
35	15,2	11,7
40	15,7	10,7
45	18,2	10,9
50	18,9	11,6
55	18,5	9,1
60	20,6	11,6

Tabelle B.4: Gemittelte Anzahl der fehlklassifizierten Daten bei der Fuzzy-Clusteranalyse des Weindatensatzes (Attribute 7,10,13) mit dem FMLE und drei Clustern. Wahrscheinlichkeit für fehlende Werte MCAR 5%. p: clusterspezifische Wahrscheinlichkeit für fehlende Werte.

7, 10 und 13 verwendet. Der Datensatz wurde hinsichtlich jedes Attributs auf den Wertebereich $[0, 10]$ skaliert. Bei jeder der drei Klassen des Weindatensatzes wurde für ein Attribut eine klassenspezifische Wahrscheinlichkeit für fehlende Werte erzeugt, indem mit einer Wahrscheinlichkeit von $p_i\%$ dieses Attribut nicht beobachtet wurde. Zusätzlich wurden in dem Datensatz fehlende Werte „missing completely at random“ mit einer Wahrscheinlichkeit von 5% generiert. In Tabelle B.4 ist die über 10 Datensätze gemittelte Anzahl der fehlklassifizierten Daten für die Fuzzy-Clusteranalyse des Weindatensatzes mit dem FMLE mit 3 Clustern angegeben.

Die Datensätze wurden mit folgenden Verfahren durch Varianten des FMLE klassifiziert:

- *Verfahren I:* Die Clusterprototypen werden nach der „available case“-Methode berechnet. Die Zugehörigkeitsgrade werden unter Berücksichtigung aller beobachteten Attributwerte geschätzt. Zusätzlich werden Daten mit fehlenden Werten bei der Fuzzy-Clusteranalyse geringer gewichtet.

- *Verfahren II:* Es wurde der wahrscheinlichkeitsbasierte Ansatz zur Berücksichtigung einer clusterspezifischen Wahrscheinlichkeit für fehlende Werte verwendet.

Literaturverzeichnis

[1] Aeberhard, S., Coomans, D. und de Vel, O.: Comparison of Classifiers in High Dimensional Settings. Tech Rep. 92—02, Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland, 1992.

[2] Anderson, J.A.: An Introduction to Neural Networks. MIT Press, Cambridge, MA, 1995.

[3] Babu, G.P. und Murty, M.N.: Clustering with Evolutionary Stategies. Pattern Recognition, 27, 321—329, 1994.

[4] Bacher, J.: Clusteranalyse, Oldenbourg Verlag, München, Wien, 1996.

[5] Backer, E. und Jain, A.K.: A clustering performance measure based on fuzzy set decomposition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 3 (1), 66—74, 1981.

[6] Ball, G. und Hall, D.A.: A Clustering Technique for Summarizing Multivariate Data. Behavioral Science, 12, 153–155, 1967.

[7] Bandemer, H. und Näther, W.: Fuzzy Data Analysis. Kluwer Academic Publishers,1992.

[8] Bandemer, H.: Ratschläge zum mathematischen Umgang mit Ungewißheit — Reasonable Computing. Teubner, Stuttgart, 1997.

[9] Barni, M., Cappellini, V. und Mecocci, A.: Comments on „a possibilistic approach to clustering“. IEEE Trans. on Fuzzy Systems, 4(3), 393—396, 1996.

[10] Bell, E.T.: Men of Mathematics. 5th printing, Simon and Schuster, New York, 1966.

[11] Bensaid, A.M., Hall, L.O., Bezdek, J.C. und Clarke, L.P.: Partially Supervised Clustering for Image Segmentation. Pattern Recognition, 29 (5), 859—871, 1996.

[12] Berthold, M. und Hand, D.J.(Ed.): Intelligent Data Analysis. Springer Verlag, Berlin, 1999.

[13] Bezdek, J.C.: Fuzzy Mathematics in Pattern Classification. Dissertation Cornell University, Ithaca, NY, USA, 1973.

[14] Bezdek, J.C.: A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2(1), 1—8, 1980.

[15] Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York, 1981.

