Institut für Parallele und Verteilte Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Masterarbeit

# Einbindung von Domänenexperten in die interaktive Verfeinerung von Clustering-Resultaten

Jannis Rapp

| | |
|---|---|
| **Studiengang:** | Informatik |
| **Prüfer/in:** | Prof. Dr. rer. nat. habil. Holger Schwarz |
| **Betreuer/in:** | Dipl.-Inf. Michael Behringer,<br>Dennis Tschechlov, M. Sc. |
| **Beginn am:** | 15. Juni 2022 |
| **Beendet am:** | 15. Dezember 2022 |

## Kurzfassung

Die heute verfügbare Datenmenge bietet für Unternehmen neue Möglichkeiten durch die Datenanalyse, etwa zur Verbesserung von Geschäftsprozessen oder zur Erschließung neuer Kunden. Eine populäre Form der Datenanalyse ist die Clusteranalyse, die sich mit der Gruppierung von Daten beschäftigt. In vielen Fällen ist jedoch die Clusteranalyse von externem Domänenwissen abhängig, weshalb die automatisierte Anwendung von Clustering-Verfahren keine zufriedenstellende Resultate erzeugt. Aus diesem Grund bietet es sich an, Domänenexperten mit ihrem implizit vorhandenen Domänenwissen direkt interaktiv in den Analyseprozess zu integrieren. Vorhandene interaktive Ansätze eignen sich allerdings aufgrund des Anforderungsprofils von Domänenexperten nicht und beschränken die mögliche Nutzung von Domänenwissen während der Clusteranalyse. Für eine zielführende Einbindung muss der Domänenexperte die Möglichkeit haben, während der Clusteranalyse sowohl Parameter zu spezifizieren, als auch das Resultat schrittweise zu verfeinern und auf den Anwendungsfall anzupassen. In dieser Arbeit wird ein Konzept für diese Einbindung von Domänenexperten zur interaktiven Verfeinerung von Clustering-Resultaten vorgestellt. Hierzu wird ein Prozessmodell zur umfassenden Integration eines Domänenexperten in die Clusteranalyse entwickelt und prototypisch implementiert. Eine umfangreiche Evaluation auf Basis von vier synthetisch generierten Datensätzen zeigt, dass die Kombination verschiedener Ansätze zu genaueren Ergebnissen in weniger Rechenschritten führt.

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Tabellenverzeichnis

# Verzeichnis der Algorithmen

# 1 Motivation

Die weltweit erzeugt und gespeicherte Datenmenge steigt rasant an [RGR18]. Auf diesen Daten können verschiedene Methoden zur Datenanalyse angewendet, um neue Einsichten zu erlangen [Dav06]. Davon machen zunehmend mehr Unternehmen Gebrauch und nutzen diese Daten für die Erzielung von Wettbewerbsvorteilen, etwa durch die Optimierung von Geschäftsprozessen oder die Vorhersage der Kundennachfrage [Dav06]. Eine gängige Form der Analyse stellt die Aufteilung der Instanzen der Daten in Gruppen mit ähnlichen Charakteristiken dar, was als Clustering bezeichnet wird [BHR+20]. Die automatisierten Durchführung von Clustering-Verfahren gruppiert die Instanzen basierend auf der Struktur der Daten. Das bekannteste Beispiel dieser automatisierten Clustering-Verfahren ist der weitverbreitet k-Means Algorithmus, welcher Daten in eine festgelegte Anzahl (k) an Gruppen unterteilt [Mac67].

Die Bildung von möglichst „guten“ Gruppierungen hängt allerdings häufig vom Anwendungsfall der Analyse und der Domäne der Daten ab [SZS16]. Diese inhärent subjektive Eigenschaft des Clustering-Problems ist für klassische unüberwachte Verfahren problematisch, da für die Erstellung bestmöglicher Resultate die Ziele des Analysten während der Gruppierung der Daten berücksichtigt werden müssen [SZS16]. Die Identifikation von Gruppierungen, welche den Anforderungen des Nutzers entsprechen, kann daher in der Regel nicht automatisiert durchgeführt werden, sondern benötigt externes Domänenwissen.

Einen populären Lösungsansatz für dieses Problem stellen interaktive Clustering-Verfahren dar, welche den Analyst in den Clustering-Prozess einbinden [AA20; BBM04a; CDH+16; LKC+12]. Der Analyst benutzt bei diesen Ansätzen sein Domänenwissen, um die Gruppierung der Instanzen während der Clusteranalyse zu beeinflussen. Zur Beeinflussung der Gruppierung wird von diesen Ansätzen beispielsweise die Definition von Constraints [BBM04a; OY11; WD10] oder die Anwendung interaktiver Operationen [CDH+16; LKC+12; SZS16] (beispielsweise das Zusammenführen mehrerer Cluster) verwendet. Daraus geht die Eignung von sogenannten Domänenexperten, Nutzer mit tiefgreifender Expertise innerhalb ihrer Domäne, als ideale Anwendergruppe interaktiver Ansätze hervor.

Vorhandene interaktive Ansätze berücksichtigen allerdings nicht die Bedürfnisse von Domänenexperten und beschränken außerdem durch die mangelnde Flexibilität der angebotenen Interaktionen die Fähigkeit des Domänenexperten, seine Expertise in die Erstellung des Clustering-Resultates einfließen zu lassen.

In dieser Arbeit wird daher ein Ansatz zur Integration der Domänenexperten in einen interaktiven Clustering-Prozess eingeführt. Der Ansatz ermöglicht die flexible Einbringung von Domänenwissen durch die Kombination unterschiedlicher Interaktionsmöglichkeiten. Hierzu werden zunächst die besonderen Fähigkeiten und Bedürfnisse der Domänenexperten definiert. Im Folgenden werden zuerst die Eigenschaften von Domänenexperten spezifiziert und Anforderungen an die Integration von Domänenexperten in die Clusteranalyse eingeführt.

## 1.1 Zielgruppe

Das Ziel der Arbeit ist die Einbindung von Domänenexperten in die interaktive Verfeinerung von Clustering-Resultaten. Die Zielgruppe der Domänenexperten besitzt besondere Fähigkeiten, Kenntnisse und Bedürfnisse, weshalb zunächst die Eigenschaften von Domänenexperten klar definiert werden müssen.

Die besonderen Kenntnisse der Domänenexperten materialisieren sich (vor allem) in der Form von Domänenwissen. Domänenwissen ist eine entscheidende Ressource in Analysen und wird in der wissenschaftlichen Literatur „Domänenwissen“ oftmals ohne konkrete Definition verwendet, was zu Unklarheiten führen kann [Ale92]. Im Folgenden wird daher zunächst der Begriff des Domänenwissens definiert und anschließend zur Abgrenzung der Zielgruppe des Domänenexperten verwendet.

### 1.1.1 Domänenwissen

Im Allgemeinen ist Domänenwissen der implizit vorhandene Wissensbestand eines Individuums innerhalb eines bestimmten Fachbereichs (einer Domäne), welcher den in der Allgemeinheit vorhandenen Wissensbestand überschreitet [Ale92].

Zum Beispiel ist die Beantwortung der Frage „Was ist eine Marketingkampagne?“ mit Wissen, welches einer durchschnittlichen Person innewohnt, möglich. Die Frage kann deshalb ohne die Verwendung von spezifischem Domänenwissen beantwortet werden. Das Gegenteil zeigt die Beantwortung der Frage „Wie können erfolgreiche Marketingkampagnen entwickelt werden“, welche nur mit tiefgreifenderem Wissen im Bereich des Marketings (beispielsweise über die Eigenschaften verschiedener Werbemaßnahmen und der beworbenen Produkte) beantwortet werden kann. Die Beantwortung der zweiten Frage benötigt somit vorhandenes Wissen, welches im Allgemeinen den Wissensbereich eines Individuums ohne vorheriger Erfahrung im Marketingbereich überschreitet und daher Domänenwissen darstellt [Ale92].

Im Rahmen dieser Arbeit wird Domänenwissen als spezialisiertes Wissen über die Domäne, aus der die zu analysierenden Daten erhoben wurden, definiert. Dies beinhaltet insbesondere Wissen über inhärente Eigenschaften der Daten (beispielsweise die Struktur oder Abhängigkeiten innerhalb der Daten) und über Eigenschaften der gewünschten Gruppierung der Daten (beispielsweise die Anzahl der gewünschten Kundengruppen oder die Eigenschaften einzelner Gruppen).

### 1.1.2 Domänenexperte

Als Domänenexperte werden Individuen bezeichnet, denen eine substantielle Menge an Domänenwissen in einem bestimmten Fachbereich (im Rahmen dieser Arbeit der Bereich, aus welchem die Daten erhoben wurden) innewohnt. Diesen Domänenexperten lassen sich aufgrund ihrer Expertise innerhalb ihrer Domäne besondere Eigenschaften zuschreiben [GC88]. Glaser und Chase [GC88, Kapitel 0] identifizieren sieben Eigenschaften:

1. Domänenexperten besitzen herausragende Fähigkeiten innerhalb der Domäne ihrer Expertise. Die Lösung von Problemen außerhalb ihrer Domäne führt (im Allgemeinen) zu durchschnittlichen Ergebnissen.

2. Domänenexperten können einfach sinnvolle Muster in domänenspezifischen Daten finden.
3. Domänenexperten können Aufgaben innerhalb ihrer Domäne schnell und mit geringer Fehleranzahl lösen.
4. Domänenexperten besitzen ein besseres Erinnerungsvermögen für domänenspezifische Szenarien.
5. Domänenexperten besitzen ein tiefgreifenderes Verständnis von domänenspezifischen Problemen.
6. Domänenexperten besitzen die Fähigkeit, aufwendige und hochwertige Analysen von domänenspezifischen Problemen durchzuführen.
7. Domänenexperten können ihre eigenen Fehlerquellen, ihr nicht vorhandenes Wissen über einen Sachverhalt und Stellen ihrer Arbeit, welche eine erneute Überprüfung erfordern, eigenständig identifizieren.

Diese Eigenschaften bilden die Basis für die in dieser Arbeit verwendete Rolle des Domänenexperten. Die Rolle des Domänenexperten wird daher innerhalb der Arbeit als spezialisierter Nutzer mit herausragenden Eigenschaften zur Analyse und Lösung von Problemen (Abschnitt 1.1.2, 2-7) innerhalb des Anwendungsbereiches (seiner Domäne) der Clusteranalyse verwendet, welcher jedoch als nicht technische Nutzer (im Allgemeinen) keine Expertise im Bereich der Datenanalyse oder Informatik besitzt.

## 1.2 Anforderungen an die Funktionalität

Domänenexperten arbeiten mit einer Vielzahl verschiedener Fragestellungen, welche mit der Hilfe von Clustering beantwortet werden können. Ein System zur Unterstützung von Domänenexperten bei der Durchführung der Clusteranalysen muss dem Domänenexperten daher flexibel bei der Beantwortung der unterschiedlichen Fragestellungen assistieren können. Für diese Unterstützung lassen sich Anforderungen identifizieren, welche von einem solchen System zur effektiven Einbindung von Domänenexperten in Clusteranalysen erfüllt werden müssen.

Im Nachfolgenden werden zuerst die identifizierten Anforderungen vorgestellt. Die Notwendigkeit der eingeführten Anforderungen wird anschließend mithilfe eines exemplarischen Anwendungsfalles, dem die verschiedenen Anforderungen zugrunde liegen, veranschaulicht.

### 1.2.1 Identifizierte Anforderungen

Der Domänenexperte besitzt für die Bearbeitung der Fragestellungen relevantes Domänenwissen. Das Domänenwissen kann verschiedene Formen (beispielsweise Abhängigkeiten innerhalb der Daten oder Eigenschaften mit Relevanz für die Aufteilung der Daten) annehmen. Damit der Domänenexperte dieses Domänenwissen für die Verfeinerung von Clustering-Resultaten nutzen kann, muss er (interaktiv) in das Clustering-Verfahren eingebunden werden. Die Anforderungen an eine solche Einbindung des Domänenexperten in Clustering-Verfahren gehen sowohl aus den Fragestellungen (weshalb werden die Clusteranalysen durchgeführt) an Domänenexperten als auch aus den Charakteristiken der Nutzergruppe der Domänenexperten (Abschnitt 1.1.2) hervor.

Folgende Anforderungen werden im Rahmen der Arbeit identifiziert und als Anspruch an das eingeführte Konzept zur Einbindung des Domänenexperten verwendet:

A1 Eigenständige Durchführung der Clusteranalyse

Der Domänenexperte muss in der Lage sein, eine gegebene Fragestellung eigenständig mithilfe seiner inhärenten Fähigkeiten (Abschnitt 1.1.2) zu bearbeiten. Die Erstellung und Verfeinerung des Clustering-Resultates muss daher ohne die Verwendung tiefgreifender technischer Kenntnisse (welche der Domänenexperte nicht besitzt, Abschnitt 1.1.2) umsetzbar sein.

A2 Kompatibilität mit generischen Anwendungsfällen

Die Domänenexperten als Zielgruppe der Arbeit (Abschnitt 1.1) besitzen ihre Expertisen in unterschiedlichen Domänen. Daraus ergibt sich der Bedarf, Clusteranalysen innerhalb verschiedener Domänen durchzuführen. Das System zur Einbindung des Domänenexperten muss daher generisch mit den verschiedenen Anwendungsfällen der Domänenexperten durchführbar sein und darf nicht auf einen bestimmten Anwendungsfall (beispielsweise Kundendaten oder Dokumente) spezialisiert sein.

A3 Wiederholte Anwendung auf (ähnlichen) Aufgaben

Die Durchführung von Clusteranalysen durch Domänenexperten findet oftmals in einem wirtschaftlichen Kontext statt. In einem solchen Kontext existieren häufig ähnliche oder sich wiederholende Fragestellungen, da sich das Ergebnis einer Analyse beispielsweise mit der Zeit (möglicherweise durch die Akquise neuer Kunden) verändern kann. Der Domänenexperte muss daher Clusteranalysen unter Umständen mehrmals auf Basis derselben oder einer ähnlichen Fragestellung anwenden. Dem Domänenexperten stehen zum Zeitpunkt der Analyse außerdem möglicherweise bereits (eventuell aus vorherigen Durchläufen) gespeicherte Ressourcen (beispielsweise bereits geclusterte Daten oder vorhandene Regeln über die Aufteilung) zur Verfügung. Der Domänenexperte muss diese in den Clustering-Prozess einbringen können, um ein bestmögliches Resultat erzielen zu können.

A4 Verfeinerung des Clustering-Resultates

Der Domänenexperte hat aufgrund seiner Expertise eine Vorstellung über die Form, welche das Clustering-Resultat annehmen soll. Für die Erstellung von optimalen Clustering-Resultaten muss der Domänenexperte in der Lage sein, das Clustering-Resultat entsprechend abändern zu können, um diese Vorstellung zu reflektieren. Die Verfeinerung des Clustering-Resultates kann so von der Expertise des Domänenexperten profitieren. Dem Domänenexperten fehlt für die präzise Anwendung der Verfeinerung allerdings die Expertise im Bereich der Datenanalyse, wodurch es zu Fehlern oder verpassten Gelegenheiten (Abschnitt 1.1.2, 1) während der Durchführung der Clusteranalyse kommen kann. Das System muss den Domänenexperten daher zusätzlich während der Clusteranalyse aktiv bei der Vermeidung dieser Fehler und der Identifizierung von Gelegenheiten zur Verfeinerung des Clustering-Resultates unterstützen.

Der Domänenexperte besitzt außerdem eine bedeutende Menge an Domänenwissen (Abschnitt 1.1.2), die manuelle Eingabe des kompletten Domänenwissens während der Clusteranalyse ist aber (zeit- und aufwandstechnisch) nicht realisierbar. Das System zur Einbindung

des Domänenexperten muss deshalb in der Lage sein, selektiv für das Clustering-Resultat besonders nützliche Informationen abzufragen, um die Menge des zur Verfeinerung benötigten Domänenwissens zu reduzieren.

A5 Kommunikation des Clustering-Resultates

Das Verständnis des Clustering-Resultates ist grundlegend für die Einschätzung der Qualität sowie die Identifikation möglicher Probleme, welche eine weitere Verfeinerung erfordern. Aufgrund ihrer fehlenden Expertise im Bereich der Datenanalyse (Abschnitt 1.1.2) kann von Domänenexperten keine eigenständige (tiefgreifende) Analyse des Clustering-Resultates vorausgesetzt werden. Das System muss Domänenexperten daher aktiv über das aktuelle Clustering-Resultat und die damit verbundene Aufteilung der Daten informieren. Die Granularität der gegebenen Informationen muss dem Domänenexperten Schlussfolgerungen über das aktuelle Clustering-Resultat auf Basis seiner Expertise ermöglichen.

### 1.2.2 Exemplarischer Anwendungsfall

In dieser Arbeit wird eine typische Problemstellung, welche nachfolgend vorgestellt wird, als laufendes Beispiel zur Veranschaulichung von Interaktionen mit dem Domänenexperten innerhalb des Konzeptes verwendet.

Die Problemstellung befasst sich mit der Segmentierung von Kundendaten im Rahmen einer Marketingkampagne und veranschaulicht die zuvor eingeführten Anforderungen (Abschnitt 1.2.1) an die Einbindung von Domänenexperten in die Verfeinerung von Clusteranalysen. Das Ergebnis der Segmentierung hängt stark von der Zielgruppe (basierend auf dem vermarkteten Produkt und der Art der Werbemaßnahme) der Marketingkampagne ab. Die Verwendung klassischer Clustering-Verfahren (beispielsweise k-Means) ist daher aufgrund ihrer fehlenden Anpassungsfähigkeit nicht ausreichend. Des Weiteren soll das Ergebnis der Segmentierung von der Expertise des Marketingteams (Abschnitt 1.2.1, A4) profitieren, welches langjährige Erfahrung mit der Betreuung der Kunden hat und sich daher bestens mit den Kundendaten und der Abwicklung von Marketingkampagnen auskennt.

Der Domänenexperte für das Beispiel ist folglich ein Marketingexperte und ist als solcher für das Management von Marketingkampagnen einer Firma verantwortlich. Der Marketingexperte kennt sich bestens mit Marketingstrategien und Kundendaten aus, verfügt jedoch (wie für Domänenexperten typisch) über keinerlei tiefgreifendere technische Expertise im IT- oder Datenanalyse-Bereich. Für die Durchführung der Clusteranalyse benötigt er deshalb (gemäß Abschnitt 1.2.1, A1) einen Clustering-Ansatz, welchen er ohne tiefgreifendere technische Expertise anwenden kann. Als Aufgabe erhält der Marketingexperte das Management einer neuen Marketingkampagne für einen Energydrink. Hierfür soll er eine Clusteranalyse verwenden, um die Kunden zu segmentieren und mit gezielten Marketing-Maßnahmen (beispielsweise Gutscheinen) anzusprechen. Ein auf Domänenexperten zugeschnittenes Clustering-Verfahren (Abschnitt 1.2.1, A2) muss diesen Anwendungsfall des Marketingexperten unterstützen.

Dieser Anwendungsfall besteht entsprechend aus den drei nachfolgenden Schritten:

1. Der Identifikation der Zielgruppe der Marketing-Maßnahme (wer soll von der Werbung angesprochen werden) für das Produkt.

2. Die Segmentierung der Kundendaten auf Basis der Charakteristiken der gewünschten Zielgruppe.
3. Der weiteren Verfeinerung des Clustering-Resultates bis ein zufriedenstellendes Ergebnis erreicht wird.

Ein Beispiel für mögliche Kundendaten stellt Tabelle 1.1 dar. Die Kundendaten setzen sich aus personenbezogenen Daten (beispielsweise Alter, Geschlecht, Lokalität) sowie konsumbezogenen Daten (beispielsweise Umsatz, Anzahl der Einkäufe in einem bestimmten Zeitraum, bevorzugte Produktkategorien) existierender Kunden zusammen.

| Personenbezogen | | | | Konsumbezogen | |
|---|---|---|---|---|---|
| Geschlecht | Alter | Breitengrad | Längengrad | Umsatz (€) | #Einkäufe |
| Männlich | 24 | 51.4508 | 7.0131 | 1955 | 26 |
| Weiblich | 16 | 49.4878 | 8.4661 | 653 | 35 |
| Männlich | 43 | 50.8333 | 12.9167 | 1521 | 42 |
| Männlich | 31 | 52.4000 | 13.0667 | 759 | 14 |
| Weiblich | 17 | 48.3984 | 9.9916 | 4755 | 70 |
| Weiblich | 57 | 48.2119 | 9.0239 | 99 | 1 |

**Tabelle 1.1:** Exemplarische Kundendaten mit Unterteilung der Attribute in personenbezogene und konsumbezogene Attribute.

Der Marketingexperte möchte nun auf Basis der Kundendaten herausfinden, welche Kundengruppen sich besonders für eine bestimmte Marketingkampagne eignen. Hierfür identifiziert er zuerst die wichtigen Charakteristiken der Zielgruppe (Schritt 1).

*Beispiel Charakteristiken von Zielgruppen:* Der Marketingexperte entscheidet sich für das Bewerben eines Energydrinks zur Verteilung von Flyern als Marketingmaßnahme. Der Domänenexperte weiß, dass die Lokalität der Kunden bei der Verteilung von Flyern eine hohe Bedeutung einnimmt. Eine Charakteristik einer idealen Zielgruppe ist daher eine hohe geografische Dichte der Kunden. Eine weitere Charakteristik leitet der Marketingexperte aus dem vermarkteten Produkt (einem Energydrink) ab. Der Marketingexperte identifiziert für Energydrinks das Alter der Kunden (da junge Kunden tendenziell verstärkt an Energydrinks interessiert sind) als weitere wichtige Charakteristik.

Für die Segmentierung der Kundendaten (Schritt 2) muss er Cluster bilden, welche typische Charakteristiken der für ihn (abhängig von der gewählten Marketingkampagne, dem Produkt und der Werbemaßnahme) idealen Zielgruppe aufweisen. Der Marketingexperte muss hierfür sein Wissen über die wichtigen Charakteristiken der Zielgruppen in das Clustering-Verfahren einbringen (Abschnitt 1.2.1, A4).

Das initiale Resultat der Clusteranalyse erfüllt jedoch nicht die Erwartungen des Marketingexperten. Für diese Feststellung muss der Marketingexperte das vom System erstellte Clustering-Resultat verstehen (Abschnitt 1.2.1, A5) können. Der Marketingexperte passt das Clustering-Resultat (Schritt 3) daher weiter an (Abschnitt 1.2.1, A4), bis ein für ihn zufriedenstellendes Endergebnis erreicht wird.

Abschließend soll der Marketingexperte bereits wenige Monate später eine konsekutive Marketingkampagne durchführen. Die Kundendaten, welche der Marketingkampagne zugrunde liegen, haben sich in dieser Zeit allerdings (leicht) verändert. Der Marketingexperte profitiert zur Bewältigung dieser Aufgabe von einem Clustering-Ansatz, welche mit der wiederholten Anwendung von Aufgaben (Abschnitt 1.2.1, A3) umgehen kann.

Die Einbindung der Expertise des Marketingexperten und die Anpassung des Clustering-Resultates sind bei der Verwendung von klassischen Clustering-Verfahren nicht möglich. Diese klassischen Clustering-Verfahre bieten keine Möglichkeit zur direkten Beeinflussung der Clustering-Resultate. Eine Anwendung des populären k-Means Algorithmus erlaubt beispielsweise alleinig die Anpassung der Cluster-Anzahl (k) durch die komplette Neuberechnung des Clustering-Resultates ohne die Möglichkeit Zwischenergebnisse weiter zu verfeinern. Der Marketingexperte benötigt daher einen (interaktiven) Ansatz zur Verfeinerung der Clustering-Resultate, welcher die identifizierten Anforderungen (Abschnitt 1.2.1) erfüllt, um seine Clusteranalyse durchführen zu können.

## 1.3 Ziele

Damit Domänenexperten unter Verwendung ihres Domänenwissens eigenständige Clusteranalysen durchführen können, müssen sie in die interaktive Verfeinerung von Clustering-Resultaten eingebunden werden. Diese Integration muss in der Lage sein, die Anforderungen des Domänenexperten (Abschnitt 1.2.1) zu erfüllen und den Domänenexperten während der Durchführung der Clusteranalyse interaktiv am Verfeinerungsprozess zu beteiligen. Im Rahmen der Arbeit wird auf Basis der Anforderungen des Domänenexperten konformes Konzept für die Einbindung von Domänenexperten eingeführt. Die wesentlichen Beiträge der Arbeit sind:

- Die Definition eines Prozessmodells für die Einbindung von Domänenexperten in Clusteranalysen sowie die Spezifikation der einzelnen Komponenten des Prozessmodells. Dieses Prozessmodell kombiniert existierende interaktive und nicht interaktive Ansätze zu einer holistischen Clustering-Lösung und definiert die Integration von Maßnahmen zur Unterstützung von Domänenexperten.
- Die Demonstration des eingeführten Konzeptes durch einen Prototyp, welcher Domänenexperten die Durchführung interaktiver Clusteranalysen unter der Verwendung ihres Domänenwissens ermöglicht.
- Die Evaluation des eingeführten Konzeptes mittels des Prototyps anhand eines umfassenden Vergleichs zwischen unterschiedlichen interaktiven und nicht interaktiven Techniken auf synthetisch generierten Datensätzen.

## 1.4 Aufbau der Arbeit

Zunächst werden in Kapitel 2 grundlegende Begriffe, die für das Verständnis der folgenden Kapitel notwendig sind, erklärt. Weiterhin werden in diesem Kapitel verwandte Forschungsarbeiten mit ähnlichen Zielsetzungen betrachtet. Kapitel 3 befasst sich mit der interaktiven Einbindung des Domänenexperten in die Durchführung von Clusteranalysen. Im Zuge dessen wird ein Konzept für die Integration von Domänenexperten den Clustering-Prozess entwickelt. Darauf aufbauend wird in Kapitel 4 eine Prototyp des entwickelten Konzeptes präsentiert. Der entwickelte Prototyp wird anschließend in Kapitel 5 evaluiert und vorhandenen Ansätzen gegenübergestellt. Abschließend werden in Kapitel 6 die Ergebnisse dieser Arbeit zusammengefasst und Möglichkeiten zur Weiterentwicklung aufgezeigt.

# 2 Grundlagen und verwandte Arbeiten

Dieses Kapitel erklärt grundlegende Begriffe und Methoden, welche für das spätere Verständnis des vorgestellten Konzeptes zur Einbindung von Domänenexperten in den Clustering-Prozess vorausgesetzt werden. Anschließend werden die in der wissenschaftlichen Literatur bereits vorhandenen Ansätze zur Unterstützung von Domänenexperten identifiziert und vom eingeführten Konzept differenziert.

## 2.1 Clustering

Das Clustering von Daten ist ein unüberwachtes Lernverfahren zur Gruppierung der Instanzen eines Datensatzes [JMF99; Mad12; SPG+17]. Ein unüberwachtes Lernverfahren verwendet während des Lernprozesses (im Gegensatz zu überwachten Lernverfahren) alleinig die gegebenen Daten, ohne auf weitere Informationen über die Korrektheit des Resultates angewiesen zu sein [HTF09, Kapitel 14]. Der Lernprozess ist deshalb auf die inhärenten Eigenschaften der Daten (z.B. Dichteverteilung) für die Inferenz von Ergebnissen angewiesen [JMF99; Mad12; SPG+17].

Das Clustering-Problem besteht darin, eine Menge Instanzen eines Datensatzes in mehrere Cluster zu unterteilen [JMF99; Mad12; SPG+17]. Das Ziel der einzelnen Cluster ist die Zusammenfassung von Instanzen mit ähnlichen Eigenschaften zu einer Kontextgruppe [JMF99; Mad12; SPG+17]. Für die Anwendung von Clustering-Verfahren identifiziert Jain [Jai10] hauptsächlich drei Gründe:

1. *Erkundung von Datenstrukturen*, um Einsicht in Daten zu erhalten, Abnormalitäten zu finden, Eigenschaften von Gruppen zu identifizieren und Hypothesen aufzustellen.
2. *Gruppierung*, zur Identifizierung von Ähnlichkeiten zwischen Instanzen.
3. *Kompression*, zur Beschreibung und Organisation von Daten.

Für die Lösung des Clustering-Problems schlägt die Literatur zahlreiche Clustering-Techniken zur Bildung von Clustern vor [Jai10]. Diese unterscheiden sich grundsätzlich in der verwendeten Vorgehensweise, welche zur Bildung der Cluster verwendet wird [FAT+14]. Die existierenden Clustering-Techniken werden basierend auf ihrer technischen Vorgehensweise zur Clusterbildung von Fahad et al. [FAT+14] in fünf Kategorien aufgeteilt [FAT+14]. Diese Kategorien setzen sich aus *partitionierenden*, *hierarchischen*, *dichtebasierten*, *rasterbasierten* und *modellbasierten* Clustering-Techniken zusammen. Im Rahmen dieser Arbeit sind vor allem die *partitionierenden* und die *hierarchischen* Clustering-Ansätze von Bedeutung. Die beiden Clustering-Ansätze werden daher im Folgenden vorgestellt.

*Partitionierende-Ansätze* unterteilen Daten direkt in eine festgelegte Anzahl an Clustern durch die (iterative) Optimierung einer Zielfunktion (beispielsweise den Euklidischen Abstand) [FR98; Jai10; SPG+17]. Der Clustering-Algorithmus verschiebt zu diesem Zweck, solange Instanzen zwischen den gebildeten Clustern, bis ein (lokales) Optimum der Zielfunktion gefunden wird [FAT+14; SPG+17].

*Hierarchische-Ansätze* ordnen die Daten in einer hierarchischen Struktur (abhängig von ihrer Nähe zueinander) an [FAT+14]. Das initiale Cluster (welches alle Instanzen beinhaltet) wird innerhalb der Hierarchie schrittweise in kleinere Gruppen unterteilt [FAT+14]. Die Literatur unterteilt hierarchische Ansätze weiter in divisive (top-down) und agglomerative (bottom-up) Ansätze [FAT+14; FR98; Jai10; SPG+17]. Der Unterschied zwischen divisiven und agglomerativen Verfahren ist die Richtung, in welche die hierarchische Struktur erstellt wird [FR98; Jai10; SPG+17]. Die divisiven Verfahren beginnen mit einem Cluster, welches alle Daten umfasst und unterteilen dieses Cluster schrittweise weiter auf, während agglomerative Ansätze mit vielen Clustern (typischerweise ein Cluster pro Instanz) beginnen und schrittweise mehrere Cluster zusammenführen (mergen) [FR98; Jai10; SPG+17]. Die Verwendung von agglomerativen Verfahren, führt (im Allgemeinen) zu besseren Ergebnissen, benötigt allerdings auch mehr Speicherplatz und Rechenleistung als die Anwendung divisiver Verfahren[FR98].

Die Auswahl der Clustering Technik und eines zugehörigen Algorithmus kann das Clustering-Resultat entscheidend beeinflussen [JMF99; SPG+17; TFS21]. Bei der Wahl des Algorithmus schließt die Literatur die Existenz eines universell „besten“ Algorithmus klar aus und stellt einen klaren Zusammenhang zwischen der Performance und der Art der Daten/des Anwendungsfalls fest [JMF99; LWG11; SPG+17]. Damit ein möglichst optimales Clustering-Resultat erreicht wird, muss daher abhängig vom Anwendungsfall ein geeigneter Clustering-Algorithmus ausgewählt werden [JMF99]. Eine genaue Differenzierung der Güte verschiedener Clustering-Algorithmen für allgemeine Anwendungsfälle stellt allerdings selbst die wissenschaftliche Literatur vor eine Herausforderung, da das Ergebnis der Evaluation des Clustering-Resultats stark von den Daten und der Art der Evaluation abhängt [LWG11]. Den populärsten Clustering-Algorithmus stellt der k-Means Algorithmus dar [Jai10]. Ein partitionierender Ansatz, welcher sich durch seinen einfachen Aufbau, Effizienz und bewährte Qualität der Clustering-Resultate in diversen Anwendungsbereichen auszeichnet [Jai10]. Für allgemeine Anwendungsfälle wird daher aufgrund schwierigen Differenzierung weiterhin (trotz der Existenz zahlreicher neuerer Clustering-Verfahren) oftmals eine Variationen des k-Means Algorithmus verwendet [Jai10].

## 2.2 Kommunikation von Clustering-Resultaten

Die menschliche Wahrnehmung limitiert den Vergleich von Clustering-Resultaten auf den zwei- bis dreidimensionalen Raum [BPB+16; CGSQ11]. Für hochdimensionale Ergebnisse stellt das Verständnis und die Interpretation von Clustering-Resultaten daher eine Herausforderung dar [BHTM22; BPB+16; DFMR20; ESG+21]. Die Bedeutsamkeit dieser Herausforderung leitet sich aus der Notwendigkeit des Verständnisses von Clustering-Resultaten durch Domänenexperten für die Realisierung belastbarer Analysen ab [BHTM22]. Es existieren daher mehrere Ansätze, um diese Herausforderung zu adressieren und den Domänenexperten bei dem Verständnis von Clustering-Resultaten zu unterstützen. Im Folgenden wird ein Überblick über diese Ansätze gegeben:

**Visualisierung:** Die oftmals zur Visualisierung von Clustering-Resultaten verwendeten Scatterplots eignen sich hervorragend zur Interpretation zweidimensionaler Ergebnisse [BHR+20; CGSQ11]. Mithilfe geeigneter Visualisierungstechniken (z.B. Parallele Koordinaten, Scatterplot-Matritzen) können Visualisierungen jedoch auch Einsicht in mehrdimensionale Clustering-Resultate verschaffen [BPB+16; CGSQ11]. Die erzielte Einsicht in Clustering-Resultate kann durch die Integration interaktiver Techniken in Visualisierungen (z.B. Drill-down/Drillup für hierarchische Clustering Verfahren) zusätzlich vertieft werden [CD18; SS02].
Ein weiterer Ansatz ist die Verwendung von Visualisierungen, um nicht das Clustering-Resultat selbst, sondern den Entscheidungsprozess („weshalb wird eine Instanz einem bestimmten Cluster zugewiesen“) abzubilden [BOW18; FGS13; LXY00]. Hierfür können Entscheidungsbäume verwendet werden, die die Zuweisung von Instanzen klar nachvollziehbar machen [BOW18; DFMR20; FGS13; LXY00]. Die Entscheidungsbäume können entweder von der angewendeten Clustering Technik generiert werden oder nachträglich anhand des Clustering-Resultats mithilfe von überwachtem Lernen trainiert werden [CD18; FGS13; LXY00].

**Dimensionsreduktion:** Die Unterstützung des Domänenexperten bei dem Verständnis von Clustering-Resultaten kann durch die Reduktion der zu betrachtenden Dimensionen realisiert werden [CEM+15; ESG+21]. Für die Reduktion der Dimensionalität kann zwischen zwei Ansätzen (der Feature-Extraktion und der Feature-Selektion) unterschieden werden [TAL14]. Die Feature-Extraktion kann durch mathematische Projektionen (z.B. PCA, Random-Projection) der originalen Features in niedrig-dimensionale Räume umgesetzt werden [CEM+15]. Bei der Anwendung von Feature-Extraktions-Ansätzen geht allerdings die originale Bedeutung der einzelnen Dimensionen aufgrund der Kombination mehrerer Dimensionen während der Projektion verloren [CEM+15; DFMR20; TAL14]. Dies erschwert die Interpretation des Clustering-Resultats durch den Domänenexperten [CGSQ11; DFMR20; TAL14]. Die Feature-Selektion selektiert (eine Teilmenge) besonders relevante Dimensionen des Datensatzes und schränkt so die Anzahl der zu betrachtenden Dimensionen ein [TAL14]. Die Selektion der relevanten Dimensionen wird auf Basis der Relevanz einzelner Dimensionen für das Ergebnis getroffen [ESG+21; TAL14]. Durch das „Aussortieren“ weniger/nicht relevanter Informationen erhöhen solche Ansätze das Verständnis von Clustering-Resultaten, ohne die Interpretationsfähigkeit einzelner Dimensionen zu beeinflussen [BHTM22; CEM+15]. Die Verwendung von Ansätzen zur Feature-Selektion ist daher in Bezug auf die Lesbarkeit und Interpretierbarkeit von Resultaten überlegen, leidet allerdings auch unter einem größeren Informationsverlust (da nicht relevante Dimensionen komplett verworfen werden) gegenüber der Feature-Extraktion [BHTM22; CEM+15; ESG+21; TAL14].

**Metrische Evaluation:** Außerhalb von Methoden zur Dimensionsreduktion können Domänenexperten durch eine metrische Abstraktion der Qualität von Clustering-Resultaten in Form von Indikatoren (z.B. Silhouetten [Rou87], Dunn-Index [Dun73]) unterstützt werden. Aufgrund des unüberwachten Charakters des Clustering-Problems können diese Indikatoren jedoch nur eine Abschätzung der Qualität anhand von Merkmalen wie der Kompaktheit und Separation von Clustern erreichen [HJK17]. Die Güte, mit welcher ein Indikator die Qualität eines Clustering-Resultates abschätzen kann, ist daher von der Art der zugrundeliegenden Daten abhängig [HJK17] und berücksichtigt nicht die inhärente Subjektivität (Abschnitt 2.4) von Clustering-Resultaten [BHR+20; BHTM22]. Alleinstehend sind Indikatoren daher nicht

zur Kommunikation von Clustering-Resultaten geeignet [BHTM22], können aber (vor allem bei Kombination mehrerer Indikatoren) bei der Evaluation von Clustering-Resultaten helfen [HJK17].

## 2.3 Active-Learning

Die Verwendung von unüberwachten Lernverfahren (beispielsweise klassische Clustering-Verfahren) ist auf die aus den Daten lernbaren Merkmale beschränkt [FZL13]. Die Anwendung klassischer überwachter Lernverfahren kann zu verbesserten Resultaten führen, erfordern allerdings vollständig gelabelte Datensatze, um ihre Modelle zu trainieren [AKG+14; DH08; FZL13; KG20; Set09; SW10]. In vielen Fällen ist die Sammlung einer ausreichenden Menge an gelabelten Daten jedoch nicht praktikabel oder zu kostspielig [AKG+14; FZL13; Set09; SW10]. In solchen Fällen können halb-überwachte Ansätze wie Active-Learning Verfahren verwendet werden [AKG+14; FZL13; KG20; Set09; SW10]. Active-Learning Verfahren bestehen aus einem Orakel, einem Query System und einem Machine Learning Modell [AKG+14; SW10]. Das Orakel (oftmals ein Nutzer) kennt die korrekten Label aller Instanzen und kann nach diesen befragt werden [AKG+14; Set09]. Während des Lernprozesses wird das Orakel von diesem Query System zu den Labeln einzelner Instanzen befragt [AKG+14; Set09]. Auf Basis dieser Antworten wird daraufhin ein Machine Learning Modell trainiert [AKG+14; Set09]. Ziel ist die Generation eines hochakkuraten Modells mit möglichst wenig Abfragen des Orakels [Set09]. Kritisch ist in diesem Zusammenhang die gezielte Selektion der abzufragenden Instanzen durch das Query System, um ein Modell zu generieren, das deutlich performanter ist, als bei einer rein zufälligen Auswahl von Instanzen [FZL13; Set09]. Das zentrale Problem des Query Systems ist daher die Identifikation der Instanzen, welche bei Abfrage die Genauigkeit des Modells maximieren [AKG+14; FZL13; Set09; SW10]. Für die Auswahl der abzufragenden Instanzen ergeben sich drei grundlegende Szenarien [KG20; Set09]:

**Membership Query synthesis-basiert:** Anhand des gelernten Modells generiert das Query System neue (nicht im Datensatz enthaltene) Instanzen nahe der gelernten Entscheidungsgrenze [KG20; Set09].

