technische universität dortmund

itpl IT in Produktion und Logistik

Masterarbeit

zum Thema

Systematische Analyse von Datenreduktionsverfahren als Vorbereitung für Data Mining im Kontext der Wissensentdeckung in Datenbanken

zur Erlangung des akademischen Grades

Master of Science (M. Sc.)

Eingereicht von:	Christoph Klon
Matrikelnummer:	231866
Studiengang:	Maschinenbau

Ausgabedatum:	09.11.2023
Abgabedatum:	26.04.2024

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

Technische Universität Dortmund
Fakultät Maschinenbau
Fachgebiet IT in Produktion und Logistik

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Im heutigen Zeitalter des digitalen Fortschritts ist eine exponentielle Zunahme der Datengenerierung und -speicherung zu verzeichnen (Manyika et al., 2011). Dennoch besteht die Herausforderung darin, aus der schieren Menge an Daten echten Nutzen zu ziehen (Cukier und Mayer-Schoenberger, 2013). Es wird argumentiert, dass der Wert von Daten nicht in ihrer Quantität liegt, sondern in der Fähigkeit, durch analytische Prozesse tiefgreifende Einsichten und Wissen zu generieren, die strategische Entscheidungen in Unternehmen und der Wissenschaft unterstützen (Davenport und Harris, 2012). Die Disziplin der Wissensentdeckung in Datenbanken hat sich demnach in den letzten Jahrzehnten als entscheidender Forschungsbereich etabliert, der darauf abzielt, das verborgene Wissen innerhalb gespeicherter Daten zu erschließen (Han et al., 2011). Dieses interdisziplinäre Feld verbindet verschiedene Techniken, unter anderem der aus Statistik, der künstlichen Intelligenz und dem Datenamangement, um aus großen Datenmengen wertvolles Wissen zu extrahieren (Kelleher und Tierney, 2018).

Infolge des erheblichen Wachstums und der Konsolidierung der datengestützten Wissensentdeckung haben sich mit dem Ziel einheitliche Standards zu etablieren, verschiedene Vorgehensmodelle zur Wissensentdeckung in Datenbanken entwickelt (Azevedo und Santos, 2008). Trotz der Vielfalt an Modellen zur Wissensentdeckung in Datenbanken spielt modellübergreifend die Datenvorverarbeitung eine zentrale Rolle und gilt als unerlässlich, um qualitativ hochwertige Ergebnisse mittels Data Mining zu erzielen und somit letztendlich die Erschließung neuen Wissens zu ermöglichen (García et al., 2015). Gerade in heutigen Zeiten der zunehmenden Datenmengen spielt die Datenreduktion als Teil der Datenvorverarbeitung eine zentrale Rolle, um Daten in angemessener Zeit zu analysieren oder Analysen überhaupt zu ermöglichen (Han et al., 2011). In den vergangenen Jahrzehnten wurden aus diesem Grund zahlreiche Verfahren zur Reduktion von Daten entwickelt, die darauf abzielen, trotz der Reduktion relevante Informationen und Muster für die Analyse zu bewahren. Neben der Auswahl einer Vielzahl von Data-Mining-Algorithmen zur Analyse der Daten, stehen Anwender nun ebenfalls vor der Herausforderung, Verfahren der Vorverarbeitung und im speziellen der Datenreduktion auszuwählen. Die Wahl von Verfahren erfordert dabei nicht nur explizites Domänenwissen bezüglich der Daten, sondern auch ein tiefgreifendes technisches Verständnis der Verfahren selbst (Guyon und Elisseeff, 2003). Obwohl verschiedene Vorgehensmodelle zur Wissensentdeckung in Datenbanken die Datenreduktion als Vorbereitung für das Data Mining aufführen, bleibt die genaue Umsetzung und die Auswahl geeigneter Datenreduktionsverfahren unklar.

Um dieser Problematik zu begegnen verfolgt die vorliegende Arbeit das Ziel, die umfangreichen Entwicklungen im Bereich der Datenreduktionsverfahren systematisch zu analysieren, zu strukturieren und praktikable Kriterien zur Auswahl dieser Verfahren zu formulieren, um eine Bewertung der Verfahren für den Kontext der Wissensentdeckung in Datenbanken zu ermöglichen und somit eine Unterstützung für die Auswahl von Datenreduktionsverfahren zu bieten.

Um dieses Ziel zu erreichen, werden zunächst im einführenden Teil technische und kon-

textuelle Grundlagen der Datenreduktion erläutert. Dies umfasst eine grundlegende Betrachtung von Datenbanken, um die daraus resultierenden Datensätze zu differenzieren, die für den Kontext der Wissensentdeckung in Datenbanken relevant sind. Dabei werden sowohl Unterschiede in den Datensätzen erläutert als auch grundlegende analytisch relevante Aspekte sowie Datensatzeigenschaften differenziert betrachtet. Anschließend erfolgt die kontextuelle Einführung der Arbeit in das Themengebiet der Wissensentdeckung in Datenbanken. Dabei werden verschiedene Vorgehensmodelle dargelegt, um ein Verständnis von Wissensentdeckungsprozessen zu vermitteln. Des Weiteren werden die zentralen Phasen der Wissensentdeckung in Datenbanken eingehend dargelegt, insbesondere die Datenvorverarbeitung und das Data Mining. Die Ausführungen zur Datenvorverarbeitung dienen zur Referenz für eine klare Abgrenzung der Datenreduktion von anderen Vorverarbeitungsschritten, da diese in der Literatur nicht immer konsistent definiert sind. Da eine Datenreduktion oft mit der Relevanz bestimmter Dateninhalte im Zusammenhang mit anschließenden Analyseprozessen einhergeht, erfolgt ebenfalls eine grundlegende Einführung in das Data Mining, bei der verschiedene Paradigmen und Aufgaben des Data Minings differenziert werden. Schließlich werden Grundlagen zur Datenreduktion erörtert. Zunächst werden verschiedene Kategorisierungen zu Methoden der Datenreduktion anhand der Ansätze verschiedener Autoren dargelegt, um eine einheitliche Definition der Datenreduktion für den Verlauf dieser Arbeit zu entwickeln. Auf dieser Basis erfolgt die Darstellung verschiedener Methoden der Datenreduktion sowie einzelner repräsentativer Verfahren aus diesen Methoden. Diese dienen zum einen dem grundlegenden Verständnis für die weiteren Ausführungen dieser Arbeit und zum anderen als Referenz für spätere Entwicklungen.

Um die übergeordnete Zielsetzung zu erreichen, erfolgt im Hauptteil dieser Arbeit die Identifikation und Analyse von Datenreduktionsverfahren. Hierzu wird eine systematische Literaturrecherche durchgeführt in der wissenschaftliche Publikationen, die sich mit Datenreduktionsverfahren befassen, analysiert werden. Dabei werden zum einen, allgemeine Kriterien zur Auswahl von Datenreduktionsverfahren identifiziert und zum anderen die in den Publikationen behandelte Verfahren strukturiert erfasst. Um eine differenzierte Betrachtung der Datenreduktionsverfahren im Kontext der Wissensentdeckung in Datenbanken zu ermöglichen, werden die formulierten Kriterien kontextspezifisch angepasst. Basierend auf dieser kontextualisierten Betrachtung werden spezifische Kriterienausprägungen definiert, die aus den Ergebnissen der Literaturrecherche abgeleitet sind. Ausgehend von dieser Strukturierung werden repräsentative Verfahren vorgestellt und anhand der entwickelten Kriterien bewertet. Die Ergebnisse dieser Bewertungen werden übersichtlich zusammengefasst, um Anwendern eine fundierte Unterstützung bei der Auswahl von Datenreduktionsverfahren im spezifischen Kontext der Wissensentdeckung in Datenbanken zu bieten.

Schließlich erfolgt die Evaluation der erarbeiteten Ergebnisse durch die exemplarische Anwendung verschiedener Datenreduktionsverfahren anhand eines konkretes Fallbeispiels. Nach dieser praktischen Überprüfung werden die Ergebnisse der Arbeit eingehend diskutiert und ein Fazit gezogen. Abschließend wird eine Zusammenfassung der wesentlichen Erkenntnisse präsentiert, gefolgt von einem Ausblick auf mögliche zukünftige Forschungsarbeiten in diesem Bereich.

2 Kontextuelle und technische Grundlagen der Datenreduktion

Im diesem Kapitel werden sämtliche kontextuellen und technischen Grundlagen der Datenreduktion eingeführt, die für das Verständnis und die Erreichung der übergeordneten Zielstellung von Relevanz sind. Zunächst werden in Abschnitt 2.1 Grundlagen im Zusammenhang mit Daten, Datenbanken und Datensätzen erläutert. Anschließend erfolgt in Abschnitt 2.2 eine Einführung in die Wissensentdeckung in Datenbanken, wobei zentrale Aspekte wie das Data Mining und die Datenvorverarbeitung genauer beleuchtet werden. Schließlich werden in Abschnitt 2.3 grundlegende Unterscheidungen und Methoden der Datenreduktion erörtert.

2.1 Daten

Um den Begriff der Daten zu verstehen und zu definieren, insbesondere die Verbindung zu Informationen und Wissen, die für das Verständnis der späteren Ausführungen zur Wissensentdeckung in Datenbanken essentiell sind, wird auf das Konzept der Wissenstreppe von North (2016) Bezug genommen. Diese ist auszugsweise in Abbildung 2.1 dargestellt.

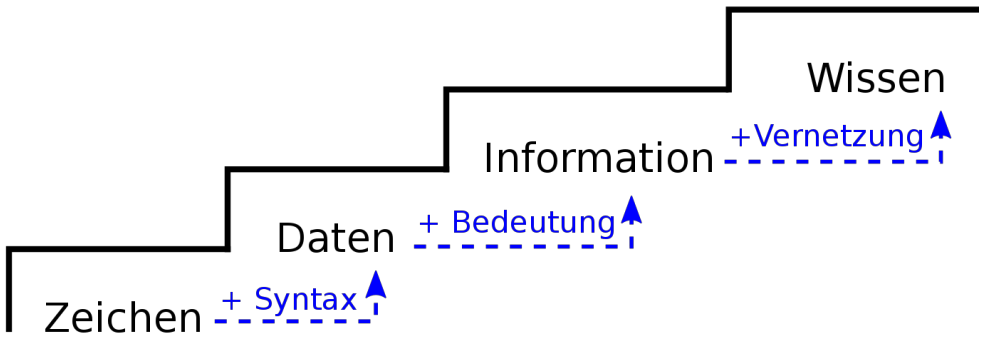


Abbildung 2.1: Ausschnitt der Wissenstreppe nach North (2016, S. 37)

Laut North (2016) lassen sich Daten zu Zeichen und Informationen abgrenzen. Der Autor beschreibt, dass Zeichen demnach einzelne Buchstaben, Ziffern oder Sonderzeichen sind und diese die unterste Ebene der Wissenstreppe bilden. Durch das Einfügen einer strukturellen Ordnung, bekannt als Syntax, werden diese Zeichen zu Daten. Durch die Zuweisung von Bedeutung zu den Daten werden diese zu Informationen. Wissen entsteht dann, wenn diese Informationen kontextualisiert oder mit Erfahrungen verknüpft werden. Weiterhin definiert North (2016, S. 37) Wissen als einen „Prozess der zweckdienlichen Vernetzung von Informationen“ sodass Wissen „als Ergebnis der Verarbeitung von Informationen durch das Bewusstsein“ entsteht. Probst et al. (2012) erweitern diese Perspektive,

indem sie hervorheben, dass Wissen sämtliche Kenntnisse und Fertigkeiten einschließt, die zur Problemlösung herangezogen werden. Weiterhin macht North (2016) auf die Bedeutung von Wissen für die Erzielung von Wettbewerbsfähigkeit und Wettbewerbsvorteilen aufmerksam, die über die in Abbildung 2.1 dargestellten Stufen hinausgeht.

Neben der Abgrenzung von Daten zu Zeichen, Informationen und Wissen wird auch die Struktur der Daten als wesentlich Eigenschaft betrachtet (Loshin, 2013). Krcmar (2015) teilt Daten in *strukturiert*, *semistrukturiert* und *unstrukturiert* ein. Strukturierte Daten beschreibt der Autor als in einem festen Schema organisiert, oft in Tabellenform, wodurch sie gut analysiert werden können. Semi-strukturierte Daten besitzen eine gewisse organisatorische Eigenschaften, wie Tags oder Markierungen, die ihre Analyse erleichtern (Pokorny, 2013). Ein Beispiel sind XML-Dokumente, in denen Daten in Tags wie Person und Adresse organisiert werden, wobei die genaue Anzahl und Art der Inhalte flexibel ist (Halevy et al., 2006). Unstrukturierte Daten hingegen, wie Texte, Bilder oder Videos, haben kein vordefinierte Struktur und erfordern komplexere Methoden der Datenverarbeitung, um nutzbare Informationen zu extrahieren (Loshin, 2013).

Die Datenmengen, aus denen potenziell Wissen generiert werden kann, haben in verschiedenen Branchen in den vergangenen Jahren eine exponentielle Zunahme erfahren (Schwab, 2017). Als Treiber dieser Entwicklung wird die fortschreitende Digitalisierung und Vernetzung im Kontext der vierten industriellen Revolution gesehen (Schwab, 2017). In dieser Dynamik haben Daten zunehmend den Charakter eines ökonomischen Gutes erlangt (Loshin, 2013). Daraus resultiert die Notwendigkeit eines umfassenden Datenmanagements, welches die Gewinnung, Nutzung und Steuerung von Daten gewährleistet und auf diese Weise einen Beitrag zur Wertschöpfung leistet (Chen et al., 2012). Ein kritischer Faktor innerhalb des Datenmanagements ist die Speicherung, die primär in Datenbanken realisiert wird (Chen et al., 2012). Diese werden im anschließenden Abschnitt erörtert.

2.1.1 Datenbanken

Nach Schicker (2017) ist eine Datenbank eine Sammlung von Daten, die untereinander in einer logischen Beziehung stehen und von einem eigenen Datenbankverwaltungssystem (Database Management System) verwaltet werden. Demnach werden Daten, die nicht zusammengehören, getrennt verwaltet, etwa in separaten Datenbanken (Schicker, 2017). Eine Datenbank zusammen mit einem Datenbankverwaltungssystem bildet das Datenbanksystem (Kemper und Eickler, 2015). Das Datenbankverwaltungssystem bietet eine logische Schnittstelle, um Anwendern Zugriffe, Manipulationen der Daten zu ermöglichen. Diese Schnittstelle ist entscheidend für die effiziente und effektive Interaktion mit den Daten (Silberschatz, 2010). Die Einteilung von Datenbanken erfolgt in der Regel anhand der zugrundeliegenden Datenmodelle, die die Struktur der Daten festlegen (Silberschatz, 2010). Die Wahl des Datenmodells hat direkte Auswirkungen auf die Art und Weise, wie Informationen gespeichert, abgerufen und organisiert werden (Kemper und Eickler, 2015).

