technische universität dortmund

IT in Produktion und Logistik

Bachelorarbeit

Systematische Analyse der Behandlung von Rauschen in der Datenvorverarbeitung bei der Wissensentdeckung in Datenbanken

Zur Erlangung des Grades Bachelor of Science (B. Sc.)

Name:	Akin Recber
Matrikelnummer:	162625
Studiengang:	Maschinenbau
Ausgabedatum:	16.11.2022
Abgabedatum:	13.02.2023

Erstprüfer:	Dr.-Ing. Dipl.-Inform. Anne Antonia Scheidler
Zweitprüfer:	M. Sc. Florian Hochkamp

Technische Universität Dortmund

Fakultät Maschinenbau

Fachgebiet IT in Produktion und Logistik

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

KDD	Knowledge Discovery in Database
KI	Künstliche Intelligenz
AI	Artificial Intelligence
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
kNN	k-Nächste-Nachbar

1 Einleitung

Wissensentdeckung in Datenbanken oder auf Englisch *Knowledge Discovery in Databases* (KDD) gewinnt immer mehr an Bedeutung, da es Unternehmen ermöglicht, profitable Muster und Trends aus ihren bestehenden Datenbanken zu erkennen (Larose und Larose 2014). Nach Runkler (2016) werden die Daten in vielen verschiedenen Bereichen eines Unternehmens verwendet, wie z. B. industrielle Prozessdaten, Geschäftsdaten, Textdaten und strukturierte Daten oder auch Bild- und Videodaten. Jeden Tag werden dabei Petabytes an Daten erzeugt und gespeichert, was zu einer enormen Menge an Informationen führt (Luengo et al. 2020). Aufgrund der hohen Datenmenge ist eine manuelle Auswertung der Daten nicht möglich und daher nutzen viele Unternehmen Data-Mining-Verfahren, um ihre Datenbanken zu analysieren. Data Mining wird in der Literatur oft als Synonym für KDD verwendet. In dieser Arbeit jedoch wird Data Mining als einzelne Phase im KDD-Prozess gesehen. Der KDD-Prozess lässt sich in 5 Phasen unterteilen, welche Datenselektion, Datenvorverarbeitung, Transformation, Data Mining und Interpretation und Evaluierung sind (Fayyad et al. 1996). Dabei ist laut Aggarwal (2015) die Phase der Datenvorverarbeitung vielleicht die wichtigste Phase im KDD-Prozess, da die Datenvorverarbeitung die Datenqualität verbessert und diese wiederum sichert die Qualität der Ergebnisse der Data-Mining-Phase. Oft ist es so, dass die Datenqualität in realen Anwendungen nicht gut genug ist, da die Daten Fehler oder Rauschen enthalten (Runkler 2016). Besonders das Auftreten von verrauschten Daten in einem Datensatz kann die Vorhersageergebnisse erheblich beeinträchtigen (Gupta und Gupta 2019). Die Frage nach einem geeigneten Verfahren zur Behandlung von verrauschten Daten ist für viele Unternehmen von besonderem Interesse. Den Schwerpunkt der Arbeit möchte ich daher in der Behandlung von Rauschen in der Datenvorverarbeitung legen.

Das Hauptziel dieser Arbeit ist es, die Suche nach geeigneten Data-Mining-Verfahren zur Behandlung von Rauschen in der Datenvorverarbeitung, durch eine systematische Literaturrecherche und die Entwicklung eines Entscheidungsmodells zu unterstützen. Dazu wird das Hauptziel in zwei Teilziele unterteilt. Das erste Teilziel dieser Arbeit umfasst eine strukturierte Literaturrecherche, um eine umfangreiche Übersicht der nutzbaren Data-Mining-Verfahren zur Behandlung von Rauschen bei der Datenvorverarbeitung zu erstellen. Das zweite Teilziel behandelt den Entwurf eines Entscheidungsmodells, welches auf Basis von aufgabenbezogenen Anforderungen die Auswahl der Data-Mining-Verfahren unterstützen soll.

Um das Hauptziel zu erreichen, wird die Arbeit wie folgt gegliedert: Als erstes wird in Kapitel 2 der Stand der Technik bezüglich der Wissensentdeckung in Datenbanken thematisiert. Damit wird zum einen die Relevanz der Datenvorverarbeitung hervorgehoben und zum anderen die

Problematik von Rauschen eingeführt. Darauf folgt in Kapitel 3 die Methodik der systematischen Literaturrecherche. Die systematische Literaturrecherche leitet den eigentlichen Hauptteil dieser Arbeit ein. Dazu werden zunächst die Ein- und Ausschlusskriterien festgelegt und anhand der Fragestellung geeignete Suchterme abgeleitet. Die Suchergebnisse werden dann zusammengefasst und in Kapitel 4 dargestellt. Dabei werden die die Data-Mining-Verfahren zunächst anhand der Literaturrecherche kategorisiert. Nach der Kategorisierung folgt eine genaue Beschreibung der einzelnen Data-Mining-Verfahren, die zur Behandlung von verrauschten Daten geeignet sind. Dieses Vorgehen ermöglicht es dem Leser weitere Data-Mining-Verfahren, die in dieser Arbeit nicht behandelt wurden, einzuordnen und schafft zugleich eine strukturierte Übersicht. Mithilfe der Informationen, die aus den vorangegangenen Kapiteln gewonnen wurden, wird die Entwicklung des Entscheidungsmodells thematisiert. Um den Aufbau des Modells zu begründen, müssen zunächst Anforderungen an das Modell abgeleitet werden. Im Anschluss an den Aufbau des Modells folgt schließlich die Evaluation. Die Evaluation wird exemplarisch an einem Beispiel erläutert. Damit sind die beiden Teilziele erreicht und es folgt eine kurze Zusammenfassung der Ergebnisse sowie ein Ausblick auf zukünftige Forschungsthemen.

2 Wissensentdeckung in Datenbanken

Wissensentdeckung in Datenbanken ist eine relevante Aufgabe vieler Unternehmen. Mithilfe von Data-Mining-Verfahren können die Unternehmen ihre Datenbanken analysieren und wichtige Informationen erhalten. Diese Informationen werden dann verwendet, um z. B. Vorhersagen zu treffen oder den Geschäftsprozess zu optimieren. Um den Prozess der Wissensentdeckung in Datenbanken zu verstehen, werden im Rahmen dieses Kapitels die erforderlichen Grundlagen sowie die Probleme der Wissensentdeckung in Datenbanken erläutert beziehungsweise verdeutlicht. Dazu werden zunächst Grundbegriffe wie zum Beispiel Daten oder Wissen erklärt. Anschließend wird das Vorgehensmodell von Fayyad vorgestellt. Im Anschluss, wird die Phase der „Datenvorverarbeitung“ noch etwas detaillierter behandelt, da auf Grundlage dieser Phase die Untersuchung der Forschungsfrage dieser Arbeit erfolgt.

2.1 Grundbegriffe

Zu Beginn dieses Kapitels werden zunächst einige Grundbegriffe eingeführt. Begriffe werden in der Literatur oft unterschiedlich definiert, daher werden in diesem Unterkapitel die Definitionen der Grundbegriffe, die für diese Arbeit relevant sind, zunächst vorgestellt, sodass feststeht, welche Definitionen für diese Arbeit verwendet werden.

2.1.1 Daten und Datentypen

Daten:

Besonders oft werden in der Literatur die Begriffe Daten, Informationen und Wissen zusammen dargestellt und in einer Hierarchie angeordnet (siehe Abbildung 2-1). Auch Bodendorf (2006) benutzt diese Hierarchie und definiert folgendermaßen:

„*Daten werden aus Zeichen eines Zeichenvorrats nach definierten Syntaxregeln gebildet*“ (Bodendorf 2006, S. 8).

Das bedeutet, dass die unterste Ebene dieser Hierarchie die Zeichen sind. Dabei können Zeichen einzelne Buchstaben oder Ziffern sein (Bodendorf 2006). Die Anordnung dieser Zeichen anhand von bestimmten Syntaxregeln bilden dann die Daten.

Bodendorf führt seinen Gedanken fort und sagt, dass Daten erst dann zu Informationen werden, wenn ihnen eine Bedeutung (Semantik) zugeordnet wird und diese Zuordnung dadurch erfolgt, dass die Daten mit einem Begriff oder einer Vorstellung aus der realen Welt oder theoretischer Art assoziiert werden. Daten können also als eine Darstellungsform von

Informationen gesehen werden. Erst durch eine Interpretation der Daten, können Informationen entstehen.

Bei dem Begriff Wissen ist Bodendorf (2006) der Ansicht, dass Wissen durch Verknüpfung von Informationen entsteht. Er ist der Meinung, dass für die Verknüpfung eine Kenntnis über die Informationen notwendig ist, die beschreibt, wie die einzelnen Informationen sich vernetzen lassen. Cleve und Lämmel (2020) ergänzen diese Aussage und definieren Wissen als Fähigkeit, Informationen zu nutzen. Aus diesen beiden Aussagen lässt sich ableiten, dass Wissen immer an Personen gebunden sein muss und abhängig von deren bereits vorhandenen Wissen ist, da *Informationen* auch unterschiedlich interpretiert werden können und somit auch unterschiedlich genutzt werden. Ein Beispiel wäre die Zahlenfolge *17658348976*. Während Person A anhand der Anordnung dieser Zahlen eine Mobilfunknummer ohne führende Null erkennen könnte, würde Person B diese Zahlenfolge nicht deuten können, weil Person B zum Beispiel keinen Bezug oder Hintergrundwissen zu Mobiltelefonen hat.

Einige Definitionen behandeln die Begriffe *Daten* und *Informationen* synonym. In dieser Arbeit ist die Wissensentdeckung in Datenbanken jedoch ein zentraler Punkt und daher sollten die Begriffe klar voneinander differenziert werden. Die Betrachtungsweise von Bodendorf ist für diese Arbeit sinnvoll, da das Ziel *Wissen*, welches bei der Wissensentdeckung in Datenbanken angestrebt wird, nicht synonym mit den Begriffen *Information* oder *Daten* verstanden werden soll. Bei der Wissensentdeckung ist folglich das Ziel, Informationen die ein Nutzen haben beziehungsweise die eine Entscheidungsmöglichkeit anbieten, zu erhalten und nicht um Daten die Informationen lediglich repräsentieren.

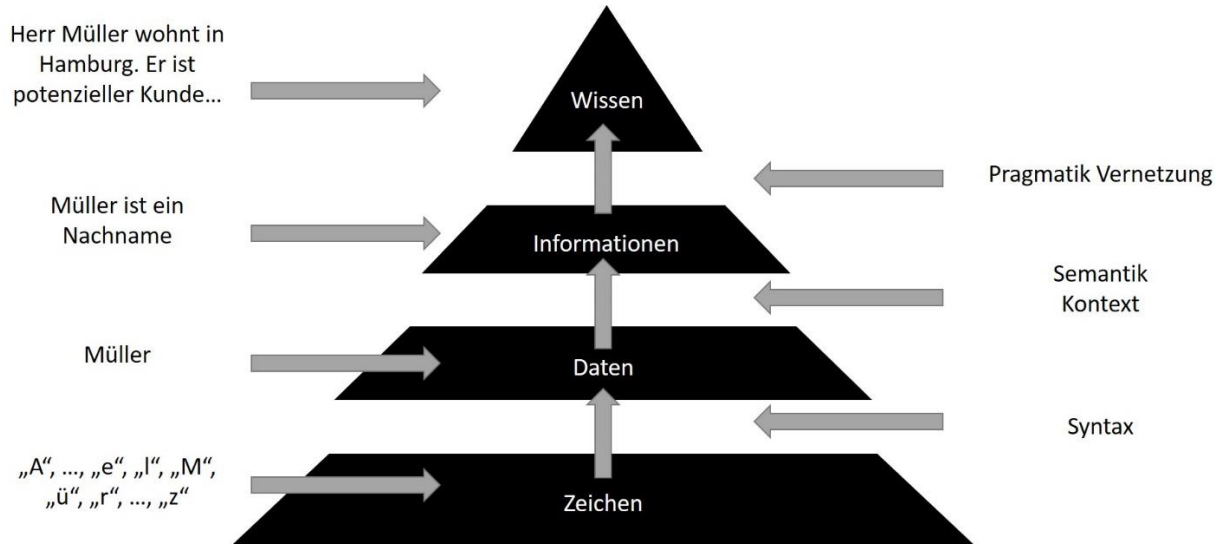

Abbildung 2-1: Begriffs hierarchie nach Bodendorf 2006

Neben den verschiedenen Begriffen, wie Daten Information und Wissen, existieren es unter dem Begriff Daten auch unterschiedliche Datentypen. Laut Cleve und Lämmel (2020) lassen sich Daten in nominale, ordinale und metrische Datentypen kategorisieren. Diese Kategorisierung begründen die Autoren mit den beiden Kriterien *Ordnen* und *Rechnen*.

Unter den nominalen Daten versteht man Daten, die keine feste Rangfolge haben. Nominale Daten können also nur verglichen werden, indem die Gleichheit der Daten geprüft wird (Cleve und Lämmel 2020). Die nominalen Daten können auch eine numerische Form aufweisen, diese numerische Form hat dann aber keine mathematische Bedeutung und es ist somit nicht möglich, mit diesen Daten Rechenoperationen durzuführen (Bramer 2020). Ein Beispiel für nominale Daten ist die Angabe eines Geschlechts – männlich oder weiblich.

Ordinale Daten sind ähnlich zu nominalen Daten mit dem Unterschied, dass sie eine feste Rangfolge haben (Bramer 2020). Ein Beispiel für ordinale Daten wären Schulnoten. Diese lassen sich von sehr gut = 1 bis ungenügend = 6 anordnen, jedoch macht die Rechnung 1 + 6 = 7 bezüglich Schulnoten keinen Sinn, da die Schulnote 7 nicht existiert.

Metrische Daten hingegen lassen sich wie reelle Zahlen Ordnen und mit Ihnen können auch Rechenoperationen durchgeführt werden (Cleve und Lämmel 2020). Metrische Daten wären zum Beispiel Temperaturangaben in Kelvin oder Streckenlängen in Metern. Die Autoren führen weiter aus und sagen, dass metrische Daten sich in diskrete und kontinuierliche Daten unterteilen lassen. Die Autoren erläutern den Unterschied so, dass diskrete Daten nur schrittweise größenveränderlich sich und kontinuierliche Daten jeden Zahlenwert innerhalb des Definitionsbereichs annehmen können.

Außerdem ist die Umwandlung eines Datentyps in einen anderen Datentyp möglich und sogar in einigen Fällen zwingend erforderlich, da zum Beispiel bestimmte Data-Mining-Verfahren wie das *k-Nearest-Neighbour-Verfahren* vorzugsweise mit kontinuierlichen metrischen Daten arbeiten (Cleve und Lämmel 2020). Die Umwandlung zwischen zwei Datentypen wird auch Transformation bezeichnet.

In der Literatur finden sich noch viele andere Datentypen beziehungsweise andere Kategorisierungsansätze. Für diese Arbeit soll aber die oben beschriebenen Definitionen genügen, da eine ausführliche Darstellung aller Datentypen nicht im Sinne der Beantwortung der Fragestellung dieser Arbeit ist.

2.1.2 Modell und Vorgehensmodell

Wie bereits in Kapitel 1 erläutert, ist ein Teilziel dieser Arbeit, der Entwurf eines Entscheidungsmodells. Darüber hinaus wird in dieser Arbeit ein *Vorgehensmodell* der *Wissensentdeckung in Datenbanken* vorgestellt. Daher ist es notwendig, zuerst den Begriff *Modell* beziehungsweise *Vorgehensmodell* einzuführen und zu definieren. Bandow (2009) nutzt zunächst die Definition nach den VDI Richlinien 3633 und sagt:

„Als Modell wird eine vereinfachte Nachbildung eines existierenden oder gedachten Systems mit seinen Prozessen in einem anderen begrifflichen oder gegenständlichen System verstanden, welches sich hinsichtlich der untersuchungsrelevanten Eigenschaften nur innerhalb eines vom Untersuchungsziel abhängigen Toleranzrahmens vom Vorbild unterscheidet“ (Bandow 2009, S. 21).

Die Definition zeigt, dass ein *Modell* sich anhand der Begriffe *Funktion* und *Ziel* beschreiben lässt. So ist laut Bodendorf eine Hauptfunktion von Modellen, die Abbildung realer Systeme und das Ziel eines Modells, die Komplexität von realen Systemen zu reduzieren und die Realität durch Darstellungsformen besser verständlich zu machen. Außerdem behauptet der Autor, dass Modelle sich in drei Hauptmerkmale unterteilen lassen. Diese Merkmale werden in Abbildung 2-2 dargestellt.