**Stream-basiert:** Das System erhält einen Datenfluss (Stream) oder eine Sequenz ungelabelter Instanzen und entscheidet, während die Daten eintreffen, ob das Orakel befragt werden soll [KG20; Set09].

**Pool-basiert:** Das System erhält eine Menge nicht gelabelter Daten und muss aus dieser die Instanzen zur Befragung des Orakels selektieren [KG20; Set09].

Im Rahmen dieser Arbeit ist aufgrund der Aufgabenstellung vor allem das pool-basierte Szenario von Bedeutung.

Das pool-basierte Szenario kann durch zahlreiche Strategien zur Selektion der bestmöglichen Instanzen verwirklicht werden [FZL13; Set09; SW10]. Diese Strategien unterscheiden sich in der Methodik, mit welcher der „Wert“, der die Abfrage einzelner Instanzen schafft, bewertet wird [FZL13; Set09; SW10]. Einige Beispiele solcher Methodiken sind uncertainty sampling (selektiert die Instanz, bei welcher sich das Modell am unsichersten ist), density-basierte Gewichtung (selektiert nicht nur nach uncertainty, sondern gewichtet aufgrund von Ausreißern zusätzlich nach Repräsentativität der Instanzen) und expected model change (selektiert Instanzen,

bei welchen sich das Modell (erwartet) am stärksten verändert) [KG20; Set09; SW10]. Die Leistung einzelner Selektionsstrategien hängt vom Anwendungsfall ab und folglich existiert keine im Allgemeinen „beste“ Strategie [AKG+14].

## 2.4 Interaktives Clustering

Techniken zur Lösung des Clustering-Problems (Abschnitt 2.1) verwenden oftmals interaktive Elemente während des Clustering-Prozesses [AA20; AL14; BHR+20; BPB+16; CD18; CDH+16]. Diese Sektion fasst die Motivation hinter der Verwendung von interaktiven Elementen während des Clustering-Prozesses zusammen und grenzt verschiedene Arten von Interaktionen während des Clustering-Prozesses voneinander ab. Grundlage der in dieser Sektion vorgestellten Kategorisierungen ist die Untersuchung existierender interaktiver Clustering-Ansätze von Bae et al. [BHR+20]. Diese identifizieren vier grundsätzliche Ziele der Integration interaktiver Elemente in den Clustering-Prozess:

**Verbesserung von Clustering-Resultaten:** Zielt auf die Kollaboration von Mensch und Maschine, um die Qualität der Clustering-Resultate zu verbessern. Motiv der Kollaboration ist die Nutzung von Domänenwissen zur Unterstützung des Clustering-Prozesses durch fachkundige Expertise.

**Einsicht in den Clustering-Prozess erhalten:** Ziel ist es, dem Domänenexperten zu erklären, was während dem Clustering-Prozess passiert und wie das Clustering-Resultat zustande kommt.

**Interessante Daten finden:** Identifikation besonders wichtiger Daten und relevanter Zusammenhänge in Daten sowie Verifikation von Hypothesen.

**Subjektivität des Clustering-Resultates:** Die Qualität von „Lösungen“ einer Clustering-Aufgabe hängt von dem Kontext (Erwartungen und Bedürfnissen des Nutzers) der Analyse ab. Aufgabe der Interaktion ist es daher, die für den Nutzer „beste Lösungen“ aus der Menge der möglichen Lösungen zu finden.

Interaktionsmöglichkeiten zur Erfüllung dieser Ziele unterteilen Bae et al. [BHR+20] in drei Kategorien, basierend auf der Art, durch welche mit dem Clustering-Prozess interagiert wird:

**Interaktion mit Parametern:** Der Nutzer passt Parameter (z.B. Anzahl der Cluster, Gewichtungen) des Clustering-Prozesses an. Der Domänenexperte muss die Entscheidung treffen, welche Parameter anzupassen sind, um das Resultat zu verfeinern.

**Interaktion mit dem Ergebnis:** Bei dieser Interaktionsart wird dem Nutzer zuerst ein Clustering-Resultat präsentiert. Der Nutzer kann dieses mittels bereitgestellter Operationen (beispielsweise die Zuweisung einer Instanz zu einem spezifischen Cluster) weiter verfeinern. Im Gegensatz zu Ansätzen, bei welchen der Domänenexperte mit Parametern interagiert, modelliert der Nutzer bei dieser Interaktionsart das gewünschte Ergebnis, welches anschließend vom System interpretiert und umgesetzt wird und nimmt keine direkten Anpassungen an Einstellungen des Clustering-Verfahrens vor.

**System-initiierte Abfrage von Informationen:** Das System übernimmt eine aktive Rolle während des Clustering-Prozesses und lenkt die Interaktionen des Nutzers, um beispielsweise die korrekte Zuweisung einzelner Instanzen abzufragen.

Die Interaktionsmöglichkeiten der einzelnen Kategorien haben unterschiedliche Stärken und Schwächen [BHR+20]. Durch die Kombination interaktiver Möglichkeiten mehrerer Kategorien können sie sich gegenseitig ergänzen [BHR+20]. Beispielsweise ermöglichen Interaktionen mit Parametern viel Kontrolle über das Clustering-Resultat, benötigen allerdings auch ein tiefgreifenderes Verständnis über den angewendeten Clustering-Algorithmus, als Interaktionen anderer Kategorien (beispielsweise mit dem Ergebnis) [BHR+20]. Im Rahmen dieser Arbeit werden die vorgestellten Kategorien verwendet, um vorhandene Ansätze zu kategorisieren und das Ausmaß an Interaktionsmöglichkeiten im vorgestellten Konzept einzuordnen.

## 2.5 Verfeinerung von Clustering-Resultaten

Die wissenschaftliche Literatur identifiziert eine Vielzahl an Methodiken zur Verfeinerung von Clustering-Resultaten [AA20; BBM04a; BHR+20; SZS16; XJC14]. Dieser Abschnitt präsentiert zwei grundlegende Arten der Verfeinerung, welche für das Verständnis des später eingeführten Konzeptes eine tragende Rolle spielen.

### 2.5.1 Constraint-basiertes Clustering

Eine Möglichkeit, Clustering-Verfahren zu verfeinern, ist die Definition von Constraints auf den Daten, welche zur Clusterbildung verwendet werden [BHR+20; CB17; CDG+17; LJJ07; MD18; THLN01]. Diese Constraints stellen Restriktionen dar, die bei der Bildung des Clustering-Resultates beachtet werden müssen [CB17; CDG+17; LJJ07]. Solche Restriktionen können auf verschiedenen Ebenen des Clustering-Verfahrens definiert werden [MD18; THLN01]. Grundlegend lassen sich diese Ebenen unterteilen in die Ebene der Instanzen (z.B. Beziehung zwischen mehreren Instanzen), Ebene der Cluster (z.B. Restriktion von Kardinalitäten) und Ebene der Parameter (z.B. maximale Anzahl an Clustern) [MD18; THLN01]. Constraints können auf Basis von existierendem Domänenwissen (Abschnitt 1.1.1) definiert werden und ermöglichen so die Nutzung von Hintergrundinformationen zur Verbesserung des Clustering-Resultates [CB17; CDG+17].

Häufig verwendete Constraints definieren Instanzen, welche zusammen in einem Cluster zugeordnet werden sollen (Must-Link) und Instanzen, welche unterschiedlichen Clustern (Cannot-Link) zugeordnet werden sollen [AA20; BHR+20; LJJ07; MD18]. Das mithilfe der Constraints definierte paarweise Verhältnis zwischen den Instanzen wird während der Anwendung des Clustering-Verfahrens verwendet, um die Aufteilung der Cluster zu beeinflussen.

Ein Beispiel, wie das Clustering-Resultat durch die Verwendung von Must-Link und Cannot-Link Constraints beeinflusst werden kann, stellt Abbildung 2.1 dar. Diese zeigt ein initiales Clustering-Resultat in a) mit der Anpassung dieses Resultats in b) und c), nachdem durch Hinzufügen von Constraints die vertikale Separierung der Cluster erzwungen wurde. Für die vertikale Separierung unter Verwendung von Constraints existieren mehrere Möglichkeiten. Exemplarisch werden in b) und c) zwei mögliche Alternativen zur Separierung der Cluster unter der Verwendung von Must-Link Constraints b) und Cannot-Link Constraints c) dargestellt.

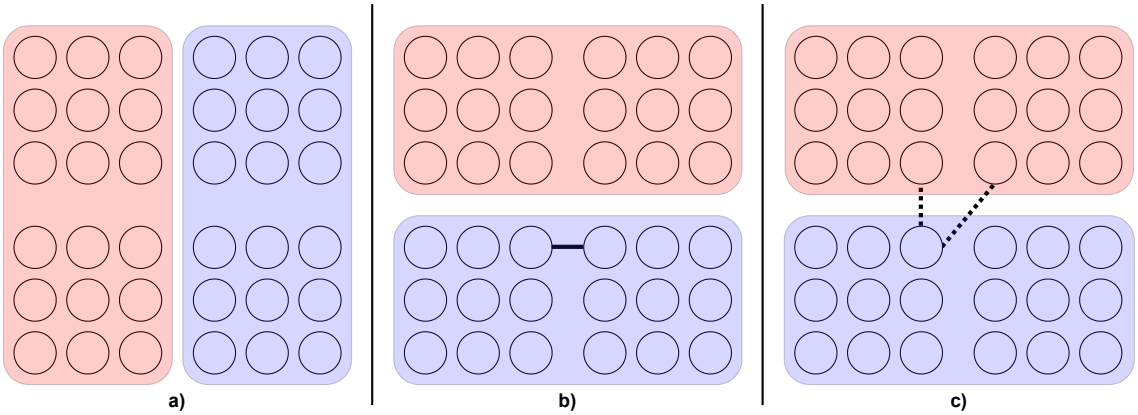



**Abbildung 2.1:** Beispiel der Anwendung von b) Must-Link Constraints (durchgezogene Linie) und c) Cannot-Link Constraints (gestrichelte Linie) zur Beeinflussung des ursprünglichen Cluster-Resultats a) (Beispiel basierend auf [AA20]).

### 2.5.2 Active-Clustering

Active-Clustering kombiniert Constraint-basiertes Clustering (Abschnitt 2.5.1) mit Active-Learning (Abschnitt 2.3) zur Verfeinerung von Clustering-Resultaten [BBM04a; BHR+20]. Bei dieser Kombination werden Query-Strategien aus der Active-Learning Domäne verwendet, um „intelligent“ Constraints für ein Constraint-basiertes Clustering Verfahren zu erzeugen [AA20; BBM04a; KG20; VD16; WD10]. Ein generisches Beispiel für die Verwendung von Active-Clustering-Verfahren zur Erzeugung von Constraints wird in Abbildung 2.2 gezeigt.

Der Clustering-Algorithmus erhält in diesem Beispiel zuerst einen Datensatz (optional mit bereits vorhandenen Constraints) zur Berechnung des Clustering Resultates. Abhängig von der Art der verwendeten Constraints (beispielsweise Restriktion einzelner Instanzen, paarweise Beziehung zwischen Instanzen oder Instanz Triple) wird ein Orakel (häufig der Nutzer), bei Active-Clustering zu der Zugehörigkeit einer Instanz oder dem Verhältnis mehrerer Instanzen zueinander, befragt [KG20; WD10]. Häufig verwenden Active-Clustering-Verfahren das Verhältnis zwischen zwei Instanzen durch die Definition von Must-Link und Cannot-Link Constraints (Abschnitt 2.5.1), welche aus der paarweisen Abfrage von Instanzen (gehören die Instanzen demselben Cluster an oder nicht) generiert werden können [AA20; BBM04a; WD10; XJC14]. Auf Basis der Rückmeldungen des Orakels werden dann Constraints generiert, welche genutzt werden, um das Ergebnis (iterativ) zu verfeinern[AA20; XJC14]. Das gegebenen Beispiel (Abbildung 2.2) zeigt die Abfrage des Nutzers, sowie die Generation neuer Constraints auf Basis der Interaktion (Beantwortung der System-Abfrage), welche zur Erstellung neuer Regeln und anschließend zur Aktualisierung des Clustering-Algorithmus verwendet werden. Die Herausforderung der Abfragestrategie ist die Identifizierung der (paarweisen) Instanzen, die bei der Abfrage die Verbesserung des Cluster-Resultats maximieren [AL14; WD10]. Aufgrund der unüberwachten Natur des Clustering-Problems (Abschnitt 2.1) kann die Auswirkung (Verbesserung) der Anwendung einer bestimmten Regel nicht präzise bestimmt werden [BBM04a; KG20; WD10; XJC14]. Für die Auswahl der abzufragenden Constraints können von Active-Clustering-Verfahren daher verschiedene Strategien zur Selektion der bestmöglichen Instanzen (maximale Verbesserung des Clustering-Resultates) verwendet werden. Die Strategien verwenden zur

Selektion der abzufragenden Instanzen beispielsweise den Abstand zwischen Instanzen [BBM04a], die erwartete Fehlerreduktion [WD10] oder Instanzen, für welche das Clustering-Modell einen hohen Grad an Unsicherheit bezüglich ihrer Zuweisung aufweist [XJC14].

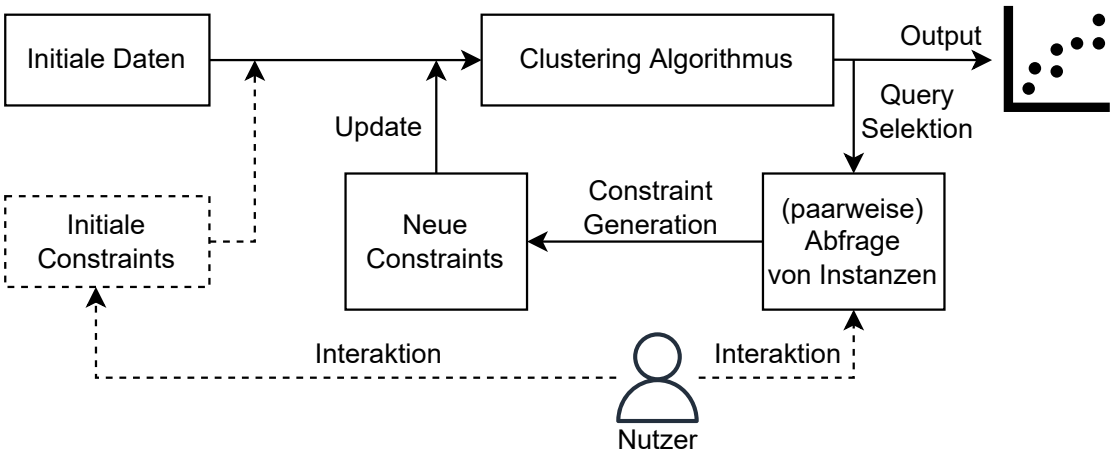



**Abbildung 2.2:** Prozessmodell eines generischen Constraint-basierten Active-Clustering Ablaufes (basierend auf [AA20; KG20; XJC14]).

Ansätze, welche Active-Clustering verwenden, unterscheiden sich im Wesentlichen in drei Aspekten:

**Art der generierten Constraints:** Verschiedene Active-Clustering Methoden verwenden unterschiedliche Arten von Constraints, welche aus den Nutzerabfragen generiert werden [AA20; BBM04a; DH08; EDSN11; VD16; XJC14].

**In der Query Strategy:** Kongruent zu Query-Strategien von Ansätzen der übergeordneten Active-Learning Domäne (Abschnitt 2.3) existieren zahlreiche Strategien zur Selektion der für die Abfrage bestmöglichsten Instanzen für Active-Clustering [KG20].
Im Besonderen können Verfahren, welche Instanzen paarweise abfragen (z.B. für häufig verwendete Must-Link, Cannot-Link Constraints), weiter in sample-based und sample-pair-based Methoden unterteilt werden [XJC14]. Diese unterscheiden sich durch die Art, auf welche Paare generiert werden. Bei sample-based Methoden wird zuerst eine (interessante) Instanz ausgewählt, für welche darauffolgend eine für die Abfrage möglichst optimale zweite Instanz (lokale Optimierung ausgehend von einer zufällig gewählten Instanz) gesucht wird [XJC14]. Die Methoden, welche für die Selektion der paarweisen Instanzen mit sample-pair-based Verfahren selektieren, suchen direkt nach dem für die Abfrage besten Paar zweier Instanzen (globale Optimierung über alle möglichen Paare) [XJC14]. Auf diese Weise kann das beste Paar gefunden werden, was allerdings auch zusätzliche Rechenleistung benötigt [XJC14].

**In der Clustering Implementierung:** Die von der Implementierung verwendete Clustering Technik (z.B. partitionierend oder dichtebasiert) sowie der zur Implementierung der Clustering Technik verwendete Clustering-Algorithmus (z.B. k-means, DBSCAN) [KG20; SPG+17; XJC14].

## 2.6 Verwandte Arbeiten

In der Literatur werden zahlreiche Ansätze vorgeschlagen, um Domänenexperten in Clustering-Verfahren zu integrieren [**interact_steering_hierarch_clustm clustrophile2**; BHR+20; CDG+17] und bei dem Verständnis von Clustering-Resultaten zu unterstützen [BHTM22; ESG+21; SS02].

Die verwandten Arbeiten zur Einbindung des Domänenexperten in Clustering-Verfahren lassen sich in die bereits vorgestellten Interaktionskategorien (Abschnitt 2.4) zur Verfeinerung von Clustering-Resultaten unterteilen. Der Domänenexperte wird von diesen Ansätzen durch die eigenständige Anpassung von Parametern (beispielsweise durch die Gewichtung von Attributen) des Clustering-Verfahrens oder die Modellierung des Clustering-Resultates (beispielsweise durch das Zusammenführen von Clustern) interaktiv in die Verfeinerung des Resultates eingebunden. Außerdem werden Ansätze, bei welchen das System die Interaktion initiiert (beispielsweise bei Active-Clustering), diskutiert. Zusätzlich zu den interaktiven Ansätzen baut die Arbeit auf Ansätzen auf, welche vorhandene Ressourcen (beispielsweise Daten oder Constraints) zur Verbesserung von Clustering-Resultaten nutzen. Neben den Ansätzen zur Verfeinerung des Clustering-Resultates werden relevante Ansätze zur Kommunikation des Clustering-Resultates mit dem Domänenexperten behandelt.

Im Rahmen dieses Abschnittes werden zuerst verwandte Arbeiten zur Verfeinerung von Clustering-Resultaten (bezogen auf die Anforderungen des Domänenexperten aus Abschnitt 1.2.1, A1-A4) vorgestellt. Anschließend wird die zur Einbindung des Domänenexperten notwendige Kommunikation (Abschnitt 1.2.1, A5) der Resultate diskutiert. Die Relevanz der vorgestellten Ansätze für das Konzept der Arbeit wird mithilfe der Anforderungen des Domänenexperten an die Einbindung in Clustering-Verfahren (Abschnitt 1.2.1) begründet.

### 2.6.1 Verfeinerung von Clustering-Resultaten

Die relevanten Arbeiten dieses Abschnittes beteiligen Domänenexperten interaktiv an Clustering-Verfahren. Die Relevanz der Arbeiten leitet sich aus den Bedürfnissen des Domänenexperten ab, sein Domänenwissen in Clustering-Resultate einfließen zu lassen (Abschnitt 1.2.1, A4).

Der Abschnitt behandelt außerdem standardmäßig nicht-interaktive Ansätze (beispielsweise Seeding oder Constraints), welche in Kombination mit interaktiven Verfahren eingesetzt werden können, um zusätzlich (zu eventuell gebotenen Interaktionen) vorhandene Ressourcen bei der Berechnung des Clustering-Resultates zu berücksichtigen und somit das korrespondierende Bedürfnis (Abschnitt 1.2.1, A3) des Domänenexperten zufriedenzustellen.

Im Folgenden werden zuerst die relevanten interaktiven Ansätze aus den eingeführten Interaktionskategorien vorgestellt und anschließend ergänzende Ansätze zur Berücksichtigung vorhandener Ressourcen dargestellt.

#### Interaktive Beteiligung an Clustering-Verfahren

Die relevanten interaktiven Ansätze werden nach ihren zugehörigen Interaktionskategorien (Abschnitt 2.4) aufgeführt.

*Anpassung von Parametern.* Die relevanten Ansätze aus der Interaktionskategorie der Anpassung von Parametern ermöglichen Domänenexperten, die für die Anwendung von Clustering-Verfahren verwendete Konfiguration anzupassen [BHR+20; CD18]. Die Elemente, mit welchen interagiert wird, gestalten sich vielzählig und können sich etwa auf die Parameter des Clustering-Algorithmus (beispielsweise die Cluster-Anzahl), den Clustering-Algorithmus selbst oder eine zugrundeliegende Abstandsmetrik beziehen [BHR+20; CD18].

Einen interessanten Ansatz dieser Kategorie stellen Marco Cavallo und Çağatay Demiralp mit Clustrophile 2 [CD18], welcher Domänenexperten schrittweise durch die Anwendung von Clustering-Verfahren führt und die Festlegung von Parametern durch metrische Indikatoren und Beschreibungen vereinfacht. Der Domänenexperte kann bei diesem Ansatz iterativ Parameter abändern und die erstellten Cluster-Resultate untereinander vergleichen [CD18].

Mehrere Ansätze verwenden zusätzlich Gewichtungs-Parameter, um den Einfluss ausgewählter Eigenschaften (beispielsweise Attribute oder Dokument-Themen) auf das Clustering-Verfahren zu regulieren [BBR14; CLRP13]. Der Domänenexperte verwendet bei diesen Ansätzen sein Domänenwissen zur Erstellung von Gewichten, welche (basierend auf seiner Erfahrung) die Relevanz einzelner Dateneigenschaften für seine aktuelle Clusteranalyse reflektieren [BBR14; CLRP13]. Der (dokumentspezifische) Ansatz von Choo et al. [CLRP13] ermöglicht Domänenexperten beispielsweise, den Einfluss einzelner Topics (Themenbereiche) auf die Unterteilung von Dokumenten festzulegen. Das Domänenwissen des Experten kann außerdem für die isolierte Betrachtung interessanter Attributskombinationen während der Clusteranalyse verwendet werden [HBS+16].

Bei der Festlegung der Parameter kann der Domänenexperte von automatisierten Empfehlungen des Systems unterstützt werden. So kann beispielsweise eine Annäherung der optimalen Cluster-Anzahl mithilfe der Elbow-Methode erzielt [BHR+20; BHTM22] oder anhand mathematischer Eigenschaften der Daten ein geeigneter Clustering-Algorithmus vorgeschlagen werden [CD18].

*Interaktion mit dem Clustering-Resultat.* Die relevanten Ansätze aus der Interaktionskategorie der Interaktion mit dem Clustering-Resultat ermöglichen dem Domänenexperten die direkte Modellierung von Änderungen, ohne eigenständig die Parameter des Clustering-Verfahrens anpassen zu müssen [BHR+20]. Diese Interaktionskategorie umfasst eine Vielzahl relevanter Ansätze mit unterschiedlichen Vorgehensweisen [ABV13; CDH+16; EFS11; LKC+12; SZS16; TPRH11].

Die Interaktionsmöglichkeiten dieser Kategorie werden oftmals in der Form von Operationen, welche vom Domänenexperten auf das berechnete Clustering-Resultat angewendet werden können, umgesetzt [ABV13; CDH+16; EFS11; LKC+12; TPRH11]. Die Anwendung der Operationen ermöglicht dem Domänenexperten, den Clustering-Prozess mithilfe seines Domänenwissens zu „lenken“, um das Clustering-Resultat seinen Vorstellungen anzupassen [CDH+16; EFS11]. Die verwandten Arbeiten identifizieren mehrere mögliche dieser anwendbaren Operationen [ABV13; EFS11; LKC+12; TPRH11]. Diese umfassen das Aufsplitten/Mergen einzelner Cluster, die Zuweisung einzelner Instanzen zu Clustern, das Entfernen spezifischer Cluster und die Neuberechnung des Clustering-Resultates (beispielsweise mit unterschiedlichen Cluster-Zentren) [ABV13; CDH+16; EFS11; LKC+12; TPRH11].

Einen grundlegend unterschiedlichen Ansatz zur Anpassung von Clustering-Resultaten ist die Option der (teilweisen) Ablehnung von Clustering-Resultaten [LKS+15; SZS16]. Bei dieser Vorgehensweise bewertet der Domänenexperte, ob er einzelne Cluster annimmt/ablehnt [SZS16] oder mehrere Clustering-Resultate miteinander vergleicht [LKS+15]. Das System merkt sich die Merkmale

der abgelehnten Cluster und erstellt daraufhin auf Basis der gemerkten Merkmale ein neues Clustering-Resultat mit Clustern, welche sich (möglichst stark) von den zuvor abgelehnten Clustern unterscheiden [SZS16]. Der Domänenexperte verwendet bei diesen Ansätzen die wiederholte Ablehnung oder Bewertung, um das Clustering-Resultat seinen Vorstellungen anzupassen.

*System-initiierte Interaktionen.* Innerhalb der relevanten Ansätze aus der Interaktionskategorie der System-initiierten Interaktionen leitet das System eigenständig die Anwendung von Interaktionen ein [BHR+20]. Dieser Interaktionskategorie gehören sowohl Active-Clustering Ansätze [AA20; BBM04a; XJC14] als auch Ansätze, welche Empfehlungen zur Unterstützung von Domänenexperten verwenden [CD18; DFB11], an. Für diese Arbeit sind die Active-Clustering-Verfahren (Abschnitt 2.5.2) aufgrund ihrer interaktiven Einbindung von Domänenexperten durch die Abfrage von Informationen zur Generierung von Constraints (Abschnitt 2.5.1) besonders relevant. Diese Abfrage befragt den Domänenexperten „intelligent“ nach für das System wichtigen Informationen und erfüllt daher das Bedürfnis des Domänenexperten, vom System aktiv bei der Bereitstellung relevanter Informationen unterstützt zu werden (Abschnitt 1.2.1).

In der Literatur lassen sich zahlreiche Active-Clustering Ansätze identifizieren, welche Active-Learning verwenden, um den Domänenexperten „intelligent“ über das Verhältnis (Abschnitt 2.5.2) selektierter Instanzen zu befragen [AA20; AL14; BBM04a; WD10; XJC14]. Diese Ansätze ergänzen (reines) Constraint-basiertes Clustering durch die zusätzliche Erstellung von Constraints während der Clusteranalyse [AA20; BBM04a; XJC14]. Hierfür wird der Domänenexperte nach Informationen befragt [AA20; BBM04a; XJC14]. Einen populären Ansatz stellen Basu et al. [BBM04a], welche mit PCKMeans eine paarweise Constraint-basierte Version (Must-/Cannot-Link) des populären k-Means Algorithmus einführen und auf Basis von „Farthest-first traversal“ (beruhend auf dem maximalen Abstand) Instanzen zur Abfrage auswählen [BBM04a]. Der PCKMeans Algorithmus wird im Rahmen dieser Arbeit exemplarisch in adaptierter Form in den entwickelten Prototypen (Kapitel 4) integriert. Aufbauend auf PCKMeans existiert mit MPCKMeans [BBM04b] eine Erweiterung, welche zusätzlich zu den generierten Constraints eine Abstands-Matrix auf Basis der Antworten des Domänenexperten erlernt [BBM04b]. Diese Abstands-Matrix wird daraufhin während der Aktualisierung der Cluster-Zentren in der Berechnung des Abstandes zwischen Instanzen verwendet [BBM04b].

Die Ansätze, welche den Domänenexperten durch Empfehlungen unterstützen, beobachten die Aktionen des Domänenexperten zur Erzeugung von Handlungsempfehlungen und Hinweisen [CD18; DFB11]. Ein solcher Ansatz ist iCluster [DFB11], welcher, basierend auf dem Verhalten des Domänenexperten (den bereits getätigten Zuweisungen), aktiv mithilfe von Machine-Learning, Empfehlungen für die weitere Zuweisung von Dokumenten zu Clustern generiert. Die Integration von System-initiierten Empfehlungen wird außerdem von Clustrophile 2 verwendet, um den Domänenexperten auf mögliche Probleme des Clustering-Resultates hinzuweisen und Änderungen vorzuschlagen [CD18].

#### Einbindung vorhandener Ressourcen

Das Ziel vieler Ansätze ist die Einbindung von bereits vor der Anwendung des Clustering-Verfahrens vorhandener Ressourcen, um das erzeugte Clustering-Resultat zu verbessern [BBM02; DR05; MD18; WCRS01; YWL+21]. Die Relevanz dieser Ansätze für diese Arbeit leitet sich aus dem Bedürfnis des Domänenexperten ab (Abschnitt 1.2.1, A3), eventuell vorhandene Ressourcen in das

interaktive Clustering-Verfahren einfließen zu lassen. Der Gebrauch von nicht interaktiven Ansätzen kann diese Anforderung erfüllen und in Kombination mit interaktiven Verfahren als Ausgangspunkt für die weitere (interaktive) Verfeinerung des Clustering-Resultates verwendet werden.

Den populärsten Ansatz zur Einbindung vorhandener Ressourcen stellen Constraint-basierte (Abschnitt 2.5.1) Clustering-Verfahren, welche Informationen über das Verhältnis zwischen Instanzen in das Clustering-Verfahren einbinden [DR05; MD18; WCRS01]. Diese Ansätze erweitern traditionelle Clustering-Verfahren wie beispielsweise k-Means [WCRS01] oder Linkage-based Clustering [MD18] mit der Fähigkeit, bereits vorhandene Ressourcen in Form von Constraints während des Clustering-Prozesses zu berücksichtigen. Der Nachteil dieser Constraint-basierten Ansätze ist die direkte Abhängigkeit von dem zugrundeliegenden Clustering-Verfahren (der Ansatz ist an den verwendeten Clustering-Algorithmus gebunden), weshalb Liu et al. [LJJ07] mit BoostCluster eine vom Algorithmus unabhängige Lösung für Constraint-basiertes Clustering vorschlagen. Bei BoostCluster wird die Repräsentation der Daten selbst auf Basis von gegebenen Constraints angepasst [LJJ07]. Diese angepasste Repräsentation ist mit beliebigen Clustering-Verfahren (beispielsweise dem klassischen k-Means Algorithmus) einsetzbar [LJJ07]. Die verschiedenen Constraint-basierten Ansätze unterscheiden sich durch die von ihnen verwendeten Constraints. Existierende Ansätze verwenden für Constraints unter anderem die Zuweisung einzelner Instanzen zu festen Clustern [SI09], Instanz Tripel für hierarchische Constraints [VD16; ZL11] sowie die Restriktion von Parametern (z.B. die maximale Anzahl an Clustern) [BHR+20]. Die in der wissenschaftlichen Literatur am häufigsten verwendeten Constraints bilden allerdings paarweise Must-/Cannot-Link Constraints (Abschnitt 2.5.1) [AA20; BHR+20; LJJ07; MD18; WCRS01].

Zusätzlich zu den Constraint-basierten Ansätzen zur Einbindung bereits vorhandener Ressourcen können diese in Form von bereits (teilweise) gelabelten Daten [BBM02] und Ontologien [YWL+21] (Datenbanken mit Informationen über verschiedene Konzepte und ihre Beziehung zueinander) verwendet werden. Der Einsatz dieser Daten ermöglicht es dem Domänenexperten eine Teilmenge an bereits gelabelten Daten zur Initialisierung des Clustering-Algorithmus zu verwenden, um ein besseres Clustering-Resultat zu erzielen [BBM02]. Durch die Einbindung von Ontologien (über die Beziehung von Konzepten) verfeinern Yang et al. [YWL+21] mit ReVision die Erstellung von Cluster-Hierarchien aus den Daten. Das ReVision System kombiniert außerdem die Verwendung von vorhandenen Ressourcen (Ontologien) mit der Interaktion des Domänenexperten durch die interaktive Verfeinerung der erstellten Cluster-Hierarchie [YWL+21].

### 2.6.2 Repräsentation und Kommunikation des Clustering-Resultates

Damit der Domänenexperte interaktiv in die Verfeinerung von Clustering-Resultaten eingebunden werden kann, muss dieser zuerst das Resultat verstehen (Abschnitt 1.2.1, A5), um informierte Entscheidungen über die Anwendung von Interaktionen zu treffen. Dieser Abschnitt stellt relevante Arbeiten zur Repräsentation und Kommunikation der erstellten Clustering-Resultate vor, welche zur Unterstützung des Domänenexperten bei der informierten Anwendung von Interaktionen verwendet werden können.

Den klassischen Ansatz zur Darstellung von Clustering-Resultaten stellen traditionelle Visualisierungsmethoden wie Scatter-plots, Heatmaps und Parallele Koordinaten dar [BPB+16; CD18; HBS+16; SS02; TPRH11]. Der Einsatz dieser Visualisierungen wird von einigen Arbeiten durch

die Integration von Interaktionsmöglichkeiten in die Visualisierung oder der visuellen Kodierung zusätzlicher Informationen in die Darstellung erweitert [BPB+16; CD18; HBS+16; SS02; TPRH11]. Die in der Literatur identifizierten Interaktionsmöglichkeiten umfassen die Selektion der darzustellenden Cluster, die Auswahl der verwendeten Projektionsmethode, die detailliertere Ansicht einzelner Bereiche des Resultates und den direkten Vergleich mehrerer (unterschiedlicher) Clustering-Resultate [BHR+20; BPB+16; SS02]. Die visuelle Kodierung von Informationen wird von vielen Ansätzen zur Integration von zusätzlichen Eigenschaften des Clustering-Resultates in die Darstellung verwendet [CGSQ11; HBS+16; YWL+21]. So wird beispielsweise von Cao et al. [CGSQ11] in ihrem Icon-basierten Visualisierungs-Ansatz die visuelle Kodierung verwendet. Diese nutzen die Größe, Form, Position und Farbe der Visualisierung, um die Anzahl an Instanzen und Information über die Qualität sowie die Ähnlichkeit der Cluster darzustellen [CGSQ11].

Außerdem existieren Ansätze, welche die visuelle Kodierung verwenden, um Qualitätsindikatoren (beispielsweise den Dunn-Index oder Silhouetten) direkt in die Visualisierung des Clustering-Resultates integrieren [HBS+16]. Die Anpassung von Visualisierungen, um die Klarheit der Darstellung für große Datenmengen zu erhalten, stellt eine weitere Verbesserung klassischer Visualisierungen dar [JLJC05]. Für die Bewältigung großer Datenmengen verwenden Johansson et al. [JLJC05] beispielsweise Transferfunktionen in Parallelen Koordinatendiagrammen, um optische Unordnung zu reduzieren, wichtige Charakteristiken des Clustering-Resultates hervorzuheben.

Neben der klassischen visuellen Darstellung des Clustering-Resultates lassen sich in der Literatur außerdem Ansätze identifizieren, welche dem Domänenexperten Clustering-Resultate erklären und ihn auf diese Art bei der informierten Anwendung von Interaktionen unterstützen [BHTM22; BOW18; DFMR20; ESG+21; FGS13; LXY05]. Einen Ansatz zur Erklärung von Clustering-Resultaten stellt die Verwendung von Entscheidungsbäumen dar, welche die Entstehung von Clustern beschreibt [BOW18; DFMR20; FGS13; LXY05]. Der Domänenexperte kann die Entscheidungsbäume nutzen, um die Zuweisung einzelner Instanzen zu Clustern (anhand der Verzweigungen des Baumes) nachzuvollziehen [BOW18; DFMR20; FGS13; LXY05]. Die Generierung der Entscheidungsbäume wird von den Ansätzen entweder in das eigentlichen Clustering-Verfahren integriert [FGS13; LXY05] oder erfolgt nachträglich durch das Erlernen des Entscheidungsbaumes auf Basis der berechneten Cluster [CD18; DFMR20]. Der Ansatz von Liu et al. [LXY05] ermöglicht dem Domänenexperten außerdem die Interaktion mit dem erzeugten Entscheidungsbaum durch Pruning (Herausschneiden nicht relevanter Knoten) des erstellten Baumes.

Die Verwendung von Entscheidungsbäumen eignet sich für die Erklärung der Zuweisung der einzelnen Instanzen zu den Clustern [BOW18; DFMR20; FGS13; LXY05]. Allerdings erlauben diese keinen detaillierten Einblick in die Bedeutung der Cluster [BHTM22]. Um die Bedeutung zu erklären, verwenden einige Ansätze [BHTM22; ESG+21] die Identifikation wichtiger Features. Diese Ansätze beschreiben den Cluster-Inhalt durch die Identifikation wichtiger Features, welche innerhalb der einzelnen Cluster oder im Rahmen des gesamten Clustering-Prozesses eine tragende Rolle spielen [BHTM22; ESG+21]. Zusätzlich zu einem Ranking der Features nach ihrer Bedeutung stellen Behringer et al.[BHTM22] außerdem Vorgehen zur Festlegung der Anzahl bedeutsamer Features zur Beschreibung von Clustern durch die Definition von Schwellenwerten oder die Verwendung der Elbow-Methode vor.

### 2.6.3 Zusammenfassung

Die in der Literatur vorhandenen Arbeiten zur Unterstützung von Domänenexperten bei der Durchführung von Clusteranalysen sind zahlreich. Die von existierenden Ansätzen gebotene Unterstützung erfüllt allerdings nicht die identifizierten Anforderungen (Abschnitt 1.2.1) an die effektive Einbindung von Domänenexperten zur interaktiven Verfeinerung von Clustering-Resultaten. Für die Nichteignung der vorhandenen Ansätze lassen sich mehrere Gründe identifizieren.

Ein Grund ist die häufige Fokussierung einzelner Interaktionskategorien mit dem Ziel, vorhandene Interaktionen zu verbessern oder eine neue Form der Interaktion einzuführen, wobei andere Kategorien nicht berücksichtigt werden. Der Domänenexperte muss allerdings während der Durchführung der Clusteranalyse zusätzlich zu Interaktionen zur eigenständigen Verfeinerung des Resultates auch durch System-initiierte Interaktionen aktiv unterstützt werden (Abschnitt 1.2.1 A4). Eine solche Unterstützung kann jedoch von Ansätzen, welche sich auf eine Interaktionskategorie fokussieren, nicht gewährleistet werden. Es wird deshalb ein hybrider Ansatz benötigt, welcher Interaktionen aus verschiedenen Interaktionskategorien (Abschnitt 2.4) kombiniert und so die Anforderungen des Domänenexperten erfüllt.

Dieser Ansatz muss während der Clusteranalyse Nutzer-initiierte und System-initiierte Interaktionen vereinigen. Die Kombination der Interaktionskategorien wird allerdings bereits von Bae et al. [BHR+20] in seiner Zusammenfassung des aktuellen wissenschaftlichen Forschungsbestandes von interaktiven Clustering-Ansätzen als Forschungslücke identifiziert. Den einzigen vorgestellten Ansatz, welcher ein solches hybrides Interaktionsmodell unterstützt, stellen Cavallo und Demiralp [CD18] mit Clustrophile 2 vor, bei welchem der Domänenexperte vom System auf mögliche Verbesserungen und Fehler hingewiesen wird.

Des Weiteren fordert der Domänenexperte die Einbindung bereits vor der Analyse vorhandener Ressourcen (Abschnitt 1.2.1, A3) in das Clustering-Verfahren. Die vorgestellten verwandten Ansätze (Abschnitt 2.6.1) verwenden allerdings oftmals die vorhandenen Ressourcen ausschließlich in nicht interaktiven Clusteranalysen. Die Ausnahmen bilden die Ansätze aus der Active-Clustering Domäne, welche die interaktive Abfrage der Domänenexperten oft mit bereits vorhandenen Constraints initialisieren können [KG20] und Yang et al. [YWL+21], welche mit ReVision vorhandene Ontologien zum Aufbau einer Hierarchie verwenden. Die erstellte Hierarchie kann anschließend vom Domänenexperten während der Clusteranalyse interaktiv angepasst werden.