In der Praxis werden SQL- und NoSQL-Datenbanken unterschieden (Kemper und Eickler, 2015). Diese Unterscheidung basiert auf den unterschiedlichen Ansätzen zur Datenmodellierung und -speicherung, die diese beiden Arten von Datenbanken verfolgen (Date, 2013). SQL-Datenbanken speichern Daten in strukturierten Form, nach einem festen Schema und basieren auf dem Relationalmodell (Kemper und Eickler, 2015). Das Relationalmodell er-

möglicht eine effiziente Datenmanipulation und Abfrage durch die Verwendung von SQL (Structured Query Language) (Codd, 1970). In diesem Modell werden Daten ausschließlich in Form von Tabellen, auch Relationen genannt, gespeichert (Kemper und Eickler, 2015). Diese tabellarische Darstellung fördert die Übersichtlichkeit und die logische Organisation von Daten (Codd, 1970). Die Zusammenhänge zwischen den einzelnen Tabellen werden über Beziehungen hergestellt Schicker (2017). Diese Beziehungen ermöglichen die Verknüpfung und den Abgleich von Daten über verschiedene Tabellen hinweg (Chen, 1976). Jeder Tabelle wird hierzu ein eindeutiger Name zugewiesen und jeder Zeile einer Tabelle eine spezielle Spalte, auch Identifikationsschlüssel oder schlicht Schlüssel genannt, das seine eindeutige Kennungen definiert(Chen, 1976). Das Entitäts-Beziehungs-Modell (engl. Entity-Relationship-Model) ist eine bekannte, konzeptionelle Modellierungsmethodik, das zur visuellen Darstellung der Daten und deren Beziehungen in einer relationaler Datenbank verwendet wird (Kemper und Eickler, 2015). Diese Methodik unterstützt den Datenbankentwurf durch die Bereitstellung eines grafischen Überblicks über die Datenstruktur (Chen, 1976).

Alle Aktivitäten, die mit einem relationalen Datenbankverwaltungssystem durchgeführt werden wie beispielsweise Datenbankabfragen, Datenspeicherung, Design, Aktualisierung und Wartung, basieren auf einer grundlegenden Arbeitseinheit, die als Transaktion bezeichnet wird (Silberschatz, 2010). Die dauerhafte, sichere und konsistente Speicherung der Daten wird dabei durch das ACID-Transaktionsprinzip gewährleistet, welches folgend nach Haerder und Reuter (1983) dargestellt wird.

Atomizität (Atomarity) Jede Transaktion erfolgt atomar, also zusammenhängend. Entweder wird die gesamte Transaktion vollständig ausgeführt, oder die Transaktion wird überhaupt nicht vollzogen, auch nicht in Teilen.

Konsistenz (Consistency) Eine Transaktion transformiert eine relationale Datenbank von einem konsistenten Zustand in einen anderen konsistenten Zustand.

Isolation Jeder transiente Zustand der relationalen Datenbank, der durch eine spezifische Transaktion verursacht wird, ist für andere Transaktionen unsichtbar, bis er abgeschlossen (committet) ist.

Dauerhaftigkeit (Durability) Sobald eine Transaktion abgeschlossen wurde, sind ihre Ergebnisse dauerhaft.

Ein weiteres wichtiges Merkmal eines relationalen Datenbankverwaltungssystems ist die Verfügbarkeit einer spezialisierten Sprache, genannt *Structured Query Language* (*SQL*), die darauf abzielt, Zugang zu Teilen der gesamten Datenbank zu ermöglichen (Date, 2011). SQL erlaubt es dem Benutzer, Abfragen zu spezifizieren, die eine Liste relevanter Attribute und Einschränkungen zu diesen Attributen enthalten (Elmasri und Navathe, 2020). Häufig bieten Datenbankverwaltungssysteme eine grafische Benutzeroberfläche, um die Formulierung von Abfragen zu erleichtern (Celko, 2010). Die Abfrage des Benutzers wird automatisch in eine Reihe von relationalen Operationen umgewandelt, wie Join, Selektion und Projektion, die für eine zeiteffiziente und ressourceneffiziente Verarbeitung optimiert und vom Datenbankverwaltungssystem ausgeführt werden (Elmasri und Navathe, 2020). SQL bietet auch die Möglichkeit, Daten zu aggregieren, indem Funktionen wie Summierung, Durchschnitt, Zählung, Maximum und Minimum berechnet werden (Date, 2011). Diese Fähigkeit ermöglicht es dem Benutzer, Antworten auf komplexere Abfragen zu er-

halten (Celko, 2010).

Den SQL-Datenbanken gegenüber stehen NoSQL-Datenbanken, die auf *nicht-relationalen* Ansätzen des Datenmanagements beruhen (Kemper und Eickler, 2015). Auch wenn SQL-Datenbanken im industriellen Kontext die vorherrschende Speicherform darstellen, haben NoSQL-Datenbanken in den letzten Jahren als Antwort auf die Grenzen relationaler Datenbanken und den Anforderungen moderner Web- und Mobile-Anwendungen starken Zuwachs erfahren (Strauch, 2011). Im Gegensatz zu relationalen Datenbanken, die strikte Schemata erfordern, bieten NoSQL-Datenbanken Flexibilität in Bezug auf Datenspeicherungsformate, wodurch Sie Daten in semi-strukturierter und unstrukturierter Form speichern können (Kemper und Eickler, 2015). Sie unterstützen eine Vielzahl von Datenmodellen, einschließlich Dokumenten-, Schlüssel-Wert-, Spaltenfamilien- und Graphen-Datenbanken (Tiwari, 2011). NoSQL-Datenbanken werden auch durch ihre Skalierbarkeit charakterisiert, da sie sind in der Lage sind, durch Hinzufügen von mehr Servern, horizontal zu skalieren (Sadalage und Fowler, 2012). Zudem bieten sie verbesserte Möglichkeiten für die Verarbeitung großer Datenmengen mit hoher Geschwindigkeit, wodurch sie für Anwendungen, die schnelle Lese- und Schreiboperationen erfordern geeignet sind (Cattell, 2011). Jedoch ist die Konsistenz der Daten in NoSQL Datenbanken lediglich verzögert gewährleistet, sofern eine hohe Verfügbarkeit und Ausfalltoleranz angestrebt werden (Strauch, 2011).

Zur Generierung von Wissen aus gespeicherten Daten der Datenbanken ist die Extraktion analytisch bedeutsamer Datensubsets in Form von Datensätzen meist eine gängie Praxis (Fayyad et al., 1996a). Obwohl sich die Anwendung von Techniken zur Wissensentdeckung in Datenbanken (*Knowledge Discovery in Databases*, KDD) in jüngerer Vergangenheit vermehrt auf nicht-relationale Datenbanken ausgeweitet hat (Cattell, 2011), fokussiert sich weiterhin ein beachtlicher Teil der wissenschaftlichen Forschung und praktischen Anwendungen auf Datensätze relationaler Datenbanken (Han et al., 2011). Grundlegende Einteilungen und Charakteristika von Datensätzen werden im folgenden Abschnitt erörtert.

2.1.2 Datensätze

Datensätze setzen sich aus einer Vielzahl von Datenobjekten zusammen, wobei jedes Datenobjekt eine spezifische Entität innerhalb des betrachteten Kontextes repräsentiert (Han et al., 2011). In einem Produktionskontext könnten solche Objekte beispielsweise einzelne Bauteile oder Maschinen darstellen, während in der Logistik Transportmittel, Ladungen und Lieferungen als primäre Datenobjekte fungieren (Chaudhuri und Dayal, 1997). Diese Objekte sind üblicherweise durch eine Reihe von Attributen charakterisiert, welche die spezifischen Eigenschaften der jeweiligen Entität durch Datenwerte beschreiben (Han et al., 2011). Darüber hinaus ist es in der Fachterminologie üblich, Datenobjekte mit Begriffen wie Stichproben, Beispiele, Instanzen, Datenpunkte oder schlicht Objekte zu bezeichnen (Witten et al., 2016). Sobald diese Objekte in einer relationalen Datenbank gespeichert werden, nehmen sie die Form von Datentupeln oder schlicht Tupeln an (Chaudhuri und Dayal, 1997). Hierbei entsprechen die Zeilen der Datenbanktabellen den einzelnen Tupeln, während die Spalten die zugehörigen Attribute repräsentieren (Witten et al., 2016). Zur Veranschaulichung zeigt Tabelle 2.1 einen fiktiven, relationalen Datensatz aus dem Logistikbereich. Jedes Tupel entspricht einem Transportmittel und jedes Attribut eine Eigenschaft, die das Transportmittel durch verschiedene Datenwerte beschreibt, wie zum

Beispiel die Ladungskapazität oder die Fahrzeug-ID.

Tabelle 2.1: Beispiel eines relationalen Datensatzes aus dem Logistikbereich

Transportmittel-ID	**Kapazität**	**Fahrzeugtyp**	**Standort**
1034262	5	LKW	Lager A
1052663	3	Lieferwagen	Lager B
1082246	8	Sattelschlepper	Lager C
...	...	...	...

Die Anzahl der Zeilen wird oft als die Größe eines Datensatzes bezeichnet, während die Anzahl der Attribute als die Dimension bezeichnet wird (Witten et al., 2016). Zur vereinfachten Darstellung werden Datensätze für die Datenanalyse oft abstrahiert in Form von $n \times m$ Datenmatrizen repräsentiert, wobei n die Anzahl der Instanzen eines Datensatzes entspricht und m der Anzahl der Spalten der korrespondierenden Attribute (Han et al., 2011). Die Begriffe Attribut, Merkmal und Variable werden in der wissenschaftlichen Literatur oft synonym verwendet (Bishop, 2006). Der Terminus Merkmal findet dabei meist in der Literatur des maschinellen Lernens Anwendung, wohingegen der Begriff der Variable tendenziell in der Statistik bevorzugt wird (Bishop, 2006). Im Fachbereich des Data Minings und der Datenbanken wird üblicherweise der Begriff Attribut verwendet (Han et al., 2011). Eine Menge von Attributen, die zur Beschreibung eines gegebenen Objekts verwendet wird, wird als Attributsvektor bezeichnet (Witten et al., 2016). Die Verteilung von Daten, die ein Attribut involvieren, wird als univariat bezeichnet (Bishop, 2006). Eine bivariate Verteilung bezieht sich auf zwei Attribute, während eine multivariate Verteilung sich auf Situationen bezieht, in denen ein Objekt durch mehr als zwei Merkmale beschrieben wird (Bishop, 2006).

Han et al. (2011) teilen Attribute in drei Typen ein: *nominal*, *ordinal* und *metrisch*.

Nominale Attribute repräsentieren kategorische Daten, bei denen keine Reihenfolge oder Hierarchie zwischen den Ausprägungen existiert (Han et al., 2011). Die Kategorien sind diskrete und voneinander unterscheidbare Einheiten (Han et al., 2011). Ein Beispiel hierfür sind Haarfarben (blond, brünett, schwarz) (Witten et al., 2016). Eine spezielle Form von nominalen Attributen stellen sogenannte binäre Attribute dar (Witten et al., 2016). Ein binäres Attribut ist ein nominales Attribut mit lediglich zwei Kategorien - 0 oder 1, wobei 0 typischerweise bedeutet, dass das Attribut abwesend ist und 1, dass es vorhanden ist (Witten et al., 2016). Binäre Attribute werden als Boolesche Attribute bezeichnet, wenn die Kategorien den Wahrheitswerten wahr und falsch entsprechen (Witten et al., 2016).

Ordinale Attribute sind ebenfalls kategorische Daten, bei denen Datenwerte jedoch in eine bestimmte Reihenfolge oder Rangfolge gebracht werden können (Han et al., 2011). Die genaue Differenz zwischen den Werten ist jedoch nicht definiert (Han et al., 2011). Ein Beispiel hierfür ist die Kundenzufriedenheit auf einer Skala unzufrieden, zufrieden und sehr zufrieden oder auch in Zahlen ausgedrückt von 1-3 (Witten et al., 2016).

Metrische Attribute können auf einer numerischen Skala gemessen werden (Han et al., 2011). Sie können in zwei Gruppen unterteilt werden - diskret und kontinuierlich (Han et al., 2011). Metrisch diskrete Daten repräsentieren Datenwerte, die auf einer

numerischen Skala gemessen werden, aber nur bestimmte, abzählbare Werte annehmen können, wie beispielsweise die Anzahl von Personen in einem Haushalt (Witten et al., 2016). Metrisch kontinuierliche Daten repräsentieren Werte, die auf einer numerischen Skala gemessen werden können und eine unendliche Anzahl von möglichen Werten zwischen zwei Punkten zulassen (Witten et al., 2016). Ein Beispiel hierfür ist die Körpergröße (Witten et al., 2016).

Die Unterscheidung der Attributtypen eines Datensatzes stellt einen grundlegenden Schritt in der Datenanalyse dar, da sie festlegt, welche mathematischen Methoden angewandt werden können und somit bestimmt, welche statistischen Maßzahlen berechnet werden können (Fahrmeir et al., 2016). Die aus den Attributen abgeleiteten grundlegenden Statistiken, einschließlich Maßzahlen für zentrale Tendenz, Variabilität und Korrelation, ermöglichen analytisch relevante Einblicke in wesentlichen Eigenschaften und die Verbindungen zwischen den Merkmalen (Hair et al., 2019). Wie in Tabelle 2.2 dargestellt, sind bestimmte statistische Operationen den jeweiligen Merkmalstypen zugeordnet, wobei jeder Attributstyp alle Eigenschaften und Operationen der übergeordneten Attributtypen übernimmt (Tan et al., 2019). Dies bedeutet, dass Eigenschaften oder Operationen, die für nominale, ordinale und metrische Attribute gültig sind, in einer kumulativen Weise anwendbar sind. Es ist jedoch nicht impliziert, dass die für einen Attributstyp geeigneten Operationen ohne Weiteres auf die ihm hierarchisch übergeordneten Attributstypen immer sinnvoll sind (Tan et al., 2019). Stevens (1946) definiert diese Attributstypen ursprünglich unter Berücksichtigung zulässiger Transformationen, die die Bedeutung eines Attributs beibehalten. Demnach gelten statistische Methoden für einen Attributstyp als angemessen, wenn sie nach einer solchen Transformation, die den semantischen Gehalt des Attributs unverändert lässt, konsistente Ergebnisse erzeugen. Tabelle 2.2 dient hierbei der Illustration und eine detaillierte Erörterung dieser Maßzahlen ist im Rahmen dieser Arbeit nicht vorgesehen.

Tabelle 2.2: Exemplarische Darstellung verschiedener Operationsmöglichkeiten und Maßzahlen zu Attributstypen Tan et al. (2019, S. 23)

Attributstypen	**Maßzahlen/Operationen**
Nominal	Modus, Entropie, Kontingenzkorrelation, χ^2 Test
Ordinal	Median, Perzentile, Rangkorrelation, Run-Tests, Vorzeichentests
Metrisch	Mittelwert, Standardabweichung, Pearsons Korrelation, t-, F-Tests

Neben den genannten Einteilungen von Attributstypen existieren im Bereich der Datenanalyse weitere gängige Einteilungen wie beispielsweise in qualitativ (nominal und ordnimal) und quantitativ (metrisch) oder in kontinuierliche und diskrete Attribute (Aggarwal, 2015). Neben diesen Unterscheidungen zu Attributstypen können diese auch als symmetrisch oder asymmetrisch klassifiziert werden (Han et al., 2011). Symmetrische Attribute sind solche, bei denen keine der Ausprägungen als dominierend oder signifikanter als die anderen betrachtet wird (Han et al., 2011). Asymmetrische Attribute hingegen weisen eine klare Hierarchie oder Bedeutungsunterschiede zwischen den Ausprägungen auf, wobei bestimmte Kategorien als wichtiger oder einflussreicher als andere angesehen werden (Witten et al., 2016).

