Institut für Parallele und Verteilte Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit

Feature-getriebene Darstellung von Clustering-Resultaten

Kay Roggenbuck

Studiengang:	Informatik
Prüfer/in:	PD Dr. rer. nat. habil. Holger Schwarz
Betreuer/in:	Dipl.-Inf. Michael Behringer, Manuel Fritz, M. Sc.
Beginn am:	2. Oktober 2018
Beendet am:	2. April 2018

Kurzfassung

Durch die voranschreitende Digitalisierung steigt die Menge an erzeugten und gespeicherten Daten schnell an. Um diese Datenmenge dem Menschen verständlich zu machen, werden häufig Data-Mining-Verfahren wie beispielsweise Clustering eingesetzt. Clustering-Verfahren separieren den zugrunde liegenden Datensatz in Gruppen mit Ähnlichkeiten. Da Clustering-Verfahren keine Dimensionsreduktion durchführen, besitzen die Resultate noch immer viele Dimensionen und Datenpunkte. Dies sorgt dafür, dass die Interpretation dieser Resultate für den Menschen schwer ist.

In dieser Arbeit wird ein Ansatz vorgestellt, welcher eine ausreichend detaillierte und zudem einfach verständliche Darstellung für Clustering-Resultate liefert. Um eine solche Darstellung zu ermöglichen, werden einzelne Cluster durch eine geringe Anzahl ausgewählter Informationen repräsentiert. Dabei werden Features anhand ausgewählter und neu entwickelter Metriken nach ihrer Aussagekraft für das Clustering bewertet und ausgewählt. Für die Wertebereiche dieser Features werden statistische Kenngrößen ermittelt. Weiterhin werden verschiedene Darstellungsformen dieser Ergebnisse vorgestellt, wie zum Beispiel Tabellen oder Wortwolken.

Eine Evaluation mithilfe eines Goldstandards zeigt, dass der entwickelte Ansatz für das Finden aussagekräftiger Features eine hohe Genauigkeit und eine lineare Laufzeitkomplexität besitzt.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Verzeichnis der Algorithmen

1 Motivation

Die Menge an Daten, die von Computern erfasst, gespeichert und verarbeitet werden, wächst rasant an. Nach Gantz und Reinsel betrug die Menge an erzeugten Daten im Jahre 2012 etwa 640 Milliarden Gigabyte und wird bis zum Jahr 2020 voraussichtlich auf 44 Billionen Gigabyte ansteigen. Bereits für das Jahr 2012 wurden etwa 23% der Daten als nützlich eingestuft, aber nur 0.5% der Daten analysiert [GR12]. Zudem ist zu erwarten, dass mit der Menge an erzeugten Daten auch die Menge an nützlichen Daten rasant steigt. Ein Grund für die geringe Menge an analysierten Daten ist, dass ein Mensch nur eine begrenzte Menge an Daten wahrnehmen und analysieren kann [MR09]. Vor allem durch die voranschreitende Digitalisierung steigt das Interesse an der Analyse der Daten. So können zum Beispiel in Zukunft medizinische Geräte in Echtzeit Daten von Patienten erfassen, um mit einer geeigneten Analyse frühzeitig Erkrankungen festzustellen. Zudem können durch solche Daten auch wirtschaftliche Vorteile erzielt werden, indem zum Beispiel das Verhalten von großen Kundenmassen analysiert wird, um zukünftige Trends vorhersagen zu können [GR12]. Anhand dieser Beispiele lässt sich abschätzen, welche Möglichkeiten sich durch eine geeignete Analyse dieser Daten ergeben und welche Konsequenzen eine fehlende Analyse dieser Daten haben kann. Somit steigt mit der Menge an erzeugten Daten auch die Relevanz von Hilfsmitteln, welche Daten für den Menschen leichter analysierbar machen.

Beispiele für solche Hilfsmittel sind Data-Mining-Verfahren wie Clustering oder Klassifikation. Diese unterteilen Daten durch eine Zuweisung in verschiedene Gruppen. Während bei der Klassifikation die verschiedenen Gruppen zu Beginn feststehen, werden diese beim Clustering durch den Algorithmus bestimmt. Dies sorgt dafür, dass für Clustering-Resultate die entstandene Gruppierung erkannt werden kann, jedoch ein tieferes Verständnis in diese ohne Weiteres nicht möglich ist. Demnach kann nach dem Clustering-Verfahren ohne weitere Verarbeitung keine Aussage darüber gemacht werden, worin die Cluster sich unterscheiden oder welche Aussage ein einzelnes Cluster liefert. Das Ziel dieser Arbeit ist es, einen Ansatz zu entwickeln, welcher Clustering-Resultate weiter verarbeitet und somit eine für den Menschen einfach verständliche und detaillierte Repräsentation dieser liefert.

Aufbau dieser Arbeit

Diese Arbeit ist wie folgt gegliedert:

Kapitel 2 - Grundlagen umfasst detaillierte Informationen zu Clustering-Verfahren und statistischen Mitteln diese zu vergleichen. Zudem wird ein kurzer Einblick in die menschliche Wahrnehmung und Streuungsmaße gegeben.

Kapitel 3 - Verwandte Arbeiten beschreibt dimensionsreduzierende und textuelle Ansätze zur Visualisierung von Clustering-Resultaten.

Kapitel 4 - Konzept stellt den entwickelten Ansatz dar. Dieser ermittelt aussagekräftige Features der Clustering-Resultate anhand Streuungsmaßen und Metrik-Differenzen. Für die Darstellung der aussagekräftigen Features werden statistische Kenngrößen verwendet.

Kapitel 5 - Evaluation beschreibt den Aufbau und die Durchführung der Evaluation anhand eines Goldstandards in Hinblick auf Genauigkeit und Laufzeit.

Kapitel 6 - Zusammenfassung und Ausblick fasst die Ergebnisse der Arbeit zusammen und gibt einen Ausblick auf potenzielle weitere Forschung in diesem Gebiet.

2 Grundlagen

In diesem Kapitel werden die für diese Arbeit notwendigen Grundlagen erläutert. Diese umfassen Clustering, Wahrnehmung des Menschen, Streuungsmaße und Metriken zur Ermittlung der Ähnlichkeit von Clustern.

2.1 Clustering

Clustering beschreibt die Unterteilung von Daten in Gruppen mit der Eigenschaft, dass Daten innerhalb einer Gruppe sehr ähnlich und die Daten aus verschiedenen Gruppen große Unterschiede aufweisen. Dabei werden die Gruppen durch den Algorithmus festgelegt. Das Gruppieren von Daten stellt eine wichtige Aufgabe in verschiedenen Anwendungsdomänen wie zum Beispiel Text-Mining oder Bildverarbeitung dar [GMW07].

Abbildung 2.1 stellt mögliche Gruppen von Clustering-Verfahren und deren Zusammenhänge dar. Diese Auflistung an Verfahren ist nur ein Auszug aller Möglichen, da in diesem Gebiet aktuell viel geforscht wird und somit stetig neue Verfahren entstehen. Nach Kaufman und Rousseeuw wird auf der obersten Ebene unterschieden zwischen [KR09]:

Hierarchische Verfahren werden häufig durch Dendrogramme dargestellt, wie in Abbildung 2.2. In diesem Beispiel sind auf der tiefsten Ebene die Punkte B und C, D und E, F und G zu einem Cluster zusammengefasst. In der nächst höheren Ebene werden dann A mit (B,C) und (D,E) mit

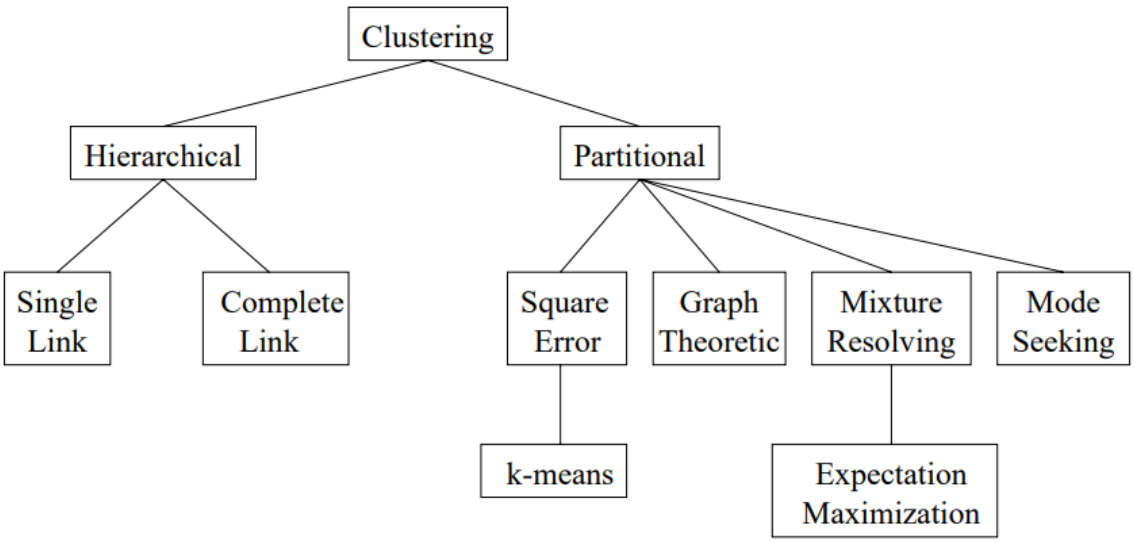


Abbildung 2.1: Unterteilung von Clustering-Verfahren [JMF99]

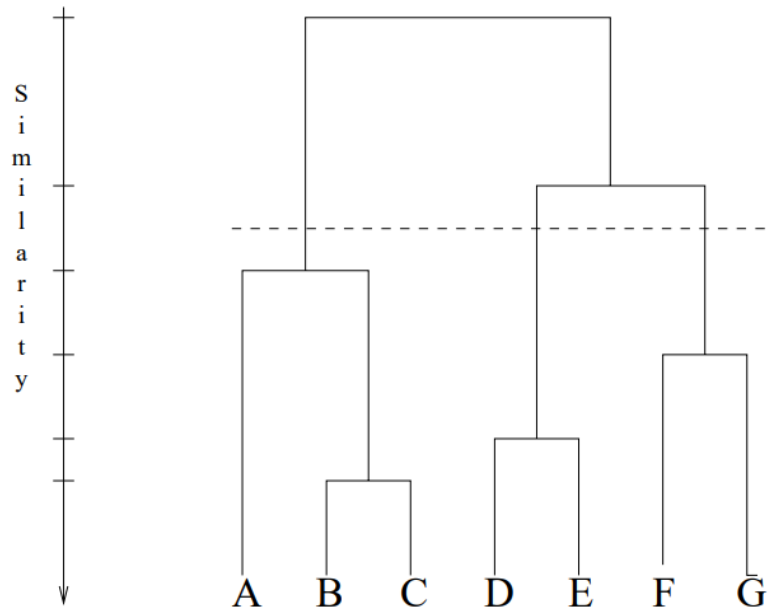


Abbildung 2.2: Beispiel eines Dendrogramms [JMF99]

(F,G) zusammengefasst. Welche Datenpunkte zusammengefasst werden hängt von der Wahl des Verfahrens ab. Der Single-Link Algorithmus wählt zum Beispiel das Minimum der Abstände zwischen den Datenpunkten der Cluster, während Complete-Link das Maximum der Abstände wählt. Der Nachteil des hierarchischen Ansatzes ist die hohe Laufzeit. So ist die Erzeugung eines Dendrograms für große Datensätze durch die lange Laufzeit nur schwer realisierbar [JMF99].

Partitionierende Verfahren unterteilen den Datensatz nur auf einer Ebene, wodurch die Auswahl der richtigen Anzahl an Clustern ein relevanter Faktor ist. Der Vorteil der Partitionierung auf einer Ebene ist die geringere Laufzeit, als die der hierarchischen Verfahren. Zudem ist das Clustering-Resultat und die Laufzeit des Verfahrens von der Auswahl der Zentroide abhängig. Um diesen Faktoren entgegenzuwirken, werden partitionierende Clustering-Verfahren meist mehrmals mit verschiedenen Positionen der Zentroide ausgeführt und mittels Metriken wird entschieden, welches dieser Clustering-Resultate am geeignetsten ist. Wie bei hierarchischen Clustering-Verfahren wird auch bei partitionierenden die Zuweisung der Punkte zu Clustern durch die Abstände der Punkte zueinander bestimmt [JMF99].

Eine weitere Gruppe außerhalb der Unterteilung von Kaufman und Rousseeuw stellen die **dichtebasierten Verfahren** dar [KR09]. Die Idee hinter diesen Verfahren liegt in der Wahrnehmung von Datenpunkten und Clustern. Datenpunkte werden intuitiv als Cluster wahrgenommen, wenn die Dichte der darin enthaltenen Datenpunkte höher ist als die der Datenpunkte außerhalb dieses Clusters [EKS+96]. Demnach werden in diesem Verfahren nur Datenpunkte geclustert, welche in einem vorgegebenen Abstand mindestens eine festgelegte Anzahl an Datenpunkten besitzen. Datenpunkte, welche diese Eigenschaft nicht erfüllen, werden nicht geclustert, da diese als Rauschen betrachtet werden. Dabei kann die Distanzfunktion zur Bestimmung des Abstands frei gewählt werden, wodurch dieser Ansatz ebenfalls für hochdimensionale Datensätze genutzt werden kann [EKS+96].

Eines der am häufigsten verwendeten Clustering-Verfahren ist das partitionierende Verfahren k-Means [JMF99]. Die Gründe dafür sind sowohl eine einfache Implementierung als auch eine geringe Laufzeit. Diese liegt mit p als Anzahl an Datenpunkten, c als die Anzahl an Clustern und t als die Anzahl an Iterationen in $O(p \cdot c \cdot t)$ [FBS19]. Der konzeptionelle Ablauf von k-Means lässt sich beschreiben durch [Llo82] [Mac+67]:

1. generiere k Zentroide $z_1, z_2, \cdots, z_k$
2. füge alle Punkte p zum nächsten Cluster CL hinzu
3. Verschiebe alle z_i, sodass sie den Mittelpunkt der Cluster CL_i darstellen
4. Falls ein z_i verschoben wurde, springe zu Schritt 2

In der Praxis wird k-Means meist öfter ausgeführt und die Summe der mittleren quadratischen Abweichungen als Metrik verwendet, um herauszufinden, welches Resultat am geeignetsten ist. Der Nachteil dabei ist, dass dieses Verfahren nur lokal konvergiert. Dies hat zur Folge, dass das Ergebnis potenziell nicht optimal ist, wodurch die Interpretierbarkeit des Clustering-Resultats zusätzlich erschwert wird [JMF99].

Durch die automatisierte Zuweisung von Daten zu dynamisch generierten Gruppen ist es meist schwer, als Nutzer diese Cluster zu interpretieren. Zudem fehlt häufig ein tiefer Einblick in die Daten, wodurch eine weitere Verarbeitung dieser Cluster von Menschen ohne weitere Werkzeuge sehr aufwendig ist [Rou87].

2.2 Ähnlichkeit von Clustern

In diesem Abschnitt werden Metriken vorgestellt, welche es ermöglichen Cluster von verschiedenen Clustering-Resultaten zu vergleichen. Dabei wird davon ausgegangen, dass der Datensatz mindestens zweimal separat geclustert wurde. Die Schwierigkeit besteht darin, dass keine Labels existieren, nach denen die Daten gruppiert werden. Stattdessen wird die Cluster-Zugehörigkeit der Datenpunkte, wie in Abschnitt 2.1 erwähnt, meist durch verschiedene Distanzmetriken bestimmt [VEB10].

Um dennoch einen Vergleich von Clustern mit nicht einheitlichen Gruppierungen durchführen zu können, sind demnach viele Vergleiche nötig. In diesem Abschnitt werden die drei Ansätze Jaccard Similarity Score, Rand Info Score und Mutual Information Score näher betrachtet. Weitere Ansätze finden sich bei Vinh et al. [VEB10].

Der erste Ansatz des Jaccard Similarity Scores liefert ein einfaches Verfahren, zwei Cluster C_i und C_j miteinander zu vergleichen [Lee00]:

$$JaccardSimilarityScore(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \tag{2.1}$$

Demnach gibt diese Gleichung an, wie viele Datenpunkte sowohl in Cluster C_i als auch in C_j sind. Normalisiert wird dies durch die Vereinigung der Cluster, wodurch der Jaccard Similarity Score im Bereich zwischen [0,1] liegt [Lee00].

Neben der Möglichkeit einzelne Cluster zu vergleichen, gibt es auch die Möglichkeit das Clustering als Gesamtes zu betrachten. Der zweite Ansatz basiert dabei auf der Betrachtung des Datensatzes als eine Menge aus Datenpunkt-Paaren. Anschließend wird betrachtet, wie die einzelnen Datenpunkte der Paare verteilt sind. Für einen Datensatz D mit p Datenpunkten ergeben sich daraus $\binom{p}{2}$ Paare. Nun werden diese Paare für die verschiedenen Clustering-Resultate R_1 und R_2 verglichen. Dabei stellen R_1 und R_2 Mengen von Clustern des Datensatzes D dar. Die folgenden Variablen geben die Summen der vier möglichen Fälle an:

- S_{00} : Summe an Paaren, die in R_1 und R_2 jeweils in verschiedenen Clustern sind
- S_{10} : Summe an Paaren, die in R_1 in verschiedenen und in R_2 im gleichen Cluster sind
- S_{01} : Summe an Paaren, die in R_2 in verschiedenen und in R_1 im gleichen Cluster sind
- S_{11} : Summe an Paaren, die in R_1 und R_2 jeweils im gleichen Cluster sind

Aus einer Kombination der paarweisen Vergleiche bildet sich der Rand Info Score [VEB10]:

$$RandInfoScore(R_1, R_2) = \frac{S_{00} + S_{11}}{\binom{p}{2}} \tag{2.2}$$

Dieser setzt das Maß der Übereinstimmungen der Clusterings über die Verteilung der Paare ins Verhältnis zur Anzahl der Paare. Somit ergeben sich auch hierfür Werte im Bereich zwischen [0,1]. Da der Rand Info Score jedoch nicht betrachtet wie wahrscheinlich es ist, dass ein Paar jeweils im selben oder in getrennten Clustern landet, sorgt dies dafür, dass in der Praxis nur selten Werte unter 0.5 erreicht werden. In der Praxis wird meist demnach der Adjusted Rand Info Score verwendet [VEB10]:

$$AdjustedRandInfoScore(R_1, R_2) = \frac{2(S_{00}S_{11} - S_{01}S_{10})}{(S_{00} + S_{01})(S_{01} + S_{11}) + (S_{00} + S_{10})(S_{10} + S_{11})} \tag{2.3}$$

Für diese Verfeinerung des Rand Info Scores kann gezeigt werden, dass er genau dann den Wert 0 annimmt, wenn der Rand Index Score dem Erwartungswert entspricht [VEB10].

Der dritte Ansatz, welcher ebenfalls komplette Clustering-Resultate vergleicht, stammt aus der Informationstheorie und heißt Mutual Information Score. Dieser errechnet sich aus den folgenden Größen:

$$s_{ij} = |C_i \cap C_j| \quad \text{mit } C_i \in R_1, C_j \in R_2 \tag{2.4}$$

$$a_k = \sum_{l=1}^{|R_2|} s_{kl} \tag{2.5}$$

$$b_k = \sum_{l=1}^{|R_1|} s_{lk} \tag{2.6}$$

$$MutualInfoScore(R_1, R_2) = \sum_{i=1}^{|R_1|} \sum_{i=1}^{|R_2|} \frac{s_{ij}}{p} log(\frac{s_{ij}/p}{a_i b_j / p^2}) \tag{2.7}$$

Die einzelnen Komponenten des Mutual Information Scores sind dabei auch als Auftrittswahrscheinlichkeiten und bedingte Wahrscheinlichkeiten der Datenpunkte in bestimmten Clustern interpretierbar. Das Konzept dieser Metrik beruht auf dem der Entropie und ermöglicht eine Aussage darüber, in welchem Maße es hilft, eines der Clusterings zu kennen, um das andere vorherzusagen. Allgemein ausgedrückt misst der Mutual Information Score damit, wie viele Informationen in beiden Clusterings übereinstimmen [VEB10][CH90].

Für die weitere Arbeit wird der Adjusted Rand Info Score benutzt.

2.3 Wahrnehmung des Menschen

Dieser Abschnitt enthält grundlegende Informationen über die Wahrnehmung des Menschen. Dafür wird betrachtet, wie viele Informationen ein Mensch gleichzeitig wahrnehmen kann. Zudem wird die Darstellung von Daten als Wortwolke genauere betrachtet.

2.3.1 Wahrnehmbare Informationen

Ein allgemeiner Ansatz, die für den Menschen wahrnehmbare Menge zu beschreiben besteht daraus, die Menge an „Information“ zu betrachten. Eine Information ist dabei eine Verallgemeinerung des Wahrgenommenen und kann somit verschiedene Ausprägungen haben, wie zum Beispiel Geräusch, Geschmack oder optische Reize. Studien zeigten, dass die wahrgenommene Information zunächst linear zu der Anzahl an vorhandenen Informationen steigt, bis ein gewisser Schwellenwert erreicht ist. Wird dieser Schwellenwert überschritten, so nähert sich der Wert der wahrgenommenen Informationen einer Asymptote an. Dies bedeutet, dass Menschen nur eine begrenzte Anzahl an Reizen gleichzeitig wahrnehmen können [Mil56].

Im Folgenden wird eine Studie vorgestellt, welche dieses Phänomen für tonale Reize zeigt. Dieses Beispiel wurde gewählt, da es repräsentativ für alle Arten von Reizen ist und von weniger Faktoren abhängt wie beispielsweise visuelle Reize. So wurden beispielsweise Probanden verschiedene Töne im Bereich zwischen 100 und 8000 Hz vorgespielt und sollten anschließend von ihnen erkannt werden. Gemessen wurden die Ergebnisse für bis zu 14 verschiedene Töne. Die Ergebnisse der Studie sind Abbildung 2.3 zu entnehmen. Die horizontale Achse des Graphen entspricht dem tatsächlichen und die vertikale Achse dem wahrgenommenen Informationsgehalt. Der Informationsgehalt ist dabei der Zweierlogarithmus der Anzahl an Tönen. Demnach wurden bis zu vier verschiedene Töne fehlerfrei erkannt. Maximal wahrgenommen wurden von den Probanden sechs verschiedene Töne [Pol52]. Daraus geht hervor, dass ein Mensch nur maximal sechs unterschiedliche Töne richtig bewerten kann. Nach Miller lassen sich diese Erkenntnisse grundsätzlich auch auf visuelle Reize übertragen. Weiterhin lässt sich feststellen, dass für Menschen die Anzahl jeglicher wahrnehmbarer Reize zwischen fünf und neun liegt [Mil56].