Die Anwendung der Clustering-Ansätze ist außerdem (unter anderem bei ReVision) häufig auf spezifische Clustering-Probleme (beispielsweise das Clustering von Dokumenten [CLRP13; LKC+12], Bildern [BBR14] oder Android-Apps [CDH+16]) spezialisiert. Diese Ansätze können nicht ohne weiteres auf allgemeine Clustering-Probleme angewendet werden und eignen sich aus diesem Grund ebenfalls nicht für die vom Domänenexperten benötigte generelle Anwendung (Abschnitt 1.2.1, A2) der Clusteranalyse.

Den vorhandenen System-initiierten Active-Clustering Ansätzen [AA20; AL14; BBM04a; WD10; XJC14] fehlt zusätzlich eine klare Definition des Interaktionsablaufs. Der Domänenexperte wird von diesen Ansätzen meist auf die Rolle eines Orakels, welches Abfragen beantwortet, reduziert. Die Interaktion (wie wird der Domänenexperte befragt) und die Darstellung des Clustering-Resultates wird von diesen Ansätzen nur selten näher spezifiziert. Die eigenständige Anwendung (nach Abschnitt 1.2.1, A1) dieser Ansätze durch den Domänenexperten stellt daher (aufgrund der nicht spezifizierten Interaktion) eine Herausforderung dar.

Zusammenfassend konnte keine Arbeit identifiziert werden, welche in der Lage ist, alle Anforderungen des Domänenexperten zu adressieren. Die einzelnen vorgestellten Arbeiten erfüllen jeweils unterschiedliche Teilbereiche der vom Domänenexperten gestellten Anforderungen. Für die Unterstützung des Domänenexperten wird daher ein Konzept benötigt, das die vorgestellten relevanten Arbeiten zu einem hybriden Interaktionsmodell (Kombination der vorgestellten Interaktionskategorien) kombiniert, um die definierten Anforderungen (Abschnitt 1.2.1) des Domänenexperten zu erfüllen. Die Rolle des Domänenexperten und der Interaktionsablauf (an welchen Stellen kann er wie mit dem System interagieren) innerhalb des Konzeptes muss zu diesem Zweck klar definiert werden. Zusätzlich müssen nach Möglichkeit relevante Ansätze für die Repräsentation und Kommunikation von Clustering-Resultaten (Abschnitt 2.6.2) integriert werden, um das Verständnis des Domänenexperten zu fördern (Abschnitt 1.2.1, A5) und eine möglichst effiziente Interaktion mit der Clusteranalyse zu ermöglichen.

# 3 Interaktive Verfeinerung von Clustering-Resultaten durch Domänenexperten

Die Integration von Domänenexperten in die interaktive Verfeinerung von Clustering-Resultaten benötigt ein System, welches den Domänenexperten in den Clustering-Prozess einbindet und adäquate Unterstützung bei der Erstellung des Clustering-Resultates bietet. Die existierenden Lösungen zur Integration von Domänenexperten in den Clustering-Prozess (Abschnitt 2.6) erfüllen allerdings nicht die identifizierten Anforderungen des Domänenexperten (Abschnitt 1.2.1). Der Domänenexperte wird von diesen Lösungen durch die gebotenen Interaktionen (Abschnitt 1.2.1, A3-A4), die fehlende Generalisierung auf seine diversen Anwendungsfälle (Abschnitt 1.2.1, A2) und der fehlenden Integration in den Clustering-Prozesses (Abschnitt 1.2.1, A1+A5) bei der Verfeinerung von Clustering-Resultaten eingeschränkt (Abschnitt 2.6). Es wird daher ein Konzept benötigt, welches die Anforderungen des Domänenexperten (Abschnitt 1.2.1) an die Integration in die interaktive Verfeinerung von Clustering-Resultaten erfüllt.

Im Rahmen dieses Kapitels wird ein Konzept zur Einbindung von Domänenexperten in die interaktive Verfeinerung von Clustering-Resultaten vorgestellt. Das Konzept kombiniert zu diesem Zweck die vorgestellten (Abschnitt 2.6.1) interaktiven Ansätze der verschiedenen Interaktionskategorien (Abschnitt 2.4). Außerdem wird die Verwendung von bereits (vor der Clusteranalyse) vorhandenen Ressourcen zur Verfeinerung des Clustering-Resultats unterstützt. Die entstehenden Clustering-Resultate werden dem Domänenexperten mithilfe der eingeführten unterstützenden Maßnahmen (Abschnitt 2.6.2) veranschaulicht und erklärt. Das Konzept ist so in der Lage, eine holistische Lösung für die Integration von Domänenexperten in den Clustering-Prozess zu stellen.

Das Konzept umfasst ein Prozessmodell für die Integration von Domänenexperten in die interaktive Verfeinerung von Clustering-Resultaten. Zusätzlich werden die einzelnen Interaktionen mit dem Domänenexperten und die Rolle des Domänenexperten innerhalb des Prozessmodells sowie die Umsetzung der einzelnen Komponenten klar definiert.

Im Folgenden wird zuerst das Prozessmodell zur Integration von Domänenexperten in die interaktive Verfeinerung von Clustering-Resultaten vorgestellt, anschließend werden die einzelnen Komponenten des Prozessmodells beschrieben und mithilfe exemplarischer Umsetzungen veranschaulicht.

## 3.1 Prozessmodell

Das Prozessmodel (Abbildung 3.1) verwendet einen hierarchischen (divisiven) Clustering-Ansatz (Abschnitt 2.1) für die Integration des Domänenexperten in den Clustering-Prozess. Der Einsatz des hierarchischen Clustering-Ansatzes erzeugt eine Cluster-Hierarchie, bei der die Instanzen einzelner

Cluster rekursiv in weitere Untergruppen (Cluster) aufgeteilt werden [FAT+14]. Ein Knoten der Cluster-Hierarchie symbolisiert daher ein Cluster, welches (außer für Blattknoten) durch die erneute Anwendung eines Clustering-Verfahrens in mehrere Unterknoten (Cluster) unterteilt wird.

Der Domänenexperte kann in diesem Prozessmodell zuerst die bereits vor der Clusteranalyse vorhandenen Ressourcen (Abbildung 3.1, B) zur Initialisierung des Clustering-Verfahrens verwenden und darauffolgend iterativ verfeinern. Diese Ressourcen bilden mithilfe von Hierarchien und Constraints Zusammenhänge innerhalb der Daten ab und ermöglichen die Verwendung von (aus vorherigen Durchläufen) erstellten Clustering-Konfigurationen. Das daraus resultierende initiale Clustering-Resultat wird anschließend durch die Anwendung von Interaktionen iterativ verfeinert. Die Kernidee dieser Verfeinerung ist, von dem tiefgreifendem Wissen des Domänenexperten Gebrauch zu machen, um die Erstellung der Cluster-Hierarchie zu lenken und iterativ Anpassungen an Hierarchie-Knoten vorzunehmen.

Die Verwendung eines hierarchischen Cluster Ansatzes ermöglicht dem Domänenexperten die Unterteilung der Problemstellung in mehrere Schritte und die Verwendung von vorhandenen Kenntnissen über hierarchische Zusammenhänge innerhalb der Daten. Zusätzlich kann die hierarchische Struktur verwendet werden, um die Berechnung von Verfeinerung-Operationen des Domänenexperten zu beschleunigen, da bei Änderungen nur eine Teilmenge (hierarchisch untergeordnete Knoten) der Daten neu berechnet werden muss.

Der Domänenexperte kann für die Verfeinerung des Clustering-Resultates mit der Hierarchie selbst und den einzelnen Anwendungen von Clustering-Verfahren innerhalb der Hierarchie interagieren. Das System unterstützt ihn während dieser Verfeinerung aktiv durch die Verwendung von Verbesserungsvorschlägen und Active-Clustering Abfragen.

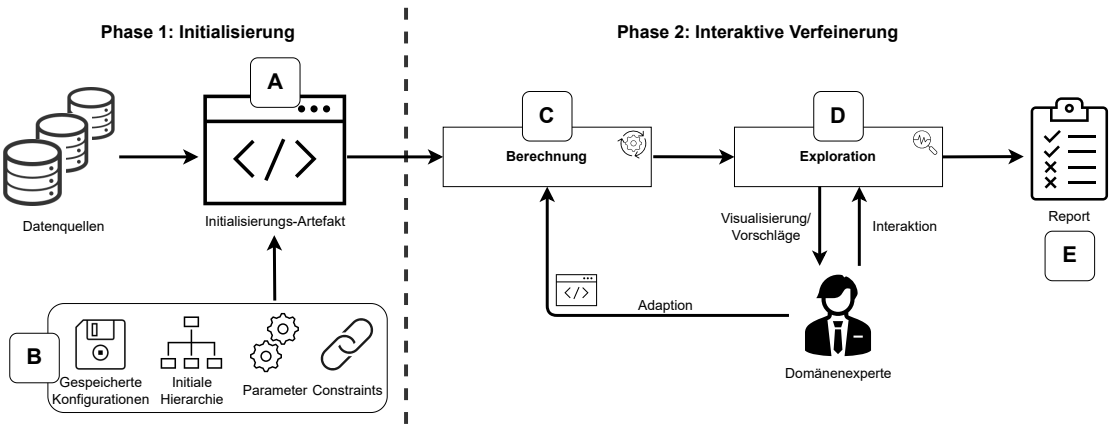



**Abbildung 3.1:** Prozessmodell für die Einbindung von Domänenexperten in Clustering Verfahren zur Verfeinerung von Clustering-Resultaten.

Grundlegend kann der Ablauf des Prozessmodells in zwei Phasen unterteilt werden. In der ersten Phase (Abbildung 3.1, A-B) werden zuerst die bereits vor der Analyse vorhandenen Ressourcen für die Erstellung eines initialen Clustering-Resultates berücksichtigt, um dieses in der darauffolgenden zweiten Phase (Abbildung 3.1, C-E) iterativ zu verfeinern. In den nachfolgenden Abschnitten werden die beiden Phasen genauer definiert.

### 3.1.1 Phase 1: Initialisierung (A, B)

In der ersten Phase des Prozessmodells (Abbildung 3.1) wird ein Initialisierungs-Artefakt (Abbildung 3.1, A) erstellt. Dieses Artefakt bildet die Basis für die in Phase zwei folgende iterative Verfeinerung durch den Domänenexperten. Hierfür müssen zuerst die für die Clusteranalyse relevanten Datenquellen identifiziert und selektiert werden. Die verwendeten Datenquellen können entweder aus dem Bestand des Domänenexperten selbst stammen oder in Anlehnung an den Datenanalyseprozess von Behringer et al. [BHM18] durch IT-Experten bereitgestellt werden.

Das Ziel ist es, bereits vorab vorhandene Informationen bei der folgenden Berechnung zu berücksichtigen, um das Clustering-Resultat zu verfeinern. Die Notwendigkeit für diese Einbindung bereits vorhandener Informationen geht aus den Anforderungen des Domänenexperten (Abschnitt 1.2.1, A3) hervor. Zusätzlich zur Identifikation relevanter Datenquellen werden für die Erstellung des Initialisierungs-Artefaktes vorhandene Ressourcen und Absichten des Domänenexperten (Abbildung 3.1, B) verwendet. Die Berücksichtigung dieser Informationen kann zur Verbesserung des initialen Clustering-Resultates führen und so eine Reduktion der in Phase 2 benötigten Iterationen bewirken. Aus dieser Reduktion folgt direkt die Verminderung der in Phase zwei benötigten Rechenleistung und des Aufwandes des Domänenexperten. Das Prozessmodell berücksichtigt vier Arten von Informationen (Abbildung 3.1, B), die zur Erstellung des Initialisierungs-Artefaktes beitragen können. Diese werden im Folgenden beschrieben und anhand der identifizierten Anforderungen des Domänenexperten (Abschnitt 1.2.1) rationalisiert sowie mithilfe eines exemplarischen Anwendungsfalles veranschaulicht.

**Gespeicherte Konfigurationen:** Das Ziel der Verwendung von gespeicherten Konfigurationen ist die Unterstützung des Domänenexperte bei der Analyse von sich wiederholenden oder ähnlichen Aufgabestellungen (siehe Abschnitt 1.2.1, Anforderung A3). Die gespeicherten Konfigurationen ermöglichen es dem Domänenexperten hierfür ein Clustering-Modell (eine Konfiguration) aus einer früheren Bearbeitung derselben oder einer ähnlichen Aufgabestellung wiederzuverwenden. Der Domänenexperte muss durch diese Wiederverwendung keine oder nur wenige zusätzliche Interaktionen während der erneuten Analyse durchführen.
*Szenario 1 (wiederkehrende Aufgaben):* Der Marketingexperte möchte im Rahmen einer Marketingkampagne den „wertvollsten“ Bestandskunden ein spezielles Angebot unterbreiten. Er entschließt sich dazu, diese Kunden mithilfe einer Clusteranalyse zu identifizieren und erstellt hierfür ein entsprechendes Clustering-Modell. Aufgrund des Erfolges der Kampagne entschließt er sich einen Monat später eine neue gleichartige Kampagne zu starten. In den vergangenen Monaten sind allerdings neue Kunden hinzugekommen. Der Marketingexperte kann in der Initialisierungsphase seine (aufgrund der vorangegangenen Marketingkampagne) existierende Konfiguration (Clustering-Modell) als Ausgangslage für die Clusteranalyse der neuen Marketingkampagne verwenden. Bei Bedarf kann er einzelne Elemente des Clustering-Modells weiter verfeinern.

**Initiale Hierarchie:** Das Ziel der Angabe von initialen Hierarchien ist die (initiale) Verwendung hierarchischer Zusammenhänge, um die anschließend benötigten Verfeinerungs-Iterationen zu reduzieren. Der Domänenexperte hat oftmals bereits eine feste Vorstellung von hierarchischen Abhängigkeiten innerhalb der aktuellen Problemstellung. In einem solchen Fall ist eine initiale grobe Modellierung der Hierarchie vorteilhaft, um die benötigten Iterationen der zweiten Phase zu reduzieren. Die Modellierung dieser hierarchischen Abhängigkeiten kann außerdem als Fundament für die weitere Integration von Vorwissen in Form von Parametern

und Constraints verwendet werden.
*Szenario 2:* Der Marketingexperte weiß, dass er Kunden, die verstärkt Online (z.B. Internetplattform) beziehungsweise Offline (z.B. lokal im Markt) einkaufen, mit verschiedenen Marketingmaßnahmen ansprechen muss. Des Weiteren will er eine Segmentierung der Kunden in wichtige und weniger wichtige Kunden vornehmen. Der Marketingexperte kann diesen Gedankengang intuitiv als Hierarchie mit vier Blattknoten modellieren, bei welcher die Daten zweimal geclustert (aufgesplittet) werden müssen. Er gruppiert die Daten in der ersten Hierarchieebene zuerst in Kunden, welche basierend auf ihrem Umsatz überwiegend Online/Offline einkaufen und unterteilt die beiden Kundengruppen darauffolgend in der zweiten Hierarchieebene basierend auf den generierten Einnahmen in mehr und weniger wertvolle Kunden.

**Constraints:** Die initiale Hierarchie kann durch die Definition von Constraints verfeinert werden. Das Ziel ist die möglichst weite Verfeinerung des initialen Clustering-Resultates durch die vom Domänenexperten geforderte Einbindung vorhandener Ressourcen (Abschnitt 1.2.1, A3) sowie der Einbindung von Domänenwissen (Abschnitt 1.2.1, A4). Der Domänenexperte kann hierzu (ähnlich zum nicht hierarchischen Ansatz von Basu et al. [BBM02]) Instanzen einem existierendem Hierarchie-Knoten (Cluster) zuweisen. Eine solche Zuweisung kann vom Domänenexperten genutzt werden, um das Clustering-Resultat des Knotens anhand der (ähnlichen) zur Verfügung gestellten Instanzen zu verbessern und um Knoten der initialen Hierarchie von Anfang an eine festgelegte Rolle (beispielsweise wichtige Kunden) zuzuteilen [BBM02]. Eine weitere Form zuweisbarer Constraints stellen vorhandene paarweise (Must-/Cannot-Link) Constraints (Abschnitt 2.5.1) und die Beschränkungen der Parameter dar. Der Domänenexperte kann hierfür beispielsweise die Wertebereiche, die Clustern zugewiesen werden können, oder die maximale Anzahl an Instanzen beschränken.

*Szenario 3.1 (Zuweisung von Instanzen):* Der Marketingexperte besitzt einen separaten (selbst gepflegten) Datensatz mit besonders wertvollen Kunden. Bei der Segmentierung aller Kundendaten möchte er die besonders wertvollen Kunden einem Cluster zuweisen. Er kann zu diesem Zweck seine bekannte Teilmenge an besonders wertvollen Kunden einem gewünschten Cluster zuweisen, um weitere Kunden mit ähnlichen Eigenschaften demselben Cluster zuzuordnen.
*Szenario 3.2 (Maximale Anzahl an Instanzen):* Der Marketingexperte weiß, dass er aufgrund von Budget-Restriktionen im Rahmen seiner Marketingkampagne nur 10.000 der Offline Kunden einen Gutschein zukommen lassen kann. Er kann hierfür die Anzahl der maximalen Instanzen des Clusters der Offline Kunden aus Szenario 2.1 auf 10.000 beschränken.

**Parameter:** Es besteht die Anforderung des Domänenexperten, sein Domänenwissen bei der Durchführung der Clusteranalyse einzubringen (siehe Abschnitt 1.2.1, Anforderung A4). Eine Möglichkeit ist die auf seinem Domänenwissen basierende Anpassung von Parametern. Er kann hierzu die Gewichtung von Attributen des Clustering-Verfahrens sowie inhärente Parameter des Clustering-Algorithmus anpassen, um diese Zusammenhänge im Clustering-Verfahren abzubilden.
*Szenario 4:* Der Marketingexperte möchte die Kunden weiterhin in wichtige und weniger wichtige Kunden segmentieren. Dem Marketingexperten ist bewusst, dass für diese Segmentierung einige Attribute (z.B. Umsatz, Anzahl der Einkäufe) von größerer Bedeutung sind als andere (z.B. Geschlecht). Er kann daher bereits in der initialen Phase die Cluster-Anzahl (zwei) festlegen und eine erste Gewichtung dieser Attribute vornehmen.

Nachdem der Domänenexperte die Selektion der Datenquellen und die Bereitstellung von Informationen abgeschlossen hat, wird vom System das Initialisierungs-Artefakt (A) erstellt. Dieses dient als Berechnungsgrundlage für den in Phase zwei folgenden iterativen Verfeinerungsprozess.

### 3.1.2 Phase 2: Interaktive Verfeinerung

Die zweite Phase befasst sich mit der iterativen Verfeinerung des Clustering-Resultates. Der Domänenexperte kann während dieser Phase die Erstellung der Cluster-Hierarchie lenken und vorhandene Einstellungen (aus früheren Iterationen oder aus Phase eins) anpassen oder entfernen. Das Ziel dieser Phase besteht aus der iterativen Verfeinerung des Clustering-Resultates, bis ein für den Domänenexperten zufriedenstellendes Clustering-Resultat erreicht ist. Die zweite Phase wird ergänzend zur ersten Phase benötigt, um die Anforderungen des Domänenexperten nach einer systemgestützten Anpassung der erstellten Clustering-Resultate zu erfüllen (Abschnitt 1.2.1, A4). Dem Domänenexperte werden für diese Verfeinerung Clustering-Resultate verständlich dargestellt und mithilfe unterstützender Maßnahmen erklärt (Abschnitt 1.2.1, A5).

Die zweite Phase beginnt mit der Berechnung eines ersten Clustering-Resultates (Abbildung 3.1, C) aus dem in Phase eins (Abschnitt 3.1.1) erstellten Initialisierungs-Artefaktes (Abbildung 3.1, A). Der Domänenexperte erhält daraufhin die Option, das erzeugte Resultat explorativ zu erkunden (Abbildung 3.1, D) und interaktiv Anpassungen vorzunehmen, um das Resultat zu verfeinern. Die explorative Erkundung und Anpassung ermöglichen es dem Domänenexperten, sich eine Übersicht über den aktuellen Stand des Clustering-Resultates zu verschaffen und, basierend auf dieser, adäquate Interaktionen zu tätigen. Das Resultat des interaktiven Verfeinerungsprozesses ist ein Report (Abbildung 3.1, E), welcher das Clustering-Resultat selbst und die vom Domänenexperten erstellte Konfiguration (das verwendete Clustering-Modell) enthält.

Im Folgenden wird die Exploration (Abbildung 3.1, D) des Clustering-Resultates durch den Domänenexperten näher spezifiziert.

#### Exploration

Ziel des Exploration-Abschnittes ist die Kommunikation des Clustering-Resultates mit dem Domänenexperten und die abschließende Identifizierung einer geeigneten Interaktion (z.B. neuen Gewichtungen, Anwendung von Systemvorschlägen) zur Verfeinerung des Resultates. Das aus der Initialisierungsphase oder der vorherigen Iteration der interaktiven Verfeinerung eingehende Clustering-Resultat (Abbildung 3.2, A) wird hierfür zuerst mit dem Domänenexperten durch die grafische Erkundung des Resultates mit dem Domänenexperten kommuniziert. Die grafische Erkundung beginnt mit der Visualisierung der Cluster-Hierarchie (Abbildung 3.2, B). Die hierarchische Visualisierung gewährt dem Domänenexperten Einsicht in die einzelnen Elemente der Cluster-Hierarchie, um das Verständnis des Domänenexperten über das derzeitige Clustering-Resultat zu fordern.

Auf Basis der gewonnenen Einsicht kann der Domänenexperte für ihn interessante Cluster identifizieren. Dabei wird der Domänenexperte zusätzlich von einer metrischen Umschreibung (D) (beispielsweise die Verteilung der Instanzen zwischen Clustern oder Qualitätsindikatoren) des Resultates unterstützt. Hat der Domänenexperte ein für ihn interessantes Cluster identifiziert, kann er dieses selektieren (Abbildung 3.2, C), um es genauer zu betrachten. Das von ihm selektierte

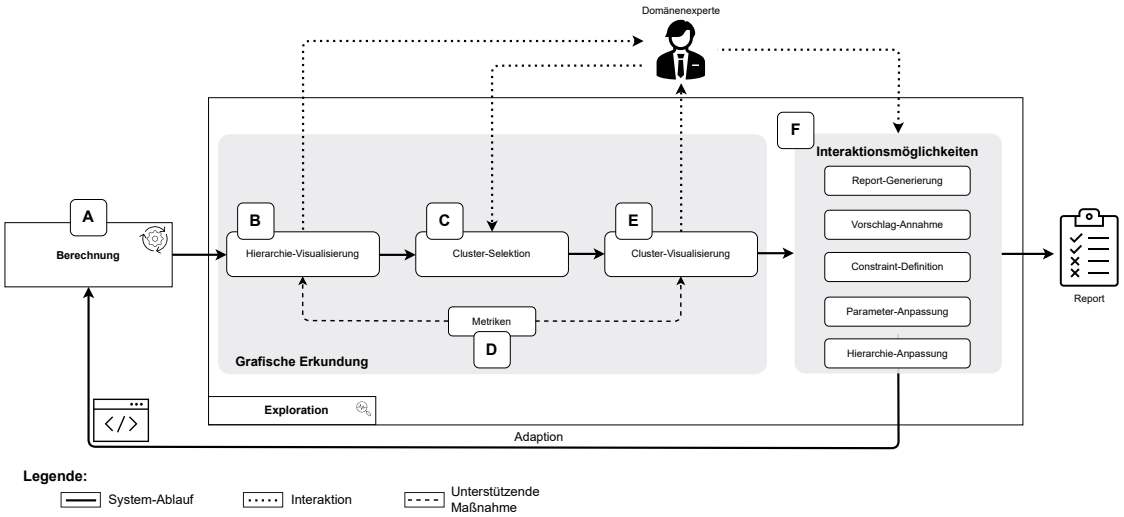



**Abbildung 3.2:** Detaillierter Ablauf des Exploration-Abschnittes aus dem übergeordneten Prozessmodell (Abbildung 3.1) zur Einbindung von Domänenexperten bei der Verfeinerung von Clustering-Resultaten.

Cluster wird daraufhin vom System visualisiert (Abbildung 3.2, E), um dem Domänenexperten einen tieferen Einblick in die für ihn relevanten Knoten der Hierarchie bereitzustellen. Die Cluster-Visualisierung zielt daher auf die Darstellung der Daten innerhalb des selektierten Clusters und die weitere Aufteilung der Daten des selektierten Clusters ab. Der Domänenexperte wird im Rahmen der Visualisierung des Clusters (Abbildung 3.2, E) erneut von einer metrischen Umschreibung (D) des Clusters (Qualitätsmetriken) und der Struktur der Daten innerhalb des Clusters unterstützt. Die Identifikation und Selektion interessanter Cluster (Abbildung 3.2, C) kann vom Domänenexperten innerhalb des explorativen Abschnittes des Prozessmodells (iterativ) beliebig oft wiederholt werden. Der Domänenexperte erhält bei wiederholten Iterationen ein Feedback zu den Anpassungen des Clustering-Resultates (Feedback-Loop) und wird so stärker in die Verfeinerung integriert. Außerdem wird der Clustering-Prozess von einer reinen Black-Box (Domänenexperte weiß nicht, was während dem Clustering-Verfahren passiert) durch die Beteiligung des Domänenexperten an den einzelnen Verfeinerungsschritten zu einem deutlich transparenterem Prozess transformiert.

Auf Basis des vom Domänenexperten selektierten Clusters (Abbildung 3.2, E) werden dem Domänenexperten verschiedene Interaktionsmöglichkeiten (F) angeboten und vom System vorgeschlagen. Der Domänenexperte kann diese Interaktionsmöglichkeiten nutzen, um das Clustering-Resultat zu verfeinern und den Exploration-Abschnitt erneut (mit aktualisiertem Clustering-Resultat) zu durchlaufen. Die erneuten Durchläufe können optional das zuletzt selektierte Cluster als initiale Cluster-Selektion übernehmen. Dieser iterative Verfeinerungsprozess kann vom Domänenexperten jederzeit durch die Erstellung eines (finalen) Berichtes des aktuellen Clustering-Resultates beendet werden.

## 3.2 Grafische Erkundung

Dieser Abschnitt befasst sich mit der Durchführung der grafischen Erkundung im Rahmen des Exploration-Abschnittes (Abbildung 3.2) zur Einbindung von Domänenexperten in die interaktive Verfeinerung von Clustering-Resultaten. Folglich wird er analog zu den Schritten des Exploration-Abschnittes in die Hierarchie-Visualisierung (Abbildung 3.2, B), die Cluster-Visualisierung (E) und zusätzlich in die differenzielle Visualisierung (Visualisierung von Unterschieden zwischen Iterationen des Exploration-Abschnittes) unterteilt. Die innerhalb des Abschnittes vorgestellten Konzepte zur visuellen Umsetzung des im Prozessmodell eingeführten Exploration-Abschnittes (Abschnitt 3.1.2) werden mithilfe von exemplarischen Umsetzungen veranschaulicht und anhand der gebotenen Unterstützung für Domänenexperten motiviert.

Eine mögliche Komposition der ganzheitlichen vorgestellten Umsetzung des Exploration-Abschnittes zeigt die (vereinfachte) Darstellung in Abbildung 3.3, in welcher die hierarchische Visualisierung (B) mit der Visualisierung des selektierten Clusters „Cluster 1“ (B.1) kombiniert wird. Im Folgenden werden die einzelnen integrierten Bestandteile zur Darstellung von Clustering-Resultaten gemäß dem Prozessmodell erklärt und die einzelnen Komponenten näher erläutert.

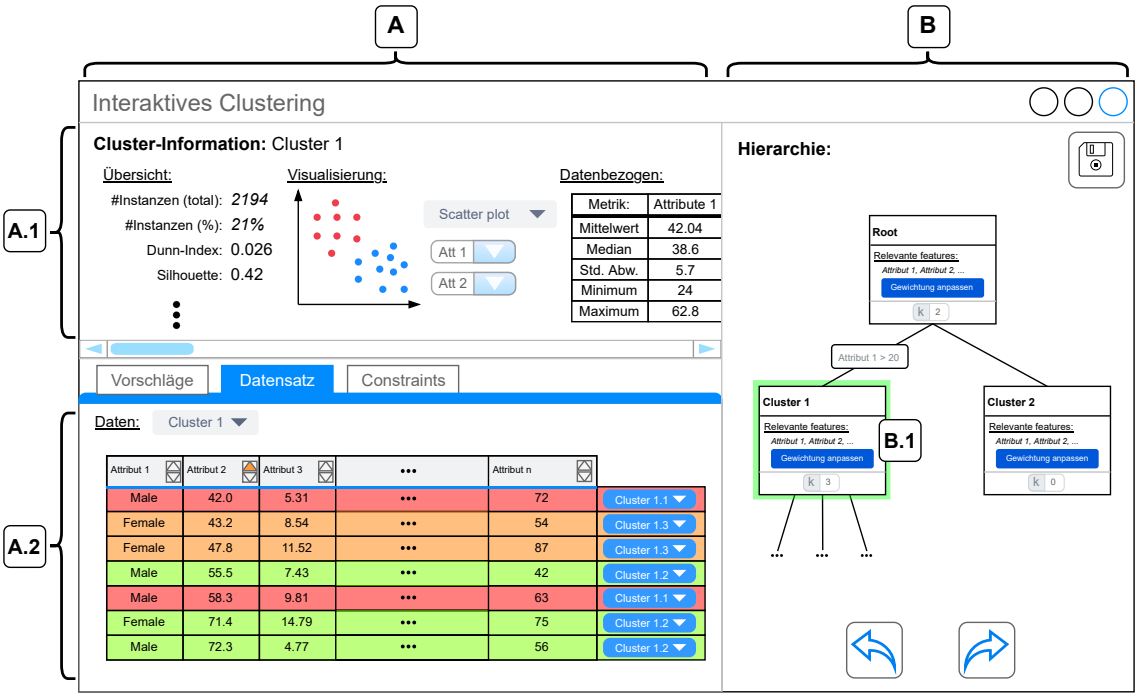



**Abbildung 3.3:** Exemplarische Umsetzung der grafischen Erkundung mit hierarchischer Visualisierung (B) und Cluster-Visualisierung (A) eines selektierten Clusters (B.1). Die Cluster-Visualisierung beinhaltet die Visualisierung clusterbezogener Informationen (A.1) sowie eine Darstellung der Instanzen innerhalb des Clusters (A.2).

### 3.2.1 Hierarchie-Visualisierung

Das Ziel der Hierarchie-Visualisierung (Abbildung 3.3, B) ist es, dem Domänenexperten Einsicht in die dem verwendeten hierarchischen Clustering-Verfahren zugrundeliegende Hierarchie zu gewähren. Das ermöglicht dem Domänenexperten die Anwendung seiner inhärenten Analysefähigkeiten (Abschnitt 1.1) zur Identifikation interessanter Knoten (Cluster) innerhalb der Cluster-Hierarchie. Der Domänenexperte kann dann diese interessanten Knoten durch die gezielte Anwendung von interaktiven Operationen verfeinern. Die klassische Darstellung von hierarchischen Clustering-Verfahrens als Dendrogramm ohne weitere Beschreibung einzelner Knoten (Cluster) ist hierfür nicht ausreichend, da diese keine Einsicht in die einzelnen Zwischenschritte bietet und lediglich die übergeordnete Hierarchie abbildet. Für das benötigte Verständnis des Clustering-Resultates (Abschnitt 1.2.1, A5) muss der Domänenexperte aber diese Zwischenschritte nachvollziehen können.

Einen (auf Dokumente spezialisierten) Ansatz zur Beschreibung dieser Zwischenschritte werden von Yang et al. [YWL+21] verwendet, welche die einzelnen Knoten durch Schlagwörter der enthaltenen Dokumente beschreiben. Das vorgestellte Konzept zur hierarchischen Visualisierung verwendet einen ähnlichen Ansatz, um einzelne Knoten innerhalb der Hierarchie darzustellen. Ein einzelner Knoten repräsentiert gleichzeitig ein Cluster des Elternknotens und die Aufteilung der Daten des Knotens in Unterknoten. Die Beschreibung dieser Knoten muss daher sowohl Informationen über die Zusammensetzung des Cluster-Inhaltes als auch über die weitere Unterteilung der Daten des Knotens enthalten. Die exemplarische Umsetzung in Abbildung 3.4 zeigt ein detailliertes Beispiel einer solchen Beschreibung. Sie enthält die Zusammensetzung der Daten innerhalb des Clusters (Abbildung 3.4, A) sowie eine Übersicht über das Clustering-Verfahren (B), welches zur weiteren Aufteilung der Daten verwendet wird. Es werden drei Eigenschaften zur Beschreibung der Zusammensetzung der Daten näher beleuchtet:

**Mengenmäßige Verteilung der Daten:** Es geht um die relative Verteilung der Datenmenge zwischen verschiedenen Clustern (enkodiert in die Breite der Verbindungslinie, C) und der Gesamtzahl an Instanzen innerhalb eines Clusters (E). Der Domänenexperte kann diese Informationen nutzen, um im Vergleich mit seinen Erwartungen auffällige Cluster zu identifizieren. Die relative Verteilung der Datenmenge zwischen verschiedenen Clustern kann ähnlich zu Yang et al. [YWL+21] mithilfe der Breite der Verbindungslinie (C) visuell enkodiert werden. *Szenario:* Der Marketingexperte weiß, dass nur ca. 10 - 20% der Kunden in der Datenbank tatsächlich als wertvoll bezeichnet werden können. Durch die Betrachtung der hierarchischen Visualisierung stellt er jedoch fest, dass die aktuelle Verteilung der Daten zwischen den beiden Unterknoten ungefähr 50:50 beträgt. Er kann daraus schlussfolgern, dass er das Clustering-Verfahren (durch Interaktionen) weiter verfeinern muss.

**Auf dem Cluster definierte Restriktionen:** Restriktionen, die die Anzahl von Instanzen oder den Wertebereich einschränken, können die Zusammensetzung eines Clusters maßgeblich beeinflussen. Die Nichtbeachtung dieser Restriktionen bei der hierarchischen Visualisierung kann daher die Interpretation des Resultates erschweren. Diese müssen deshalb in die Visualisierung integriert werden. Die exemplarische Umsetzung bildet Restriktionen, die den Aufteilung der Daten und damit die Zusammensetzung eines Clusters beeinflussen, grafisch ab (Abbildung 3.4, D).

**Identifikation relevanter Features (Cluster):** Der Inhalt von Clustern kann Domänenexperten anhand von identifizierten, für das betrachtete Cluster besonders wichtigen Attributen, verständlich erklärt werden [BHTM22]. Eine solche Beschreibung des Cluster-Inhaltes wird in Abbildung 3.4 dargestellt. Zugrunde liegt der Ansatz von Behringer et al. [BHTM22], die Cluster anhand besonders wichtiger Attribute beschreiben. Dieser Ansatz verwendet die Streuung von Werten innerhalb eines Clusters zur Identifikation von Attributen (Features), die für ein einzelnes Cluster von besonderer Bedeutung sind [BHTM22]. Diese identifizierten Features können zur Beschreibung des Cluster-Inhaltes (Abbildung 3.4, F) verwendet werden [BHTM22]. Das hat den Vorteil, dass sie die Zusammensetzung des Clusters selbst anhand einer geringen, von Menschen verarbeitbaren Menge an Informationen darstellt und eine Interpretation der Bedeutung des Clusters (was wird durch das Cluster repräsentiert) ermöglicht [BHTM22]. Zur Identifikation von interessanten Clustern kann der Domänenexperte die identifizierten Features, welche zur Umschreibung eines Clusters verwendet werden, mit der von ihm für das Cluster beabsichtigten Bedeutung abgleichen. Er kann auf diese Weise (schnell) Unterschiede zwischen dem aktuellen Clustering-Modell und seiner konzeptionellen Vorstellung identifizieren.

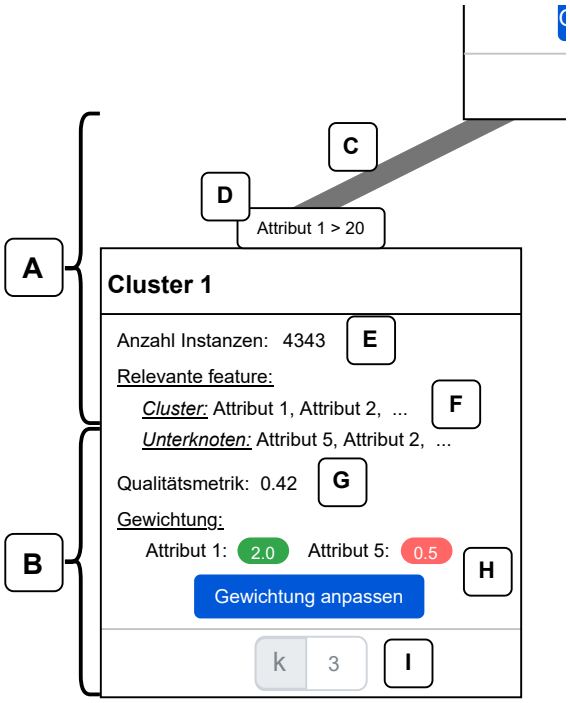



**Abbildung 3.4:** Exemplarische Umsetzung der hierarchischen Visualisierung mit Informationen über die Zusammensetzung des Clusters (A), das darauffolgende Clustering-Verfahren (B) und einzelnen Bestandteilen, die zur Umschreibung genutzt werden (C-I).

Zusätzlich zur Beschreibung des Inhaltes einzelner Knoten wird außerdem die Aufteilung des Cluster-Inhaltes umschrieben. Die Umschreibung der mit einem Knoten verbundenen Aufteilung der Daten (Abbildung 3.4, B) informiert den Domänenexperten über die Konfiguration des angewendeten

Clustering-Verfahrens und erklärt das aus der Anwendung entstandene gesamtheitliche Clustering-Resultat (Abbildung 3.4, F-G). Es werden drei weitere Eigenschaften zur Beschreibung der Aufteilung der Daten näher beleuchtet:

**Identifikation relevanter Features (Unterknoten):** Die Identifikation wichtiger Attribute innerhalb eines Clusters kann zur Inhaltsbeschreibung einzelner Cluster genutzt werden. Behringer et al. [BHTM22] stellen jedoch fest, dass die Wichtigkeit einzelner Attribute vom Anwendungsfall, in welchem diese betrachtet werden, abhängig ist. In der Verwendung von wichtigen Features zur Beschreibung des Cluster-Inhaltes kann die Streuung der Werte innerhalb des Clusters als entscheidendes Merkmal für die Identifikation wichtiger Attribute verwendet werden [BHTM22]. Für die Beschreibung der Aufteilung von Daten eines Knotens (F) ist es jedoch interessanter zu erfahren, welche Features maßgeblich zur Separierung der Cluster (Aufteilung der Daten) beitragen. Dafür stellen Behringer et al. [BHTM22] einen angepassten Ansatz, welcher die Bedeutung einzelner Attribute durch ihren Beitrag zur Separation der Ergebnis-Cluster ermittelt. Diese für die Aufteilung wichtigen Attribute können verwendet werden, um dem Domänenexperten die Clustering-Verfahren einzelner Knoten zu erklären und ihm so die Aufteilung der Daten zu vermitteln. Der Ansatz wird vom exemplarischen Beispiel in F verwendet, um die Separierung der Unterknoten zu erklären. Der Domänenexperte kann diese Information verwenden, um Clustering-Verfahren anzupassen, bei welchen aus seiner Perspektive nicht relevante Attribute einen maßgeblichen Einfluss auf das Clustering-Resultat nehmen.
*Szenario:* Der Marketingexperte möchte seine Kundendaten in marketingtechnisch mehr und weniger wertvolle Kunden unterteilen. Die Beschreibung des Clustering-Resultates anhand wichtiger Attribute informiert ihn, dass der Gesamtumsatz der Kunden das für die Separierung wichtigste Attribut stellt. Der Marketingexperte will mit seiner aktuellen Kampagne jedoch vor allem derzeit besonders aktive Kunden ansprechen. Er weiß, dass sich der Gesamtumsatz hierfür nicht eignet, da es sich bei vielen der Kunden mit hohem Gesamtumsatz um Langzeitkunden, die aktuell nicht unbedingt aktiver als andere Kunden sind, handelt. Idealerweise möchte der Domänenexperte daher den Einfluss des Gesamtumsatzes reduzieren und möglicherweise den Einfluss von relevanten Attributen wie dem Umsatz der letzten sechs Monaten erhöhen, um die Zielgruppe seiner Marketingkampagne besser zu repräsentieren.