Datensätze können spezielle Semantiken aufweisen, die beispielsweise von räumlicher oder zeitlicher Natur sind (Shekhar und Chawla, 2015). Datensätze können dahingehend in abhängigkeitsorientiert und nicht-abhängigkeitsorientiert eingeteilt werden (Han et al.,

2011). Nicht-abhängigkeitsorientierte Datensätze stellen beispielsweise Datensätze aus relationalen Datenbanken dar, die aus einzelnen Aufzeichnungen in Form von Instanzen bestehen, zwischen denen keine Abhängigkeiten bestehen (Aggarwal, 2015). In abhängigkeitsorientierten Datensätzen stehen Instanzen hingegen in einem Zusammenhang (Han et al., 2011). Diese Abhängigkeiten unterteilt Aggarwal (2015) in implizit und explizit. Datensätze mit impliziter Abhängigkeit sind beispielsweise Zeitreihen oder geografische Daten (Shekhar und Chawla, 2015). In diesen Datensätzen definieren Attribute den Kontext, auf deren Basis implizite Abhängigkeiten der Daten auftreten (Han et al., 2011). Im Falle einer Zeitreihe definieren ein oder mehrere Zeitstempel die zeitliche Abhängigkeit der Datenwerte, wie beispielsweise bei Sensormessungen (Aggarwal, 2015). Datensätze mit expliziten Abhängigkeiten beziehen sich in der Regel auf Graph- oder Netzwerkdaten, bei denen Kanten verwendet werden, um explizite Beziehungen zu spezifizieren (Aggarwal, 2015).

Es ist daher wichtig anzumerken, dass nicht alle Datensätze grundlegend in Form von Datenmatrizen vorliegen (Han et al., 2011). Komplexere Datensätze können, je nach zugrundeliegendem Datenbanktyp, aus Sequenzen (z. B. DNA, Proteine), Texten, Bildern, Audio oder Video und ähnlichem bestehen (Leskovec et al., 2020). In vielen Fällen ist es jedoch möglich, selbst wenn Daten nicht in einer Datenmatrix vorliegen, eine Transformation in diese Form zu ermöglichen (Witten et al., 2016). Beispielsweise kann aus einem Datensatz einer NoSQL-Datenbank, der Bilder enthält, eine Datenmatrix erstellt werden, bei der die Zeilen den Bildern und die Spalten Bildmerkmalen wie Farbe und Textur entsprechen (Zhou et al., 2018). Aggarwal (2015) beschreibt in diesem Kontext das Konzept der *Datentyp-Protabilität* und benennt verschiedene Methoden, um beispielsweise kategorische, Text-, Zeitreihen-, Graph-, räumliche- oder Sequenzdaten in numerische Datenmatrizen umzuwandeln, da numerische Daten die einfachsten und am weitesten erforschten für Data-Mining-Algorithmen sind und es daher besonders nützlich ist sich auf diese zu konzentrieren, wie verschiedene Daten in diesen umgewandelt werden können. Der Autor ergänzt jedoch, dass die Portierung von Datentypen in einigen Fällen zu einem Verlust an Darstellungsgenauigkeit und Ausdruckskraft führen kann.

Zusätzlich zu den dargelegten, grundlegenden Differenzierungen eines Datensatzes hinsichtlich der Attributstypen, dessen Größe und Dimensionalität, sind für die Analyse der Daten weitere Informationen relevant (Han et al., 2011). Diese Informationen beinhalten typischerweise die Existenz von fehlenden Werten (engl. Missing Values), sowie eines Labels (Witten et al., 2016). Ein Label repräsentiert ein spezifisches Attribut, das für die Klassifizierung oder Vorhersage in der Datenanalyse von Bedeutung ist (Aggarwal, 2015). Ist das Label von kategorischer Natur, so spricht man auch von einem Klassenlabel (Aggarwal, 2015). Klassenlabel implizieren die Zugehörigkeit verschiedener Datenwerte zu einer bestimmten Klasse oder Kategorie (Han et al., 2011). Die Existenz von Labels ermöglicht eine grundlegende Unterscheidung zwischen verschiedenen Paradigmen im Bereich des Data Minings, wie beispielsweise überwachtes und unüberwachtes Lernen, auf die in einem späteren Abschnitt dieser Arbeit eingegangen wird (Witten et al., 2016).

Während fehlende Werte, gerade wenn Datensätze in Form von Datenmatrizen vorliegen, einfach identifiziert werden können (Little und Rubin, 2019), können weitere analytisch relevante Informationen zu Datensätzen durch Anwendung statistischer Methoden gewonnen werden (Han et al., 2011). Im Kontext der Datenanalyse ist die Bewertung der qualitativen Eigenschaften eines Datensatzes von entscheidender Bedeutung, um die Güte der

Datenanalyse und die Verlässlichkeit der daraus resultierenden Erkenntnisse sicherzustellen (Han et al., 2011). Eine grundlegende Methode zur Beurteilung der Datenqualität ist die Analyse der Vollständigkeit der Daten, bei der geprüft wird, ob wichtige Werte fehlen oder inadäquat sind (Little und Rubin, 2019), oder die Konsistenz, bei der Daten auf Widersprüche oder Abweichungen von vorgegebenen Regelwerken untersucht werden (Witten et al., 2016). Die Genauigkeit der Daten, die misst, wie nahe die Datensätze an den tatsächlichen, realen Werten liegen, ist ebenfalls ein kritischer Faktor für die Datenqualität (Witten et al., 2016). Die Einzigartigkeit, also das Fehlen von Duplikaten, gewährleistet, dass jede Instanz eine eindeutige Informationseinheit repräsentiert und vermeidet Redundanzen (Hernández und Stolfo, 1998). Um all diese Qualitätsaspekte effektiv zu bewerten, werden oft explorative Datenanalysen (EDA) durchgeführt, die Einblicke in die Verteilung und Struktur der Daten geben und es ermöglichen, vorläufige Hypothesen über die Daten zu formulieren (Tukey, 1977). Spezielle Visualisierungstechniken wie Histogramme, Boxplots und Scatter-Plots unterstützen die Identifikation relevanter Informationen wie der Lage und Streuung sowie Ausreißern und Anomalien in den Daten (Tukey, 1977). Diese Techniken sind meist integraler Bestandteil der Disziplin der Wissensentdeckung in Datenbanken, welche den Kontext der vorliegenden Arbeit bildet und im folgenden Kapitel eingehender betrachtet wird (Fayyad et al., 1996a).

2.2 Wissensentdeckung in Datenbanken

Die Wissensentdeckung in Datenbanken stellt einen systematischen Prozess dar, der darauf abzielt, bisher unentdecktes und potenziell wertvolles Wissen aus umfangreichen Datensätzen zu extrahieren (Fayyad et al., 1996a). Fayyad et al. (1996a) definieren die Wissensentdeckung in Datenbanken erstmalig als "*nicht-trivialen Prozess der Identifizierung gültiger, neuartiger, potenziell nützlicher und letztlich verständlicher Muster in (großen) Datenmengen*". Weiterhin führen die Autoren aus, dass die Charakteristik *nicht-trivial* darauf zurückzuführen ist, dass der Prozess aus mehreren Phasen besteht, eine sorgfältige Planung, iteratives Vorgehen, fortlaufende Anpassungen und ein tiefgehendes Verständnis der Domäne erfordert. Zudem impliziert sie eine Suche oder Inferenz, die über einfache Berechnungen statistischer Maßzahlen hinausgeht (Fayyad et al., 1996a). Die Wissensentdeckung in Datenbanken integriert Methoden aus verschiedenen Disziplinen, einschließlich dem maschinellen Lernen, der Statistik und dem Datenbankmanagement (Witten et al., 2016).

Der Ursprung der Wissensentdeckung in Datenbanken liegt in der Reaktion auf die Analyse der stetig wachsenden Menge an gespeicherten Daten, die durch herkömmliche Datenanalysemethoden nicht mehr effektiv zu bewältigen war, da diese in der Regel auf einfachen Datenbankabfragen und der Erstellung standardisierter Berichte limitiert waren (Fayyad et al., 1996a). Diese Methoden zum Zweck der Umwandlung von Daten in Wissen beruhen auf der manuellen Analyse und Interpretation von Datensätzen, die langsam, sehr subjektiv und oft kognitiv nicht möglich ist (Han et al., 2011). Konventionelle Methoden verfolgen einen hypothesengetriebenen Ansatz, bei dem die Analyse durch die Verifikation von Hypothesen geleitet wird (Piatetsky-Shapiro, 1989). Piatetsky-Shapiro (1989) betonte die Notwendigkeit, über diese traditionelle Datenanalyse hinauszugehen und datengetriebene Analysemethoden zu entwickeln, um bisher unentdecktes Wissen und Erkenntnisse aus den stetig steigenden Datenmengen zu extrahieren. Datengetriebene, hypothesenfreie

Analysemethoden werden in der Wissensentdeckung in Datenbanken durch die Anwendung verschiedener Data-Mining-Algorithmen integriert, die es ermöglichen, Muster und Beziehungen in großen Datenmengen automatisch zu identifizieren (Fayyad et al., 1996a).

Die Anwendung und den Mehrwert dieser datengetriebenen Analysemethoden zeigen bereits in verschiedenen Branchen und Disziplinen, wie auch im produktionstechnischen Umfeld deutlich zu erkennen (Choudhary et al., 2009). In der Fertigungsindustrie werden Data-Mining-Vefahren genutzt, um Produktionsprozesse zu optimieren, indem sie beispielsweise Ausfallzeiten von Maschinen vorhersagen und Wartungspläne effizienter gestalten (Choudhary et al., 2009). Dies ermöglicht es Unternehmen, präventive Wartungsstrategien zu entwickeln, die unerwartete Stillstände minimieren und die Lebensdauer von Anlagen verlängern (Choudhary et al., 2009). Weiterhin werden in der Produktionsüberwachung Data-Mining-Verfahren angewendet, um Qualitätssicherungsprozesse zu verbessern (Harding et al., 2006). Durch die Analyse von Produktionsdaten können Muster identifiziert werden, die auf Qualitätsabweichungen oder -fehler hinweisen, was zu einer frühzeitigen Fehlererkennung und -behebung führt (Harding et al., 2006). Im Supply-Chain-Management ermöglichen datengetriebene Analysemethoden eine effizientere Gestaltung von Logistikprozessen (Chae, 2014). Sie unterstützen bei der Vorhersage von Lieferengpässen, optimieren Lagerbestände und verbessern die Nachfrageprognose, indem sie historische Lieferdaten und Markttrends analysieren (Chae, 2014).

Die begriffliche Abgrenzung zwischen der Wissensentdeckung in Datenbanken und dem Data Mining, obgleich in der Literatur häufig unklar getrennt, ist von signifikanter Bedeutung (Fayyad et al., 1996a). Data Mining fungiert als eine integrale Phase innerhalb der Wissensentdeckung in Datenbanken, in der Muster mittels mathematischer Algorithmen aus großen Datensätzen extrahiert werden (Han et al., 2011). Die aus dieser Phase gewonnenen Muster und Erkenntnisse erreichen jedoch noch nicht die Stufe des Wissens gemäß der Wissenstreppe, da sie ohne den Kontext der Anwendung oder das Verständnis der zugrundeliegenden Geschäftsprozesse betrachtet werden (North, 2016). Der gesamte Prozess der Wissensentdeckung in Datenbanken hingegen bietet einen breiteren, holistischen Ansatz, welcher neben der Extraktion von Mustern mittels Data Mining auch die Interpretation und Kontextualisierung der Ergebnisse einschließt (Fayyad et al., 1996a). Dieser Prozess ermöglicht schließlich die Generierung von Wissen, indem er Daten in einen sinnvollen Kontext setzt und so wertvolle Einsichten liefert, die für Entscheidungsträger von Nutzen sind (Fayyad et al., 1996a).

Eine wesentliche Phase der Wissensentdeckung in Datenbanken, um *gültige, potenziell nützliche und verständliche* Muster durch Data Mining extrahieren zu können, stellt die Datenvorverarbeitung dar (Fayyad et al., 1996a). Datenbanken in der realen Welt sind häufig durch negative Faktoren beeinträchtigt, wie das Vorhandensein von Rauschen, fehlenden Werten, irrelevanten Daten sowie deren erhebliche Größe und Dimension (Pyle, 1999a). Diese Faktoren führen dazu, dass Daten von geringer Qualität unzureichende Ergebnisse im Data Mining Prozess produzieren (Han et al., 2011). Weiterhin stellen Data-Mining-Algorithmen spezifische Anforderungen an den zu analysierenden Datensatz, bezogen auf Format, Qualität, Größe und Dimension (Witten et al., 2016). Nichterfüllung dieser Anforderungen kann dazu führen, dass eine Analyse nicht initiiert werden kann oder zu ungültigen Ergebnissen führt (Witten et al., 2016). Beispielsweise können einige Data-Mining-Algorithmen nur numerische Attributstypen verarbeiten oder sind nicht in der Lage, mit fehlenden Werten umzugehen (García et al., 2015). Andere Algorithmen

produzieren bei zu großer Dimensionalität unverständliche Ergebnisse oder sind aufgrund der Datengröße in keiner angemessenen Zeit zu analysieren (Aggarwal, 2015).

In der Wissensentdeckung in Datenbanken wird die Datenvorverarbeitung, abhängig vom gewählten Vorgehensmodell, unterschiedlich strukturiert (Cios et al., 2008). Einige Modelle gliedern die Vorverarbeitung in mehrere Phasen, um spezifische Aufgaben wie das Bereinigen von Daten, das Füllen von fehlenden Werten, die Normalisierung von Daten und die Merkmalsauswahl gezielt zu adressieren (Pyle, 1999a). Andere Modelle fassen die Datenvorverarbeitung in einer einzigen Phase zusammen, betonen jedoch die Bedeutung einer umfassenden Datenaufbereitung als Grundlage für den Erfolg der anschließenden Data-Mining-Analysen (Kurgan und Musilek, 2006). Unabhängig vom Modell ist die Datenvorverarbeitung entscheidend, da sie direkt die Qualität der extrahierten Muster und somit die Effektivität der Wissensentdeckung beeinflusst (Han et al., 2011).

Der folgende Abschnitt führt zunächst zwei bedeutende Vorgehensmodelle ein, um ein fundiertes Verständnis des Wissensentdeckungsprozesses zu fördern. Weiterhin wird der Terminus *Aktivitäten* zur Bezeichnung der spezifischen Aufgaben einer Phase verwendet.