[16] Bezdek, J.C., Coray, C., Gunderson, C. und Watson, J.: Detection and Characterization of Cluster Substructure. SIAM Journal of applied mathematics, 40, 339—372, 1981.

[17] Bezdek, J.C., Hathaway, R.J., Sabin, M.J. und Tucker, W.T.: Convergence Theory for Fuzzy c-Means: Counterexamples and Repairs. IEEE Trans. on Systems, Man, and Cybernetics, 17(5), 873—877, 1987.

[18] Bezdek, J.C. und Pal, S.K. (Hrsg.): Fuzzy Models for Pattern Recognition: methods that search for structures in data. IEEE Press, Piscataway, 1992.

[19] Bezdek, J.C., Keller, J., Krishnapuram, R. und Pal, N.R.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer, Boston, London, 1999.

[20] Bezdek, J.C. und Hathaway R.J.: Some New Ways to Cluster in Incomplete Data. In: Sincak, P. und Vascak, J. (Hrsg.): Quo Vadis Computational Intelligence?: New Trends and Approaches in Computational Intelligence. Physia-Verlag, Heidelberg, 190—208 , 2000.

[21] Blake, C.L. und Merz, C.J.: UCI Repository of machine learning databases. `http:`
`www.ics.uci.edu/~mlearn/MLRepository.html`, 1998.

[22] Bock, H.H.: Automatische Klassifikation: Vandenhoeck & Ruprecht, Göttingen, Zürich, 1974.

[23] Borgelt, C. und Kruse, R.: Evaluation Measures for Learning Probabilistic and Possibilistic Networks. Proc. 6th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE'97), Vol. 2, 1034—1038, Barcelona, Spain, 1997.

[24] Borgelt, C., Gebhardt, J. und Kruse, R.: Chapter F1.2: Inference Methods. In: Ruspini, E., Bonissone, P. und Pedrycz W. (Hrsg.): Handbook of Fuzzy Computation. Institute of Physics Publishing Ltd., Bristol, United Kingdom, 1998.

[25] Borgelt, C. und Timm, H.: Advanced Fuzzy Clustering and Decision Tree Plug-Ins for DataEngine. In: Azvine, B., Azarmi, N. und Nauck, D. (Hrsg): Intelligent Systems and Soft Computing: Prospects, Tools and Applications. Springer Verlag, Berlin, Deutschland, 188—212, 2000.

[26] Borgelt, C., Timm, H. und Kruse, R.: Unsicheres und vages Wissen. In: Görz, G. Rollinger, C.-R. und Schneeberger, J. (Hrsg): Einführung in die Künstliche Intelligenz (3. Auflage). Addison Wesley, Bonn, Germany, 291—347, 2000.

[27] Borgelt, C., Timm, H. und Kruse, R.: Using Fuzzy Clustering to Improve Naive Bayes Classifiers and Probabilistic Networks. Proc. 8th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'00), IEEE Press, Piscataway, NJ, USA, 2000.

[28] Borgelt, C.: Data Mining with Graphical Models. Dissertation Fakultät für Informatik, Otto von Guericke Universität Magdeburg, 2000.

[29] Borgelt, C., Timm, H. und Kruse, R.: Probabilistic Networks and Fuzzy Clustering as Generalizations of Naive Bayes Classifiers. In: Reusch, B. und Temme, K.-H. (Hrsg.): Computational Intelligence in Theory and Practice. Physica Verlag, Heidelberg, Deutschland, 121—138, 2001.

[30] Breiman, L., Friedman, J.H., Olshen, R.A. und Stone, C.J.: Classification and Regression Trees. Wadsworth International, Belmont, CA, 1984.

[31] Buck, S.F.: A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society, 22(B), 302—306, 1960.

[32] The CRISP-DM Process Model. http://www.crisp-dm.org/.

[33] Chi, Z., Yan, H., Pham, T.: Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition. World Scientific, Singapore, New Jersey, London, 1996.

[34] Davé, R.N.: Use of the Adaptive Fuzzy Clustering Algorithm to Detect Lines in Digital Images. Proc. of the SPIE — The International Society for Optical Engineering, 1192, 600—611, 1989.