**Qualitätsmetrik(en):** Die Qualität des Clustering-Resultates einzuschätzen ist schwierig und kann je nach Anwendungsfall stark vom subjektiven Empfinden des Domänenexperten abhängen [LWG11]. Cluster-interne Qualitätsmetriken (Abschnitt 2.2) versuchen, die Qualität des Clustering-Resultates anhand der Kompaktheit und Separierung von Clustern abzuschätzen und können so einen Indikator für die Qualität des Resultates stellen [HJK17]. Ein solcher Indikator kann dem Domänenexperten helfen, die Qualität des Ergebnisses einfach einzuschätzen. Die exemplarische Umsetzung bindet einen solchen Qualitätsindikator (Abbildung 3.4, G) in die hierarchische Visualisierung ein.
Ein Problem bei der Einbindung dieser Qualitätsmetriken in die hierarchische Visualisierung von Knoten stellt die Selektion einer geeigneten Metrik dar, da die wissenschaftliche Literatur eine Vielzahl von Qualitätsmetriken identifiziert [HJK17]. Dies verhindert die Selektion einer besten Metrik für die generischen Anwendungsfälle (Abschnitt 1.2.1, A2) des Domänenexperten [HJK17]. Die Integration von Qualitätsmetriken in die hierarchische Visualisierung

kann daher wahlweise entweder durch die Selektion einer der (durchschnittlich besseren) Qualitätsmetriken oder durch eine hybride Kombination (Aggregation) mehrerer Metriken zu einem Wert bewerkstelligt werden.

**Clustering-Parameter:** Damit der Domänenexperte die Anwendung des Clustering-Verfahrens verstehen kann, muss er wissen, mit welcher Konfiguration das Clustering-Verfahren eines Knotens ausgeführt wird. Es ist daher essenziell, dass die Übersicht eines Knotens, die zur weiteren Aufteilung der Daten verwendeten Parameter enthält. Die exemplarische Umsetzung der hierarchischen Visualisierung enthält deshalb die Parameter, welche vom Domänenexperten für die Gewichtung des Clustering-Verfahrens genutzt werden (Abbildung 3.4, H) sowie die Parameter, welche zur Ausführung des Clustering-Algorithmus (Abbildung 3.4, I) benötigt werden.

Die Kombination der knotenweisen Beschreibung des Cluster-Inhaltes (Abbildung 3.4, A) mit der Beschreibung der korrespondierende Aufteilung der Daten (B) ermöglicht dem Domänenexperten ein tiefgreifendes Verständnis über den hierarchischen Ablauf des Clustering-Verfahrens. Dieser erhält Einsicht in die Regeln (nach welchen Featuren wird separiert, welche Parameter werden verwendet, ...) anhand welcher die Daten beim Clustering-Verfahren aufgeteilt werden (Abbildung 3.4, B). Er kann mithilfe der Beschreibung des Cluster-Inhaltes der Unterknoten Rückschlüsse über das Ergebnis der Aufteilung ziehen. Der Domänenexperte ist so in er Lage, die Zuweisung der Daten zu Clustern ähnlich zu Ansätzen, welche Entscheidungsbäume für die Erklärung von Clustering-Resultaten verwenden [BOW18; DFMR20; FGS13; LXY05] nachzuvollziehen. Im Gegensatz zu diesen Verfahren wird die Zuteilung der Daten allerdings nicht durch feste Entscheidungsregeln, sondern für die Aufteilung relevanter Eigenschaften (beispielsweise wichtige Features) und der Beschreibung des resultierenden Ergebnisses erklärt. Das Verständnis über die Aufteilung der Daten innerhalb der Cluster-Hierarchie ermöglicht dem Domänenexperten die Identifikation von für ihn interessanten Stellen und erreicht so das übergeordnete Ziel, den Domänenexperten gemäß seiner Anforderung an die Kommunikation des Clustering-Resultates (Abschnitt 1.2.1, A5) zu unterstützen.

### 3.2.2 Cluster-Visualisierung

Die vorangegangene Visualisierung der Hierarchie (Abschnitt 3.2.1) ermöglicht dem Domänenexperten die Identifikation von für ihn interessante Stellen innerhalb der Cluster-Hierarchie zu identifizieren. Damit die Anforderung des Domänenexperten an die Kommunikation des Clustering-Resultates (Abschnitt 1.2.1, A5) erfüllt wird, muss dieser auf Basis der kommunizierten Informationen geeignete Interaktionen anwenden können. Hierzu benötigt er einen tieferen Einblick in interessante Stellen der Cluster-Hierarchie. Die Cluster-Visualisierung (Abbildung 3.2, E) bietet dem Domänenexperten eine detaillierte Übersicht über ein Cluster, welches gemäß dem Prozessmodell in der explorativen Phase als besonders interessant identifiziert und anschließend selektiert (Abbildung 3.2, C) wurde. Ziel der Cluster-Visualisierung ist daher die umfassende Darlegung der clusterbezogenen Informationen, um eine tiefgreifendere Analyse des Cluster-Inhaltes sowie des Clustering-Resultates durch den Domänenexperten zu ermöglichen und somit die Anforderungen des Domänenexperten an die Kommunikation von Clustering-Resultaten (Abschnitt 1.2.1, A5) umfassend zu erfüllen. Eine mögliche Umsetzung der Cluster-Visualisierung des Prozessmodells wird von (Abbildung 3.3, A) gestellt.

Für die Darlegung der clusterbezogenen Informationen nutzt die exemplarische Umsetzung (Abbildung 3.3, A.1) eine Beschreibung der weiteren Aufteilung des Cluster-Inhaltes anhand von klassischen Cluster-Visualisierungstechniken (Abschnitt 2.2), wie beispielsweise Scatterplots sowie die Aufreihung mehrerer Qualitätsmetriken (beispielsweise Silhouetten) und Metriken der deskriptiven Statistik (beispielsweise Median oder Varianz), um den Inhalt des selektierten Clusters näher zu beschreiben. Die Verwendung der Kombination von klassischen Visualisierungstechniken und Qualitätsmetriken für die Repräsentation des Clustering-Resultates folgt dabei dem Beispiel, bereits vorhandener interaktiver Clustering Ansätze [CD18; LKS+15; YWL+21]. Deskriptive statistische Metriken erlauben zusätzlich die Umschreibung der Lage und Streuung der Daten innerhalb von Clustern [Nic07]. Der Domänenexperte kann diese Beschreibungen nutzen, um mithilfe seiner Expertise Abnormalitäten (Teile des Datensatzes, die von dem von ihm erwarteten Inhalt abweichen) zu identifizieren und informierte Entscheidungen für die weitere Aufteilung des Datensatzes in folgenden Cluster-Verfahren zu treffen. Ein Beispiel einer solchen Entscheidung stellt die Anpassung von Wertebereichen eines Knotens auf Basis von Lageeigenschaften der zugehörigen Daten.

Für die exemplarische Umsetzung wird in dieser Sektion die Verwendung von geläufigen deskriptiven Metriken vorgeschlagen. Die Übersicht über den Bereich der deskriptiven Statistik von T. Nick [Nic07] identifiziert hierfür folgende Metriken zur Beschreibung der Lage (Tendenz der Werte) und Streuung (Verteilung der Werte) eines Datensatzes:

1. Lage
    - arithmetisches Mittel
    - Median
    - Modus (häufigster Wert)
    - Quartile
2. Streuung
    - Varianz
    - Standardabweichung
    - Abstand zwischen Minimum und Maximum
    - Interquartilsabstand

Zusätzlich ist eine grafische Beschreibung der Lage und Streuung der Daten in Form eines Box-Plots möglich [Nic07]. Diese enthalten Informationen zum Median, den Quartilen, dem Interquartilsabstand sowie zu Ausreißern der Daten [Nic07].

Die Verwendung von Metriken der deskriptiven Statistik gewinnt im Rahmen des vorgestellten Konzeptes besonders an Bedeutung, da die Daten einzelner Hierarchie-Knoten beschrieben werden. Diese Daten können abhängig vom selektierten Knoten stark voneinander abweichen. Es ist daher wichtig, dem Domänenexperten die Unterschiede der Datenverteilung klarzumachen.

Außer der Umschreibung der Aufteilung der Daten und des Inhaltes (Abbildung 3.3, A.1) enthält die Cluster-Visualisierung des exemplarischen Beispiels (ähnlich zu Clustrophile 2 [CD18], welches ebenfalls eine tabellarische Darstellung der Daten implementiert) zusätzlich eine direkte Darstellung des Cluster-Inhaltes in tabellarischer Form (A.2). Der Domänenexperte kann diese Darstellung

nutzen, den Inhalt des Clusters und die weitere Zuweisung einzelner Instanzen (farblich encodiert) direkt und nicht nur in umschriebener Form (A.1) zu betrachten. Außerdem ist die Erweiterung der tabellarischen Darstellung durch interaktive Tabellen-Operationen wie beispielsweise das Sortieren, Filtern und Suchen möglich, um den betrachteten Datenausschnitt anzupassen.

### 3.2.3 Differenzielle Visualisierung

Der Explorationsschritt des Prozessmodells (Abbildung 3.2) beinhaltet explizit die hierarchische Visualisierung (B) sowie die Cluster-Visualisierung (E) als Teilschritte des Explorationsablaufes. Aufgrund des iterativen Charakters des Explorationsschrittes ergibt sich jedoch zusätzlich eine Differenz zwischen den Clustering-Resultaten, welche in den verschiedenen Durchläufen der Explorationsphase generiert werden.

Ähnlich zu existierenden interaktiven Clustering-Ansätzen, welche den Vergleich von mehreren Clustering-Resultaten untereinander ermöglichen [LKS+15], kann der Domänenexperte durch die Visualisierung der Unterschiede zwischen zwei (oder mehreren) Durchläufen des Explorationsschrittes unterstützt werden. Die Visualisierung eines vorherigen Durchlaufes ermöglicht es dem Domänenexperten, die Richtung, in welche sich das Clustering-Modell durch seine iterativen Anpassungen entwickelt, einzuschätzen und falls notwendig zu adaptieren. Die menschliche Fähigkeit, solche Änderungen zwischen aufeinanderfolgenden Visualisierungen präzise (was genau hat sich verändert) wahrzunehmen, ist allerdings inhärent schlecht [SL97]. Es ist daher notwendig, eine kombinierte Visualisierung zu erstellen, welche die Ergebnisse beider Durchläufe miteinander verbindet und so einen direkten Vergleich der Veränderungen ermöglicht.

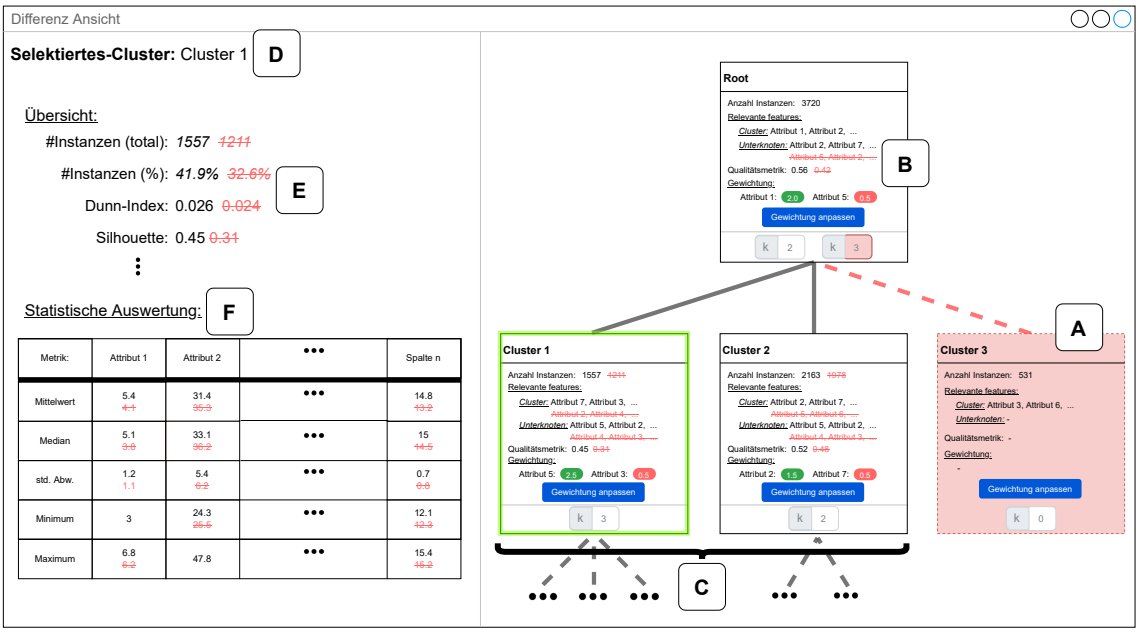



**Abbildung 3.5:** Exemplarische Umsetzung der differenziellen Visualisierung zwischen Iterationen des Explorationsschrittes mit Änderungen (rot), welche sich aus der Entfernung des Knotens „Cluster 3“ (A) in der vorangegangenen Iteration ergeben.

Eine exemplarische Umsetzung einer solchen Visualisierung zeigt Abbildung 3.5, in welcher beispielhaft (in der vorherigen Iteration des Prozessmodells) der Knoten „Cluster 3“ (A) gelöscht wurde. Die Änderungen der Hierarchie werden von der exemplarischen Umsetzung durch den rot hervorgehobenen Knoten (Abbildung 3.5, A) symbolisiert, welcher in der vorherigen Iteration vom Domänenexperten entfernt wurde. Zusätzlich werden die Änderungen der Werte, die zur Beschreibung der hierarchischen Visualisierung verwendet werden (durch die rote Hervorhebung der vorherigen Werte), visualisiert (Abbildung 3.5, B-C). Dies umfasst zum einen eine Änderung der für die Aufteilung der Daten relevanten Attribute, Qualitätsmetriken und Anzahl der Cluster aus dem Elternknoten (Abbildung 3.5, B) sowie zum anderen die Darstellung des Effektes der Änderung in den verbleibenden Kindknoten (C). Der Domänenexperte kann außerdem Unterschiede innerhalb des von ihm selektierten Clusters (Abbildung 3.5, D) näher betrachten. Er kann so auf Änderungen der Metriken (Abbildung 3.5, E) und die geänderte Werteverteilung (Lage und Streuung, F) reagieren.

Identifiziert der Domänenexperte eine Veränderung in der differenziellen Visualisierung, welche von seiner Erwartung abweicht, kann diese angepasst oder rückgängig gemacht werden. Beispielsweise stellt der Marketingexperte (mithilfe der differenziellen Visualisierung) fest, dass sich die Anzahl der Kunden im Cluster „VIP Kunden“ durch seine letzte Anpassung unerwartet stark verändert hat und macht diese daraufhin rückgängig.

## 3.3 Interaktionsmöglichkeiten

Diese Sektion befasst sich mit den Interaktionsmöglichkeiten (Abbildung 3.2, F), welche dem Domänenexperten für die iterative Verfeinerung des Clustering-Resultates zur Verfügung stehen. Die vorgestellten Interaktionen ermöglichen die Verbesserung des Clustering-Resultates sowie die Adressierung des subjektiven Aspektes (was ist für den Domänenexperten „korrekt“) von Clustering-Resultaten. Es wird so direkt die Anforderung des Domänenexperten nach der interaktiven Verfeinerung von Clustering-Resultaten unter Verwendung seines Domänenwissens (Abschnitt 1.2.1, A4) angesprochen. Der Domänenexperte benötigt für die Erfüllung seiner Anforderung an die Durchführung der Verfeinerung (Abschnitt 1.2.1, A4) Interaktionsmöglichkeiten, sowohl um die Erstellung des Clustering-Resultates zu beeinflussen als auch zur direkten Anpassung von Clustering-Resultaten. Außerdem muss er während des Verfeinerungsprozesses durch vom System eingeleiteten Interaktionen für die effiziente Abfrage von Informationen und die Vermeidung von Fehlern unterstützt werden (Abschnitt 1.2.1, A4). Für die umfassende Erfüllung dieser Anforderung müssen daher Interaktionen aus allen drei von Bae et al. [BHR+20] identifizierten Interaktionskategorien (Interaktion mit Parametern, Interaktion mit Ergebnis und System initiierte Interaktionen) kombiniert und ins überliegende Prozessmodell integriert werden. Der Domänenexperte kann durch diese Kombination der Interaktionen sein Domänenwissen bestmöglich in die Verfeinerung des Clustering-Resultates einfließen lassen.

Die Anwendung der Interaktionsmöglichkeiten baut auf den Einsichten (beispielsweise über die Cluster-Hierarchie und den Cluster-Inhalte) des Domänenexperten über das aktuelle Clustering-Resultat aus der vorangegangenen grafischen Erkundung (Abschnitt 3.2) auf. Dem Domänenexperten dienen diese Einsichten als Entscheidungsgrundlage für die Art und Lokalität der anzuwendenden Interaktionsmöglichkeit. Der Domänenexperte kann hierfür bei ReVision [YWL+21], einem Ansatz

zum Clustering von Dokumenten, die Themenbereiche der einzelnen Dokument-Cluster einer Hierarchie bearbeiten. Ähnlich dazu kann der Domänenexperte einzelne Stellen der hierarchischen Visualisierung (Abschnitt 3.2.1), auf welche er eine Interaktion anwenden möchte, selektieren.

Eine beispielhafte Integration der Interaktionsmöglichkeiten in die Hierarchie-Visualisierung (Abschnitt 3.2.1) zeigt die exemplarische Umsetzung in Abbildung 3.6.

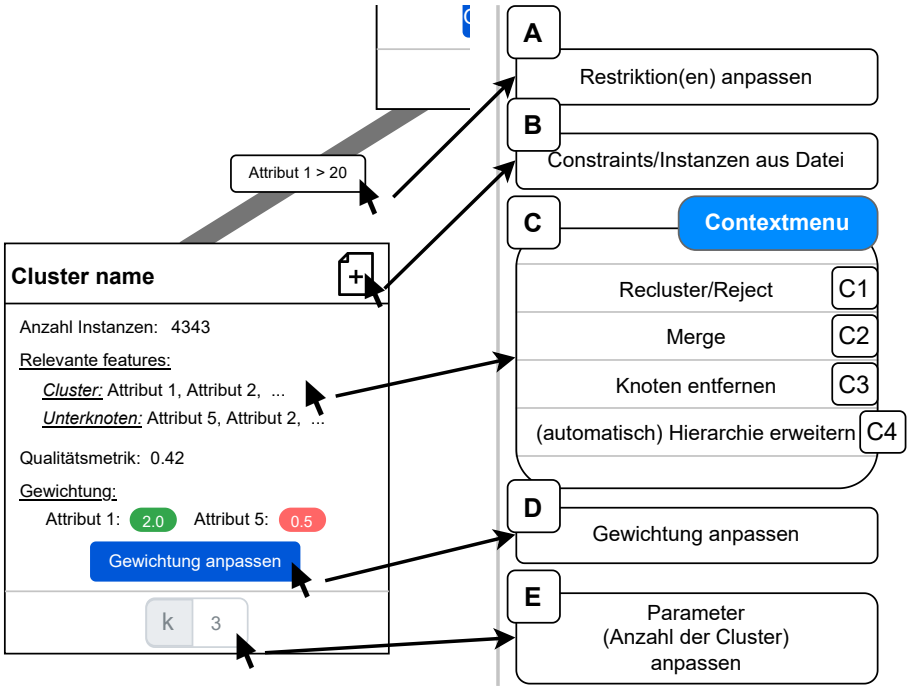



**Abbildung 3.6:** Übersicht über die mögliche Integration der Interaktionen in die exemplarische Umsetzung der hierarchische Visualisierung (Abbildung 3.4) mit den verschiedenen Interaktionsmöglichkeiten A-E und den korrespondierenden Interaktionspunkten des Models (Cursor).

Die in diesem Abschnitt vorgestellten Interaktionsmöglichkeiten sind (nach ihren zugehörigen Interaktionskategorien) in Interaktionsmöglichkeiten zur Interaktion mit den Cluster-Parametern (Abschnitt 3.3.1), Interaktionsmöglichkeiten zur direkten Anpassung des Clustering-Resultates (Abschnitt 3.3.2) und vom System eingeleitete Interaktionsmöglichkeiten (Abschnitt 3.3.3) unterteilt. In den nachfolgenden Abschnitten wird die Integration der einzelnen Interaktionsmöglichkeiten in das Prozessmodell sowie deren Anwendung durch den Domänenexperten näher erörtert, anhand von exemplarischen Umsetzungen visualisiert und mit bereits vorgestellten Elementen des Konzeptes (dem Prozessmodell und der grafischen Erkundung) verknüpft.

### 3.3.1 Interaktion mit Parametern

Die Bestimmung von Parametern des Clustering-Verfahrens sowie die wiederholte Neuberechnung des Clustering-Resultates zur Optimierung der Parameter wird von vielen interaktiven Clustering-Ansätzen unterstützt [BHR+20]. Für die effektive Verfeinerung des Clustering-Resultates durch

den Domänenexperten (Abschnitt 1.2.1, A4) ist die Fähigkeit, die Parameterkonfigurationen der Clustering-Verfahren anpassen zu können, unabdingbar. Aufgrund der hierarchischen Natur des Prozessmodells und der daraus resultierenden knotenweisen Anwendung von Interaktionen ergibt sich im Rahmen dieser Arbeit allerdings ein abweichendes Szenario, bei welchem der Domänenexperte die Festlegung der Parameter (knotenweise) in einem hierarchischen Kontext vornimmt. Die einzelnen Parameter, die bei jedem dieser Knoten angegeben werden müssen (und weiter optimiert werden können), hängen vom zugrundeliegenden Clustering-Algorithmus ab. Die feste Auswahl des Clustering-Algorithmus ist jedoch schwierig, da die Qualität des Resultates verschiedener Algorithmen vom Anwendungsfall abhängig ist. Der Domänenexperte muss daher zwischen verschiedenen Algorithmen zur Lösung des Clustering-Problems auswählen können, um ein bestmögliches Resultat zu erzielen.

Im Folgenden wird die Wahl des Clustering-Algorithmus mit der Festlegung der korrespondierenden Parameter und die Gewichtung von Attributen genau erklärt.

#### Wahl des Clustering-Algorithmus und der Parameter

Die Auswahl eines geeigneten Clustering-Algorithmus kann die Güte des Clustering-Resultates beeinflussen [Jai10; LKS+15]. Die informierte Auswahl des Clustering-Algorithmus, welcher für die Berechnung des Resultates verwendet wird, erfordert jedoch eine Expertise in dem Bereich der Datenanalyse. Diese kann von Domänenexperten im Allgemeinen nicht vorausgesetzt werden. Dieses Selektionsproblem wird von den vorgestellten vorhandenen Arbeiten (Abschnitt 2.6) durch den Vergleich der Clustering-Resultate unterschiedlicher Algorithmen [LKS+15] oder durch die automatisierte Auswahl eines geeigneten Clustering-Algorithmus auf Basis der verwendeten Daten [CD18] adressiert.

Im Rahmen der Umsetzung des Prozessmodells (Abschnitt 3.1) wird eine Kombination beider Methoden vorgeschlagen, bei welcher der Domänenexperte durch einen vorgeschlagenen Clustering-Algorithmus bei der Selektion unterstützt wird, diesen Clustering-Algorithmus allerdings während der iterativen Verfeinerung ändern kann. Der Domänenexperte kann die Auswirkung dieser der Änderung während der iterativen Verfeinerung mithilfe der differenziellen Visualisierung (Abschnitt 3.2.3) vergleichsbasiert nachvollziehen, um die Eignung des Clustering-Algorithmus einzuschätzen. Der Domänenexperte muss die auf die Algorithmen bezogenen Parameter bei einem Wechsel daher eventuell neu festlegen. Aufgrund des (im Vergleich zur Anwendung anderer interaktiver Maßnahmen) relativ großen Aufwandes, der mit der Bestimmung des am „besten“ geeigneten Algorithmus verbunden ist, wird die Auswahl des Clustering-Algorithmus für die gesamte erstellte Cluster-Hierarchie (anstatt für einzelne Knoten) übernommen.

Verbunden mit der Wahl des Clustering Algorithmus ist die Auswahl der vom Clustering-Algorithmus benötigten Parameter. Der Domänenexperte kann diese im Explorationsschritt (Abbildung 3.1, D) des Prozessmodells iterativ anpassen. Hierfür kann der Domänenexperte die verwendeten Parameter knotenweise in der hierarchischen Visualisierung festlegen. Die Anpassung ermöglicht es dem Domänenexperten, die vorhandene Cluster-Hierarchie anzupassen oder durch die weitere Unterteilung von Blattknoten zu erweitern.

Die exemplarische Umsetzung von Interaktionsmöglichkeiten auf Knotenebene (Abbildung 3.6) zeigt in E exemplarisch eine Option für die Anpassung des Parameters „k“ (der Clusteranzahl).

#### Gewichtung von Attributen

Der Domänenexperte kann während der Bearbeitung der Hierarchie möglicherweise für die Gruppierung der Daten besonders relevante Attribute identifizieren, welche basierend auf seiner Expertise einen stärkeren oder weniger starken Einfluss auf die Aufteilung der Daten im Rahmen des Clustering-Verfahrens eines Knotens nehmen sollen. Er muss dieses Domänenwissen gemäß seinen Anforderungen (Abschnitt 1.2.1, A4) in die Clusteranalyse einbringen können.

Der interaktive dokumentbasierte Clustering-Ansatz iVisClustering [LKC+12] erlaubt dem Domänenexperten den Einfluss verschiedener Themen (Topics) auf die Gruppierung der Dokumente durch eine Gewichtung zu regulieren. Ähnlich zu diesem Ansatz wird zur Unterstützung des Domänenexperten die Gewichtung einzelner Attribute eingeführt, um dem Domänenexperten die Möglichkeit zu geben, seine Expertise in der Cluster-Hierarchie abzubilden. Eine mögliche Einbindung der Gewichtung von Attributen in die hierarchische Visualisierung zeigt die exemplarische Umsetzung aus Abbildung 3.6 in D. Im folgenden Beispiel wird die Verwendung von Domänenwissen zur Gewichtung von Attributen veranschaulicht, bevor ein Konzept zur Umsetzung einer Maßnahme zur Gewichtung eingeführt wird.
*Beispiel-Szenario:* Der Marketingexperte möchte eine Marketingkampagne für einen neuen Joghurt durchführen. Die Zielgruppe der Marketingkampagne sollen besonders wichtigen Kunden (basierend auf ihrem Umsatz) verschiedener Regionen bilden. Für die Segmentierung der Kunden weis der Marketingexperte in diesem Szenario aufgrund seiner Expertise, dass der Umsatz der Produktkategorie „Milchprodukte“ sich aufgrund der Nähe zum vermarkteten Produkt besser für die Separierung der Kunden eignet als andere Kategorien. Der Marketingexperte erhöht daher den Einfluss (die Gewichtung) des Umsatzes aus der Kategorie „Milchprodukte“. Das erstellte Clustering-Resultat reflektiert die Expertise des Marketingexperten.

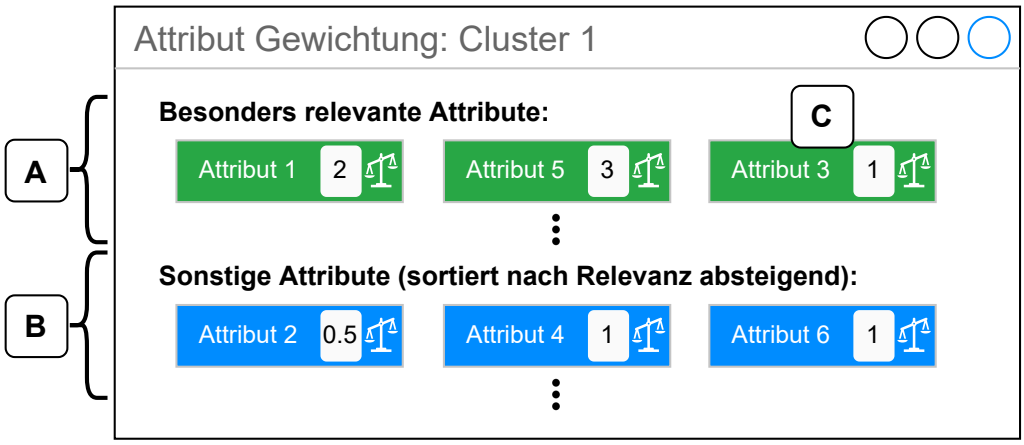



**Abbildung 3.7:** Exemplarischer Dialog für die Gewichtung einzelner Attribute eines Knotens in Kombination mit der Identifikation besonders wichtiger Attribute von Behringer et al. [BHTM22], um die Erklärbarkeit des Dialoges zu verbessern.

Die Wahl der passenden Gewichtungen für Attribute ist für den Domänenexperten nicht trivial. Für die effektive Anwendung von Gewichtungen muss der Domänenexperte die Auswirkung der angewendeten Gewichtung auf das Clustering-Resultat einschätzen können. Diese Auswirkung muss dem Domänenexperten daher vom System kommuniziert (Abschnitt 1.2.1, A5) werden. Hierfür

kann die Gewichtung einzelner Attribute mit der Identifikation besonders wichtiger Attribute von Behringer et al. [BHTM22] kombiniert werden, um dem Domänenexperten die Identifizierung von passenden Gewichten zu vereinfachen. Die Kombination erlaubt es dem Domänenexperten, den Einfluss des Attributes auf das Clustering-Resultat anhand der berechneten Relevanz des Attributes einzuschätzen. Der Domänenexperte kann so die Gewichtung der Attribute anpassen, bis er mit dem Einfluss, welche einzelne Attribute auf das Endergebnis nehmen, zufrieden ist.

Eine exemplarische Umsetzung dieser Kombination zeigt Abbildung 3.7, in welcher der Domänenexperte eine Übersicht über die besonders relevanten Attribute (A) und deren Gewichtung (C) erhält. Zusätzlich kann der Domänenexperte anhand der nach Relevanz sortierten sonstigen Attribute (B) die relative Bedeutung einzelner Attribute einschätzen. In dem Fall, dass die Relevanz eines Attributes nicht seiner Vorstellung entspricht, kann er die Gewichtung anpassen.
*Beispiel-Szenario (Kombination):* Der Marketingexperte stellt fest, dass die Kategorie „Milchprodukte“ nach seiner vorherigen Gewichtung nun das wichtigste Attribut der Kategorie „Sonstige Attribute“ (B) stellt. Der Marketingexperte möchte jedoch, dass der Umsatz der Milchprodukte eine tragende Rolle spielt, also den besonders relevanten Attributen zugeordnet wird. Er erkennt daher direkt, dass er die Gewichtung weiter erhöhen muss.

### 3.3.2 Interaktion mit dem Ergebnis

Der Domänenexperte kann das auf Basis der festgelegten Parameter (Abschnitt 3.3.1) berechnete Clustering-Resultat durch die direkte Interaktion mit Elementen des Resultates weiter verfeinern. Diese Interaktionsart basiert auf der Modellierung gewünschter Änderungen, welche im folgenden vom System interpretiert und umgesetzt werden [BHR+20]. Der Domänenexperte kann diese Modellierung nutzen, um das Clustering-Resultat mithilfe seiner Expertise in die von ihm gewünschte Richtung zu lenken [BHR+20]. Dadurch wird die Anforderung des Domänenexperten nach der direkten Anpassung der Clustering-Resultate mithilfe seines Domänenwissens (Abschnitt 1.2.1, A4) adressiert. Für die Anwendung der Operationen muss der Domänenexperte, ähnlich wie bei der Änderung von Parametern, einen Knoten, auf welchem die Operation angewendet werden soll, auswählen. Die Anwendung einzelner Operationen muss damit erneut in die hierarchische Visualisierung sowie die Cluster-Visualisierung integriert werden.

Die identifizierten Interaktionsmöglichkeiten dieser Kategorie werden basierend auf dem Gegenstand der Interaktion (Constraints, Cluster-Inhalte, Hierarchie) in drei Abschnitte unterteilt. In den folgenden Abschnitten werden die einzelnen Interaktionen mit dem Clustering-Resultat und deren Integration in den explorativen Schritt des Prozessmodells näher erklärt.

#### Constraints

Der Domänenexperte hat basierend auf seinem Domänenwissen eine grobe Vorstellung über das Clustering-Resultat. Diese Vorstellung ist oftmals unpräzise, ermöglicht es dem Domänenexperten allerdings trotzdem Clustering-Resultate durch die Entfernung von für den Domänenexperten unrealistischen Sachverhalten einzuschränken. Der Domänenexperte besitzt zu diesem Zweck beispielsweise Informationen über mögliche Attributwerte oder Menge an Instanzen, welche einem Cluster innewohnen können. Dieses Wissen muss vom Domänenexperten verwendet werden können, um die Anforderung an die Verfeinerung des Clustering-Resultates zu erfüllen (Abschnitt 1.2.1, A4).

Hierfür können Constraints verwendet werden, um mithilfe der Expertise des Domänenexperten Wertebereiche von Clustering-Resultaten zu beschränken, die Anzahl von Instanzen innerhalb von Clustern zu beschränken sowie Kenntnisse über die (im Rahmen des vorgestellten Konzeptes paarweise) Beziehung von Instanzen in die Berechnung des Clustering-Resultates einfließen zu lassen.

Die Einschränkung der Instanzanzahl von Clustern folgt Höppner und Klawonn [ZWL10], welche Constraints zur Restriktion der Instanzmenge verwenden. Diese mengenbezogenen Constraints schränken die mögliche relative Instanzverteilung zwischen mehreren Clustern oder die maximale Instanzanzahl einzelner Cluster ein [ZWL10]. Eine ähnliche Einschränkung von Mengen bei der Zuweisung von Instanzen kann den Domänenexperten auch bei der iterativen Modellierung der Cluster-Hierarchie unterstützen. Der Domänenexperte kann für diese Einschränkung (analog zu Höppner und Klawonn [ZWL10]) die relative Aufteilung der Instanzen eines Knotens der Cluster-Hierarchie sowie die maximale Anzahl an Instanzen eines Knotens festlegen. Der Domänenexperte kann hierdurch beispielsweise eine aus dem Kontext der Anwendung abgeleitete Gleichverteilung der Instanzen zwischen mehreren Clustern oder Beschränkungen der Menge einzelner Zielgruppen abbilden [ZWL10].

Zusätzlich zu den mengenbasierten Restriktionen kann der Domänenexperte außerdem durch die Definition von Regeln den Wertebereich der einem Cluster zugehörigen Instanzen einschränken, um für ihn unrealistische Werte auszuschließen. Die zuweisbaren Werte können für die Umsetzung der Werte-Restriktion beispielsweise vom Domänenexperten mithilfe von Vergleichsoperatoren (größer, kleiner, gleich, ungleich) gefiltert werden. Dies ermöglicht es dem Domänenexperten, seine Expertise über die Verteilung von Attributwerten in bestimmten Zielgruppen bei der Erstellung des Clustering-Resultates zu berücksichtigen. Ein ähnliches Konzept zur Einschränkung von Attributwerten in einem hierarchischen Clustering Konzept konnte in der Literatur nicht identifiziert werden.

Eine mögliche Einbindung der Restriktion der Instanzmengen und Wertebereiche von Clustern in die hierarchische Visualisierung wird in Abbildung 3.6 (A) dargestellt. Der Domänenexperte kann bei dieser exemplarischen Umsetzung mit den Verbindungslinien zwischen einem Cluster-Knoten und seinem Elternknoten interagieren, um die möglichen Werte sowie die Anzahl der zugewiesenen Instanzen zu beschränken. Die vom Domänenexperten definierten Restriktionen werden von der exemplarischen Umsetzung in die hierarchische Visualisierung integriert (z.B. Abbildung 3.6, „Attribut 1 > 20“).

*Szenario (Mengen- und Werte-Restriktionen):* Der Marketingexperte soll ein neu aufgestelltes Getränkesortiment vermarkten und möchte hierfür selektierten Kunden ein Probierpaket zukommen lassen. Die teilnehmenden Kunden möchte er mithilfe einer Clusteranalyse identifizieren. Die Anzahl der vorhandenen Probierpakete ist dabei beschränkt. Der Marketingexperte muss außerdem zwischen Probierpaketen mit und ohne alkoholischen Getränken unterscheiden. In diesem Szenario kann der Domänenexperte Mengenbeschränkungen nutzen, um die Anzahl der Instanzen innerhalb der Cluster der wichtigen Kunden auf die Anzahl vorhandener Probierpakete zu beschränken. Zusätzlich kann der Marketingexperte die Wertebeschränkungen „Alter >= 18“ und „Alter < 18“ verwenden, um den die verschiedenen Arten der Probierpakete (alkoholisch/nicht-alkoholisch) in der von ihm modellierten Cluster-Hierarchie wiederzuspiegeln.

### Interaktion mit der Cluster-Hierarchie

Der Domänenexperte besitzt Domänenwissen, über die hierarchischen Zusammenhänge der Daten und des erstellten Clustering-Resultates. Dieses Domänenwissen muss er in das Clustering-Resultates integrieren können (Abschnitt 1.2.1, A4). Hierfür muss er die erstellte Cluster-Hierarchie an die ihm bekannten hierarchischen Zusammenhänge anpassen können. Die alleinige Interaktion mit den Parametern der Knoten ist hierfür nicht ausreichend, da diese keine Interpretation der (komplexen) Intention des Domänenexperten ermöglichen [BHR+20]. Es werden daher Interaktionen benötigt, welche es dem Domänenexperten ermöglichen, seine Intention (beispielsweise spezifisches Cluster entfernen) zur Anpassung des Clustering-Resultates zu verwenden [BHR+20]. In diesem Abschnitt wird die Integration solcher Interaktionsmöglichkeiten in das Prozessmodell besprochen, um eine entsprechende Anpassung der Cluster-Hierarchie durch den Domänenexperten zu ermöglichen. Die vorgestellten Interaktionsmöglichkeiten werden außerdem in die exemplarische Umsetzung der hierarchische Visualisierung integriert.