2.2.1 Vorgehensmodelle

Im Kontext der Wissensentdeckung in Datenbanken haben sich im Laufe der Zeit unterschiedliche Vorgehensmodelle herausgebildet, die darauf abzielen, den Erfolg von KDD-Projekten zu gewährleisten (Azevedo und Santos, 2008). Laut Cios et al. (2008) kann die Unterteilung dieser Modelle in industrielle und akademische Ansätze eine nützliche Perspektive bieten, um die unterschiedlichen Schwerpunkte und Anwendungen innerhalb der Wissensentdeckung in Datenbanken zu verstehen. Industrielle Modelle sind typischerweise auf die Lösung konkreter Probleme innerhalb spezifischer Industriebereiche ausgerichtet, während akademische Modelle stärker auf theoretische Grundlagen und allgemeine Methoden fokussiert sind (Cios et al., 2008). Trotz dieser grundsätzlichen Unterscheidung zeigen sich die Modelle flexibel in ihrer Anwendbarkeit, sodass akademische Ansätze im industriellen Kontext genutzt werden können und umgekehrt (Cios et al., 2008). Um ein tieferes Verständnis von Wissensentdeckungsprozessen zu erhalten und gleichzeitig eine allgemeine, aber dennoch fokussierte Betrachtung der Datenreduktion innerhalb dieser Prozesse zu ermöglichen, werden das akademische Modell nach Fayyad et al. (1996a) und das industrielle Modell nach Chapman et al. (2000) beschrieben und verglichen. Diese Modelle haben in der Literatur an Popularität gewonnen und wurden bereits in zahlreichen KDD-Projekten angewendet (Cios et al., 2008).

KDD-Prozess nach Fayyad et al. (1996a)

In den 1990er Jahren schlagen Fayyad et al. (1996a) ein wegweisendes Vorgehensmodell zur Wissensentdeckung in Datenbanken vor. Dieses Modell, das als erstes seiner Art betrachtet werden kann, hat sich seitdem als grundlegender Ansatz für Wissensentdeckungsprozesse etabliert und diente als Inspiration für nachfolgende Vorgehensmodelle (Maimon und Rokach, 2010). Fayyad et al. (1996a) integrieren dabei wesentliche Abläufe von Brachman und Anand (1996) und erweitern diese zu einem umfassenden Vorgehensmodell. Dieses Vorgehensmodell, welches oft auch als *KDD-Prozess* bezeichnet wird, umfasst neuen Schritte

oder auch Phasen genannt (Fayyad et al., 1996a). Fünf dieser Schritte erfordern direkte Operationen an Daten. Vier weitere unterstützende Schritte komplettieren das Modell, die zum Projekterfolg beitragen. Abbildung 2.2 zeigt den schematischen Ablauf des KDD-Prozesses, welche die fünf datenbezogenen Schritte darstellt.

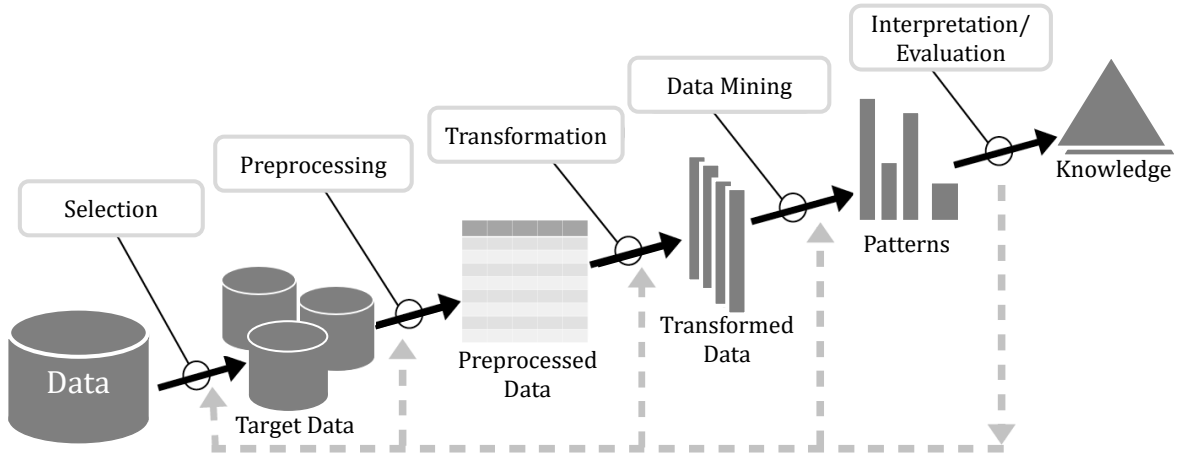


Abbildung 2.2: KDD-Prozess nach Fayyad et al. (1996a, S. 3)

1. *Domänenverständnis und Zieldefinition*: Dieser Schritt beinhaltet das Erlernen des relevanten Vorwissens und der Ziele des Endbenutzers des entdeckten Wissens.

2. *Datenselektion*: Hier wird ein Ziel-Datensatz festgelegt. Der Datenanalyst wählt eine Teilmenge von Attribute und Instanzen aus, die für die Durchführung von Entdeckungsaufgaben verwendet werden. Dieser Schritt umfasst in der Regel das Abfragen von Datenbanken, um die gewünschte Teilmenge auszuwählen.

3. *Datenbereinigung und -vorverarbeitung*: Dieser Schritt besteht darin, Ausreißer zu entfernen, mit Rauschen und fehlenden Werten in den Daten umzugehen und Informationen zu Zeitsequenzen und bekannten Abhängigkeiten zu berücksichtigen.

4. *Datenreduktion und -projektion*: Dieser Schritt besteht darin, nützliche Attribute durch Anwendung von Dimensionsreduktions- und Transformationsmethoden zu finden und eine invariante Repräsentation der Daten zu finden.

5. *Auswahl der Data-Mining-Aufgabe*: Hier passt der Datenanalyst die in Schritt 1 definierten Ziele an eine bestimmte Data Mining Aufgabe an, wie z.B. Klassifikation, Regression, Clusterbildung.

6. *Auswahl des Data-Mining-Algorithmus*: Der Datenanalyst wählt konkrete Algorithmen zur Suche nach Mustern in den Daten aus und entscheidet, welche Modelle und Parameter der verwendeten Algorithmen geeignet sein könnten.

7. *Data Mining*: Dieser Schritt erzeugt Muster in einer bestimmten repräsentativen

Form, wie z.B. Klassifizierungsregeln, Entscheidungsbäume, Regressionsmodelle, Trends.

8. *Interpretation der Ergebnisse*: Hier führt der Analyst die Visualisierung der extrahierten Muster und Modelle durch und die Visualisierung der Daten basierend auf den extrahierten Modellen.

9. *Anwendung des erschlossenen Wissens*: Der letzte Schritt besteht darin, das entdeckte Wissen in das Leistungssystem zu integrieren und es den interessierten Parteien zu dokumentieren und zu berichten. Dieser Schritt kann auch die Überprüfung und Lösung potenzieller Konflikte mit zuvor angenommenem Wissen umfassen.

   (Fayyad et al., 1996a)

Die Autoren dieses Modells erklären, dass normalerweise eine Reihe von Schleifen zwischen zwei beliebigen Schritten ausgeführt werden, geben jedoch keine spezifischen Details an (Cios et al., 2008). Das Modell bietet eine detaillierte technische Beschreibung in Bezug auf die Datenanalyse, mangelt jedoch wie eingangs beschrieben an einer Berücksichtigung geschäftlichen Aspekte (Cios et al., 2008). Dokumentierte Anwendungen des Modells finden lassen sich in der Literatur in verschiedener Bereichen finden, einschließlich dem Ingenieurwesen, der Medizin, der Produktion, dem E-Business und der Softwareentwicklung (Cios et al., 2008).

CRISP-DM nach Chapman et al. (2000)

Das CRISP-DM-Modell (Cross-Industry Standard Process for Data Mining) wurde Ende der 1990er Jahre durch eine gemeinsame Initiative von vier Unternehmen entwickelt: Integral Solutions Ltd., einem Anbieter von kommerziellen Data-Mining-Lösungen, NCR, einem Technologieunternehmen spezialisiert auf Datenbanktechnologien, DaimlerChrysler aus der Automobilbranche und OHRA, einem Versicherungsunternehmen (Cios et al., 2008). Das Modell ist in sechs Phasen untergliedert, welche die systematische Herangehensweise verdeutlicht und in der Fachliteratur wie in Abbildung 2.3 zu sehen abstrahiert dargestellt wird (Wirth und Hipp, 2000).

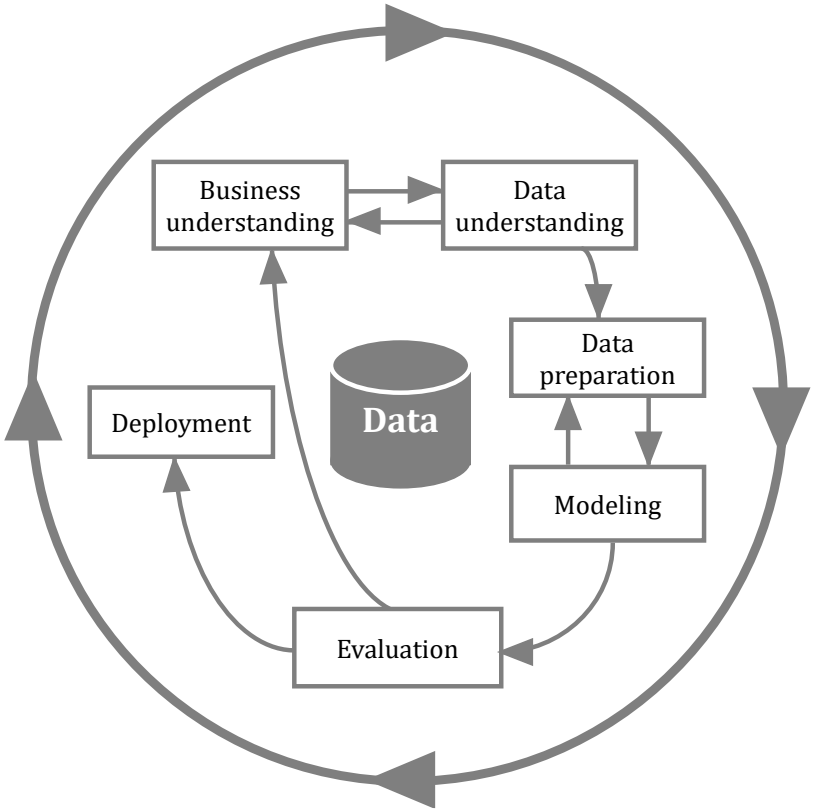


Abbildung 2.3: CRISP-DM nach Chapman et al. (2000, S. 6)

1. *Geschäftsverständnis*: Dieser Schritt konzentriert sich auf das Verständnis der Ziele und Anforderungen aus geschäftlicher Sicht. Er wandelt diese in Data-Mining-Problembeschreibungen um und entwirft einen vorläufigen Projektplan zur Erreichung der Ziele.

2. *Datenverständnis*: Dieser Schritt beginnt mit der anfänglichen Datensammlung und dem Vertrautmachen mit den Daten. Zu den spezifischen Zielen gehören die Identifizierung von Datenqualitätsproblemen, erste Einblicke in die Daten und die Erkennung interessanter Datensubsets.

3. *Datenvorbereitung*: Dieser Schritt umfasst alle Aktivitäten, die benötigt werden, um den endgültigen Datensatz zu erstellen, der im nächsten Schritt mittels Data-Mining-Algorithmen analysiert werden. Aufgaben der Datenaufbereitung werden nach Beschreibung der Autoren in diesem Modell meist mehrfach und nicht in einer bestimmten Reihenfolge durchgeführt. Zu den Aufgaben gehören die Auswahl von Tabellen, Instanzen und Attributen sowie die Umwandlung und Bereinigung von Daten für Modellierungswerkzeuge beziehungsweise Data-Mining-Algorithmen.

4. *Modellierung*: An diesem Punkt werden verschiedene Modellierungstechniken ausgewählt und angewendet. Modellierung beinhaltet üblicherweise den Einsatz mehrerer Methoden für denselben Data-Mining-Problemtyp und die Kalibrierung ihrer Parameter auf optimale Werte. Da einige Methoden ein spezifisches Format für Eingabedaten erfordern können, ist oft eine Wiederholung des vorherigen Schrittes notwendig.

5. *Evaluation*: Nachdem ein oder mehrere Modelle erstellt wurden, die aus der Perspek-

tive der Datenanalyse von hoher Qualität sind, wird das Modell aus der Perspektive der Geschäftsziele bewertet. Eine Überprüfung der ausgeführten Schritte zur Konstruktion des Modells wird ebenfalls durchgeführt. Ein Schlüsselziel ist es, festzustellen, ob wichtige Geschäftsprobleme nicht ausreichend berücksichtigt wurden. Am Ende dieser Phase sollte eine Entscheidung über die Verwendung der Data-Mining-Ergebnisse getroffen werden.

6. *Bereitstellung*: Nun muss das entdeckte Wissen organisiert und so präsentiert werden, dass der Kunde es nutzen kann. Abhängig von den Anforderungen kann dieser Schritt einfach sein, wie die Erstellung eines Berichts oder komplex, wie die Implementierung eines wiederholbaren Wissensentdeckungsprozesses.

   (Wirth und Hipp, 2000)

Das Modell zeichnet sich durch einen leicht verständlichen Wortschatz und einer guten Dokumentation aus, in dem alle Schritte in Unterabschnitte unterteilt werden und weitere Details bereitstellen (Cios et al., 2008). Zudem weist das Modell durch Schleifen zwischen mehreren der Schritte ebenfalls eine stark iterative Natur auf (Wirth und Hipp, 2000). Im Allgemeinen ist es ein sehr weit verbreitetes Modell, hauptsächlich aufgrund seiner Verankerung in praktischer, industrieller, realweltlicher Erfahrung der Wissensentdeckung (Cios et al., 2008). Zudem berücksichtigt das Modell industriell relevante Aspekte für den Projekterfolg, einschließlich solcher zeitlicher Natur. (Cios et al., 2008). Das CRISP-DM Modell wurde unter anderem in Bereichen wie Medizin, Ingenieurwesen, Marketing und Verkauf eingesetzt (Cios et al., 2008).

Vergleich von Vorgehensmodellen

Neben den genannten Vorgehnsmodellen existieren eine Vielzahl weiterer Modelle wie das nach Cabena et al. (1998), Anand et al. (1998), Cios et al. (2000) oder Haglin et al. (2005). Generell herrscht in der Wissensentdeckungsgemeinschaft der Konsens, dass kein universell bestes Modell existiert, da jedes Modell Stärken und Schwächen hat, basierend auf dem Anwendungsbereich und den spezifischen Zielen (Cios et al., 2008). Oft bestehen jedoch modellübergreifend Überschneidungen verschiedener Phasen, wenn auch anders benannt. Demnach beschreiben Kurgan und Musilek (2006) in einem Vergleich der beiden, in diesem Kapitel eingeführten, sowie drei weiteren, bekannten Vorgehensmodellen ein generisches Modell, dass diese Modelle in den Phasen *Domänenverständnis*, *Datenverständnis*, *Datenvorbereitung*, *Data Mining*, *Evaluation* sowie *Konsolidierung und Anwendung des Wissens* vereint. Wesentliche, vereinende Aktivitäten der Datenvorbereitung werden im folgenden Abschnitt erläutert.