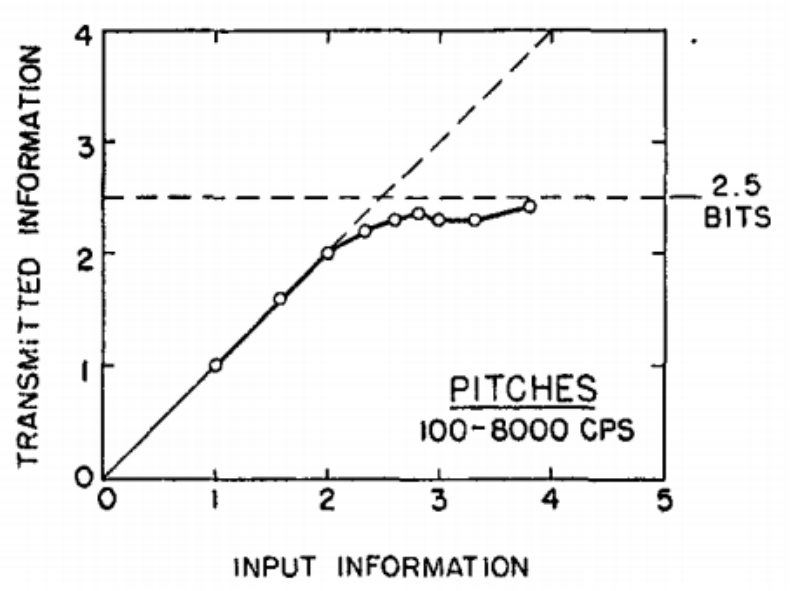


Abbildung 2.3: Der Verlauf der wahrgenommenen Töne in Abhängigkeit der tatsächlich gespielten Töne [Pol52]

2.3.2 Wortwolken

Wortwolken sind eine häufige Darstellungsform für Informationen im Internet [LZT09]. Dabei gibt es verschiedene Varianten dieser Darstellung, welche mit verschiedenen Parametern arbeiten, wie zum Beispiel Schriftgröße, Schriftfarbe oder der Abstand der Wörter. In einer Studie wurde untersucht, wie sich diese verschiedenen Parameter von Wortwolken auf die Wahrnehmung des Menschen auswirken. Dafür wurden Probanden verschiedene Wortwolken-Typen mit verschiedenen Parametern präsentiert. Eine in der Studie verwendete Wortwolke ist in Abbildung 2.4 zu sehen. Generiert wurden diese Wortwolken zufällig und besaßen zwischen 139 und 147 Wörter pro Wortwolke. In diesen sollten sie jeweils die zehn, für sie am wichtigsten erscheinenden Wörter, heraussuchen. Dabei sollten sie sich auf das Erscheinungsbild der Wortwolke und Wörter konzentrieren und nicht auf die Bedeutung des Worts. Um die Probanden nicht zu beeinflussen, wurde der Begriff der Wichtigkeit nicht genauer erläutert. Nach dieser Studie wurde die Auswahl der Wörter am meisten durch die Schriftgröße beeinflusst. Dabei wurden bereits kleine Unterschiede zwischen den Schriftgrößen wahrgenommen. Demnach kann mittels der Schriftgröße eine große Bandbreite an unterschiedlicher Relevanz oder Häufigkeit dargestellt werden. Zudem wurde auch die Schriftdicke häufig von den Probanden als Auswahlkriterium genutzt. Dies könnte beispielsweise genutzt werden, um einzelnen Wörtern aussagekräftiger darzustellen als andere [BGN08].

Abbildung 2.4: Beispielhafte Wortwolke der vorgestellten Studie [BGN08]

2.4 Streuungsmaße

Streuungsmaße liefern eine Metrik zur Charakterisierung von Stichproben. Aus diesen Werten lässt sich ablesen, wie stark die Werte der betrachteten Menge variieren. Im Allgemeinen wird eine Metrik als Streuungsmaß S bezeichnet auf einer Stichprobe X, wenn sie die folgenden Eigenschaften erfüllt [Von93]:

i. $\forall\, x_i, x_j \in X : x_i = x_j \Rightarrow S(X) = 0$

ii. $\exists\, x_i, x_j \in X : x_i \neq x_j \Rightarrow S(X) > 0$

iii. Invariant gegenüber konstanter Verschiebung aller Elemente

iv. X' entsteht aus X durch die Ersetzung eines Elements x_i durch x_j wobei die Summe der Abweichungen von x_j größer ist als die Summe der Abweichungen von $x_i \Rightarrow S(X') \geq S(X)$

Aus Eigenschaft ii lässt sich ableiten, dass Streuungsmaße nur Abstände oder mittlere Abweichungen messen, jedoch keine Richtungen. Eigenschaft 3 beschreibt die Unabhängigkeit der Streuungsmaße gegenüber ihrer Lage im Koordinatensystem. Dies gilt jedoch nicht für Veränderungen der Abstände

zwischen einzelnen Elementen. Die Charakteristik der Streuungsmaße als Maß für den Abstand zum Mittelwert oder der Abstände der Elemente wird in Eigenschaft 4 deutlich. Um die Abstände zwischen den Elementen messen zu können, ist zudem eine metrische Skala erforderlich [Von93][EKT08].

Streuungsmaße werden unterschieden in die zwei Klassen:

1. Abstand zweier Kennzahlen
2. Abstand der Elemente von Referenzmaßen

In den folgenden Abschnitten werden diese zwei Klassen der Streuungsmaße genauer betrachtet.

2.4.1 Abstand zweier Kennzahlen

Für Klasse 1 werden für gewöhnlich Ordnungszahlen gewählt. Dabei stellt das einfachste Beispiel die Spannweite dar. Diese wird berechnet durch

$$Spannweite(X) = max(X) - min(X) \tag{2.8}$$

Durch die Betrachtung des Minimums und Maximums ist dieses Streuungsmaß sehr anfällig für Ausreißer [EKT08]. Ein weiteres Beispiel für Klasse 1 stellt der Quartilsabstand dar. Berechnet wird dieser durch

$$Quartilsabstand(X) = x_{0.75} - x_{0.25} \tag{2.9}$$

Dabei ist x_i der Wert der Stichprobe, sodass $i \cdot 100\%$ der Werte der Stichprobe kleiner oder gleich x_i sind. Nach dieser Definition gibt der Quartilsabstand den Bereich an, in dem sich die mittleren 50% der geordneten Stichprobe befinden. Dies sorgt dafür, dass der Quartilsabstand eine hohe Resistenz gegen Ausreißer liefert. Häufig wird der Quartilsabstand mit dem Median ins Verhältnis gesetzt, was zu der Definition des Quartilsdispersionskoeffizients führt [EKT08]:

$$Quartilsdispersionskoeffizient(X) = \frac{x_{0.75} - x_{0.25}}{Median(X)} \tag{2.10}$$

2.4.2 Abstand der Elemente von Referenzmaßen

Für Klasse 2 der Streuungsmaße wird der direkte Abstand der Elemente der Stichprobe zu einer Referenz gemessen. Ein Beispiel dieser Klasse stellt die Varianz dar [EKT08]:

$$Varianz(X) = (\sum_{i=1}^{|X|} \frac{(x_i - Durchschnitt(X))^2}{|X|}) \tag{2.11}$$

Die Varianz ist eine besonders aussagekräftige Metrik für die Streuung der Elemente einer Stichprobe, da sie die Abstände der Elemente zum arithmetischen Mittel quadriert aufsummiert. Durch die Quadrierung wird gewährleistet, dass positive und negative Abstände dieselbe Auswirkung

auf das Ergebnis haben und sich nicht gegenseitig aufheben. Die Quadrierung sorgt aber auch für eine schwerere Interpretierbarkeit der Ergebnisse, da auch die Einheit der Daten in quadrierter Form vorliegt. Um dies zu verhindern, kann stattdessen die Standardabweichung betrachtet werden [EKT08]:

$$Standardabweichung(X) = \sqrt{Varianz(X)} \tag{2.12}$$

Diese kehrt den Prozess der Quadrierung der Einheit um und sorgt für intuitivere Ergebnisse. Wird diese Größe mit dem Mittelwert ins Verhältnis gesetzt, so ergibt sich daraus der Variationskoeffizient [EKT08]:

$$Variationskoeffizient(X) = \frac{Standardabweichung(X)}{Durchschnitt(X)} \tag{2.13}$$

Neben der mittleren Abweichung vom arithmetischen Mittel kann auch beispielsweise die Abweichung vom Median betrachtet werden:

$$Medianabweichung(X) = (\sum_{i=1}^{|X|} \frac{|x_i - Median(X)|}{|X|}) \tag{2.14}$$

2.4.3 Laufzeit

Der lineare Zusammenhang zwischen der Anzahl an Datenpunkten und der Laufzeit der Metriken geht für die Varianz, Standardabweichung und Variationskoeffizient direkt aus den Formeln 2.11 und 2.12 hervor. Um die Laufzeitkomplexität der Medianabweichung und des Quartilsdispersionskoeffizienten zu bestimmen, ist es nötig herauszufinden, welche Laufzeitkomplexität das Finden des k-ten Elements einer sortierten Liste hat. Dies ist eine Verallgemeinerung der Probleme, um den Median oder das untere oder obere Quartil zu finden und schätzt damit beide Probleme nach oben ab. Dabei ist zu beachten, dass die Eingabeliste nicht zwangsläufig bereits in sortierter Form vorliegt. Dies ist mit Hilfe des SELECT-Algorithmus von Floyd und Rivest's [BFP+73] ebenfalls mit p als Anzahl der Elemente der Liste in $O(p)$ möglich.

3 Verwandte Arbeiten

Die Visualisierung von Clustering-Resultaten ist über verschiedene Ansätze möglich. So ist eine Darstellung von Clustern in zweidimensionalen Räumen durch das Anzeigen der Datenpunkte in einem Koordinatensystem einfach realisierbar. Im dreidimensionalen Raum ist dies ebenfalls noch verständlich möglich, sofern die Betrachtungsrichtung geändert werden kann. In diesen Fällen sind anhand der Achsen die Wertebereiche einzelner Features in den Clustern ablesbar.

Für höherdimensionale Räume ist eine Darstellung von Clustering-Resultaten in Koordinatensystemen ohne weitere Verarbeitung für Menschen nicht verständlich. In diesen Fällen bietet sich eine Dimensionsreduktion auf zwei oder drei Dimensionen an.

Eine weitere Möglichkeit Clustering-Resultate aus hochdimensionalen Räumen darzustellen liefert ein textueller Ansatz. Hierbei werden aus den Daten der Cluster spezielle Werte errechnet, welche dann dem Nutzer angezeigt werden können.

3.1 Dimensionsreduktion

Zunächst wird der Ansatz der Dimensionsreduktion anhand der Verfahren „Principal Component Analysis“ und „t-Distributed Stochastic Neighbor Embedding“ genauer betrachtet.

3.1.1 Principal Component Analysis

Principal Component Analysis (PCA) liefert für die Eingabe eines multidimensionalen Datensatzes eine Ausgabe mit weniger Dimensionen. Die Ziele von PCA sind unter anderem die Vereinfachung, Reduzierung, Modellierung und Klassifizierung der Daten. In der Praxis wird häufig eine Dimensionsreduktion des multidimensionalen Raums auf einen zwei- oder drei-dimensionalen gewählt. Durch den Ansatz der Dimensionsreduktion wird es ermöglicht, große Datensätze mittels PCA zu visualisieren [WEG87]. Diese Visualisierung enthält häufig Gruppierungen der Datenpunkte aus dem Datensatz. Durch diese Gruppierung ist es möglich Ausreißer zu erkennen, da diese zu keiner der Gruppen zugeordnet werden und somit einen sichtbaren Abstand zu allen Gruppen aufweisen. Ein Beispiel für eine Visualisierung eines PCA-Resultats ist Abbildung 3.1 zu entnehmen. Zu sehen ist darin eine Unterteilung des Datensatzes in zwei Gruppen, dargestellt durch Kreise und Dreiecke. Der Datenpunkt mit der Nummer 7 stellt einen Ausreißer dar, da dieser einen großen Abstand zu den anderen Datenpunkten aufweist [Olk02].

Der Nachteil ist dabei die Laufzeitkomplexität der Berechnung der Dimensionsreduktion. Diese wird durch das Lösen eines Gleichungssystems berechnet, mit f als die Anzahl an Features und p als die Anzahl an Datenpunkten, in $O(f^2 \cdot p)$, und die Berechnung der Zerlegung der Eigenwerte in $O(f^3)$. Damit liegt die Gesamtlaufzeit von PCA in $O(f^2 p + f^3)$ [ACLS12].

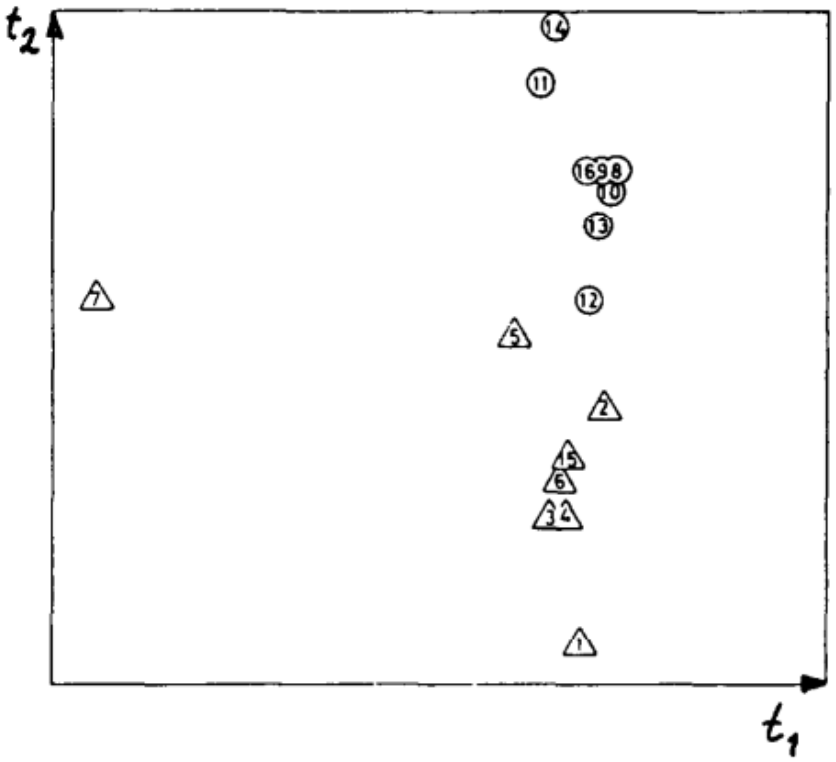


Abbildung 3.1: Visualisierung eines Datensatzes nach der Dimensionsreduktion mittels PCA [WEG87]

Ein weiterer Nachteil der Visualisierung von PCA ist die Dimensionsreduktion. Vor allem für hochdimensionale Daten kann eine Reduktion auf zwei oder drei Dimensionen eine zu starke Einschränkung darstellen. So können durch die starke Begrenzung der Dimensionen potenziell Abhängigkeiten der Features zueinander nicht visualisiert werden. Dies kann dazu führen, dass Visualisierungen entstehen, bei denen keine Gruppierungen der Daten zu erkennen sind [May08].

3.1.2 t-Distributed Stochastic Neighbor Embedding

„t-Distributed Stochastic Neighbor Embedding“ (t-SNE) ist eine Abwandlung des „Stochastic Neighbor Embedding“ Algorithmus (SNE) von Hinton und Roweis aus dem Jahr 2002 „Stochastic neighbor embedding“ [MH08]. Dieser Algorithmus liefert für einen hochdimensionalen Datensatz als Eingabe, eine visuelle Ausgabe je nach Präferenz in zwei oder drei Dimensionen. Im Gegensatz zu SNE ist t-SNE einfacher zu optimieren und liefert bessere visuelle Ausgaben. In Abbildung 3.2 ist als Beispiel eine Visualisierung von handgeschriebenen Zahlen des MNIST Datensatzes[1] mittels t-SNE dargestellt. Die Zahlen sind in diesem Datensatz als 28 mal 28 Pixel große Bilder gespeichert, wobei die Pixel den Dimensionen des Datensatzes entsprechen. Demnach reduziert t-SNE in diesem Beispiel 784 Dimensionen auf 2 [MH08].

[1] http://yann.lecun.com/exdb/mnist/index.html

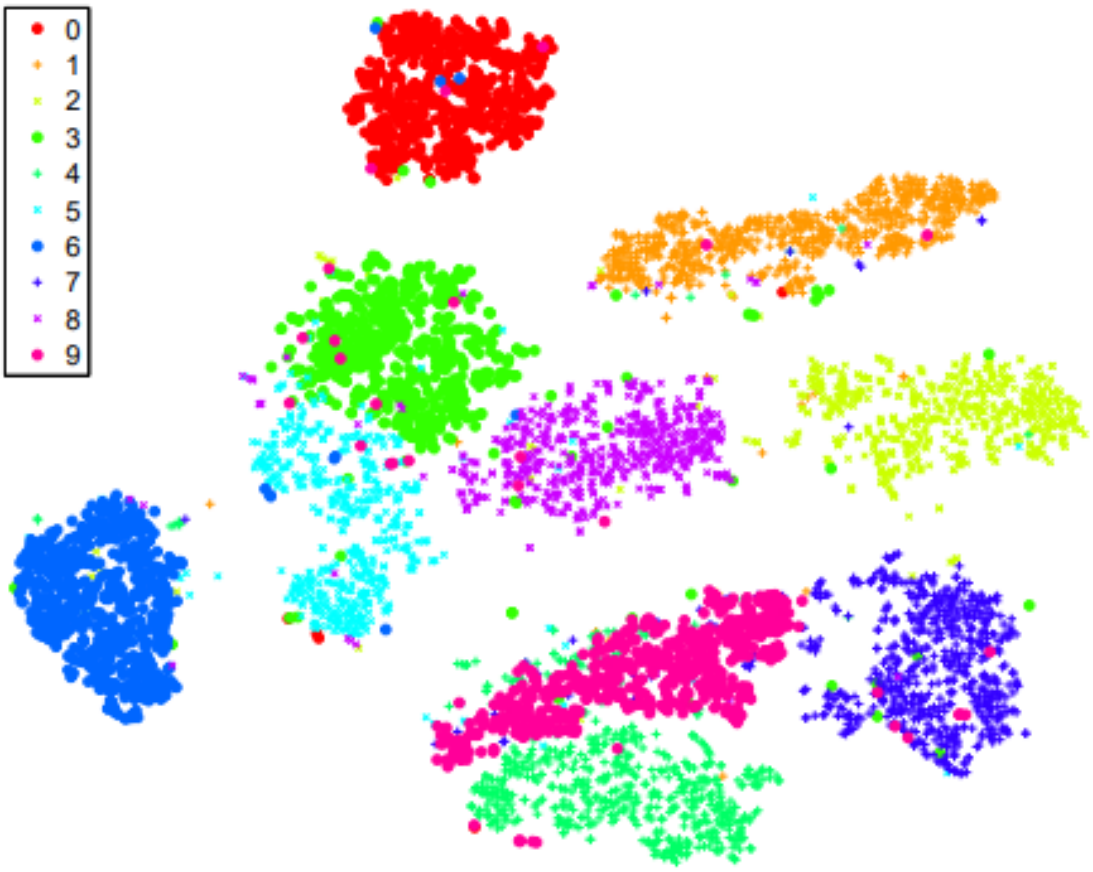


Abbildung 3.2: Visualisierung des MNIST Datensatzes nach der Dimensionsreduktion mittels t-SNE [MH08]

Kritisch zu betrachten ist die Ausgabe von t-SNE. Sie liefert visuell einen Zusammenhang der Daten, jedoch ist es schwierig genauere Informationen herauszulesen. Wie gut in Abbildung 3.2 zu sehen ist, werden die unterschiedlichen Zahlen durch verschiedene Gruppen gut ersichtlich. Jedoch ist es zum Beispiel nur schwer bis gar nicht möglich herauszufinden, welche Features maßgeblich zu der zugrunde liegenden Gruppierung geführt haben. Somit ist es schwierig die Ausgabe von t-SNE oder die Cluster, welche der Algorithmus als Eingabe bekommen hat, zu bewerten und weiter zu verarbeiten.

Ein weiterer Nachteil von t-SNE ist die Laufzeit. Wird der Standard-Algorithmus verwendet, so hat dieser eine Laufzeit in $O(p^2)$, mit p als die Anzahl an Datenpunkten [MH08]. Zudem kann gezeigt werden, dass t-SNE nicht für jede Eingabe gegen ein globales Optimum konvergiert. Somit kann es passieren, dass für nur geringfügig andere Eingabedaten stark unterschiedliche Ausgaben erzeugt werden [MH08].

3.1.3 Zusammenfassung

Im Allgemeinen liefern PCA und t-SNE Verfahren zur Dimensionsreduktion von hochdimensionalen Datensätzen. Um eine Visualisierung der Ergebnisse zu ermöglichen, werden die Dimensionen meist auf zwei oder drei Dimensionen reduziert. Dies liefert häufig eine sichtbare Separierung in einzelne Gruppen, wie beispielsweise in Abbildung 3.2 zu sehen. Die Separierung entsteht dabei durch die verschiedenen Abstände der Datenpunkte in verschiedenen Dimensionen. Dieses entspricht genau der Zuweisung von Datenpunkten zu Clustern (vgl. 2.1). Somit ist PCA und t-SNE auch als eine Visualisierung von Clustering-Resultaten interpretierbar. Der Nachteil dieser Verfahren ist die Kommunikation gegenüber dem Nutzer. Visuell werden die Cluster ersichtlich, doch genaue Informationen zu den Features oder Datenpunkten sind in diesen Verfahren nicht erkennbar. Zudem besitzen diese Verfahren eine lange Laufzeit, wodurch diese für große Datensätze nicht praktikabel eingesetzt werden können.

3.2 Data-Mining-Werkzeuge

Im folgenden Abschnitt wird der Ansatz der textuellen Beschreibung von Clustering-Resultaten betrachtet. Demnach werden die Cluster und Datenpunkte anhand von Statistiken und Kenngrößen anstatt in einem Koordinatensystem dargestellt. Dies wird häufig von Mining-Werkzeugen zusätzlich angeboten, um dem Nutzer eine detaillierte Auskunft über die Daten zu ermöglichen. Vorgestellt wird für diesen Ansatz zunächst Waikato Environment for Knowledge Analysis (Weka), als ein Beispiel für ein häufig verwendetes Data Mining Werkzeug. Anschließend wird der IBM Clustering Visualizer betrachtet, welcher eine detaillierte Visualisierung von Cluster-Resultaten liefert.

3.2.1 Waikato Environment for Knowledge Analysis

Waikato Environment for Knowledge Analysis (Weka)[2] ist ein häufig verwendetes Machine-Learning-Werkzeug, welches verschiedene Aufgaben wie Data-Mining, Klassifizierung, Clustering und Visualisierung vereint [HFH+09]. Auf der Seite des Clusterings ermöglicht Weka die Ausführung von Clustering-Algorithmen, wie zum Beispiel k-Means. Für die Clustering-Resultate werden anstatt einer Visualisierung der Datenpunkte in Koordinatensystemen Statistiken errechnet und dargestellt. Abbildung 3.3 zeigt ein Resultat von Weka auf einem Beispieldatensatz mit fünf Features a0 bis a5 und 100 Datenpunkten. Wie in der obersten Zeile zu sehen ist, wurde der Datensatz in drei Cluster unterteilt. Falls im Datensatz zu einem Teil der Daten bereits ein Clustering angegeben ist, so geben die weiteren Einträge bei den Features an, wie viele dieser Datenpunkte diesem Cluster zugewiesen wurden. Die unteren Einträge aus Abbildung 3.3 geben an, wie die prozentuale Verteilung der Datenpunkte auf die Cluster ist. Zudem bietet Weka eine Option, die aussagekräftigsten Features für den Datensatz zu finden. Dies wird erreicht, indem zu Beginn ein Feature als Referenzklasse gewählt und anschließend die restlichen Features mittels Machine-Learning-Verfahren bewertet werden. Möchte man dies auch für das Clustering anwenden, so muss es zunächst manuell vor dem Clustering eingestellt werden [HFH+09].