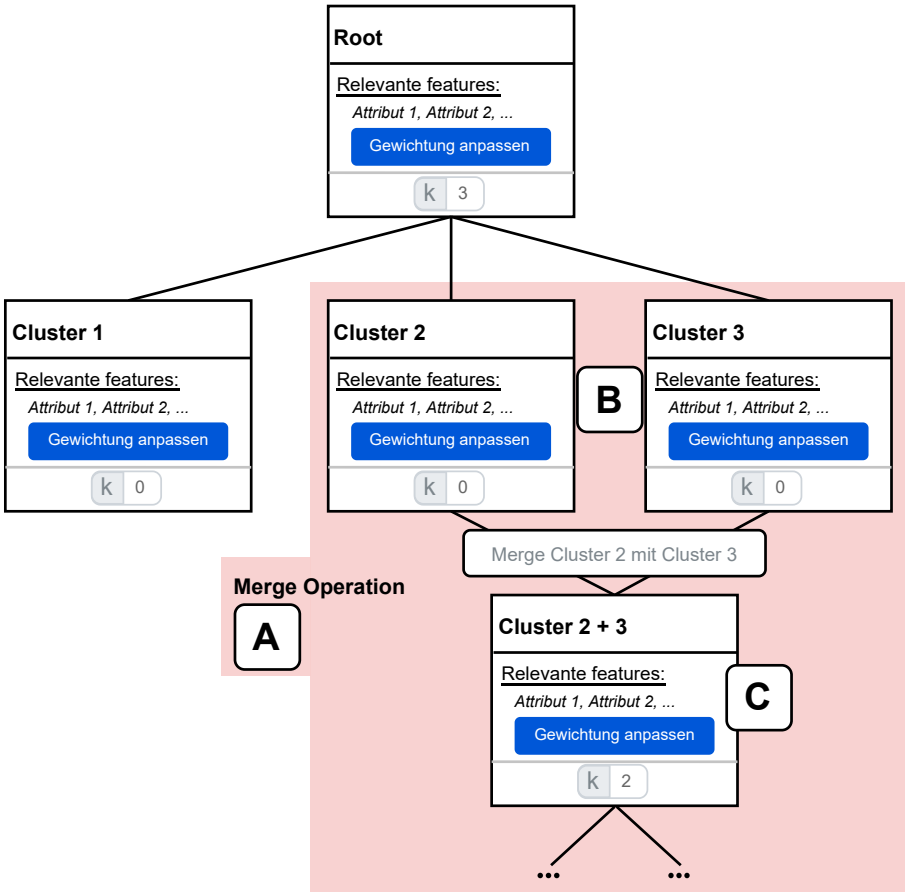



**Abbildung 3.8:** Exemplarische Integration der Merge-Operation (A) in die hierarchische Visualisierung mit Anwendung auf die Cluster 2, 3 (B) und dem daraus resultierenden Cluster (C).

Eine populäre Interaktion mit dem Clustering-Resultat stellt das Splitten (aufteilen) und das Mergen (zusammenführen) von Clustern [BHR+20]. Im Rahmen des vorgestellten Prozessmodells kann die Split Operation einzelner Cluster vernachlässigt werden, da diese implizit durch den hierarchischen Aufbau des Prozessmodells gegeben ist. Die Aufsplittung eines Clusters wird bei dem vorgestellten Prozessmodel direkt durch die nachfolgende Ebene gestellt und kann vom Domänenexperten standardmäßig angepasst werden. Im Gegensatz zum Aufsplitten einzelner Cluster stellt das Mergen mehrerer Cluster eine Erweiterung der vorhandenen Interaktionen dar. Der Domänenexperte kann die Merge Operation nutzen, um mehrere unterbelegte Cluster mit ähnlichem Inhalt zu mergen [ABV13].

Für die Anwendung einer Merge Operation kann der Domänenexperte direkt mit dem von ihm identifizierten Zielknoten aus der hierarchischen Visualisierung interagieren (exemplarisch in Abbildung 3.6). Das Resultat der Anwendung der Merge Option muss in die Hierarchie-Visualisierung eingebunden werden, um den weiteren Ausbau der Hierarchie nach dem Mergen mehrerer Knoten zu ermöglichen und die Auswirkung der Merge Operation (was genau wird zusammengeführt, wie sieht das Ergebnis aus) einschätzen zu können. Eine Möglichkeit für die Einbindung der Merge Operation in die hierarchische Visualisierung zeigt Abbildung 3.8 (A), in welcher die beiden Cluster „Cluster 2“ und „Cluster 3“ (B) zu einem neuen Cluster (C) zusammengeführt werden. Der Domänenexperte kann in diesem Beispiel weiterhin den Inhalt der beiden Ursprungscluster (Abbildung 3.8, B) inspizieren und so die Zusammensetzung des neu gebildeten Clusters (C) nachvollziehen. Er kann außerdem das resultierende Cluster (C) weiter aufteilen, und so die Cluster-Hierarchie fortführen.

Eine weitere Möglichkeit, die Hierarchie anzupassen, ist die Entfernung einzelner Knoten [LKC+12]. Der Domänenexperte kann von dieser Interaktionsmöglichkeit Gebrauch machen, wenn er ein bestimmtes Cluster aus dem gesamten Cluster-Resultat entfernen möchte [LKC+12].

Abschließend muss der Domänenexperte zusätzlich in Fällen, in welchen er keine klare Vorstellung über die Anpassung der Cluster-Hierarchie hat, vom System unterstützt werden (Abschnitt 1.2.1, A4). Diese Unterstützung kann durch das Vorschlagen einer Cluster-Hierarchie durch das System realisiert werden. Hierfür kann eine automatisierte Erstellung von Teilen der Cluster-Hierarchie verwendet werden.

Die automatische Generation der Cluster-Hierarchie kann durch die Anwendung von mehrfachverzweigten (die meisten hierarchischen Clustering-Verfahren sind binär verzweigt, d. h. Aufteilung in jeweils zwei weitere Knoten) Cluster-Verfahren bewerkstelligt werden [BTH12]. Einen populären Ansatz für die Erstellung solcher Mehrfachverzweigter Hierarchien stellen Blundell et al. [BTH12] mit „Bayesian Rose Trees“, einem auf „Bayesian hierarchical clustering“ basierten Ansatz für die Generierung einer Hierarchie mit beliebiger Verzweigungsstruktur dar.

Eine exemplarische Umsetzung der automatisierten Hierarchie-Generation zeigt Abbildung 3.9, welche ausgehend von „Cluster 1“ (A) die weitere Generation von zwei Hierarchieebenen darstellt. Der Domänenexperte kann hierfür in einem Dialog (B) die gewünschte Anzahl an Hierarchieebenen, welche er generieren möchte (in der Abbildung exemplarisch zwei), spezifizieren. Darauf folgend erhält er vom System eine (automatisch erstellte) mögliche Fortsetzung der Cluster-Hierarchie (C) mit der zuvor spezifizierten Anzahl (zwei) an Hierarchie-Ebenen.

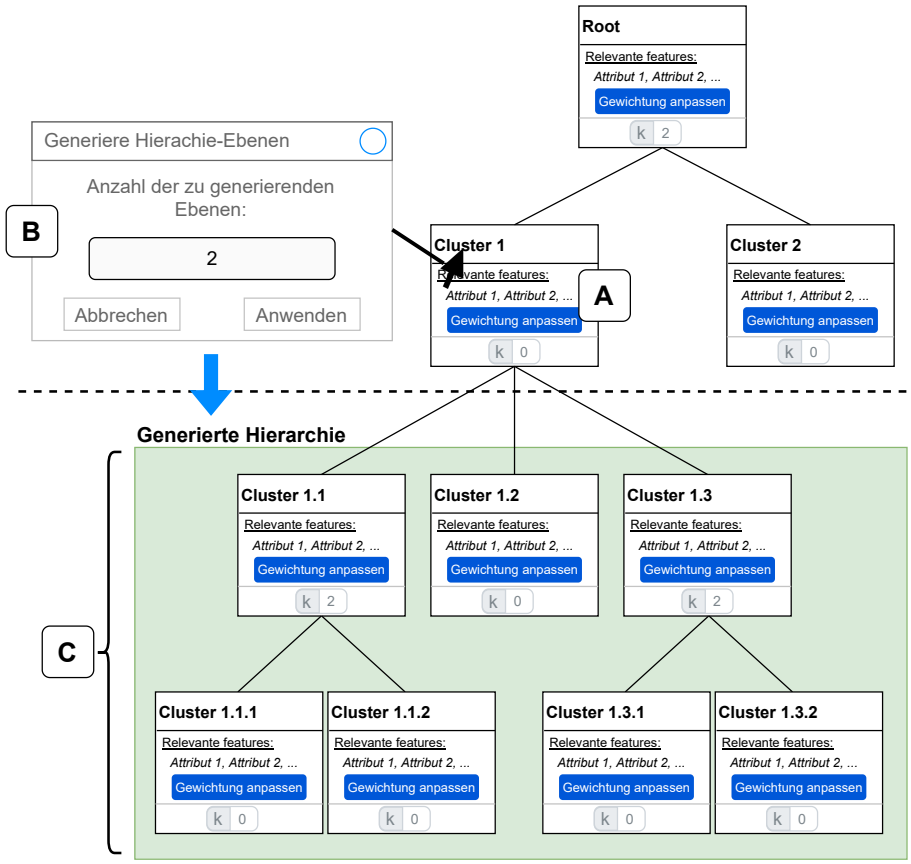



**Abbildung 3.9:** Exemplarische Umsetzung der automatisierten Generierung einer Cluster-Hierarchie mit der beispielhaften Generation von zwei Hierarchie-Ebenen (B) vom Ausgangscluster (A) und der aus der Operation resultierenden Hierarchie (C).

### Bearbeitung von Cluster-Inhalten

Der Domänenexperte hat durch sein Domänenwissen oftmals Vorstellungen über den Inhalt einzelner Cluster. Damit die Anforderungen des Domänenexperten (Abschnitt 1.2.1, A4) erfüllt werden können, muss der Domänenexperte fähig sein, das Clustering-Resultat auf Basis dieser Vorstellungen anzupassen. Hierzu werden in diesem Abschnitt Interaktionen zur Bearbeitung von Cluster-Inhalten eingeführt. Der Domänenexperte kann diese Interaktionen nutzen, um Cluster mit für ihn „inkorrekten“ (das heißt nicht seiner Vorstellung entsprechenen) Inhalten zu verfeinern. Für die Identifikation der Cluster kann der Domänenexperte die Informationen der Cluster-Visualisierung (Abschnitt 3.2.2) des Prozessmodells verwenden und diese mit seinen Vorstellungen abgleichen.

Eine Möglichkeit für eine solche Verfeinerung stellen Srivastava et al. [SZS16] durch die Option Cluster, welche nicht der Vorstellung des Domänenexperten entsprechen, „abzulehnen“ dar. Basierend auf den zuvor abgelehnten Clustern generiert das System neue Cluster, welche sich (nach Möglichkeit) von zuvor abgelehnten Clustern unterscheiden [SZS16]. Im Rahmen des Prozessmodells kann der Domänenexperte eine solche Operation verwenden, um einzelne Knoten der Hierarchie, welche nicht seiner Vorstellung entsprechen, erneut zu berechnen, um eine unterschiedli-

che Aufteilung zu erzwingen und so das modellierte Cluster-Modell seiner Vorstellung anzupassen. Die mögliche Integration der Ablehnung einzelner Knoten wird exemplarisch in Abbildung 3.6 (C1), als Option in einem Kontextmenü eines Hierarchie-Knotens dargestellt.

Eine weitere Möglichkeit für die Korrektur von inkorrekt zugewiesenen Instanzen ist die manuelle Neuzuweisung von Instanzen zu einem anderen Cluster [BFDL10; CD18; CDG+17]. Die Zuweisung einzelner Instanzen kann in die Cluster-Visualisierung (Abschnitt 3.2.2) des derzeit vom Domänenexperten betrachteten (selektierten) Clusters integriert werden. Eine exemplarische Integration stellt die tabellarische Darstellung der Daten in Abbildung 3.2 (A.2), in welcher der Domänenexperte die Zuweisung einzelner Instanzen anpassen kann, dar. Basierend auf den Zuweisungen des Domänenexperten kann das System außerdem automatisiert weitere ähnliche Instanzen identifizieren und zuweisen [BHR+20].

Abschließend kann der Domänenexperte den Inhalt von Clustern durch die Verwendung bereits gelabelter Instanzen als Seeds spezifizieren [BBM02]. Dies ermöglicht es dem Domänenexperten, den gewünschten Inhalt von Clustern anzugeben, um so die Bedeutung einzelner Cluster festzulegen [BBM02]. Die daraus resultierende Verbesserung des Clustering-Resultates demonstriert Sugato et al. [BBM02] mithilfe eines nicht interaktiven Clustering-Ansatzes. Im Rahmen des interaktiven Prozessmodells kann Seeding integriert werden, um dem Domänenexperten die exakte Definition des gewünschten Inhaltes einzelner Knoten der Cluster-Hierarchie zu ermöglichen. Die exemplarische Übersicht für die Integration von Interaktionsmöglichkeiten in die hierarchische Visualisierung (Abbildung 3.6) integriert Seeding als Interaktionsmöglichkeit (B) durch die direkte Zuordnung vorhandener Instanzen zu einem gewünschten Knoten der Cluster-Hierarchie. Aufgrund des hierarchischen Clustering Ansatzes des Prozessmodells muss die interaktive Anwendung des Seeding-Verfahrens in die darüberliegenden Ebenen der Clustering-Hierarchie propagiert werden (da jede Instanz eines Unterknotens der Cluster-Hierarchie implizit auch ein Element des zugehörigen Elternknotens ist).

### 3.3.3 Systemvorschläge

Im Rahmen des Prozessmodells werden systemgenerierte Vorschläge verwendet, um die Anforderung des Domänenexperten nach einer aktiv vom System gestützten Verfeinerung des Clustering-Resultates (Abschnitt 1.2.1) zu erfüllen. Hierfür wird der Domänenexperte vom System eigenständig nach relevanten Informationen befragt und auf mögliche Anpassungen hingewiesen. Die Vorschläge bieten dem Domänenexperten einfach anwendbare Änderungsoperationen zur Anpassung des derzeit selektierten Clusters (Abbildung 3.2, C) und unterstützen ihn zusätzlich bei der Identifikation potenzieller Möglichkeiten zur Verfeinerung des Clustering-Resultates. Für die Erstellung der Systemvorschläge wird zum einen Active Clustering (Abschnitt 2.5.2) für die „intelligente“ Abfrage des paarweisen Verhältnisses zwischen Instanzen und der darauffolgenden Generation von Constraints verwendet. Zum anderen wird der Domänenexperte durch Vorschläge zur automatisiert ausführbaren Anwendung bereits vorgestellter (Abschnitt 3.3.1, Abschnitt 3.3.2) Interaktionen auf etwaige Verbesserungsmöglichkeiten hingewiesen.

Eine exemplarische Umsetzung von Systemvorschlägen zeigt Abbildung 3.10, in welcher der Domänenexperte vom System Vorschläge (A) für das von ihm selektierte Cluster („Cluster 1“) gestellt bekommt. Diese Umsetzung beinhaltet zum einen die Verwendung von Active Clustering

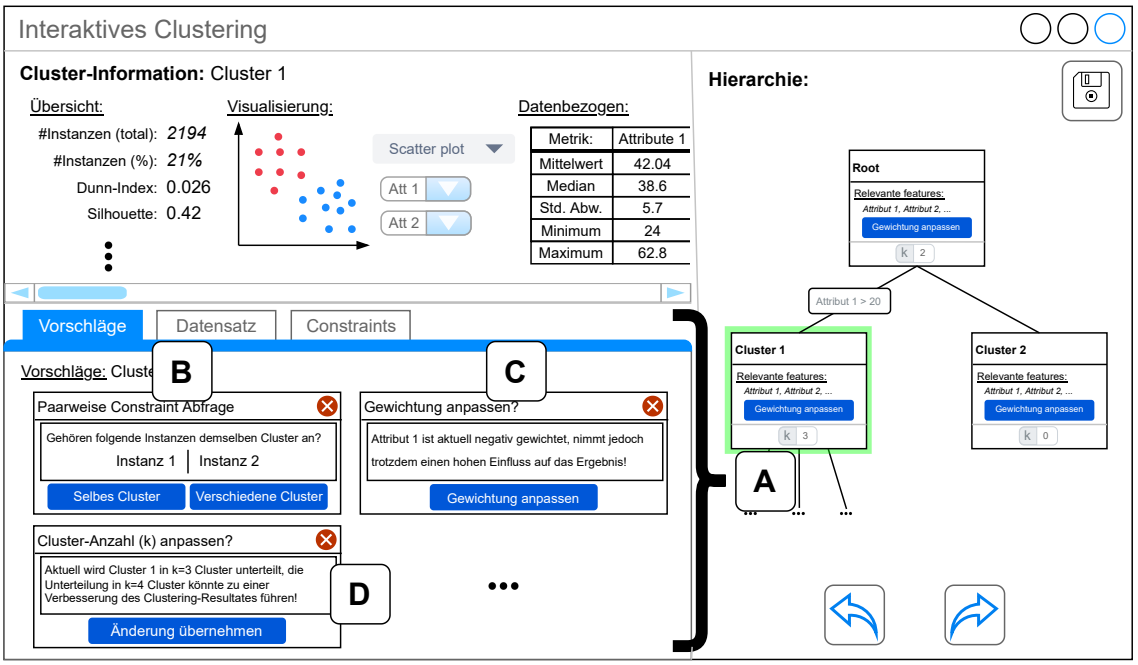



**Abbildung 3.10:** Exemplarische Integration von Systemvorschlägen für selektierte Cluster in die Cluster-Visualisierung (Abschnitt 3.2.2) mit den erzeugten Vorschlägen (A), eingebundenem Active Clustering (B) sowie weiteren Vorschlägen (C, D) zur Anpassung der vorhandenen Cluster-Konfiguration.

zur Abfrage des paarweisen Verhältnisses zweier Instanzen (Abbildung 3.10, B) und zum anderen Vorschläge, welche den Domänenexperten auf mögliche Anpassungen des Clustering-Resultates hinweisen (C-D).

In den folgenden Abschnitten werden mögliche Systemvorschläge genauer besprochen. Die vorgestellten Systemvorschläge werden hierfür nach dem Ziel der eingeleiteten Interaktion (die Durchführung von Active-Clustering, die Zuweisung von Instanzen, das Hinweisen auf Abweichungen von Präferenzen und die Anpassung von Parametern) in vier Abschnitte unterteilt.

### Active-Clustering

Der Domänenexperte hat einen Bedarf (Abschnitt 1.2.1, A4), vom System aktiv über wichtige Informationen befragt zu werden. Die Verwendung von Active-Clustering ermöglicht eine solche Befragung zur effizienten Erzeugung paarweiser (Must-/Cannot-Link) Constraints (Abschnitt 2.5.2). Diese Constraints können daraufhin zur Verfeinerung des Clustering-Resultates verwendet werden[AA20; BBM04a; KG20]. Die genaue Einbindung des Domänenexperten als Orakel (wie werden die Informationen abgefragt) wird von Active-Clustering Ansätzen jedoch oftmals nicht genau definiert (Abschnitt 2.6). Dies erschwert die Integration in die interaktive Verfeinerung, da ein einfacher Vergleich der einzelnen Attributwerte für hochdimensionale Daten für Menschen unzureichend ist. Für die Integration von Active-Clustering als Systemvorschlag wird für die Einbindung des Domänenexperten eine Kombination aus der Visualisierung der abgefragten Instanzen mithilfe einer durch Dimensionsreduktion erstellten 2D Projektion (ähnlich zu Okabe und

Yamada [OY11]) und der bereits in der hierarchischen Visualisierung verwendeten Identifikation besonders wichtiger Attribute (von Behringer et al. [BHTM22]) verwendet. Der Domänenexperte erhält bei dieser Kombination zunächst die Möglichkeit, die relevanten Attribute der abgefragten Instanzen miteinander zu vergleichen, wodurch die Komplexität des Vergleichs reduziert wird.

Die Kombination (exemplarisch in Abbildung 3.11) verwendet hierzu die visuelle Darstellung der abgefragten Instanzen (Abbildung 3.11, B) mithilfe einer 2D Projektion [OY11], um eine übersichtliche Darstellung zu gewährleisten. Der Domänenexperte verliert bei der alleinigen Verwendung dieser Projektion allerdings die Möglichkeit, sein Domänenwissen über die einzelnen Attribute der Daten anzuwenden (Abschnitt 2.2). Die Projektion muss daher zusätzlich durch den Vergleich einzelner Instanz-Werte anhand der identifizierten besonders wichtigen Features (Abbildung 3.11, A) ergänzt werden. Die für die Aufteilung der Daten besonders wichtigen Features tragen maßgeblich zur Separierung einzelner Instanzen bei [BHTM22]. Es ist daher in vielen Fällen bereits ausreichend, bei der Abfrage der Instanzen die Werte dieser Attribute miteinander zu vergleichen. Dem Domänenexperten werden zum Vergleich daher zuerst die als besonders wichtig identifizierten Attribute (Abbildung 3.11, C) präsentiert. Bei Bedarf können zusätzlich die übrigen Attribute hinzugezogen werden (Abbildung 3.11, D).

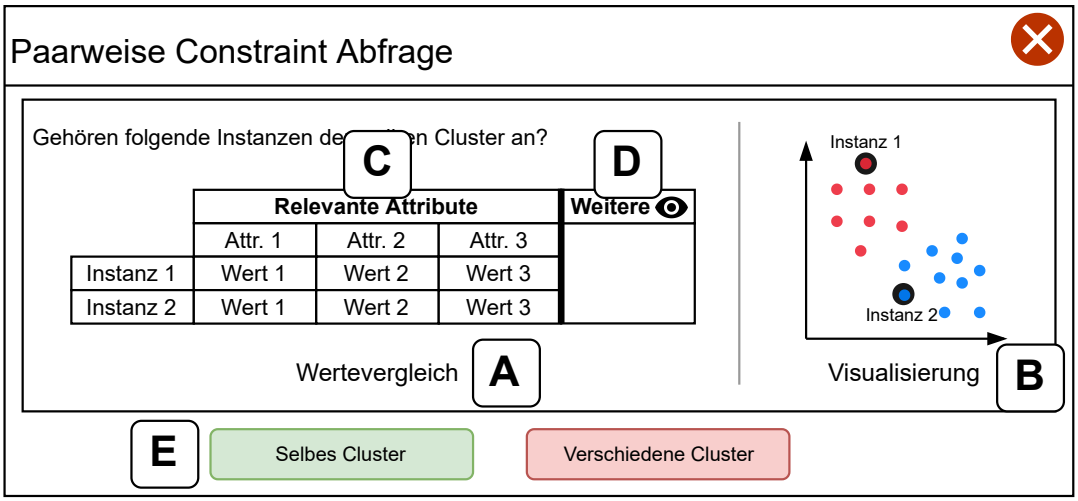



**Abbildung 3.11:** Exemplarische Einbindung des Domänenexperten in Active-Clustering-Verfahren mithilfe der Erzeugung von paarweisen Instanz-Abfragen als Systemvorschläge. Das Beispiel enthält den Vergleich einzelner Attributwerte anhand der identifizierten wichtigen Features (A) sowie eine 2D Projektion der Daten mit hervorgehobenen Instanzen (B) und eine Möglichkeit zur Beantwortung der Abfrage (E).

Abschließend kann der Domänenexperte, nach ausreichendem Vergleich beider Instanzen, den Systemvorschlag beantworten (E). Basierend auf der Antwort werden vom System die entsprechenden Must-/Cannot-Link Constraints („Selbes Cluster“ -> Must-Link oder „Verschiedene Cluster“ -> Cannot-Link) erstellt.

#### Zuweisung von Instanzen

Die bereits vorgestellte interaktive Operation der Fehlerkorrektur durch die manuelle Zuweisung einzelner Instanzen zu Clustern (Abschnitt 3.3.2) kann durch die systeminitiierte Selektion geeigneter Instanzen ergänzt werden [CD18; IMN+07]. Hierfür weist das System den Domänenexperten eigenständig auf (die für die Zuweisung) geeignete Instanzen hin, wodurch die Anforderung des Domänenexperten nach einer aktiven Unterstützung während der Anwendung von Operationen (Abschnitt 1.2.1, A4) erfüllt wird. Die vom System für die manuelle Zuweisung vorgeschlagenen Instanzen können sich beispielsweise aus Ausreißern [IMN+07] oder Instanzen mit hoher Unsicherheit [CD18] zusammensetzen. Aufgrund der des Prozessmodells zugrundeliegenden iterativen Anpassung der Hierarchie ist zusätzlich die Abfrage von Instanzen, welche bei Anpassungen zwischen Iterationen häufig zwischen verschiedenen Clustern hin- und hergeschoben werden, möglich.

Ähnlich zu den Systemvorschlägen des Active-Clustering in Abschnitt 3.3.3 wird der Domänenexperte von vorhandenen Ansätzen ([CD18; IMN+07]) durch eine 2D/3D Projektion bei der Zuweisung der Instanzen zu korrekten Clustern unterstützt. Diese Projektionen können erneut analog zu den Systemvorschlägen des Active-Clustering mit den zur Beschreibung der Hierarchie identifizierten besonders relevanten Attributen kombiniert werden. Anders als bei der exemplarischen Umsetzung des Wertevergleichs der Active-Clustering Vorschläge (Abbildung 3.11, A) müssen bei der Selektion des „korrekten“ Clusters allerdings nicht die Werte zweier Instanzen verglichen werden. Stattdessen können die Werte relevanter Attribute der Instanz statistischen Auswertungen (z.B. dem Median) gegenübergestellt werden, um eine Entscheidungshilfe bei der Zuweisung (welches Cluster passt am besten zu der Instanz) stellen zu können.

#### Übereinstimmung mit Präferenzen

Die Gewichtung einzelner Attribute (Abschnitt 4.3.1) ermöglicht es dem Domänenexperten, den Einfluss, welchen einzelne Attribute auf die Separierung der Daten nehmen, direkt anzupassen. Der Domänenexperte kann diese Gewichtung nutzen, um seiner Vorstellung entsprechend Attribute stärker oder weniger stark in das Clustering-Verfahren einzubringen. Neben der Bearbeitung des Einflusses der Attribute verwendet der vorgestellte Ansatz die Identifikation für die Aufteilung von Clustern wichtiger Attribute von Behringer et al. [BHTM22], um den Einfluss, welchen Attribute auf die Aufteilung der Daten nehmen, abzuschätzen und besonders wichtige Attribute zu identifizieren.

Die Grundidee der Generation von Systemvorschlägen auf Basis der Präferenzen des Domänenexperte ist, dass der Domänenexperte bei einer hohen Attribut-Gewichtung einen erhöhten Einfluss des gewichteten Attributes auf das Clustering-Resultat (analog niedrigerer Einfluss für niedrige Gewichtungen) erwartet. Die im Konzept integrierte Identifikation wichtiger Attribute bewertet die Bedeutung, welche jedes Attribut für die Erstellung des Clustering-Resultates übernimmt [BHTM22]. Die vom System erstellte Bewertung der Attribute kann daher mit den Erwartungen des Domänenexperten abgeglichen werden. Im Fall einer Diskrepanz zwischen der Erwartung des Domänenexperten und dem vom System identifizierten Einfluss wird die Intention des Domänenexperten (weshalb die Gewichtung angewendet wurde) nicht erfüllt. Das System kann daher auf Basis dieser Diskrepanz einen Vorschlag zur Anpassung der Gewichtung erzeugen,

um den Domänenexperten auf diesen Missstand hinzuweisen. Dieser Vorschlag unterstützt den Domänenexperten bei der effektiven Anwendung von Interaktionen und trägt so zur Erfüllung der Anforderungen des Domänenexperten (Abschnitt 1.2.1, A4) bei.

Ein Beispiel für eine solche Empfehlung zeigt die Abbildung 3.10 (C), in welcher der Domänenexperte das Gewicht des Attributes „Attribut 1“ bereits reduziert hat, dieses allerdings weiterhin einen hohen Einfluss auf die Aufteilung der Daten nimmt. Der Domänenexperte erhält daher von dem System die Empfehlung, die Gewichtung des Attributes weiter anzupassen.

#### Anpassung von Parametern

Eine weitere Möglichkeit für die Erzeugung von Systemvorschlägen stellen automatisiert identifizierte, für den angewandten Clustering-Algorithmus (möglichst) optimale Parameter, dar. Diese identifizierten Parameter können dem Domänenexperten vom System als Alternative zu den derzeit verwendeten Parametern vorgeschlagen werden. Der Domänenexperte kann so in Fällen, in welchen er sich (aufgrund seiner fehlenden Analyse-Erfahrung) über die Eignung der gewählten Parameter unsicher ist, unterstützt werden. Die Notwendigkeit einer solchen Unterstützung geht aus der Anforderung des Domänenexperten, Hilfe vom System bei der Vermeidung von Fehlern zu erhalten (Abschnitt 1.2.1, A4), hervor.

Eine automatisierte Identifikation wird beispielsweise von Iorio et al. [IMN+07]) verwendet, welche verschiedene Parameterkonfigurationen auf Datenausschnitten testen und mithilfe von Qualitätsindikatoren (Dunn Index, Silhouette) die „Beste“ auswählen. Für die automatisierte Berechnung der Parameter eines selektierten Knotens können vom System daher analog hierzu verschiedene Parameterkonfigurationen auf einem Datenausschnitten getestet und mithilfe von Qualitätsindikatoren ausgewählt werden. Eine exemplarische Umsetzung für einen solchen Vorschlag stellt die Abbildung 3.10 (D), in welchem dem Domänenexperten ein anderer Wert für den Parameter k des k-Means Algorithmus vorgeschlagen wird, dar.

# 4 Implementierung

Im Rahmen dieser Arbeit wurde das vorgestellte Konzept zur Unterstützung von Domänenexperten bei der iterativen Verfeinerung von Clustering-Resultaten (Kapitel 3) prototypisch implementiert, um die Umsetzung des Konzeptes zu demonstrieren.


**Abbildung 4.1:** Übersicht über den zugrundeliegenden Clustering Communicator. Die Abbildung zeigt die Möglichkeit der klassischen Visualisierung von Clustering-Resultaten (A) sowie die Beschreibung einzelner Cluster-Inhalte (B) mithilfe der Identifizierung wichtiger Feature (C).

Der Prototyp bietet dem Domänenexperten eine Reihe der, für die Integration in das Prozessmodell vorgestellten Interaktionsmöglichkeiten (Abschnitt 3.3) zur Verfeinerung von Clustering-Resultaten. Die implementierten Interaktionsmöglichkeiten kombinieren die vorgestellten Interaktionskategorien (Abschnitt 2.6.1) der Parameter- und Ergebnis-Anpassung mit System initiierten Interaktionen und stellen so ein hybrides Interaktionsmodell dar.
Die Basis für die Implementierung bietet der an der Universität Stuttgart entwickelte Clustering Communicator (Abbildung 4.1), der Clustering-Resultate anhand der Identifikation wichtiger Feature (Abbildung 4.1, B) sowie der klassischen Visualisierung von Clustering-Resultaten (Parallele Koordinaten und Histogramme, A) für Domänenexperten leichter verständlich macht.

Das User-Interface des entwickelten Prototypen zeigt Abbildung 4.2 anhand eines exemplarischen Anwendungsfalles, in welchem ein Domänenexperte den Prototyp nutzt, um Kunden basierend auf verschiedenen Einkaufskategorien zu segmentieren. Die Übersicht zeigt die entsprechend des vorgestellten Prozessmodells eingeführte hierarchische Visualisierung (Abschnitt 3.2.1) der Daten (A), aus welcher der Domänenexperte derzeit das Cluster „freizeit_enthusiasten“ (C) selektiert hat, um dieses in der Cluster-Visualisierung (B) näher zu betrachten.

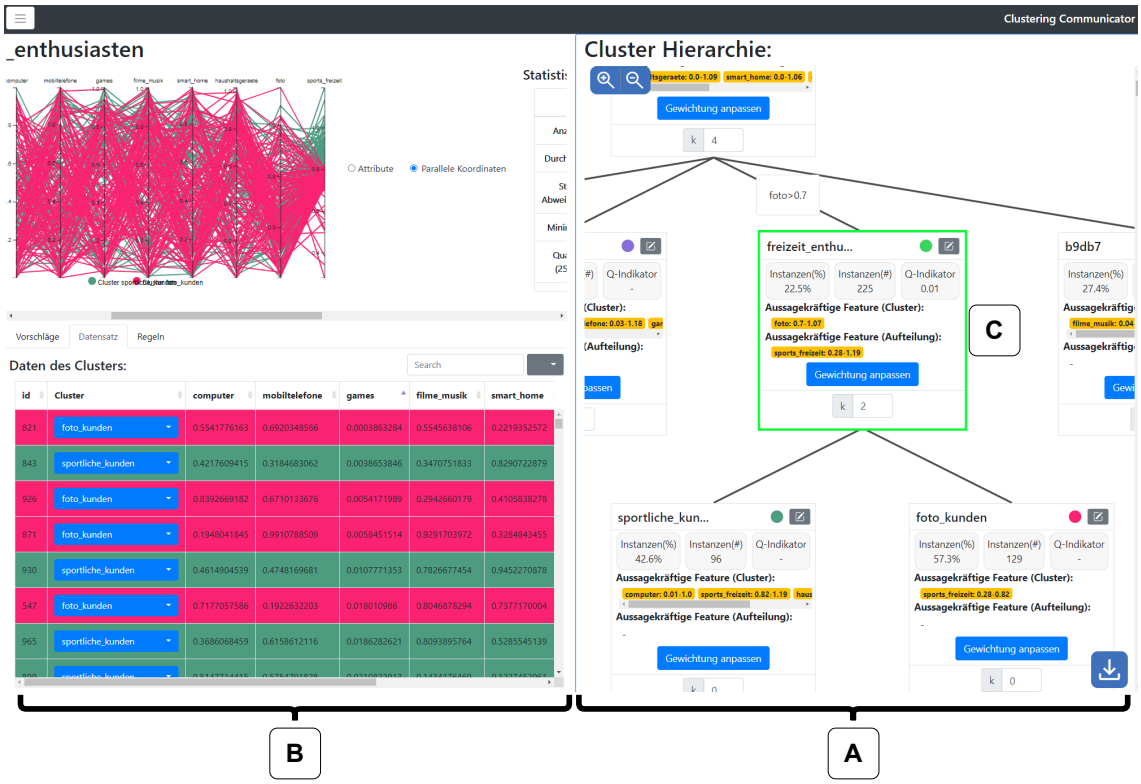



**Abbildung 4.2:** Übersicht über den entwickelten Prototyp mit der hierarchischen Visualisierung (A) und der Cluster-Visualisierung (B) des derzeit selektierten Clusters (C).

In den folgenden Sektionen wird der Aufbau des Prototyps sowie die integrierten Elemente zur Kommunikation des Clustering-Resultates und der Interaktion mit dem Clustering-Modell zur Verfeinerung des Resultates vorgestellt. Hierfür wird zuerst auf die Architektur des Prototyps eingegangen.

Anschließend wird die Kommunikation des Clustering-Resultates mithilfe der Hierarchie- und Cluster-Visualisierung (Abschnitt 4.2) diskutiert. Abschließend wird die Implementierung der interaktiven Möglichkeiten (Abschnitt 4.3), welche dem Domänenexperten zur Verfeinerung der Clustering-Resultate zur Verfügung stehen, diskutiert.

## 4.1 Architektur

Der implementierte Prototyp verwendet ein JavaScript-basiertes Frontend, welches für die Interaktion mit dem Domänenexperten sowie die Repräsentation von Clustering-Resultaten zuständig ist. Das Frontend kommuniziert über das MQTT-Protokoll mit einem Python-basierten Backend-Service, welches für die Berechnung der Clustering-Resultate und zugehöriger Metriken zur Umschreibung der Resultate zuständig ist.

Der Aufbau des Prototyps (Abbildung 4.3) mit den einzelnen enthaltenen Komponenten und der Kommunikation (Abbildung 4.3, blau) zwischen diesen Komponenten wird im Folgenden erklärt. Hierzu wird zuerst die Funktionsweise der einzelnen Komponenten vorgestellt und anschließend anhand eines exemplarischen Ablaufes veranschaulicht.

#### Funktionsweise der Komponenten

Dieser Abschnitt erklärt die Funktionsweise der einzelnen Komponenten. Es werden hierfür zuerst die Komponenten des JavaScript-Frontends (Abbildung 4.3, A) und anschließend die Komponenten des Python-Backend (Abbildung 4.3, B) beschrieben.

*JavaScript-Frontend (A).* Das JavaScript-Frontend (Abbildung 4.3, A) ist für die Repräsentation des im Python-Backend (Abbildung 4.3, B) berechneten Clustering-Resultates und für die Interaktion mit dem Domänenexperten verantwortlich. Die Umsetzung des JavaScript-Frontends orientiert sich am Ablauf des Prozessmodells (Abschnitt 3.1) und ist dementsprechend in eine Komponente für die hierarchische Visualisierung der Hierarchie (Abbildung 4.3, A1) und in eine Komponente für die Visualisierung des derzeit selektierten Clusters (Abbildung 4.3, A2) unterteilt.

Hierarchie-Visualisierung (A1): Die Komponente für die Hierarchie-Visualisierung beinhaltet die Logik für die Darstellung der Cluster-Hierarchie und die Interaktion mit Hierarchie-Elementen. Für die Modellierung der Hierarchie enthält die Komponente daher Teilkomponenten für die Repräsentation der einzelnen Hierarchie-Elemente (Knoten und Kanten) sowie eine Teilkomponente, um die Interaktion mit dem Domänenexperten zu unterstützen.

Die Teilkomponenten visualisieren die aus der Berechnung über die vorhandenen Cluster vorhandenen (beispielsweise wichtige Attribute oder Qualitätsindikatoren) Informationen. Außerdem werden (von den Teilkomponenten) die auf den Clustern angewendeten Interaktionen an die Kommunikations-Schnittstelle (Abbildung 4.3, A3) des Frontends übermittelt. Auf Basis der übermittelten Interaktion wird extern ein aktualisiertes Clustering-Resultat berechnet. Die Hierarchie-Komponente aktualisiert daher lediglich die dargestellte Cluster-Hierarchie, basierend auf den von der Kommunikations-Schnittstelle erhaltenen neu berechneten Clustering-Resultaten.

Cluster-Visualisierung (A2): Die Komponente für die Cluster-Visualisierung ist für die detaillierte Darstellung des derzeit selektierten Clusters zuständig. Die Komponente enthält daher die Logik zur Darstellung des Clusterinhaltes mithilfe von Diagrammen (Histogrammen und Parallelen Koordinaten), Statistiken und in tabellarischer Form. Zusätzlich präsentiert die Komponente die vom System generierten Vorschläge zur Verfeinerung des Clustering-Resultates und verarbeitet die Interaktionen mit diesen.