[35] Davé, R.N.: Fuzzy Shell-Clustering and Application to Circle Detection in Digital Images. Intern. Journal General Systems, 16, 343—355, 1990.

[36] Davé, R.N. und Bhaswan, K.: Adaptive Fuzzy c-Shells Clustering and Detection of Ellipses. IEEE Trans. Neural Networks, 3(5), 643—662, 1992.

[37] Davé, R.N.: Characterization and Detection of Noise in Clustering. Pattern Recognition Letters, 12, 657—664, 1991.

[38] Davé, R.N. und Krishnapuram, R.: Robust Clustering Methods: A Unified View, IEEE Transactions on Fuzzy Systems, 5(2), 270—293, 1997.

[39] Davé, R.N. und Sen, S.: On Generalizing the Noise Clustering Algorithms. Proc. of the 7th Fuzzy Systems Association World Congress (IFSA97), vol. III, 205—210, 1997.

[40] Dempster, A.P., Laird, N.M. und Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, 39(B), 1—38, 1977.

[41] Dixon, J.K.: Pattern Recognition with partly missing data. IEEE Transactions on Systems, Man, and Cybernetics, 9(6), 617—621, 1979.

[42] Dubois, D. und Prade, H.: Possibility Theory. Plenum Press, New York, NY, USA 1988.

[43] Dubois, D. und Prade, H.: The three semantics of fuzzy sets. Fuzzy Sets and Systems, 90(2), 141—150, 1997.

[44] Duda, R. und Hart, P.: Pattern Classification and Scene Analysis. Wiley, New York, 1973.

[45] Dunn, J.: A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact, Well Seperated Clusters. Journal of Cybernetics, 3(3), 32—57, 1973.

[46] Dumitrescu, D., Lazzerini, B. und Jain, L.C.: Fuzzy Sets and their Application to Clustering and Training. CRC Press, Boca Raton, London, New York, 2000.

[47] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. und Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. MIT Press, Cambridge, London, 1996.

[48] Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179—188, 1936.

[49] Frigui, H. und Krishnapuram, R.: A Comparison of Fuzzy Shell-Clustering Methods for the Detection of Ellipses. IEEE Trans. on Fuzzy Systems, 4(2), 193—199, 1996.

[50] Frigui, H. und Krishnapuram, R.: Clustering by Competitive Agglomeration. Pattern Recognition, 30 (7), 1109—1119, 1997.

[51] Gath, I. und Geva, A.B.: Unsupervised Optimal Fuzzy Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11, 773—781, 1989.

[52] Gebhardt, J. und Kruse, R.: POSSINFER — A Software Tool for Possibilistic Inference. In: Dubois, D., Prade, H. und Yager, R. (Hrsg.) Fuzzy Set Methods in Information Engineering: A Guided Tour of Applications, Wiley, 1995.

[53] Gentsch, P.: Data Mining Tools: Vergleich marktgängiger Tools. WHU Koblenz, Deutschland, 1999.

[54] Grabmeier, J., Buhmann, J., Kruse, R., Timm, H.: Segmentierende und clusterbildende Methoden. In: Hippner, H., Küsters, U.L. und Meyer, M. (Hrsg.): Handbuch Data Mining im Marketing. Knowledge Discovery in Marketing Databases. Vieweg, Wiesbaden, 2001.

[55] Gunderson, R.: An adaptive FCV clustering algorithm. International Journal of Man-Machine Studies, 19, 97—104, 1983.

[56] Gustafson, E.E. und Kessel, W.C.: Fuzzy Clustering with a Fuzzy Covariance Matrix, IEEE CDC, San Diego, Californien, 761—766, 1979.

[57] Hathaway, R.J. und Bezdek J.C.: Optimization of Clustering Criteria by Reformulation. IEEE Trans. on Fuzzy Systems, 3(2), 241—245, 1995.

[58] Hathaway, R.J., Overstreet, D.D. und Bezdek J.C.: Fuzzy c-Means Clustering of partially missing data sets. Proceedings of SPIE — Applications and Science of Computational Intelligence, 4055, 159—165, 2000.