[2] https://www.cs.waikato.ac.nz/ml/weka/

```
                Cluster
Attribute         0       1       2
                (0.26)  (0.38)  (0.36)
==================================
a0
  false       25.9344 10.1092  9.9564
  true         1.8053 29.8574 28.3373
  [total]     27.7398 39.9666 38.2937
a1
  false        9.4211 14.0859 22.4929
  true        18.3186 25.8806 15.8007
  [total]     27.7398 39.9666 38.2937
a2
  false       25.3279 16.0219 12.6502
  true         2.4119 23.9447 25.6434
  [total]     27.7398 39.9666 38.2937
a3
  false       12.7091 23.0927 20.1981
  true        15.0307 16.8738 18.0955
  [total]     27.7398 39.9666 38.2937
a4
  false       15.6077 38.9327  1.4595
  true         12.132  1.0338 36.8341
  [total]     27.7398 39.9666 38.2937
class
  c0          26.4735  1.3075 37.2189
  c1           1.1364 22.8215  1.0421
  c2           1.1299 16.8375  1.0327
  [total]     28.7398 40.9666 39.2937


Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       28 ( 28%)
1       38 ( 38%)
2       34 ( 34%)
```

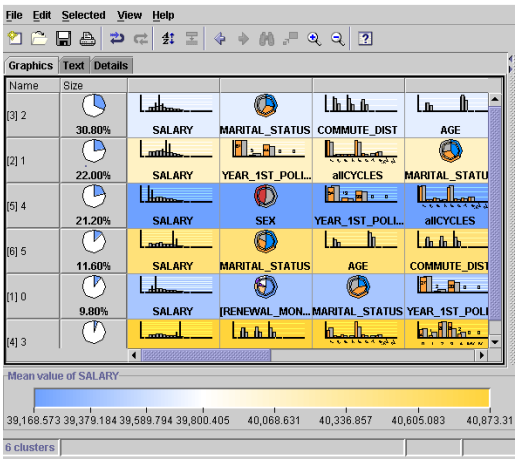
Abbildung 3.3: Eine beispielhafte Angabe für ein Clustering-Resultat in Weka

Im Allgemeinen liefert Weka ein vielseitiges Data Mining Werkzeug, welches auch Aufgaben im Bereich Clustering durchführen kann. Somit kann beispielsweise ermittelt werden, wie die Datenpunkte auf die einzelnen Cluster verteilt sind. Dabei liefert WEKA im Allgemeinen eine Beschreibung des Modells und keine genaue Darstellung des Clustering-Resultats.

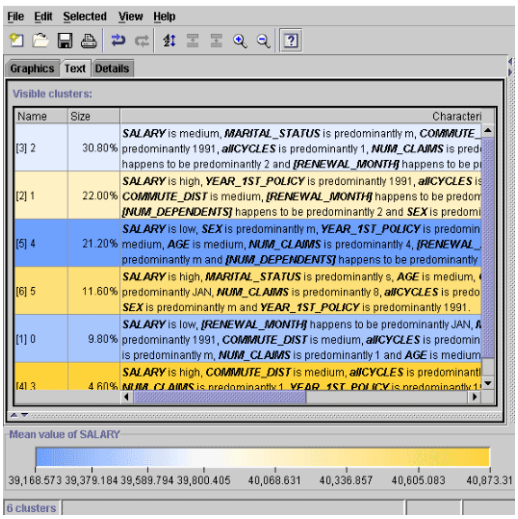
3.2.2 IBM Clustering Visualizer

Der IBM Clustering Visualizer[3] ist ein Unterprogramm des IBM Intelligent Miners. Er ist ein Werkzeug zur Kommunikation von Clustering-Resultaten und bietet drei verschiedene Anwendungen: *Graphics View*, *Text View* und *Details View*. Erläutert werden die einzelnen Anwendungen anhand des Beispiels aus Abbildung 3.4. Dieses zeigt die Oberfläche des IBM Clustering Visualizer für einen Datensatz von Personendaten mit Features wie Gehalt, Geschlecht oder Alter.

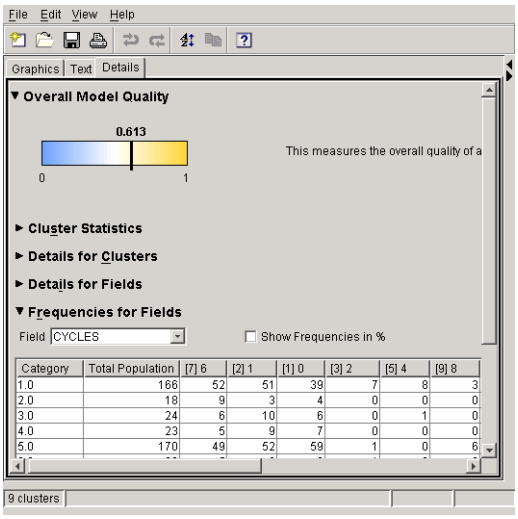
[3] https://www.ibm.com/support/knowledgecenter/en/SSEPGG_11.1.0/com.ibm.im.visual.doc/c_introducing_the_clusteirng_visualizer.html



(a) Graphics View



(b) Text View



(c) Detail View

Abbildung 3.4: Die verschiedenen Ansichten des IBM Clustering Visualizer [Cen]

Die Graphics View (Abbildung 3.4a) liefert eine Visualisierung mittels verschiedener Diagramme für jedes Cluster, wobei jedes Cluster durch eine Zeile repräsentiert wird. Aus diesen lassen sich Tendenzen der Verteilungen einzelner Features innerhalb dieses Clusters identifizieren. Zudem gibt die Hintergrundfarbe einer Zeile einen Mittelwert eines gewählten Features wieder. In dem Beispiel aus Abbildung 3.4a bedeutet zum Beispiel eine dunkelblaue Färbung der Zeile ein Jahresgehalt von etwa 39.000 $ und eine stark gelbe Färbung ein Jahresgehalt von etwa 41.000 $.

Die Text View (Abbildung 3.4b) bietet eine Angabe von Tendenzen der einzelnen Features innerhalb eines Clusters mittels Schlüsselwörter wie „medium“ oder „high“. Wie bei der Graphics View wird auch hier jedes Cluster in einer Zeile repräsentiert und die Hintergrundfarbe der Zeile gibt Auskunft über den Mittelwert eines Features. Für jedes Cluster werden 15 Features mit Name und Tendenz dargestellt. Welche Features dargestellt werden, richtet sich nach der Auswahl des „field-sorting modes“. Dieser kann in den Sortiereinstellungen geändert werden. Bei der Angabe der Tendenzen unterscheidet das Werkzeug zwischen Kategorischen-, kontinuierlich Numerischen- und Diskret Numerischen-Daten. Bei den kategorischen Daten wird dem entsprechenden Feature

der Wert zugewiesen, welcher am Häufigsten vertreten ist in diesem Cluster. Ein Beispiel hierfür ist Abbildung 3.4b „Marital_Status is predominantly m“. Die Kontinuierlich Numerischen-Daten werden auf die drei Worte „low“, „medium“ und „high“ abgebildet, wie zum Beispiel „Salary is medium“ wie in Spalte 2 der Tabelle 3.4 zu sehen. Für diskrete Numerische-Daten wird das Feature auf einen einzelnen String abgebildet. Dieser String ist der Wert, welcher am häufigsten in dem Feature in dem Cluster vorkommt. In Abbildung 3.4b ist ein Beispiel hierfür die Angabe „Year_1st_Policy is predominantly 1991“.

Die Details View liefert eine genaue Angabe verschiedener Werte eines einzelnen Clusters. Wie in Abbildung 3.4c zu sehen, liefert diese Ansicht einige Statistiken für das Cluster und dessen einzelner Features. Dabei müssen die gewünschten Metriken und Anzeigeoptionen vom Nutzer selber eingestellt werden. Zudem ist diese Übersicht auf ein Cluster begrenzt [Cen].

3.2.3 Zusammenfassung

WEKA und IBM Clustering Visualizer stellen zwei Beispiele für Data-Mining-Werkzeuge dar, welche eine textuelle Beschreibung von Clustering-Resultaten liefern. WEKA liefert eine Beschreibung des Modells und nur wenige Angaben zu einzelnen Clustern. Der IBM Clustering Visualizer ermöglicht die Angabe verschiedener Graphen oder Statistiken zu Features innerhalb der Cluster. Diese Angaben sind generell auf einer groben Ebene gehalten, wie zum Beispiel „Salary is medium“ in Abbildung 3.4b. Genauere Angaben über Features und Cluster können nur gefunden werden, wenn diese manuell gesucht werden. Im Allgemeinen liefern diese Werkzeuge damit grobe Angaben zu Clustern, jedoch keine einfache Darstellung der relevanten Informationen für einen Analysten.

4 Konzept

Clustering-Verfahren stellen eine einfache Methode dar, große Datenmengen zu gruppieren. In diesem Kapitel wird ein Ansatz zur Darstellung für Clustering-Resultate entwickelt, welcher dem Nutzer Einblicke in die Cluster und deren Unterschied ermöglicht. Zudem wird dieser alle Clustering-Algorithmen unterstützen, da die Berechnungen auf den einzelnen Clustern stattfinden und somit nicht relevant ist, wie es zu diesen Clustering-Resultaten kam.

Um diesen Ansatz zu erläutern, wird zunächst die zugrunde liegende Problemstellung vorgestellt. Hierbei werden die Herausforderungen aufgezeigt, welche bewältigt werden müssen. Dabei wird zunächst darauf eingegangen, wie andere Ansätze zur Visualisierung von Clustering-Resultaten arbeiten und welche Vor- und Nachteile diese besitzen. Dies ermöglicht es, Verbesserungspotenzial zu identifizieren und im folgenden Ansatz umzusetzen.

4.1 Problemstellung

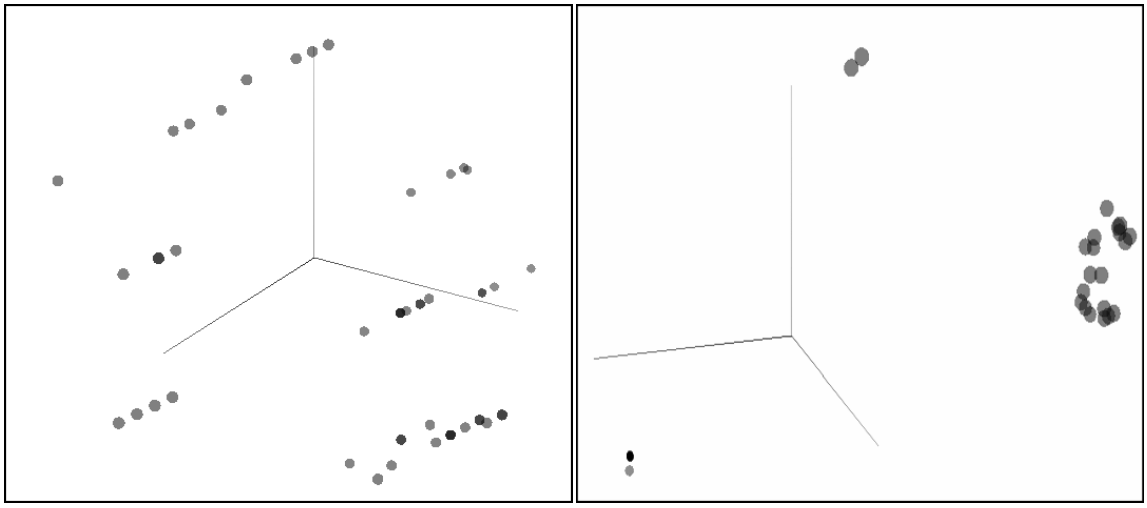
Durch die stetig steigende Menge an generierten und gespeicherten Daten wird deren Interpretierbarkeit durch Analysten immer schwieriger. Um dennoch Aussagen über die Daten treffen zu können, werden Data-Mining-Verfahren wie zum Beispiel Clustering eingesetzt. Clustering-Verfahren liefern eine Gruppierung der Daten und benötigen dafür kein Domänenwissen. Das Problem dabei ist, dass Clustering-Resultate von voluminösen Daten noch immer eine unüberschaubare Menge an Daten und Dimensionen pro Cluster besitzen. Demnach ist eine weitere Verarbeitung dieser Clustering-Resultate nötig, damit diese Daten für einen Analysten bewertbar sind.

Bisherige Verfahren zur Darstellung von Clustering-Resultaten basieren auf zwei verschiedenen Ansätzen:

Dimensionsreduktion:
Hochdimensionale Datensätze werden in diesem Ansatz durch eine Dimensionsreduktion für den Menschen verständlich in einem Koordinatensystem visualisiert.

textuelle Beschreibung:
Data-Mining-Werkzeuge verwenden eine textuelle Darstellung statistischer Metriken zur Repräsentation von Clustering-Resultaten.

Dimensionsreduktion: Zwei häufig verwendete Beispiele für diesen Ansatz sind PCA und t-SNE (vgl. Abschnitt 3.1). Diese liefern für die Eingabe eines hochdimensionalen Datensatz eine Ausgabe in zwei oder drei Dimensionen. Dabei entsprechen die Dimensionen der Ausgabe nicht einer Auswahl der Eingabedaten, sondern einer Kombination dieser. Daraus ergibt sich, dass die Achsen des resultierenden Graphen nur schwer interpretierbar sind. Dies führt dazu, dass, wie in Abbildung 4.1a und 4.1b zu sehen, eine Gruppierung der Daten erkennbar, ein genauer Einblick in



(a) PCA **(b)** t-SNE

Abbildung 4.1: Visualisierung eines Datensatzes mit PCA und t-SNE

den Wertebereich einzelner Gruppen aber nicht möglich ist. Zudem entsteht durch die komplexe Berechnung der Achsen eine quadratische Laufzeit der Algorithmen, wodurch die Anwendung auf großen Datensätzen sehr zeitintensiv ist (vgl. Abschnitt 3.1).

textuelle Beschreibung: Ein bekanntes Beispiel für diesen Ansatz ist WEKA. Dieser ermöglicht es, Statistiken des Datensatzes oder einzelner Cluster zu ermitteln und textuell darzustellen. Im Allgemeinen ist der Fokus von WEKA aber die Beschreibung des Modells und nicht die der einzelnen Cluster. Mehr Informationen über Cluster liefert der IBM Clustering Visualizer. Mit diesem Werkzeug sind kategorische Angaben zu Features innerhalb Clustern möglich, wie zum Beispiel „Salary is medium“. Ein Nachteil beider Werkzeuge ist die fehlende detaillierte Beschreibung und Darstellung der Cluster (vgl. Abschnitt 3.2).

Somit liefert weder die Dimensionsreduktion noch die textuelle Beschreibung eine detaillierte und einfach verständliche Visualisierung von Clustering-Resultaten. So werden in beiden Ansätzen entweder zu viele oder zu abstrakte Informationen dargestellt. Dies sorgt für eine schwere Interpretierbarkeit der Resultate. Demnach kann ein Analyst in diesen ohne weitere Verarbeitung nur aufwendig Informationen darüber erhalten, welche Features maßgeblich zu dem Clustering-Resultat geführt haben oder welche Wertebereiche die Features besitzen.

Szenario

In diesem Abschnitt wird diskutiert, welche Eingabedaten möglich sind und wie die gewünschten Resultate für den Analysten aussehen. Um Clustering-Resultate verstehen und damit weiterarbeiten zu können, sind im Allgemeinen folgende Informationen für den Analysten relevant:

I1 **Cluster-Beschreibung:**
Um Clustering-Resultate analysieren zu können, ist es wichtig, Informationen über einzelne Cluster zu erhalten. Somit ist erkennbar, welche Features in diesen Clustern aussagekräftig sind und welche Wertebereiche diese besitzen.

I2 **Datensatz-Beschreibung:**
Eine detaillierte Beschreibung des kompletten Datensatzes wird erreicht, indem für alle Cluster dieselben Features angezeigt werden. Ein Analyst erhält damit ein Überblick über das gesamte Clustering-Resultat und somit auch über den kompletten Datensatz.

I3 **Unterschiede der Cluster:**
Einen tieferen Einblick in das Clustering-Resultat wird durch die Betrachtung der Unterschiede der Cluster erreicht. Dies liefert Informationen darüber, welche Features potenziell zu der Separierung der Cluster geführt haben.

Dies wird unter anderem anhand eines Szenarios erläutert, welches auszugsweise Tabelle 4.1 zu entnehmen ist. Ein relevanter Faktor wird dabei die Aussagekraft der Features sein. Diese bestimmt, wie stark der Einfluss eines Features auf das Clustering-Resultat war.

Generell enthalten alle möglichen Eingabedaten bereits ein Clusterlabel für die Datenpunkte. Dies ist in der Tabelle an der ersten Stelle zu sehen. Um Berechnungen auf den restlichen Features und die Vergleichbarkeit von Clustern zu ermöglichen, wird im Folgenden von normierten numerischen Daten ausgegangen. Eine Normierung bedeutet in diesem Fall, dass der Wertebereich alle Features gleich ist. Dies stellt keine Einschränkung an die Datensätze dar, da dies für einen Wert x mit ursprünglichem Wertebereich $[a_{min}, a_{max}]$ und neuem Wertebereich $[b_{min}, b_{max}]$ beispielsweise durch folgende Formel errechnet werden kann:

$$normalisierung(x) = (x - (a_{min} - b_{min}))\frac{b_{max} - b_{min}}{a_{max} - a_{min}} \tag{4.1}$$

In Tabelle 4.2 ist der Auszug aus Tabelle 4.1 auf den Wertebereich [0,9] normiert.

Für numerische Daten sind für die Eingabedaten verschiedene Charakteristika möglich. So könnten die Werte der Features wie etwa das Alter oder die Größe aus Tabelle 4.2 durch das Clustering-Verfahren separiert sein. Dies bedeutet, der Wertebereich eines Features innerhalb eines Clusters überschneidet sich nur gering oder gar nicht mit denen der anderen Cluster. Gilt dies für mehrere Features des Datensatzes, so kann von einer Abhängigkeit der Features zueinander ausgegangen werden. Im Allgemeinen legt eine solche Separierung nahe, dass derartige Features maßgeblich für die Zuweisung der Daten in Cluster waren und somit eine hohe Aussagekraft über das Clustering haben. Demnach erhält der Analyst einen tieferen Einblick in die Cluster, wenn er Features betrachtet, deren Wertebereiche sich nur gering überschneiden. Dieser Zusammenhang zwischen Cluster-Zuweisung und Wertebereich der Features wird deutlicher, je weniger sich die Wertebereiche überschneiden. Zusätzlich steigt die Aussagekraft über die Cluster-Verschiedenheit, wenn die Überschneidung der Wertebereiche sinkt.

Dem entgegengesetzt ist es möglich, dass Features keine Abhängigkeit zu den anderen Features aufweisen und deren Wertebereiche sich zwischen den Clustern stark überschneiden. Diese gelten im Folgenden als zufällig. Ein Beispiel hierfür ist in Tabelle 4.2 das Gewicht, welches weder eine Abhängigkeit mit dem Alter, der Größe oder der Nationale Identitätsnummer aufweist. Generell lässt sich nicht ausschließen, dass mehrere solche Features eine Abhängigkeit zueinander besitzen, jedoch lässt sich daraus feststellen, dass das Clustering dies nicht widerspiegelt. Demnach haben solche Features eine geringe bis keine Aussagekraft über das Clustering. Die Aussagekraft solcher Features ist für den Analysten im Hinblick auf die Cluster und deren Verschiedenheit somit gering.

Clusterlabel	Alter	Größe	Gewicht	Nationale Identitätsnummer	...
0	14	1.50	57	5	...
	18	1.65	63	5	...
	16	1.60	69	5	...
	17	1.60	81	5	...
	17	1.70	64	5	...
1	19	1.70	58	5	...
	22	1.73	73	5	...
	19	1.71	62	5	...
	21	1.80	91	5	...
2	23	1.80	67	5	...
	24	1.84	69	5	...
	23	1.88	92	5	...
	23	1.82	82	5	...
	24	1.90	84	5	...
	24	1.83	71	5	...
⋮	⋮	⋮	⋮	⋮	⋱

Tabelle 4.1: Szenario nach einem Clustering-Verfahren

Clusterlabel	Alter	Größe	Gewicht	Nationale Identitätsnummer	...
0	0.0	0.000	0.000	0.000	...
	3.6	3.375	1.543	0.000	...
	1.8	2.250	3.086	0.000	...
	2.7	2.250	6.171	0.000	...
	2.7	4.500	1.800	0.000	...
1	4.5	4.500	0.257	0.000	...
	7.2	5.175	4.114	0.000	...
	4.5	4.752	1.286	0.000	...
	6.3	6.750	8.743	0.000	...
2	8.1	6.750	2.571	0.000	...
	9.0	7.650	3.086	0.000	...
	8.1	8.550	9.000	0.000	...
	8.1	7.200	6.429	0.000	...
	9.0	9.000	6.943	0.000	...
	9.0	7.425	3.600	0.000	...
⋮	⋮	⋮	⋮	⋮	⋱

Tabelle 4.2: Szenario nach einer Normierung auf den Wertebereich [0,9]

Eine weitere mögliche Charakteristik eines Features könnte sein, dass es konstant über den kompletten Datensatz ist. In Tabelle 4.2 ist ein Beispiel hierfür die nationale Identitätsnummer. Entstehen kann ein solches Feature zum Beispiel durch eine einseitige Auswahl von Daten. Eine weitere Begründung für eine solche Konstante kann fehlerhafte Software sein, die in ein Feature Initial-Werte schreibt, diese aber nie ändert. Dies zeigt beispielhaft, dass ein solches Feature nur eine geringe, bis keine Aussagekraft über das Clustering hat. Demnach haben auch konstante Features nur eine geringe Relevanz für den Analysten.

4.2 Aufbau des Lösungsansatzes

In diesem Abschnitt wird anhand ausgewählter Schritte eine Darstellung von Clustering-Resultaten entwickelt. Diese verschafft einem Analysten einen genaueren Einblick in die Cluster und den zugrunde liegenden Datensatz. Da dieser Datensatz beliebig viele Datenpunkte und Dimensionen hat, ist es nötig diese auf ein Maß zu beschränken, welches für den Analysten verständlich ist (vgl. Abschnitt 2.3.1). Dies führt zu den drei Schritten:

S1 **Feature-Selektion:**
Hierbei werden mittels Metriken Features bewertet und die aussagekräftigen ermittelt.

S2 **Statistische Kenngröße ermitteln:**
Durch die Berechnung von Kenngrößen der Features werden die Informationen der Daten auf zwei Werte pro ermitteltem Feature reduziert.

S3 **Ergebnisse darstellen:**
Es wird eine einfach verständliche Darstellung der Clustering-Resultate anhand der Ergebnisse aus den vorherigen Schritten geliefert.

Abbildung 4.2 zeigt den konzeptionellen Ablauf der Schritte S1, S2 und S3. In den folgenden Abschnitten werden diese Schritte nach ihrer Ausführungsreihenfolge genauer betrachtet.

4.2.1 Feature-Selektion

Das Ziel des ersten Schritts ist es, dem Analysten nur noch solche Features anzuzeigen, welche nach Abschnitt 4.1 aussagekräftig für das Clustering-Resultat oder einzelne Cluster sind. Allgemein wird dieses Verfahren als Feature-Selektion bezeichnet. Dies sorgt für die Notwendigkeit von Metriken, welche folgende Eigenschaften erfüllen:

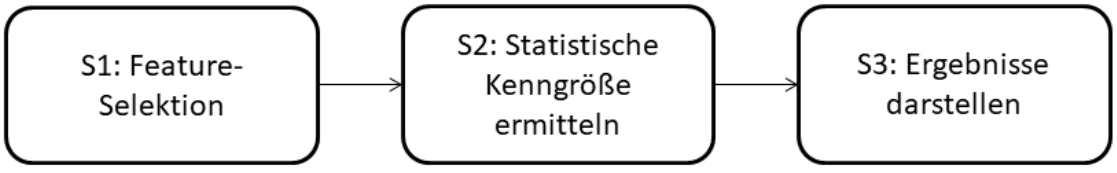


Abbildung 4.2: Prozessablauf des Ansatzes

E1 **bewerten Features nach Aussagekraft:**
Aussagekräftige Features werden von zufälligen unterschieden.

E2 **funktionieren für reelle Zahlen:**
Alle numerische Features erhalten durch die Metriken vergleichbare Resultate.

E3 **geringe Laufzeit:**
In Hinblick auf die Problemstellung aus 4.1 wird eine geringe Laufzeit gefordert.

Um die Auswahl an möglichen Metriken einzugrenzen, wird zunächst betrachtet, welche eine hohe Vergleichbarkeit von Clustern ermöglichen. Viele Clustering-Verfahren nutzen eine Minimierung des Abstands zwischen den Punkten für die Zuweisung zu Clustern (vgl. Abschnitt 2.1). Dies legt nahe, dass Streuungsmaße geeignete Metriken zur Bewertung von Clustern sind (vgl. Abschnitt 2.4.3). Somit sind Features welche einen geringen Wert eines Streuungsmaßes erhalten, aussagekräftig über das Clustering-Resultat, da sie maßgeblich für die Separierung der Cluster gesorgt haben. Demnach besitzen Streuungsmaße die Eigenschaft E1. Zudem kann gezeigt werden, dass Streuungsmaße für gewöhnlich eine lineare Laufzeit besitzen (E3) und für positive und negative reelle Zahlen funktionieren (E2) (vgl. Abschnitt 2.4).