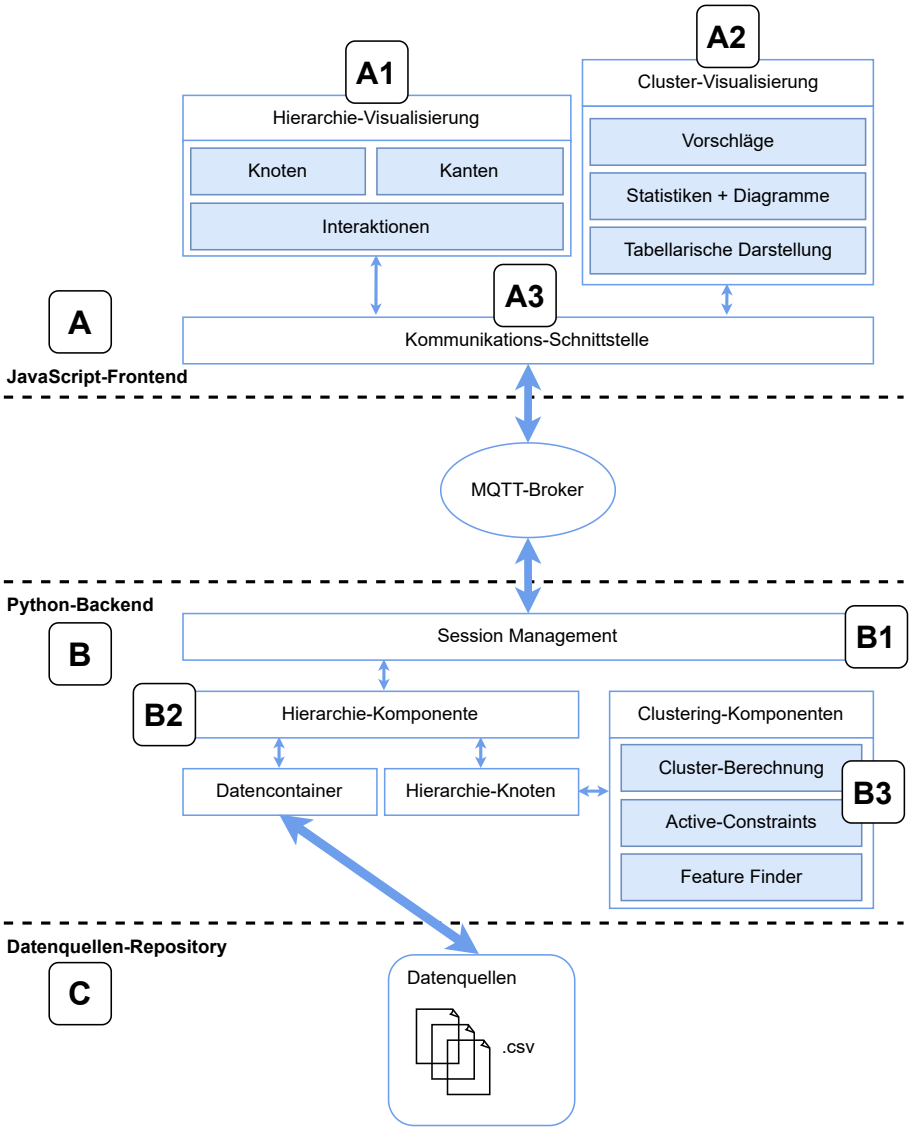



**Abbildung 4.3:** Aufbau des Prototyps mit Frontend- und Backendschicht (A, B), den einzelnen Komponenten innerhalb der Schichten (A1-A3, B1-B3) sowie Aufzeigen der Kommunikation zwischen verschiedenen Komponenten (blau).

Die Komponente für die Cluster-Visualisierung führt (erneut) keine eigenständigen Berechnungen durch, sondern visualisiert lediglich die von der Kommunikations-Schnittstelle erhaltenen Informationen.

Kommunikations-Schnittstelle (A3): Die Kommunikations-Schnittstelle ist für die Übermittlung von Informationen zwischen dem JavaScript-Frontend und dem Python-Backend verantwortlich. Für die Übertragung der Informationen wird das MQTT-Protokoll verwendet. Bei der parallelen Anwendung mehrerer Clusteranalysen können die Kommunikations-Schnittstellen über IDs jeweils einer Analyse-Instanz zugeordnet werden.

*Python-Backend (B).* Das Python-Backend (Abbildung 4.3, B) ist für die Berechnungen des im JavaScript-Frontend (Abbildung 4.3, A) modellierten Clustering-Modells zuständig. Das Python-Backend umfasst die folgenden Komponenten:

Session Management (B1): Ein integraler Teil des Python-Backends ist die Unterstützung mehrerer parallel ausgeführter Analyse-Instanzen (gleichzeitig ausgeführte Clusteranalysen) durch die serverseitige Verwaltung von Sitzungen (Sessions) basierend auf einer der Frontend-Instanz zugewiesenen Session-ID. Die Session Management Komponente speichert zu diesem Zweck den aktuellen Sitzungszustand (Session State) der existierenden Analysen und ordnet eingehende Anfragen ihrer korrespondierenden Sitzung zu. Zusätzlich wird von der Session Management Komponente die korrekte sequentielle Abarbeitung der Anfragen der einzelnen Sitzungen sichergestellt.

Hierarchie-Komponente (B2): Die Hierarchie-Komponente bildet das verwendete Clustering-Modell ab und stellt dadurch die Grundlage für die Berechnung des Clustering-Resultates. Für die Abbildung des Clustering-Modells reflektiert die Hierarchie-Komponente die im Frontend modellierte Cluster-Hierarchie, mithilfe von Hierarchie-Knoten, welche über eine ID dem korrespondierenden interaktiven Knoten der JavaScript-Frontend Sitzung zugeordnet werden können. Die einzelnen Hierarchie-Knoten Instanzen enthalten die Konfiguration (Parameter, Regeln und Constraints), welche zur Berechnung der Clustering-Resultate benötigt werden.

Clustering-Komponenten (B3): Die Clustering-Komponenten erhalten einzelnen Knoten der Hierarchie als Input und berechnen auf Basis der Parameter und Dateninstanzen des Knotens das Clustering-Resultat der folgenden Hierarchie-Ebene. Zusätzlich zu der Berechnung des Clustering-Resultates werden von den Clustering-Komponenten die für die Visualisierung des JavaScript-Frontends relevanten Informationen (beispielsweise die für die Aufteilung wichtigen Feature und die abzufragenden Constraints) berechnet.

Datencontainer: Der Daten-Container verwaltet die zur Analyse verwendeten Daten und ermöglicht den einfachen Zugriff auf unterschiedliche Repräsentationen (beispielsweise als NumPy-Array oder Pandas Dataframe) der Daten.

*Datenquellen-Repository (C).* Das Datenquellen-Repository ist die Sammlung der zur Analyse verfügbaren Datenquellen. Als Datenquellen verwendet der Prototyp CSV-Dateien. Eine Erweiterung zur Unterstützung von Datenbanken oder zusätzlichen Dateiformaten ist jedoch möglich.

#### Exemplarischer Ablauf

Der exemplarische Ablauf verwendet einen Marketingexperten, welcher die Gewichtung eines Knotens der Hierarchie anpasst, um das Zusammenspiel der vorgestellten Komponenten zu demonstrieren. Der Marketingexperte interagiert zur Abänderung der Gewichtung zuerst mit dem gewünschten Knoten der Hierarchie. Diese Interaktion wird vom JavaScript Frontend (Abbildung 4.3, A) ausgeführt. Der Knoten, mit welchem der Domänenexperte interagiert, ist eine Teilkomponente der Hierarchie-Visualisierung. Nach der Bestätigung der Änderung wird die Anwendung der Interaktion über die Kommunikations-Schnittstelle (Abbildung 4.3, A3) des Frontends über das MQTT-Protokoll an das Python-Backend (B) übertragen.

Die Session Management Komponente (Abbildung 4.3, B1) des Python-Backends wird daraufhin über die Anwendung der Interaktion informiert und ordnet die Interaktion der bereits im Backend existierenden Sitzung des Marketingexperten zu. Anschließend verarbeitet die Hierarchie-Komponente (Abbildung 4.3, B2) des Python-Backends die Interaktion durch die Anpassung der Gewichtung des (mit der Knoten-ID) korrespondierenden Hierarchie-Knotens. Auf die Anpassung folgend wird der Hierarchie-Knoten (und die eventuell in der Hierarchie untergeordneten Knoten) neu berechnet. Für die Neuberechnung werden die Clustering-Komponenten (Abbildung 4.3, B3) verwendet. Aus der Berechnung resultiert eine aktualisierte Aufteilung der Instanzen, neu identifizierte für die Aufteilung wichtige Features (Feature Finder), die neuen paarweisen Instanzen zur Abfrage (Active-Constraints) sowie statistische Informationen.

Die aktualisierten Informationen werden über die Session Management Komponente (Abbildung 4.3, B1) an die korrespondierende Frontend-Instanz (Abbildung 4.3, A) übermittelt. Die Frontend Instanz verwendet diese Informationen zur Aktualisierung der Hierarchie Komponente (Abbildung 4.3, A1). Der Domänenexperte betrachtet nun das aktualisierte Clustering-Resultat, welches seine angepasste Gewichtung reflektiert.

## 4.2 Implementierung der interaktiven Exploration

Diese Abschnitt befasst sich mit der Implementierung des interaktiven Exploration-Schrittes (Abbildung 3.2) aus dem eingeführten Prozessmodell zur Unterstützung von Domänenexperten bei der interaktiven Verfeinerung von Clustering-Resultaten (Abschnitt 3.1). Dem Prozessmodell folgend wird die Beschreibung der Implementierung des interaktiven Exploration-Schrittes in die Implementierung der hierarchischen Visualisierung des Clustering-Resultates (Abschnitt 3.2.1) und die Implementierung der Visualisierung des vom Domänenexperten selektierten Clusters (Cluster-Visualisierung Abschnitt 3.2.2) unterteilt.

Die Grundlage für die implementierten Elemente zur Kommunikation der Clustering-Resultate stellt der an der Universität Stuttgart entwickelte Clustering Communicator (Abbildung 4.1), welcher die Beschreibung einzelner Cluster (beispielsweise Abbildung 4.1, C) anhand der Identifikation wichtiger Attribute (B) von Behringer et al. [BHTM22] sowie die klassische Visualisierung des Clustering-Resultates (mithilfe von Histogrammen und Parallelen Koordinaten, A) in nicht hierarchischen Anwendungsfällen unterstützt, dar.

In den folgenden Abschnitten werden die implementierten Elemente der beiden Bestandteile des Exploration-Schrittes dargestellt und näher erläutert.

### 4.2.1 Hierarchie Visualisierung

Die Implementierung der Hierarchie-Visualisierung stellt die Cluster-Hierarchie als Baumstruktur mit knotenweiser Beschreibung des Knoteninhaltes und der weiteren Aufteilung dar. Die Integration der hierarchischen Visualisierung in die prototypische Implementierung ist in Abbildung 4.2 (A), zu sehen. Der Domänenexperte kann einzelne Knoten der Hierarchie selektieren, um diese (mithilfe der Cluster-Visualisierung) genauer zu betrachten. Das vom Domänenexperten selektierte Cluster

wird in der Implementierung grün hervorgehoben. Ein Beispiel für die Hervorhebung eines vom Domänenexperten selektierten Clusters zeigt Abbildung 4.2, in welcher der Domänenexperte das Cluster „freizeit_enthusiasten“ (C) für die nähere Betrachtung selektiert.

### Hierarchie-Darstellung

Die implementierte Umsetzung der Cluster-Hierarchie wird in Abbildung 4.4 dargestellt. Die Implementierung vermittelt dem Domänenexperten die schrittweise hierarchische Aufteilung der Daten ausgehend von einem Root-Knoten (B) und visualisiert die einzelnen Cluster der Hierarchie als Knoten mit Verbindungslinie zum übergeordneten Cluster-Knoten, welcher als Datengrundlage für die Clustering basierte Aufteilung der Daten in die nächste Ebene dient. Der Domänenexperte kann die Daten innerhalb einzelner Knoten beliebig unterteilen. Ein Beispiel stellt Abbildung 4.4 (E), in welcher das aus der Aufteilung des „root“ Clusters (B) resultierende Cluster „cluster_1“ durch erneutes Clustering in die beiden Cluster „cluster_1.1“ und „cluster_1.2“ aufgeteilt wird, dar.

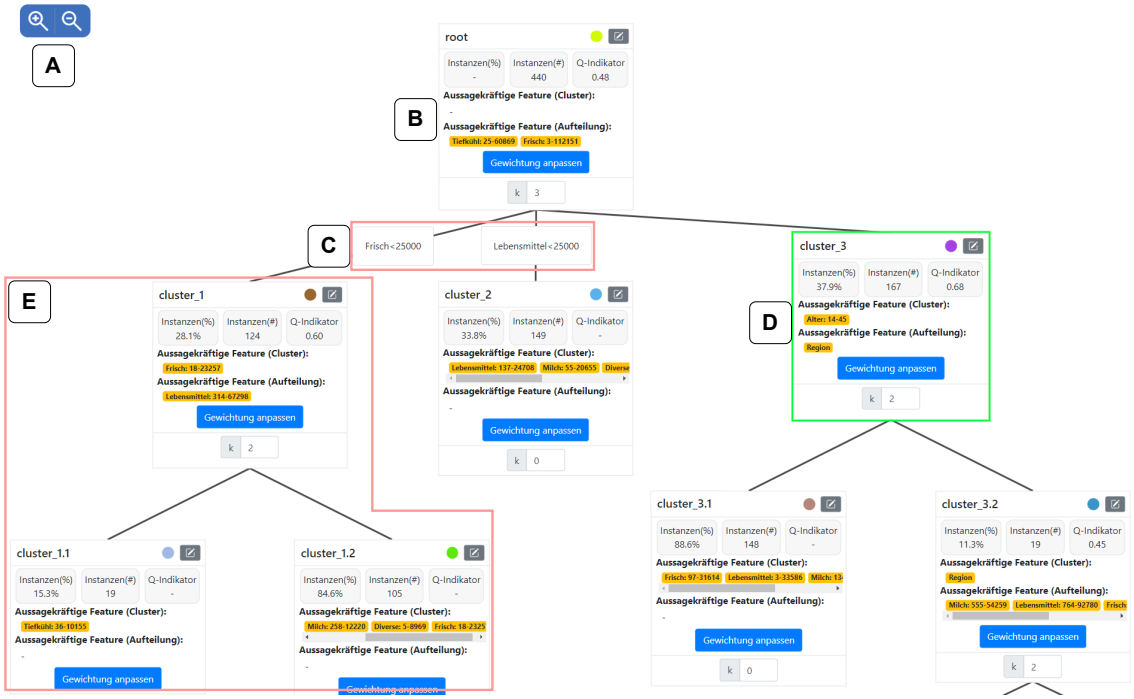



**Abbildung 4.4:** Übersicht über die implementierte Darstellung der Cluster-Hierarchie mit Root-Knoten (B), der Demonstration der weiteren Aufteilung einzelner Cluster (E) und der Repräsentation von definierten Regeln (C). Die Übersicht zeigt außerdem die implementierte Zoom-Option (A) sowie die Verschiebung von Knoten (D).

Die implementierte Darstellung der Hierarchie stellt außerdem dem eingeführten Konzept folgend die vom Domänenexperten definierten Regeln zur Eingrenzung der Wertebereiche einzelner Cluster dar. Für diese Darstellung werden die auf Clustern definierten Regeln (analog zum eingeführten Konzept) auf den Verbindungslinien zu den Elternknoten dargestellt. Eine Demonstration dieser Darstellung zeigt Abbildung 4.4 (C), in welcher der Domänenexperte den Wertebereich der Attribute „Frisch“ und „Lebensmittel“ einschränkt.

Zusätzlich kann der Domänenexperte die Darstellung der Hierarchie anpassen. Dem Domänenexperten steht für die Anpassung zum einen eine Zoom-Option (Abbildung 4.4, A) zur Verfügung, mit der er die dargestellte Hierarchie vergrößern oder verkleinern kann, um den Überblick über das Cluster-Modell zu behalten. Zum anderen kann der Domänenexperte die Organisation der Hierarchie-Knoten nach dem Drag-and-drop Prinzip anpassen. Ein Beispiel für die manuelle Organisation von Knoten mithilfe der Verschiebung einzelner Knoten zeigt der Knoten „cluster_3“ (D), welcher durch den Domänenexperten relativ zu den anderen Knoten derselben Hierarchie-Ebene nach oben-rechts verschoben wurde. Bei der Organisation einzelner Knoten werden vom System automatisch die Kindknoten des verschobenen Knotens (für das Beispiel D die beiden Knoten „cluster_3.1“ und „cluster_3.2“) relativ zum Elternknoten verschoben, um die konsistente Darstellung der untergeordneten Hierarchie zu gewährleisten.

#### Hierarchie-Knoten

Die einzelnen Knoten der implementierten hierarchischen Visualisierung verwenden mehrere der in Abschnitt 3.2.1 diskutierten Elemente, um den Inhalt des Knotens sowie die weitere Aufteilung der Instanzen, mit dem Domänenexperten zu kommunizieren. Der Aufbau eines Knotens der implementierten hierarchischen Visualisierung wird von Abbildung 4.5 demonstriert.

Die Knoten beinhalten einen vom Domänenexperten anpassbaren Namen mit zugehöriger (ebenfalls anpassbarer) Farbe (Abbildung 4.5, A) zur Organisation und Identifikation der Knoten. Der Inhalt eines Knotens wird durch die Verteilung der Instanzen und die Identifikation wichtiger Attribute von Behringer et al. [BHTM22] beschrieben. Die Implementierung verwendet hierfür zum einen die relative Anzahl an Instanzen, die dem Kindknoten aus dem Elternknoten zugewiesen werden (Abbildung 4.5, B) und zum anderen die absolute Anzahl an Instanzen (C) eines Knotens. Die Identifikation wichtiger Attribute wird von der Implementierung (Abbildung 4.5, E) verwendet, um den Inhalt eines Clusters mithilfe von innerhalb eines Clusters besonders wichtigen Attributen zu beschreiben. Zusätzlich werden für die Aufteilung der Daten besonders wichtige Attribute verwendet, welche die resultierenden Cluster voneinander abgrenzen [BHTM22], um die weitere Unterteilung eines Cluster-Knotens zu beschreiben. Für die identifizierten wichtigen Attribute (E) wird dem Domänenexperten außerdem der Wertebereich der Attribute angegeben, um das Verständnis des Domänenexperten zu fördern und eine einfache Definition von Regeln zur Anpassung des Wertebereiches wichtiger Attribute zu ermöglichen.

Abschließend erhält ein Knoten die aktuelle Konfiguration der Cluster-Parameter (Abbildung 4.5, F), welche zur weiteren Aufteilung genutzt werden. Aufgrund des auf k-Means basierten Clustering-Algorithmus besteht diese Konfiguration für den implementierten, angepassten PCKMeans-Algorithmus aus dem Parameter „k“ zur Festlegung der Cluster-Anzahl.

### 4.2.2 Cluster-Visualisierung

Die implementierte Cluster-Visualisierung ermöglicht es Domänenexperten, dem eingeführten Konzept (Abschnitt 3.2.2) entsprechend einzelne selektierte Cluster der Hierarchie detaillierter zu betrachten. Einen Überblick über den Prototyp mit implementierter Cluster-Visualisierung zeigt Abbildung 4.2 (B), mit der detaillierten Darstellung des selektierten Clusters (C) durch eine tabellarische Ansicht und Parallele Koordinaten. Die einzelnen Elemente der implementierten

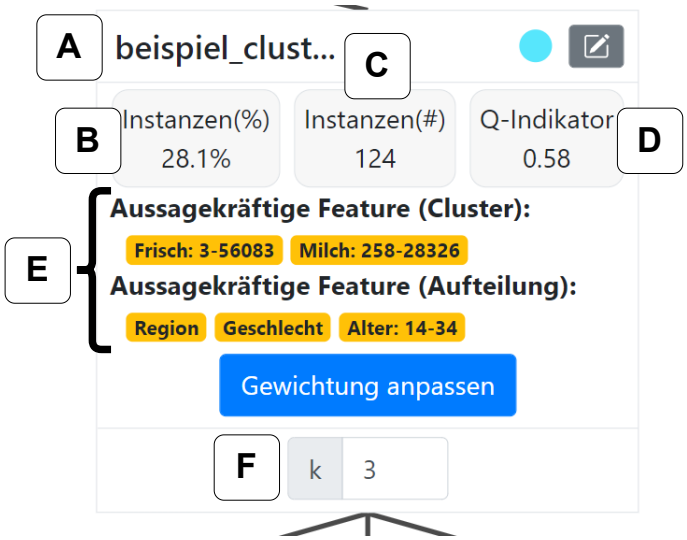



**Abbildung 4.5:** Aufbau eines Knotens der implementierten hierarchischen Visualisierung mit Knotenname (A), Beschreibung des Knoteninhaltes (B, C, E) und Beschreibung der weiteren Aufteilung des Knotens (D-F)

Cluster-Visualisierung werden von Abbildung 4.6 veranschaulicht. Die implementierten Elemente können in die Beschreibung des Clustering-Modells, die Verwendung klassischer Visualisierungen, die statistische Zusammenfassung und die tabellarische Darstellung der Daten aufgeteilt werden.

Die Beschreibung des Clustering-Modells (Abbildung 4.2, A) beinhaltet Informationen zur Hierarchie-Ebene des selektierten Clusters sowie mengenbezogene Informationen zur Aufteilung der Daten aus übergeordneten Ebenen. Die Güte der weiteren Aufteilung des selektierten Clusters wird außerdem mithilfe von Qualitäts-Indikatoren umschrieben. Für die Umschreibung verwendet die Implementierung, die in der populären SKlearn[1] Bibliothek inkludierten (Silhouettenkoeffizient, Calinski-Harabasz, Davies-Bouldin) Qualitätsindikatoren für Clustering-Verfahren.

[1] https://scikit-learn.org/stable/

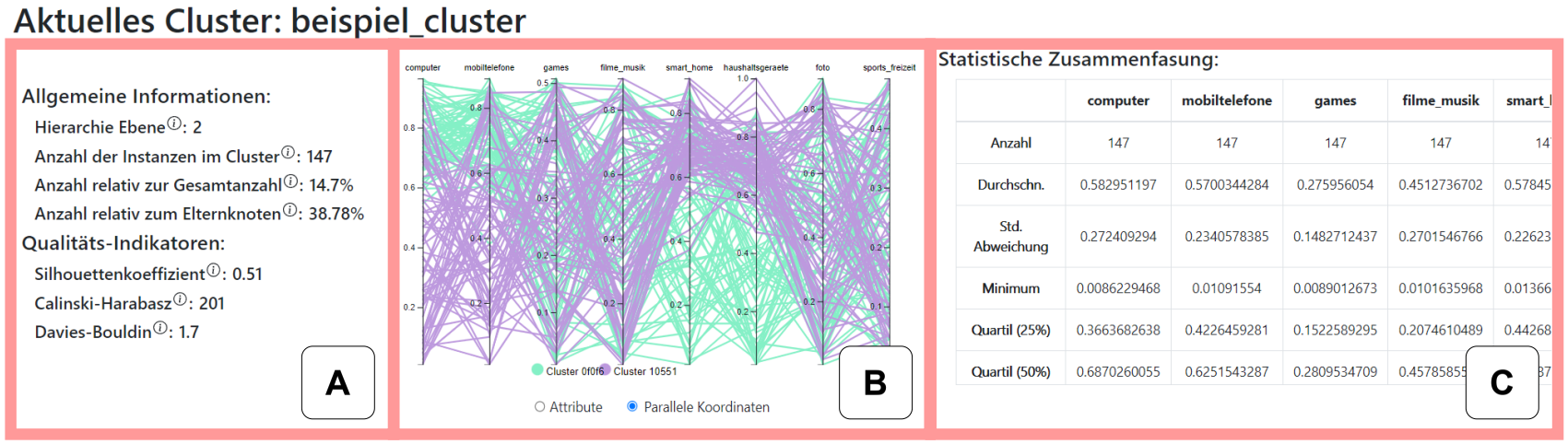
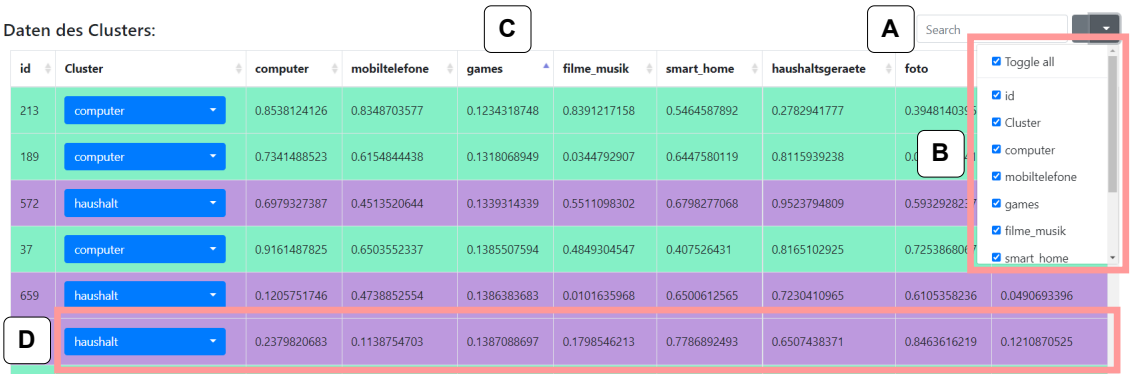



**Abbildung 4.6:** Detaillierter Überblick über die einzelnen Elemente der Cluster-Visualisierung des implementierten Prototyps (Abbildung 4.2, B) mit der Visualisierung von Cluster-Informationen und Qualitätsindikatoren zur Beschreibung der weiteren Aufteilung der Daten (A), klassischen Visualisierungen des Clustering-Resultates (B) sowie eine statistische Zusammenfassung des Cluster-Inhaltes (C).

Die verwendeten klassischen Visualisierungen (Abbildung 4.2, B) wurden von dem existierenden Clustering Communicator (Abbildung 4.1), welcher die Basis für die Implementierung stellt, übernommen. Die klassischen Visualisierungen beinhalten Parallele Koordinaten und Histogramme zum Vergleich einzelner Attribute zwischen den Kindknoten des selektierten Clusters.

Neben der Umschreibung des Clustering-Modells und der Aufteilung der Daten enthält die Implementierung eine statistische Zusammenfassung des selektierten Clusters (Abbildung 4.2, C), welche den Cluster-Inhalt beschreibt. Diese statistische Zusammenfassung verwendet deskriptive Statistiken (Std. Abweichung, Mean, ...), um die Lage und Streuung der Werte innerhalb des selektierten Clusters zu beschreiben.




| id | Cluster | computer | mobiltelefone | games | filme_musik | smart_home | haushaltsgeraete | foto |
|---|---|---|---|---|---|---|---|---|
| 213 | computer | 0.8538124126 | 0.8348703577 | 0.1234318748 | 0.8391217158 | 0.5464587892 | 0.2782941777 | 0.394814039 |
| 189 | computer | 0.7341488523 | 0.6154844438 | 0.1318068949 | 0.0344792907 | 0.6447580119 | 0.8115939238 | 0.0 |
| 572 | haushalt | 0.6979327387 | 0.4513520644 | 0.1339314339 | 0.5511098302 | 0.6798277068 | 0.9523794809 | 0.59329282 |
| 37 | computer | 0.9161487825 | 0.6503552337 | 0.1385507594 | 0.4849304547 | 0.407526431 | 0.8165102925 | 0.725386800 |
| 659 | haushalt | 0.1205751746 | 0.4738852554 | 0.1386383683 | 0.0101635968 | 0.6500612565 | 0.7230410965 | 0.6105358236 | 0.0490693396 |
| | haushalt | 0.2379820683 | 0.1138754703 | 0.1387088697 | 0.1798546213 | 0.7786892493 | 0.6507438371 | 0.8463616219 | 0.1210870525 |

**Abbildung 4.7:** Tabellarische Darstellung des Cluster-Inhaltes mit der Sortierung von Attributen (C), Suchfunktion (A), Ein-/Ausblendung von Attributen (B) und der manuellen Zuweisung von Instanzen (D).

Zusätzlich enthält die Cluster-Visualisierung eine tabellarische Ansicht (Abbildung 4.7) der im selektierten Cluster enthaltenen Instanzen. Die tabellarische Darstellung der Daten enthält außerdem die weitere Aufteilung des selektierten Clusters durch die farbige Kodierung der Zeilen und das zugewiesene Label der Instanz als extra Spalte. Das verwendete Label sowie die zur Kodierung verwendete Farbe entsprechen dem jeweils korrespondierenden Knotenname/-farbe (Abschnitt 4.2.1) aus der hierarchischen Visualisierung.
Der Domänenexperte kann zusätzlich mit der tabellarischen Ansicht interagieren, um gezielte Abschnitte der Daten zu betrachten. Ihm stehen hierfür Optionen zur Sortierung der Attribute (Abbildung 4.7, C) und der Suche nach bestimmten Werten (A) sowie dem Ein-/Ausblenden einzelner Attribute (B) zur Verfügung. Für die Umsetzung der tabellarischen Ansicht wird das Bootstrap-Table[2] Modul verwendet.

## 4.3 Implementierte Interaktionsmöglichkeiten

Die Implementierung stellt dem Domänenexperten neben der Visualisierung des Clustering-Resultates (Abschnitt 4.2) eine Teilmenge der im Konzept vorgestellten interaktiven Möglichkeiten (Abschnitt 3.3) zur Verfeinerung des Clustering-Resultates zur Verfügung. Die Anwendung von

[2] https://bootstrap-table.com/

Interaktionsmöglichkeiten richtet sich nach dem eingeführten Prozessmodell (Abschnitt 3.1). Der Domänenexperte erstellt daher zuerst in der Initialisierungsphase eine initiale Hierarchie und wählt den zu analysierenden Datensatz und den gewünschten Clustering-Algorithmus aus.


**Abbildung 4.8:** Abschluss der Initialisierungsphase mit der Auswahl des zu analysierenden Datensatzes (A) sowie des Clustering-Algorithmus (B).

Bei der Erstellung der initialen Hierarchie kann der Domänenexperte in der prototypischen Implementierung die Parameter „k“ einzelner Knoten festlegen. Anschließend muss der Domänenexperte einen Datensatz sowie den gewünschten Clustering-Algorithmus auswählen. Der Prototyp implementiert diese Auswahl mithilfe der in Abbildung 4.8 abgebildeten Menüoption. Der Domänenexperte kann in diesem Menü einen Datensatz aus der Sammlung der existierenden Datenquellen auswählen (A) und anschließend den gewünschten Clustering-Algorithmus (B) auswählen. Die Auswahl des Clustering-Algorithmus beschränkt sich allerdings auf eine (für den Einsatz im Prototyp) adaptierte Version des PCKMeans Algorithmus. Die Erweiterung des Prototyps durch die Implementierung zusätzliche Algorithmen ist jedoch möglich.

Nach der Initialisierung des Clustering-Verfahrens stehen dem Domänenexperten Interaktionsmöglichkeiten zur weiteren Verfeinerung des Clustering-Resultates zur Verfügung. Die Teilmenge der im Prototyp implementierten Interaktionsmöglichkeiten deckt alle drei der grundlegenden Interaktionskategorien ab. Die implementierten Interaktionsmöglichkeiten der verschiedenen Interaktionskategorien werden in den folgenden Abschnitten vorgestellt.

### 4.3.1 Anpassung von Parametern

Die prototypische Implementierung unterstützt die Anpassung der Clustering-Parameter einzelner Knoten durch die direkte Anpassung der Parameter des verwendeten Clustering-Algorithmus und die Anpassung der vom Clustering-Algorithmus verwendeten Gewichtung.

Für die Anwendung der Interaktionsmöglichkeiten kann der Domänenexperte direkt mit dem gewünschten Zielknoten der Cluster-Hierarchie interagieren. Die möglichen Interaktionen mit einem Knoten zeigt Abbildung 4.9. Der Domänenexperte kann diese Interaktionsmöglichkeiten (B) nutzen, um den Clustering-Parameter („k“) des PCKMeans Algorithmus direkt anzupassen.

Für die Gewichtung einzelner Attribute kann der Domänenexperte die Interaktion mit dem Knoten (A) verwenden, um einen Gewichtungs-Dialog zu öffnen, welcher die aktuellen Gewichte und die Möglichkeit zur Anpassung der Gewichtung enthält. Der implementierte Dialog zur Anpassung der Gewichtung (Abbildung 4.10) kombiniert dem Konzept entsprechend die Gewichtung einzelner Attribute () mit der Identifikation wichtiger Feature von Behringer et al.[BHTM22], um die Auswirkung der Gewichtung auf die Aufteilung der Daten darzustellen. Der implementierte Dialog (Abbildung 4.10) für die Gewichtung der Attribute sortiert die Attribute hierfür nach ihrer relativen Wichtigkeit (beginnend von Abbildung 4.10, A absteigend) und hebt besonders wichtige Attribute

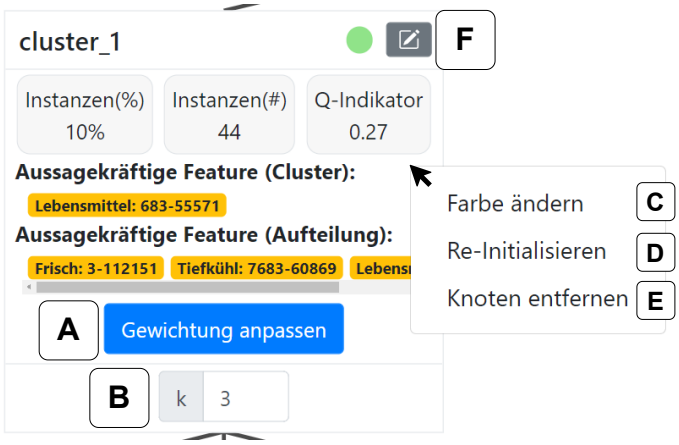



**Abbildung 4.9:** Interaktionsmöglichkeiten eines Cluster-Knotens des Prototyps, mit Möglichkeiten zur Anpassung der Parameter (A, B) den Optionen des Kontextmenüs (C-E) und der Möglichkeit die Benennung des Knotens anzupassen (F).

grün hervor. Der Domänenexperte kann den Wert der Gewichtung einzelner Attribute anpassen (beispielsweise die Festlegung des Gewichts des Attributes „Milch“ auf 0 in B), um deren Einfluss auf die Aufteilung der Daten zu regulieren. Durch die identifizierte relative Wichtigkeit kann die Auswirkung seiner angepassten Gewichtung direkt anhand der veränderten relativen Wichtigkeit der Attribute nachvollziehen. Die Erwartung ist, dass Attribute mit reduzierter Gewichtung in der Reihe der nach Einfluss auf die Aufteilung der Daten sortierten Attribute nach hinten verschoben werden (umgekehrt für erhöhte Werte) und der Domänenexperte so den Einfluss der Attribute durch die Verschiebung der relativen Wichtigkeiten nachvollziehen kann.

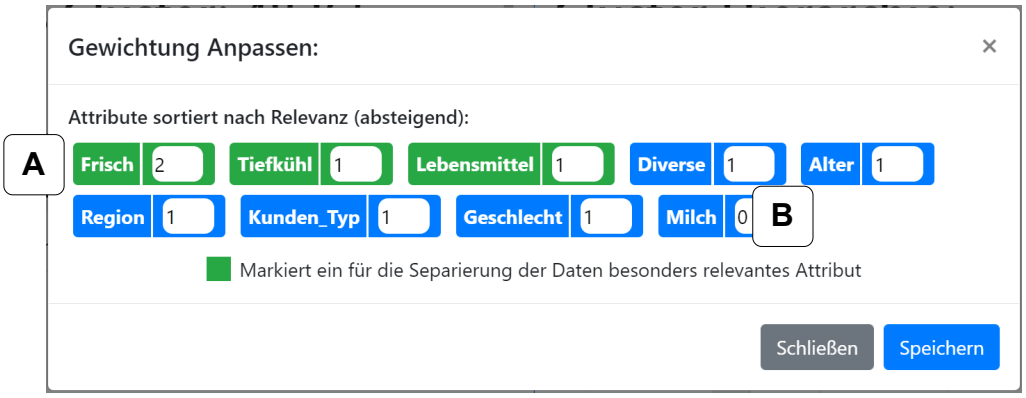



**Abbildung 4.10:** Dialog zur Anpassung der Gewichtung sortiert nach identifizierter Wichtigkeit der Attribute (von A absteigend) und der möglichen Anpassung der Gewichtung (B). Die als besonders wichtig identifizieren Attribute werden grün hervorgehoben.

### 4.3.2 Interaktion mit Clustering-Resultat

Aus der Interaktions-Kategorie der Interaktion mit dem Clustering-Resultat implementiert der Prototyp die Restriktion von Wertebereichen einzelner Cluster, die manuelle Zuweisung einzelner Instanzen sowie Operationen zur Manipulation von Knoten der modellierten Hierarchie aus den im Konzept vorgestellten (Abschnitt 3.3.2) Interaktionsmöglichkeiten.

Der Domänenexperte kann mithilfe dieser Interaktionsmöglichkeiten direkte Änderungen an den von der prototypischen Implementierung berechneten Clustering-Resultaten vornehmen. In den folgenden Abschnitten wird die Implementierung der einzelnen Interaktionsmöglichkeiten dieser Kategorie erläutert.

#### Werte-Restriktionen

Der Domänenexperte kann in der prototypischen Implementierung den im Konzept eingeführten Interaktionsmöglichkeiten (Abschnitt 3.3) folgend, die Wertebereiche für Instanzen einzelner Knoten der Cluster-Hierarchie einschränken. Für die Einschränkung der Wertebereiche kann der Domänenexperte in der prototypischen Implementierung mit der Cluster-Hierarchie (Abschnitt 4.2.1) interagieren.

Der Ablauf der Interaktion zur Erstellung von Regeln für die Beschränkung des Wertebereiches wird von Abbildung 4.11 veranschaulicht. Die Restriktion des Wertebereiches eines Clusters wird vom Domänenexperten durch die Öffnung eines Kontextmenüs (A) auf der mit dem Cluster in der Cluster-Hierarchie korrespondierenden Verbindungslinie eingeleitet. Der Domänenexperte kann darauffolgend mithilfe eines Dialoges die Werte einzelner Attribute einschränken.

In der Abbildung 4.11 (B) nutzt der Domänenexperte den Dialog (B) exemplarisch zur Einschränkung der Attribute (B1) „Geschlecht“, „Alter“ und „Kunden_Typ“. Dem Domänenexperten stehen für die Einschränkung der Attribute verschiedene Operatoren (B2) zum Vergleich (größer, kleiner, gleich, ungleich) des Attributwertes einzelner Instanzen mit einem vom Domänenexperte gestellten Vergleichswert (B3) zur Verfügung. Der Domänenexperte kann außerdem jederzeit bestehende Restriktionen entfernen (B4) sowie weitere Restriktionen hinzufügen (C), um den Wertebereich weiter anzupassen.

Abschließend werden die vom Domänenexperten erstellten Restriktionen gespeichert und bei der Berechnung des Clustering-Resultates berücksichtigt. Die angewendeten Restriktionen werden dem Domänenexperten in der Visualisierung der Cluster-Hierarchie auf der Verbindungslinie des eingeschränkten Clusters (D) dargestellt. Eine exemplarische Übersicht über eine Cluster-Hierarchie mit angewandten Werte-Restriktionen wird in Abbildung 4.4 (C) gezeigt.