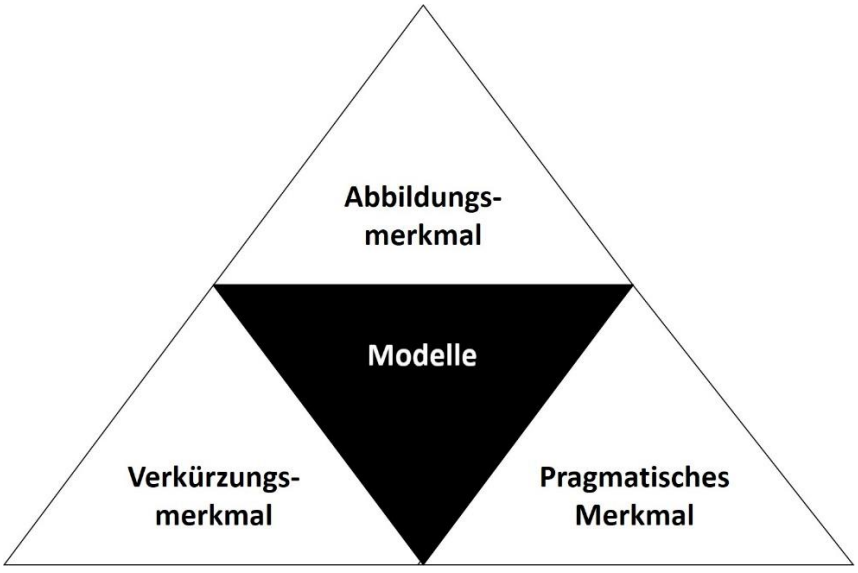

Abbildung 2-2 Allgemeine Merkmale von Modellen nach Bandow (2009)

Unter Abbildungsmerkmal wird wie bereits erwähnt, die Repräsentation der Realität verstanden. Mit dem Verkürzungsmerkmal ist gemeint, dass die Modelle nicht alle Informationen erfassen, sondern nur solche die für den Ersteller beziehungsweise Benutzer

relevant sind (Bandow 2009). Mit dem pragmatischen Merkmal ist gemeint, dass die Modelle nicht fest zugeordnet werden, sondern lediglich, für unterschiedliche Zwecke erstellt werden.

Der Begriff Vorgehensmodell wird in der Ingenieurswissenschaft häufig benutzt, um die Reihenfolge von bestimmten Prozessen abzubilden (Bandow 2009). Oft wird auch der Begriff Prozessmodell als synonym für Vorgehensmodell verwendet.

2.2 Vorgehensmodell von Fayyad

Nachdem die grundlegenden Begriffe im vorherigen Unterkapitel definiert wurden, wird in diesem Unterkapitel genauer auf die Wissensentdeckung in Datenbanken eingegangen. Es existieren viele verschiedene Theorien und Vorgehensmodelle zur Wissensentdeckung. Zu den beiden bekanntesten Vorgehensmodellen gehören zum einen das *Cross Industry Standard Process for Data Mining* (CRISP-DM) (Wirth und Hipp 2000) und zum anderen das Vorgehensmodell *Knowledge Discovery in Databases* (Fayyad et al. 1996). Da in dieser Arbeit der Schwerpunkt auf der Datenvorverarbeitung liegt, wird lediglich das Fayyad-Vorgehensmodell behandelt, da die Phasen der Datenvorverarbeitung im Vorgehensmodell von Fayyad et al. detaillierter beschrieben werden als im CRISP-DM-Modell. Das Vorgehensmodell von Fayyad et al., welches oft auch als KDD-Prozess oder Wissenentdeckungsprozess bezeichnet wird, lässt sich in 5 Phasen unterteilen (siehe Abbildung 2-3).

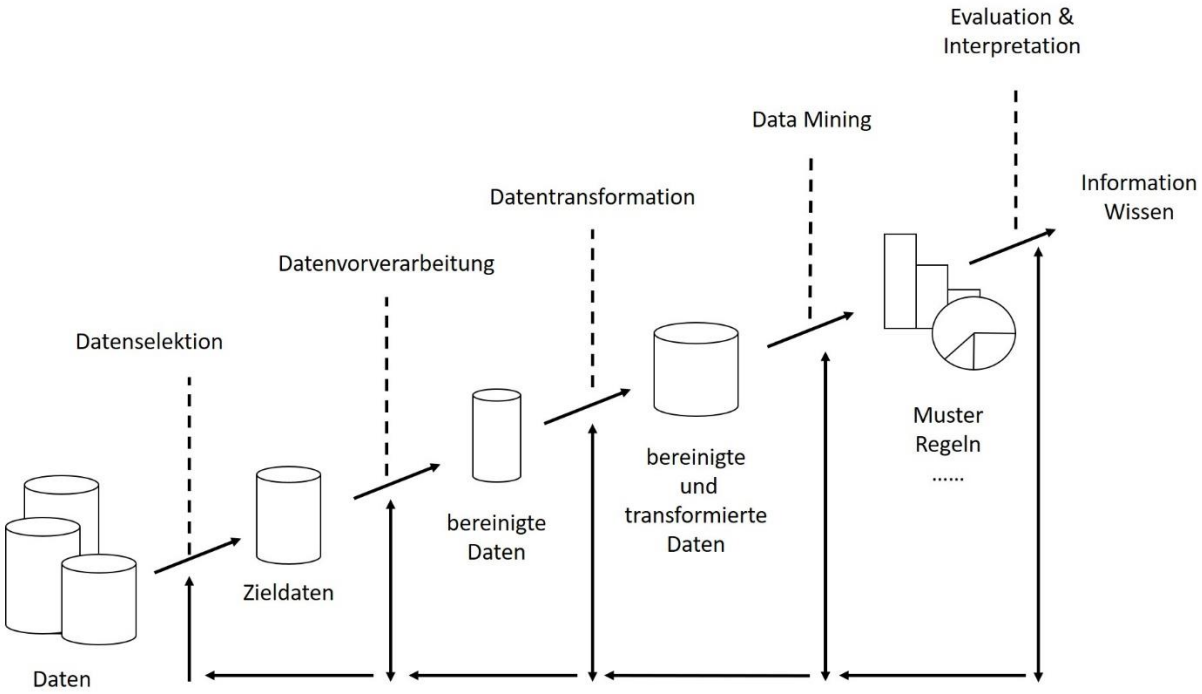

Abbildung 2-3 KDD-Prozess nach Fayyad et. al 1996, S. 10

In der ersten Phase werden die für die Analyse benötigten Daten, ausgehend aus der Zielvorstellung und dem vorhandenen Domänenwissen aus den Datenquellen ausgewählt (Sharafi 2013). Dabei können nach Cleve und Lämmel auch Daten von externen Quellen exportiert und genutzt werden, wenn diese dem Anwender als nützlich erscheinen. Diese erste Phase im KDD-Prozess wird auch *Datenselektion* genannt und hat das Ziel, die relevanten Daten für den Wissensentdeckungsprozess in einen Zieldatenbestand zu überführen, welcher für die darauffolgenden Phasen verwendet wird. Dies bedeutet, dass die Datenselektion die weiteren Phasen auch negativ beeinflussen kann, wenn relevante Daten nicht ausgewählt werden beziehungsweise Daten ausgewählt werden, die unwirksam für den KDD-Prozess sind. Nach Cleve und Lämmel können bei der Überführung der Daten in den Zieldatenbestand auch rechtliche sowie technische Probleme auftreten, da dem Anwender zum Beispiel die Zugriffsrechte oder die notwendigen Kapazitäten fehlen. Die Autoren führen weiter aus, dass bei diesen Problemen die Datenselektion neu überdacht werden muss oder die Möglichkeit besteht, dass nur bestimmte Teilmengen des Datenbestandes ausgewählt werden, um die Probleme zumindest teilweise zu umgehen.

Die in den verschiedenen Datenquellen vorkommenden Daten sind zudem oft fehlerbehaftet, verrauscht oder ungünstig skaliert (Runkler 2016). Daraus folgt, dass auch die Daten im Zieldatenbestand von diesen *Qualitätsmängeln* betroffen sind. Die Qualitätsmängel müssen vor der Datenanalyse beseitigt werden, um bestmögliche Ergebnisse bei dem Wissensentdeckungsprozess zu erzielen. Laut Cleve und Lämmel enthalten 5 % der Attribute eines realen Datenbestandes falsche Angaben. Die Entfernung von Qualitätsmängeln beziehungsweise das Korrigieren von falschen Angaben werden in der 2. Phase des KDD-Prozesses, welche auch *Datenvorverarbeitung* genannt wird, durchgeführt. Besonders die Entfernung von Rauschen sehen Fayyad et al. als elementar wichtig an, da ein hohes Maß an Rauschen das Erkennen von Mustern durch Data-Mining-Verfahren erheblich beeinträchtigt. Neben der Entfernung der Qualitätsmängel ist die Aufbereitung der Daten durch Standardisierung sowie die Zusammenfassung aller benötigten Daten, zum Beispiel zu einer einzigen Datenmatrix, ebenfalls Aufgaben der Datenvorverarbeitung (Runkler 2016). Dazu betont Runkler, dass die Datenvorverarbeitung einen erheblichen Anteil an Arbeits- und Zeitaufwand im KDD-Prozess einnimmt, während Cleve und Lämmel diese Aussage ergänzen und die Datenvorverarbeitung als *schwierigste Aufgabe* im KDD-Prozess bezeichnen.

Nach der Datenvorverarbeitung folgt die dritte Phase, die *Datentransformation*. Das Ziel der Datentransformation ist es, den bereinigten Datenbestand für die jeweiligen Data-Mining-Verfahren anzupassen, sodass die Data-Mining-Algorithmen benutzt werden können. Einige Data-Mining-Verfahren können zum Beispiel besser mit numerischen Daten umgehen,

während andere mit nominalen Daten besser arbeiten können (Cleve und Lämmel 2020). Außerdem wird durch die Datentransformation die effektive Anzahl der betrachteten Variablen sowie die Datendimension des Datenbestandes reduziert (Sharafi 2013). Einige Methoden der Transformation sind zum Beispiel die Diskretisierung, Normierung, Generalisierung oder Binärcodierung.

In der vierten Phase werden die eigentlichen Analysen zur Wissensentdeckung durchgeführt. Diese Phase heißt *Data Mining*. Mithilfe von Data-Mining-Verfahren können Muster in den zuvor vorverarbeiteten Daten gefunden werden und aus diesen Mustern wiederum dann Informationen beziehungsweise Wissen abgeleitet werden (Petersohn 2005). Dazu muss der Anwender für die jeweilige Aufgabenstellung ein geeignetes Data-Mining-Verfahren auswählen. Die Herausforderung jedoch besteht darin, dass alle Data-Mining-Verfahren einzigartig sind und dadurch eine Verallgemeinerung beziehungsweise eine aufgabenbezogene Zuordnung der Data-Mining-Verfahren schwierig ist (Aggarwal 2015). Aggarwal betont, dass es bestimmte Data-Mining-Formulierungen gibt, die immer wieder im Zusammenhang mit verschiedenen Anwendungen verwendet werden. Dabei bezeichnet er diese Formulierungen als *Superprobleme* beziehungsweise Bausteine des Data-Mining-Prozesses. Nach Sharafi werden bei der Datenanalyse mithilfe von Data-Mining-Verfahren häufig zwei unterschiedliche Arten von Zielen erfolgt. Zum einen die *Verifikation*, bei der bereits vorhandene Hypothesen untersucht werden und zum anderen die *Entdeckung* die Hauptaufgabe, das Finden neuer Muster ist.

In der letzten Phase werden die gefundenen Muster visualisiert und interpretiert. Dabei sollen nach Fayyad et al. die gefundenen Muster die Anforderungen *Gültigkeit*, *Neuartigkeit*, *Nützlichkeit* und Verständlichkeit erfüllen, um neues Wissen darzustellen. Das daraus abgeleitete Wissen kann dann entweder direkt angewendet oder erst einmal dokumentiert werden (Cleve und Lämmel 2020). Die Dokumentation wird dann entweder an die Domänen weiter gegeben oder der komplette Prozess wird mit anderen Parametereinstellungen durchgeführt und mit der Dokumentation verglichen, um so die Ergebnisse zu optimieren (Sharafi 2013).

2.3 Datenvorverarbeitung

Wie bereits im vorherigen Unterkapitel erläutert, ist die Datenvorverarbeitung besonders relevant für die Analyseergebnisse, da die Datenvorverarbeitung sicherstellt, dass die Data-Mining-Verfahren zum einen richtig funktionieren und zum anderen die Qualität der Ergebnisse aufgrund der höheren Datenqualität verbessert werden. In diesem Unterkapitel wird daher die Datenvorverarbeitung detaillierter dargelegt. Es existieren mehrere Verfahren der

Datenvorverarbeitung, die sich nach Petersohn in *Verfahrensunabhängigen* und *Verfahrensabhängigen Methoden* unterscheiden. Um die Relevanz der Datenvorverarbeitung hervorzuheben, werden in Abschnitt 2.3.1 zunächst die Fehlerarten in Datenbanken vorgestellt und erläutert, wobei der Fokus auf das Rauschen gelegt wird. In den darauffolgenden Abschnitten 2.3.2 und 2.3.3 werden dann die Verfahrensunabhängigen sowie Verfahrensabhängigen Methoden vorgestellt.

2.3.1 Fehlerarten

In realen Datensätzen können aus verschiedenen Gründen fehlerhafte Werte erfasst werden, die durch Messfehler, subjektive Einschätzungen und Fehlfunktionen oder Missbrauch von automatischen Aufzeichnungsgeräten entstehen können (Bramer 2020). Nach Runkler (2016) lassen sich Fehler in zufällige und systematische Fehler unterteilen. Als zufälliger Fehler nennt Runkler (2016) Mess- und Übertragungsfehler und führt weiter aus, dass diese Fehler als additives Rauschen modelliert werden können. Die Ursache für systematische Fehler können fehlerhafte Formeln in der Berechnung von Merkmalen, falsche Kalibrierungen von Messgeräten oder eine falsche Skalierung sein (Runkler 2016). In der Literatur werden oft die Fehlerarten, fehlende *Werte, inkonsistente Werte, Ausreißer* und *Rauschen* als Datenqualitätsmängel aufgezeigt. In dieser Arbeit sollen nur verrauschte Daten betrachtet werden, wobei der Übergang von verrauschten Daten zu Ausreißern in der Literatur nicht klar getrennt ist. Oft werden diese beiden Begriffe synonym verwendet. Tatsächlich können viele Data-Mining-Verfahren zur Behandlung von Rauschen auch zur Behandlung von Ausreißern verwendet werden, wie zum Beispiel die Cluster-Analyse-Verfahren. Im Rahmen dieser Arbeit ist es jedoch sinnvoll, diese beiden Begriffe zu trennen, da der Schwerpunkt dieser Arbeit auf den Behandlungsmöglichkeiten von Rauschen liegt und einige Erkennungsmethoden für Ausreißer nicht für verrauschte Daten geeignet sind. Daher ist zunächst wichtig, den Unterschied zwischen Rauschen und Ausreißer zu erläutern.

Als *Rauschen* oder auf Englisch *Noise* wird oft ein Fehler bezeichnet, der zufällig aufgezeichnet wird und den Wert einer Variable verändert (Aggarwal 2015). Darüber hinaus ergänzt Bramer diese Aussage damit, dass Rauschen zwar ein falsch aufgezeichneter Wert ist, jedoch für den vorliegenden Datensatz gültig sein kann. Diese Eigenschaft der Gültigkeit kann zu Problemen bei der Identifizierung von verrauschten Daten führen, beispielsweise wenn der numerische Wert *1,45* als *14,5* eingegeben wird. Ungültige Werte wie zum Beispiel der Wert *1,45X* sind leichter zu identifizieren und korrigieren und sind daher eher unproblematisch (Bramer 2020). Dabei sind verrauschte Daten als Daten zu verstehen, die

Rauschen besitzen. Nach García et al. (2015) können im Datensatz zwei Arten von Rauschen auftreten:

1. *Klassenrauschen (Class Noise):* Klassenrauschen tritt auf, wenn Daten falsch beschriftet werden. Dabei kann es viele Ursachen für die Entstehung von Klassenrauschen geben wie zum Beispiel Subjektivität während des Beschriftungsprozesses, Dateneingabefehler oder unzureichende Informationen, die zur Beschriftung der einzelnen Beispiele verwendet werden. Das Klassenrauschen lässt sich selbst noch einmal in zwei verschiedenen Arten von Klassenrauschen aufteilen:
    - *Widersprüchliche Beispiele:* Dabei handelt es sich um Beispiele, die Doppelt in einem Datensatz vorkommen, aber unterschiedliche Klassenbezeichnungen besitzen.
    - *Fehlklassifizierungen:* Diese Beispiele sind Daten, die mit einer Klassifizierung beschriftet werden, die sich von ihrer wahren Bezeichnung unterscheiden.
2. *Attributrauschen (Attribute Noise):* Attributrauschen bezieht sich dabei auf die Verfälschung der Attributwerte. Fehlerhafte Attributwerte, fehlende oder unbekannt Attributwerte, unvollständige Attribute oder sogenannte *egal*-Werte sind Beispiele für Attributrauschen.

Unter Ausreißer wiederum werden Daten bezeichnet, die stark von den anderen Daten abweichen. Nach Hawkins (1980) ist die Abweichung so groß, dass der *Verdacht* entsteht, dass die Abweichung durch einen anderen Mechanismus entstanden ist. Jedoch ist zu erwähnen, dass die Abweichung ein korrekt erfasster Wert sein kann und somit eine wichtige Information darstellt (Cleve und Lämmel 2020). Ausreißer, die falsche oder fehlerhafte Werte besitzen, werden oft als Rauschen markiert und sind somit aus dem Datenbestand gelöscht (Cleve und Lämmel 2020). Solche Fehler können zum Beispiel durch fehlerhafte Eintragung in der Datenbank entstehen, dabei kann die Entstehung dieser Fehler zufällig oder systematisch sein (Runkler 2016).

In dieser Arbeit liegt der Schwerpunkt daher auf den Daten, die einen Fehler aufweisen und somit als Rauschen bezeichnet werden.

2.3.2 Verfahrensunabhängige Methoden

Bei den verfahrensunabhängigen Methoden der Datenvorverarbeitung sind die im Anschluss angewendeten Data-Mining-Verfahren nicht von Bedeutung (Petersohn 2005). Daher können diese Methoden bereits vor der Auswahl der Data-Mining-Verfahren angewendet werden (Cleve und Lämmel 2020). Unter den verfahrensunabhängigen Methoden sind beispielsweise die *Datenintegration, Datenbereinigung* und *Datenreduktion* zu verstehen.