Damit ein Streuungsmaß für die verschiedenen Features eines Datensatzes vergleichbar ist, wird vorausgesetzt, dass die Features der Daten bereits auf den gleichen Wertebereich normierter sind. Dass dies keine große Einschränkung darstellt, wird in Abschnitt 4.1 diskutiert. Durch diese Normierung sind Streuungsmaße verwendbar, die absolute und relative Streuung der Features messen.

Für die weitere Veranschaulichung anhand des Szenarios wird im Folgenden exemplarisch die Varianz als Metrik verwendet. Da diese Metrik ein Streuungsmaß ist und für die weiteren Betrachtungen dieselben Eigenschaften wie andere Streuungsmaße besitzt, wird sie im Folgenden als Repräsentant verwendet. Tabelle 4.3 zeigt die Anwendung der Varianz auf dem Datensatz des Szenarios. In dieser und allen folgenden Tabellen werden pro Cluster die zwei geringsten Metrikergebnisse markiert. Hierbei wird die Varianz pro Cluster auf allen Features berechnet. Besonders auffällig ist dabei die Varianz der nationalen Identitätsnummer. Da diese im kompletten Datensatz konstant ist, ergibt die Varianz dafür den kleinstmöglichen Wert null. Dieses Verhalten gilt aber nicht nur für die Varianz, sondern für alle Streuungsmaße, da diese von dem Abstand der Punkte zueinander abhängig sind. Demnach wäre dieses Feature nach diesen Metriken sehr aussagekräftig. Eine Konstante über den gesamten Datensatz besitzt demnach auch in allen Clustern den gleichen Wertebereich. Würde dies als aussagekräftig gewählt und anschließend visualisiert werden, so würde jedes Cluster die gleichen Werte anzeigen. Dies hat den Nachteil, dass ein Analyst dadurch weder einzelne Cluster gut verstehen, noch die Unterschiede zwischen den Clustern erkennen kann.

Clusterlabel	Var(Alter)	Var(Größe)	Var(Gewicht)	Var(Nat. Ident.)
0	**1.490**	2.228	4.295	**0**
1	1.367	**0.765**	10.811	**0**
2	**0.202**	0.603	5.483	**0**

Tabelle 4.3: Varianz des Datensatzes aus dem Szenario

Um aussagekräftige Features zu finden, ist es somit sinnvoll, dass diese Features innerhalb eines Clusters nur wenige verschiedene und über den ganzen Datensatz hinweg viele verschiedene Werte annehmen. Aus der Sicht der Streuungsmaße werden nach dieser Anforderung Features gesucht, welche über alle Cluster hinweg eine große, und innerhalb eines Clusters eine kleine Streuung aufweisen. Daraus ergibt sich eine zusätzliche Eigenschaft, welche eine Metrik erfüllen muss:

E4 **Behandlung von Konstanten:**
Die Metrik muss für ein Feature minimal oder maximal sein, wenn dieses innerhalb eines Clusters eine kleine und über den ganzen Datensatz eine große Streuung aufweist.

Dies sorgt dafür, dass verschiedene Cluster verschiedene Wertebereiche des Features besitzen. Features, welche durch solche Metriken ausgewählt werden, ermöglichen dem Analysten sowohl die Betrachtung einzelner Cluster als auch die Unterschiede der verschiedenen Cluster. Nach dieser Anforderung wird neben den Streuungsmaßen noch eine veränderte Form dieser eingeführt:

$$MetrikDiff_{Cluster\ i,Feature\ j} = \left|Metrik_{Cluster\ i,Feature\ j}\right| - \left|Metrik_{Cluster_{all},Feature\ j}\right| \quad (4.2)$$

Abgesehen von wenigen Ausnahmen sind Streuungsmaße für alle Eingaben positiv, wodurch für diese die Berechnung der Beträge der einzelnen Elemente nicht notwendig ist. Um ein korrektes Ergebnis für alle Streuungsmaße zu garantieren, werden diese in Formel 4.2 jedoch mit angegeben. Nach dieser Gleichung wird $MetrikDiff$ genau dann gering, wenn $Metrik_{Cluster\ i,Feature\ j}$ gering und $Metrik_{Cluster_{all},Feature\ j}$ groß ist. Somit erfüllt diese Gleichung auch die zusätzliche Eigenschaft E4, welche von den Metriken gefordert wird. Um die aussagekräftigsten Features zu finden, müssen die Metriken für jedes Feature des Clusters errechnet und schließlich die Features mit den geringsten Werten ausgewählt werden. Am Beispiel der Varianz-Differenz ist in Tabelle 4.4 zu erkennen, dass die Features Alter, Größe und Gewicht im Vergleich zu den Ergebnissen der Varianz nahezu um dieselbe Konstante verschoben wurden, während die Bewertung der nationalen Identitätsnummer sich nicht geändert hat. Damit sind beispielsweise für Cluster 2 die Features Alter und Größe sehr, das Gewicht wenig und die nationale Identitätsnummer nicht aussagekräftig für ihre Cluster. Ein Analyst erlangt somit für die Features Alter und Größe Einblicke in die einzelnen Cluster und deren Verschiedenheit.

Wie in Tabelle 4.4 zu sehen, ist für Cluster 1 die Bewertung für das Gewicht schlechter als für die nationale Identitätsnummer. Dies entsteht dadurch, dass die Varianz des Gewichts innerhalb Cluster 1 höher ist, als die des ganzen Datensatzes. Demnach bewerten die Metrik-Differenzen Konstanten potenziell besser als Features mit einer großen Streuung innerhalb eines Clusters. Dies kann durch eine leicht veränderte Form der Metrik-Differenzen gelöst werden:

Clusterlabel	VarDiff(Alter)	VarDiff(Größe)	VarDiff(Gewicht)	VarDiff(Nat. Ident.)
0	**-6.876**	**-4.044**	-3.624	0
1	**-6.999**	**-5.507**	2.892	0
2	**-8.164**	**-5.669**	-2.436	0

Tabelle 4.4: Varianz-Differenz des Datensatzes aus dem Szenario

Clusterlabel	VarDiff(Alter)	VarDiff(Größe)	VarDiff(Gewicht)	VarDiff(Nat. Ident.)
0	**-6.876**	**-4.044**	-3.624	0
1	**-6.999**	**-5.507**	-2.892	0
2	**-8.164**	**-5.669**	-2.436	0

Tabelle 4.5: Adjusted Varianz-Differenz des Datensatzes aus dem Szenario

$$\text{AdjMetrikDiff}_{Cluster\ i,Feature\ j} = -\left|Metrik_{Cluster\ i,Feature\ j} - Metrik_{Cluster_{all},Feature\ j}\right| \quad (4.3)$$

Im Unterschied zur $MetrikDiff$ sind, wie in Tabelle 4.5 zu sehen, alle Werte der $AdjMetrikDiff$ negativ. Demnach wird $AdjMetrikDiff$ genau dann gering, wenn einer der Werte $Metrik_{Cluster\ i,Feature\ j}$ und $Metrik_{Cluster_{all},Feature\ j}$ gering und einer groß ist. Dies sorgt dafür, dass Konstanten mit dem größtmöglichen Wert 0 bewertet werden. Der Nachteil der $AdjMetrikDiff$ ist, dass auch Features mit einer großen Streuung innerhalb eines Clusters und einer kleinen über den ganzen Datensatz als aussagekräftig bewertet werden. Demnach liefert die $MetrikDiff$ eine Ordnung der Features nach ihrer Aussagekraft, während die $AdjMetrikDiff$ in seltenen Fällen Features als aussagekräftiger bewertet, welche nach der Definition aus Abschnitt 4.1 dies nicht sind. Für die folgenden Beispiele wird deshalb auf die Metrik-Differenzen nach Formel 4.2 zurückgegriffen.

Auswahl der Features

Im Anschluss an die Berechnung der Metriken auf den Features folgt die Auswahl der Features, welche für die einzelnen Cluster gewählt werden. In Anlehnung an die Informationen I1 (Cluster-Beschreibung), I2 (Datensatz-Beschreibung) und I3 (Unterschiede der Cluster) aus Abschnitt 4.1, liefert dies drei unterschiedliche Ansätze:

Pro Cluster:
Durch die Auswahl aussagekräftiger Features pro Cluster erhält ein Analyst detaillierte Angaben zu I1.

Häufigste aller Cluster:
Mit einer Auswahl von Features, welche für alle Cluster aussagekräftig ist, werden Einblicke in I2 ermöglicht.

Unterschiede der Cluster:
Durch eine geeignete Wahl an Features, welche die Unterschiede der Cluster darstellen, werden einem Analysten Informationen zu I3 dargestellt.

Pro Cluster: Hierbei werden für alle Cluster separat die Features mit den geringsten Werten der Metriken gewählt. Dadurch wird eine individuelle Beurteilung der einzelnen Cluster erreicht, da verschiedene Cluster auch potenziell verschiedene Features als aussagekräftig bewerten. Dies sorgt aber auch dafür, dass die Vergleichbarkeit der Cluster abnimmt. Am Beispiel der Tabelle 4.4 bedeutet

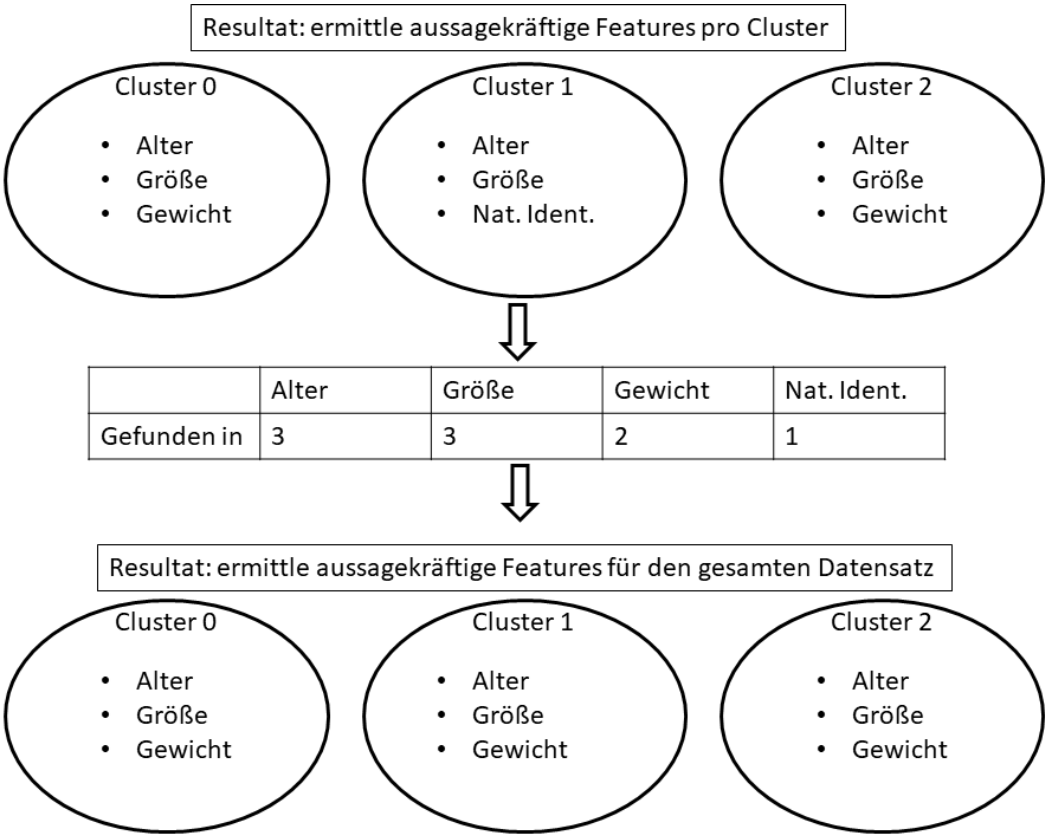


Abbildung 4.3: Schematische Darstellung zur Ermittlung der aussagekräftigsten Features für den ganzen Datensatz

dies für die Annahme, dass drei Features gewählt werden, dass für Cluster 0 und Cluster 2 die Features Alter, Größe und Gewicht und für Cluster 1 Alter, Größe und nationale Identitätsnummer angezeigt werden.

Häufigste aller Cluster: Aus dem vorherigen Ansatz lässt sich dieser Ansatz ableiten. Dieses Verfahren wird in Abbildung 4.3 dargestellt. Hierfür wird zunächst wie in dem vorherigen Ansatz für jedes Cluster einzeln bestimmt welche Features aussagekräftig sind. Anschließend wird für alle Features ermittelt, in wie vielen Clustern es als aussagekräftig erkannt wurde. Nun werden für alle Cluster die Features gewählt, welche über alle Cluster hinweg am häufigsten als aussagekräftig erkannt wurden. Dies führt dazu, dass für alle Cluster dieselben Features als aussagekräftig bestimmt werden. Für den Datensatz aus dem Szenario werden nach Abbildung 4.3 in diesen Ansatz für alle Cluster die Features Alter, Größe und Gewicht ausgewählt. Demnach werden durch diesen Ansatz einzelne Cluster eventuell schlechter dargestellt als durch den vorherigen. Jedoch ist die Vergleichbarkeit einzelner Cluster größer und einfacher, da für alle Cluster dieselben Features ausgewählt werden.

Unterschiede der Cluster: Bei den vorherigen Ansätzen ist es möglich, dass mehrere Cluster exakt dieselben Werte für ein Feature anzeigen. Dies tritt zum Beispiel auf, wenn für zwei Cluster ein Feature ausgewählt wurde, welches bei beiden dieselbe Konstante ist. In diesem Fall wäre es einem Analysten nicht möglich herauszufinden, worin die Cluster sich unterscheiden. Um dies zu

beheben, wird eine weitere Verfeinerung des Ansatzes pro Cluster verwendet. So werden zunächst die Metriken für alle Features pro Cluster berechnet. Nun wird für die Wertebereiche der Features zusätzlich bestimmt, wie groß die Überschneidung zwischen den Clustern ist. Es existieren somit zwei Metriken für jedes Feature innerhalb eines Clusters; Streuung und Überschneidung. Um die Auswahl der Features zu vereinfachen, werden diese beiden Metriken kombiniert. Für die kombinierte Metrik für das Cluster c_i und das Feature f_j ergibt sich:

$$kombMetrik c_i, f_j = a \cdot Streuung_{c_i,f_j} + b \cdot Überschneidung_{c_i,f_j} \tag{4.4}$$

Die Parameter a und b geben die Gewichtung der einzelnen Metriken an. Um diese Metrik für den Vergleich von Features nutzen zu können, müssen die Werte für a und b über alle Cluster und Features konstant sein. Diese kombinierte Metrik wird, wie die bisherigen betrachteten Metriken, minimal, wenn die Aussagekraft eines Features nach Abschnitt 4.1 maximal ist. Für die folgenden Eigenschaften werden nur positive Werte für die Parameter a und b betrachtet. Für negative Werte ergeben sich die gegenteiligen Aussagen.

Für $b = 0$ und $a > 0$ entspricht die Reihenfolge der Features, geordnet nach den Metrikergebnissen, von $kombMetrik$ genau der Reihenfolge der Streuungsmaßen und Metrik-Differenzen. Dabei ist zu beachten dass für $a \neq 1$ sich die absoluten Werte der Features bei $kombMetrik$ zu denen der Streuungsmaßen oder Metrik-Differenzen unterscheiden.

Für $a = 0$ und $b > 0$ entspricht das Ergebnis genau der Überschneidung der Wertebereiche der Features. Dies ist nach Abschnitt 4.1 ebenfalls eine valide Metrik, um herauszufinden, welche Features maßgeblich zu einem Clustering geführt haben.

Werden sowohl $a = 0$ als auch $b = 0$ gesetzt, wird $kombMetrik$ für alle Features und Cluster 0 und liefert somit keine sinnvolle Bewertung der Features.

Der letzte Fall $a > 0$ und $b > 0$ entspricht dem, wofür diese Metrik eingeführt wurde. Damit kann bestimmt werden, welche der beiden Metriken stärker und welche schwächer gewichtet wird. Zudem können Resultate für verschiedene Werte von a und b erzeugt und durch einen Analysten verglichen werden.

Anzahl ausgewählter Features

Für alle drei Ansätze ist relevant, wie viele Features angezeigt werden. Dies lässt sich in folgende drei Herangehensweisen unterscheiden:

Statisch:
Diese Herangehensweise ermittelt eine konstante Anzahl an aussagekräftigen Features.

Schwellenwert:
Hiermit wird lediglich ein Schwellenwert festgelegt. Ausgewählt werden anschließend alle Features, welche eine Metrik unter diesem Schwellenwert aufweisen.

Dynamisch:
Durch den Verlauf der Metrikergebnisse für die einzelnen Features wird dynamisch festgelegt, wie viele Features ausgewählt werden.

Hybrid:
Durch eine Kombination der verschiedenen Herangehensweisen entsteht ein Hybrid.

Statisch: Durch Studien wurde herausgefunden, dass ein Mensch zwischen 5 und 9 Informationen gleichzeitig wahrnehmen kann (vgl. Abschnitt 2.3.1). Dies liefert die Möglichkeit, die Anzahl auszuwählender Features konstant zwischen 5 und 9 festzusetzen. Dadurch wird garantiert, dass für jeden möglichen Datensatz nur so viele Features angezeigt werden, wie wahrgenommen werden können. Der Nachteil dieser Herangehensweise ist, dass Features gewählt werden, welche nur wenig oder keine Aussagekraft über das Clustering-Resultat haben. Dies passiert, wenn die Anzahl aussagekräftiger Features geringer als die Konstanten ist.

Schwellenwert: Durch die Wahl eines geeigneten Schwellenwerts wird in dieser Herangehensweise die Anzahl anzuzeigender Features dynamisch ermittelt. Dieser kann dabei durch einen Analysten eingestellt werden oder wird vom Programm vorgeschlagen. Nach der Festlegung dieses Schwellenwerts werden alle Features gewählt, deren Metrik diesen Wert unterschreitet. Zu beachten ist, dass die Wahl eines geeigneten Schwellenwerts auch von der gewählten Normierung der Daten abhängt. Bei einem geeigneten Schwellenwert werden demnach genau die aussagekräftigen Features ausgewählt. Jedoch führt eine zu hohe Wahl dazu, dass zu viele Features gewählt werden und die Ergebnisse für einen Analysten nur schwer interpretierbar sind. Bei einem zu geringen Wert kann es passieren, dass teilweise gar keine Features gewählt werden, sofern alle Features diesen Schwellenwert nicht unterschreiten.

Dynamisch: Diese Herangehensweise wird durch eine Verfeinerung der vorherigen erreicht. In dieser Herangehensweise werden die Features nach dem Wert der Metrik aufsteigend sortiert. Nun wird für jedes Paar benachbarter Features berechnet, wie groß der Unterschied der Metriken dieser Features ist. Ändert sich der Wert der Metrik stark von seinem Vorgänger, so lässt sich daraus schließen, dass die Aussagekraft des nächsten Features um ein Vielfaches geringer ist als die seiner Vorgänger. In Abbildung 4.4 ist der beschriebene Graph für Cluster 1 aus dem Datensatz des Szenarios zu sehen. Daraus geht hervor, dass der Unterschied der Metriken zwischen Alter und Größe um ein Vielfaches kleiner ist, als die zwischen nationale Identitätsnummer und Größe. Dies lässt sich anhand des Knicks an dieser Stelle erkennen. Je stärker der Knick zu erkennen ist, desto größer ist der Unterschied der Metriken. Daraus lässt sich schließen, dass Alter und Größe eine vergleichbare Aussagekraft haben, während das Gewicht eine wesentlich kleinere besitzt. Da die Features nach der Größe ihrer Metrik sortiert wurden, sind auch alle weiteren Features nach dem Gewicht um ein vielfaches weniger aussagekräftig, als die ersten zwei. Daraus folgt, dass für dieses Cluster die ersten zwei Features gewählt werden.

Hybrid: Zudem ist eine Kombination dieser Ansätze möglich. So ist zum Beispiel die Herangehensweise des Schwellenwerts mit der statischen kombinierbar. Hierfür gibt die Konstante an, wie viele Features mindestens angezeigt werden. Unterschreiten noch weitere Features den Schwellenwert, so werden noch weitere angezeigt. Nach den Erkenntnissen der Wahrnehmung ist es zusätzlich möglich, eine obere Grenze von zum Beispiel maximal 9 auszuwählenden Features festzulegen (vgl. Abschnitt 2.3.1).

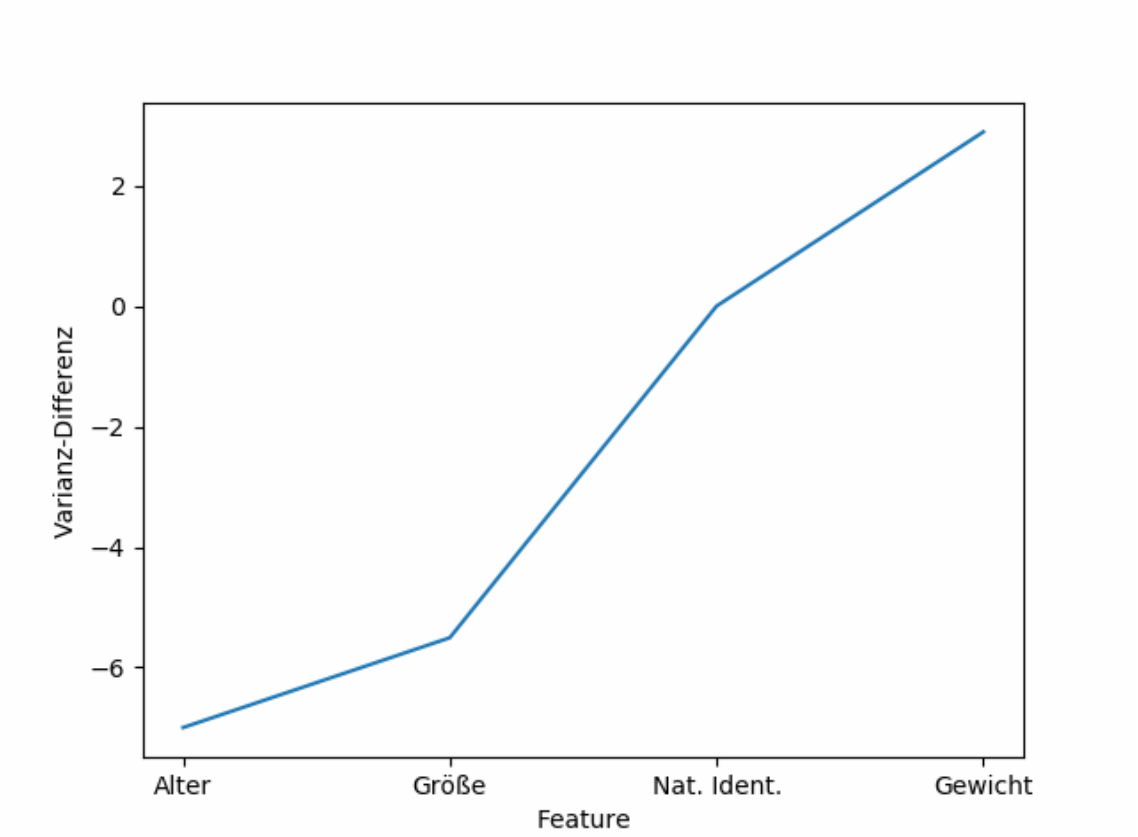


Abbildung 4.4: Verlauf der Varianz-Differenz aus Cluster 1

4.2.2 Statistische Kenngröße ermitteln

Dieser Schritt dient zur Ermittlung statistischer Kenngrößen der aussagekräftigen Features. Clustering-Verfahren sorgen lediglich für eine Gruppierung der Daten, jedoch bleiben alle Datenpunkte erhalten. Damit einzelne Cluster für den Analysten interpretierbar sind, ist es notwendig, die Menge an Informationen, die über die Features angezeigt werden, auf ein Maß zu reduzieren, welches für ihn leicht verständlich ist (vgl. Abschnitt 2.3.1). Umgesetzt wird dies durch ausgewählte statistische Kenngrößen, wie zum Beispiel Minimum/Maximum oder oberes/unteres Quartil. Diese Kenngrößen werden für alle Cluster für jedes der im vorherigen Schritt gewählten Features berechnet. Dies liefert eine übersichtliche Repräsentation der Werte eines Clusters mit zwei Kenngrößen pro Feature und Cluster. Tabelle 4.6 zeigt für den Datensatz aus 4.1 das Minimum und Maximum für die Features, welche im vorherigen Abschnitt durch die Hybrid-Herangehensweise pro Cluster als die aussagekräftigsten identifiziert wurden.