[59] Hathaway, R.J. und Bezdek J.C.: Fuzzy c-Means Clustering of Incomplete Data. IEEE Trans. on Systems, Man, and Cybernetics - Part B, 31(5), 735—744, 2001.

[60] Hathaway, R.J., Hu, Y. und Bezdek J.C.: Local Convergence of Tri-Level Alternating Optimization. Neural, Parallel, Sci. Computat., 9, 19—28, 2001.

[61] Heckerman, D., Geiger, D. und Chickering, D.M.: Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Machine Learning, 20, 197—243, Kluwer, 1995.

[62] Heitjan, D.F.: Application of random effects pattern mixture models for missing data in longitudinal studies. Psychological Methods, 2(1), 64—78, 1997.

[63] Hita, K. und Iwama, K.: Application of modified FCM with additional data to area divison of images. Inf. Sci., 45(2), 213—230, 1988.

[64] Höppner, F.: Fuzzy Shell Clustering Algorithms in Image Processing — fuzzy c-rectangular and 2-rectangular shells. IEEE Trans. on Fuzzy Systems, 5(4), 599-613, 1997.

[65] Höppner, F., Klawonn, F., Kruse, R. und Runkler, T.: Fuzzy Cluster Analysis, Wiley, Chichester, New York, 1999. Aktualisierte Version der deutschen Ausgabe: Höppner, F., Klawonn, F. und Kruse, R.: Fuzzy-Clusteranalyse. Verfahren für die Bilderkennung, Klassifikation und Datenanalyse, Vieweg, Braunschweig, 1997.
`http://fuzzy.cs.uni-magdeburg.de/clusterbook`

[66] Jensen, F.V.: An Introduction to Bayesian Networks. Springer Verlag, New York, 1996.

[67] Kersten, P.R.: Implementation Issues in the Fuzzy c-Medians Clustering Algorithm. Proc. of the 6th IEEE International Conference on Fuzzy Systems (FUZZIEEE 97), 957—962, Spanien, 1997.

[68] Klawonn, F. und Kruse, R.: Automatic Generation of Fuzzy Controllers by Fuzzy Clustering. Proc. IEEE Int. Conf. on Systems, Man and Cybernetics, 2040-2045, Vancouver, 1995.

[69] Klawonn, F., Kruse, R. und Timm, H.: Fuzzy Shell Cluster Analysis. In: Della Riccia, G., Lenz, H.-J. und Kruse, R.(Hrsg): Learning Networks and Statistics, 105—120, Springer Verlag, New York, Berlin, 1997.

[70] Klawonn, F. und Keller, A.: Fuzzy Clustering with Evolutionary Algorithms. Intelligent Systems, 13, 975—991, 1998.

[71] Klir, J. und Yuan, B.: Fuzzy Sets and Fuzzy Logic — Theory and Applications, Prentice Hall, New Jersey, 1995.

[72] Krishnapuram, R. und Freg, C.-P.: Fitting an Unknown Number of Lines and Planes to Image Data through Compatible Cluster Merging. Pattern Recognition, 25(4), 385—400, 1992.

[73] Krishnapuram, R., Frigui, H. und Nasroui, O.: The Fuzzy C Spherical Shells Algorithm: A New Approach. IEEE Trans. Neural networks, 3(5), 663—670, 1992.

[74] Krishnapuram, R., Frigui, H. und Nasroui, O.: The Fuzzy C Quadric Shell Clustering Algorithm and the Detection of Second-Degree Curves. Pattern Recognition Letters, 14, 545—552, 1993.

[75] Krishnapuram, R. und Keller, J.: A Possibilistic Approach to Clustering, IEEE Transactions on Fuzzy Systems, pp. 98—110, (1) 1993.

[76] Krishnapuram, R. und Keller, J.: Fuzzy and Possibilistic Clustering Methods for Computer Vision. In: Mitra, S., Gupta, M. und Krsake, W. (Hrsg.): Neural and Fuzzy Systems. SPIE Institute Series, Vol. IS 12, 133—159, 1994.