Die Datenintergration wird dann verwendet, wenn Daten aus verschiedenen Quellen zusammengeführt werden müssen und hat das Ziel, eine idealerweise konsistente Tabelle mit schlüssigen Datensätzen zu erstellen (Cleve und Lämmel 2020). Wird der Integrationsprozess nicht ordnungsgemäß ausgeführt, kann es zu redundanten oder inkonsistenten Daten führen (García et al. 2015).

Nach der Datenintegration folgt meistens die Datenbereinigung, da die Daten nach dem Integrationsprozess noch viele verschiedene Fehler wie Ausreißer oder Rauschen besitzen. Die Bereinigung des Datensatzes von diesen Fehlern ist von großer Bedeutung, da fehlerhafte Daten die Ergebnisse der Data-Mining-Verfahren verfälschen können, die dann zu unzuverlässige oder sogar falsche Resultate führen (Cleve und Lämmel 2020). Nach Cleve und Lämmel lassen sich verrauschte Daten durch folgende Verfahren behandeln:

- *Binning*: Bei diesem Verfahren werden die Daten in kleine Gruppen, sogenannte *Bins* aufgeteilt und durch Mittelwerte oder Grenzwerte ersetzt. Dies hat einen glättenden Effekt auf die Daten. Binning-Verfahren werden auch als *Datenglättungsverfahren* oder auf Englisch *Data Smoothing Methods* bezeichnet.
- *Regression*: Bei dieser Methode werden die Daten durch eine mathematische Funktion beschrieben. Die verrauschten Datenwerte werden dann durch die Funktionswerte ersetzt.
- *Cluster-Analyse*: Die Cluster-Analyse ist besonders effektiv bei Ausreißern. Ähnliche Werte werden zu einem Cluster vereint. So werden Ausreißer von den anderen Daten getrennt.

Neben den in Abschnitt 2.3.1 genannten Fehlerarten können auch zu große Datensätze die Leistung der Data-Mining-Verfahren bei der Datenanalyse erheblich beeinträchtigen (Petersohn 2005). Durch die Datenreduktion kann die Dimension eines Datensatzes reduziert werden. Dabei erfolgt die Reduzierung nach Petersohn zum einen durch die Verringerung der Anzahl der Attribute (Aggregation und Dimensionsreduktion) und zum anderen durch die Verringerung der Anzahl der Datensätze (Stichprobenziehung).

2.3.3 Verfahrensabhängige Methoden

Die verfahrensabhängigen Methoden sind keine typischen Datenvorverarbeitungsschritte wie die zuvor behandelten unabhängigen Methoden. Verfahrensabhängige Methoden der Datenvorverarbeitung können erst dann angewendet werden, wenn ein Data-Mining-Verfahren zur Datenanalyse festgelegt ist. Die vorliegenden Daten können trotz intensiver Datenvorverarbeitung oft nicht mit Data-Mining-Verfahren analysiert werden, da diese einen inkompatiblen Datentyp aufweisen (Cleve und Lämmel 2020). Die Daten müssen daher transformiert werden und einen homogenen Datentyp aufweisen. Die *Datentransformation* hat daher die Aufgabe, die Daten so umzuwandeln, dass die jeweilige Data-Mining-Verfahren damit arbeiten können (Cleve und Lämmel 2020). Besondere Relevanz besitzen Datentransformationsverfahren wie *Diskretisierung*, *Normalisierung* und *Skalierung. Da* die Datentransformation jedoch als ein Zwischenschritt zwischen Datenvorverarbeitung und Data Mining gesehen werden kann, wird diese nicht weiter erläutert.

3 Methodik der systematischen Literaturrecherche zur Behandlung von Rauschen

Neues Wissen kann nur durch Kombination und Interpretation von vorhandenem Wissen entstehen und daher spielt die Literaturrecherche eine große Rolle in der Wissenschaft (vom Brocke et al. 2009). Die in dieser Arbeit als Ziel deklarierte Untersuchung zur Behandlung von Rauschen in der Datenvorverarbeitung mithilfe von Data-Mining-Verfahren soll auf Grundlage bereits in der Literatur und Forschung bestehender Data-Mining-Verfahren erfolgen. Um die einzelnen Verfahren zur Behandlung von Rauschen zu identifizieren, bedarf es einer systematischen Literaturrecherche. Neben der Darstellung der Ergebnisse ist es ebenfalls sinnvoll, die Vorgehensweise der systematischen Literaturrecherche zu erläutern, um weitere Recherchen bei ähnlichen Forschungsfragen zu unterstützen und redundante Untersuchungen zu vermeiden. In diesem Kapitel wird daher die systematische Literaturrecherche dieser Arbeit vorgestellt. Dabei orientiert sich die Vorgehensweise an das etablierte System von vom Brocke et al. (2009). Die systematische Literaturrecherche nach vom Brocke et al. (2009) lässt sich dabei in fünf Phasen unterteilen (siehe Abbildung 3-1).

Die erste Phase legt den Umfang der Literaturrecherche fest, während die zweite Phase eine Konzeptdarstellung des Themenbereiches aufzeigt. Dabei wird durch die Berücksichtigung des bereits vorhandenen Wissens die Datenquellen und Suchbegriffe sowie Ein- und Ausschlusskriterien bestimmt. Die eigentliche Suche wird dann in der dritten Phase durchgeführt. In Phase vier wird dann die relevante Literatur selektiert und analysiert, während in Phase fünf die Ergebnisse anhand einer Übersicht festgehalten werden. Die einzelnen Phasen werden für diese Arbeit in den folgenden Unterkapiteln dargestellt.

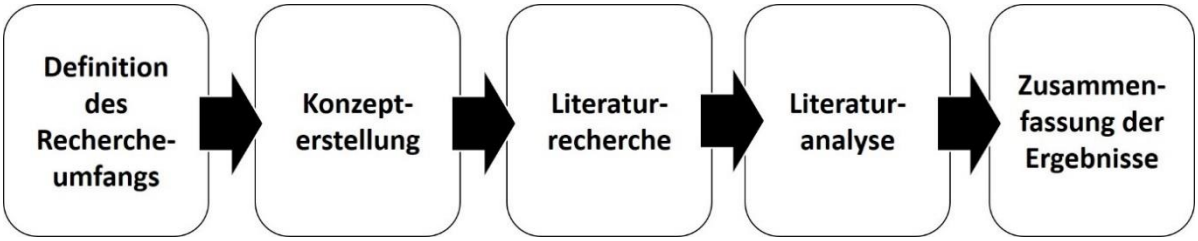

Abbildung 3-1 Vorgehensweise der systematischen Literaturrecherche in Anlehnung an vom Brocke (2009)

3.1 Taxonomie der Arbeit

Um den Umfang der Literaturrecherche zu bestimmen, schlägt vom Brocke et al. (2009) eine Taxonomie nach Cooper (1988) vor. Dazu ist die Arbeit anhand der Merkmale *Fokus, Ziel, Organisation, Zielgruppe, Darstellung* und *Grad der Abdeckung* zu klassifizieren (vom Brocke et al. 2009). Diese Klassifizierung ist in Tabelle 3-1 dargestellt.

Tabelle 3-1 Definition des Untersuchungsumfangs nach Cooper (1988)

Merkmal	Kategorien
Fokus	Forschungsergebnisse, Forschungsmethoden, Theorien, Praktiken oder Anwendungen
Ziel	Integration, Kritik, Identifikation Zentraler Aspekte
Organisation	historisch, konzeptionell, methodisch
Zielgruppe	Spezialisierte Wissenschaftler, allgemeine Wissenschaftler, Praktiker, öffentliches Publikum
Darstellung	Neutrale Darstellung, Vertretung eines Standpunktes
Grad der Abdeckung	Vollständig, vollständig selektiv, repräsentativ, zentrale Abdeckung

Werden die einzelnen Merkmale auf diese Arbeit angewendet, so kann der Fokus dieser Arbeit durch das Thema der Arbeit abgeleitet werden und liegt in der Behandlung von Rauschen in der Datenvorverarbeitung mithilfe von Data-Mining verfahren. Das Ziel dieser systematischen Literaturrecherche ist es, einen Überblick über mögliche Data-Mining-Verfahren zur Behandlung von Rauschen in der Datenvorverarbeitung zu schaffen. Dabei ist die Literaturrecherche methodisch gegliedert, das bedeutet, dass Abstrakte mit gleichen Methoden gemeinsam betrachtet werden. Die Ergebnisse werden neutral und sachlich dargestellt und dabei wird vorzugsweise ein Fachpublikum angesprochen. Damit eine aussagekräftige Anzahl an relevanter Literatur erschlossen werden kann, wird jede Form von Literatur in die Recherche mit einbezogen.

3.2 Konzepterstellung des Themenbereiches

Um die Literaturrecherche durchführen zu können, ist es zunächst wichtig, einen Überblick über bereits vorhandenes Wissen zum Thema zu bekommen und daraus dann eine Suchstrategie in Form von Suchbegriffen sowie Ein- und Ausschlusskriterien abzuleiten (vom Brocke et al. 2009). Hierzu wurde die Suchmaschine der Universitätsbibliothek Dortmund benutzt und einzelne Begriffe, die sich aus der Forschungsfrage ableiten lassen wie zum Beispiel *Wissensentdeckung, Datenvorverarbeitung* oder *Data Mining eingegeben.* Die Ergebnisse dieser ersten Recherche wurden teilweise in Kapitel 2 aufgeführt. Anhand der in Kapitel 2 aufgezeigten Ergebnisse sowie der Forschungsfrage konnten Schlüsselbegriffe wie *Rauschen (Noise), Klassenrauschen (Class Noise), Attributrauschen (Attribute Noise),*

Wissensentdeckung in Datenbanken (Knowledge Discovery in Databases), Data Mining, Regression, Cluster-Analyse und Binning abgeleitet werden. Eine Gesamtübersicht der Suchbegriffe ist im Anhang in der Tabelle A-1 zu sehen. Anhand der Schlüsselbegriffe ist zu erkennen, dass sowohl deutsch- als auch englischsprachige Literatur betrachtet wird. Darüber hinaus werden Synonyme der einzelnen Begriffe ebenfalls verwendet. In dieser Arbeit werden dabei Datenbanken sowie Publikationen aus dem technischen Bereich durchsucht. Dazu gehören zum Beispiel *Ieee Xplore Digital Libary, Sciencedirect*, *Springer Link, Engineering Source* und *Hanser-EBA.* Außerdem werden lediglich Veröffentlichungen zwischen 1996 und 2022 analysiert, damit alle Literaturwerke seit der Veröffentlichung des Vorgehensmodells von Fayyad et al. (1996) berücksichtigt werden können.

3.3 Literaturrecherche

Mithilfe des Konzepts aus dem vorherigen Abschnitt kann die eigentliche Literaturrecherche durchgeführt werden. Dazu werden die Suchbegriffe einzeln in die Suchmaschine eingegeben und die Anzahl an Resultate notiert. Durch Kombination der Suchbegriffe mithilfe von sogenannten Suchoperatoren wie *AND* (∧) oder *OR* (∨) kann die Anzahl der Resultate reduziert werden, sodass am Ende eine überschaubare Anzahl an relevanter Literatur vorhanden ist. Der Suchprozess ist daher als iterativer Prozess zu verstehen. Es wurden mehrere Suchvorgänge durchgeführt. Dabei wurden immer andere finale Suchbegriffe ausgewählt oder das Suchfeld wurde geändert. Die folgende Tabelle zeigt exemplarisch das Vorgehen der Datenbankabfrage.

Tabelle 3-1 Beispiel eines Suchvorgangs

Operator	Suchfeld	Suchbegriff	Anzahl der Resultate
	Alle Felder	Data Mining	5112508
AND	Alle Felder	Data Mining ∧ Data Preprocessing	179859
AND	Alle Felder	Data Mining ∧ Data Preprocessing ∧ Noise	87895
AND	Alle Felder	Data Mining ∧ Data Preprocessing ∧ Noise ∧ Identification	62128
AND	Alle Felder	Data Mining ∧ Data Preprocessing ∧ Noise ∧ Identification ∧ Binnig	4555

Werden die Resultate in der Tabelle 3-1 betrachtet, so ist zu sagen, dass die Anzahl der Resultate in der letzten Zeile mit 4555 noch zu hoch ist. Um dieses Problem zu lösen, ohne weitere Suchbegriffe hinzuzufügen, wurden zwei weitere Möglichkeiten angewandt. Zum einen wurde das Suchfeld von *Alle Felder* auf *Titel* verändert. Dadurch beschränkt sich die Suche nur auf Literaturwerke die alle genannten Suchbegriffe im Titel Beinhalten würden. Zum anderen wurden Suchphrasen verwenden. Bei Suchphrasen werden mehrere Suchbegriffe mit Anführungszeichen begrenzt, dadurch wird die Reihenfolge der einzelnen Suchbegriffe berücksichtigt. Um diese Vorgehensweise besser zu versehen, sind im Anhang einige weiter Suchvorgänge dargestellt.

3.4 Literaturanalyse

Um relevante Literaturwerke zu identifizieren wurde eine Literaturanalyse durchgeführt. Dazu wurden bei jedem Iterationsschritt die Titel der ersten 100 Ergebnisse betrachtet. Titel, die bereits vielversprechend wirkten, wurden dann weiter analysiert. Dazu wurde als Nächstes der *Abstract* gelesen um einen Überblick zu bekommen, welche Themen in den jeweiligen Literaturwerken behandelt werden. Wurde bereits im Abstract das Rauschen in der Datenvorverarbeitung angesprochen, so wurde diese Literatur sofort als Relevant angesehen. Die Literaturwerk von Cleve und Lämmel (2020), Aggarwal (2015), Sharafi (2013) und Runkler (2016) wurden als erstes gelesen, da diese bereits viele Suchbegriffe abgedeckt haben und bei mehreren Verschiedenen Suchvorgängen immer wieder ermittelt werden konnten. Darüber hinaus wurde anhand dieser Literaturwerke zusätzlich eine Rückwärtssuche durchgeführt. Das bedeutet, dass die Quellen in den Literaturwerken ebenfalls berücksichtigt wurden.

Durch die Suchvorgänge in der Suchmaschine der Universitätsbibliothek Dortmund und mithilfe der Rückwärtssuche konnten 40 relevante Literaturwerke ermittelt werden. Diese sind Übersichtlich in der Tabelle A-6 im Anhang dargestellt. Die Tabelle A-6 zeigt zudem bereits, welch Verfahren zur Behandlung von Rauschen in der Datenvorverarbeitung in den jeweiligen Literaturwerken thematisiert werden. Die Ergebnisse werden im folgenden Kapitel erläutert.

4 Durchführung der Literaturrecherche und Kategorisierung der Data-Mining-Verfahren

Durch den Prozess der systematischen Literaturrecherche konnten verschiedene Literaturwerke identifiziert und analysiert werden, die zur Beantwortung der Forschungsfrage relevant sind. Die durch die Literaturrecherche ermittelten Verfahren weisen unterschiedliche Vorgehensweisen auf, um verrauschte Daten zu behandeln. Aus der Gesamtübersicht in Tabelle A-6 lässt sich ableiten, dass die Verfahren sich in die Kategorien Binning- und Smoothing- sowie Regressions-, Klassifikation-, Cluster-Analyse- und Filter-Verfahren einteilen lassen. Die Regressionsanalyse, Klassifikation und Cluster-Analyse sind üblicherweise Data-Mining-Verfahren die in der Data-Mining-Phase des KDD-Prozesses genutzt werden. Also überwiegend zu Erkennung von Mustern innerhalb von Datensätzen. Durch die Literaturrecherche konnte gezeigt werden, dass diese Verfahren auch besonders gut geeignet sind, um in der Datenvorverarbeitungsphase verrauschte Daten zu behandeln.

In diesem Kapitel werden zu den jeweiligen Kategorien die ermittelten Verfahren vorgestellt und erläutert. Darüber hinaus werden die Aussagen der Autoren verglichen und die Unterschiede und Gemeinsamkeiten erarbeitet. Außerdem wird zu Beginn jeder Kategorie ein Auszug aus Tabelle A-6 dargestellt, in der alle Literaturwerke gezeigt werden, die die jeweilige Kategorie thematisieren. Dadurch wird eine übersichtlichere Darstellung innerhalb der Kategorien gewährleistet.

4.1 Binning und Smoothing-Verfahren

Tabelle 4-1 Literaturübersicht zu den Binning- und Smoothing-Verfahren

Autor (Jahr)	Titel	Vorgehen zur Behandlung von Rauschen	Data-Mining-Verfahren	Art der Publikation
Aggarwal (2015)	Data mining – The Textbook	Glätten; Identifizieren und Löschen; Filtern	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse; Filter-Verfahren	Monographie
Alpaydın (2022)	Maschinelles Lernen	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse	Monographie
Cios et al. (2007)	Data Mining – A Knowledge Discovery Approach	Glätten; Identifizieren und Löschen	Binning/Smoothing; Clusteranalyse	Monographie
Cleve und Lämmel (2020)	Data Mining	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse	Monographie

Islam et al. (2013)	Speckle Noise Reduction From Ultrasound	Glätten	Binning/Smoothing	Zeitschriftenaufsatz
Kessler (2006)	Multivariate Datenanalyse für die Pharma	Glätten	Binning/Smoothing	Monographie
Liu et al. (2002)	Discretization: An Enabling Technique	Glätten	Binning/Smoothing	Zeitschriftenaufsatz
Runkler (2016)	Data Analytics	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse Klassifikation	Monographie

Um Rauschen zu entfernen, werden oft Verfahren verwendet, die verrauschte Daten glätten. Zu diesen Verfahren gehören die sogenannte Binning- und Smoothing-Verfahren. Im ersten Abschnitt dieses Unterkapitels wird zunächst das Binning erklärt. Im Anschluss daran werden die Smoothing-Verfahren, *Moving-Average Smoothing* und *Exponential Smoothing* vorgestellt und erläutert.