Kritisch zu betrachten ist hierbei die Auswahl der geeigneten Kenngröße. So liefert eine Angabe des Minimums und Maximums eines Features innerhalb eines Clusters eine Spanne, zwischen der sich alle Werte befinden, doch ist dies auch sehr für Ausreißer anfällig. Das obere und untere Quartil agiert genau gegenteilig. Es ist robuster gegen Ausreißer, da diese Maße prozentual aus dem Datensatz errechnet werden, doch werden nicht alle Werte des Features durch dieses Maß

Clusterlabel	Alter		Größe	
	Minimum	Maximum	Minimum	Maximum
0	14	18	1.50	1.70
1	19	22	1.70	1.80
2	23	24	1.80	1.90

Tabelle 4.6: Tabellarische Repräsentation des Datensatzes aus dem Szenario anhand Minimum und Maximum

Clusterlabel	Alter		Größe	
	u.Quartil	o.Quartil	u.Quartil	o.Quartil
0	16	17	1.6	1.65
1	19	21.5	1.705	1.765
2	23	24	1.82	1.88

Tabelle 4.7: Tabellarische Repräsentation des Datensatzes aus dem Szenario anhand unterem und oberen Quartil

repräsentiert. Die Quartile der aussagekräftigen Features des Datensatzes aus dem Szenario sind Tabelle 4.7 zu entnehmen. Allgemein lässt sich daraus ableiten, dass die Wahl der Kenngröße von den gewünschten Ergebnissen abhängt und damit abhängig vom Analysten und der Domäne ist.

4.2.3 Ergebnisse darstellen

In diesem Abschnitt werden Möglichkeiten zur Kommunikation der Daten gegenüber einem Analysten vorgestellt. In den vorherigen Schritten S1 und S2 wurde berechnet, welche Features und zudem welche Werte für diese Features pro Cluster angezeigt werden. Ziel dieses Schritts ist es, diese Ergebnisse anhand einer geeigneten Darstellung zu visualisieren. Um bei der Darstellung der Daten für beliebig viele Dimensionen dennoch detailliert zu sein, wird eine textuelle Darstellung gewählt. Diese ist zum Beispiel in Form einer Tabelle oder durch eine Wortwolke realisierbar. Die Tabelle 4.6 zeigt eine mögliche Repräsentation der ermittelten Ergebnisse für das Beispiel des Szenarios aus Abschnitt 4.1.

Für den Ansatz der Wortwolke können, entsprechend Abbildung 4.5, die verschiedenen Features so dargestellt werden, dass die Schriftgröße der Feature-Namen einen Einblick in die Aussagekraft des Features ermöglicht. Somit könnten Features, welche in Schritt 1 als aussagekräftig ermittelt wurden, größer dargestellt werden, als nicht aussagekräftige. Um die Werte aus Schritt zwei zusätzlich zu visualisieren, können diese anhand eines Tooltips dargestellt werden. Dabei ermöglichen Wortwolken durch die verschiedenen Schriftgrößen ein intuitives Verständnis dafür, welche Features aussagekräftiger sind als andere (vgl. Abschnitt 2.3.2). Dies ist ohne Weiteres in einer Darstellung als Tabelle nur schwer zu erkennen. Dem entgegengesetzt ist bei diesem Ansatz der Vergleich der direkten Werte der Features erschwert, da ein Analyst sich diese Werte merken muss.

Durch eine angepasste Variante der Wortwolke können auch die Werte aus Schritt zwei für alle Features gleichzeitig visualisiert werden (Abbildung 4.6). Anhand dieser Darstellung ist ebenfalls erkenntlich, dass verschiedene Formen von Wortwolken möglich sind. Mit dieser Wortwolke lassen

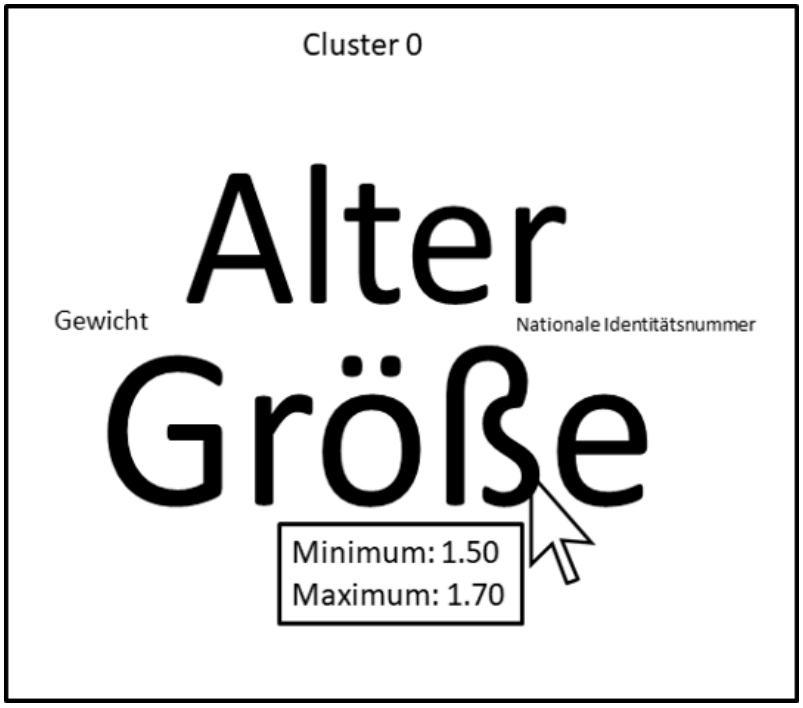
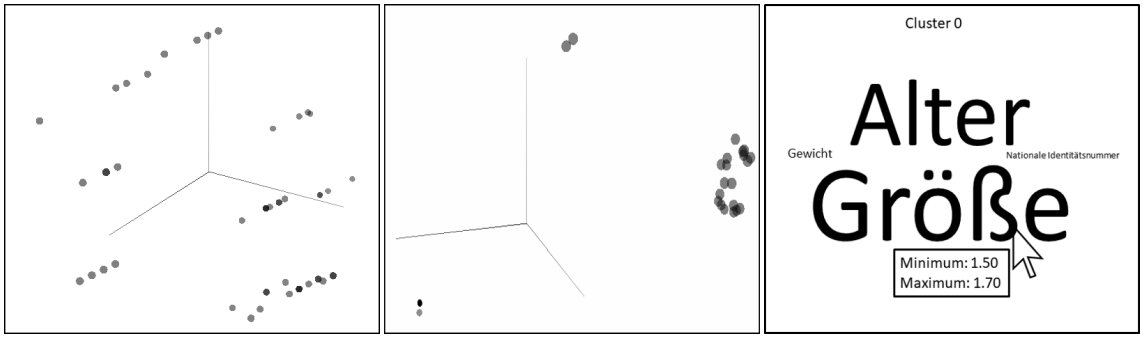


Abbildung 4.5: Visualisierung von Clusters 0 des Szenarios anhand einer Wortwolke

sich die Wertebereiche verschiedener Cluster mit geringerem Aufwand vergleichen. Jedoch sorgt die größere Anzahl angezeigter Werte dafür, dass die Schriftgrößen der Features verkleinert werden müssen. Um dennoch die Aussagekraft der Features anhand der Schriftgröße darzustellen, werden demnach Features potenziell sehr klein dargestellt.

Abschließend ist in Abbildung 4.7 ein Vergleich der Darstellungen von PCA, t-SNE und einer aus diesem Ansatz generierten Wortwolke mit Tooltip zu sehen. Darin wird deutlich, dass die Wortwolke eine detailliertere und übersichtlichere Darstellung der Clustering-Resultate liefert, als die anderen Verfahren. So ist in der Wortwolke durch die verschieden großen Schriften der Feature-Namen intuitiv feststellbar, welche Features aussagekräftiger sind als andere. Dies lässt sich aus den Visualisierungen von PCA und t-SNE nicht herauslesen. Zudem kann in den visuellen Ansätzen keine Aussage über die Wertebereiche der Features getroffen werden. Dies ist in der Wortwolke möglich, indem für ein Feature beispielsweise Minimum und Maximum oder unteres und oberes Quartil angezeigt wird.

Abbildung 4.6: Veränderte Version der Wortwolke



(a) PCA **(b)** t-SNE **(c)** Wortwolke

Abbildung 4.7: Vergleich der Darstellungen von Clustering-Resultaten durch PCA, t-SNE und eine Wortwolke

5 Evaluation

In diesem Kapitel wird der Ansatz aus Kapitel 4 evaluiert. Der Fokus der Evaluation ist Schritt eins (Feature-Selektion), da die folgenden Schritte auf diesem aufbauen. Werden in diesem Schritt nicht aussagekräftige Features ausgewählt, so ist eine detaillierte Darstellung der Clustering-Resultate nicht möglich. Das Ziel dieser Evaluation ist es, eine Metrik zu finden, welche für Datensätze mit verschiedenen Charakteristiken performant ist und mit einer hohen Genauigkeit aussagekräftige Features auswählt.

5.1 Versuchsaufbau

In diesem Abschnitt werden die Details zum Versuchsaufbau und dessen Durchführung erläutert. Zunächst wird auf einen synthetischen Goldstandard eingegangen, welcher es ermöglicht, den vorgestellten Ansatz in Hinblick auf die Genauigkeit und Laufzeit zu evaluieren. Zusätzlich wird angegeben, welche Hard- und Software verwendet wurde, um sowohl den Goldstandard zu generieren, den Ansatz umzusetzen und die Evaluation durchzuführen.

5.1.1 Generierung des Goldstandards

Synthetisch generierte Goldstandards werden für die Evaluation relevant sein, da an ihnen getestet wird, wie der entwickelte Ansatz arbeitet. Für die Evaluation wird auf synthetische Daten zurückgegriffen, da für Datensätze aus der realen Welt für gewöhnlich nicht bekannt ist, welche Features eine hohe Aussagekraft haben. Dies sorgt dafür, dass bei diesen Datensätzen häufig kein Vergleich von gewünschter und tatsächlicher Ausgabe möglich ist. Um diesen Vergleich für die Evaluation zu ermöglichen, wird bei der Generierung des Goldstandards festgelegt, wie viele und welche Features aussagekräftig für das Clustering sind. Die Berechnung der aussagekräftigen Features geschieht ohne dieses Wissen und wird anschließend mit den bekannten Resultaten verglichen.

Erzeugung

Bei der Erzeugung der Daten wird zunächst die Anzahl Cluster und Anzahl Datenpunkte festgelegt. Dabei werden die Datenpunkte gleichmäßig auf die Cluster verteilt und innerhalb der Cluster erzeugt. Dadurch entsteht ein simuliertes Clustering-Resultat. Dabei werden die zu erzeugenden Features unterschieden in:

aussagekräftig:
Für die Generierung dieser Features wird eine mehrdimensionale Gaussverteilung verwendet.

zufällig:
Über den kompletten Wertebereich werden diese Features gleichverteilt zufällig erzeugt.

konstant:
Für alle Datenpunkte besitzen diese Features denselben Wert.

Die mehrdimensionale Gaussverteilung der aussagekräftigen Features sorgt für eine Abhängigkeit dieser Features. Zudem ist zu beobachten, dass Messungen aus der realen Welt häufig einen zu dieser Verteilung ähnlichen Verlauf bilden [Wei09]. Beispiele für solche Features aus dem Datensatz des Szenarios in Abschnitt 4.1 sind das Alter und die Größe von Menschen. Durch eine gleichverteilung der zufälligen Features wird eine Unabhängigkeit der Features untereinander erzeugt. Ein solches Feature könnte nach dem Datensatz aus Abschnitt 4.1 etwa das Gewicht sein. Die konstanten Features nehmen einen konstanten Wert über den ganzen Datensatz an. Solche Features könnten zum Beispiel bei einer schlechten Wahl an Probanden auftreten und ist im Datensatz aus 4.1 mit der nationalen Identitätsnummer vertreten.

Die Anzahl der jeweiligen Features wird initial festgelegt und bleibt für jeden Datenpunkt des Datensatzes gleich. Zudem besitzen alle Features den gleichen Wertebereich, wodurch die Daten bereits normiert generiert werden. Für c Cluster, p Datenpunkte und den Rauschwert r lässt sich der Ablauf der Generierung konzeptionell wie folgt beschreiben:

1. generiere c Cluster
2. füge in jedes Cluster $\frac{p}{c}$ viele Datenpunkte ein (mit aussagekräftigen, zufälligen und konstanten Features)
3. erstelle $p \cdot r$ viele verrauschte Datenpunkte ohne Cluster-Zugehörigkeit (mit ausschließlich zufälligen Features)
4. clustere alle Datenpunkte neu mit c Clustern

Das Clustering-Resultat nach Schritt 1 wird im Folgenden als **initiales Clustering-Resultat** und das nach Schritt 4 als **verrauschtes Clustering-Resultat** bezeichnet. Für diese wird der adjusted mutual info score (AMIS) errechnet, um eine Aussage darüber zu treffen, wie nah das verrauschte Clustering-Resultat an dem initialen ist.

Durchführung

Für die Evaluation wurden Datensätze mit folgenden Parametern erzeugt:

- Anzahl an Punkten p: 100, 1000, 10000, 100000, 1000000
- Anzahl an Clustern c: 5, 25, 50
- Anzahl an Features f: 10, 100
- Rauschen r: 0.33, 0.66, 0.99

Dabei wurden für die entsprechenden Kombinationen sowohl eine Version des Datensatzes mit Konstanten und eine Version ohne erzeugt. Über alle Datensätze hinweg wurden fünf aussagekräftige Features erzeugt. Evaluiert werden diese Datensätze mit allen Streuungsmaßen aus Abschnitt 2.4 außer der Spannweite und dem Quartilsabstand. Diese Metriken werden im Folgenden als „Standardmetriken“ bezeichnet. Zudem werden für diese Metriken die entsprechenden Metrik-Differenzen betrachtet. Für die Evaluation wird für jeden Datensatz und jede Metrik die Laufzeit gemessen und die Genauigkeit errechnet.

5.1.2 Hardware und Software

Durchgeführt wurde die Evaluation unter Windows 10 mit einem Intel Core i5 4690 Prozessor und einer Samsung SSD 840 EVO. Für die Umsetzung des Goldstandards aus Abschnitt 5.1.1 wurde Python mit den Bibliotheken scikit-learn 0.20.1[1] für die Erzeugung der zufälligen Daten und numpy 1.15.4[2] für die Speicherung und Verarbeitung der Daten verwendet. Die Berechnung der Metriken wurde mit der Bibliothek Apache Commons Math 3.4 durchgeführt.

5.2 Evaluation der Genauigkeit

In diesem Abschnitt wird die Genauigkeit der Standardmetriken und Metrik-Differenzen untersucht. Diese ermöglicht es zu messen, wie viele aussagekräftige Features von den Metriken gefunden werden. Diese Betrachtung kann als generelle Bewertung des entwickelten Ansatzes gesehen werden, da die Schritte S2 und S3 darauf aufbauen. Als Genauigkeit einer Metrik wird im Folgenden bezeichnet, wie viele von den bekannten aussagekräftigen Features gefunden wurden.

Wie in Abschnitt 4.2.1 festgestellt unterscheiden sich die Resultate der Standardmetriken und Metrik-Differenzen stark, wenn Konstanten in dem Datensatz enthalten sind. Demnach ist die folgende Evaluation in diese zwei Fälle unterteilt. Zunächst werden Datensätze mit Konstanten verwendet und darin die Genauigkeit der verschiedenen Metriken gemessen. Im Anschluss werden Datensätze ohne Konstanten betrachtet, um zu sehen, welche Unterschiede sich dadurch ergeben. Für beide Fälle wird die Evaluation unterteilt in:

Durchschnittsgenauigkeit:
Dies liefert eine Bewertung der Metriken auf der Ebene der Datensätze.

Betrachtung einzelner Cluster:
Hierbei wird eine detaillierte Analyse der Metriken ermöglicht, indem einzelne Cluster betrachtet werden.

Einfluss des Clustering-Verfahrens:
In diesem Teil der Evaluation wird untersucht, wie stark verschiedene Metriken von der Güte des Clustering-Verfahrens abhängen.

[1] https://scikit-learn.org/stable/index.html
[2] http://www.numpy.org/

Durchschnittsgenauigkeit: Diese gibt an, zu wie viel Prozent die aussagekräftigen Features in einem Datensatz ausgewählt wurden. Errechnet wird dies für c Cluster durch folgende Formel:

$$Durchschnittsgenauigkeit_{Metrik\ m,Feature\ f} = (\sum_{i=1}^{c} \frac{gefundene_{cluster\ i,m,f}}{bekannteProCluster}) \tag{5.1}$$

Die Durchschnittsgenauigkeit wird genutzt, um Metriken auf der Ebene der Datensätze zu unterscheiden. Liefert eine Metrik bereits bei dieser Analyse wesentlich schlechtere Ergebnisse als andere, so kann diese Metrik für die weiteren Analysen vernachlässigt werden.

Betrachtung einzelner Cluster: Hierzu werden die Ergebnisse der Metriken auf einzelnen Clustern genauer betrachtet. Der Unterschied zu der Betrachtung der Durchschnittsgenauigkeit wird exemplarisch an einem Datensatz mit zehn Clustern erläutert, mit fünf nach Abschnitt 4.1 aussagekräftigen Features. Die Betrachtung zweier Metriken A und B ergibt, dass beide eine Durchschnittsgenauigkeit von 80% auf diesem Datensatz erzielen. Eine genauere Betrachtung der einzelnen Cluster ergibt jedoch, dass

- Metrik A in 10 Clustern 4 von 5 aussagekräftige Features findet
- Metrik B in 8 Clustern 5 von 5 und in 2 Clustern 0 von 5 aussagekräftige Features findet

Damit bewertet Metrik B zwei Cluster falsch, während Metrik A eine gute Bewertung für alle zehn Cluster liefert. Dies zeigt, dass die Durchschnittsgenauigkeit gut geeignet ist, um Metriken grob zu unterscheiden, doch erzielen mehrere Metriken ähnliche oder gleiche Ergebnisse so müssen diese genauer betrachtet werden.

Einfluss des Clustering-Verfahrens: Um den Einfluss des Clustering-Verfahrens zu analysieren wird untersucht, wie stark die Durchschnittsgenauigkeit der Metriken (Formel 5.3) vom Clustering-Resultat abhängt. Dafür wird aus den in Abschnitt 5.1.1 beschriebenen Datensätzen der Verlauf der Durchschnittsgenauigkeit der Metriken in Abhängigkeit des AMIS betrachtet. Dieser misst die Überschneidungen des initialen Clustering-Resultats mit dem des verrauschten Clustering-Resultats (vgl. Abschnitt 5.1.1). Das initiale Clustering-Resultat wird als optimales Resultat betrachtet, da nach diesem die Abhängigkeiten der Features und somit die Separierung der Datenpunkte in die Cluster festgelegt wurde. Ein hoher AMIS steht für eine hohe Übereinstimmung zwischen den zwei Clustering-Resultaten (vgl. Abschnitt 2.2). Ein niedriger Wert dieser Metrik bedeutet demnach, dass die zwei Clustering-Resultate nur eine geringe Übereinstimmung haben und das Clustering-Resultat tendenziell nicht optimal ist, da es die initialen Abhängigkeiten der Features nur gering widerspiegelt.

5.2.1 Goldstandards mit Konstanten

In diesem Abschnitt wird die Genauigkeit der in Abschnitt 5.1.1 beschriebenen Versuchsdurchführung anhand des Goldstandards mit Konstanten evaluiert.

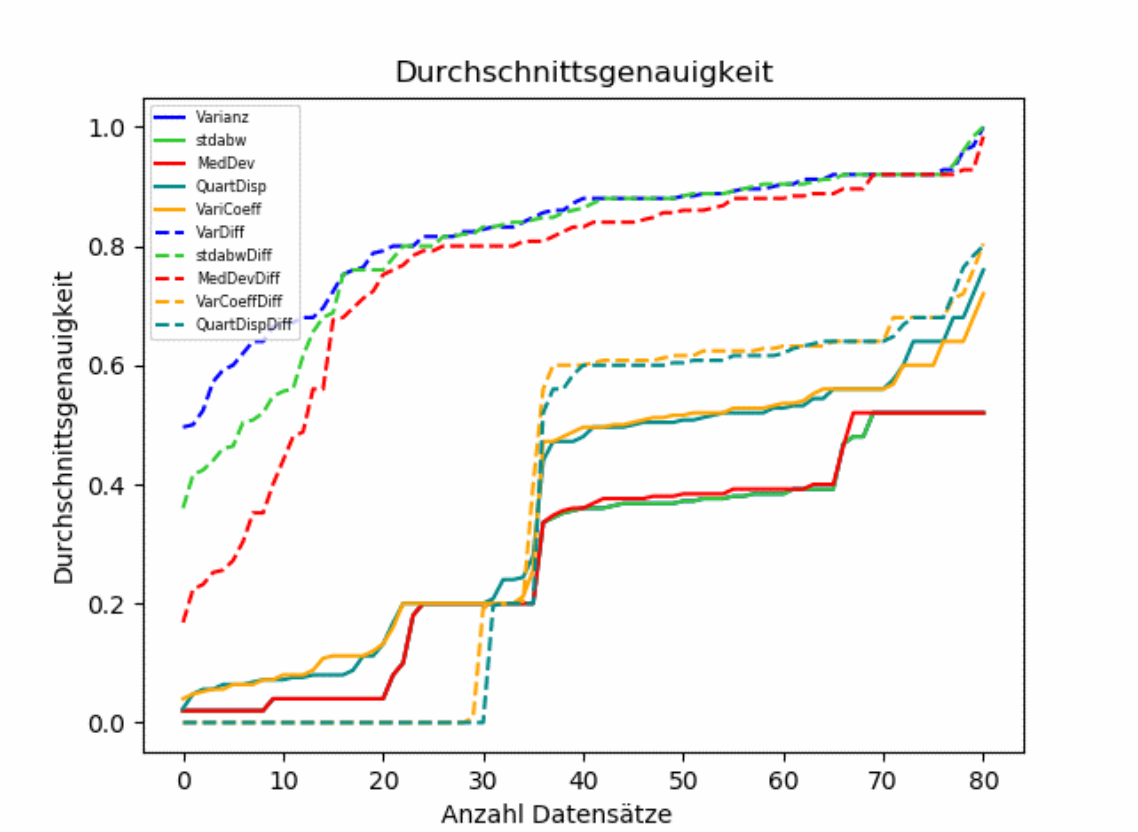


Abbildung 5.1: Durchschnittsgenauigkeit der Metriken in Abhängigkeit der Anzahl an Datensätzen mit Konstanten für die maximal diese Genauigkeit erreicht wurde

Durchschnitt aller Cluster

Abbildung 5.1 wurde aus den Resultaten auf den Datensätzen erstellt und zeigt den Verlauf der Durchschnittsgenauigkeit der Metriken in Abhängigkeit der Anzahl an Datensätzen. Durch eine Sortierung nach der Durchschnittsgenauigkeit pro Metrik wird erreicht, dass jeder Punkt der Graphen angibt, wie viele Datensätze maximal eine bestimmte Durchschnittsgenauigkeit erreicht haben. Unterschieden wird dabei nach Abschnitt 2.4 in Standardmetriken und Metrik-Differenzen. Dabei sind die Standardmetriken als durchgezogene und die Metrik-Differenzen als gestrichelte Linien dargestellt.