[77] Krishnapuram, R., Frigui, H. und Nasroui, O.: Fuzzy and Possibilistic Shell Clustering Algorithms and Their Application to Boundary Detection and Surface Approximation — Part I & Part II. IEEE Trans. on Fuzzy Systems, 3(1), 29—60, 1995.

[78] Krishnapuram, R. und Keller, J.M.: The Possibilistic c-Means Algorithm: Insights and Recommendations. IEEE Trans. on Fuzzy Systems, 4(3), 385—393, 1996.

[79] Kruse, R,, Schwecke, E. und Heinsohn, J.: Uncertainty and Vagueness in Knowledge-based Systems. Serie Artificial Intelligence, Springer Verlag, Berlin, 1991.

[80] Kruse, R., Gebhardt, J. und Klawonn, F.: Foundations of Fuzzy Systems, Wiley, Chichester, New York, 1994.

[81] Kruse, R., Gebhardt, J. und Klawonn, F.: Fuzzy-Systeme, 2. Auflage, Teubner Verlag, Stuttgart, 1995.

[82] Kruse, R., Borgelt, C. und Nauck, D.: Fuzzy Data Analysis: Challenges and Perspectives. Proc. 8th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE'99), IEEE Press, Piscataway, NJ, USA, 1211—1216, 1999.

[83] Kruse, R., Borgelt, C. und Nauck, D.: Problems and Prospects in Fuzzy Data Analysis. In: Azvine, B., Azarmi, N. und Nauck, D. (Hrsg.): Intelligent Systems and Soft Computing: Prospects, Tools and Applications. Springer Verlag, Berlin, Deutschland, 188—212, 2000.

[84] Lauritzen, S.L. und Spiegelhalter, D.J.: Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. Journal of the Royal Statistical Society, Series B, 2, 50, 157—224, 1988.

[85] Lauritzen, S.L.: Graphical Models. Oxford University Press, Oxford, England, 1997.

[86] Little, R.J.A. und Rubin, D.A.: Statistical analysis with missing data. John Wiley and Sons, New York, 1987.

[87] Mahalanobis, P.C.: On Tests an Measures of Groups Divergence I. Journal of the Asiatic Society of Benegal, 26, 541, 1930.

[88] Mahalanobis, P.C.: On the generalized distance in statistics. Proc. Nat. Inst. Sci. India 2, 49—55, 1936.

[89] Mangasarian, O.L. und Wolberg, W.H.: Cancer diagnosis via linear programming. SIAM News, Vol. 23, Nr. 5, 1—18, 1990.

[90] McKay, M.D., Beckman, R.J. und Conover, W.J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21(2), 239—245, 1979.

[91] Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3. Auflage,Springer Verlag, Berlin, Heidelberg, New York, 1996.

[92] Miyamoto, S., Takata, O. und Umayahara, K.: Handling Missing Values in Fuzzy-c Means. Proc. of the Third Asian Fuzzy Systems Symposium, 139—142, 1998.

[93] Miyamoto, S.: Fuzzy Sets in Information Retrieval and Cluster Analysis, Kluwer, Dordrecht, Boston, 1990.

[94] Mirkin, B.: Mathematical Classification and Clustering, Kluwer, Dordrecht, Boston, 1996.

[95] Mucha, H.-J.: Clusteranalyse mit Mikrocomputern. Akademie-Verlag, Berlin, 1992.

[96] Nakhaeizadeh, G.: Data Mining: Theoretische Aspekte und Anwendungen. Physica Verlag, Heidelberg, Deutschland, 1998.

[97] Nauck, D., Klawonn, F. und Kruse, R.: Foundations of Neuro-Fuzzy Systems. Wiley, Chichester, 1997. Aktualisierte Version der deutschen Ausgabe: Nauck, D., Klawonn, F. und Kruse, R.: Neuronale Netze und Fuzzy-Systeme, 2. erweiterte Auflage. Vieweg, Wiesbaden, 1996.

[98] Nürnberger, A. und Timm, H.: OR Software: Data Engine. OR Spektrum, 21(3), 305—313, Springer Verlag, 1999.

[99] Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (2nd edition). Morgan Kaufman, New York, 1992.