Bei den Binning-Verfahren wird bei einem nach der Größe sortierten Datensatz ein Glättungseffekt erzielt, indem die einzelnen Werte mit ihren Nachbarwerten kommunizieren (HAN et al. 2012). Dazu werden die einzelnen Werte zunächst in einzelne Gruppen sogenannte *Bins* aufgeteilt (Cleve und Lämmel 2020). Die Aufteilung der Bins kann dabei durch zwei unterschiedliche Methoden erfolgen. Zum einen können die Werte in *k* gleich große Bins mit gleicher Anzahl von Elementen aufgeteilt werden (*Equal-Frequency*) und zum anderen in *k* Bins mit gleich großen Intervallen (*Equal-Width*) (Luengo et al. 2020). Nach der Aufteilung in Bins, werden die einzelnen Werte innerhalb der Bins durch andere Werte ersetzt, um den Glättungseffekt zu erzielen. Dabei gibt es erneut zwei verschiedenen Möglichkeiten diese Werte zu ersetzen. Bei der ersten Möglichkeit werden die Werte durch den Mittelwert oder den Median ersetzt, während bei der zweiten Möglichkeit die Werte durch die Randwerte, also den kleinsten und den größten Wert innerhalb eines Bins ersetzt werden. Um das Glätten durch Binning zu veranschaulichen, wird das folgende Beispiel aufgeführt, bei dem der Datensatz *{13, 9, 26, 39, 33, 26, 20, 29, 30}* Daten für Preise (in Euro) verschiedener Produkte darstellt. Als Erstes müssen die einzelnen Werte nach Ihrer Größe sortiert werden. Der sortierte Datensatz ergibt sich dann zu *{9, 13, 20, 26, 26, 29, 30, 33, 39}.* Anschließend erfolgt die Einteilung in Bins und die Glättung durch Ersetzen der Werte innerhalb der jeweiligen Bins. Die Einteilung in Bins und das Ersetzen der Werte ist im Folgenden exemplarisch dargestellt:

- *Einteilung in gleich große Bins*
    - *Bin 1: {9, 13, 20}*
    - *Bin 2: {26, 26, 29}*
    - *Bin 3: {30, 33, 39}*
- *Ersetzen durch Mittelwert*
    - *Bin 1: {14, 14, 14}*
    - *Bin 2: {27, 27, 27}*
    - *Bin 3: {34, 34, 34}*
- *Ersetzen durch Grenzwert*

- *Bin 1: {9, 9, 20}*
- *Bin 2: {26, 26, 29}*
- *Bin 3: {30, 30, 39}*

In der Wissenschaft treten Daten oft in Form von Zeitreihen auf, wobei eine Zeitreihe als eine (zeitlich) geordnete Folge von Messungen beziehungsweise Beobachtungen einer Größe zu verstehen ist (Schlittgen und Streitberg 2001). Eine Zeitreihe könnte zum Beispiel eine Tabelle sein, bei der die Umsätze eines Unternehmens für jeden Monat aufgelistet sind (siehe Tabelle 4-1). Ein häufig genutztes Verfahren zur Glättung von Zeitreihen ist das Moving-Average Smoothing (Bhattacharyya 2022). In der deutschsprachigen Literatur wird das Moving-Average auch als *gleitender Durchschnitt* bezeichnet. Um die Zeitreihe zu glätten und dadurch die verrauschten Werte zu entfernen, werden bei dem Moving-Average Smoothing die Daten der Zeitreihe in überschneidende Bins aufgeteilt (Aggarwal 2015). Durch die Überschneidung der Bins werden die Verluste, die bei den im vorherigen Abschnitt behandelten Binning-Verfahren entstehen können, reduziert (Aggarwal 2015). Nach Aggarwal (2015) besteht der Hauptunterschied darin, dass Bins nicht nur bei Zeitstempeln, die an den Grenzen der Bins befindlich sind, erstellt werden, sondern bei allen Zeitstempeln in einer Reihe ein Bin erstellt wird. Ähnlich wie bei den Binning-Verfahren, wird der glättende Effekt durch das Einsetzen der arithmetischen Mittelwerte in den jeweiligen Bins erzielt. Als Beispiel kann die Tabelle 4-2 betrachtet werden. So wird bei einer Periodenlänge von drei Monaten zunächst der gleitende Durchschnitt für die Monate Januar – März ermittelt. Es wird also berechnet

$$\frac{1}{3} \cdot (13.000\ € + 9.000\ € + 8.000\ €) = 10.000\ €.$$

Der nächste gleitende Durchschnitt wird dann berechnet, indem der älteste Wert (hier: Januar = 13.000 €) weggelassen wird und dafür ein weiterer Wert (hier: April = 16.000 €) hinzugenommen wird. Daher wird im nächsten Schritt der arithmetische Mittelwert von den Monaten Februar – März berechnet. Das Vorgehen wird solange wiederholt, bis alle Werte der Zeitreihe berücksichtigt werden.

Tabelle 4-2 Beispiel einer Zeitreihe

Januar	Februar	März	April	Mai
13.000 €	9.000 €	8.000 €	16.000 €	24.000€

Der gleitende Durchschnitt wird in einer Echtzeitanwendung erst nach dem letzten Zeitstempel des Fensters verfügbar (Aggarwal 2015). Daher führen nach Aggarwal gleitende Durchschnitte

zu Verzögerungen in der Analyse und außerdem gehen aufgrund von Randeffekten einige Punkte am Anfang der Reihe verloren. Darüber hinaus betont der Autor, dass größere Bin-Größen zu besseren Glättungseffekten führen.

Das Exponential Smoothing oder auf Deutsch die exponentielle Glättung ist eine Form des Moving-Average-Verfahren, bei der die jüngsten und früheren Beobachtungen unterschiedlich gewichtet werden (Bhattacharyya 2022). Dazu wird der geglättet Wert y'_i als Linearkombination aus dem aktuellen Wert y_i und dem zuvor geglätteten Wert y'_{i-1} dargestellt (Aggarwal 2015). Mithilfe von $\alpha \in (0,\ 1)$ folgt dann nach Aggarwal:

$$y'_i = \alpha \cdot y_i + (1-\alpha) \cdot y'_{i-1} \tag{4.1}$$

Wird $\alpha = 1$ gewählt, so liegt kein Glättungseffekt vor und die Zeitreihe bleibt unverändert, während für $\alpha = 0$ der größte Glättungseffekt entsteht (Aggarwal 2015). Der Autor erklärt, dass die Formel (4.1) sich in eine exponentiell abklingende Summe überführen lässt und daher dieser Ansatz als exponentielles Glätten bezeichnet wird. Die Formel

$$y'_i = (1-\alpha)^i \cdot y'_0 + \alpha \cdot \sum_{j=1}^{i} y_j \cdot (1-\alpha)^{i-j}, \tag{4.2}$$

welche durch eine rekursive Substitution der Gleichung (4.1) entsteht, zeigt die exponentiell abnehmenden Gewichte und ergründet daher die Namensgebung dieses Verfahrens (Aggarwal 2015; Krämer 2015).

Es ist zu erkennen, dass alle drei Verfahren dieses Unterkapitels aufeinander aufbauen. Das Moving-Average-Smoothing berücksichtigt im Gegensatz zum Binning die Übergänge zwischen den einzelnen Bins und das Exponential-Smoothing wiederum berücksichtigt im Gegensatz zum Moving-Average zusätzlich die Aktualität der Beobachtungen.

4.2 Regression

Tabelle 4-3 Literaturübersicht zu den Regressionsverfahren

Autor (Jahr)	Titel	Vorgehen zur Behandlung von Rauschen	Data-Mining-Verfahren	Art der Publikation
Aggarwal (2015)	Data mining – The Textbook	Glätten; Identifizieren und Löschen; Filtern	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse; Filter-Verfahren	Monographie
Alpaydın (2022)	Maschinelles Lernen	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse	Monographie
Cleve und Lämmel (2020)	Data Mining	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse	Monographie
Fayyad et al. (1996)	From Data Mining to Knowledge	Glätten; Identifizieren und Löschen	Regressionsanalyse; Clusteranalyse	Zeitschriftenaufsatz
Ghavami (2020)	Big data analytics methods	Glätten	Regressionsanalyse	Monographie

Otte et al. (2020)	Von Data Mining bis Big Data	Glätten	Regressionsanalyse	Monographie
Pyle (1999)	Data preparation for data mining	Glätten	Regressionsanalyse	Monographie
Raja (2022)	Data Mining and Machine Learning Applications	Glätten; Identifizieren und Löschen	Regressionsanalyse; Klassifikation; Clusteranalyse	Monographie
Runkler (2016)	Data Analytics	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse Klassifikation	Monographie

Wie bereits in Abschnitt 2.3.2 erwähnt, können verrauschte Daten auch durch Regression identifiziert werden. Regressionen werden hauptsächlich für Prognosen von numerischen Werten verwendet, können jedoch auch zum Erkennen von zusammenhängen in der Datenvorverarbeitung genutzt werden (Cleve und Lämmel 2020). Die Fehlerreduzierung wird durch Schätzung der Beziehung, die zwischen Variablen bestehen kann, erzielt (Raja 2022). In diesem Unterkapitel werden einige Regressions-Verfahren vorgestellt.

Das Ziel der linearen Regression ist es eine lineare Funktion zu finden die am besten eine Reihe von Daten approximiert (Ghavami 2020). Um dieses Ziel zu erreichen wird bei der linearen Regression davon ausgegangen, dass ein linearer Zusammenhang zwischen den Daten und deren Funktionswerte besteht und dieser daher durch eine Gerade beschrieben werden kann (Cleve und Lämmel 2020). Dabei werden die Funktionswerte oft als Outputvariable und die Datenwerte als Inputvariable bezeichnet (Aggarwal 2015). Das bedeutet, dass bei einer linearen Regression nur Daten betrachtet werden können die durch zwei Variablen beschrieben werden können (Pyle 1999). Die bestimmte Regressionsfunktion $\hat{y} = f(x)$ hat nach Cleve und Lämmel (2020) das Ziel den Fehler zwischen realen und berechneten Werten zu minimieren. Dazu wird der quadratische Fehler untersucht, damit positive und negative Abweichungen sich nicht gegenseitig kompensieren können (Cleve und Lämmel 2020). In Abbildung 4-1 ist eine lineare Regression beispielhaft dargestellt. Wie in Abbildung 4-1 zu erkennen ist, weichen die Daten für A und B deutlich von den anderen Daten ab. Diese Daten würde dann als verrauschte Daten interpretiert werden und müssen durch Werte, die mithilfe der Regressionsfunktion berechnet werden können, ersetzt werden (Cleve und Lämmel 2020).

Eine erweiterte Form der linearen Regression ist die *multivariante lineare Regression (Pyle 1999)*. Bei der multivariaten linearen Regression wird angenommen, dass die Output-Variablen y_i als gewichtete Summe von verschiedenen Inputvariablen$x_1, \dots, x_n$ und Rauschen in Form einer linearen Funktion beschrieben werden (Alpaydın 2022). Dadurch dass mehrere Variablen existieren, ist die Approximation genauer (Pyle 1999). Jedoch sollte die Anzahl der Inputvariablen nicht zu hoch gewählt werden, da das Modell sonst instabil werden kann (Ghavami 2020).

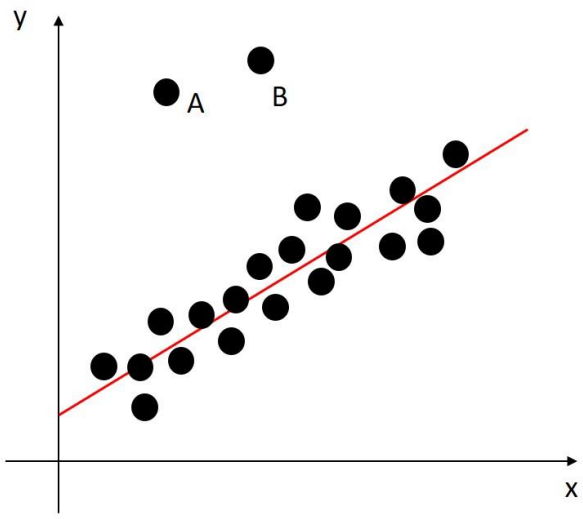

Abbildung 4-1 Beispiel einer linearen Regression

Wird die lineare Regression aus dem vorherigen Abschnitt betrachtet, stellt sich die Frage, wie dieses Verfahren reagieren würde, wenn die Anzahl an verrauschten Daten deutlich höher (ähnlich wie in Abbildung 4-1 die Punkte A und B) wäre als lediglich zwei Einträge. Nach Runkler (2016) ist die gewöhnliche lineare Regression besonders anfällig für solche Punkte, da diese den Fehler stark beeinflussen. Daraus folgt, dass bei einer hohen Anzahl an verrauschten Daten, die Fehlerfunktion die verrauschten Daten approximieren würde und diese dann fälschlicherweise als richtig interpretiert werden könnten. Um dennoch diese Datenbestände mithilfe der Regression analysieren zu können, ist eine lineare Regression mit robuster Fehlerfunktion notwendig (Runkler 2016). Als Beispiel für eine robuste Fehlerfunktion schlägt Runkler die *Summe der kleinsten quadratischen Fehler* oder auf Englisch *Least Trimmed Squares* vor. Bei dieser Fehlerfunktion werden die quadratischen Fehler der Größe nach sortiert und nur die m kleinsten berücksichtigt (Runkler 2016). Dadurch, dass nur die m kleinsten Fehlerwerte berücksichtigt werden, werden verrauschte Werte nicht berücksichtigt und somit wird auch die Fehlerfunktion nicht beeinflusst.

4.3 Klassifikation

Tabelle 4-4 Literaturübersicht zu den Klassifikationsverfahren

Autor (Jahr)	Titel	Vorgehen zur Behandlung von Rauschen	Data-Mining-Verfahren	Art der Publikation
Aggarwal (2015)	Data mining – The Textbook	Glätten; Identifizieren und Löschen; Filtern	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse; Filter-Verfahren	Monographie
Alpaydın (2022)	Maschinelles Lernen	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse	Monographie
Bramer (2020)	Principles of Data Mining	Identifizieren und Löschen	Klassifikation; Clusteranalyse	Monographie
Cleve und Lämmel (2020)	Data Mining	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse	Monographie

Frochte (2021)	Maschinelle Lernen	Identifizieren und Löschen	Klassifikation; Clusteranalyse	Monographie
Kantardzic (2011)	Data Mining	Identifizieren und Löschen	Klassifikation; Clusteranalyse	Monographie
Larose und Larose (2014)	Discovering Knowledge in Date	Identifizieren und Löschen	Klassifikation; Clusteranalyse	Monographie
Lee et al. (2020)	Automatic Bridge Design Parameter Extraction for Scan-to-BIM	Identifizieren und Löschen	Klassifikation	Zeitschriftenaufsatz
Luengo et al. (2020)	Big Data Preprocessing	Identifizieren und Löschen; Filtern	Klassifikation; Filter-Verfahren	Monographie
Runkler (2016)	Data Analytics	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse Klassifikation	Monographie

Die in den Unterkapiteln 4.1 und 4.2 beschriebenen Verfahren sind für numerische Datentypen geeignet. In diesem Unterkapitel wird die Klassifikation vorgestellt. Die Klassifikation ist üblicherweise eine Aufgabe die erst in der Data-Mining-Phase des KDD-Prozesses relevant ist. Durch die Literaturrecherche konnte jedoch ermittelt werden, dass einige Klassifizierungs-Verfahren auch in der Lage sind verrauschte Daten zu identifizieren. Bei den Klassifikationsverfahren ist das Ziel anhand bestehender Klassen eines Datensatzes, einem neuen Element dessen Klassenzugehörigkeit unbekannt ist, eine Klasse zuzuordnen.

Eines der bekanntesten Klassifizierungs-Verfahren ist das k-Nearest-Neighbors oder auf Deutsch k-Nächste-Nachbar-Verfahren. In der Literatur wird überwiegend das Verfahren auch als kNN-Algorithmus bezeichnet. In Abbildung 4-2 ist die Vorgehensweise an einem Beispiel dargestellt.