#### Zuweisung von Instanzen

Der implementierte Prototyp ermöglicht es dem Domänenexperten, einzelne Instanzen eines Knotens manuell einem Unterknoten (Cluster) fest zuzuweisen. Die Zuweisung erfolgt über die tabellarische Darstellung des Cluster-Inhaltes der implementierten Cluster-Visualisierung (Abschnitt 4.2.2). Für

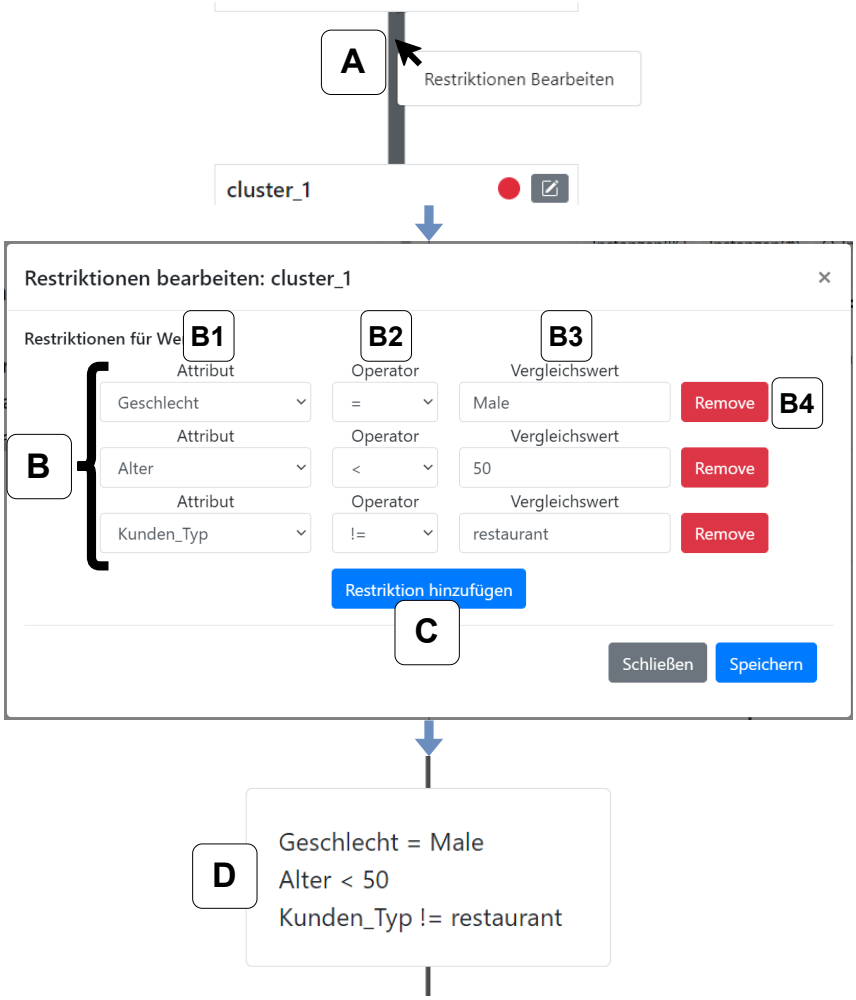



**Abbildung 4.11:** Exemplarische Anwendung von Werte-Restriktionen in der prototypischen Implementierung beginnend mit dem Öffnen eines Kontextmenüs (A), der darauf folgenden Erstellung von Regeln (B, mit Optionen B1-4) sowie der abschließenden Visualisierung der erstellten Regel auf der Verbindungslinie der Cluster-Hierarchie (D).

die manuelle Zuweisung einer Instanz verwendet der Domänenexperte das in die tabellarische Darstellung integrierte Dropdown-Menü (Abbildung 4.7), um einen Unterknoten (Cluster) des selektierten Knotens auszuwählen, zu welchem er die Instanz zuweisen möchte.

Das Python-Backend berechnet daraufhin das Clustering-Resultat neu und aktualisiert die Centroide des PCKMeans Algorithmus basierend auf der zugewiesenen Instanz.

#### Manipulation von Hierarchie-Knoten

Der Prototyp beinhaltet Interaktionen, welche direkt auf einzelne Knoten der Cluster-Hierarchie angewandt werden können, um das Clustering-Resultat zu verfeinern. Die Anwendung dieser Interaktionen erfolgt über die Verwendung des in die hierarchische Visualisierung integrierten Kontextmenüs (Abbildung 4.9) der einzelnen Hierarchie-Knoten.

Der Domänenexperte kann dieses Kontextmenü nutzen, um das Clustering-Verfahren neu zu initialisieren (die Centroide neu zu verteilen) und ein alternatives Clustering-Resultat zu erstellen. Außerdem kann der Domänenexperte einen spezifischen Knoten der Cluster-Hierarchie entfernen (E).

Zusätzlich wird dem Domänenexperten eine kosmetische Operation zur Änderung der mit einem Knoten assoziierten Farbe (C) gestellt.

### 4.3.3 System-initiiert

Der System initiierte Teil der Interaktionen mit dem Domänenexperten wird von der prototypischen Implementierung gemäß dem erarbeiteten Konzept (Abschnitt 3.3) durch die Verwendung von System generierten Vorschlagen für den aktuell selektierten Knoten umgesetzt. Die Darstellung der System initiierten Vorschläge wird vom Prototyp in die Cluster-Visualisierung integriert (Abbildung 4.12, A) und zählt zusätzlich zu den vorhandenen Visualisierungen die verfügbaren (A1, A2) Verbesserungsvorschläge auf.

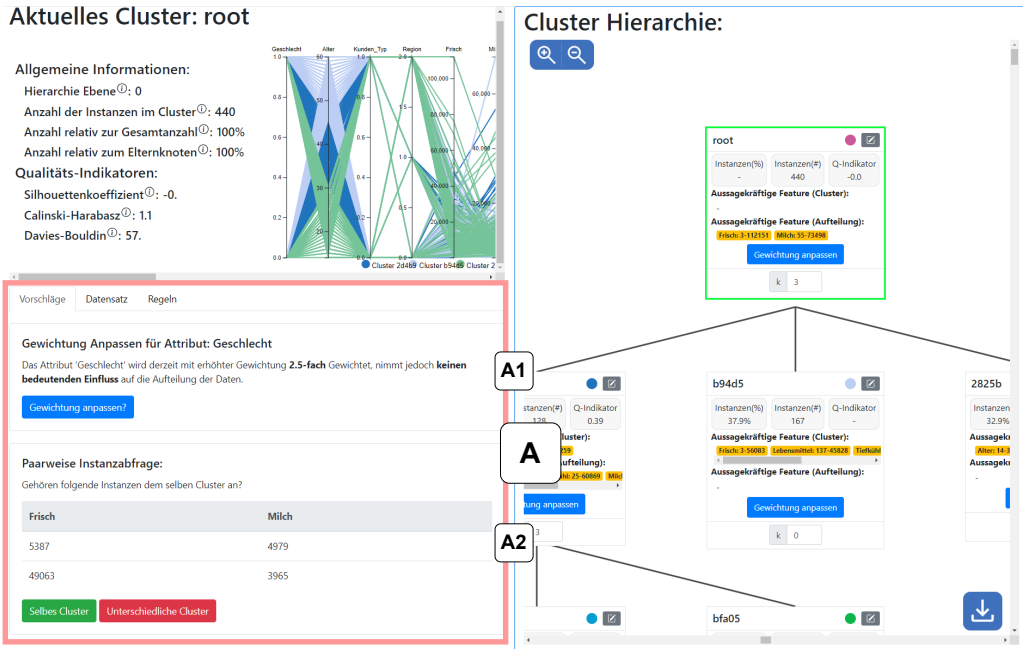



**Abbildung 4.12:** Übersicht über die Einbindung von Systemvorschlägen in die prototypische Implementierung (A) mit den für das selektierte Cluster „root“ generierten Vorschlägen (A1, A2).

Die aus dem Konzept (Abschnitt 3.3) implementierten Typen an Vorschlägen setzen sich zum einen aus der paarweisen Abfrage einzelner Instanzen unter der Verwendung von Active-Clustering zur Erstellung von Must-/Cannot-Link Constraints und zum anderen aus der auf der Identifikation wichtiger Attribute basierten Abänderung von Gewichtungen zusammen. Die einzelnen implementierten Systemvorschläge werden in den folgenden Abschnitten erläutert.

### Active-Clustering

Für die Verwendung von Active-Clustering fragt die prototypische Implementierung den Domänenexperten mit Vorschlägen (Abbildung 4.13) zum paarweisen Verhältnis von Instanzen des selektierten Clusters ab. Den Vergleich der abgefragten Instanzen kann der Domänenexperte anhand der Attributwerte der abgefragten Instanzen vornehmen. Dem eingeführten Konzept folgend wird der Domänenexperte bei dem Vergleich der einzelnen Attributwerte durch die Identifikation wichtiger Attribute unterstützt.

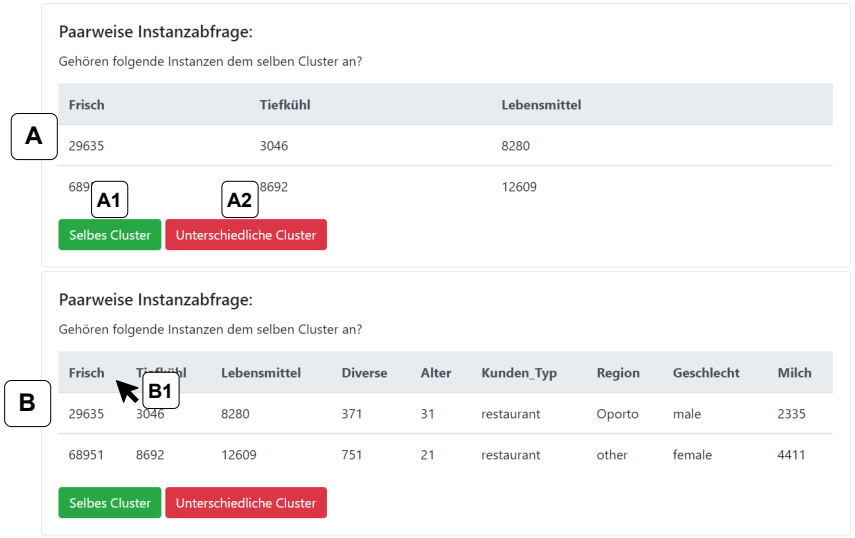



**Abbildung 4.13:** Umsetzung der Active-Clustering Abfrage von paarweisen Instanzen eines Knotens mit dem Vergleich der Instanzen auf Basis der identifizierten besonders wichtigen Attribute (A) sowie der Interaktion (B1) für den umfassenden Vergleich aller Attributwerte (B) und die Beantwortung der Abfrage (A1, A2).

Der Domänenexperte erhält aufgrund dieser Unterstützung initial nur die Attributwerte der für die Aufteilung der Daten besonders wichtigen Attribute angezeigt (Abbildung 4.13, A), um den Vergleich der beiden Instanzen (durch Komplexitätsreduktion der zu vergleichenden Daten) zu vereinfachen. Die Darstellung der zum Vergleich verwendeten Attributwerte kann vom Domänenexperten auf Wunsch erweitert werden (B), um alle Attribute der Daten (und nicht nur die als besonders relevant identifizierten) einzuschließen. Der Domänenexperte kann hierzu mit der Maus über die angezeigten Werte hovern (B1), um eine Übersicht über alle Attribute der Daten zu erhalten.

Abschließend kann der Domänenexperte nach dem Vergleich der Attributwerte der abgefragten Instanzen die Abfrage durch die Betätigung der zugehörigen Schaltflächen „Selbes Cluster“ (A1) oder „Unterschiedliches Cluster“ (Abbildung 4.13, A2) abhängig von der (vom Domänenexper-

ten) ermittelten Zugehörigkeit der Instanzen beantworten. Basierend auf den Antworten werden vom Python-Backend (Abschnitt 4.1) die entsprechenden Must-/Cannot-Link Constraints zur Verfeinerung des Clustering-Resultates erstellt.

#### Anpassung der Gewichtung

Die prototypische Implementierung enthält außerdem Vorschläge zur Anpassung der vom Domänenexperten festgelegten Gewichtung (dem Konzept Abschnitt 3.3.3 folgend) basierend auf den für die Aufteilung der Daten besonders wichtigen Attributen. Der Domänenexperte wird von der prototypischen Implementierung durch die vom System generierten Vorschläge auf Diskrepanzen zwischen der Intention seiner Gewichtung (den Einfluss von Attributen zu regulieren) und Clustering-Resultaten, welche den regulierten Einfluss des Attributes nicht reflektierten, hingewiesen. Diese Hinweise helfen dem Domänenexperten, ineffektive Gewichtungen zu identifizieren und zu korrigieren.

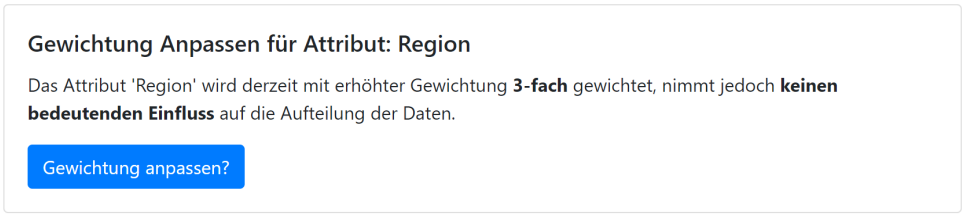



**Abbildung 4.14:** Vorschlag zur Anpassung der Gewichtung basierend auf identifizierten wichtigen Feature und der vom Domänenexperten festgelegten Gewichtung der Attribute

Die Implementierung der Vorschläge zur Anpassung der Gewichtung wird exemplarisch von Abbildung 4.14 gezeigt, in welcher der Domänenexperte das Attribut „Region“ stark erhöht (dreifach) gewichtet, das Clustering-Resultat durch diese Gewichtung allerdings nicht wie gewollt beeinflusst wird, da das Attribut „Region“ für das Clustering-Resultat (von der im Prototyp an mehreren Stellen verwendeten Identifikation wichtiger Attribute) nicht als wichtig für die Aufteilung der Daten identifiziert wird. Der Domänenexperte erhält vom Systemvorschlag des Prototyps daher die Option, die Gewichtung des Attributes durch eine Schaltfläche (Abbildung 4.14) anzupassen. Die Verwendung der Schaltfläche „Gewichtung anpassen„ ermöglicht es dem Domänenexperten mithilfe eines Dialoges, die Gewichtung des Attributes anzupassen.

## 4.4 Clustering-Algorithmus

Die prototypische Implementierung kann gemäß dem eingeführten Konzept mit verschiedenen unterliegenden Clustering-Algorithmen verwendet werden und enthält eine Option zur Auswahl des zu verwendeten Algorithmus während der Initialisierung (Abbildung 4.8, B) des Clustering-Verfahrens.

Im Rahmen der Masterarbeit wurde exemplarisch eine angepasste Version des PCKMeans [BBM04a] Clustering-Algorithmus in den Prototyp implementiert. Der PCKMeans Algorithmus wurde aufgrund seiner Popularität (in der Kategorie der Constraint basierten Clustering-Algorithmen), seiner aus dem Aufbau auf k-Means resultierenden Simplizität und der Existenz einer für den

Algorithmus vorgestellten Active-Learning Strategie zur „intelligenten“ Generation von Must-/Cannot-Link Constraints ausgewählt. Die Implementierung verwendet die vorhandene Python Implementierung des PCKMeans Algorithmus und der korrespondierenden Active-Learning Strategie von Jakub Švehla [Šve18], welche für den Einsatz in der Scikit-learn Bibliothek entwickelt wurde und passt diese an das im Konzept diskutierte interaktive hierarchische Clustering-Szenario an.

Der folgende Abschnitt bespricht die Implementierung des Algorithmus und die zur Integration in das interaktive Konzept der Arbeit vorgenommenen Anpassungen.

### 4.4.1 Implementierung des PCKMeans Algorithmus

Die Implementierung des (angepassten) PCKMeans Algorithmus besteht aus einer Initialisierungsphase (Algorithmus 4.1), in welcher (ähnlich zum klassischen k-Means Algorithmus) initiale Centroide gefunden werden sowie einer Update-Phase (Algorithmus 4.2) zur Berechnung der (lokal) optimalen Cluster-Zentren. Eine vereinfachte Version der angepassten Algorithmen wird anhand von Algorithmus 4.1 und Algorithmus 4.2 gezeigt.

#### Initialisierungsphase

Die Initialisierung des Clustering-Algorithmus (Algorithmus 4.1) bekommt die Anzahl der Cluster (k), die Daten (X) und die bereits aus vorherigen Iterationen der iterativen Verfeinerung existierenden Cluster übergeben (Zeile 1) und berechnet auf Basis dieser Informationen k Cluster-Zentren.

**Algorithmus 4.1** Initialisierung (KMeans++)

```
 1: function InitialisiereCluster(k, X, existingClusters)
 2:     centroids ← []
 3:     // Verwendung von Cluster-Zentren aus vorherigen Iteration
 4:     for all cluster ∈ existingClusters do
 5:         temp ← GetCentroid(cluster)
 6:         centroids.add(temp)
 7:     end for
 8:     // Wähle restliche Cluster mit KMeans++
 9:     while centroids.length < k do
10:         temp ← GetCentroidKmeansPlusPlus(centroids, X)
11:         centroids.add(temp)
12:     end while
13:     return centroids
14: end function
```

Die Initialisierungsphase (Findung der Cluster-Zentren) des originalen PCKMeans Algorithmus wurde bei der Implementierung komplett durch die Initialisierung des populären KMeans++ [AV07] (eine Verbesserung zur originalen KMeans Initialisierung) Algorithmus (Algorithmus 4.1, Zeilen 8-13) ausgetauscht. Dieser Austausch wurde vorgenommen, da der originale PCKMeans Algorithmus die Cluster-Zentren auf Basis vorhandener Must-/Cannot-Link Costraints auswählt. Diese Auswahl der Cluster-Zentren geht allerdings von der Existenz von bereits vorhandener Constraints aus,

welche bei dem interaktiven Szenario des Prototyps nicht gegeben ist, da alle Constraints vom Domänenexperten erst nach erstmaliger Berechnung des Resultates durch Interaktionen erstellt werden. Für die Anwendung in der prototypischen Implementierung ist die originale Initialisierung des PCKMeans daher nicht geeignet. Die erprobte KMeans++ Initialisierung bietet aufgrund ihrer Constraint unabhängigen Berechnung von Cluster-Zentren für den KMeans Algorithmus eine geeignete Alternative für die Implementierung in den Prototyp.

Zusätzlich erweitert die Implementierung die existierende Initialisierung durch die Wiederverwendung (aus vorherigen Iterationen der iterativen Verfeinerung) vorhandener Cluster-Zentren (Zeilen 3-7), um die Bedeutung der einzelnen Hierarchie-Knoten bei der Neu-Anwendung des Cluster-Verfahrens beizubehalten. Das Resultat ist, dass bei einer Erhöhung der Cluster-Anzahl (durch den Domänenexperten) die vorhandenen Cluster-Zentren wiederverwendet werden und nur die (aus der Änderung) neu hinzukommenden Cluster-Zentren mithilfe der KMeans++ Initialisierung berechnet werden. Diese Wiederverwendung von Cluster-Zentren wird von dem nachfolgenden Beispiel veranschaulicht:

*Beispiel-Szenario:* Der Marketingexperte unterteilt die Daten initial in drei Cluster. Die (drei) benötigten Cluster-Zentren werden zuerst mithilfe des KMeans++ Verfahrens bestimmt. Er betrachtet den Inhalt der resultierenden Cluster und beschließt daraufhin eine Aufteilung in vier (anstatt der vorherigen drei) Cluster. Der Algorithmus verwendet die bereits berechneten Cluster-Zentren der vorherigen Iteration erneut und verwendet die KMeans++ Initialisierung, um auf Basis der drei vorhandenen Zentren ein viertes Cluster-Zentrum zu errechnen.

#### Cluster-Berechnung

Die Berechnung des Clustering-Resultates erfolgt nach Algorithmus 4.2 und folgt der Berechnung des Clustering-Resultates des PCKMeans Algorithmus, mit Erweiterungen zur Umsetzung der interaktiven Maßnahmen der Werterestriktionen, Instanz-Zuweisung und Gewichtung von Attributen.

Für die Berücksichtigung der Gewichtung einzelner Attribute wurde zuerst die Distanzfunktion des ursprünglichen PCKMeans Algorithmus (Algorithmus 4.2, Zeilen 20-24) um einen Parameter (gw, Zeile 19) für die Gewichtungen der Attribute erweitert. Die Berechnung des Abstandes (Zeile 20) multipliziert den Abstand zwischen den Attributwerten der Instanz (instance, Zeile 20) und dem Cluster-Zentrum (c, Zeile 20) mit den Gewichten (gw) der einzelnen Attribute. Die restliche Abstandsfunktion folgt der ursprünglichen PCKMeans Implementierung und berechnet „Bestrafungen“ basierend auf Must-/Cannot-Link Constraints (Zeile 22).

Die eigentliche Berechnung des Clustering-Resultates (Zuweisung der Label) wird von der CalcCluster Funktion (Zeile 1) durchgeführt. Der Clustering-Algorithmus beginnt analog zum ursprünglichen Algorithmus mit der Berechnung eines Graphen (Zeile 4) basierend auf den existierenden Constraints. Daraufhin werden die Cluster-Zentren und Label analog zum ursprünglichen Algorithmus iterativ aktualisiert (Zeilen 6-15). Die Implementierung berechnet während dieser Aktualisierungsdurchläufe allerdings ein Set an möglichen Cluster-Zentren (candCentroids, Zeile 9) basierend auf den vorhandenen vom Domänenexperten vorgenommenen Einschränkungen. Dieses Set entfernt Cluster, für welche die aktuell betrachtete Instanz eine Restriktion von Attributwerten (vrest) nicht erfüllt und reduziert die Anzahl an möglichen Clustern für Instanzen, welche vom Domänenexperten direkt einem Cluster zugewiesen wurden, auf ein einzelnes (das zugewiesene) Cluster.

**Algorithmus 4.2** Cluster-Berechnung (Angepasst von PCKMeans)

```
1:  function CalcClusterLabels(centroids, X, ml, cl, vrest, gw, zw)
2:      // X: Daten, ml/cl: Must-/Cannot-Link Constraints, vrest: Werte Restriktionen,
3:      // gw: Attribut Gewichtungen, zw: Zugeweisene Instanzen
4:      constN ← generateConstNeighborhoods(ml, cl)
5:      labels_{∀i∈X} ← −1
6:      while not converged do
7:          for all instance ∈ X do
8:              // Überprüfe welche Cluster in Frage kommen
9:              candCentroids ← GetCandidateClusters(instance, vrest, zw, centroids)
10:             // Berechne Cluster Zuweisung
11:             labels_instance ← argmin_{c∈candCentroids} Dist(instance, c, constN, labels, gw)
12:         end for
13:         // Berechne neue Cluster-Zentren
14:         centroids ← UpdateCentroids(X, labels)
15:     end while
16:     return labels, centroids
17: end function
18:
19: function Dist(instance, c, constN, labels, gw)
20:     distance ← 1/2 * ∑ ((instance − c) * gw)^2
21:     // Überprüfe Constraints
22:     penalty ← calcPenalties(labels, constN, c)
23:     distance ← distance + constraints
24:     return distance
25: end function
```

Die Zuweisung des Cluster-Labels während des Aktualisierungsschrittes (Zeile 11) weist der Instanz das beste Cluster aus der zuvor berechneten Menge an möglichen Clustern (Zeile 9 candCentroids) zu. Dies verhindert die Zuweisung von Instanzen zu Clustern, welche gegen Einschränkungen des Domänenexperten (Abschnitt 4.3.2) verstoßen.

Abschließend werden die Cluster-Zentren analog zum ursprünglichen (PCKMeans) Algorithmus am Ende jeder Iteration Aktualisierungsschrittes (Zeile 14) upgedatet. Die Schleife wird so lange wiederholt, bis der Algorithmus konvergiert oder eine Anzahl an maximalen Iterationen (nicht im vereinfachten Algorithmus abgebildet) erreicht ist.

# 5 Evaluation

Dieses Kapitel befasst sich mit der Evaluation des Prototyps (Kapitel 4) im Hinblick auf die gebotene Unterstützung für die Verfeinerung von Clustering-Resultaten. Die Auswertung verwendet mehrere Evaluations-Szenarien (Abschnitt 5.1.1), um verschiedene Vorgehensweisen von Domänenexperten bei der Auswertung der Daten gegenüberzustellen.

Im Folgenden wird der verwendete Versuchsaufbau dargestellt. Danach werden die Resultate der Evaluation diskutiert.

## 5.1 Versuchsaufbau

In der Praxis basiert die Clusteranalyse auf der langjährigen Erfahrung des Domänenexperten, die für die Erstellung des Clustering-Resultates ausschlaggebend ist (Kapitel 1). Dabei ergeben sich für den Domänenexperten mehrere Möglichkeiten, seine Erfahrungen in die Clusteranalyse einzubringen. Die Auswertung des Konzeptes verwendet basierend auf diesen Möglichkeiten Evalutions-Szenarien, die in echten Anwendungsfällen typischerweise durchlaufen werden. Diese Szenarien repräsentieren unterschiedliche Verhaltensweisen von Domänenexperten und werden einander gegenübergestellt. Ausgewertet werden diese Evaluations-Szenarien auf synthetisch generierten Datensätzen mit bekannten ground-truth Clustern.

Nachfolgend werden zuerst die verwendeten Evaluations-Szenarien vorgestellt. Anschließend wird der Prozess für die Generierung der synthetischen Datensätze erklärt. Abschließend wird die Vorgehensweise, welche für die Auswertung der Daten genutzt wird, beschrieben.

### 5.1.1 Evaluation-Szenarien

Die Evaluations-Szenarien setzen sich aus vier unterschiedlichen Szenarien zur Anwendung von Clusteranalysen auf einem Datensatz zusammen. Der Kerngedanke ist der Vergleich von (interaktiven) Möglichkeiten des Prototyps, welche dem Domänenexperten zur Auswertung der Daten zur Verfügung stehen.

Die verwendeten Evaluations-Szenarien werden im Einzelnen vorgestellt.

**Szenario 1: Baseline (k-Means++)** Der Domänenexperte wendet ein klassisches Clustering-Verfahren, den k-Means++ Clustering-Algorithmus ohne jegliche Interaktion (mit korrekter Cluster-Anzahl) auf den gegebenen Datensatz an. Dieses Szenario wird als Baseline für die weiteren (interaktiven) Szenarien, welche Interaktionen zur Verfeinerung des Clustering-Resultates anwenden, verwendet. Der k-Means++-Algorithmus baut ebenso wie der entwickelte Prototyp auf den populären k-Means-Algorithmus auf. Die erzeugten Clustering-Resultate bilden deshalb Richtwerte für die weiteren interaktiven Anwendungs-Szenarien.

**Szenario 2: Active-Clustering (PCKMeans)** Das zweite Szenario betrachtet einen Domänenexperten, der sein Wissen nicht auf eigene Initiative in die Clusteranalyse einbringt, sondern allein die paarweisen Instanzabfragen des Systems beantwortet. Das Szenario entspricht der Anwendung von Active-Clustering (Abschnitt 3.3.3) beziehungsweise der Beantwortung der Active-Clustering Abfragen des Prototyps (Abschnitt 4.3.3). Das Clustering-Resultat dieses Szenarios ist deshalb mit der Anwendung des PCKMeans-Algorithmus unter Verwendung von auf Basis der Abfragen erzeugten Constraints gleichzusetzen.

**Szenario 3: Restriktion des Wertebereiches (prototypische Implementierung)** Das dritte Szenario betrachtet einen Domänenexperten, der sein Domänenwissen eigenständig in den Clustering-Prozess einbringt. Hierzu nutzt der Domänenexperte sein Domänenwissen, um während des Clustering-Prozesses Regeln zur Einschränkung von Attributwerten zu erstellen. Der Effekt der einzelnen Regeln hängt vom Clustering-Resultat, das verfeinert wird, ab. Dies wirft die Frage auf, welche Auswirkung die Variation der Anwendungsreihenfolge der Regeln auf das Clustering-Resultat nimmt. Deshalb betrachtet dieses Szenario zusätzlich variierende Anwendungsreihenfolgen.

**Szenario 4: Hierarchiesche Anwendung (prototypische Implementierung)** Das vierte Szenario simuliert das Verhalten eines Domänenexperten, welcher die prototypische Implementierung zur Modellierung einer Cluster-Hierarchie verwendet. Anschließend an die Modellierung der Hierarchie werden erneut Regeln zur Einschränkung von Attributwerten angewendet und ausgewertet. Es wird daher in diesem Szenario nicht nur der Effekt der modellierten Hierarchie, sondern auch die Kombination der modellierten Hierarchie mit der Anwendung von Regeln ausgewertet.

### 5.1.2 Generation des synthetischen Datensatzes

Die generierten Daten (exemplarisch in Tabelle 5.1) setzen sich aus Kundendaten zusammen, welche durch die Anwendung einer Clusteranalyse in vier fiktive Kundengruppen (ground-truth Cluster) segmentiert werden. Als Kundengruppen werden Technik-Enthusiasten, Freizeit-Enthusiasten, Automatisierungs-Enthusiasten und Outdoor-Enthusiasten verwendet. Diese Kundengruppen werden durch die unterschiedlicher Ausprägungen der Daten simuliert. Hierfür bilden die Daten den fiktiven Umsatz von Kunden in acht verschiedenen Produktkategorien (Computer, Mobiltelefone, Games, Filme & Musik, Smart Home, Haushaltsgeräte, Foto und Sport & Freizeit) ab. Die Gruppierbarkeit der Daten in diese Kundengruppen wird durch die Verwendung von unterschiedlichen Verteilungen (abhängig von der Kundengruppe) während der Generierung der Daten realisiert.

Die einzelnen Kundengruppen interessieren sich jeweils für zwei sich nicht überschneidende Produktkategorien (Attribute der Daten) besonders stark. Für die interessanten Produktkategorien einer Kundengruppe wird bei der Generierung eine andere Verteilung (Normalverteilung) als für die nicht interessanten Produktkategorien (Gleichverteilung) eingesetzt. Die Attribute, für welche sich die einzelnen Kundengruppen interessieren, werden im Folgenden aufgelistet:

1. Technik-Enthusiasten
    a) Computer
    b) Mobiltelefone

2. Freizeit-Enthusiasten
    a) Games
    b) Filme & Musik
3. Automatisierungs-Enthusiasten
    a) Smart Home
    b) Haushaltsgeräte
4. Outdoor-Enthusiasten
    a) Foto
    b) Sport & Freizeit

Die Instanzen des Datensatzes werden einzeln mit normalverteilten interessanten und gleichverteilten uninteressanten Attributen generiert. Bei der Generierung einer Instanz der Daten wird zuerst eine Kundengruppe (jede Kundengruppe wird gleichhäufig selektiert) ausgewählt. Die Werte der für die gewählte Kundengruppe **interessanten** Attribute werden daraufhin zufällig **normalverteilt** und mit einem Durchschnitt der Verteilung von 0.8 sowie einer Standardabweichung von 0.1 gewählt. Anschließend werden alle für die Kundengruppe **nicht interessanten** Attributwerte zufällig **gleichverteilt** zwischen 0 und 1 gewählt.
Abschließend wird die zur Generierung der Instanz verwendete Kundengruppe als korrektes Label (Tabelle 5.1 Label-Attribut mit f_enth –> Freizeit-Enthusiast) zur Evaluation von Clustering-Resultaten angehängt.

| Computer | Mobiltelefone | Games | Filme & Musik | Smart Home | Haushaltsgeräte | Foto | Sport & Freizeit | Label |
|---|---|---|---|---|---|---|---|---|
| 0.643 | 0.404 | 0.826 | 0.817 | 0.241 | 0.170 | 0.505 | 0.083 | f_enth |
| 0.654 | 0.921 | 0.689 | 0.535 | 0.658 | 0.423 | 0.465 | 0.000 | t_enth |
| 0.546 | 0.214 | 0.591 | 0.572 | 0.606 | 0.822 | 0.013 | 0.640 | a_enth |
| 0.756 | 0.334 | 0.350 | 0.417 | 0.896 | 0.765 | 0.424 | 0.631 | a_enth |
| 0.660 | 0.754 | 0.008 | 0.473 | 0.225 | 0.759 | 0.703 | 0.000 | t_enth |

**Tabelle 5.1:** Exemplarischer Ausschnitt aus den zur Evaluation verwendeten Daten mit verwendetem Schema (Computer bis Sport & Freizeit) und dem für die Generation der Daten verwendeten Label.

Für die Auswertung der Daten werden drei dieser Datensätze mit variierender Anzahl (1000, 5000, 10000) an Instanzen und normiertem Wertebereich (0-1) verwendet.

Die Evaluation in einem hierarchischen Kontext (Evaluations-Szenario 4) verwendet außerdem einen weiteren Datensatz, in welchem die Kundengruppe der Outdoor-Enthusiasten weiter in die Gruppen der Foto-Enthusiasten und die der Sport & Freizeit-Enthusiasten (welche sich jeweils beide verstärkt für ihre namensgebende Produktkategorie interessieren) unterteilt wird. Diese Unterteilung passt die Normalverteilung der beiden für Outdoor-Enthusiasten interessanten Produktkategorien (Foto und Sport & Freizeit) wie folgt an:

1. Foto-Enthusiasten
    a) Foto: Normalverteilt mit Durchschnitt 0.8 –> 0.85

   b) Sport & Freizeit: Normalverteilt mit Durchschnitt 0.8 –> 0.75

2. Sport & Freizeit
   a) Foto: Normalverteilt mit Durchschnitt 0.8 –> 0.75
   b) Sport & Freizeit: Normalverteilt mit Durchschnitt 0.8 –> 0.85

Die angepassten Wahrscheinlichkeitsverteilungen innerhalb der Kundengruppe der Outdoor-Enthusiasten ermöglichen die Modellierung einer Hierarchie, bei welcher der Outdoor-Enthusiast in zwei weitere (exemplarisch in Abbildung 5.1) Untergruppen unterteilt wird.

### 5.1.3 Auswertungsmethodik

Die Auswertung einzelner Evaluations-Szenarien erfolgt durch die schrittweise Anwendung von Interaktionen. Das Zwischenergebnis sowie die für die Berechnung der einzelnen Schritte des Clustering-Resultates benötigte Rechenzeit, wird gespeichert. Für die Evaluation wurde ein Notebook mit Windows 11 Pro (Version 21H2) Betriebssystem, AMD Ryzen 7 PRO 4750U Prozessor und 16.00GB Arbeitsspeicher genutzt. Aufgrund der Eigenschaft des zugrundeliegenden k-Means-Algorithmus gegen ein lokales Minimum zu konvergieren, wird für die Gegenüberstellung der Werte der Durchschnitt aus fünf automatisch angewendeten Testdurchläufen gebildet. Für die Gegenüberstellung von Clustering-Resultaten wird die Accuracy auf Basis der (aus der Generation der Daten) bekannten korrekten Label berechnet.

**Die Berechnung der Accuracy** eines Clustering-Resultates vergleicht die aus der Generation des Datensatzes verfügbaren korrekten Labeln der einzelnen Instanzen des Clustering-Resultates. Die verschiedenen aus dem berechneten Clustering-Resultat resultierenden Label der Instanzen (normalerweise in numerischer Form 0, 1, 2, ...) werden für die Berechnung der Accuracy den korrekten Labeln (aus der Erstellung der Daten) der Daten zugeordnet. Anschließend wird die Accuracy anhand der Übereinstimmung zwischen den zugeordneten Labeln und den tatsächlich korrekten Labeln berechnet.

**Die automatisierte Anwendung von Interaktionen** setzt sich aus der Beantwortung von Active-Clustering Abfragen (Evaluations-Szenario 2) und der Anwendung von Regeln zur Restriktion von Wertebereichen (Evaluations-Szenario 3-4) zusammen.

Die Beantwortung von Active-Clustering Abfragen ist durch die Verwendung der aus der Generierung der Daten (Abschnitt 5.1.2) bekannten ground-truth Cluster trivial, da für die Beantwortung der paarweisen Abfrage nur die korrekten Label der beiden abgefragten Instanzen miteinander verglichen (gleiches Label –> gehören ins selbe Cluster) werden müssen.

Die automatisierte Anwendung von Regeln gestaltet sich schwieriger. Der Experte hat in der Praxis Erfahrungen gemacht und hat dadurch geeignete Regeln im Kopf. Für eine automatisierte Anwendung müssen diese aber zuerst identifiziert werden. Zur Identifikation dieser Regeln werden die aus der Normalverteilung der interessanten Attribute (Abschnitt 5.1.2) hervorgehenden Verteilungsunterschiede der Kundengruppen verwendet. Diese Verteilungsunterschiede ermöglichen die Abgrenzung einzelner Kundengruppen aufgrund der durchschnittlich deutlich höheren Attributwerte der interessanten Attribute. Dadurch können Instanzen anderer Kundengruppen mit einem niedrigeren Attributwert ausgeschlossen werden. Aufgrund der Gleichverteilung der Attributwerte

(Abschnitt 5.1.2) ist ein kompletter Ausschluss der anderen Kundengruppen allerdings nicht möglich. Das Ergebnis bilden Regeln der Form „Kundengruppe X, interessantes Attribut Y > Minimum der normalverteilten Werte“.

*Beispiel Identifikation von Regeln:* In diesem Beispiel wird exemplarisch die Identifikation der Regeln der Kundengruppe Technik-Enthusiasten betrachtet. Diese besitzen die beiden interessanten Attribute „Computer“ und „Mobiltelefone“, welche anders verteilt sind als die übrigen Attribute. Die Minima dieser Attribute sind exemplarisch 0.56 für das Attribut „Computer“ und 0.53 für das Attribut „Mobiltelefone“. Daraus leiten sich die Regeln ab:

1. Computer >= 0.56
2. Mobiltelefone >= 0.53

Die erstellten Regeln werden während der automatische Anwendung in den Evaluations-Szenarien 3-4 in verschiedenen Reihenfolgen angewandt. Für die Anwendung der Regeln in unterschiedlichen Reihenfolgen wird eine Reihe von (greedy) Selektionsverfahren verwendet:

Optimale Selektion: Selektiert die nächste Regel, für welche im aktuellen Clustering-Resultat am meisten Verstöße existieren. Ein Verstoß wird als Instanz mit einem gegen den Wertebereich der Regel verletzenden Attributwert definiert.

Zufällige Selektion: Selektiert die nächste Regel zufällig aus der Menge der noch nicht verwendeten Regeln.

Worst-Case Selektion: Selektiert die nächste Regel, für welche im aktuellen Clustering-Resultat am wenigsten Verstöße (Gegenteil zur optimalen Selektion) existieren.

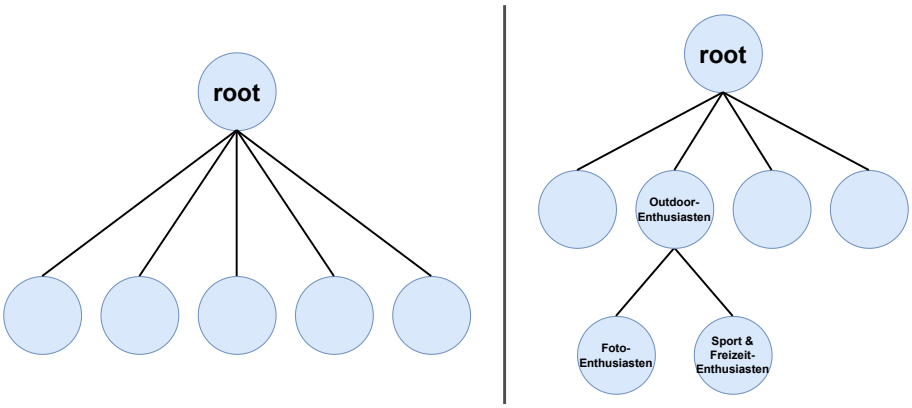



**Abbildung 5.1:** Schematischer Evaluationsaufbau zur Auswertung des vierten Szenarios, mit der verwendeten Konfiguration für die nicht nicht-hierarchische (links) sowie die hierarchische (rechts) Auswertung.