2.2.2 Datenvorverarbeitung

Datenbanken sind in der Praxis oft durch negative Aspekte beeinflusst, wie das Auftreten von Rauschen, fehlenden Werten, irrelevanten Daten sowie signifikante Größe und Dimension (Cios et al., 2008). Diese Aspekte tragen dazu bei, dass Daten niedriger Qualität zu unzureichenden Ergebnissen im Data-Mining führen (Cios et al., 2008). Um hochwertige Ergebnisse durch Data-Mining erzielen zu können bedarf es deshalb einer Vorverarbeitung

der Daten (García et al., 2015). Trotz unterschiedlicher Vorgehensmodelle in der Wissensentdeckung in Datenbanken lassen sich generische Vorverarbeitungsaktivitäten zusammenfassen (García et al., 2015). Die wesentlichen Aktivitäten der Datenvorverarbeitung, die *Datenselektion und -integration, Datensäuberung, Datenreduktion* und die *Datentransformation*, werden folgend dargelegt (Cleve und Lämmel, 2020; Han et al., 2011; García et al., 2015).

Datenselektion und -integration

Die angemessene Selektion von Daten ist entscheidend für die Datenanalyse (García et al., 2015). Man bezeichnet diesen Vorgang als Datenselektion, bei dem Daten oft aus diversen Datensätzen oder Datenbanken stammen und zu einem einheitlichen Datensatz zusammengefasst werden (Cleve und Lämmel, 2020). Diese Aktivität kann bereits in den frühen Phasen Herausforderungen bergen (Doan et al., 2012). Ein konkretes Beispiel hierfür ist die Situation, in der jede Filiale eines Unternehmens ihre eigene Datenbank führt, was zu Inkonsistenzen führen kann, da die Datenbanken unterschiedliche Strukturen und Interpretationen von Attributen aufweisen können (Lenzerini, 2002). Die zentrale Aufgabe der Datenselektion besteht darin, multiple Datensätze so zu vereinen, dass am Ende eine konsistente und logisch stimmiger Datensatz entsteht (Rahm und Bernstein, 2001). García et al. (2015) und Cleve und Lämmel (2020) beschreiben in diesem Zusammenhang verschiedene Methoden um bei der Integration der Daten Redundanzen zu vermeiden.

Datenbereinigung

Die Datenbereinigung ist wichtiger Schritt in Wissensentdeckungsprozessen, in der die Qualität der Daten verbessert wird und somit zu verlässlicheren Ergebnissen im Data-Mining führt (García et al., 2015). Sie beinhaltet das Identifizieren und Beheben von fehlenden Werten, Rauschen sowie Ausreißern (García et al., 2015). Fehlende Werte können durch verschiedene Imputationsmethoden behandelt werden, die darauf abzielen, die Lücken in den Daten auf plausible Weise zu füllen (Garcia-Laencina et al., 2010). Ausreißer sind Datenpunkte, die sich signifikant von den restlichen Daten unterscheiden, sei es aufgrund von Messfehlern, Fehlern bei der Dateneingabe oder natürlicher Variabilität, wohingegen rauschende Daten Datenpunkte sind, die zufällige oder irrelevante Schwankungen enthalten, die dem Signal Rauschen hinzufügen (Cios et al., 2008). Die Behandlung von Ausreißern kann durch Methoden wie das Trimmen oder Winsorisieren erfolgen, bei denen extreme Werte entweder entfernt oder durch weniger extreme Werte ersetzt werden (Han et al., 2011). Rauschen in Daten kann durch Glättungsverfahren reduziert werden, wie zum Beispiel durch gleitende Durchschnitte oder Filtertechniken, die darauf abzielen, die Signalqualität zu verbessern, indem irrelevante Fluktuationen minimiert werden (Han et al., 2011).

Datenreduktion

Die Datenreduktion ist ein wichtiger Schritt im Prozess der Datenvorverarbeitung, der darauf abzielt, das Datenvolumen zu minimieren, während gleichzeitig so viel analytisch

relevante Informationen wie möglich erhalten bleiben (Han et al., 2011). Dieser Schritt ist von entscheidender Bedeutung, da er nicht nur die Effizienz von Data-Mining-Algorithmen verbessert sondern und in Situationen sehr großer Datensätze die Analyse erst ermöglicht, sondern auch dazu beiträgt die Interpretation der Ergebnisse des Data-Minings zu verbessern (Pyle, 1999b). Durch die Anwendung verschiedener Datenreduktionsverfahren werden Datensätze in kompaktere Repräsentationen überführt, ohne die Integrität der Daten zu beeinträchtigen (Pyle, 1999b). Die Datenreduktion stellt den Fokus der vorliegenden Arbeit dar und wird im darauffolgenden Teilkapitel genauer beleuchtet.

Datentransformation

Datentransformationen sind von entscheidender Bedeutung, um Data-Mining-Algorithmen ausführen zu können oder sinnvolle Ergebnisse aus ihnen zu erzielen. Wie von Han et al. (2011) betont wird, müssen Rohdaten oft in eine geeignete Form gebracht werden, bevor sie von den Algorithmen verarbeitet werden können. Dies beinhaltet bespielsweise die Transformation kontinuierlicher Daten, die zunächst diskretisiert werden müssen, um von Algorithmen wie Entscheidungsbäumen verarbeitet werden zu können (García et al., 2015). Oft müssen zur Gewährleistung gültiger Data-Mining-Ergbnisse zudem Daten normalisiert oder skaliert werden Han et al. (2011). Während die Normalisierung darauf abzielt, Daten auf einen bestimmten Bereich zu bringen oder eine spezifische Verteilung zu erzielen, zielt die Skalierung darauf ab, die Skalen der Datenattribute anzugleichen, um Verzerrungen aufgrund von Unterschieden in den Einheiten oder Größenordnungen zu vermeiden Han et al. (2011).

Alle diskutierten Vorverarbeitungstechniken haben das gemeinsame Ziel, die Analyse zu ermöglichen und Ergebnisse des Data-Minings gültig und verständlich zu gestalten (García et al., 2015). Im Folgenden werden grundlegende Konzepte des Data-Minings erläutert bevor schließlich die Datenreduktion in den Fokus rückt.

2.2.3 Data Mining

Das Data-Mining stellt die zentrale Phase von Wissensentdeckungsprozessen dar, bei dem durch den Einsatz diverser Algorithmen verborgene Strukturen und Muster in Datensätzen identifiziert werden (Han et al., 2011). In der Fachliteratur findet sich eine umfangreiche Dokumentation verschiedener Data-Mining-Algorithmen, die in zahlreichen Anwendungsfällen zur Anwendung kommen (Witten et al., 2016). Diese Algorithmen erfüllen unterschiedliche Funktionen, je nachdem, welche spezifischen Aufgaben sie lösen sollen (Fayyad et al., 1996b). Diese werden in der Literatur meist in *deskriptive* und *vorhersagende* Aufgaben eingeteilt (García et al., 2015). Oft erfolgt die Einteilung jedoch auch in die Paradigmen des *überwachten-* und des *unüberwachten Lernens*, basierend auf der Existenz von Labelinformationen, wie in 2.1.2 bereits dargelegt. (Bishop, 2006). Abbildung 2.4 klassifiziert die gängigsten Data-Mining-Aufgaben gemäß diesen Lernparadigmen, worauf im Folgenden kurz eingegangen wird (Alpaydin, 2020).

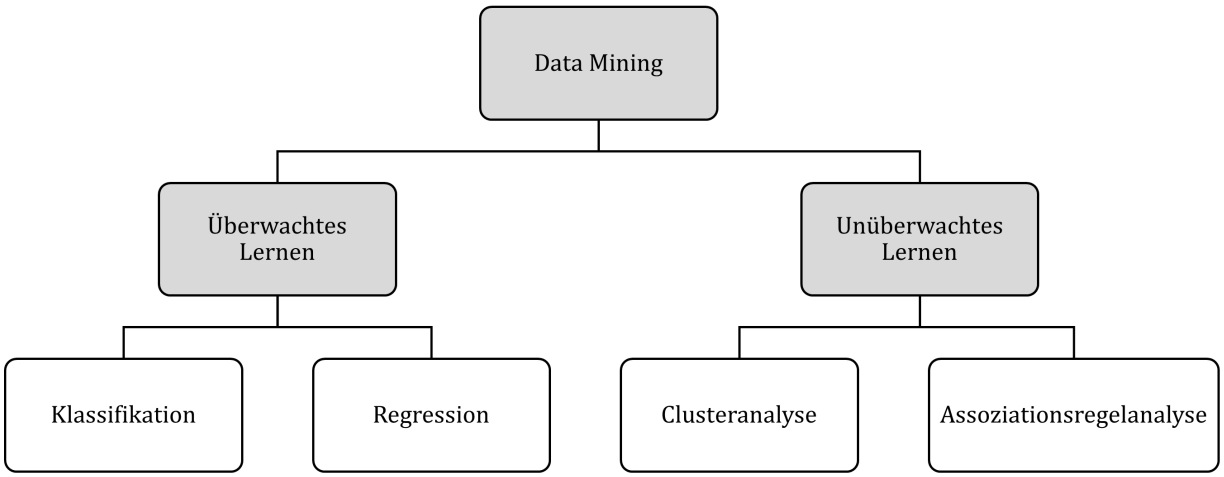


Abbildung 2.4: Einteilung gängiger Data-Mining-Aufgaben

Überwachtes Lernen

Das überwachte Lernen hat das Ziel, die fundamentalen Zusammenhänge zwischen Eingabevariablen (Inputs) und einem Zielmerkmal (Label, Output) in einem Datensatz zu identifizieren (Bishop, 2006). Diese Beziehungen werden durch ein Modell repräsentiert, das dazu dient, die in den Daten enthaltenen Muster zu charakterisieren und zu erklären (Alpaydin, 2020). Ein solches Modell ermöglicht die Prognose des Zielmerkmals auf Basis neuer Eingabedaten, indem es auf diesen gelernten Beziehungen fußt (Hastie et al., 2009). Deshalb werden die Aufgaben des supervised Learnings auch den prädiktiven beziehungsweise vorhersagenden Aufgaben zugeordnet (Witten et al., 2016). Gängige Aufgaben des überwachten Lernens umfassen die Klassifikation und Regression (Murphy, 2012).

Klassifikation Bei der Klassifikation sind die Werte des Zielattributs endlich und kategorisch, was bedeutet, dass es eine beschränkte Anzahl an Klassen oder Kategorien gibt, denen eine Instanz zugeordnet werden kann (Witten et al., 2016). Ein Klassifikationsmodell wird daraufhin trainiert, einer unbekannten Instanz eine dieser vordefinierten Klassen basierend auf einem Satz von Trainingsdaten zuzuweisen (Bishop, 2006). Das Wesen der Klassifikation manifestiert sich in der Fähigkeit, Instanzen anhand ihrer Attribute zu unterscheiden und auf dieser Grundlage verlässliche Vorhersagen zu treffen (Murphy, 2012).

Regression Wenn das Zielattribut in einer numerischen und kontinuierlichen Form vorliegt, wird dies als eine Regressionsaufgabe definiert (Witten et al., 2016). Das Ziel dabei ist, eine Funktion zu entwickeln, die die Beziehung zwischen Eingabevariablen und dem Zielattribut möglichst präzise abbildet, um daraufhin Werte des Zielattributs für neue, unbekannte Datensätze vorhersagen zu können (Hastie et al., 2009).

Unüberwachtes Lernen

Im unüberwachten Lernen stehen ausschließlich Eingabedaten ohne zugehörige Ausgabe in Form eines Labels zur Verfügung (Bishop, 2006). Das Ziel hierbei ist die Erkennung von Mustern, Anomalien, Beziehungen und Ähnlichkeiten innerhalb der Daten (Alpaydin, 2020). Daher werden die Aufgaben des unüberwachten Lernens den deskriptiven Aufgaben zugeordnet (Hastie et al., 2009). Zu den gängigen Aufgaben in diesem Bereich zählen das Clustering und die Assoziationsregelanalyse. Clustering kann insbesondere auch dazu verwendet werden, Labels für einen Datensatz zu generieren, die dann zur Entwicklung eines prädiktiven Modells im Rahmen des überwachten Lernens herangezogen werden können (Murphy, 2012).

Clustering Das Ziel des Clusterings besteht darin, Instanzen in Gruppen, sogenannte Cluster, zu unterteilen, wobei die Elemente innerhalb eines Clusters ähnlicher zueinander sind als zu Elementen in anderen Clustern (Jain et al., 1999). Diese Methoden stützen sich beispielsweise auf die Berechnung von Distanzen oder Ähnlichkeitsmaßen zwischen den Datenpunkten, um Untergruppen mit gemeinsamen Merkmalen zu erkennen, können jedoch auch auf der Dichte von Datenpunkten oder anderen Konzepten beruhen (Xu und Wunsch, 2015).

Assoziationsregeln Während das Clustering darauf abzielt, Instanzen auf Grundlage ihrer Ähnlichkeit in Gruppen einzuteilen, fokussiert sich die Analyse von Assoziationsregeln auf die Identifizierung von Abhängigkeiten zwischen Ereignissen, um Regeln zu formulieren, die die Beeinflussung des Auftretens bestimmter Ereignisse durch andere beschreiben (Agrawal et al., 1993). Ein prominentes Beispiel hierfür ist die Untersuchung von Einkaufskörben, um Produkte zu ermitteln, die häufig zusammen gekauft werden (Hipp et al., 2000).

Weitere Paradigmen und Aufgaben des Data-Minings

Neben den gängigen Paradigmen des überwachten und unüberwachten Lernens existieren weitere Data-Mining-Paradigmen und -Aufgaben, die sich nicht eindeutig diesen Kategorien zuordnen lassen (Witten et al., 2016). Das *semi-überwachte* Lernen kombiniert Elemente aus dem überwachten und unüberwachten Lernens, indem es Modelle mit einer Mischung aus gelabelten und nicht gelabelten Daten trainiert wird, um die Modellgenauigkeit in Situationen zu verbessern, in denen gelabelte Daten begrenzt sind (Zhu und Goldberg, 2009). *Bestärkendes Lernen* unterscheidet sich grundlegend von überwachtem und unüberwachtem Lernen, da es auf der Interaktion mit einer Umgebung basiert, um durch Belohnungen oder Strafen zu lernen, welche Aktionen in verschiedenen Situationen am besten sind (Sutton und Barto, 2018).

Nachdem der Kontext dieser Arbeit, die Wissensentdeckung und Datenbanken, eingeführt wurde und Differenzierungen hinsichtlich der Datenvorverarbeitung sowie unterschiedlicher Data-Mining-Paradigmen und -Aufgaben vorgenommen wurden, die für das Verständnis und als Referenz für spätere Ausführungen dienen, folgt nun eine Einführung in den Fokus dieser Arbeit - der Datenreduktion.

2.3 Methoden der Datenreduktion

Nachdem in den vorherigen Abschnitten eine Einführung in die Wissensentdeckung in Datenbanken und speziell die Datenvorverarbeitung sowie des Data-Minings gegeben wurde, rückt nun der Fokus auf die Datenreduktion. In diesem Zusammenhang werden zunächst die Ziele, generelle Vor- und Nachteile, sowie gängige Kategorisierungen von Methoden der Datenreduktion erörtert und repräsentative Datenreduktionsverfahren vorgestellt.