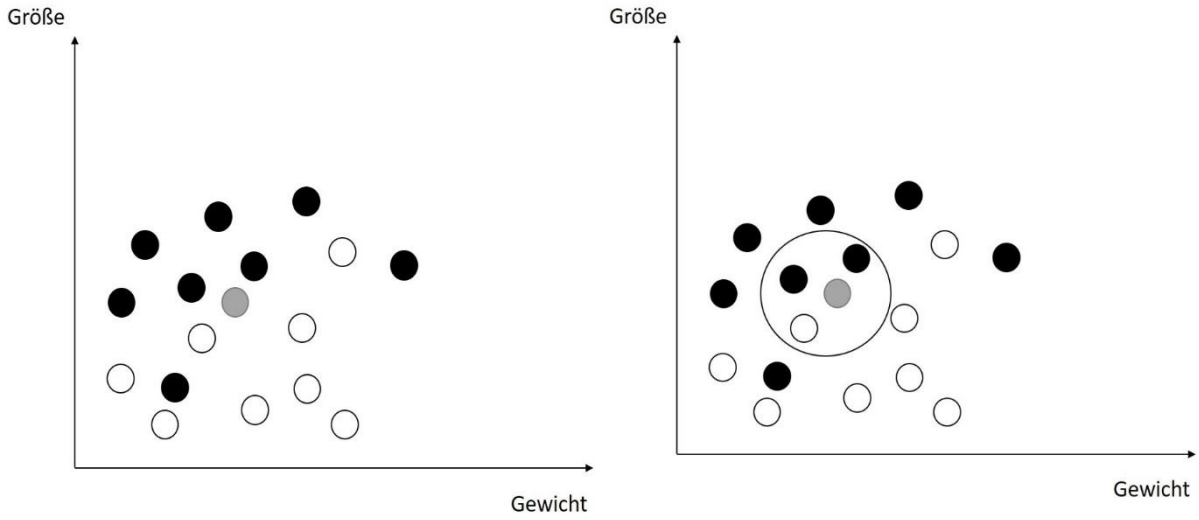

Abbildung 4-2 Beispiel für den kNN-Algrithmus

Als Beispiel kann ein Datensatz betrachtet werden, in dem zwei unterschiedliche Katzenrassen auf ihre Größen und ihr Gewicht untersucht werden. Die Katzenrasse A sind als dunkele Punkte und die Katzenrasse B als helle Punkte in Abbildung 4-2 dargestellt. Wenn anschließend die Größe und das Gewicht einer Katze bekannt ist, dessen Rasse aber noch ungeklärt ist, wird diese zunächst anhand der Daten für Größe und Gewicht in das Koordinatensystem eingetragen. In Abbildung 4-2 entspricht dies dem grauen Punkt. Als nächstes wird ein k Bestimmt welche dies Anzahl der zu untersuchenden Nachbarn festlegt. In diesem Beispiel wurde für k = 3 gewählt. Schließlich werden die drei ähnlichsten beziehungsweise nächsten Nachbarn des grauen Punktes betrachtet und anhand der Anzahl der Rassenzugehörigkeiten der Nachbarpunkte bestimmt, zu welcher Rasse der neue Datenpunkt angehören muss. In

diesem Fall sind zwei dunkele Punkte und ein heller punkt als Nachbarpunkte zu bestimmen. Dementsprechend ist der neue Datenpunkt der Katzenrasse A zuzuordnen.

Um mit dem kNN-Algorithmus die Verrauschten Daten zu erkennen, wird der Algorithmus für mehrere verschiedene Werte von k ausgeführt, bis erkenntlich ist, welche Punkte zu einer Klasse zugeordnet werden können und welche nicht. Punkte die zu einer Klasse gehören bezeichnen die Autoren Lee et al. (2020) als *Inlier* und Punkte außerhalb dieser Klasse als *outlier* (Ausreißer) beziehungsweise Rauschen bezeichnet. Dabei wird erneut deutlich, dass viele Autoren die Begriffe Ausreißer und Rauschen synonym verwenden. Darüber hinaus wird in der Literatur einheitlich die Ähnlichkeit mit dem Abstand zwischen den Punkten quantifiziert. Je näher die Punkte zueinander sind, desto ähnlicher sind sie daher auch zueinander. Dabei wird der Abstand mit der euklidischen Distanzfunktion berechnet. In der Literatur behauptet die Mehrheit der Autoren, dass das kNN-Verfahren für verrauschte Daten ungeeignet sei. Lee et al. (2020) verwendet jedoch den kNN-Algorithmus, um Klassenrauschen bei dem SCAN-to-BIM-Prozess zu identifizieren. Außerdem wird in der Literatur einheitlich erwähnt, dass das kNN-Verfahren für kleine Werte von k besonders schlechte Ergebnisse bei verrauschten Daten liefert. Lee et al. (2020) hat bei seiner Untersuchung k = 4 gewählt.

4.4 Cluster-Analyse

Tabelle 4-5 Literaturübersicht zu den Cluster-Verfahren

Autor (Jahr)	Titel	Vorgehen zur Behandlung von Rauschen	Data-Mining-Verfahren	Art der Publikation
Aggarwal (2015)	Data mining – The Textbook	Glätten; Identifizieren und Löschen; Filtern	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse; Filter-Verfahren	Monographie
Alpaydın (2022)	Maschinelles Lernen	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse	Monographie
Ataeyan und Daneshpour (2021)	Automated Noise Detection in a Database Based on a Combined Method	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz

Bramer (2020)	Principles of Data Mining	Identifizieren und Löschen	Klassifikation; Clusteranalyse	Monographie
Cios et al. (2007)	Data Mining – A Knowledge Discovery Approach	Glätten; Identifizieren und Löschen	Binning/Smoothing; Clusteranalyse	Monographie
Cleve und Lämmel (2020)	Data Mining	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse	Monographie
Ester et al. (1996)	A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz
Ester und Sander (2000)	Knowledge discovery in databases	Identifizieren und Löschen	Clusteranalyse	Monographie
Fayyad et al. (1996)	From Data Mining to Knowledge	Glätten; Identifizieren und Löschen	Regressionsanalyse; Clusteranalyse	Zeitschriftenaufsatz
Frochte (2021)	Maschinelle Lernen	Identifizieren und Löschen	Klassifikation; Clusteranalyse	Monographie
Herbold (2022)	Data-Science-Crashkurs	Identifizieren und Löschen	Clusteranalyse	Monographie

Kantardzic (2011)	Data Mining	Identifizieren und Löschen	Klassifikation; Clusteranalyse	Monographie
Kaur et al. (2010)	Effect of Noise	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz
Larose und Larose (2014)	Discovering Knowledge in Date	Identifizieren und Löschen	Klassifikation; Clusteranalyse	Monographie
Nematzadeh et al. (2020)	A hybrid model for class noise detection using k-means and classification filtering algorithms	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz
Raja (2022)	Data Mining and Machine Learning Applications	Glätten; Identifizieren und Löschen	Regressionsanalyse; Klassifikation; Clusteranalyse	Monographie
Ram et al. (2010)	A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz
Ros und Guillaume (2019)	A hierarchical clustering algorithm and an improvement	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz

	of the single linkage criterion to deal with noise			
Sloutsky et al. (2013)	Accounting for noise when clustering biological data	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz
Yin et al. (2009)	A Cluster-Based Noise Detection Algorithm	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz

Bei der Cluster-Analyse oder kurz *Clustering* werden die Daten oder Beobachtungen in kleine Gruppen (Cluster) unterteilt mit dem Ziel, dass Daten in gleichen Clustern möglichst ähnlich und Daten unterschiedlicher Cluster möglichst unähnlich sind (Ester und Sander 2000). Dabei wird die Ähnlichkeit durch Distanzfunktionen modelliert, die für Paare von Daten beziehungsweise Beobachtungen definiert sind. Dazu werden direkt und abgeleitete Eigenschaften betrachtet (Ester und Sander 2000). Daraus folgt, dass es mehrere Möglichkeiten gibt, die Daten in Cluster einzuteilen und daher wird nach Cleve und Lämmel (2020) zusätzlich eine *Qualitätsfunktion* benötigt, die eine Bewertung der Cluster-Bildung ermöglicht. Da wie in Abschnitt 2.3.1 beschrieben verrauschte Daten durch zufällig aufgezeichnete Fehler entstehen, die den Wert einer Messung beziehungsweise Beobachtung verändern können, ist es wahrscheinlich, dass die verrauschten Daten keinem Cluster zugeordnet werden und daher die Identifizierung dieser Daten kein großes Hindernis darstellt. In der Literatur ist ein breites Spektrum an Cluster-Verfahren zu finden und diese lassen sich in folgende Unterklassen unterteilen:

- Partitionierende Cluster-Bildung
- Hierarchische Cluster-Bildung
- Dichtebasierte Cluster-Bildung
- Cluster-Bildung mit Neuronalen Netzen

Die Literaturwerke aus Tabelle 4-5 nutzen besonders häufig die Partitionierende- und Dichtebasierte-Cluster-Bildung um verrauschte Daten zu entfernen.

Eine sehr bekannte Variante der partitionierenden Cluster-Bildung wird durch den k-Means-Algoritmus realisiert. Bei dem k-Means-Algorithmus werden die Cluster durch ihre Zentren repräsentiert (Herbold 2022). Das Zentrum eines Clusters wird auch Centroid genannt, wobei mit Zentrum der Schwerpunkt eines Clusters gemeint ist (Cleve und Lämmel 2020). Der k-Means-Algorithmus ist ein iterativer Prozess, der sich in vier Schritte unterteilen lässt. In Abbildung 4-3 ist die Cluster-Bildung mit dem k-means Algorithmus exemplarisch für drei Centroide (Kreuze) dargestellt. Im ersten Schritt wird die Anzahl und Position der Centroiden festgelegt. Dabei wird die Position der Centroide zu Beginn zufällig bestimmt. Im zweiten Schritt wird dann die Ähnlichkeit der Datenpunkte zu den Centroiden wie bereits erwähnt, durch eine Abstandsfunktion quantifiziert. Je kleiner der Abstand zwischen den Datenpunkten und den Centroiden ist, desto ähnlicher sind diese zueinander. Im dritten Schritt werden dann die Cluster so gebildet, dass jeder Datenpunkt zu dem Centroid zugeordnet wird, zu dem der Abstand am kleinsten ist. Dies ist in Abbildung 4-3 unten links dargestellt. Im letzten Schritt wird dann die Position der Centroiden neu berechnet. Dazu werden die Mittelwerte der x- und

y-Koordinaten aller Datenpunkte innerhalb eines Clusters berechnet. Die Mittelwerte entsprechen dann den neuen Koordinaten der Centroide. Es wird danach erneut wie in Schritt zwei die Ähnlichkeit bestimmt, sodass eine neue Zuordnung erfolgen kann. Dieses Vorgehen wird solange wiederholt, bis die Cluster sich nicht mehr verändern beziehungsweise kein Datenpunkt einem anderen Centroiden zugeordnet wird. Darüber hinaus ist es ebenfalls möglich, zuerst zufällige Cluster zu bilden und dann erst die Position der Centroiden zu berechnen

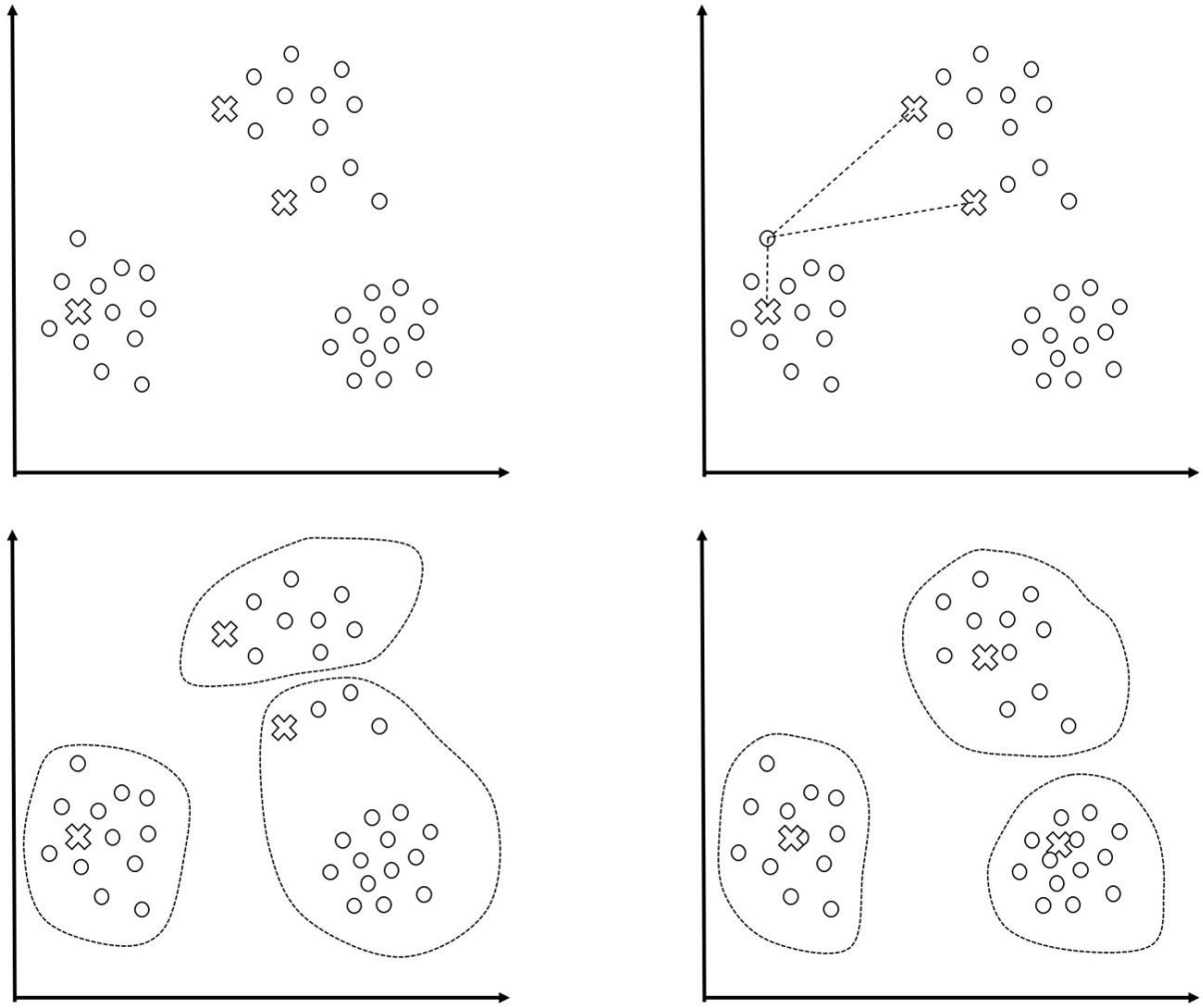

Abbildung 4-3 Beispielhafte Cluster-Bildung mit dem k-Means-Algorithmus in Anlehnung an Cleve und Lämmel (2020)

Wird die Literaturübersicht in Tabelle 4-5 betrachtet, so kann gesagt werden, dass die Literaturwerke die Funktionsweise des k-Means-Algorithmus zwar identisch beschreiben, jedoch unterschiedliche Aussagen treffen, ob der k-Means-Algorithmus zur Behandlung von verrauschten Daten geeignet ist oder nicht. Überwiegend wird der k-Means-Algorithmus als empfindlich gegen Rauschen und Ausreißer dargestellt. Nematzadeh et al. (2020) jedoch nutzt den k-Means-Algorithmus um Klassenrauschen zu identifizieren und diese anschließend

mit verschiedenen Klassifikationsverfahren zu filtern. Auch Ataeyan und Daneshpour (2021) nutzen den k-Means-Algorithmus in Kombination mit dem kNN-Verfahren und konnten dadurch bis zu 92% der verrauschten Daten aus einem Datensatz identifizieren. Darüber hinaus wird durch diese Methode auch Rauschen in Feldern mit unterschiedlichen Datentypen erkannt. In nahezu allen Literaturwerken aus Tabelle 4-5 wird statt dem k-Means-Algorithmus der k-Medoid-Algorithmus verwendet um verrauschte Daten zu behandeln. Der k-Medoid-Algorithmus wird einheitlich als Robust gegenüber zum Rauschen angesehen. Der Unterschied zwischen k-Medoid und k-Means ist, dass ein Datenpunkt als zentrales Element der einzelnen Cluster angesehen wird und nicht der Schwerpunkt. Das zentrale Element wird dann als Medoid bezeichnet. Um die Cluster-Bildung zu optimieren, wechseln die Elemente ständig ihren Status von Nichtmedoid zu Medoid und umgekehrt, bis die Qualität der Cluster sich nicht mehr verändert Cleve und Lämmel (2020).

Als dichtebasierte Cluster-Verfahren wird einheitlich das Density-Based Spatial Clustering of Applications with Noise (DBSCAN) – Verfahren vorgeschlagen. Bei dem DBSCAN-Verfahren werden die Cluster durch eine Mindest-Dichte von Punkten gebildet. Dabei werden drei verschieden Arten von Punkten unterschieden:

Tabelle 4-6 Punkte bei dem DBSCAN-Verfahren

Art der Punkte	**Erklärung**
Kernpunkte:	Diese Punkte bilden das Zentrum und gelten selbst als dicht
Randpunkte:	Diese Punkte befinden sich auf den Rändern der Cluster
Rauschpunkte:	Das sind Punkte die weder Kernpunkte noch Randpunkte sind und stellen das Rauschen dar.