Generell ist zu erkennen, dass die Standardmetriken eine maximale Genauigkeit von 75% erreichen. Dies lässt sich durch die Auswahl der Metriken begründen, denn diese sind so gewählt, dass sie eine geringe Streuung der Features bevorzugen. Mit Betrachtung der Features aus Abschnitt 5.1.1 bedeutet dies nun, dass auch die Konstante über alle Cluster hinweg als besonders aussagekräftig beurteilt wird, da diese stets eine Streuung von null aufweist. Somit finden diese Metriken drei der vier bekannten aussagekräftigen Features und zusätzlich dieses vierte Feature, somit wird eine Genauigkeit von maximal 75% erreicht. Dem entgegengesetzt erreichen die Varianz-Differenz und die Medianabweichungs-Differenz bei etwa 75% der Datensätze eine Genauigkeit von mehr als 75%. Für diese Datensätze erreichen die Varianz-Differenz und die Medianabweichungs-Differenz

höhere Genauigkeiten, als alle Standardmetriken. Ebenfalls hohe Genauigkeiten erreichen die Quartildispersions-Differenz und die Variationscoeffizienz-Differenz. Diese erreichen, wie in Abbildung 5.1 zu sehen, für etwa 55% der Datensätze eine höhere Genauigkeit als die Standardmetriken. Für die restlichen 45% der Datensätze erreichen diese Metriken eine vergleichbare Genauigkeit wie die Standardmetriken und sind für wenige Datensätze ungenauer als diese. Für die weitere Evaluation wird nach diesen Ergebnissen lediglich die Varianz- und Medianabweichungs-Differenz genauer betrachtet, da diese höhere Genauigkeiten aufweisen als die Quartildispersions-Differenz und die Variationscoeffizienz-Differenz. Die Standardmetriken können ebenfalls vernachlässigt werden, da deren Resultate auf diesen Datensätzen wesentlich ungenauer sind als die der Varianz- oder Medianabweichungs-Differenz.

Für die genauere Betrachtung wird untersucht, wie stark die Durchschnittsgenauigkeit dieser zwei Metrik-Differenzen von den verschiedenen Parametern Anzahl an Datenpunkten, Clustern oder Features abhängt. Um dies zu verdeutlichen, wurde Abbildung 5.2 aus den Resultaten der Metriken erstellt. In dieser sind Graphen der Durchschnittsgenauigkeit der Varianz-Differenz und Medianabweichungs-Differenz in Abhängigkeit der Anzahl an Datenpunkten zu sehen. Für jede Zeile dieser Tabelle wurden die Datensätze visualisiert, welche die gleiche Anzahl an Clustern haben. Die Spalten geben die Datensätze an, die dieselbe Anzahl an Features haben. Der Rauschwert aller Graphen ist 33%. Dies bedeutet, dass zum Beispiel der Graph aus Abbildung 5.2a alle Datensätze darstellt, die zehn Features, fünf Cluster und 33% Rauschen besitzen.

Allgemein lässt sich aus Abbildung 5.2 schließen, dass die Durchschnittsgenauigkeit dieser Metrik-Differenzen mit der Anzahl an Datenpunkten zusammenhängt. In jedem der sechs dargestellten Graphen ist die Tendenz zu sehen, dass die Durchschnittsgenauigkeit steigt, wenn auch die Anzahl an Datenpunkten steigt. Dies lässt sich durch die verschiedenen Arten an Features erklären. So werden spezielle Muster oder Abhängigkeiten verschiedener Features für größere Datensätze immer deutlicher, während zufällige Features für größere Datensätze tendenziell deutlicher als zufällig klassifiziert werden können. Zudem ist in den Graphen zu sehen, dass die Abhängigkeit von Anzahl an Datenpunkten und Durchschnittsgenauigkeit nur schwach bis gar nicht von der Anzahl an Clustern abhängt.

Generell lässt sich auch eine Abhängigkeit von Durchschnittsgenauigkeit und Anzahl an Clustern feststellen. Dies lässt sich besonders für die rechte Spalte aus Abbildung 5.2 feststellen. In diesen Graphen ist die Tendenz zu erkennen, dass die Metrik-Differenzen genauer werden, wenn die Anzahl an Clustern steigt. Eine Ausnahme stellt dabei der Datensatz mit 50 Clustern, 100 Features und 1000 Datenpunkten dar. Da die Aussage für sämtliche andere Datensätze zutrifft, wird dies als Ausreißer betrachtet. Für die linke Spalte lässt sich diese Tendenz ebenfalls erkennen, jedoch deutlich schwächer. Dies bedeutet, dass die Abhängigkeit der Durchschnittsgenauigkeit von der Anzahl an Clustern größer wird, je mehr Features der Datensatz besitzt. Diese Beobachtungen könnten daran liegen, dass Clustering-Resultate verwendet wurden. So wird ein Clustering auf wenigen Features mit einer höheren Wahrscheinlichkeit genau nach den aussagekräftigen Features clustern und deren Muster und Abhängigkeiten beibehalten bleiben, als bei einer hohen Anzahl an Features.

Zudem lässt sich auch eine direkte Abhängigkeit zwischen der Durchschnittsgenauigkeit und der Anzahl an Features erkennen. So fällt bei dem Vergleich der Graphen aus Abbildung 5.2 auf, dass die Graphen der linken Spalte für alle Einträge genauer sind, als die der rechten Spalte. Dies bedeutet, dass die Durchschnittsgenauigkeit dieser Metrik-Differenzen sinkt, wenn die Anzahl an Features steigt. Der Grund dafür ist, dass auch komplett zufällig generierte Features mit einer geringen

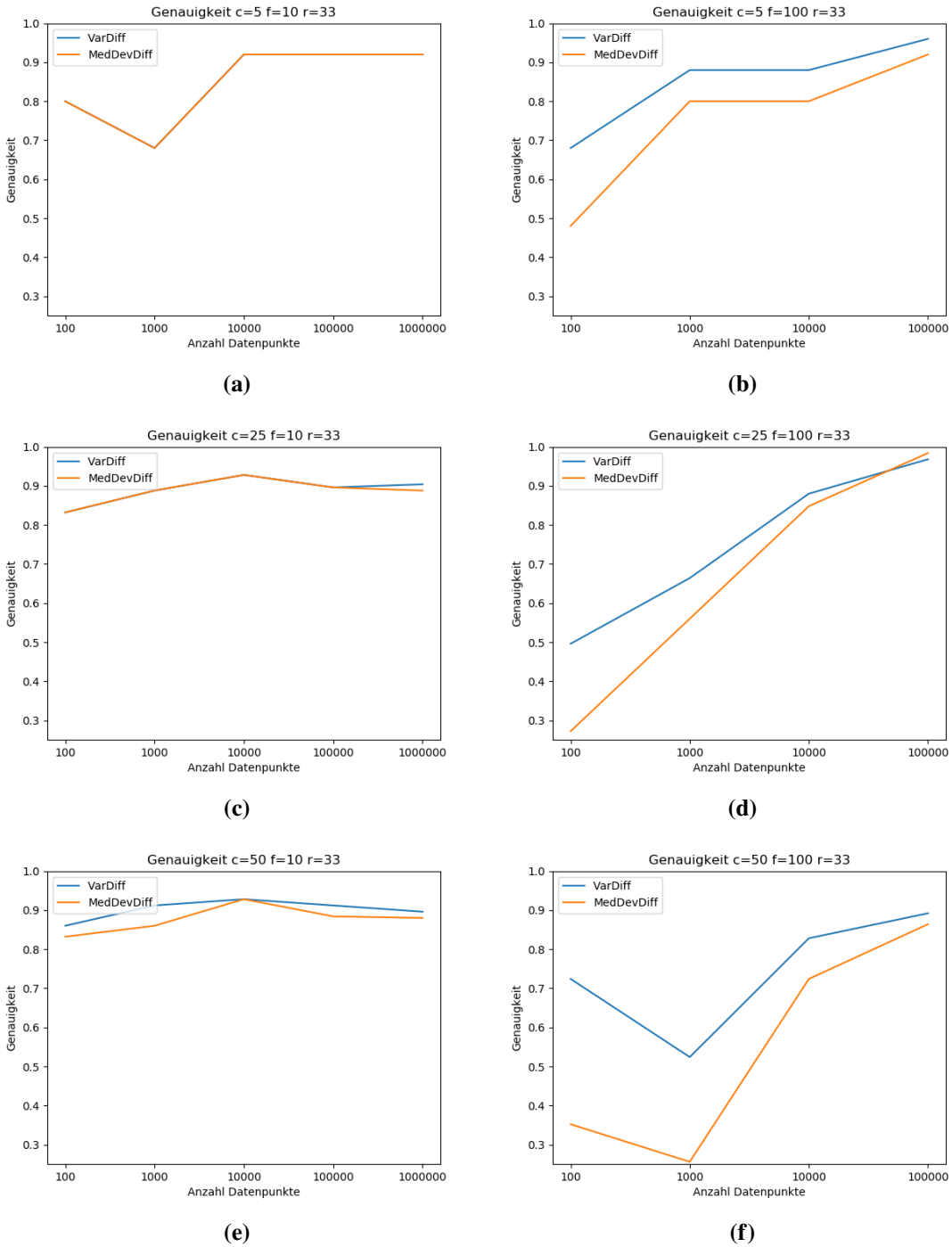


Abbildung 5.2: Vergleich der Durchschnittsgenauigkeiten von Varianz-Differenz und Medianabweichungs-Differenz in Abhängigkeit der Anzahl Datenpunkte, Cluster und Feature

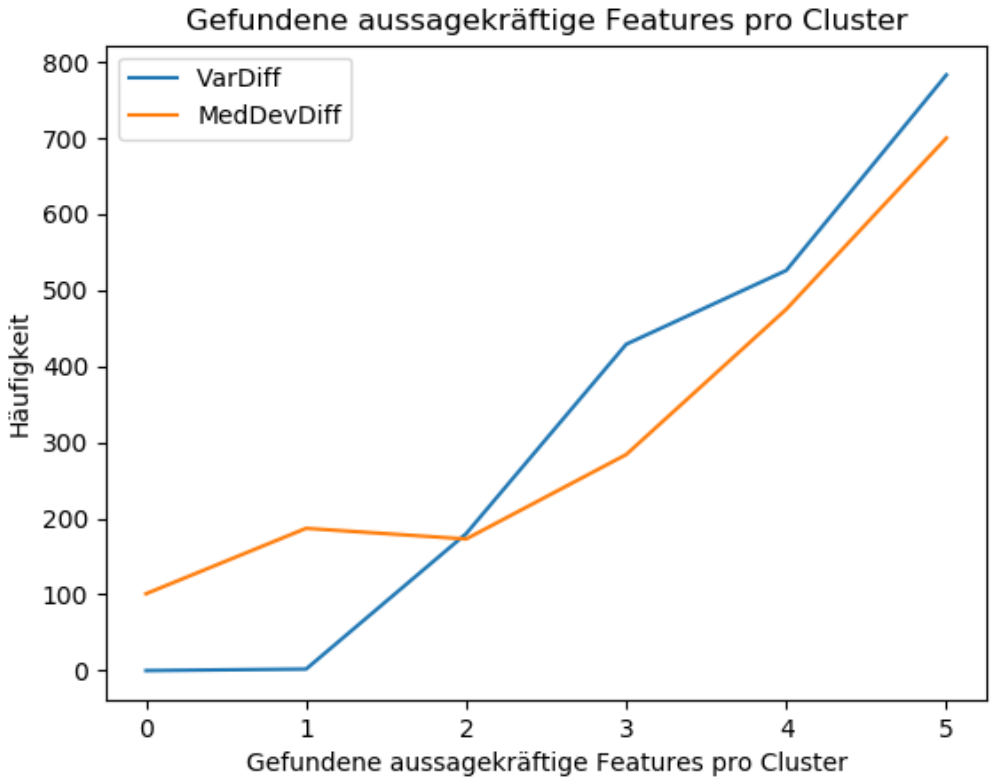


Abbildung 5.3: Anzahl gefundener aussagekräftiger Features pro Cluster über alle Datensätze mit konstanten hinweg

Wahrscheinlichkeit Abhängigkeiten oder Muster besitzen können. Je mehr Features der Datensatz enthält, desto größer ist somit die Wahrscheinlichkeit, dass eines der Features eine geringere Streuung als ein aussagekräftiges Feature besitzt und somit von dem Ansatz fälschlicherweise ausgewählt wird. Auffällig ist zudem, dass die Medianabweichungs-Differenz in den Datensätzen mit 100 Features bis zu 35% ungenauer ist als die Varianz-Differenz. Dies könnte daran liegen, dass die Medianabweichungs-Differenz anfälliger für zufällig generierte Muster ist, während die Genauigkeiten der Varianz-Differenz stabiler bleiben.

Um einen genaueren Einblick in die Vor- und Nachteile der Varianz-Differenz und Medianabweichungs-Differenz zu bekommen, wird im Folgenden untersucht, wie die Ergebnisse auf den einzelnen Clustern aussehen. Dies liefert eine detailliertere Betrachtung der Metriken als die Durchschnittsgenauigkeit.

Betrachtung einzelner Cluster

Betrachtet man lediglich die Summe gefundener aussagekräftige Features für alle Datensätze, so liegt die Medianabweichungs-Differenz mit 1920 gefundenen von 2400 existierenden aussagekräftigen Features exakt gleich mit der Varianz-Differenz. Eine genauere Betrachtung, wie oft eine bestimmte Anzahl an aussagekräftigen Features gefunden wurde, ist Abbildung 5.3 zu entnehmen. Auffällig ist,

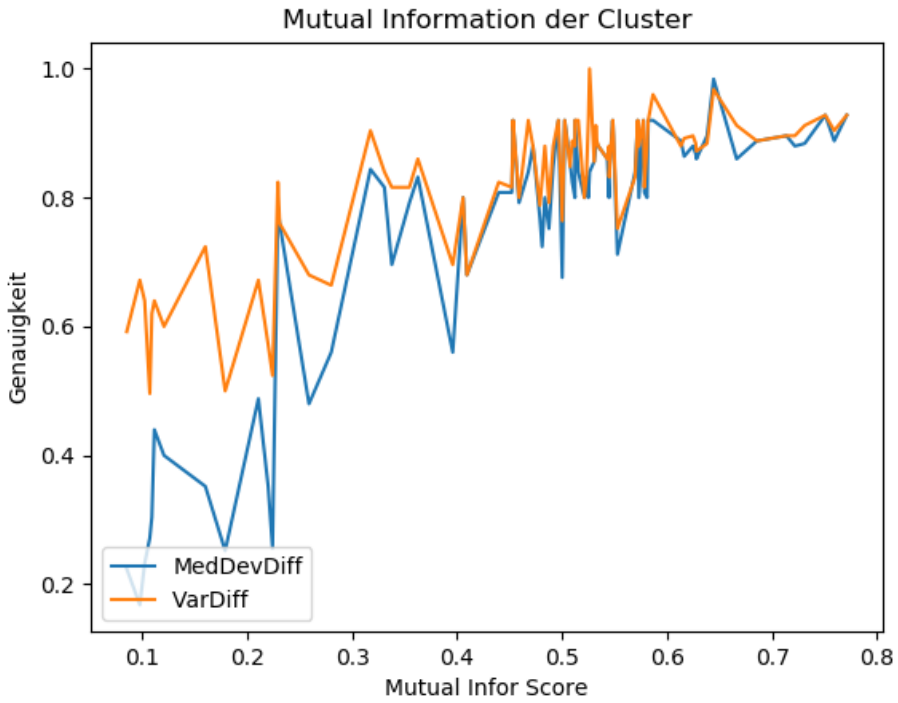


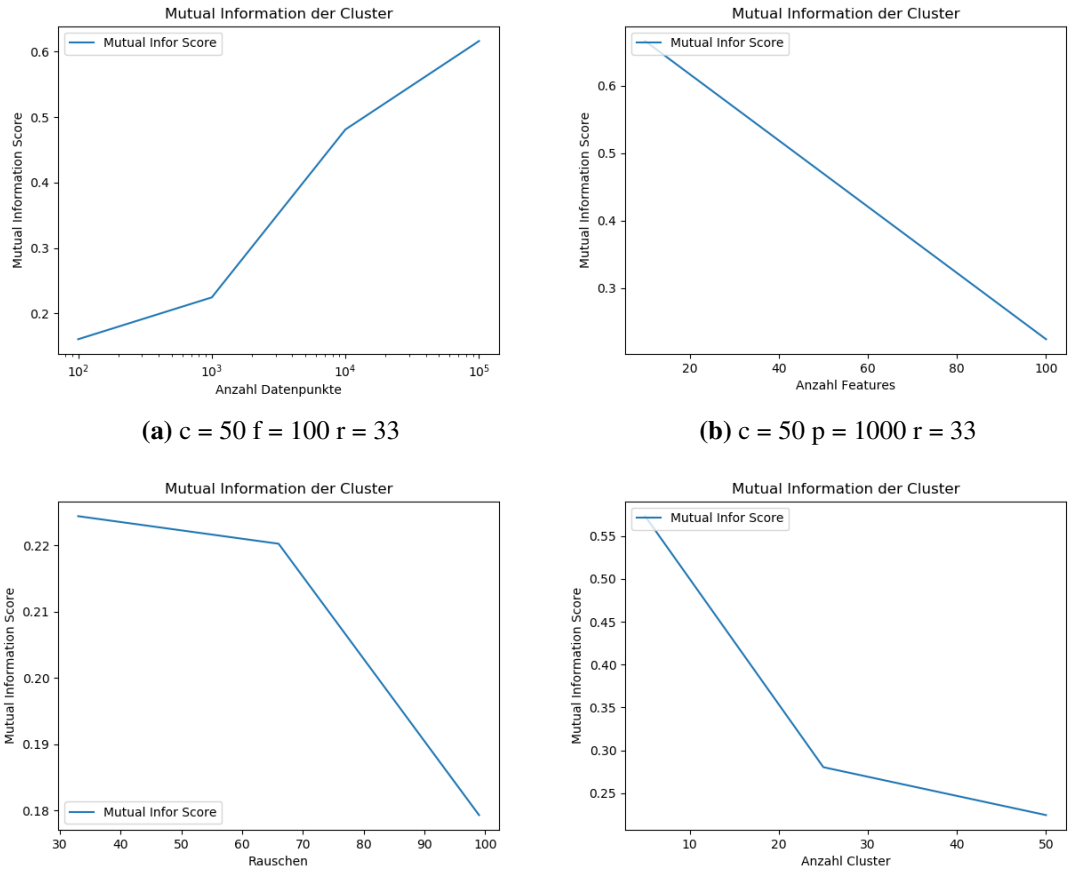
Abbildung 5.4: Verlauf der Genauigkeiten der Varianz-Differenz und Medianabweichungs-Differenz in Abhängigkeit des AMIS

dass beide Metrik-Differenzen annähernd gleich oft zwei aussagekräftige Features in einem Cluster finden. Zusätzlich ist zu sehen, dass es nahezu keine Cluster gibt, in welchen die Varianz-Differenz weniger als zwei aussagekräftige Features findet.

Die Medianabweichungs-Differenz dagegen findet für einen Teil der Cluster nur ein oder zwei aussagekräftige Features. Drei, vier und fünf aussagekräftige Features pro Cluster werden von dieser seltener gefunden, als von der Varianz-Differenz. Demnach bestätigt sich durch die Betrachtung der einzelnen Cluster die Tendenz der Durchschnittsgenauigkeit, dass die Varianz-Differenz genauer ist als die Medianabweichungs-Differenz.

Einfluss des Clustering-Verfahrens

Für beide Metrik-Differenzen ist in der Abbildung die Tendenz zu erkennen, dass die Genauigkeit der Metriken steigt, wenn der AMIS steigt. Dies bedeutet, die Metrik-Differenzen werden genauer, je näher das verrauschte Clustering-Resultat an dem initialen Clustering-Resultat ist. Wie Abbildung 5.4 aber auch zeigt, ist diese Abhängigkeit nicht eindeutig zu erkennen. Da die Datensätze sich in vielen Parametern, wie Anzahl Cluster, Datenpunkte, Features und Rauschen unterscheiden, ist es wahrscheinlich, dass diese Parameter ebenfalls Einfluss auf den AMIS haben. Dies könnte zu einer Verrauschung der Ergebnisse führen.



(a) c = 50 f = 100 r = 33

(b) c = 50 p = 1000 r = 33

(c) c = 50 f = 100 p = 1000

(d) f = 100 p = 1000 r = 33

Abbildung 5.5: Verlauf des AMIS in Abhängigkeit der Parameter Anzahl Datenpunkte (oben links), Anzahl Features (oben rechts), Rauschen (unten links), Anzahl Cluster (unten rechts)

Um dennoch abzuschätzen, wie stark der Einfluss des Clustering-Verfahrens ist, werden im folgenden Beispiele von Verläufen des AMIS in Abhängigkeit der verschiedenen Parameter angegeben und genauer betrachtet.

In Abbildung 5.5 ist zu sehen, dass der AMIS sinkt, wenn die Anzahl an Clustern oder Anzahl an Features steigt. Dies könnte daran liegen, dass in einem Datensatz mit einer großen Anzahl an Features die Wahrscheinlichkeit für zufällige Gruppierungen von Daten steigt. Durch die zusätzlich hohe Anzahl an Clustern steigt damit auch die Wahrscheinlichkeit, dass eine solche zufällige Gruppierung als Cluster gewählt wird. Zusätzlich lässt sich feststellen, dass der AMIS steigt, wenn die Anzahl an Datenpunkten steigt. Diese Beobachtungen des Einflusses der Anzahl Datenpunkte und Anzahl Features entsprechen auch denen aus der Analyse der Durchschnittsgenauigkeit. Damit lässt sich feststellen, dass es einen Zusammenhang zwischen diesen Parametern, dem AMIS und der Durchschnittsgenauigkeit gibt. Für die Anzahl an Clustern lässt sich eine gegenteilige Beobachtung feststellen. Während der AMIS für eine steigende Anzahl an Clustern sinkt, so steigt, wie eingangs in diesem Abschnitt gezeigt, die Durchschnittsgenauigkeit dafür. Dies lässt die Vermutung zu, dass

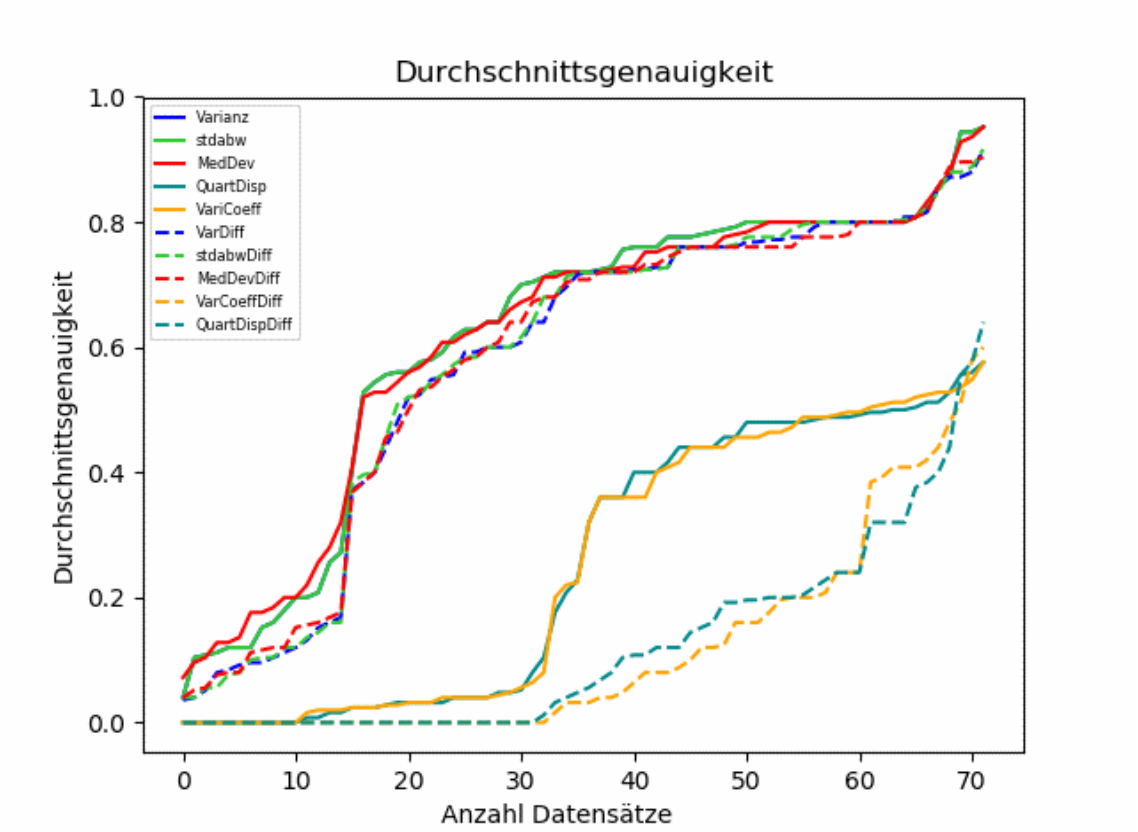


Abbildung 5.6: Durchschnittsgenauigkeit der Metriken in Abhängigkeit der Anzahl an Datensätzen ohne Konstanten für die maximal diese Genauigkeit erreicht wurde

die Anzahl an Cluster die Genauigkeit stärker beeinflusst, als ein verrauschtes Clustering. Demnach ist die Genauigkeit des Ansatzes für eine große Anzahl an Clustern noch immer hoch, auch wenn das Clustering vermeintlich schlecht ist.