**Die Durchführung der hierarchischen Clusteranalyse** des vierten Evaluations-Szenarios verwendet zusätzlich eine modellierte Cluster-Hierarchie. Hierfür wird die weitere Unterteilung der Kundengruppe Outdoor-Enthusiasten in die beiden Untergruppen der Foto-Enthusiasten und der Sport & Freizeit-Enthusiasten verwendet. Die daraus resultierende Hierarchie wird schematisch

in Abbildung 5.1 (rechts), dargestellt. Bei der Durchführung wird die Verwendung dieser hierarchischen Struktur der Daten der direkten Gruppierung der Daten in fünf Kundengruppen (Abbildung 5.1, links) gegenübergestellt.

## 5.2 Evaluationsergebnisse

Dieser Abschnitt befasst sich mit den Resultaten der Evaluation. Betrachtet werden die nach Abschnitt 5.1.3 erstellten durchschnittlichen Resultate aus fünf Durchläufen. Die Auswertung der Ergebnisse wird auf Basis des untersuchten Sachverhaltes in drei Abschnitte unterteilt.

### 5.2.1 Untersuchung der Ansätze

In Abbildung 5.2 werden drei Evaluations-Szenarien gegenübergestellt und die Genauigkeit und Rechenzeit der Ansätze abhängig von der Zahl der angewendeten Interaktionsschritte miteinander verglichen. Die aus dem ersten Szenario resultierende Baseline des k-Means++ Resultates dient als Indikator für die Accuracy eines klassischen, nicht interaktiven Ansatzes. Die beiden anderen interaktiven Evaluations-Szenarien übertreffen diese Baseline durch die schrittweise Verfeinerung des Clustering-Resultates. Eine Ausnahme stellt die Anwendung des PCKMeans Algorithmus auf dem Datensatz mit 10000 Instanzen dar, in welcher keine Verbesserung erkennbar ist. Der Grund hierfür könnte ein aus der höheren Instanzanzahl abgeleiteter Bedarf nach einer größeren Menge an Constraints sein.

Der Einsatz von Regeln zur Einschränkung der Attributwerte (Szenario 3) verbessert die Accuracy signifikant und übertrifft den Constraint-basierten Ansatz (Szenario 2) deutlich. Bereits die Anwendung von wenigen Regeln führt auf den getesteten Datensätzen zu einer deutlichen Steigerung der erzielten Accuracy. Es ist außerdem schon ab dem ersten Interaktionsschritt eine deutlich erhöhte Rechenzeit für die Berechnung des Active-Clustering Resultates (Szenario 2) erkennbar. Dieser Zeitunterschied ist bei der Verwendung großer Datenmengen besonders ausgeprägt. Der Unterschied der Rechenzeit lässt sich auf den zusätzlich benötigten Aufwand zur Identifizierung von paarweisen Instanzen für die Constraint-Abfrage zurückführen.

Auffällig ist darüber hinaus die beim Active-Clustering Ansatz (PCKMeans) deutlich erhöhte Menge an zur Verbesserung der Accuracy benötigten Interaktionen. Der Domänenexperte muss bei dem Active-Clustering Ansatz aus diesem Grund einen erhöhten Aufwand für die Verfeinerung des Resultates betreiben, als bei der Regel-basierten Verfeinerung des Prototyps. Die Reduktion des Aufwandes könnte auf die Anwendung von Regeln auf gesamtheitliche Cluster, anstatt auf einzelne (paarweise) Instanzen zurückzuführen sein. Es ergibt sich aber die Möglichkeit, anhand der (vom Prototyp unterstützten) Kombination der beiden Ansätze, durch die Verwendung ihres unterschiedlichen Domänenwissens, zu profitieren.

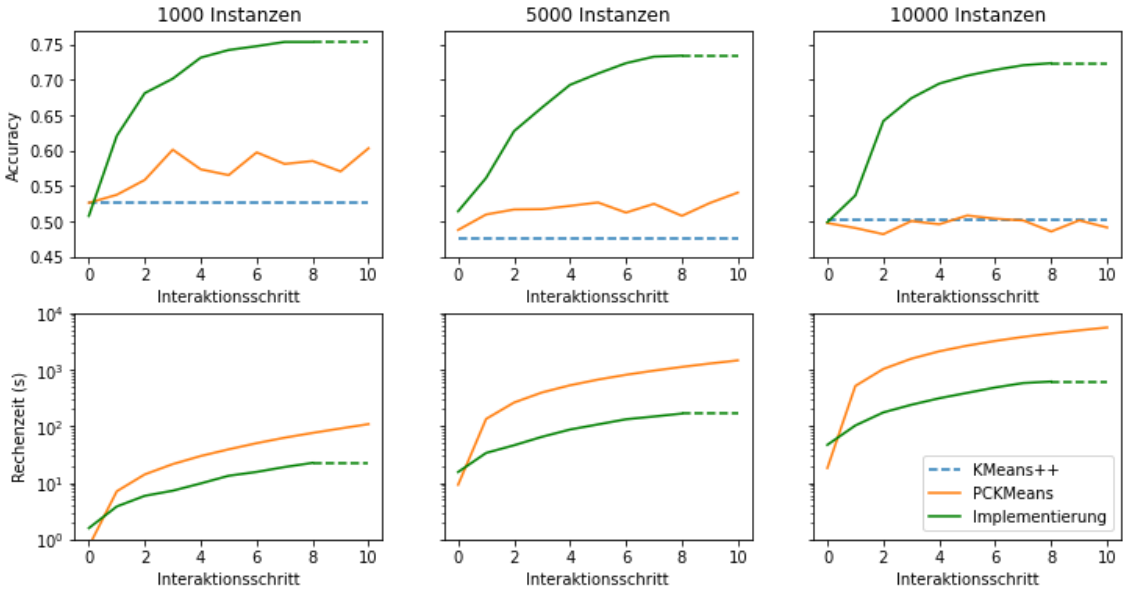



**Abbildung 5.2:** Gegenüberstellung der Accuracy und verwendeten Rechenzeit von k-Means++ (Szenario 1), PCKMeans (Szenario 2) und Implementierung (Szenario 3, optimale Selektionsstrategie) auf drei Datensätzen mit Zusammenfassung von jeweils einer Regel und zehn paarweisen Constraints zu einem Interaktionsschritt. Gestrichelte Linien symbolisieren Zeiträume ohne Interaktion, welche auf die Erschöpfung der anwendbaren Regeln oder den verwendeten Clustering-Ansatz (k-Means++) zurückzuführen sind.

### 5.2.2 Untersuchung der Selektionsstrategien

Dieser Abschnitt betrachtet variierende Anwendungsreihenfolgen von Regeln. Hierzu wird von der Abbildung 5.3 die Durchführung des dritten Evaluations-Szenarios (Abbildung 5.2) unter Verwendung der verschiedenen eingeführten Selektionsstrategien (Abbildung 5.3) genauer betrachtet. Verglichen wird die Anwendung der Regeln nach der optimalen, worst-case und zufälligen Selektionsstrategie. Es wird außerdem die Anwendung von Regeln, welche das 5% Quartil anstatt des Minimums der Daten zur Abgrenzung der Kundengruppen verwenden (orange), ausgewertet.

Die Auswertung zeigt, dass die gleiche maximale Accuracy unabhängig von der Selektionsreihenfolge der Regeln erreicht wird. Durch die gewählte Selektionsreihenfolge wird allerdings die Geschwindigkeit, mit der die Accuracy des Clustering-Resultates ansteigt, beeinflusst. Die Selektion der Regeln nach der optimalen Selektionsstrategie (grün) verbessert die Accuracy am schnellsten. Bei der Verwendung der worst-case (rot) Selektionsstrategie steigt die Accuracy zunächst am langsamsten an. Es zeigt sich aber, dass sogar die Auswahl der Regeln in der ungünstigsten Reihenfolge (worst-case Selektion) bereits nach drei Interaktionen in jedem Fall zu einer deutlichen Verbesserung der Accuracy führt. Der Erfolg der Verfeinerung ist daher nicht maßgeblich von der Anwendungsreihenfolge der Regeln durch den Domänenexperten abhängig. Die Accuracy der zufälligen Selektionsstrategie (blau) bewegt sich zwischen den Accuracy-Werten der worst-case und best-case Strategie. Dies zeigt, dass ein Domänenexperte auch bei willkürlichen Reihenfolgen mit bereits wenigen Regeln eine deutliche Verbesserung erzielen kann.

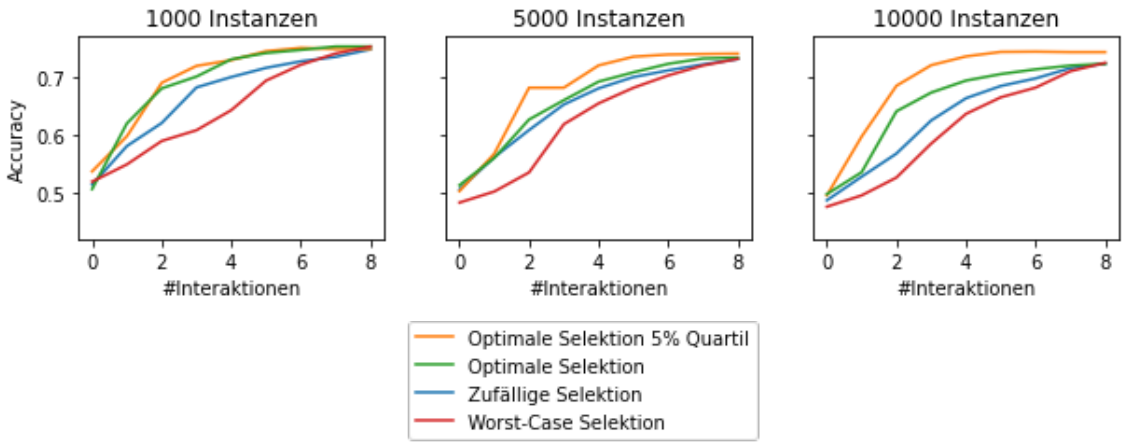



**Abbildung 5.3:** Ausführlicher Vergleich von verschiedenen Strategien zur Selektion der Anwendungsreihenfolge der Regeln, mit der optimale, zufälligen und worst-case Selektion (Abschnitt 5.1.3) sowie der Verwendung des 5% Quartils (orange) anstatt des Minimums zur Einschränkung der Wertebereiche.

Auffällig ist außerdem eine höhere maximal erreichte Accuracy bei der Verwendung der Regeln, welche das 5% Quartil (orange) nutzen, auf den großen (5000/10000 Instanzen) Datensätzen. Dies ist wahrscheinlich auf die Verwendung der Normalverteilung zur Generierung der Daten zurückzuführen, aus welcher bei einer großen Anzahl generierter Instanzen mit hoher Wahrscheinlichkeit einige extrem kleine Werte hervorgehen. Derartige Werte senken die Effizienz bei der Verwendung des Minimums zur Erstellung der Regeln. Daraus geht der negative Einfluss von Extremwerten auf die Anwendung von Regeln hervor, sowie die Möglichkeit, diesen durch die Verwendung von nicht exakt alle korrekten Instanzen umfassenden (beispielsweise durch die Ausgrenzung von Extremwerten) Regeln zu mitigieren. Diese Regeln werden von Domänenexperten verwendet, welche die Definition der Regeln nicht auf Basis der (unbekannten) korrekten Instanzen, sondern mithilfe ihres Domänenwissens durchführen.

### 5.2.3 Untersuchung der Hierarchie

Dieser Abschnitt untersucht die Verwendung der Cluster-Hierarchie und die Anwendung von Regeln auf die Hierarchie. Als Hierarchie wird entsprechend Abschnitt 5.1.3 die in Abbildung 5.1 dargestellte weitere Unterteilung der Kundengruppe Outdoor-Enthusiasten verwendet. Für die Auswertung wird in Abbildung 5.4 die Anwendung der Regeln sowohl in der hierarchischen als auch im nicht hierarchischen Clusteranalyse gegenübergestellt.

Bei der Durchführung der Clusteranalyse führt die Verwendung der Cluster-Hierarchie zu einem besseren initialen Clustering-Resultat (orange gestrichelt), als die nicht hierarchische Anwendung (grün gestrichelt). Diese höhere initiale Accuracy ist möglicherweise auf die (für den Clustering-Algorithmus deutlich) einfachere Unterteilung der Daten in vier anstatt fünf Cluster auf der ersten Hierarchie-Ebene (Abbildung 5.1) zurückzuführen. Zusätzlich führt die Verwendung der Hierarchie bei der Anwendung von Regeln (pink) zu einer schnelleren Annäherung an den maximal erreichten Accuracy-Wert. Die Ausprägung des maximal erreichte Accuracy-Wertes wird allerdings durch die Verwendung der Cluster-Hierarchie nicht signifikant beeinflusst.

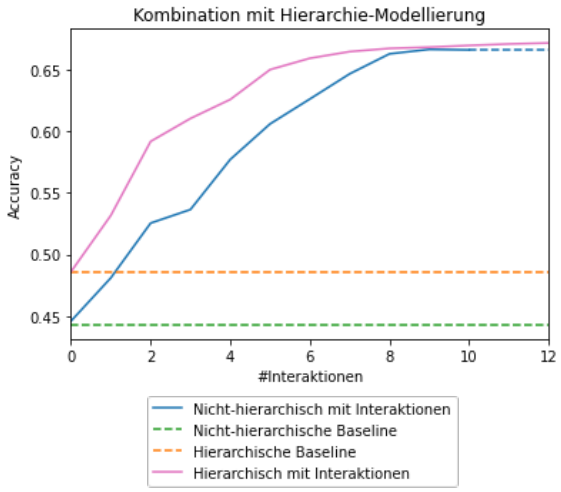



**Abbildung 5.4:** Auswertung der Cluster-Hierarchie in Kombination mit Regeln zur Begrenzung des Wertebereiches (Evaluations-Szenario 4) sowie mit den initialen Accuracy-Werten (gestrichelt) der hierarchischen und nicht hierarchischen Auswertung.

Die Modellierung bekannter hierarchischer Abhängigkeiten wirkt sich auf den ausgewerteten Daten in jedem Fall positiv auf das Clustering-Resultat aus. Der Domänenexperte erreicht auf diesen Daten unabhängig von der Anzahl der ihm bekannten Regeln immer ein besseres Clustering-Resultat durch die Integration der hierarchischen Abhängigkeiten in die Clusteranalyse. Bei der hierarchischen Durchführung ergibt sich aufgrund der Unterteilung der Problemstellung zusätzlich die Möglichkeit, das Resultat weiter zu verfeinern. Hierfür kann die Aufteilung der ersten Hierarchie-Ebene (siehe Abbildung 5.1), etwa basierend auf den zwei für die Kundengruppen der folgenden Hierarchie-Ebene wichtigen Attributen (Foto und Sport & Freizeit), gewichtet werden.

# 6 Zusammenfassung und Ausblick

Die weltweit produzierte und verfügbare Datenmenge steigt rasant an. Unternehmen können diese Daten analysieren und sich so Wettbewerbsvorteile schaffen, etwa durch die Optimierung von Geschäftsprozessen. Daraus resultiert ein steigender Bedarf, neue Datenanalyse-Lösungen zu entwickeln. Eine populäre Form der Datenanalyse ist die Clusteranalyse. Die automatisierte Anwendung von Clustering-Verfahren ist allerdings unzureichend, da Clustering-Resultate in vielen Fällen von externem Domänenwissen abhängen. Aus diesem Grund bietet es sich an, Domänenexperten mit ihrem implizit vorhandenen Domänenwissen direkt interaktiv in den Analyseprozess zu integrieren. Vorhandene Ansätze eignen sich jedoch aufgrund des Anforderungsprofils von Domänenexperten nicht und schränken Domänenexperten während der Analyse ein. Deshalb wird eine neue Lösung zur interaktiven Verfeinerung von Clustering-Resultaten benötigt.

In dieser Arbeit wurde deshalb ein an den Bedürfnissen eines Domänenexperten ausgerichtetes Prozessmodell entwickelt. Das eingeführte Prozessmodell bietet einen von Clustering-Algorithmen unabhängige Vorgehensweise für die Integration von Domänenexperten in den Analyseprozess. Damit die effektive Integration des Domänenexperten gewährleistet wird, wurden die Eigenschaften und Funktionsweisen der im Prozessmodell verwendeten Komponenten aufgezeigt und mithilfe von exemplarischen Umsetzungen veranschaulicht. Hierfür wurden populäre interaktive Ansätze zur Durchführung von Active-Clustering, Anwendung von Bearbeitungsoperationen, Nutzung von vorhandenen Ressourcen und kommunikative Ansätze zur Erklärung von Clustering-Resultaten integriert, um das Anforderungsprofil des Domänenexperten zu erfüllen. Das resultierende Prozessmodell ermöglicht Domänenexperten sowohl die eigenständige, als auch eine vom System eingeleitete Verfeinerung des Clustering-Resultates mithilfe der Generierung clusterspezifischer Handlungsvorschläge.

Auf Basis des ausgearbeiteten Prozessmodell wurde ein Prototyp zur interaktiven Verfeinerung von Clustering-Resultaten entwickelt. Der Prototyp kombiniert mehrere interaktive Clustering-Ansätze und bestätigt die Realisierbarkeit des eingeführten Konzeptes. Zur Verfeinerung der Clustering-Resultate ermöglicht der Prototyp die Erstellung von Hierarchien, Anwendung von Operationen, Änderung von Parametern und die vom System eingeleitet Verfeinerung mithilfe von Anpassungsvorschlägen, etwa durch Active-Clustering.

Abschließend wurde mithilfe des Prototyps eine umfangreiche Evaluation auf Basis vier synthetisch generierter Datensätze durchgeführt. Bei dieser wurden anhand von Evaluationsszenarien verschiedene Vorgehensweisen zur interaktiven Verfeinerung ausgewertet und gegenübergestellt. Die Evaluation verwendet den nicht interaktiven k-Means++-Algorithmus und den PCKMeans-Algorithmus unter Einsatz von Active-Clustering als Baseline für die Qualität der Resultate. Dementsprechend wurde die Anwendung von k-Means++, PCKMeans und die Verfeinerung des Prototyps miteinander verglichen. Daraus ergab sich, dass beide interaktiven Ansätze (PCKMeans und Prototyp) die Genauigkeit des k-Means++-Algorithmus übertreffen, wobei der neu entwickelte Prototyp die höchste Genauigkeit mit weniger Interaktionsschritten und geringerer Rechenzeit erzielt.

Dabei führte die Anwendung von Regeln in jedem Fall zu einer Verbesserung der Genauigkeit. Änderungen der Reihenfolge beeinflussten jedoch das Ausmaß der Verbesserung. Die beste Lösung für die evaluierten Szenarien ist die Kombination verschiedener Ansätze des Prototyps, welcher die beiden Baselines (k-Means++, PCKMeans) deutlich übertraf.

## Ausblick

Das Prozessmodell wurde entwickelt, um den Domänenexperten bei der Durchführung von Clusteranalysen im Rahmen realer Anwendungsszenarien zu unterstützen. In die Evaluation des Prototyps wurden aber bisher keine Domänenexperten eingebunden, sondern stattdessen deren Verhalten simuliert. Es ergibt sich daher der Bedarf, die Leistungsfähigkeit des Konzepts unter Verwendung von Domänenexperten anhand realer Aufgabenstellungen genauer zu bewerten. Für diese Auswertung ist die Durchführung einer Nutzerstudie nötig, welche die Resultate verschiedener Domänenexperten bei der Anwendung unterschiedlicher interaktiver Ansätze gegenüberstellt, um so qualitatives Feedback zu erhalten.

Zusätzlich kann das eingeführte Konzept zur Bewältigung sehr großer Datenmengen erweitert werden. Aus diesen Datenmengen ergibt sich die Herausforderung, die Interaktionsfähigkeit des Systems, trotz signifikant erhöhter Rechenzeiten, weiterhin zu gewährleisten. Zur Adressierung dieser Herausforderung kann eventuell die Berechnung auf Samples der Daten angewendet oder verteilt ausgeführt werden. Bei der Verwendung von Samples müssen die entstehenden Qualitätsverluste berücksichtigt und nach Möglichkeit, etwa durch „intelligente“ Samplingstrategien, mitigiert werden. Damit die verteilte Berechnung des Resultates unterstützt werden kann, muss eine Anpassung des Prozesses zur parallelen Berechnung von Teilresultaten vorgenommen werden. Hierfür muss vor allem auf mögliche Geschwindigkeitseinbußen geachtet werden, welche die Interaktion mit dem Domänenexperten negativ beeinflussen können.

Abschließend lässt sich das in dieser Arbeit vorgestellte interaktive Clustering-Verfahren für Domänenexperten in den übergeordneten Datenanalyseprozess integrieren. Die Grundlage für eine solche Integration wird bereits von Behringer et al. [BHM18] durch die Definition eines Datenanalyseprozesses aus der Perspektive von Domänenexperten gestellt. Diese Integration kann einerseits in Form einer Funktionserweiterung, welche weitere Schritte des Datenanalyseprozesses unterstützt oder andererseits als Einbindung des Konzepts in bereits vorhandene (Teil-)Umsetzungen des Datenanalyseprozesses (beispielsweise die Verwendung von Mashups zur Modellierung von Datenflüssen) erfolgen.

# Literaturverzeichnis


[AA20] W. Atwa, A. Almazroi. „Active Selection Constraints for Semi-supervised Clustering Algorithms“. In: *International Journal of Information Technology and Computer Science* (Dez. 2020), S. 23–30 (zitiert auf S. 13, 25–28, 31, 32, 34, 60).

[ABV13] P. Awasthi, M. F. Balcan, K. Voevodski. „Local algorithms for interactive clustering“. In: *Journal of Machine Learning Research* (Dez. 2013), S. 1–35 (zitiert auf S. 30, 57).

[AKG+14] C. C. Aggarwal, X. Kong, Q. Gu, J. Han, S. Y. Philip. „Active learning: A survey“. In: *Data Classification: Algorithms and Applications*. 2014, S. 599–634 (zitiert auf S. 24, 25).

[AL14] W. Atwa, K. Li. „Active query selection for constraint-based clustering algorithms“. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2014), S. 438–445 (zitiert auf S. 25, 27, 31, 34).

[Ale92] P. A. Alexander. „Domain Knowledge: Evolving Themes and Emerging Concerns“. In: *Educational Psychologist* (Jan. 1992), S. 33–51 (zitiert auf S. 14).

[AV07] D. Arthur, S. Vassilvitskii. „K-Means++: The Advantages of Careful Seeding“. In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Jan. 2007, S. 1027–1035 (zitiert auf S. 83).

[BBM02] S. Basu, A. Banerjee, R. J. Mooney. „Semi-Supervised Clustering by Seeding“. In: *Proceedings of the Nineteenth International Conference on Machine Learning*. Juli 2002, S. 27–34 (zitiert auf S. 31, 32, 40, 59).

[BBM04a] S. Basu, A. Banerjee, R. J. Mooney. „Active semi-supervision for pairwise constrained clustering“. In: *SIAM Proceedings Series* (2004), S. 333–344 (zitiert auf S. 13, 26–28, 31, 34, 60, 82).

[BBM04b] M. Bilenko, S. Basu, R. J. Mooney. „Integrating Constraints and Metric Learning in Semi-Supervised Clustering“. In: *Proceedings of the Twenty-First International Conference on Machine Learning*. Juli 2004, S. 81–88 (zitiert auf S. 31).

[BBR14] M. Babaee, R. Bahmanyar, G. Rigoll. „Interactive clustering for SAR image understanding“. In: *EUSAR 2014; 10th European Conference on Synthetic Aperture Radar*. Juli 2014, S. 1–4 (zitiert auf S. 30, 34).

[BFDL10] S. Basu, D. Fisher, S. Drucker, H. Lu. „Assisting Users with Clustering Tasks by Combining Metric Learning and Classification“. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Juli 2010 (zitiert auf S. 59).

[BHM18] M. Behringer, P. Hirmer, B. Mitschang. „A Human-Centered Approach for Interactive Data Processing and Analytics“. In: *Enterprise Information Systems*. Apr. 2018, S. 498–514 (zitiert auf S. 39, 98).

[BHR+20] J. Bae, T. Helldin, M. Riveiro, S. Nowaczyk, M.-R. Bouguelia, G. Falkman. „Interactive Clustering: A Comprehensive Review“. In: *ACM Computing Surveys* (Jan. 2020), S. 1–39 (zitiert auf S. 13, 23, 25–27, 29–34, 50, 51, 54, 56, 57, 59).

[BHTM22] M. Behringer, P. Hirmer, D. Tschechlov, B. Mitschang. „Increasing Explainability of Clustering Results for Domain Experts by Identifying Meaningful Features“. In: *Proceedings of the 24th International Conference on Enterprise Information Systems*. Mai 2022, S. 364–373 (zitiert auf S. 22–24, 29, 30, 33, 45, 46, 53, 54, 61, 62, 70, 72, 76).

[BOW18] D. Bertsimas, A. Orfanoudaki, H. Wiberg. „Interpretable Clustering via Optimal Trees“. In: *arXiv* (Dez. 2018) (zitiert auf S. 23, 33, 47).

[BPB+16] L. Boudjeloud-Assala, P. Pinheiro, A. Blansché, T. Tamisier, B. Otjacques. „Interactive and iterative visual clustering“. In: *Information Visualization* (Juli 2016), S. 181–197 (zitiert auf S. 22, 23, 25, 32, 33).

[BTH12] C. Blundell, Y. Teh, K. Heller. „Bayesian Rose Trees“. In: *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* (März 2012) (zitiert auf S. 57).

[CB17] T. V. Craenendonck, H. Blockeel. „Constraint-based clustering selection“. In: *Machine Learning* (Okt. 2017), S. 1497–1521 (zitiert auf S. 26).

[CD18] M. Cavallo, Ç. Demiralp. „Clustrophile 2: Guided Visual Clustering Analysis“. In: *IEEE Transactions on Visualization and Computer Graphics* (Apr. 2018), S. 267–276 (zitiert auf S. 23, 25, 30–34, 48, 52, 59, 62).

[CDG+17] A. Coden, M. Danilevsky, D. Gruhl, L. Kato, M. Nagarajan. „A Method to Accelerate Human in the Loop Clustering“. In: Juni 2017, S. 237–245 (zitiert auf S. 26, 29, 59).

[CDH+16] S. Chang, P. Dai, L. Hong, C. Sheng, T. Zhang, E. H. Chi. „AppGrouper: Knowledge-graph-based Interactive Clustering Tool for Mobile App Search Results“. In: *Proceedings of the 21st International Conference on Intelligent User Interfaces*. März 2016, S. 348–358 (zitiert auf S. 13, 25, 30, 34).

[CEM+15] M. B. Cohen, S. Elder, C. Musco, C. Musco, M. Persu. „Dimensionality reduction for k-means clustering and low rank approximation“. In: *Proceedings of the Annual ACM Symposium on Theory of Computing*. Juni 2015, S. 163–172 (zitiert auf S. 23).

[CGSQ11] N. Cao, D. Gotz, J. Sun, H. Qu. „DICON: interactive visual analysis of multidimensional clusters“. In: *IEEE transactions on visualization and computer graphics* (2011), S. 2581–2590 (zitiert auf S. 22, 23, 33).

[CLRP13] J. Choo, C. Lee, C. K. Reddy, H. Park. „UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization“. In: *IEEE Transactions on Visualization and Computer Graphics* (Dez. 2013), S. 1992–2001 (zitiert auf S. 30, 34).

[Dav06] T. Davenport. „Competing on Analytics“. In: *Harvard business review* 84 (Feb. 2006) (zitiert auf S. 13).

[DFB11] S. M. Drucker, D. Fisher, S. Basu. „Helping Users Sort Faster with Adaptive Machine Learning Recommendations“. In: *Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction - Volume Part III*. Sep. 2011, S. 187–203 (zitiert auf S. 31).

[DFMR20] S. Dasgupta, N. Frost, M. Moshkovitz, C. Rashtchian. „Explainable k-Means and k-Medians Clustering“. In: *Proceedings of the 37th International Conference on Machine Learning*. Juli 2020 (zitiert auf S. 22, 23, 33, 47).

[DH08] S. Dasgupta, D. Hsu. „Hierarchical sampling for active learning“. In: *Proceedings of the 25th international conference on Machine learning - ICML '08*. 2008, S. 208–215 (zitiert auf S. 24, 28).

[DR05] I. Davidson, S. S. Ravi. „Clustering With Constraints: Feasibility Issues and the k -Means Algorithm“. In: *Proceedings of the 2005 SIAM International Conference on Data Mining*. Apr. 2005, S. 138–149 (zitiert auf S. 31, 32).

[Dun73] J. C. Dunn. „A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters“. In: *Journal of Cybernetics* (Jan. 1973), S. 32–57 (zitiert auf S. 23).

[EDSN11] B. Eriksson, G. Dasarathy, A. Singh, R. Nowak. „Active Clustering: Robust and Efficient Hierarchical Clustering using Adaptively Selected Similarities“. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Apr. 2011, S. 260–268 (zitiert auf S. 28).

[EFS11] U. Erra, B. Frola, V. Scarano. „An Interactive Bio-inspired Approach to Clustering and Visualizing Datasets“. In: *2011 15th International Conference on Information Visualisation*. Juli 2011, S. 440–447 (zitiert auf S. 30).

[ESG+21] C. A. Ellis, M. S. E. Sendi, E. P. T. Geenjaar, S. M. Plis, R. L. Miller, V. D. Calhoun. „Algorithm-Agnostic Explainability for Unsupervised Clustering“. In: *arXiv* (Mai 2021) (zitiert auf S. 22, 23, 29, 33).

[FAT+14] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, A. Bouras. „A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis“. In: *IEEE Transactions on Emerging Topics in Computing* (Sep. 2014), S. 267–279 (zitiert auf S. 21, 22, 38).

[FGS13] R. Fraiman, B. Ghattas, M. Svarc. „Interpretable clustering using unsupervised binary trees“. In: *Advances in Data Analysis and Classification* (März 2013), S. 125–145 (zitiert auf S. 23, 33, 47).

[FR98] C. Fraley, A. E. Raftery. „How many clusters? Which clustering method? Answers via model-based cluster analysis“. In: *Computer Journal* (1998), S. 586–588 (zitiert auf S. 22).

[FZL13] Y. Fu, X. Zhu, B. Li. „A survey on instance selection for active learning“. In: *Knowledge and Information Systems* (Juni 2013), S. 249–283 (zitiert auf S. 24).

[GC88] R. Glaser, W. G. Chase. *The Nature of Expertise*. Hrsg. von M. T. Chi, R. Glaser, M. J. Farr. Psychology Press, 1988 (zitiert auf S. 14).

[HBS+16] M. Hund, D. Böhm, W. Sturm, M. Sedlmair, T. Schreck, T. Ullrich, D. A. Keim, L. Majnaric, A. Holzinger. „Visual analytics for concept exploration in subspaces of patient groups“. In: *Brain Informatics* (Dez. 2016), S. 233–247 (zitiert auf S. 30, 32, 33).

[HJK17] J. Hämäläinen, S. Jauhiainen, T. Kärkkäinen. „Comparison of Internal Clustering Validation Indices for Prototype-Based Clustering“. In: *Algorithms 2017, Vol. 10, Page 105* (Sep. 2017), S. 105 (zitiert auf S. 23, 24, 46).

[HTF09] T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009 (zitiert auf S. 21).

[IMN+07] F. Iorio, G. Miele, F. Napolitano, G. Raiconi, R. Tagliaferri. „An Interactive Tool for Data Visualization and Clustering“. In: *Knowledge-Based Intelligent Information and Engineering Systems*. 2007, S. 870–877 (zitiert auf S. 62, 63).

[Jai10] A. K. Jain. „Data clustering: 50 years beyond K-means“. In: *Pattern Recognition Letters* (Juni 2010), S. 651–666 (zitiert auf S. 21, 22, 52).

[JLJC05] J. Johansson, P. Ljung, M. Jern, M. Cooper. „Revealing structure within clustered parallel coordinates displays“. In: *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*. Okt. 2005, S. 125–132 (zitiert auf S. 33).

[JMF99] A. K. Jain, M. N. Murty, P. J. Flynn. „Data Clustering: A Review“. In: *ACM Computing Surveys (CSUR)* (Sep. 1999), S. 264–323 (zitiert auf S. 21, 22).

[KG20] P. Kumar, A. Gupta. „Active Learning Query Strategies for Classification, Regression, and Clustering: A Survey“. In: *Journal of Computer Science and Technology* (Juli 2020), S. 913–945 (zitiert auf S. 24, 25, 27, 28, 34, 60).

[LJJ07] Y. Liu, R. Jin, A. K. Jain. „BoostCluster: Boosting clustering by pairwise constraints“. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Aug. 2007, S. 450–459 (zitiert auf S. 26, 32).

[LKC+12] H. Lee, J. Kihm, J. Choo, J. Stasko, H. Park. „iVisClustering: An Interactive Visual Document Clustering via Topic Modeling“. In: *Computer Graphics Forum* (Juni 2012), S. 1155–1164 (zitiert auf S. 13, 30, 34, 53, 57).

[LKS+15] S. L'Yi, B. Ko, D. Shin, Y.-J. Cho, J. Lee, B. Kim, J. Seo. „XCluSim: A visual analytics tool for interactively comparing multiple clustering results of bioinformatics data“. In: *BMC Bioinformatics* (Aug. 2015), S. 1–15 (zitiert auf S. 30, 48, 49, 52).

[LWG11] U. Luxburg, R. Williamson, I. Guyon. „Clustering: Science or Art?“ In: *JMLR: Workshop and Conference Proceedings* (Juli 2011), S. 6579 (zitiert auf S. 22, 46).

[LXY00] B. Liu, Y. Xia, P. S. Yu. „Clustering Through Decision Tree Construction“. In: *Proceedings of the ninth international conference on Information and knowledge management - CIKM '00* (Okt. 2000) (zitiert auf S. 23).

[LXY05] B. Liu, Y. Xia, P. Yu. „Clustering Via Decision Tree Construction“. In: *Foundations and Advances in Data Mining*. Sep. 2005, S. 97–124 (zitiert auf S. 33, 47).

[Mac67] J. B. MacQueen. „Some Methods for Classification and Analysis of MultiVariate Observations“. In: *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1967, S. 281–297 (zitiert auf S. 13).

[Mad12] T. S. Madhulatha. „An Overview on Clustering Methods“. In: *IOSR Journal of Engineering* (Mai 2012), S. 719–725 (zitiert auf S. 21).

[MD18] X. Ma, S. Dhavala. „Hierarchical Clustering with Prior Knowledge“. In: *arXiv* (Juni 2018) (zitiert auf S. 26, 31, 32).

[Nic07] T. G. Nick. „Descriptive Statistics“. In: *Topics in Biostatistics*. 2007, S. 33–52 (zitiert auf S. 48).

[OY11] M. Okabe, S. Yamada. „An Interactive Tool for Human Active Learning in Constrained Clustering“. In: *Journal of Emerging Technologies in Web Intelligence* (Feb. 2011), S. 20–27 (zitiert auf S. 13, 61).

[RGR18] D. Reinsel, J. Gantz, J. Rydning. *The Digitization of the World From Edge to Core*. Techn. Ber. IDC, Nov. 2018 (zitiert auf S. 13).

[Rou87] P. J. Rousseeuw. „Silhouettes: A graphical aid to the interpretation and validation of cluster analysis“. In: *Journal of Computational and Applied Mathematics* (1987), S. 53–65 (zitiert auf S. 23).

[Set09] B. Settles. *Active Learning Literature Survey*. Computer Sciences Technical Report. University of Wisconsin–Madison, Jan. 2009 (zitiert auf S. 24, 25).

[SI09] Y. Sato, M. Iwayama. „Interactive Constrained Clustering for Patent Document Set“. In: *Proceedings of the 2nd International Workshop on Patent Information Retrieval*. Nov. 2009, S. 17–20 (zitiert auf S. 32).

[SL97] D. J. Simons, D. T. Levin. „Change blindness“. In: *Trends in Cognitive Sciences* (Okt. 1997), S. 261–267 (zitiert auf S. 49).

[SPG+17] A. Saxena, M. Prasad, A. Gupta, N. Bharill, o. Patel, A. Tiwari, M. Er, C.-T. Lin. „A Review of Clustering Techniques and Developments“. In: *Neurocomputing* (2017), S. 664–681 (zitiert auf S. 21, 22, 28).

[SS02] J. Seo, B. Shneiderman. „Interactively exploring hierarchical clustering results“. In: *Computer* (Aug. 2002), S. 80–86 (zitiert auf S. 23, 29, 32, 33).

[Šve18] J. Švehla. *active-semi-supervised-clustering*. Sep. 2018. URL: https://github.com/datamole-ai/active-semi-supervised-clustering (zitiert auf S. 83).

[SW10] L. L. Sun, X. Z. Wang. „A survey on active learning strategy“. In: *2010 International Conference on Machine Learning and Cybernetics, ICMLC 2010* (2010), S. 161–166 (zitiert auf S. 24, 25).

[SZS16] A. Srivastava, J. Zou, C. Sutton. „Clustering with a Reject Option: Interactive Clustering as Bayesian Prior Elicitation“. In: *arXiv* (Feb. 2016) (zitiert auf S. 13, 26, 30, 31, 58).

[TAL14] J. Tang, S. Alelyani, H. Liu. „Feature Selection for Classification: A Review“. In: *Data Classification: Algorithms and Applications*. 2014, S. 37–64 (zitiert auf S. 23).

[TFS21] D. Tschechlov, M. Fritz, H. Schwarz. „AutoML4Clust: Efficient AutoML for Clustering Analyses“. In: *Proceedings of the 24th International Conference on Extending Database Technology*. März 2021, S. 343–348 (zitiert auf S. 22).

[THLN01] A. K. Tung, J. Han, L. V. Lakshmanan, R. T. Ng. „Constraint-based clustering in large databases“. In: *Database Theory*. 2001, S. 405–419 (zitiert auf S. 26).

[TPRH11] C. Turkay, J. Parulek, N. Reuter, H. Hauser. „Interactive Visual Analysis of Temporal Cluster Structures“. In: *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*. Juni 2011, S. 711–720 (zitiert auf S. 30, 32, 33).

[VD16] S. Vikram, S. Dasgupta. „Interactive Bayesian Hierarchical Clustering“. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning*. Juni 2016, S. 2081–2090 (zitiert auf S. 27, 28, 32).

[WCRS01] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl. „Constrained K-Means Clustering with Background Knowledge“. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. 2001, S. 577–584 (zitiert auf S. 31, 32).

[WD10] X. Wang, I. Davidson. „Active Spectral Clustering“. In: *2010 IEEE International Conference on Data Mining*. Dez. 2010, S. 561–568 (zitiert auf S. 13, 27, 28, 31, 34).

[XJC14] C. Xiong, D. Johnson, J. J. Corso. „Active Clustering with Model-Based Uncertainty Reduction“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Feb. 2014), S. 5–17 (zitiert auf S. 26–28, 31, 34).

[YWL+21] W. Yang, X. Wang, J. Lu, W. Dou, S. Liu. „Interactive Steering of Hierarchical Clustering“. In: *IEEE Transactions on Visualization and Computer Graphics* (Okt. 2021), S. 3953–3967 (zitiert auf S. 31–34, 44, 48, 50).

[ZL11] L. Zheng, T. Li. „Semi-supervised Hierarchical Clustering“. In: *2011 IEEE 11th International Conference on Data Mining*. Dez. 2011, S. 982–991 (zitiert auf S. 32).

[ZWL10] S. Zhu, D. Wang, T. Li. „Data clustering with size constraints“. In: *Knowledge-Based Systems* (Dez. 2010), S. 883–889 (zitiert auf S. 55).

Alle URLs wurden zuletzt am 13. 12. 2022 geprüft.

**Erklärung**

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

---

Ort, Datum, Unterschrift