Das DBSCAN-Verfahren wird einheitlich in der Literatur als sehr robust gegenüber Rauschen angesehen und wird von Ram et al. (2010) sogar als *Bahnbrechend* bezeichnet. Jedoch sehen Ram et al. (2010) Probleme bei den Dichteunterschieden innerhalb der Cluster. Wird die Tabelle A-6 betrachtet, so kann festgestellt werden, dass die Cluster-Verfahren besonders oft verwendet werden um verrauschte Daten zu identifizieren und zu behandeln und das dabei die Effektivität bei dem DBSCAN-Verfahren am größten ist

4.5 Filter-Verfahren

Tabelle 4-7 Literaturübersicht zu den Filter-Verfahren

Autor (Jahr)	Titel	Vorgehen zur Behandlung von Rauschen	Data-Mining-Verfahren	Art der Publikation
Aggarwal (2015)	Data mining – The Textbook	Glätten; Identifizieren und Löschen; Filtern	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse; Filter-Verfahren	Monographie
Bhattacharyya und Ghosh (2022)	Noise Filtering for Big Data	Filtern	Binning/Smoothing; Filter-Verfahren	Zeitschriftenaufsatz
Frénay und Verleysen (2014)	Classification in the presence of label noise: a survey	Filtern	Filter-Verfahren	Zeitschriftenaufsatz
García‐Gil et al. (2019)	From Big to Smart Data: Iterative ensemble filter for noise filtering in Big Data classification	Filtern	Filter-Verfahren	Zeitschriftenaufsatz

Luengo et al. (2020)	Big Data Preprocessing	Identifizieren und Löschen; Filtern	Klassifikation; Filter-Verfahren	Monographie
Wang et al. (2014)	Data mining based noise diagnosis and fuzzy filter design for image processing	Filtern	Filter-Verfahren	Zeitschriftenaufsatz
ZHU und WU (2004)	Class Noise vs. Attribute Noise: A Quantitative Study of Their Impacts	Filtern	Filter-Verfahren	Zeitschriftenaufsatz

Als letzte Kategorie konnten durch die systematische Literaturrecherche die Kategorie der Filter-Verfahren ermittelt werden. Filter-Verfahren werden in der Literatur einheitlich als besonders gut geeignete Verfahren gegen Klassenrauschen gesehen. Dabei kann ein Filter als eine Ansammlung von vielen Klassifizierungsalgorithmen gesehen werden. Diese Algorithmen lernen das Verhalten von verrauschten Daten anhand von Trainingsdaten. In der Literatur werden die Filter-Verfahren oft unterschiedlich kategorisiert. Frénay und Verleysen (2014) teilen die Filter-Verfahren in klassifizierende, abstimmende und partitionierende Verfahren auf. Luengo et al. (2020) teilen die Algorithmen in zwei unterschiedliche Kategorien auf. Zum einen in die Verfahren die Rauschen Entfernen und zum anderen in Verfahren die Rauschen Filtern. Zu den entfernenden Verfahren werden die *Homogeneous-Ensemble*-, Heterogeneous-Ensemble- und ENN-Filter- Verfahren genannt. Dabei ist das Homogeneous-Ensemble- und das Heterogeneous-Ensemble-Verfahren gleichzusetzen mit den partitionierenden Verfahren von Frénay und Verleysen (2014). Der ENN-Filter basiert dabei auf dem kNN-Algoritmus aus Kapitel 4.3. Die filternden Verfahren basieren auf verschiedenen Variationen des kNN-Algorithmus. Die Anwendung von Filter-Verfahren wird in der Literatur vorrangig in der Klassifizierung von Big Data angewendet. Wang et al. (2014) zeigt, dass die Filter-Verfahren auch für Bilddaten verwendet werden können. Durch die Anwendung verschiedener Filter konnten Wang et al. (2014) Bildrauschen in medizinischen CT-Aufnahmen entfernen.

Es konnte durch die Literaturrecherche keine bevorzugten Filter-Verfahren ermittelt werden. Aufgrund der hohen Anzahl an unterschiedlichen Filter-Verfahren wird im Rahmen dieser Bachelorarbeit, daher nicht weiter auf die Funktionsweisen einzelner Verfahren im Detail eingegangen. Für detaillierte Ausführungen von Filter-Verfahren wird auf die Literaturwerke aus Tabelle 4-7 verwiesen.

Abschließend zur Literaturrecherche soll noch erwähnt werden, dass viele Verfahren außerhalb der genannten Kategorien ermittelt werden konnten. Aufgrund der geringen Dichte der Literaturwerke zu diesen Verfahren, wurde der Schwerpunkt jedoch nur auf die Kategorien Binning und Smoothing, Regressionsanalyse, Klassifikation, Clusteranalyse und Filter-Verfahren gelegt.

5 Modellentwurf zur Auswahl geeigneter Data-Mining-Verfahren für spezifische Fragestellungen

In diesem Kapitel wird die Erarbeitung eines Modells in Form einer Entscheidungsmatrix beschrieben, welches die Auswahl geeigneter Data Mining Verfahren zur Behandlung von Rauschen unterstützen soll. Dazu wird zunächst im ersten Unterkapitel 5.1 auf Grundlage der Erkenntnisse der systematischen Literaturrecherche Anforderungen an das Modell abgeleitet. Im Unterkapitel 5.2 wird dann der Aufbau der Entscheidungsmatrix begründet sowie die Funktionsweise erläutert. Im letzten Unterkapitel 5.3 wird dann die Leistung dieser Arbeit kritisch diskutiert, um ein Fazit zu erstellen. Dabei wird insbesondere die Methodik der systematischen Literaturrecherche sowie der Modellentwurf betrachtet.

5.1 Ableiten von Anforderungen an das Modell

Um die Entscheidungsfindung durch ein Modell zu unterstützen, ist es zunächst notwendig, die Anforderungen an das Modell zu ermitteln. Aus den Ergebnissen der systematischen Literaturrecherche ist ersichtlich, dass einige Data-Mining-Verfahren bestimmten Einschränkungen unterliegen. Darüber hinaus können Datensätze unterschiedliche Eigenschaften wie zum Beispiel verschiedene Datentypen oder unterschiedliche Datenmengen aufweisen. Es ist daher sinnvoll, die Anforderungen an das Modell anhand der Eigenschaften, die ein Datensatz aufweisen kann und den Ergebnissen der systematischen Literaturrecherche aus Kapitel 4 abzuleiten.

Durch die Ergebnisse der systematischen Literaturrecherche in Kapitel 4 geht hervor, dass einige Data-Mining-Verfahren nur mit bestimmten Datentypen kompatibel sind. Zum Beispiel wäre die Auswahl von Binning- und Smoothing-Verfahren nur bei Datensätzen, die metrische Daten aufweisen, sinnvoll. Dies liegt daran, dass der glättende Effekt durch das Ersetzen von Mittelwerten entsteht. Ein Mittelwert bei nominalen Daten wie zum Beispiel das Geschlecht einer Person existiert jedoch nicht und dementsprechend wäre für diesen Fall die Binning- und Smoothing-Verfahren ungeeignet. Als erste Anforderung lässt sich damit der Datentyp ableiten. Dabei ist zu beachten, dass der Datentyp bei Data-Mining-Verfahren aufgeteilt werden muss in Input- und Outputvariablen, da diese nicht immer identisch sind wie zum Beispiel bei dem kNN-Verfahren (siehe Kapitel 4).

Eine weitere Eigenschaft von Datensätzen ist die Datenmenge beziehungsweise die Datendimension. In Kapitel 4 wurde ausgeführt, dass Binning- und Smoothing-Verfahren große Datenmengen effizient verarbeiten können, während das kNN-Verfahren oder hierarchische Cluster-Verfahren eher für kleinere Datensätze geeignet sind. Die Datenmenge hat folglich einen großen Einfluss auf die Ergebnisse der Data-Mining-Verfahren.

Problematisch ist es jedoch, die Datenmenge zu quantifizieren. Durch die systematische Literaturrecherche konnten keine Informationen ermittelt werden, ab welcher Anzahl an Elementen ein Datensatz als groß oder klein zu bezeichnen ist. Um die Größe von Datensätzen zu quantifizieren, bieten sich zwei Möglichkeiten an. Die erste Möglichkeit ist, durch eine experimentelle Anwendung der einzeln Verfahren auf unterschiedlich Datendimensionen einen Grenzwert für die Quantifizierung der Datenmenge zu ermitteln. Eine zweite Möglichkeit ist es, anhand der Ergebnisse aus Kapitel 4 einen Grenzwert zu schätzen. Da eine experimentelle Untersuchung den Rahmen dieser Bachelorarbeit überschreiten würde, wird daher die zweite Möglichkeit verwendet und der Grenzwert wird geschätzt. Aus der Literaturrecherche ist zu ermitteln, dass die hierarchischen Cluster-Verfahren nur für wenige Tausend Elemente geeignet sind und daher wird in dieser Arbeit als Grenzwert 1000 angenommen. Das heißt, dass ein Datensatz, der 1000 Elemente oder weniger besitzt als klein bezeichnet werden kann. Datensätze mit mehr als 1000 Elementen werden dann folglich als große Datensätze angesehen.

Eine zusätzliche Anforderung an das Modell ist die Art des Rauschens. Wie bereits in Abschnitt 2.3.1 erwähnt, kann Rauschen in Form von Klassenrauschen oder Attributrauschen erscheinen. Aus Kapitel 4 geht hervor, dass einige Verfahren nur Attributrauschen behandeln können wie zum Beispiel das Binning-Verfahren. Wenn die Art des Rauschens nicht berücksichtigt werden würde, könnte das Modell aufgrund der anderen Eigenschaften ein Verfahren auswählen welches gar nicht in der Lage wäre, das Rauschen zu identifizieren oder zu entfernen. Es stellt sich daher zusätzlich die Frage, ob diese Anforderung sogar höher gewichtet werden sollte als zum Beispiel die Datenmenge.

Neben den Datentypen und der Art des Rauschens ist die Art des Datensatzes ebenfalls entscheidend. Daten könne in Form einer Zeitreihe eines Bildes oder als Text vorliegen. Wie in Kapitel 4 ersichtlich, sind einige Verfahren besser für bestimmte Arten von Datensätzen geeignet als andere Verfahren. Beispielsweise sind die im Gegensatz zu den Binning- und Smoothing-Verfahren, die Klassifikationsverfahren eher ungeeignet für Zeitreihen.

Die in diesem Abschnitt abgeleiteten Anforderungen bilden die Kriterien zur Bewertung der einzelnen Data-Mining-Verfahren. Dabei können die Anforderungen wir bereits erwähnt auch unterschiedlich gewichtet werden. Wie die Anforderungen sowie die Gewichtung der Anforderungen in das Entscheidungsmodell eingebaut werden, wird im nachfolgenden Abschnitt erläutert. Die Anforderungen werden zur besseren Übersicht in der Tabelle 5-1 noch einmal dargestellt.

Tabelle 5-1 Anforderunge an das Entscheidungsmodell

Nr.	Anforderung
1)	Datentyp für Input- und Outputvariable
2)	Datenmenge
3)	Art des Rauschens
4)	Art der Daten

5.2 Entwicklung und Evaluation des Modells

Nachdem die Anforderungen an das Entscheidungsmodell im vorherigen Unterkapitel bestimmt wurden, wird in diesem Unterkapitel der Aufbau des Modells erläutert. Dazu wird die äußerliche Darstellung sowie die Funktionsweise erklärt und durch ein kurzes Beispiel die Entscheidungsfindung exemplarisch vorgeführt. Am Ende dieses Unterkapitels wird das Modell kritisch bewertet indem die Vor- und Nachteile des Modells aufgezeigt werden.

Als Entscheidungsmodell wird in dieser Arbeit eine Entscheidungsmatrix gewählt. Eine Entscheidungsmatrix hat eine tabellarische Form, in der die Kriterien, die Zeilen und die einzelnen Verfahren die Spalten der Tabelle bilden. Für diese Arbeit wären daher als Zeilen die in Tabelle 5-1 dargestellten Anforderungen und als Spalten die einzelnen Data-Mining-Verfahren aus Kapitel 4 anzusehen. Darüber hinaus wird jeweils eine Spalte für den Datensatz und die Gewichtung hinzugefügt. Jede Spalte mit Ausnahme der Spalte für die Gewichtung wird dann anhand der Kriterien bewertet, dabei kann die Bewertung in Form von Zahlen erfolgen. Mithilfe der Spalte für die Gewichtung ist es möglich, die Entscheidungsmatrix so zu erweitern, dass die Kriterien unterschiedlich gewertet beziehungsweise gewichtet werden. Durch eine Gewichtung können unlogische Schlussfolgerungen vermieden werden. Beispielsweise könnte ein Verfahren vorgeschlagen werden, welches zwar sehr viele Kriterien erfüllen kann, aber nicht in der Lage ist, metrische Datentypen zu analysieren. Eine Gewichtung des Datentyps würde dieses Problem lösen, da die Zeile, die den Datentyp bewertet, mehr Punkte erzielen würde und somit die Entscheidung so beeinflusst, dass das Verfahren, welches nicht für den Datentyp geeignet ist, auch weniger Punkte erhalten würde.

Um die Funktionsweise des Entscheidungsmodells noch einmal zu veranschaulichen, wird mithilfe der Abbildung 5-1 exemplarisch ein Beispiel vorgeführt. Um die Entscheidungsauswahl übersichtlich darzustellen, wurden bewusst einige Verfahren aus Kapitle 4 weggelassen.

Anforderungen:		Datensatz	Gewichtung	Binning	Lineare Regression	k-Medoid	DBSCAN	kNN
Datentyp Inputvariable	Nominal	0	1	0	0	1	1	1
	Ordinal	0	1	0	0	1	1	1
	Metrisch	1	1	1	1	1	1	1
Datentyp Outputvariable	Nominal	0	1	0	0	1	1	1
	Ordinal	0	1	0	0	1	1	0
	Metrisch	1	1	1	1	0	0	0
Datenmenge	Kleiner Datensatz	0	1	0	1	0	0	1
	Großer Datensatz	1	2	1	0	1	1	0
Art des Rauschens	Klassen-rauschen	0	1	0	0	1	1	1
	Attribut-rauschen	1	2	1	1	0	0	0
Art der Daten	Zeitreihen-daten	1	2	1	1	1	1	0
	Bilddaten	0	1	0	0	1	1	1
	Textdaten	0	1	0	0	1	1	1
		Ergebnis:		**8**	**6**	**5**	**5**	**1**

Abbildung 5-1 Beispielhafte Anwendung des Entscheidungsmodells

Nachdem die Anforderungen und die zur Auswahl stehenden Verfahren in die Entscheidungsmatrix eingetragen wurden, wird der Datensatz sowie die einzelnen Data-Mining-Verfahren mithilfe der Entscheidungsmatrix bewertet. Beispielsweise wird bei der Datenbank eine *1* eingetragen, falls das Kriterium auf die Eigenschaften des vorliegenden Datensatzes zutrifft und eine *0* falls das Kriterium nicht zutrifft beziehungsweise diese Information nicht vorhanden ist. Bei den Data-Mining-Verfahren wird geprüft, ob die Verfahren für die einzelnen Kriterien geeignet sind oder nicht. Dementsprechend wird dann auch entweder eine *1* oder eine *0* eingetragen. In Abbildung 5-1 ist zu sehen, dass ein Datensatz vorliegt, der metrische Inputvariablen besitzt und ebenfalls einen metrischen Datentyp für die Outputvariablen erwartet. Darüber hinaus ist zu erkennen, dass es sich bei dem Datensatz um einen großen Datensatz handelt. Die Größe des Datensatzes wurde dabei mit dem Grenzwert aus dem Unterkapitel 5.1 quantifiziert und daraus folgt, dass der Datensatz mehr als 1000 Elemente beinhaltet. Zusätzlich ist bekannt, dass die Daten in Form einer Zeitreihe vorliegen und dass nur Attributrauschen behandelt werden muss. Nachdem der Datensatz und die einzelnen Data-Mining-Verfahren bewertet wurden, werden alle Zeilen markiert, in der die Spalte der Datenbank eine *1* aufweist (siehe Abbildung 5-1). Anschließend werden nur die

markierten Zeilen berücksichtig. In jeder Zeile wird dann das Produkt aus Gewichtung und eingetragenem Wert gebildet. Das Ergebnis wird dann aus der Summe der innerhalb einer Spalte errechneten Produkte gebildet und wird dann in der untersten Zeile der jeweiligen Verfahren eingetragen. Das Verfahren welches den höchsten Wert in der Ergebniszeile aufweist ist dann am besten für den vorliegenden Datensatz und die Aufgabe Rauschen zu behandeln geeignet. Für das Beispiel errechnet sich das Ergebnis für das Binning-Verfahren, indem $1 \cdot 1 + 1 \cdot 1 + 2 \cdot 1 + 2 \cdot 1 + 2 \cdot 1 = 8$ gerechnet wird (Siehe Abbildung 5-1). Aus Abbildung 5-1 geht hervor, dass das Binning-Verfahren am besten und das kNN-Verfahren am schlechtesten für die Behandlung von Rauschen für den beispielhaften Datensatz geeignet ist. Bei einem anderen Datensatz würden dann andere Zeilen Berücksichtigt werden, sodass dann ein anderes Verfahren den höchsten Wert in der Ergebniszeile aufweisen würde.

Durch die exemplarische Durchführung des Entscheidungsmodells sowie der näheren Betrachtung von Abbildung 5-1 werden schnell die Vor- und Nachteile dieses Modells deutlich. Vorteilhaft ist besonders der einfache Aufbau in Form einer Entscheidungsmatrix. Der Aufbau lässt sich problemlos an neue Aufgaben anpassen und ist theoretisch unendlich erweiterbar. Wenn beispielsweise ein neues Data-Mining-Verfahren zur Behandlung von Rauschen entwickelt wurde, muss dieses lediglich in eine neue Spalte eingetragen und anhand der Kriterien bewertet werden. Diese Erweiterungen lassen sich folglich auch bei den Anforderungen beziehungsweise Kriterien realisieren, da in diesem Fall analog die Zeilen erweitert werden müssten. Darüber hinaus, ist das Modell aufgrund seiner einfachen Form problemlos in Data-Mining-Algorithmen implementierbar. Besonders die Bewertung der einzelnen Data-Mining-Verfahren durch Binärzahlen zeigt dies deutlich. Eine Software könnte daher problemlos dieses Entscheidungsmodell vor der Datenanalyse vorschalten und dann anhand der Ergebnisse das beste Verfahren zur Datenanalyse auswählen. Ein weiterer Vorteil ist, dass die Data-Mining-Verfahren anhand der Eigenschaften des vorliegenden Datensatzes bewertet werden. Somit wird gewährleistet, dass das Data-Mining-Verfahren für den vorliegenden Datensatz auch geeignet ist. Außerdem wird durch die Spalte *Gewichtung* eine einfache Möglichkeit geboten, die Kriterien unterschiedlich zu bewerten. Dadurch lässt sich das Modell an die aufgabenbezogenen Herausforderungen noch genauer anpassen. Des Weiteren ist es möglich die Anforderungen so anzupassen, dass mehrere Eigenschaften gleichzeitig berücksichtigt werden können. Beispielsweise kann ein Datensatz sowohl metrische als auch nominale Werte besitzen. Dadurch, dass die Anforderung an den Datentyp detaillierter aufgeteilt wurde, ist es problemlos möglich Datensätze mit mehreren Datentypen mit der Entscheidungsmatrix zu bewerten. Ein weiterer Vorteil des Entscheidungsmodells ist, dass durch einfache Anpassungen das Modell auch in anderen Phasen der Wissensentdeckung benutzt werden kann. Beispielsweise bei der Auswahl der Data-Mining

Verfahren zur Erkennung von Mustern innerhalb eines Datensatzes. Außerdem ist das Entscheidungsmodell insgesamt sehr leicht zugänglich und erfordert kein weiteres Fachwissen um es verwenden zu können.