Ziele sowie allgemeine Vor- und Nachteile der Datenreduktion

Das kontinuierliche Anwachsen der Datenmengen stellt eine Herausforderung für die Analyse der Daten dar (García et al., 2015). Außerdem verursacht eine hohe Anzahl an Merkmalen in großen Datensätzen den Fluch der Dimensionalität, welcher bedeutende Rechenressourcen erfordert, um nutzbare Muster zu identifizieren (García et al., 2015). Die Datenreduktion ist ein entscheidender Schritt in Wissensentdeckungsprozessen, da sie dazu beiträgt, die Effizienz der Datenanalyse erheblich zu verbessern (Han et al., 2011). Durch die Verringerung der Datengröße können Algorithmen zur Datenverarbeitung und -analyse schneller ausgeführt werden, was die Gesamteffizienz des Data-Mining-Prozesses steigert (Leskovec et al., 2020). Zudem ermöglichen Datenreduktionsverfahren erhebliche Kosteneinsparungen, indem sie den Bedarf an Speicherplatz und Rechenleistung reduzieren (Leskovec et al., 2020). Eine weitere wichtige Rolle der Datenreduktion liegt in der Verbesserung der Datenqualität, indem Rauschen und irrelevante Informationen entfernt werden, was zu präziseren Analyseergebnissen führt (Han et al., 2011). Darüber hinaus erleichtert die Reduzierung der Datenmenge die Visualisierung und Interpretation der Daten, wodurch komplexe Datensätze zugänglicher und verständlicher werden (Han et al., 2011).

Trotz dieser Vorteile besteht der Hauptnachteil von Datenreduktionsverfahren im potenziellen Verlust wichtiger Informationen, was die Ergebnisse der Datenanalyse beeinträchtigen kann (Leskovec et al., 2020). Die Komplexität einiger Reduktionsmethoden kann ebenfalls eine Herausforderung darstellen, da ein tiefes Verständnis der Daten und der eingesetzten Techniken erforderlich ist (Han et al., 2011). Ein weiteres Risiko ist die Übervereinfachung der Daten, die zu Fehlinterpretationen und fehlerhaften Entscheidungen führen kann (Han et al., 2011). Schließlich können Datenreduktionsmethoden auch eigene Verzerrungen einführen und so die Analyseergebnisse beeinflussen (Leskovec et al., 2020).

Kategorisierungen zu Methoden der Datenreduktion

In der existierenden Fachliteratur zum Data Mining haben sich vielfältige Ansätze zur Systematisierung von Methoden und Strategien der Datenreduktion etabliert. Im folgenden Abschnitt werden einige dieser gebräuchlichen Systematisierungsansätze vorgestellt, um die verschiedenen Methoden anschließend in einer strukturierten Form einführen zu können.

Han et al. (2011) klassifizieren Methoden der Datenreduktion in zwei Hauptkategorien, der Dimensionsreduktion und Numerositätsreduktion. Zur Dimensionsreduktion zählen sie sämtliche Aktivitäten zur Reduzierung der Anzahl der Merkmale eines Datensatzes.

Hierzu nennen sie Methoden der Merkmalsauswahl sowie Kompressionsmethoden durch Projektionen des Merkmalsraums auf einen Merkmalsraums geringerer Dimension. Die Numerositätsreduktion hingegen erklären sie als Strategie der Datenreduktion, in der die Anzahl der Instanzen reduziert wird. Hierfür nennen unterscheiden sie parametrische Verfahren und nichtparametrischer Modelle. Beide Ansätze zielen darauf ab, eine reduzierte Repräsentation des Datensatzes zu erhalten, die zwar deutlich geringer im Umfang ist, aber dennoch dieselben oder nahezu dieselben analytischen Ergebnisse liefert.

Aggarwal (2015) folgen einer Klassifizierung der Datenreduktionsmethoden, die mit der von Han et al. (2011) vergleichbar ist, indem sie diese in die beiden Hauptkategorien der Verringerung der Anzahl von Instanzen und der Reduktion der Dimensionen einteilen. Darüber hinaus differenzieren sie diese Hauptkategorien weiter in drei spezifische Methodiken, von denen zwei der Merkmalsreduktion zugeordnet werden, während eine Methode der Reduktion der Anzahl von Instanzen dient. Die erste Kategorie, Datenstichprobierung, beinhaltet Verfahren, bei denen aus einem umfangreichen Datensatz eine repräsentative Stichprobe extrahiert wird, um eine verkleinerte Datenmenge zu erzeugen. Die zweite Kategorie, Merkmalsauswahl, konzentriert sich auf die Selektion einer relevanten Teilmenge von Merkmalen aus den Gesamtdaten, die für spezifische analytische Zwecke geeignet sind. Als dritte Kategorie wird die Datenreduktion durch Dimensionsreduktion, ähnlich wie bei Han et al. (2011) vorgestellt.

García et al. (2015) erkennen die eben vorgestellten, allgemein anerkannte und weit verbreitete Konvention an, Methoden der Datenreduktion in die Kategorien der Instanzenreduktion und der Merkmalsreduktion einzuteilen. Sie erweitern diese jedoch um eine Kategorie, indem sie Methoden, die oft dem Bereich der Datentransformation zugeordnet sind, wie Diskretisierung und Binning, zur Datenreduktion hinzuzählen. Dies begründen sie damit, dass sie alle Methoden, die die Datenanalyse durch zeitliche Effizienzsteigerung verbessern, zur Datenreduktion zählen. Sie differenzieren dabei zwischen der zweckmäßigen Transformation von kontinuierlichen in kategorische Werte zur Anforderungserfüllung eines Data Mining-Algorithmus und der optimalen Diskretisierung von Daten zur Minimierung der Analysezeit. Letztere ordnen sie der Datenreduktion zu, da sie eine direkte Auswirkung auf die Beschleunigung des Analyseprozesses hat. Darüber hinaus unterteilen sie jede der drei Strategien weiter in einfache und fortgeschrittene Methoden. Beispielsweise wird das Sampling von Datensätzen als eine einfache Methode zur Reduktion der Instanzen angeführt, während Verfahren der Instanzenauswahl als fortgeschritten betrachtet werden, da sie die Auswahl der Instanzen begründet durch die Analyse der Daten erfolgt.

Cleve und Lämmel (2020) folgen ebenfalls den beiden eingangs vorgestellten Einteilungen und gliedern die Methoden der Datenreduktion in zwei Hauptkategorien der Reduktion der Datenmenge und der Reduktion der Dimensionalität. In ihrer Arbeit beschreiben sie Methoden, die diesen beiden Kategorien zugeordnet werden können, und erörtern diese.

Im Rahmen dieser Arbeit wird die von Aggarwal (2015), Han et al. (2011) und Cleve und Lämmel (2020) beschrieben, in der wissenschaftlichen Literatur zum Data Mining weit verbreitete Ansicht der Datenreduktion adaptiert, welche davon ausgeht, dass die Datenreduktion sämtliche Methoden einschließt, die auf eine Verkleinerung der Größe eines Datensatzes in Bezug auf die Instanzen oder Merkmale abzielen. Die Diskretisierung sowie verwandte Aktivitäten werden in der vorliegenden Arbeit, abweichend von den Darlegungen von García et al. (2015), nicht als Methoden der Datenreduktion, sondern explizit als Verfahren der Transformation gemäß dem KDD-Prozess oder als Elemente der

Merkmalskonstruktion im Rahmen der Datenvorbereitung des CRISP-DM-Modells klassifiziert. Folgend werden Methoden der Merkmalsreduktion sowie der Instanzenreduktion dargelegt.

2.3.1 Merkmalsreduktion

Daten mit hoher Dimensionalität können die Anforderungen an den Speicherplatz und die Rechenkosten für Datenanalysen erheblich erhöhen (Lee und Kang, 2018). Die Merkmlasreduktion ist eine der wirkungsvollsten Strategien, um die zuvor beschriebenen Probleme anzugehen (Maaten, 2014). Sie lässt sich hauptsächlich in zwei Methoden unterteilen, der *Dimensionsreduktion* und der *Merkmalsauswahl* (Rodriguez und Gutierrez, 2019). Die Dimensionsreduktion projiziert die ursprünglichen hochdimensionalen Merkmale auf einen neuen Merkmalsraum mit geringer Dimensionalität (De Backer et al., 1998). Der neu konstruierte Merkmalsraum ist dabei üblicherweise eine lineare oder nichtlineare Kombination der ursprünglichen Merkmale (Kim und Choi, 2021). Die Merkmalsauswahl hingegen wählt direkt eine Teilmenge relevanter Merkmale aus (Liu und Wang, 2022).

Sowohl die Merkmalsauswahl und die Dimensionsreduktion tragen maßgeblich zur Steigerung der Lernleistung, zur Effizienzsteigerung bei der Berechnung, zur Reduktion des Speicherplatzbedarfs und zur Entwicklung robusterer Modelle bei (Guyon und Elisseeff, 2003). Beide Ansätze sind demnach wirkungsvolle Methoden, um die Komplexität von Datensätzen durch die Reduktion der Merkmale zu verringern (Maaten und Hinton, 2008). Insbesondere wird die Dimensionsreduktion favorisiert, wenn die Ausgangsdaten keine direkt interpretierbaren Eigenschaften für spezifische Lernalgorithmen aufweisen, da die Dimensionsreduktion, durch die Generierung eines neuen Attributsets, die direkte Analyse erschweren kann, da die originale semantische Bedeutung der Attribute verloren geht (Bengio et al., 2013). Im Gegensatz dazu behält die Merkmalsauswahl bestimmte originale Attribute bei und konserviert somit deren semantische Integrität, was die Modelle leichter lesbar und interpretierbar macht (Guyon und Elisseeff, 2003). Diese Eigenschaft macht die Merkmalsauswahl besonders attraktiv für Einsatzgebiete in denen der erhalt der ursprünglichen Merkmale wichtig ist (Mladenic und Grobelnik, 1999). Es ist hervorzuheben, dass die Dimensionsreduktion und Merkmalsauswahl auch in Szenarien von Bedeutung ist, in denen die Merkmalsdimensionen nicht extrem hoch sind, indem sie beispielsweise die Lernleistung verbessert, Überanpassung verhindert und Rechenkosten minimiert (Maaten und Hinton, 2008).

In der wissenschaftlichen Literatur wird die Dimensionsreduktion zum Teil äquivalent zur *Merkmalsextraktion* angesehen wird (Jolliffe, 2006; Guyon und Elisseeff, 2003). Die Äquivalentssetzung von Dimensionsreduktion und Merkmalsextraktion kann in der wissenschaftlichen Diskussion konfliktreich sein, da beide Prozesse unterschiedliche Zielsetzungen verfolgen, obwohl sie eng miteinander verbunden sind (Aggarwal, 2015). Die Dimensionsreduktion, als eine Form der Datenreduktion, zielt darauf ab, die Menge an Daten zu verkleinern, indem sie die wesentlichen Informationen beibehält. Dieser Prozess ist insbesondere in der Handhabung von umfangreichen Datensätzen nützlich und steigert die Effizienz von Algorithmen (Aggarwal, 2015; Han et al., 2011). Im Gegensatz dazu ist die Merkmalsextraktion ein entscheidender Schritt, um Rohdaten in eine analysierbare Form zu überführen, insbesondere im Kontext der Datentyp-Portabilität (Kapitel 2.2.2) (Aggarwal, 2015; García et al., 2015). Dies beinhaltet die Umwandlung von Daten in ein neues

Set von Merkmalen, das für spezifische analytische Aufgaben relevant ist (Aggarwal, 2015; García et al., 2015).

Die Gleichsetzung wird im weiteren Verlauf der Arbeit nicht verfolgt, da sie unterschiedliche Aspekte der Datenverarbeitung betreffen. Während die Dimensionsreduktion auf die Verringerung der Datenmenge abzielt, konzentriert sich die Merkmalsextraktion auf die Umwandlung und Anreicherung der Daten für spezifische Analysezwecke.

Dimensionsreduktion

Die Dimensionsreduktion ist die Transformation von Daten hoher Dimensionalität in eine aussagekräftige Repräsentation reduzierter Dimensionalität (Aggarwal, 2015). Die Reduktion der Daten-Dimensionalität vereinfacht nicht nur die Verarbeitung und Analyse der Daten, sondern verbessert auch ihre Übersichtlichkeit und Handhabung (Esling und Agon, 2012). Dieser Ansatz ermöglicht die Transformation komplexer Datensätze in einfachere Formate, die leichter zu verstehen, zu speichern und zu verarbeiten sind, ohne wichtige Informationen zu verlieren (Van Der Maaten et al., 2009). Idealerweise sollte die reduzierte Repräsentation eine Dimensionalität aufweisen, die der intrinsischen Dimensionalität der Daten entspricht (Jolliffe, 2006). Die intrinsische Dimensionalität von Daten ist die minimale Anzahl von Parametern, die benötigt wird, um die beobachteten Eigenschaften der Daten zu erklären (Lee und Verleysen, 2007). Die Dimensionsreduktion ist in vielen Bereichen wichtig, da sie den Fluch der Dimensionalität und andere unerwünschte Eigenschaften von hochdimensionalen Räumen mildert (Van Der Maaten et al., 2009). Als Ergebnis verbessert die Dimensionsreduktion zum Teil die Klassifizierung und ermöglicht Visualisierung und Kompression von hochdimensionalen Daten (Maaten und Hinton, 2008).

In der Literatur werden diese Methodiken oft weiter in lineare und nicht-lineare Methoden unterteilt (Abdi und Williams, 2010). Lineare Methoden zielen darauf ab, hochdimensionale Daten auf eine niedrigere Dimension zu projizieren, während gleichzeitig versucht wird, so viel wie möglich von der ursprünglichen Datenvarianz zu bewahren (Jolliffe, 2006). Eine der bekanntesten Techniken in diesem Bereich ist die Hauptkomponentenanalyse (engl. Principle Component Analyse, PCA), die eine orthogonale Transformation durchführt, um die Daten auf eine Reihe von linear unabhängigen Dimensionen (Hauptkomponenten) zu projizieren, die die meiste Varianz der Daten erklären (Abdi und Williams, 2010). Ein weiterer Ansatz ist die Lineare Diskriminanzanalyse (LDA), die darauf abzielt, Datenpunkte so zu projizieren, dass die Trennbarkeit zwischen verschiedenen Klassen maximiert wird, indem die Varianz zwischen den Klassen maximiert und die Varianz innerhalb der Klassen minimiert wird (Mika et al., 1999). Trotz der Reduktion der Daten kann mittles der genannten Verfahren zum Teil die Klassifizierungsleistung verbessert werden (Martínez und Kak, 2001). Beide Methoden, PCA und LDA, bieten robuste Werkzeuge für die lineare Dimensionsreduktion, die es ermöglichen, komplexe Datensätze effizienter zu analysieren und zu interpretieren (Fukunaga, 1990).