[100] Pedryzc, W.: Algorithms of Fuzzy Clustering with Partial Supervision. Pattern Recognition Letters, 3, 13—20, 1985.

[101] Pedryzc, W. und Waletzky, J.: Fuzzy-Clustering with Partial Supervision. IEEE Trans. on Systems, Man, and Cybernetics - Part B, 27(5), 787—795, 1997.

[102] Quinlan, J.R.: Induction of Decision Trees. Machine Learning, 1, 81—106, 1986.

[103] Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufman, 1993.

[104] Rojas, R.: Theorie der Neuronalen Netze: Eine systematische Einführung. Springer Verlag, Berlin, 1993.

[105] Runkler, T. und Bezdek, J.C.: Alternating Cluster Estimation: A new tool for clustering and function approximation. IEEE Trans. on Fuzzy Systems, 7(4), 377—393, 1999.

[106] Ruspini, E.H.: A New Approach to Clustering. Information Control, 15(1), 22—32, 1969.

[107] Sachs, L.: Angewandte Statistik, 8. Auflage, Springer Verlag, Berlin, Heidelberg, New York, 1997.

[108] Sato, M., Sato, Y. Jain, L.C.: Fuzzy Clustering Models and Applications, Physica Verlag, Heidelberg, New York, 1997.

[109] Schafer, J.L.: Analysis of Incomplete Multivariate Data, Chapman & Hall, London, 1997.

[110] Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379—423, 1948.

[111] Späth, H.: Cluster-Formation und -Analyse. Oldenbourg, München, 1983.

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

[113] Strackeljahn, J., Behr, D. und Kocher, T.: Fuzzy-Pattern Recognition for Automatic Detection of Different Teeth Substances. Fuzzy Sets and Systems, 85, 275—286, 1997.

[114] Strackeljahn, J. und Weber, R.: Quality Control an Maintenance. In: Zimmermann, H.J. (Hrsg.): Practical Applications of Fuzzy Technologies. 161—184, Kluwer, Boston, London, 1999.

[115] Stutz, C.: Partially Supervised Fuzzy c-Means Clustering with Cluster Merging. Proc. of the 6th Europ. Congress on Intelligent Techniques and Soft Computing, 1725—1729, 1998.

[116] Stutz, C.: Anwendungsspezifische Fuzzy-Clustermethoden. Dissertation Fakultät für Informatik, Technische Universität München, Deutschland, 1999.

[117] Thorndike, R.L.: Who belongs in the family? Psychometrika 18, 18, 267-276, 1953.

[118] Timm, H.: Fuzzy-Methoden zur Erkennung von komplexen Objekten in Bildern. Diplomarbeit Technische Universität Braunschweig, 1996.

[119] Timm, H., Kruse, R., Nauck, D. und Klawonn, F.: Flexible Fuzzy Clustering for Data Analysis as a Plug-In Library for Data Engine. Proc. 1st International Data Analysis Symposium, 67—71, Aachen, Deutschland, 1997.

[120] Timm, H. und Klawonn, F.: Object Recognition with Fuzzy Shell Cluster Analysis. Proc. 5th European Congress on Intelligent Techniques and Soft Computing (EUFIT '97), 1039—1043, Aachen, Deutschland, 1997.

[121] Timm, H.: A Fuzzy Cluster Analysis Plug-In for DataEngine. Proc. 6th European Congress on Intelligent Techniques and Soft Computing (EUFIT '98), 1304—1308, Aachen, Deutschland, 1998.

[122] Timm, H. und Klawonn, F.: Classification of Data with Missing Values. Proc. 6th European Congress on Intelligent Techniques and Soft Computing (EUFIT '98), 1304—1308, Aachen, Deutschland, 1998.

[123] Timm, H. und Kruse, R.: Fuzzy Cluster Analysis with Missing Values. Proc. 17th International Conf. of the North American Fuzzy Information Processing Society (NAFIPS98), 242—246, Pensacola, FL, USA, 1998.

[124] Timm, H. und Kruse, R.: Fuzzy-Clusteranalyse mit DataEngine. Proc. 4. DataEngine User Meeting, Aachen, Germany, 1998.