Neben den Vorteilen sind jedoch auch einige Nachteile deutlich erkennbar. Beispielsweise ist in Tabelle 5-1 in der Ergebniszeile zu sehen, dass das k-Medoid- und das DBSCAN-Verfahren beide fünf Punkte erzielt haben. Wenn dies die Höchstpunktzahl gewesen wäre, hätte das Modell diese zwei Verfahren vorschlagen beziehungsweise gleichwertig gesehen. Dieses Ergebnis ist jedoch nicht akzeptabel, da der Anwender durch das Modell höchstens die Verfahren eingrenzt, die geeignet sein könnten, aber nicht in der Lage ist, eine eindeutige Entscheidung zu treffen. Um dieses Problem zu lösen, müssen Anforderungen abgeleitet werden, sodass alle Verfahren sich untereinander unterscheiden. Nach Abbildung 5-1 würde das k-Medoid und das DBSCAN-Verfahren sich immer identisch verhalten, da nur einige wenige Anforderungen abgleitet wurden. Darüber hinaus sind weitere Probleme bei den Anforderungen zu erkennen. Neben den Eigenschaften eines Datensatzes, die im Unterkapitel 5.1 aufgeführt wurden, existieren oft weitere Eigenschaften, die hier nicht berücksichtigt wurden. Beispielsweise können Daten in Form von Bildern oder Textabschnitten vorliegen und die Anforderungen dieser Datenmenge sind deutlich unterschiedlich im Gegensatz zu numerischen Datensätzen. Vor jeder Nutzung müssen daher die Anforderungen erst abgeleitet beziehungsweise ermittelt werden, wodurch der Zeitaufwand erheblich erhöht werden kann. Außerdem ist es nicht immer möglich, eine Anforderung zu quantifizieren wie zum Beispiel die Datenmenge, die im Unterkapitel 5.1 bereits thematisiert wurde. Um solche Anforderungen quantifizieren zu können, sind mehrere Experimente im Voraus durchzuführen und zu analysieren. Dadurch kann ermittelt werden, bei welcher Dimension von Daten die Verfahren ungenaue Ergebnisse erzeugen. Ein weiteres Problem ist die eingeführte Gewichtung. Es müsste wie bei der Datenmenge analysiert werden, welche Anforderung stärker gewichtet werden sollten. Dazu müsste bei jedem neuen Datensatz diese Untersuchung erneut durchgeführt werden, da die Datensätze unterschiedlich sein können. Es ist erkennbar, dass das Modell selbst eine Art Vorverarbeitung benötigt. Diese kann aufgrund der Variationen der Eigenschaften eines Datensatzes sehr umfangreich sein.

Abschließend lässt sich sagen, dass das Modell durchaus die Entscheidungsfindung unterstützen kann. Die übersichtliche Darstellungsform sowie die hohe Anpassungsfähigkeit an die aufgabenbezogenen Herausforderungen sind besonders positive Merkmale des Modells. Jedoch müssten vorab mehrere Informationen über die verschiedenen Eigenschaften eines Datensatzes sowie die Stärken und Schwächen der Data-Mining-Verfahren bekannt sein, um ein eindeutiges Ergebnis zu gewährleisten. Ohne eine zeitintensive Untersuchung der möglichen Anforderungen sowie die Quantifizierung der Anforderungen selbst ist es jedoch

nicht möglich, mit dem Modell eine Entscheidung zu treffen. Schlussendlich lässt sich sagen, dass das Modell ein hohes Potenzial aufweisen kann, wenn die nötigen Erkenntnisse durch Experimente erlangt werden, jedoch zum jetzigen Zeitpunkt nicht empfehlenswert ist.

5.3 Diskussion und Fazit

Im letzten Teil dieser Arbeit werden die Ergebnisse zur Beantwortung der Forschungsfrage aus den vorherigen Kapiteln zusammengefasst und Interpretiert. Weiterhin werden die Limitationen dieser Arbeit aufgezeigt. Dazu erfolgt eine kritische Auseinandersetzung der Methodik dieser Arbeit. Schließlich wird die erbrachte Leistung in dieser Arbeit noch einmal reflektiert und zusammengefasst.

In dieser Arbeit wurde die Behandlung von Rauschen in der Datenvorverarbeitung durch eine systematische Literaturrecherche untersucht. Die Ergebnisse der Literaturrecherche konnten in einer Literaturübersicht festgehalten werden. Anhand der Ergebnisse konnte darüber hinaus ein Modell in Form einer Entscheidungsmatrix entwickelt werden, welches die Entscheidung bei der Auswahl der Verfahren für einen vorliegenden Datensatz unterstützt. Um die Behandlung von Rauschen mithilfe einer systematischen Literaturrecherche zu untersuchen, wurde mithilfe von Kapitel 2, welcher den Stand der Technik dieser Arbeit repräsentiert, Suchbegriffe abgeleitet. Anhand dieser Suchbegriffe und der Kombination der Suchbegriffe durch sogenannte Operatoren war es möglich relevante Literaturwerke zu identifizieren. Die Literaturübersicht am Ende von Kapitel 3 zeigt, dass die Verfahren die zur Behandlung von Rauschen geeignet sind, sich in die Kategorien Binning, Regression, Cluster-Analyse, Klassifikation und Filter-Verfahren unterteilen lassen. Wird diese Erkenntnis mit der Aussage von Cleve und Lämmel (2020) aus dem Abschnitt 2.3.2 verglichen, so ist ersichtlich, dass durch die systematische Literaturrecherche zwei weitere Kategorien, nämlich die Klassifikations- und Filter-Verfahren ermittelt werden konnten. Eine weitere Kategorisierungsmöglichkeit der Verfahren wird in der Literatur anhand der Art des Rauschens vorgenommen. Wie in Abschnitt 2.3.1 kann Rauschen in Form von Attributrauschen oder Klassenrauschen auftreten. Die Verfahren werden in der Literatur dementsprechend auch in Verfahren zur Behandlung von Klassenrauschen oder in Verfahren zur Behandlung von Attributrauschen unterteilt. Aus den Ergebnissen der Literaturrecherche wird darüber hinaus deutlich, dass die einzelnen Verfahren drei unterschiedliche Vorgehen zur Behandlung von Rauschen aufweisen. Die erste Möglichkeit ist das Glätten der verrauschten Attribute in dem die Werte durch errechnete Mittelwerte ersetzt werden. Zu diesen Verfahren gehören die Binning- und Smoothing-Verfahren sowie die Regressionsverfahren. Die zweite Möglichkeit ist die Identifizierung und anschließende Löschung der verrauschten Daten. Dazu zählen die Cluster- und Klassifikationsverfahren. Die letzte Möglichkeit verrauschte Daten zu behandeln

ist die direkte Ausgrenzung der verrauschten Daten. Dies wird mithilfe der Filter-Verfahren aus Kapitel 4.5 ermöglicht. Anhand dieser Erkenntnis lässt sich sagen, dass die Verfahren sich auch in die Kategorien glättende, identifizierende und filternde Verfahren unterteilen lassen. Mit der Literaturrecherche konnte daher ein weiterer Kategorisierungsansatz zur Behandlung von Rauschen ermittelt werden. Darüber hinaus wird durch die Betrachtung der Anzahl der Literaturen, die jede Kategorie aufweist deutlich, dass die Cluster-Analyse besonders häufig zur Behandlung von Rauschen vorgeschlagen wird. Bei der Cluster-Analyse ist auffällig, dass diese zwar in der Literatur oft vorgeschlagen wird um verrauschte Daten zu entfernen, jedoch bei der näheren Beschreibung der einzelnen Cluster-Verfahren lediglich die Identifikation von Ausreißern erläutert wird. Daraus lässt sich ableiten, dass auch Verfahren zur Erkennung von Ausreißern betrachtet werden sollten um Rauschen zu behandeln. Weiterhin konnte in Kapitel 4.4 ermittelt werden, dass der k-Means-Algorithmus in der Literatur als ungeeignet für verrauschte Daten gilt. Jedoch nutzen Tang und Khoshgoftaar (2004) den Algorithmus als Rauschfilter um Klassenrauschen zu behandeln indem sie eine einfache Anpassung an dem Algorithmus durchgeführt haben. Eine ähnliche Erkenntnis konnte bei dem kNN-Verfahren in Kapitel 4.3 ermittelt werden. Auch das kNN-Verfahren wurde in der Literatur überwiegend als ungeeignet deklariert, aber Lee et al. (2020) verwenden dass kNN-Verfahren zur Entfernung von Rauschen bei dem *Scan-to-BIM-Prozess*. Sowohl der k-Means-Algorithmus als auch der kNN-Algorithmus werden häufig in der Datenanalyse verwendet, jedoch nicht um verrauschte Daten zu identifizieren. Dass diese Verfahren auch in der Lage sind Rauschen zu behandeln, ist eine relevante Erkenntnis, die durch die systematische Literaturrecherche gezeigt werden konnte. Der Standpunkt der Literaturwerke die der Meinung sind, dass das k-Means-Verfahren und das kNN-Verfahren ungeeignet zur Behandlung von Rauschen sei, muss auf Grundlage der Erkenntnisse dieser Arbeit noch einmal hinterfragt werden. Darüber hinaus ist aufgefallen, dass kein Literaturwerk aus der Gesamtübersicht alle Verfahren die in dieser Arbeit ermittelt werden konnten beinhaltet. Während einige Autoren zum Beispiel Cleve und Lämmel (2020) nur die Binning-, Regressions- und Cluster- Verfahren erwähnen, liegt der Fokus bei anderen Autoren zum Beispiel bei García et al. (2015) auf die Filter-Verfahren. Neben den Unstimmigkeiten in der Literatur sowie den Kategorisierungsansätzen konnten durch die Literaturrecherche auch einige Limitationen der Verfahren aufgezeigt werden. Jedoch wurden die Limitationen nur in einer überschaubaren Anzahl an Literaturwerken angesprochen. Besonders in Zeitschriftenaufsätzen wurde die Auswahl der Verfahren nicht begründet, sondern lediglich die Vorgehensweise der Verfahren erläutert. Die Limitation sind jedoch von besonderer Bedeutung, da nur anhand dieser es möglich ist die Eignung eines Verfahrens zur Behandlung von Rauschen für einen vorliegenden Datensatz zu quantifizieren. Die Wichtigkeit dieser Limitationen ist darüber hinaus in Kapitel 5 bei der Modellentwicklung ersichtlich. Mithilfe des Entscheidungsmodells werden die Fähigkeiten der einzelnen Verfahren in Bezug

auf die Eigenschaften eines Datensatzes bewertet. Die Fähigkeiten der Verfahren können wiederum mit ihren Limitationen eingegrenzt werden. Die wenigen Anforderungen die ermittelt wurden zeigen, dass unter diesen Umständen das k-Medoid- und das DBSCAN-Verfahren völlig gleichwertig gesehen werden können, obwohl diese beiden Verfahren unterschiedliche Vorgehensweisen aufzeigen. Es müssen daher mehrere Anforderungen abgeleitet werden, damit die Verfahren sich in den Fähigkeiten unterscheiden. Außerdem ist die Anforderung der Datenmenge problematisch, da diese wie bereits in Kapitel 5.1 erläutert nicht ohne weiteres quantifiziert werden kann. Ein weiterer Faktor der die Ergebnisse des Entscheidungsmodells beeinflusst ist die Gewichtung der Anforderungen. Dies wird deutlich, wenn das Beispiel aus Kapitel 5.2 geringfügig verändert wird. Wenn zum Beispiel die Outoutvariable des Datensatzes von metrisch auf nominal geändert werden würde, so würde das Binning-Verfahren einen Punkt verlieren und die Verfahren k-Means und DBSCAN einen Punkt dazu gewinnen. Das Modell würde trotzdem das Binning-Verfahren mit sieben Punkten den anderen beiden Verfahren vorziehen, obwohl das Binning-Verfahren nicht fähig ist einen nominalen Datentyp als Outputvariable auszugeben und die anderen beiden Verfahren dafür normalerweise besser geeignet wären. Die Gewichtungen der einzelnen Anforderungen wurden in den Literaturwerken nicht behandelt. In Abbildung 5-1 wurde die Gewichtung zur Veranschaulichung willkürlich gewählt und basiert nicht auf Ergebnissen aus der Literaturrecherche.

Werden die Ergebnisse aus Kapitel 4 und 5 genauer betrachtet, so können einige Limitationen dieser Arbeit festgestellt werden. Besonders bei dem Entscheidungsmodell ist es offensichtlich, dass neben der systematischen Literaturrecherche auch empirische Forschungen nötig sind um die einzelnen Verfahren effizienter vergleichen zu können. Eine empirische Forschung würde jedoch den Rahmen dieser Bachelorarbeit überschreiten. Darüber hinaus ist ersichtlich, dass je nach Art der Daten auch andere Anforderungen abgeleitet werden müssen. Aufgrund der hohen Anzahl an Anforderungen wurde im Rahmen dieser Arbeit lediglich ein Beispiel zur Demonstration der Funktionsweise erläutert. Darüber hinaus wurde in dieser Arbeit nur die Binning- und Smoothing- sowie Regressions-, Cluster-, Klassifikations- und Filterverfahren betrachtet. In der Literaturrecherche konnten jedoch auch Verfahren identifiziert werden die zu keiner der zugehörigen Kategorien gehören. Auch hier wurde aufgrund des Umfangs auf die nähere Untersuchung dieser Verfahren verzichtet. Trotz der genannten Limitationen, konnten in dieser Arbeit einige Erkenntnisse abgeleitet werden. Es wurde zum Beispiel gezeigt, dass die Verfahren k-Means und kNN doch geeignet sein können um verrauschte Daten zu behandeln. Darüber hinaus wurde ein weiterer Kategorisierungsansatz ermittelt in der sich die Verfahren einordnen lassen. Auch das Entscheidungsmodell sollte nicht vernachlässigt werden. In der Literaturrecherche konnten

kaum Hinweise ermittelt werden, die die Schwächen und Stärken der Verfahren anhand der Eigenschaften von Datensätzen zeigen. Für ein Unternehmen kann es jedoch durchaus interessant sein zu wissen welches Verfahren am besten geeignet ist um die Daten zu bereinigen. Die Eigenschaft, dass das Entscheidungsmodell mit Binärzahlen arbeitet ist nicht zu unterschätzen, da das Modell, wenn es in einem ausgereiften Zustand ist, dann sogar die Entscheidung automatisch treffen könnte. Für ein Unternehmen würde das bedeuten, dass der Datensatz automatisch von dem Entscheidungsmodell analysiert wird und dann das Verfahren auswählt, welches am besten geeignet ist um die verrauschten Daten zu entfernen.

Insgesamt lässt sich sagen, dass in dieser Arbeit die relevantesten Verfahren zur Behandlung von Rauschen in der Datenvorverarbeitung ermittelt und in einer Gesamtübersicht dargestellt werden konnten. Des Weiteren konnte ein erster Ansatz in Richtung Bewertung der Verfahren sowie Entscheidungsfindung bei der Auswahl der Verfahren durch das Entscheidungsmodell geschaffen werden.

6 Zusammenfassung und Ausblick

Das Ziel dieser Arbeit war es, Verfahren zur Behandlung von Rauschen in der Datenvorverarbeitung bei der Wissensentdeckung in Datenbanken zu ermitteln und die Entscheidung bei der Auswahl der Verfahren zu unterstützen. Dazu wurde zunächst in Kapitel 2 die Wissensentdeckung in Datenbanken thematisiert sowie alle nötigen Definitionen zum Verständnis der Thematik eingeführt. Zusätzlich wurde die Problematik der Wissensentdeckung in Datenbanken durch das Rauschen erläutert. In Kapitel 3 wurde dann die Methodik dieser Arbeit vorgestellt. Die Methodik dieser Arbeit war eine systematische Literaturrecherche. Die systematische Literaturrecherche wurde mithilfe von Suchbegriffen durchgeführt, die durch die Definitionen und Erläuterungen aus Kapitel 2 abgeleitet wurden. Am Ende der systematischen Literaturrecherche konnte eine Gesamtübersicht der relevanten Literaturwerke erstellt werden. In Kapitel 4 wurden dann die Verfahren anhand der Literaturübersicht in Kategorien unterteilt und beschrieben. Darüber hinaus wurden innerhalb der Kategorien die Aussagen der Autoren miteinander verglichen. Nachdem die Verfahren zur Behandlung von Rauschen in der Datenvorverarbeitung in Kapitel 3 und 4 ermittelt wurden, wurde in Kapitel 5 ein Entscheidungsmodell konzipiert, um die Auswahl eines Verfahrens bei aufgabenbezogenen Herausforderungen zu unterstützen. Dazu wurden zunächst Anforderungen an das Modell in Kapitel 5.1 erarbeitet. In Kapitel 5.2 wurde dann der Aufbau des Models in Form einer Entscheidungsmatrix erläutert. Das Modell beurteilt die einzelnen Verfahren aus Kapitel 4 anhand der abgeleiteten Anforderungen. Zusätzlich wird der vorliegende Datensatz ebenfalls anhand der Anforderungen bewertet, um so einen Bezug zwischen Verfahren und Datensatz zu erstellen. Die Bewertung erfolgt durch ein Punktesystem und die einzelnen Anforderungen können dabei zusätzlich unterschiedlich gewichtet werden. Es konnte jedoch festgestellt werden, dass zwei Verfahren aufgrund der abgeleiteten Anforderungen sich für alle möglichen Datensätze immer identisch verhalten.