5.2.2 Goldstandards ohne Konstanten

Dieser Abschnitt evaluiert die Durchführung aus Abschnitt 5.1.1 für Datensätze ohne Konstanten.

Durchschnitt aller Cluster

In Abbildung 5.6 ist der Verlauf der Durchschnittsgenauigkeit der Metriken in Abhängigkeit der Anzahl an Datensätzen ohne Konstanten, die diese oder eine geringere erreicht haben. Darin sind Standardmetriken mit durchgehenden, Metrik-Differenzen mit gestrichelten Linien dargestellt.

Hier ist deutlich zu erkennen, dass Varianz, Standardabweichung und Medianabweichung sowie deren Differenzen die höchsten Genauigkeiten erzielen. Diese sind bei etwa 80% der Datensätze bei mehr als 50% und für etwa 60% der Datensätze bei mehr als 75%. Auffällig ist, dass die Durchschnittsgenauigkeit dieser Standardmetriken für alle Datensätze höher oder gleich ist, wie

die der entsprechenden Metrik-Differenz. Für etwa 70% der Datensätze beträgt der Unterschied weniger als 5%. Bei den restlichen 30% der Datensätze steigt der Unterschied auf bis zu 20% an. Dies legt nahe, dass der Unterschied der Genauigkeit zwischen den Standardmetriken und den Metrik-Differenzen für bestimmte Parameter steigt. Der Quartilsdispersionskoeffizient und Variationskoeffizient erreichen bei etwa 45% der Datensätze eine Genauigkeit von mehr als 50%. Dabei steigt die Durchschnittsgenauigkeit dieser Metriken für keinen Datensatz über 70%. Deren Differenzen erreichen für etwa 85% der Datensätze eine Durchschnittsgenauigkeit von 20% oder weniger. Für einzelne Datensätze wird die Genauigkeit der Differenzen besser als die der Standardmetriken, jedoch liegt sie noch immer unter 75%. Im Folgenden werden nach diesen Resultaten die Varianz und Medianabweichung sowie deren Differenzen genauer betrachtet. Die Standardabweichung sowie deren Differenz muss nicht zusätzlich betrachtet werden, da die Genauigkeiten nahezu identisch zur Varianz und Varianz-Differenz sind.

In Abbildung 5.7 sind die Durchschnittsgenauigkeiten der Metriken Varianz, Medianabweichung, Varianz-Differenz und Medianabweichungs-Differenz in Abhängigkeit der Anzahl an Datenpunkte zu sehen. Die verschiedenen Graphen stehen dabei für verschiedene Werte der Parameter Anzahl an Cluster und Anzahl an Features. Dabei ist in einer Spalte die Anzahl an Features und in einer Zeile die Anzahl an Clustern konstant. Der Rauschwert aller Graphen beträgt 33%.

Im Allgemeinen lassen sich in dieser Abbildung für die Standardmetriken und Metrik-Differenzen ähnliche Abhängigkeiten feststellen, wie bei der Betrachtung mit Konstanten aus Abschnitt 5.2.1. Demnach werden alle betrachteten Metriken genauer, je größer die Anzahl an Datenpunkten im Datensatz ist. Und auch für die Anzahl an Features gelten dieselben Abhängigkeiten für Datensätze ohne Konstanten; die Metriken werden ungenauer für eine größere Anzahl an Features. Dabei ist zu beachten, dass für eine große Anzahl an Datenpunkten die Genauigkeiten der Metriken für 10 und 100 Features nahezu identisch sind.

Im Gegensatz zu den Datensätzen mit Konstanten lässt sich für diese Datensätze nur eine schwache, bis gar keine Abhängigkeit der Durchschnittsgenauigkeit zu der Anzahl an Clustern feststellen. Die Tendenz der Graphen ist dabei, dass die Durchschnittsgenauigkeit schlechter wird für eine größere Anzahl an Clustern. Dies entspricht dem Gegenteil zu den Datensätzen mit Konstanten.

Generell ist in allen Graphen zu sehen, dass die Standardmetriken einen ähnlichen Verlauf zu den Metrik-Differenzen aufweisen. Dabei sind die Standardmetriken, abgesehen von wenigen Ausnahmen, genauer als deren Differenzen. Für 10 Features beträgt der Unterschied der Standardmetriken und Metrik-Differenzen weniger als 5%. Einzige Ausnahme ist der Datensatz mit 50 Clustern, 10 Features und 100 Datenpunkten. In diesem ist der Unterschied der Metriken etwa 15%. Da dieser nur einer von zwölf Datensätzen ist, wird dieser als ein Ausreißer betrachtet. Für 100 Features werden die Unterschiede der Standardmetriken und Metrik-Differenzen mit etwa 5% bis 10% deutlicher. Demnach wird der Unterschied der Durchschnittsgenauigkeit der Metrik-Differenzen und Standardmetriken größer, je größer die Anzahl an Features wird. Dabei ist zu beachten, dass in Abbildung 5.7 die Anzahl an Features der rechten Spalte zehnmal größer ist als die der linken Spalte. Dabei steigt der Unterschied zwischen den Metriken um etwa 5%. Der Einfluss der Anzahl an Features auf die Durchschnittsgenauigkeit ist somit nicht groß, für große Änderungen aber erkennbar.

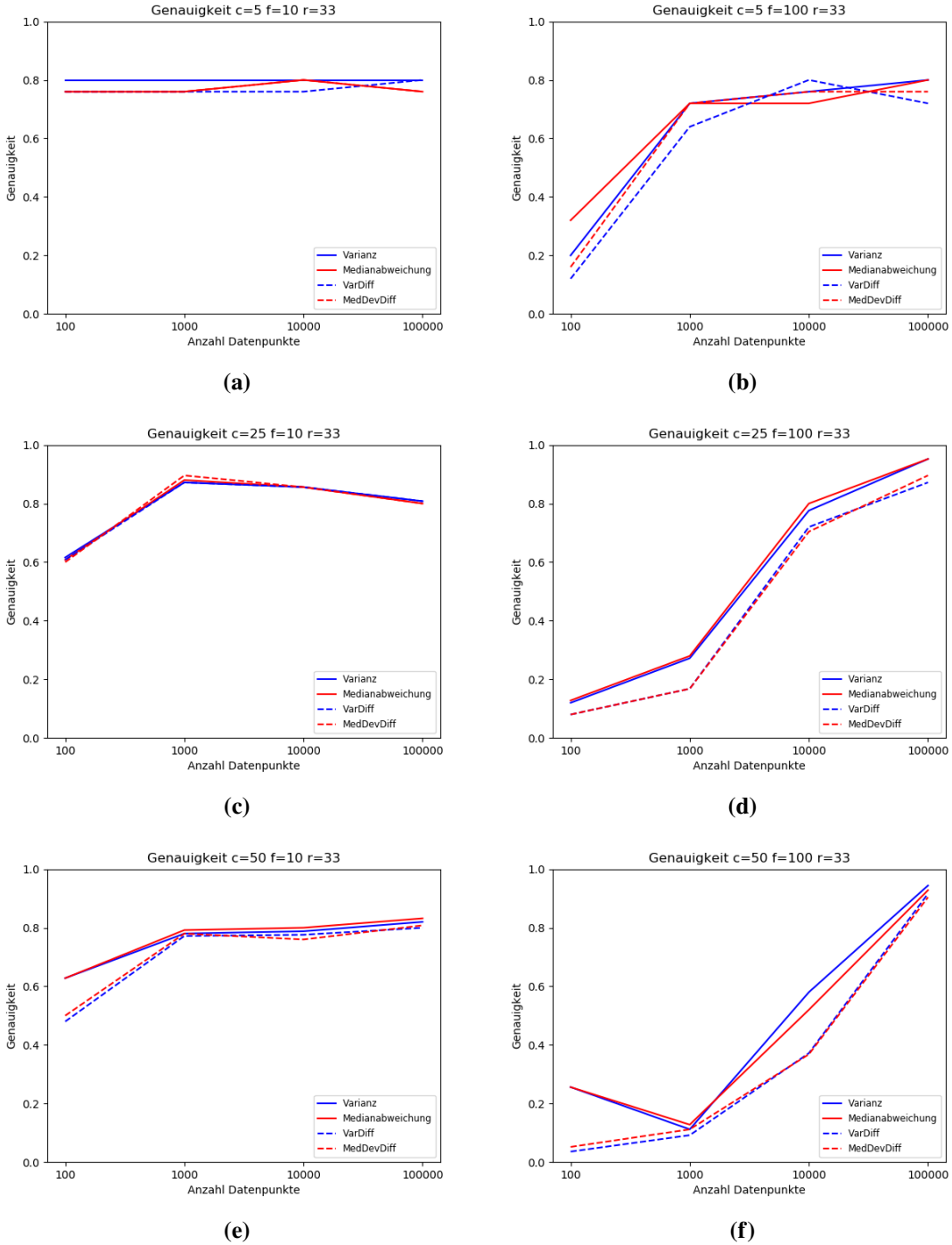


Abbildung 5.7: Vergleich der Durchschnittsgenauigkeiten von Varianz, Medianabweichung, Varianz-Differenz und Medianabweichungs-Differenz in Abhängigkeit der Anzahl Datenpunkte, Cluster und Features

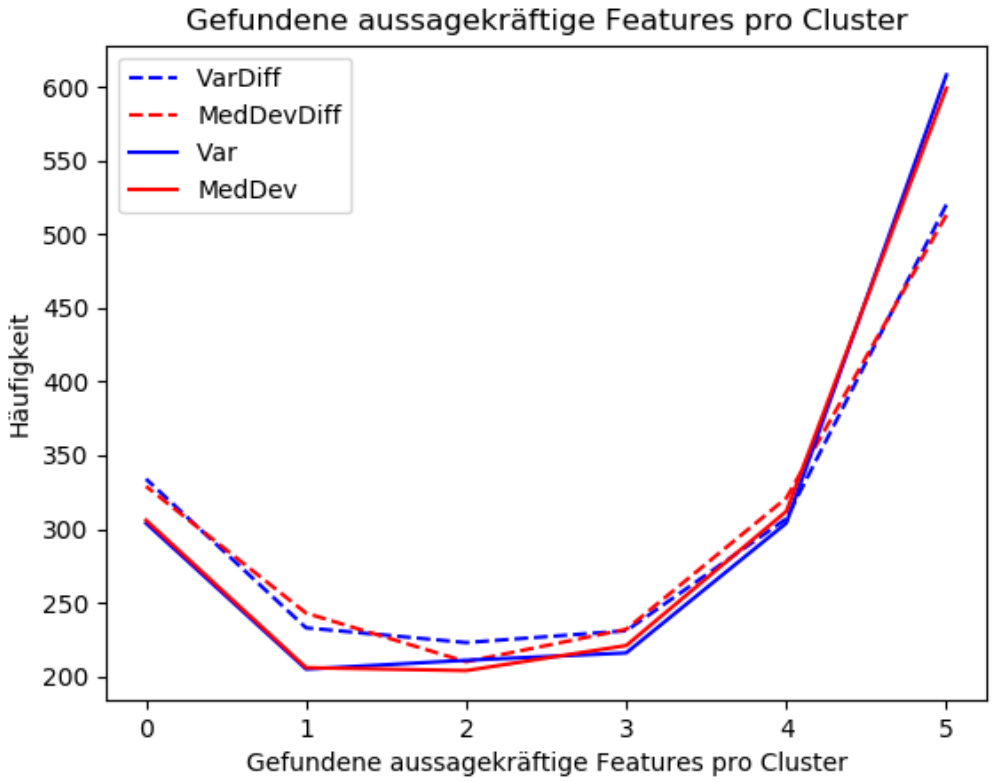
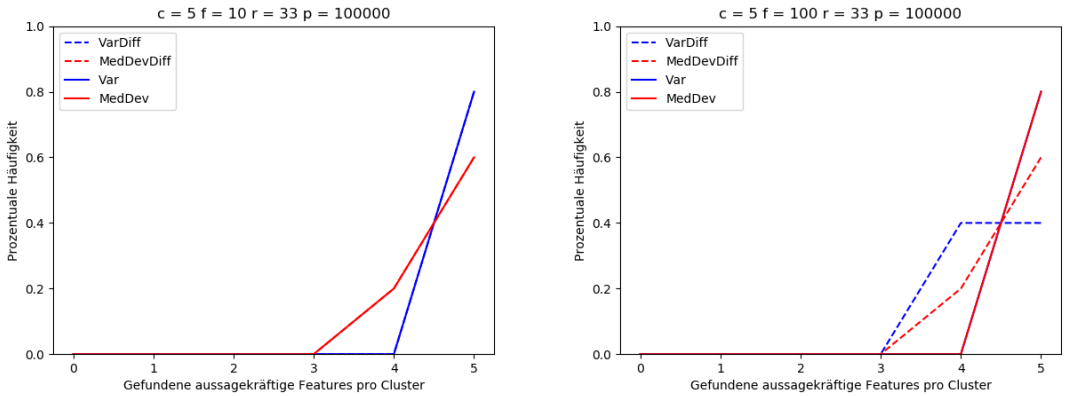


Abbildung 5.8: Anzahl gefundener aussagekräftiger Features pro Cluster über alle Datensätze ohne konstanten hinweg

Betrachtung einzelner Cluster

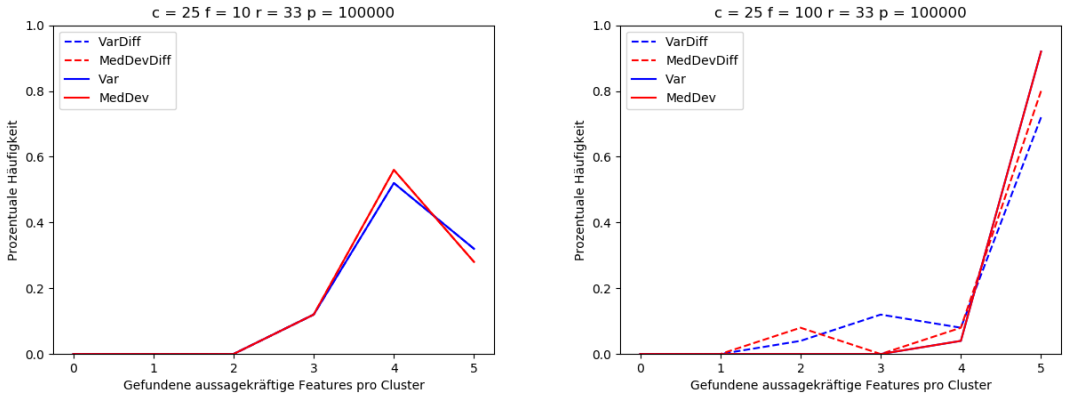
Für diese Betrachtung wurde zunächst Abbildung 5.8 erzeugt um darzustellen, wie oft über alle Datensätze ohne Konstanten, die verschiedenen Metriken eine bestimmte Anzahl an aussagekräftigen Features pro Cluster gefunden haben. Dabei haben alle dargestellten Metriken am häufigsten alle 5 aussagekräftige Features pro Cluster gefunden. Auffällig ist dabei, dass die Kurven der Metrik-Differenzen für weniger als 5 gefundene aussagekräftige Features pro Cluster über den Kurven der Standardmetriken liegen. Um die Abhängigkeiten der Parameter der Datensätze auf die gefundenen aussagekräftigen Features pro Cluster zu untersuchen, wurde zusätzlich Abbildung 5.9 erzeugt.

Abbildung 5.9 zeigt den Verlauf der Anzahl gefundener aussagekräftiger Features pro Cluster und Metrik in Abhängigkeit der prozentualen Häufigkeit in diesem Datensatz. Eine Spalte der Graphen entspricht dabei einer konstanten Anzahl an Features und eine Zeile einer konstanten Anzahl an Clustern. Über alle Datensätze hinweg sind die Rauschwerte 33% und die Anzahl Datenpunkte 100000. Der Rauschwert wurde so gewählt, da er damit demselben entspricht wie bei der Betrachtung der Durchschnittsgenauigkeit. Zudem wurde die Anzahl an Datenpunkten möglichst hoch gewählt, da nach der Durchschnittsgenauigkeit dafür die Unterschiede zwischen den Metriken am größten sind. Dabei ist auffällig, dass in der linken Spalte der Unterschied zwischen den Standardmetriken und Metrik-Differenzen kaum bis gar nicht erkennbar ist. Sichtbar wird dieser



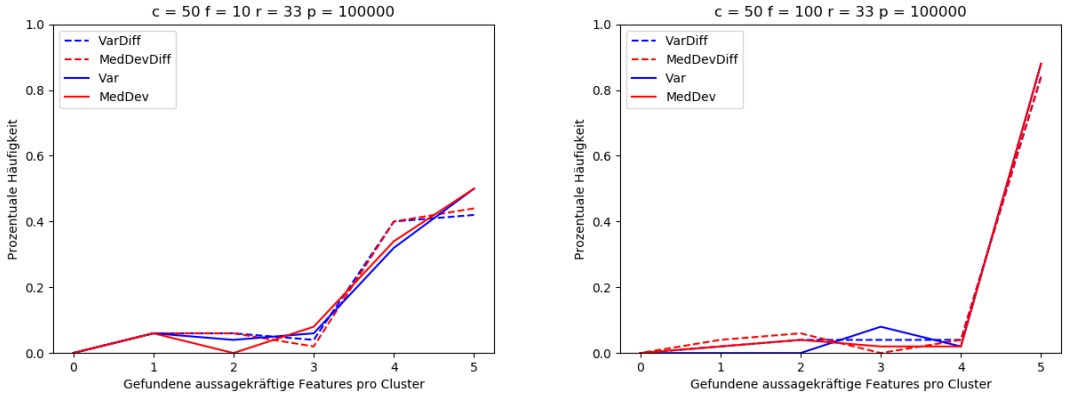
(a)

(b)



(c)

(d)



(e)

(f)

Abbildung 5.9: Diese Graphen zeigen Verläufe der gefundenen aussagekräftigen Features pro Cluster in Abhängigkeit der prozentualen Häufigkeit

nur bei dem Datensatz mit 10 Features und 50 Clustern, dabei entsprechen die Ergebnisse der Standardmetriken etwa den gemittelten Ergebnissen der Metrik-Differenzen. Die Anzahl gefundener aussagekräftiger Features bleibt somit nahezu gleich.

Für die rechte Spalte, und somit 100 Features, ist der Unterschied zwischen den Standardmetriken und den Metrik-Differenzen deutlich erkennbar. Dabei fällt auf, dass über alle dargestellten Datensätze hinweg die Metrik-Differenzen seltener 5 aussagekräftige Features pro Cluster finden als die Standardmetriken. Dass die Ergebnisse der Metrik-Differenzen dennoch nicht wesentlich ungenauer sind, als die der Standardmetriken wird an einer entsprechend höheren Anzahl von 4 und 3 gefundenen aussagekräftigen Features pro Cluster deutlich.

Zudem lässt sich durch diese Abbildung die Abhängigkeit der Genauigkeit von der Anzahl an Features genauer feststellen. Um dies auf der Ebene der einzelnen Cluster zu betrachten, werden die Graphen spaltenweise verglichen. Daraus ergibt sich, dass alle betrachteten Metriken für 100 Features bis zu 50% häufiger 5 aussagekräftigen Features in einem Cluster gefunden haben als bei 10 Features.

Allgemein lässt sich sowohl aus der Betrachtung der Durchschnittsgenauigkeit als auch aus der Betrachtung einzelner Cluster schließen, dass die Standardmetriken genauer sind als die Metrik-Differenzen. Dabei hängen die Unterschiede dieser zwei Klassen nur gering von den Parametern der Datensätze ab. Somit lässt sich auch für größere Datensätze ohne Konstanten vermuten, dass die Standardmetriken zu genaueren Ergebnissen führen.

Einfluss des Clustering-Verfahrens

Analog zu der Betrachtung der Datensätze mit Konstanten aus Abschnitt 5.2.1 wurde für die Datensätze ohne Konstanten Abbildung 5.10 erzeugt. Diese enthält den Verlauf der Varianz und Medianabweichung in Abhängigkeit des AMIS. Die Betrachtung der entsprechenden Metrik-Differenzen wurde bereits in Abschnitt 5.2.1 diskutiert. Dabei besagt ein hoher AMIS eine große Übereinstimmung zwischen dem initialen Clustering-Resultat und dem verrauschten Clustering-Resultat aus Abschnitt 5.1.1. Somit lässt sich in dieser Abbildung einen positiven Zusammenhang zwischen der Genauigkeit und dem AMIS feststellen. Aufgrund der starken Ähnlichkeit dieser Ergebnisse zu denen der Metrik-Differenzen wird für eine genauere Untersuchung auf Abschnitt 5.2.1 verwiesen.

5.3 Evaluation der Laufzeit

Für die Laufzeit wird untersucht, wie stark sich die verschiedenen Parameter Anzahl Datenpunkte, Anzahl Cluster und Anzahl Features auf die Laufzeit verschiedener Metriken innerhalb des Ansatzes auswirken. Die Anzahl verrauschter Datenpunkte kann vernachlässigt werden, da diese aus Sicht der Laufzeit nur die Anzahl an Datenpunkte erhöht. Es genügt somit, die Auswirkung der Anzahl an Datenpunkten auf die Laufzeit zu betrachten.