[125] Timm, H. und Klawonn, F.: Different Approaches for Fuzzy Cluster Analysis with Missing Values, Proceedings of 7th European Congress on Intelligent Techniques & Soft Computing, Aachen, Germany, Proceedings auf CDROM, 1999.

[126] Timm, H.: Fuzzy Cluster Analysis of Classified Data. Proc. Joint 9th IFSA World Congress and 20th International Conf. of the North American Fuzzy Information Processing Society (NAFIPS01), Vancouver, Canada, 2001.

[127] Timm, H., Borgelt, C., Döring, C. und Kruse, R.: Fuzzy Cluster Analysis with Cluster Repulsion. Proc. European Symposium on Intelligent Technologies (EUNITE), Tenerife, Spain, 2001.

[128] Timm, H. und Kruse, R.: A Modification to Improve Possibilistic Fuzzy Cluster Analysis. angenommen für 2002 IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE'02), Honolulu, HI, USA, 2002.

[129] Titterington, D., Smith, A. und Makov. U.: Statistical Analysis of Finite Mixture Distributions. Wiley, London, 1985.

[130] Tizhoosh, H.R.: Fuzzy-Bildverarbeitung — Einführung in Theorie und Praxis. Springer Verlag, Berlin Heidelberg, New York, 1998.

[131] Xie, X.L. und Beni, G.: A Validity Measure for Fuzzy Clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 13(8), 841—847, 1991.

[132] Zadeh, L.A.: Fuzzy Sets. Information Control, 8, 338—353, 1965.

[133] Zadeh, L.A.: The Concepts of a Linguistic Variable and its Application to Approximate Reasoning. Information Sciences, 8, 199—249, 301—357, 9, 43—80, 1975.

[134] Zadeh, L.A.: Fuzzy Sets as a Basis for a Theory of Possibility. Fuzzy Sets and Systems, 1, 3—28, 1978.

[135] Zimmermann, H.J. (Hrsg.): Fuzzy Technologien — Prinzipien, Werkzeuge, Potentiale. VDI Verlag, Düsseldorf, 1993.

[136] Zimmermann, H.J. (Hrsg.): Datenanalyse — Anwendung von DataEngine mit Fuzzy Technologien und Neuronalen Netzen. VDI Verlag, Düsseldorf, 1995.

[137] Zimmermann, H.J. (Hrsg.): Neuro + Fuzzy — Technologien - Anwendungen. VDI Verlag, Düsseldorf, 1995.

[138] Zimmermann, H.J. (Hrsg.): Practical Applications of Fuzzy Technologies. Kluwer Academic Publishers, Boston, Dordrecht, London, 2000.

Curriculum Vitae

Name:	Heiko Friedrich Timm
Anschrift:	Barnser Str. 2 D-29581 Gerdau
Geburtstag:	7. Mai 1970
Geburtsort:	Göttingen
Familienstand:	ledig
Eltern:	Dr. Fritz Timm Ulrike Timm, geborene Schmidtmadel
Schulbildung:	1976 – 1980 Besuch der Grundschule in Mainz und Remscheid 1980 – 1989 Besuch des Leibniz-Gymnasiums in Remscheid
Schulabschluß:	Allgemeine Hochschulreife
Wehrdienst:	1. Juni 1989 – 30. September 1990 in Lüneburg
Studium:	1990 – 1996 an der Technischen Universität Carolo-Wilhelmina zu Braunschweig
Oktober 1990:	Immatrikulation für den Studiengang Informatik
September 1992:	Vordiplom im Studiengang Informatik
November 1996:	Diplom im Studiengang Informatik
Schwerpunkte:	Künstliche Intelligenz/Wissensverarbeitung Technische Informatik (integrierte Schaltungen) Theoretische Informatik
Anwendungsfach:	BWL-Operations Research

Berufliche Tätigkeit:

Seit dem 1. Dezember 1996
wissenschaftlicher Mitarbeiter am Institut für Wissens- und Sprachverarbeitung der Otto-von-Guericke-Universität Magdeburg (Arbeitsgruppe Neuronale Netze und Fuzzy-Systeme, Prof. Dr. Rudolf Kruse).