In dieser Arbeit ist erkennbar, dass besonders die Modellerstellung ausbaufähig ist, da durch die Literaturrecherche zu wenig Anforderungen abgeleitet werden konnten. Eine Anforderung an das Modell ist die Datenmenge. Die Datenmenge konnte jedoch nicht durch die Literaturrecherche quantifiziert werden. Daher wird vorgeschlagen, die Verfahren experimentell an verschiedenen Datensätzen zu untersuchen, um so zum einen mehrere Anforderungen ableiten zu können und zum anderen einen Schwellwert für die Datenmenge zu ermitteln. Außerdem werden dadurch zusätzlich die Stärken und Schwächen der Verfahren in Bezug auf die Eigenschaften von Datensätzen analysiert. Die Verfahren sollten auf alle möglichen Kombinationen der Anforderungen untersucht werden. Darüber hinaus wurde aufgrund des Umfangs auf eine Validierung des Entscheidungsmodells verzichtet, sodass dort ebenfalls eine Untersuchung zu empfehlen ist. Darüber hinaus zeigen die Ergebnisse der systematischen

Literaturrecheche, dass auch Verfahren, die zur Identifizierung von Ausreißern gedacht sind auch für die Rauscherkennung nützlich sein können wie zum Beispiel die Cluster-Verfahren. Es ist daher auch zu empfehlen, in diesem Gebiet weiter zu forschen.

7 Literaturverzeichnis

Aggarwal, Charu C. (2015): Data mining. The textbook. Cham: Springer.

Alpaydın, Ethem (2022): Maschinelles Lernen. 3., aktualisierte und erweiterte Auflage. Berlin, Boston: De Gruyter Oldenbourg (De Gruyter Studium). Online verfügbar unter https://www.degruyter.com/isbn/9783110740141.

Ataeyan, Mahdieh; Daneshpour, Negin (2021): Automated Noise Detection in a Database Based on a Combined Method. In: *Stat., optim. inf. comput.* 9 (3), S. 665–680. DOI: 10.19139/soic-2310-5070-879.

Bandow, Gerhard (2009): "Das ist gar kein Modell!". Unterschiedliche Modelle und Modellierungen in Betriebswirtschaftslehre und Ingenieurwissenschaften. Wiesbaden: Gabler (Springer eBook Collection Business and Economics).

Bhattacharyya, Souvik (2022): Noise Filtering for Big Data Analytics: De Gruyter.

Bhattacharyya, Souvik; Ghosh, Koushik (2022): Noise Filtering for Big Data: De Gruyter.

Bodendorf, Freimut (2006): Daten- und Wissensmanagement. 2., aktualisierte und erweiterte Auflage. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg (Springer-Lehrbuch).

Bramer, Max (2020): Principles of Data Mining. London: Springer London.

Cios, Krzystof J.; Pedrycz, Witold; Swiniarski, Roman W.; Kurgan, Lukasz A. (2007): Data mining. A knowledge discovery approach. New York, NY: Springer. Online verfügbar unter https://swbplus.bsz-bw.de/bsz254756778err.htm.

Cleve, Jürgen; Lämmel, Uwe (2020): Data Mining. 3. Auflage. Berlin: De Gruyter (De Gruyter Studium).

Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996): A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, S. 226–231.

Ester, Martin; Sander, Jörg (2000): Knowledge discovery in databases. Techniken und Anwendungen. Berlin, Heidelberg: Springer.

Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996): From Data Mining to Knowledge Discovery in Databases. In: *AI Magazine* (17), S. 37–54.

Frénay, Benoît; Verleysen, Michel (2014): Classification in the presence of label noise: a survey. In: *IEEE transactions on neural networks and learning systems* 25 (5), S. 845–869. DOI: 10.1109/TNNLS.2013.2292894.

Frochte, Jörg (2021): Maschinelles Lernen. Grundlagen und Algorithmen in Python. 3., überarbeitete und erweiterte Auflage. München: Hanser (Plus.Hanser-Fachbuch).

García, Salvador; Luengo, Julián; Herrera, Francisco (2015): Data Preprocessing in Data Mining. Cham: Springer International Publishing; Imprint; Springer (Intelligent Systems Reference Library, 72).

García‐Gil, Diego; Luque‐Sánchez, Francisco; Luengo, Julián; García, Salvador; Herrera, Francisco (2019): From Big to Smart Data: Iterative ensemble filter for noise filtering in Big Data classification. In: *Int J Intell Syst* 34 (12), S. 3260–3274. DOI: 10.1002/int.22193.

Ghavami, Peter (2020): Big data analytics methods. Analytics techniques in data mining, deep learning and natural language processing. 2nd edition. Boston, Berlin: De Gruyter (Business & economics).

Gupta, Shivani; Gupta, Atul (2019): Dealing with Noise Problem in Machine Learning Datasets: A Systematic Review. In: *Procedia Computer Science* 161, S. 466–474. DOI: 10.1016/j.procs.2019.11.146.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline (2012): Data mining. Concepts and techniques. 3rd ed. Amsterdam, Boston, Heidelberg, Amsterdam, Boston, Heidelberg: Morgan Kaufmann; Elsevier (The Morgan Kaufmann series in data management systems).

Hawkins, D. M. (1980): Identification of Outliers. Dordrecht: Springer (Springer eBook Collection Mathematics and Statistics).

Herbold, Steffen (2022): Data-Science-Crashkurs. Eine interaktive und praktische Einführung. Heidelberg: dpunkt.verlag.

Islam, Aminul; Talukder,, Mehedi Hasan; Hasan, Mosaddik (2013): Speckle Noise Reduction From Ultrasound Image Using Modified Binning Method and Fuzzy Inference System. In: *International Conference on Advances in Electrical Engineering*.

Kantardzic, Mehmed (2011): Data Mining. Concepts, Models, Methods, and Algorithms. Somerset: John Wiley & Sons Incorporated. Online verfügbar unter https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=697640.

Kaur, Amaninder; Kumar, Pankaj; Kumar, Paritosh (2010): Effect of Noise on the Performance of Clustering Techniques. Conference on Computational and Statistical Science (ICCSS 2010). In: *2010 International Conference on Networking and Information Technology (ICNIT 2010)*.

Kessler, Waltraud (2006): Multivariate Datenanalyse für die Pharma-, Bio- und Prozessanalytik. Ein Lehrbuch. [Elektronische Ressource]. Weinheim: WILEY-VCH. Online verfügbar unter http://onlinelibrary.wiley.com/book/10.1002/9783527610037.

Krämer, Walter (2015): Statistik für alle. Die 101 wichtigsten Begriffe anschaulich erklärt. Berlin: Springer Spektrum.

Larose, Daniel T.; Larose, Chantal D. (2014): Discovering knowledge in data. An introduction to data mining. 2nd ed. Hoboken, New Jersey: IEEE.

Lee, Jae Hyuk; Park, Jeong Jun; Yoon, Hyungchul (2020): Automatic Bridge Design Parameter Extraction for Scan-to-BIM. In: *Applied Sciences* 10 (20), S. 7346. DOI: 10.3390/app10207346.

Liu, Huan; Hussain, Farhad; Lim Tan, Chew; Dash, Manoranjan (2002): Discretization: An Enabling Technique 6, S. 393–423.

Luengo, Julián; García-Gil, Diego; Ramírez-Gallego, Sergio; García, Salvador; Herrera, Francisko (2020): Big Data Preprocessing. Enabling Smart Data: Springer International Publishing.

Nematzadeh, Zahra; Ibrahim, Roliana; Selamat, Ali (2020): A hybrid model for class noise detection using k-means and classification filtering algorithms. In: *SN Appl. Sci.* 2 (7). DOI: 10.1007/s42452-020-3129-x.

Otte, Ralf; Wippermann, Boris; Otte, Viktor (Hg.) (2020): Von Data Mining bis Big Data. München: Carl Hanser Verlag GmbH & Co. KG.

Petersohn, Helge (2005): Data Mining. Verfahren, Prozesse, Anwendungsarchitektur. München: Oldenbourg Verlag.

Pyle, Dorian (1999): Data preparation for data mining. San Francisco, Calif.: Morgan Kaufmann. Online verfügbar unter http://www.loc.gov/catdir/description/els033/99017280.html.

Raja, Rohit (2022): Data Mining and Machine Learning Applications. Unter Mitarbeit von Kapil Kumar Nagwanshi, Sandeep Kumar und K. Ramya Laxmi. Newark: John Wiley & Sons Incorporated. Online verfügbar unter https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6875444.

Ram, Anant; Jalal, Sunita; Jalal, Anand S.; Kumar, Manoj (2010): A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases. In: *IJCA* 3 (6), S. 1–4. DOI: 10.5120/739-1038.

Ros, Frédéric; Guillaume, Serge (2019): A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. In: *Expert Systems with Applications* 128, S. 96–108. DOI: 10.1016/j.eswa.2019.03.031.

Runkler, Thomas A. (2016): Data Analytics. Models and Algorithms for Intelligent Data Analysis. 2. Aufl. Wiesbaden: Springer Verlag.

Schlittgen, Rainer; Streitberg, Bernd (2001): Zeitreihenanalyse. 9 Aufl. [Erscheinungsort nicht ermittelbar]: Oldenbourg (Lehr- und Handbücher der Statistik).

Sharafi, Armin (2013): Knowledge Discovery in Databases: Springer Fachmedien Wiesbaden.

Sloutsky, Roman; Jimenez, Nicolas; Swamidass, S. Joshua; Naegle, Kristen M. (2013): Accounting for noise when clustering biological data. In: *Briefings in bioinformatics* 14 (4), S. 423–436. DOI: 10.1093/bib/bbs057.

Tang, W.; Khoshgoftaar, T. M. (2004): Noise identification with the k-means algorithm. In: 16th IEEE International Conference on Tools with Artificial Intelligence. 16th IEEE International Conference on Tools with Artificial Intelligence. Boca Raton, FL, USA, 15-17 Nov. 2004: IEEE Comput. Soc, S. 373–378.

vom Brocke, Jan; Simons, Alexander; Niehaves, Bjoern; Reimer, Kai; Plattfaut, Ralf; Cleven, Anne (2009): Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process. In: *In Proceedings of the 17th European Conference on Information Systems (ECIS)* (9), S. 2206–2217.

Wang, Yongfu; Wu, Gaochang; Chen, Gang; Chai, Tianyou (2014): Data mining based noise diagnosis and fuzzy filter design for image processing. In: *Computers & Electrical Engineering* 40 (7), S. 2038–2049. DOI: 10.1016/j.compeleceng.2014.06.010.

Wirth, R; Hipp, J (2000): Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining. 11th-13th April 2000, Crowne Plaza Midland Hotel, Manchester, UK.

Yin, Hua; Dong, Hongbin; Li, Yuxuan (2009): A Cluster-Based Noise Detection Algorithm. In: 2009 First International Workshop on Database Technology and Applications. 2009 First International Workshop on Database Technology and Applications, DBTA. Wuhan, Hubei, China, 25.04.2009 - 26.04.2009: IEEE, S. 386–389.

ZHU, XINGQUAN; WU, XINDONG (2004): Class Noise vs. Attribute Noise: A Quantitative Study of Their Impacts. In: *Artificial Intelligence Review* (22), S. 177–210.

Anhang

Tabelle A-1 Suchbegriffe

Suchbegriffe
Wissensentdeckung, Wissensentdeckung in Datenbanken
Knowledge Discovery, Knowledge Discovery in Databases
Datenvorverarbeitung
Data Preprocessing
Data Mining
Rauschen, Klassenrauschen, Attributrauschen
Noise, Class Noise, Attribute Noise
Binning
Regression
Smoothing
Cluster, Clustering, Clusteranalyse
Data
k-Means
k-Nearest-Neighbors
DBSCAN
Noise Filter
Data Filtering
Data Cleaning, Data Cleansing
Techniques, Methods

Tabelle A-2 Beispiel Suchvorgang Nr. 2

Operator	Suchfeld	Suchbegriff	Anzahl der Resultate
	Alle Felder	Wissensentdeckung	651
AND	Alle Felder	Wissensentdeckung + Rauschen	54

Tabelle A-3 Beispiel Suchvorgang Nr. 3

Operator	Suchfeld	Suchbegriff	Anzahl der Resultate
	Alle Felder	Rauschen	128482
AND	Alle Felder	Rauschen + Behandlung	33762
AND	Alle Felder	Rauschen + Behandlung + Datenvorverarbeitung	160

Tabelle A-4 Beispiel Suchvorgang Nr. 4

Operator	Suchfeld	Suchbegriff	Anzahl der Resultate
	Titel	Noise	1708229
AND	Titel	Noise ∧ Reduction	98017
AND	Titel	Noise ∧ Reduction ∧ Data	2417
AND	Titel	Noise ∧ Reduction ∧ Data ∧ Algorithm	102

Tabelle A-5 Beispiel Suchvorgang Nr. 5

Operator	Suchfeld	Suchbegriff	Anzahl der Resultate
	Titel	"Noise removal techniques“	169

Tabelle A-6 Ergebnisse der systmatischen Literaturrecherche

Autor (Jahr)	Titel	Vorgehen zur Behandlung von Rauschen	Data-Mining-Verfahren	Art der Publikation
Aggarwal (2015)	Data mining – The Textbook	Glätten; Identifizieren und Löschen; Filtern	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse; Filter-Verfahren	Monographie
Alpaydın (2022)	Maschinelles Lernen	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse	Monographie
Ataeyan und Daneshpour (2021)	Automated Noise Detection in a Database Based on a Combined Method	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz
Bhattacharyya und Ghosh (2022)	Noise Filtering for Big Data	Filtern	Binning/Smoothing; Filter-Verfahren	Zeitschriftenaufsatz
Bramer (2020)	Principles of Data Mining	Identifizieren und Löschen	Klassifikation; Clusteranalyse	Monographie

Cios et al. (2007)	Data Mining – A Knowledge Discovery Approach	Glätten; Identifizieren und Löschen	Binning/Smoothing; Clusteranalyse	Monographie
Cleve und Lämmel (2020)	Data Mining	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse; Klassifikation; Clusteranalyse	Monographie
Ester et al. (1996)	A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz
Ester und Sander (2000)	Knowledge discovery in databases	Identifizieren und Löschen	Clusteranalyse	Monographie
Fayyad et al. (1996)	From Data Mining to Knowledge	Glätten; Identifizieren und Löschen	Regressionsanalyse; Clusteranalyse	Zeitschriftenaufsatz
Frénay und Verleysen (2014)	Classification in the presence of label noise: a survey	Filtern	Filter-Verfahren	Zeitschriftenaufsatz
Frochte (2021)	Maschinelle Lernen	Identifizieren und Löschen	Klassifikation; Clusteranalyse	Monographie
García‐Gil et al. (2019)	From Big to Smart Data: Iterative ensemble filter for	Filtern	Filter-Verfahren	Zeitschriftenaufsatz

	noise filtering in Big Data classification			
Ghavami (2020)	Big data analytics methods	Glätten	Regressionsanalyse	Monographie
Herbold (2022)	Data-Science-Crashkurs	Identifizieren und Löschen	Clusteranalyse	Monographie
Islam et al. (2013)	Speckle Noise Reduction From Ultrasound	Glätten	Binning/Smoothing	Zeitschriftenaufsatz
Kantardzic (2011)	Data Mining	Identifizieren und Löschen	Klassifikation; Clusteranalyse	Monographie
Kaur et al. (2010)	Effect of Noise	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz
Kessler (2006)	Multivariate Datenanalyse für die Pharma	Glätten	Binning/Smoothing	Monographie
Larose und Larose (2014)	Discovering Knowledge in Date	Identifizieren und Löschen	Klassifikation; Clusteranalyse	Monographie
Lee et al. (2020)	Automatic Bridge Design Parameter Extraction for Scan-to-BIM	Identifizieren und Löschen	Klassifikation	Zeitschriftenaufsatz

Liu et al. (2002)	Discretization: An Enabling Technique	Glätten	Binning/Smoothing	Zeitschriftenaufsatz
Luengo et al. (2020)	Big Data Preprocessing	Identifizieren und Löschen; Filtern	Klassifikation; Filter-Verfahren	Monographie
Nematzadeh et al. (2020)	A hybrid model for class noise detection using k-means and classification filtering algorithms	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz
Otte et al. (2020)	Von Data Mining bis Big Data	Glätten	Regressionsanalyse	Monographie
Pyle (1999)	Data preparation for data mining	Glätten	Regressionsanalyse	Monographie
Raja (2022)	Data Mining and Machine Learning Applications	Glätten; Identifizieren und Löschen	Regressionsanalyse; Klassifikation; Clusteranalyse	Monographie
Ram et al. (2010)	A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz
Ros und Guillaume (2019)	A hierarchical clustering algorithm and an improvement	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz

	of the single linkage criterion to deal with noise			
Runkler (2016)	Data Analytics	Glätten; Identifizieren und Löschen	Binning/Smoothing; Regressionsanalyse Klassifikation	Monographie
Sloutsky et al. (2013)	Accounting for noise when clustering biological data	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz
Wang et al. (2014)	Data mining based noise diagnosis and fuzzy filter design for image processing	Filtern	Filter-Verfahren	Zeitschriftenaufsatz
Yin et al. (2009)	A Cluster-Based Noise Detection Algorithm	Identifizieren und Löschen	Clusteranalyse	Zeitschriftenaufsatz
ZHU und WU (2004)	Class Noise vs. Attribute Noise: A Quantitative Study of Their Impacts	Filtern	Filter-Verfahren	Zeitschriftenaufsatz