Linearen Methoden können komplexe, nicht-lineare Daten nicht angemessen reduzieren, ohne wesentlichen Informationsverlust (Hinton und Roweis, 2002). Daher wurden im Laufe der zeit einige nicht-lineare Dimensionreduktionsverfahren entwickelt (Lee und Verleysen, 2007). Sie spielen eine entscheidende Rolle bei der Analyse und Verarbeitung von Daten,

die in komplexen, hochdimensionalen Räumen liegen, in denen lineare Methoden nicht ausreichend sind, um die inhärente Struktur der Daten vollständig zu erfassen (Lee und Verleysen, 2007). Eine populäre Technik in diesem Bereich ist der Kernel-Trick, der bei Methoden wie der Kernel-PCA (KPCA) zum Einsatz kommt und es ermöglicht, lineare Algorithmen in einem hochdimensionalen Merkmalsraum anzuwenden, indem die ursprünglichen Daten implizit in einen höherdimensionalen Raum projiziert werden (Schölkopf et al., 1998). KPCA hat sich als effektiv erwiesen, um nichtlineare Muster in Daten zu erhalten und wird in einer Vielzahl von Anwendungen eingesetzt (Hofmann et al., 2008). Ein weiteres bedeutendes Verfahren ist t-Distributed Stochastic Neighbor Embedding (t-SNE), das darauf abzielt, hochdimensionale Datenpunkte auf zwei- oder dreidimensionale Punkte zu reduzieren (Maaten, 2014). Dabei werden ähnliche Objekte in der niedrigdimensionalen Darstellung nahe beieinander und unähnliche Objekte weiter voneinander entfernt platziert (Maaten und Hinton, 2008). t-SNE hat sich als besonders nützlich erwiesen, um Einsichten in die Datenstruktur zu gewinnen und Muster zu identifizieren, die in den ursprünglichen hochdimensionalen Daten verborgen sind und weißt somit neben der Datenreduktion einen stark explorativen Charakter auf (Maaten, 2014). Es sei hierbei angemerkt, dass durch die beschriebenen Eigenschaften verschiedener Dimensionsreduktionsarten neben der eigentlichen Datenreduktion oft weitere Ziele zur Datenexploration beziehungsweise der Visualisierung im Fokus stehen, wobei der Fokus im Rahmen der Arbeit auf der Datenreduktion liegt.

Merkmalsauswahl

Im Gegensatz zur Dimensionsreduktion, bei der die Anzahl der Merkmale durch Transformation verringert wird, zielt die Merkmalsauswahl darauf ab, die Datensätze durch das Entfernen redundanter oder für die Analyse irrelevanter Merkmale zu vereinfachen (Aggarwal, 2015). Die Merkmalsauswahl ist eine intensiv erforschte Methode in der Datenverarbeitung (Chandrashekar und Sahin, 2014). In der Fachliteratur wird berichtet, dass in den letzten Jahren über 100 neue Verfahren für die Merkmalsauswahl entwickelt wurden (Bolón-Canedo und Remeseiro, 2019). Diese Verfahren lassen sich üblicherweise in drei Hauptkategorien einteilen, Filter-, Wrapper- und Embedded-Methoden (Bolón-Canedo und Remeseiro, 2019). Jede dieser Kategorien nutzt unterschiedliche Strategien zur Auswahl der Merkmale (Chandrashekar und Sahin, 2014), die im Folgenden näher vorgestellt werden.

Wrapper-Verfahren evaluieren die Auswahl an Merkmalen auf Basis der Vorhersagekraft eines bestimmten Lernalgorithmus (Kohavi und John, 1997). Innerhalb eines vorgegebenen Lernalgorithmus führen Wrapper-Methoden in der Regel zwei Hauptoperationen aus: die Identifizierung einer Merkmalsuntergruppe und deren Evaluierung (Guyon und Elisseeff, 2003). Diese Prozedur wird fortgesetzt, indem sie zwischen Merkmalsauswahl und Leistungsbewertung wechselt, bis ein Abbruchkriterium erreicht wird (Guyon und Elisseeff, 2003). Die Auswahl an Merkmalen wird anfänglich erzeugt und anschließend durch den Lernalgorithmus, der als Blackbox fungiert, auf Basis ihrer Effektivität beurteilt (Masaeli et al., 2010). Dieser iterative Prozess wird solange fortgeführt, bis entweder die optimale Leistungsfähigkeit erzielt oder eine festgelegte Anzahl an Merkmalen ausgewählt wurde, wobei am Ende die leistungsfähigste Merkmalskombination bestimmt wird (Masaeli et al., 2010). Ein wesentliches Problem der Wrapper-Methoden ist jedoch die exponentielle Größe des Suchraums bei einer hohen Anzahl an Merkmalen, was sie für große Merkmalsräume

praktisch unanwendbar macht (Bolón-Canedo und Remeseiro, 2019). Daher werden in der Literatur verschiedene Suchstrategien wie die sequentielle Suche, die Hill-Climbing-Suche, die Best-First-Suche, die Branch-and-Bound-Suche oder genetische Algorithmen vorgeschlagen, um eine lokale optimale Lernleistung zu erzielen (Guyon und Elisseeff, 2003). Jedoch ist der Suchraum für hochdimensionale Datensätze immer noch extrem groß weshalb Wrapper-Methoden selten in der Praxis zu großen Datensätzen verwendet (Masaeli et al., 2010).

Filter-Verfahren sind unabhängig von Lernalgorithmen (Guyon und Elisseeff, 2003). Sie nutzen verschiedene Metriken, um die Wichtigkeit von Merkmalen zu bewerten (Yu und Liu, 2003). Filter-Verfahren sind typischerweise effizienter als Wrapper-Verfahren (Masaeli et al., 2010). Jedoch kann aufgrund des Fehlens eines spezifischen Lernalgorithmus, der die Merkmalsauswahlphase leitet, die Auswahl der Merkmale nicht optimal für den Ziel-Lernalgorithmen sein (Guyon und Elisseeff, 2003). Ein typisches Filter-Verfahren besteht aus zwei Schritten (Peng et al., 2005). Im ersten Schritt wird die Wichtigkeit der Merkmale entsprechend verschiedener Bewertungskriterien für Merkmale eingestuft (Bolón-Canedo und Remeseiro, 2019). Der Bewertungsprozess der Merkmalswichtigkeit kann entweder univariat oder multivariat sein (Robnik-Šikonja und Kononenko, 2003). Im univariaten Schema wird jedes Merkmal individuell eingestuft, unabhängig von anderen Merkmalen, während das multivariate Schema mehrere Merkmale auf einmal einstuft (He et al., 2005). Im zweiten Schritt eines typischen Filter-Verfahren werden niedrig eingestufte Merkmale herausgefiltert (Masaeli et al., 2010). In den vergangenen Jahrzehnten wurden verschiedene Evaluationskriterien für Filtermethoden vorgeschlagen und unterschiedlichen Kontexten angewendet (Nguyen et al., 2014).

Embedded-Verfahren integrieren die Merkmalsauswahl direkt in den Algorithmus, was es ermöglicht, sowohl die Vorzüge von Wrapper- als auch von Filtermethoden zu nutzen, ohne die Nachteile eines separaten Auswahlprozesses zu haben (Gao et al., 2016). Das Ergebnis sind sowohl das trainierte Modell als auch die ausgewählten Merkmale, die gemeinsam zurückgegeben werden (Jiang und Ren, 2011).

Abbildung 2.5 veranschaulicht die vorgestellten Methoden zur Merkmalsauwahl schematisch. Filter-Verfahren beginnen mit der Generierung von Merkmalsuntergruppen, die unabhängig von den Lernalgorithmen ausgewertet und für das Lernen verwendet werden, gefolgt von einer Modellbewertung. Wrapper-Verfahren integrieren die Merkmalsauswahl als Teil des Lernprozesses, wobei die Auswahl auf der Leistung des Lernmodells basiert, das wiederum bewertet wird. Embedded-Verfahren kombinieren Merkmalsauswahl und Lernen in einem Schritt und führen direkt zur Modellbewertung, wobei die Auswahl innerhalb des Lernalgorithmus eingebettet ist.

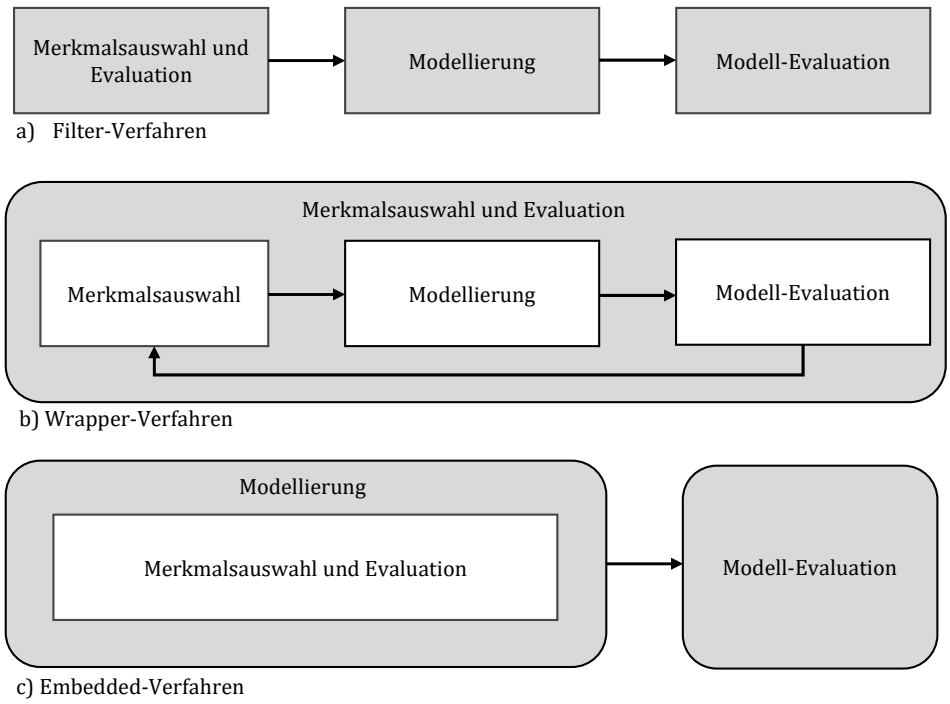


Abbildung 2.5: Schematische Darstellung von a) Filter, b) Wrapper und c) Embedded-Verfahren zur Merkmalsauswahl

Es sei zudem angemerkt, dass in der Literatur die Verfahren der Merkmalsauswahl oft in vier Kategorien klassifiziert werden, indem sie die hybriden Merkmalsauswahlsverfahren einschließen (Bolón-Canedo und Remeseiro, 2019). Hybride Verfahren können als eine Kombination mehrerer Merkmalsauswahlalgorithmen (zum Beispiel Wrapper, Filter und Embedded) angesehen werden (Yang et al., 2011). Durch die Aggregation mehrerer ausgewählter Merkmalssets aus verschiedenen Verfahren soll dadurch die Glaubwürdigkeit der ausgewählten Merkmale verbessert werden (Tang et al., 2014).

2.3.2 Instanzenreduktion

Die Instanzenreduktion, auch Reduktion der Datenmenge oder numerische Datenreduktion genannt (García et al., 2015; Aggarwal, 2015; Han et al., 2011; Cleve und Lämmel, 2020), ist ein wesentlicher Bestandteil innerhalb der Datenreduktion, der darauf abzielt, die Menge der Instanzen in einem Datensatz zu reduzieren, ohne dabei signifikante Informationen zu verlieren (Wilson et al., 2000). Durch die Verringerung der Anzahl der Instanzen können Algorithmen schneller ausgeführt werden, da weniger Daten verarbeitet werden müssen, was wiederum die Analysezeit verkürzt und Ressourcen spart (Fayyad et al., 1996b). In der Literatur findet man hierfür verschiedene Herangehensweisen und Methoden, die im Folgenden dargelegt werden.

Sampling

Samplingverfahren wurden ursprünglich für Umfragen in menschlichen Populationen entwickelt, um aus einer anfänglichen Fragestellung, die Messungen von Variablen in der Population erfordert, näherungsweise Antworten zu generieren (Cochran, 1977). Dies erfolgt durch die Entnahme von Stichproben aus der Population, die Messung interessierender Variablen in diesen Stichproben und die anschließende Extrapolation der Ergebnisse von den Stichproben auf die gesamte Population (Cochran, 1977; Lohr, 2009). Im Laufe der Zeit fanden Samplingverfahren auch in der Domäne des Data Minings Anwendung in der sie vorrangig der Datenreduktion und Analyse dienen (Han et al., 2011). Anstatt den gesamten Datensatz zu untersuchen, werden durch Samplingverfahren Teilmengen extrahiert und analysiert (Han et al., 2011).

Samplingverfahren werden meist in wahrscheinlichkeitsbasierte und nicht wahrscheinlichkeitsbasierte Verfahren eingeteilt (Lohr, 2009; Tyrer und Heyman, 2016). Wahrscheinlichkeitsbasierte Samplingverfahren integrieren einen Aspekt der zufälligen Auswahl, um sicherzustellen, dass jedes Element in der Population die gleiche Chance hat, ausgewählt zu werden (Thompson, 2012). Zu den gängigen Typen von Wahrscheinlichkeitsverfahren gehören Zufallsstichproben, systematische Stichproben, geschichtete Stichproben und Cluster-Stichproben (Särndal et al., 2003). Nicht wahrscheinlichkeitsbasierte Samplingverfahren nutzen einen Ansatz, bei dem die Stichprobe aufgrund des subjektiven Urteils des Forschers und nicht durch Zufallsauswahl bestimmt wird (Bethlehem, 2009). Zu den gängigen Typen von nicht wahrscheinlichkeitsbasierte Samplingverfahren gehören Quotenstichproben, zweckgerichtete Stichproben, Selbstselektionsstichproben und Schneeballstichproben (Särndal et al., 2003).

Ergänzend sei an dieser Stelle angemerkt, dass zufällige Stichproben auch bei der Modellierung im Data Minnig eine wesentliche Rolle spielen, mit Hilfe dessen Stichprobenstatistiken ermittelt werden, um die Qualität von Vorhersagemodellen zu beurteilen und statistische Modelle durch verschiedenen Algorithmen zu testen, wie beispielsweise dem Bootstrap-Verfahren (Efron und Tibshirani, 1994).

Instanzenauswahl

Die Instanzenauswahl stellt die komplementäre Methode zur Merkmalsauswahl dar und wird ähnlich wie diese in Filter- und Wrapper-Verfahren unterteilt (García et al., 2012). Durch die Auswahl repräsentativer Datenpunkte aus einem umfangreichen Datensatz kann die Analyse der Daten so effizienter gestaltet werden (Liu und Motoda, 2002). Filtermethoden konzentrieren sich primär auf die Verringerung des Datenvolumens, ohne dabei spezielle Data-Mining-Verfahren in Betracht zu ziehen. Diese Bewertungsansätze kennzeichnen sich durch Kriterien, die auf die statistische Repräsentativität abzielen, wie etwa die Sicherstellung, dass die Zusammensetzung der Stichprobe den Modalwert der ursprünglichen Datenbasis annähernd widerspiegelt (Reinartz, 2002). Im Gegensatz dazu setzen Wrapper-Verfahren einen deutlichen Fokus auf den Data-Mining-Prozess selbst und beurteilen die Ergebnisse auf Basis des Einsatzes eines spezifischen Data-Mining-Algorithmus, um die Instanzenauswahl voranzutreiben (Reinartz, 2002). Ein beispielhaftes Verfahren innerhalb der Wrapper-Verfahren beginnt mit der Auswahl einer anfänglichen Teilmenge von Daten, auf die dann der Data-Mining-Algorithmus angewendet wird. Anschließend werden die

so erzielten Data-Mining-Ergebnisse bewertet und die anfängliche Teilmenge schrittweise erweitert, bis die Ergebnisse des Data Minings als ausreichend gut erachtet werden (Liu und Motoda, 2002). Im Falle einer Klassifikationsaufgabe wird hierfür oft die Klassifikationsgenauigkeit herangezogen (Liu und Motoda, 2002). Diese misst den Prozentsatz der korrekt klassifizierten Instanzen eines Klassifikationsmodells und quantifiziert somit die Fähigkeit des Modells zur richtigen Vorhersage der Klassenlabels (Bishop, 2006).