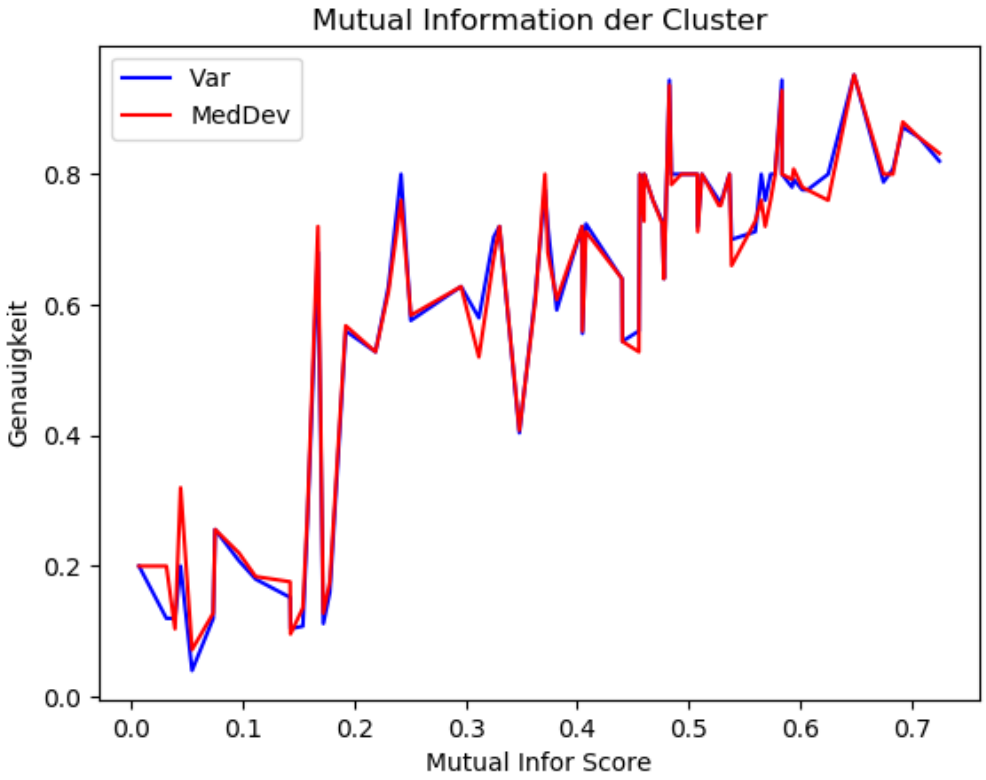


Abbildung 5.10: Verlauf der Genauigkeiten der Varianz und Medianabweichung in Abhängigkeit des AMIS

5.3.1 Ermittlung aussagekräftiger Features

Algorithmus 5.1 beschreibt die Ermittlung aussagekräftiger Features in Pseudocode. Aus diesem geht hervor, dass ein linearer Zusammenhang zwischen der Laufzeit und der Anzahl Cluster c und Anzahl Features f besteht. Zudem hängt die Laufzeit maßgeblich von der ausgewählten Metrik ab, da diese bestimmt, wie stark sich die Anzahl Datenpunkte p auf die Laufzeit auswirken. Die in dieser Arbeit ausgewählten Metriken haben alle eine lineare Laufzeit in Bezug auf die Anzahl Datenpunkte (vgl. Abschnitt 2.4). Dies ergibt eine Gesamtlaufzeit in $O(c \cdot f \cdot p)$.

Die Abbildung 5.11 zeigt Graphen für den Verlauf der Laufzeit der Metriken in Abhängigkeit der Anzahl an Datenpunkten für verschiedene Anzahlen an Cluster und Features. Die Spalten stellen darin unterschiedliche Anzahlen an Clustern und die Zeilen unterschiedlich viele Features dar. Dabei sind die Standardmetriken mit durchgezogenen und die Metrik-Differenzen mit gestrichelten Linien dargestellt. Die Laufzeit wird hierbei beschrieben durch

$$Laufzeit_{Metrik\ m,all} = (\sum_{i=1}^{c} Laufzeit_{Metrik\ m,Cluster\ i}) \tag{5.2}$$

Algorithmus 5.1 findNMeaningfulFeatures

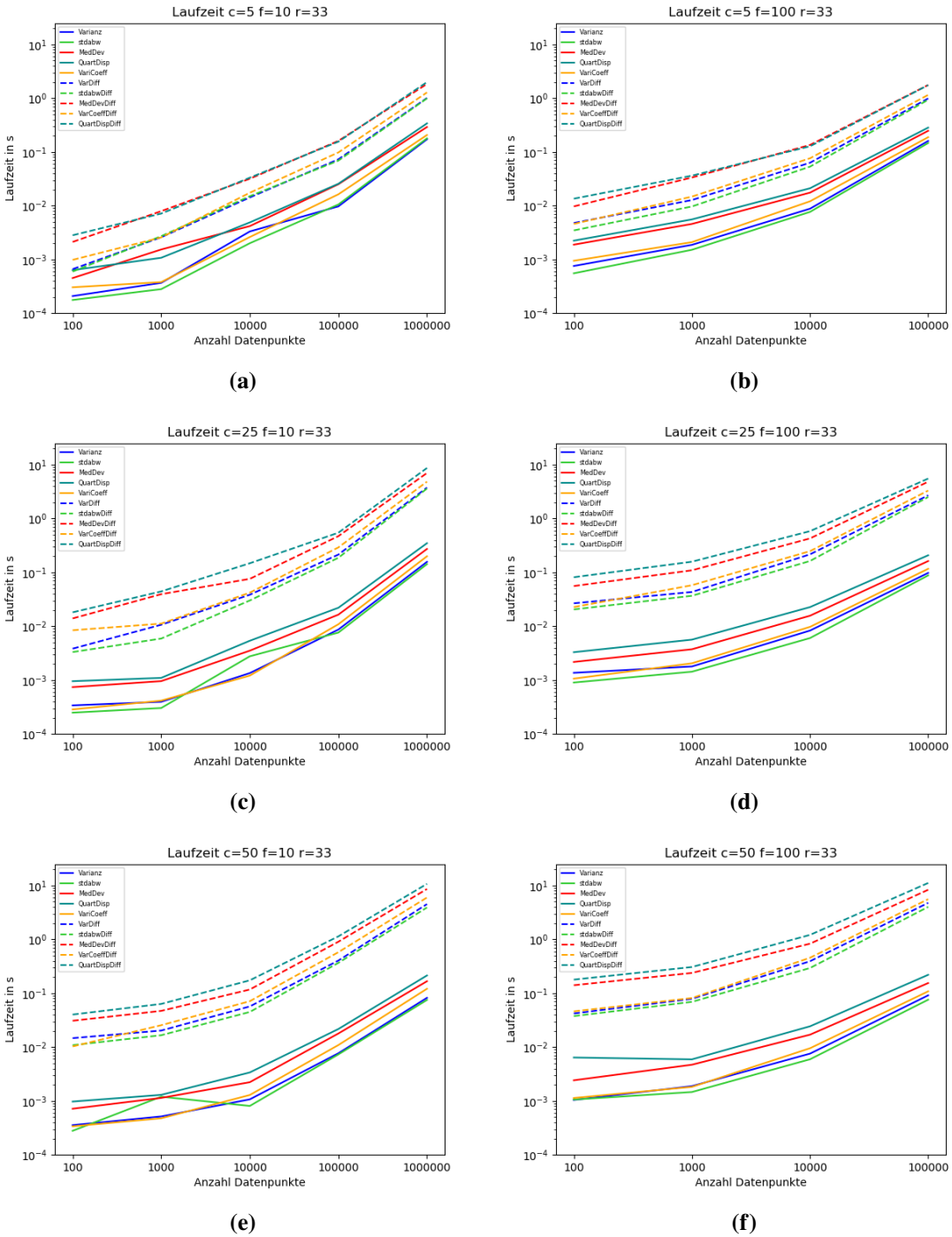
```
procedure findNMeaningfulFeatures(points,n)
    bestList ← emptyList(Size = n)
    for all c ∈ Clusters do
        for all f ∈ Features do
            points ← getPoints(c)
            m ← Metric(points, f)
            if m < one in bestList then
                replaceWorst(bestList, m)
            end if
        end for
    end for
    return bestList
end procedure
```

Diese Graphen bestätigen den linearen Zusammenhang zwischen der Anzahl an Datenpunkten und der Laufzeit. Deutlich zu sehen ist in dieser Abbildung der Unterschied der Laufzeiten zwischen den Standardmetriken und den Metrik-Differenzen. Nach Betrachtung der Formel für die Metrik-Differenzen aus Abschnitt 2.4 ist dieses Phänomen bereits ableitbar. Die Metrik-Differenzen benötigen neben der Berechnung der Standardmetriken zusätzlich die Berechnung der Metriken für den ganzen Datensatz und das Bilden der Differenz pro Cluster. Daraus ergibt sich, dass der Unterschied der Laufzeiten zwischen den Standardmetriken und den Metrik-Differenzen größer wird, je größer die Anzahl an Clustern ist. Diese Überlegung lässt sich durch den Vergleich der Graphen in den Spalten der Abbildung 5.11 bestätigen. Zudem ergibt sich aus der Definition der Metrik-Differenzen und Algorithmus 5.1 dass der Unterschied zwischen Standardmetriken und Metrik-Differenzen für eine steigende Anzahl an Features steigt. Dies liegt daran, dass für jedes dieser Features zunächst die Metrik über den ganzen Datensatz bestimmt werden muss, was bei den Standardmetriken nicht benötigt wird.

Zudem ist in allen Graphen aus Abbildung 5.11 zu sehen, dass der Verlauf der Laufzeit der Standardmetriken nahezu identisch zu dem der passenden Metrik-Differenz ist. So ist zum Beispiel die Laufzeit der Medianabweichungs-Differenz für alle Punkte um etwa eine Konstante größer, als die der Medianabweichung. Dies bedeutet, dass die Anzahl an Datenpunkten keinen negativen Einfluss auf die Metrik-Differenzen im Gegensatz zu den Standardmetriken aufweist.

Aus dem vorherigen Abschnitt geht jedoch hervor, dass die Metrik-Differenzen immer dann genauer sind, sofern sich eine Konstante in dem Datensatz befindet. Dies bedeutet, dass die Standardmetriken die gleichen Resultate wie die Metrik-Differenzen liefern, sofern vor der Berechnung der Datensatz auf Konstanten überprüft wurde. Falls dabei Konstanten gefunden wurden, so werden diese für die weitere Berechnung ignoriert. Somit sollten die Metrik-Differenzen immer dann verwendet werden, wenn

$$Laufzeit_{MetrikDiff\ m,all} < Laufzeit_{Metrik\ m,all} + TesteDatensatzAufKonstanten \quad (5.3)$$



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Abbildung 5.11: Vergleich der Laufzeiten der Standardmetriken (durchgezogen) und Metrik-Differenzen (gestrichelt) in Abhängigkeit der Anzahl Datenpunkte, Cluster und Features

Dabei ist zu beachten, dass die Laufzeit des Tests auf Konstanten von der Speicherung des Datensatzes abhängt. Zum Beispiel benötigen verschiedene Datenbanken unterschiedlich viel Zeit für eine solche Funktion. Dies sorgt dafür, dass keine allgemeine Aussage darüber möglich ist, ob das Finden der Konstanten mit anschließender Berechnung der Standardmetriken schneller ist als die Berechnung der Metrik-Differenzen. Dies muss somit für jedes System separat geprüft werden.

5.3.2 Berechnung anzuzeigender Werte

Die Laufzeit der Berechnung der anzuzeigenden Werte hängt maßgeblich von der gewählten Metrik ab. Für die in Abschnitt 4.2.2 vorgestellten Metrik Minimum/Maximum liegt die Laufzeit mit p als die Anzahl an Datenpunkten in $O(p)$. Dass dies auch für das Oberes-/Unteres-Quartil gilt, geht aus Abschnitt 4.2.1 hervor.

5.3.3 Darstellung der Ergebnisse

Sowie die Laufzeit der Berechnung der anzuzeigenden Werte ist auch die der Darstellung der Ergebnisse stark von der gewählten Art ab. Die in Abschnitt 4.2.3 thematisierten Ansätze der Tabelle und Wortwolke hängen von der Anzahl an Features und deren Anzahl an Werten die angezeigt werden ab. Da bereits die Anzahl an Features und die Anzahl an Datenpunkte reduziert wurde, kann die Laufzeit nach oben abschätzen werden durch $O(f \cdot p)$, mit p als die Anzahl an Datenpunkten und f als die Anzahl an Features.

Dies führt zu der Gesamtlaufzeit in

$$O(c \cdot f \cdot p) + O(p) + O(f \cdot p) = O(c \cdot f \cdot p) \tag{5.4}$$

6 Zusammenfassung und Ausblick

Die Menge an erzeugten Daten wächst rasant über die Jahre hinweg. Diese Masse an Daten ist für den Menschen nicht mehr wahrnehmbar, wodurch ein großer Teil dieser nicht analysiert werden kann. Dennoch ist es aus der Sicht verschiedener Bereiche wie etwa Medizin und Wirtschaft wichtig diese Daten zu verarbeiten. Dies sorgt für die Notwendigkeit von Verfahren, welche Daten verarbeiten und somit für den Menschen verständlich machen. Bisherige Ansätze zur Verarbeitung der Daten basieren auf Dimensionsreduktion oder einer textuellen Beschreibung des Modells. Bei der Dimensionsreduktion werden hochdimensionale Datensätze auf eine für den Menschen verständliche Anzahl an Dimensionen reduziert und innerhalb eines Koordinatensystems visualisiert. Darin lassen sich häufig Gruppierungen der Daten erkennen, detaillierte Informationen zu diesen Gruppen sind jedoch daraus ohne Weiteres nicht ableitbar. Die textuellen Beschreibungen finden sich häufig in Data-Mining-Werkzeugen und liefern ausgewählte Statistiken zu den Clustern. Detaillierte Informationen zu der Aussagekraft verschiedener Features oder zu den Unterschieden der Cluster sind darin nicht enthalten.

Durch den in dieser Arbeit vorgestellten Ansatz wird es ermöglicht, Clustering-Resultate zu verarbeiten und darzustellen. Dadurch erhält ein Analyst Informationen über einzelne Cluster, die Unterschiede der Cluster oder das gesamte Clustering-Resultat. Erreicht wird dies durch die Auswahl aussagekräftiger Features, welche durch Streuungsmaße bestimmt werden können. Die Auswahl der Streuungsmaße beruht auf der Herangehensweise von Clustering-Verfahren, den Abstand der Datenpunkte innerhalb der Cluster zu minimieren. Um zusätzlich zu den positiven Eigenschaften der Streuungsmaße eine Resistenz gegenüber Konstanten zu erhalten, wurden Metrik-Differenzen entwickelt. Weiterhin wurden unter Beachtung der menschlichen Wahrnehmung verschiedene Ansätze entwickelt, um die Anzahl auszuwählender aussagekräftiger Features zu bestimmen. Auf diesen Ansätzen aufbauend wurden Möglichkeiten vorgestellt, welche Details zu den ausgewählten aussagekräftigen Features darstellen.

Dieser Ansatz wurde anschließend in Hinblick auf die Genauigkeit und Laufzeit bei Verwendung verschiedener Streuungsmaße und Metrik-Differenzen gegen einen Goldstandard evaluiert. Hierbei zeigte sich, dass für Datensätze ohne Konstanten die Streuungsmaße Varianz und Medianabweichung am genausten sind. Die Varianz-Differenz und Medianabweichungs-Differenz erzielen vergleichbare Ergebnisse, sind jedoch im Allgemeinen für diese Datensätze um etwa 5 - 10% ungenauer. Eine genauere Betrachtung hat gezeigt, dass die Genauigkeit der Varianz und Medianabweichung steigt, wenn die Anzahl an Datenpunkten oder Clustern steigt. Dem entgegengesetzt sinkt die Genauigkeit, wenn die Anzahl an Features steigt. Für Datensätze mit Konstanten konnte gezeigt werden, dass von den ausgewählten Metriken Varianz-Differenz und Medianabweichungs-Differenz am genausten sind. Die Genauigkeit der Metrik-Differenzen besitzt dabei vergleichbare Abhängigkeiten gegenüber den Parametern der Datensätze, wie die Streuungsmaße.

Zudem wurde der Einfluss des Clusterings auf die Genauigkeit dieser Metriken geprüft. Dies lässt die Tendenz zu, dass die Genauigkeit des entwickelten Ansatzes ebenfalls von der Qualität des Clustering-Verfahrens abhängt. Eine genauere Aussage lässt sich jedoch auf Basis der durchgeführten Evaluation nicht treffen.

Die Betrachtung der Laufzeit ergab eine lineare Abhängigkeit von den Parametern der Datensätze. Damit ist die Laufzeit dieses Ansatzes wesentlich geringer als die der Dimensionsreduktions-Ansätze. Zudem bedingt das Berechnen der Metrik-Differenz eine höhere Laufzeit als die der Streuungsmaße. Demnach kann es für manche Systeme sein, dass das Finden von Konstanten und das anschließende Berechnen der Streuungsmaße eine geringere Laufzeit aufweist, als das berechnen der Metrik-Differenzen. Dies ist maßgeblich davon abhängig, wie lange das System für das Finden von Konstanten im Datensatz benötigt, wodurch keine allgemeine Aussage darüber möglich ist, welcher der beiden Ansätze verfolgt werden soll.

Ausblick

Neben der Betrachtung weiterer Streuungsmaße kann für die Feature-Selektion (S1) des Ansatzes die Überschneidung der Wertebereiche der Features untersucht werden. Es ist denkbar, dass eine solche Metrik vor allem geeignet ist, um Features zu finden, welche die Cluster unterscheiden. Demnach könnte diese Metrik potenziell zu Ergebnissen führen, welche tiefere Einblicke in die Clustering-Verfahren ermöglichen. Zudem lassen sich noch weitere Alternativen erarbeiten und vergleichen. So kann beispielsweise festgestellt werden, für welche Domänen welche Kenngrößen oder Darstellungsformen am geeignetsten sind. Durch diese Erkenntnisse kann das Programm ideal an die Domäne angepasst werden, wodurch die Resultate optimiert werden.

Um quantifizierte Vergleiche zwischen verwandten Arbeiten und diesem Ansatz zu erhalten, kann eine Nutzerstudie durchgeführt werden. Aufbauend auf diesen Ergebnissen lassen sich voraussichtlich weitere Darstellungsformen entwickeln, welche speziell an die Anwendung und den Analysten angepasst sind. Durch diese angepasste Darstellung könnte sowohl die nötige Zeit zum Analysieren der Daten als auch die Anzahl an Fehlinterpretationen sinken.

Ein weiterer unbeachteter Punkt ist der Zusammenhang zwischen diesem Ansatz und der Ermittlung einer geeigneten Anzahl an Clustern für einen Datensatz. So könnte ein Datensatz für verschiedene Mengen von Zentroiden geclustert und anschließend die Ergebnisse des Ansatzes verglichen werden. Möglicherweise lässt sich anhand der Ergebnisse des Ansatzes feststellen, welche Anzahl an Clustern am geeignetsten für den Datensatz waren.

Literaturverzeichnis

[ACLS12] R. Arora, A. Cotter, K. Livescu, N. Srebro. „Stochastic optimization for PCA and PLS“. In: *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2012, S. 861–868 (zitiert auf S. 25).

[BFP+73] M. Blum, R. W. Floyd, V. R. Pratt, R. L. Rivest, R. E. Tarjan. „Time bounds for selection“. In: *J. Comput. Syst. Sci.* 7.4 (1973), S. 448–461 (zitiert auf S. 23).

[BGN08] S. Bateman, C. Gutwin, M. Nacenta. „Seeing things in the clouds: the effect of visual features on tag cloud selections“. In: *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*. ACM. 2008, S. 193–202 (zitiert auf S. 20, 21).

[Cen] I. K. Center. *Clustering Visualizer*. URL: https://www.ibm.com/support/knowledgecenter/en/SSEPGG_9.7.0/com.ibm.im.visual.doc/c_introducing_the_clusteirng_visualizer.html (zitiert auf S. 30, 31).

[CH90] K. W. Church, P. Hanks. „Word association norms, mutual information, and lexicography“. In: *Computational linguistics* 16.1 (1990), S. 22–29 (zitiert auf S. 19).

[EKS+96] M. Ester, H.-P. Kriegel, J. Sander, X. Xu et al. „A density-based algorithm for discovering clusters in large spatial databases with noise.“ In: *Kdd*. Bd. 96. 34. 1996, S. 226–231 (zitiert auf S. 16).

[EKT08] H.-F. Eckey, R. Kosfeld, M. Türck. *Deskriptive Statistik*. Springer, 2008 (zitiert auf S. 22, 23).

[FBS19] M. Fritz, M. Behringer, H. Schwarz. „Quality-driven early stopping for explorative cluster analysis for big data“. In: *SICS Software-Intensive Cyber-Physical Systems* (2019), S. 1–12 (zitiert auf S. 17).

[GMW07] G. Gan, C. Ma, J. Wu. *Data clustering: theory, algorithms, and applications*. Bd. 20. Siam, 2007 (zitiert auf S. 15).

[GR12] J. Gantz, D. Reinsel. „The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east“. In: *IDC iView: IDC Analyze the future* (2012), S. 1–16 (zitiert auf S. 13).

[HFH+09] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. „The WEKA data mining software: an update“. In: *ACM SIGKDD explorations newsletter* 11.1 (2009), S. 10–18 (zitiert auf S. 28).

[HR03] G. E. Hinton, S. T. Roweis. „Stochastic neighbor embedding“. In: *Advances in neural information processing systems*. 2003, S. 857–864 (zitiert auf S. 26).

[JMF99] A. K. Jain, M. N. Murty, P. J. Flynn. „Data clustering: a review“. In: *ACM computing surveys (CSUR)* 31.3 (1999), S. 264–323 (zitiert auf S. 15–17).

[KR09] L. Kaufman, P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Bd. 344. John Wiley & Sons, 2009 (zitiert auf S. 15, 16).

[Lee00] L. Lee. „Measures of distributional similarity“. In: *CoRR* (2000) (zitiert auf S. 17).

[Llo82] S. Lloyd. „Least squares quantization in PCM“. In: *IEEE transactions on information theory* 28.2 (1982), S. 129–137 (zitiert auf S. 17).

[LZT09] S. Lohmann, J. Ziegler, L. Tetzlaff. „Comparison of tag cloud layouts: Task-related performance and visual exploration“. In: *IFIP Conference on Human-Computer Interaction*. Springer. 2009, S. 392–404 (zitiert auf S. 20).

[Mac+67] J. MacQueen et al. „Some methods for classification and analysis of multivariate observations“. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Bd. 1. 14. Oakland, CA, USA. 1967, S. 281–297 (zitiert auf S. 17).

[May08] A. L. Mayer. „Strengths and weaknesses of common sustainability indices for multidimensional systems“. In: *Environment international* 34.2 (2008), S. 277–291 (zitiert auf S. 26).

[MH08] L. v. d. Maaten, G. Hinton. „Visualizing data using t-SNE“. In: *Journal of machine learning research* 9.Nov (2008), S. 2579–2605 (zitiert auf S. 26, 27).

[Mil56] G. A. Miller. „The magical number seven, plus or minus two: Some limits on our capacity for processing information.“ In: *Psychological review* 63.2 (1956), S. 81 (zitiert auf S. 19).

[MR09] O. Maimon, L. Rokach. „Introduction to knowledge discovery and data mining“. In: *Data Mining and Knowledge Discovery Handbook*. Springer, 2009, S. 1–15 (zitiert auf S. 13).

[Olk02] G. C. S. F. I. Olkin. „Springer Texts in Statistics“. In: (2002) (zitiert auf S. 25).

[Pol52] I. Pollack. „The information of elementary auditory displays“. In: *The Journal of the Acoustical Society of America* 24.6 (1952), S. 745–749 (zitiert auf S. 19, 20).

[Rou87] P. J. Rousseeuw. „Silhouettes: a graphical aid to the interpretation and validation of cluster analysis“. In: *Journal of computational and applied mathematics* 20 (1987), S. 53–65 (zitiert auf S. 17).

[VEB10] N. X. Vinh, J. Epps, J. Bailey. „Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance“. In: *Journal of Machine Learning Research* 11.Oct (2010), S. 2837–2854 (zitiert auf S. 17–19).

[Von93] P. Von der Lippe. *Deskriptive Statistik*. Walter de Gruyter GmbH & Co KG, 1993 (zitiert auf S. 21, 22).

[WEG87] S. Wold, K. Esbensen, P. Geladi. „Principal component analysis“. In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), S. 37–52 (zitiert auf S. 25, 26).

[Wei09] C. Weigand. „Spezielle Verteilungen“. In: *Statistik mit und ohne Zufall*. Springer, 2009, S. 181–219 (zitiert auf S. 50).

Alle URLs wurden zuletzt am 30. 03. 2019 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