Eine spezielle Methode der Instanzenauswahl stellt das Prototyping dar, das darauf abzielt, einen Datensatz durch eine repräsentative Menge von Prototypen zu ersetzen (Reinartz, 2002). In der Literatur sind hierfür verschiedene Ansätze dokumentiert. Einige Ansätze konzentrieren sich auf die Auswahl von tatsächlichen Instanzen aus dem Datensatz, die als repräsentativ angesehen werden, wobei andere Verfahren neue Datenpunkte generieren, die die ursprünglichen Daten synthetisieren oder abstrahieren (García et al., 2012). Diese synthetischen Prototypen können durch verschiedene Techniken erzeugt werden, einschließlich der Mittelwertbildung, dem Clustering oder durch komplexere Algorithmen, die auf dem maschinellem Lernen basieren (Liu und Motoda, 2002; Reinartz, 2002).

Weitere Methoden der Instanzenreduktion

Cluster-Algorithmen teilen Datensätze wie bereits im Abschnitt 2.2.3 zum Data Minnig beschrieben in Gruppen oder Cluster auf, sodass Instanzen innerhalb eines Clusters einander ähnlich sind und sich von Instanzen anderer Cluster unterscheiden. Im Kontext der Datenreduktion können Clusterrepräsentationen, wie beispielsweise Instanzen der Clusterzentren, anstelle der tatsächlichen, gesamten Daten verwendet und so die Anzahl der Instanzen reduzieren, was in Anwendungen, in denen Daten in distinkte Gruppen organisiert werden können, wirksam sein kann (Berkhin, 2006). Es gibt eine Vielzahl von Clusteralgorithmen zur Definition von Clustern und zur Messung ihrer Qualität. Das Clustering gehört wie bereits dargelegt zu der häufigsten Aufgabe des unüberwachten Lernens im Data Mining. Daher werden Clusterverfahren hier aufgrund der deutlichen Überschneidungen zwischen dem Data Mining und der Datenreduktion nicht weiter behandelt. Für umfassende Abhandlungen zu dieser Thematik sei auf Jain et al. (1999), Xu und Wunsch (2005), Kriegel et al. (2011), Berkhin (2006) und Aggarwal (2013) verwiesen.

Die bislang diskutierten Methoden der Instanzenreduktion können zusammengefasst der nicht-parametrischen Instanzenreduktion zugeordnet werden (García et al., 2015; Han et al., 2011). Diese unterscheiden sich grundlegend von parametrischen Methoden, die ebenfalls eine enge Verbindung zu Data-Mining-Modellen aufweisen (Han et al., 2011). Bei parametrischen Ansätzen wird ein Modell eingesetzt, um die Daten zu approximieren, was bedeutet, dass in der Regel nur Parameter der Daten statt der tatsächlichen Daten selbst gespeichert werden müssen (Han et al., 2011). Parametrische Techniken, wie die Regression und log-lineare Modelle, gehen davon aus, dass die Datenverteilung durch eine festgelegte Anzahl von Parametern repräsentiert werden kann (Hastie et al., 2009). Diese Ansätze sind besonders effektiv, wenn es gelingt, die Datenstruktur effizient mit mathematischen Funktionen anzunähern (Han et al., 2011). Parametrische Methoden haben daher ebenfalls eine enge Beziehung zu Data-Mining-Verfahren, da sie darauf abzielen, das Wesentliche großer Datensätze zu extrahieren und werden daher nicht weiter behandelt.

Die Darstellung der verschiedenen Methoden zur Datenreduktion sowie exemplarischer

Verfahren, kombiniert mit den Differenzierungen zu Datensätzen und der Nicht-Trivialität von Wissensentdeckungsprozessen, veranschaulicht die ausgeprägte Komplexität dieser Thematik, die aus divergenten theoretischen Konzepten resultiert. Wie bereits in der Einleitung und zu Beginn dieses Kapitels erörtert wurde, ist neben der Komplexität der Methoden auch die Problematik der umfangreichen und schwer überschaubaren Literatur zu Datenreduktionsverfahren von Bedeutung. Diese stellt die Anwender vor Herausforderungen bezüglich der Eignung von Datenreduktionsverfahren in verschiedenen Kontexten. Folglich wird im anschließenden Kapitel eine systematische Literaturrecherche durchgeführt, um die umfassenden Entwicklungen zu Datenreduktionsverfahren zu analysieren, zu strukturieren und den Anwendern somit eine praktikablere Sicht auf Datenreduktionsverfahren zu ermöglichen und so eine begründete Auswahl zu Datenreduktionsverfahren zu erleichtern.

3 Identifikation und Analyse bestehender Datenreduktionsverfahren

Im zweiten Kapitel wurden die kontextuellen sowie technischen Grundlagen der Datenreduktion erörtert. Auf dieser Basis wird nun eine systematische Literaturrecherche durchgeführt. Es werden Studien analysiert, um Datenreduktionsverfahren zu identifizieren, zu strukturieren und Kriterien für die Auswahl von Datenreduktionsverfahren zu entwickeln. Diese Kriterien werden verwendet, um Datenreduktionsverfahren für den Kontext der Wissensentdeckung in Datenbanken einzuordnen und zu bewerten, um somit die Auswahl von Datenreduktionsverfahren zu erleichtern. In Abschnitt 3.1 wird die Methodik der systematischen Literaturrecherche dargelegt. In Abschnitt 3.2 werden die Ergebnisse der Recherche hinsichtlich identifizierter Auswahlkriterien zu Datenreduktionsverfahren dargelegt und anschließend im Kontext von Wissensentdeckungsprozessen betrachtet. Nachdem Ausprägungen der Kriterien festgelegt wurden erfolgt in Abschnitt 3.3 die Kategorisierung und Bewertung identifizierter Datenreduktionsverfahren anhand der entwickelten Kriterien.

3.1 Methodik der systematischen Literaturrecherche

Um das übergeordnete Ziel zu erreichen, wird nun eine systematische Literaturrecherche durchgeführt. Systematische Literaturrecherchen ermöglichen eine umfassende, transparente und reproduzierbare Basis für die Identifizierung und Analyse relevanter Literatur. Diese Methodik soll eine systematische Erfassung aller verfügbaren Erkenntnisse zu Datenreduktionsverfahren ermöglichen. Die Strukturiertheit und Methodik einer systematische Literaturrecherche ermöglicht es, klare Rechercheziele zu definieren und eine präzise Suchstrategie zu entwickeln. Dies gewährleistet, dass kein relevantes Wissen übersehen wird und potenzielle Verzerrungen in der Ergebnisgewinnung minimiert werden. Die detaillierte Dokumentation und Beschreibung des Rechercheprozesses ermöglicht es anderen Forschern, die Durchführung der Recherche zu reproduzieren und die Gültigkeit der Ergebnisse zu überprüfen. Darüber hinaus ermöglicht die Synthese der Ergebnisse eine umfassende Analyse der vorhandenen Literatur und liefert fundierte Schlussfolgerungen.

Die angewandte Methode der systematischen Literaturrecherche orientiert sich an den etablierten Leitlinien von Kitchenham und Charters (2007), die sich durch über 10.000 Zitationen in der wissenschaftlichen Informatik als verlässlich erwiesen hat. Diese methodische Vorgehensweise umfasst die präzise Planung der Recherche, die Durchführung des Suchprozesses und die Darstellung der Ergebnisse. Dabei werden zunächst die Forschungsziele klar definiert, eine detaillierte Suchstrategie entwickelt und Auswahlkriterien für die Aufnahme relevanter Literatur festgelegt. Anschließend erfolgt eine eingehende Erläuterung der Schritte zur Datenextraktion und -aggregation, gefolgt von der Durchführung der Recherche sowie der Synthese der gesammelten Erkenntnisse.

3.1.1 Planung der systematischen Literaturrecherche

In der Planung der systematischen Literaturrecherche werden zunächst die Rechercheziele festgelegt. Das Ziel dieser systematischen Literaturrecherche besteht darin, Datenreduktionsverfahren zu identifizieren und anhand praktikabler Kriterien einordnen und bewerten zu können. Um dies zu erreichen werden Veröffentlichungen untersucht, die Datenreduktionsverfahren thematisieren. Eine Beschränkung auf Publikationen spezifischer Disziplinen erscheint nicht zweckmäßig, da Verfahren der Datenreduktion interdisziplinär angewendet werden. Weiterhin wird für die Recherche die digitale Bibliothek *Scopus* genutzt, welche Publikationen von verschiedenen Verlagen wie *Elsevier*, *Springer*, dem *Institute of Electrical and Electronics Engineers (IEEE)* und der *Association for Computing Machinery (ACM)* beinhaltet. Die Suchbegriffe werden in englischer Sprache verfasst, da sie im technischen Bereich als die global vorherrschende Wissenschaftssprache gilt. Zur Konstruktion des Suchstrings wird ein in der Literatur weit verbreitetes vorgehen verwendet:

1. Ermittlung der Hauptbegriffe aus dem Rechercheziel
2. Verwendung von Synonymen und alternativen Schreibweisen der Hauptbegriffe
3. Überprüfung der obigen Schritte durch Abgleichen der Schlüsselwörter aus relevanten Forschungsarbeiten
4. Verwendung des Booleschen Operators *OR*, um alternative Schreibweisen und Synonyme zu verknüpfen sowie des Operators *AND* um die Hauptbegriffe zu verbinden

Zur Identifikation und Strukturierung der Datenreduktionsverfahren werden die allgemein formulierten Hauptbegriffe *data reduction* und *technique* festgelegt. sowie die in Abschnitt 2.3 vorgestellten Methoden der Datenreduktion. Um Kriterien zur Auswahl von Datenreduktionsverfahren identifizieren zu können, impliziert dies, dass die zu untersuchende Literatur sich vertieft mit Datenreduktionsverfahren auseinandersetzt. Demnach wird als weiterer Hauptbegriff *analysis* gewählt. Die für die Recherche verwendeten alternativen Schreibweisen sind in Tabelle 3.1 zu sehen.

Tabelle 3.1: Hauptbegriffe der Recherche sowie alternative Schreibweisen

Hauptbegriffe	Synonyme/alternative Schreibweisen
data	*dimension, attribute, feature, variable, instance, tuple, sample, prototype*
reduction	*selection*
technique	*method, algorithm*
analysis	*study, survey, review, comparison*

Einschränkung des Suchbereichs

Scopus bietet die Möglichkeit Recherchen auf den Titel zu limitieren. Auf diese Weise lässt sich Literatur aussortieren, die beispielsweise Datenreduktionsverfahren oder die Analyse derer lediglich erwähnen oder referenzieren, solche aber nicht explizit behandeln. Die Einschränkung der Recherche auf den Titel steigert somit die Relevanz und Präzision der

Suchergebnisse, da der Titel die Kerninhalte und Hauptthemen der Studien reflektiert und somit eine effiziente und zielgerichtete Vorauswahl möglich ist. Eine zusätzliche Beschränkung wird durch das Publikationsjahr vorgenommen, indem auf Publikationen ab dem Jahr 2015 fokussiert wird. Dadurch wird sichergestellt, dass die durchgeführte Literaturrecherche aktuelle Datenreduktionsverfahren erfasst und die Ergebnisse somit von unmittelbarer praktischer Relevanz sind. Berücksichtigt wird jedoch, dass ältere Verfahren, die nach wie vor relevant sind, in der wissenschaftlichen Praxis auch in neueren Veröffentlichungen diskutiert oder referenziert werden. Durch die Fokussierung auf die Literatur nach 2015 werden daher nicht nur neuere Forschungen erfasst, sondern auch ältere, grundlegende Verfahren, die weiterhin Anerkennung finden. Obwohl die Suchanfrage durch die Verwendung englischer Hauptbegriffe bereits auf englischsprachige Literatur ausgerichtet ist, kann es vorkommen, dass einige Publikationen nur englischsprachige Titel und Zusammenfassungen haben, während der Volltext in einer anderen Sprache verfasst ist. Um den Aufwand für das Aussortieren solcher Volltexte zu minimieren, wird die Suche zusätzlich auf englischsprachige Veröffentlichungen gefiltert.

3.1.2 Literaturauswahl

Trotz der bereits angeführten Beschränkungen des Suchbereichs beinhalten die Resultate der Recherche sowohl relevante als auch nicht zutreffende Publikationen. Resultate umfassen beispielsweise Arbeiten zu *MapReduce*, einem Verfahren das primär für die effiziente Verarbeitung großer Datenmengen durch deren Verteilung auf mehrere Systeme entwickelt wurde, und nicht explizit durch die Reduktion der Datenmenge mittels Datenreduktionsverfahren. Für die Erkennung und Eliminierung nicht relevanter Werke erfolgt hier die Festlegung von Ausschlusskriterien. Einige Limitierungen sind bereits in der Formulierung der Suchanfrage integriert und erfahren nachstehend eine weitere Präzisierung.

1. Formale Kriterien zur Ausschließung
    a) Der Volltext steht nicht öffentlich oder mithilfe des Zugangs der Technischen Universität Dortmund zur Verfügung.
    b) Der Volltext ist nicht in englischer Sprache verfasst.
    c) Das Erscheinungsjahr entspricht nicht dem Zeitraum 2015 bis Januar 2024.
2. Inhaltliche Kriterien zur Ausschließung
    a) Die Veröffentlichung behandelt keine Datenreduktionsverfahren, bei denen die Verarbeitung eines Datensatzeingangs zu einer kompakteren Darstellung als Ausgabedatensatz führt.
    b) Die Veröffentlichung bietet keine klaren und nachvollziehbaren Beschreibungen der angewendeten Datenreduktionsverfahren.

Die festgelegten Ausschlusskriterien werden verwendet, um die Suchergebnisse zu filtern und Studien zu selektieren, die für die Analyse relevant sind. Wenn eine Publikation den definierten Ausschlusskriterien entspricht, wird sie von der weiteren Analyse ausgeschlossen. Auf zusätzliche Qualitätskriterien wird verzichtet, da in Scopus indizierten Veröffentlichungen einer Begutachtung und Selektion durch ein unabhängiges Komitee, was bereits

auf ein entsprechendes Qualitätsniveau der Literatur schließen lässt.

3.1.3 Datenextraktion und -aggregation

Zur systematischen Organisation und detaillierten Erfassung von Informationen aus den einzelnen Studien wird ein Tabellenkalkulationsprogramm genutzt. Das weitere Vorgehen zielt darauf ab die definierten Ziele der Literaturrecherche zu erreichen. Anhand der Titel, Zusammenfassungen und Einleitungen der Studien werden vorläufige Kategorien zu den verschiedenen Methoden der Datenreduktion sowie Kriterien gebildet. Sollte eine klare Zuordnung von Studien auf dieser Grundlage nicht möglich sein, werden zusätzliche Informationen aus dem Volltext der Studien herangezogen. Diese entwickelten Kategorien können im Verlauf des Prozesses nach Bedarf aktualisiert oder verfeinert werden. Die in den Publikationen behandelten spezifischen Verfahren werden durch Schlagwörter erfasst, wobei ein iteratives Vorgehen angewendet wird, um sicherzustellen, dass alternative oder ähnliche Bezeichnungen der Verfahren standardisiert werden. Dies führt zu einer strukturierten Übersicht über die identifizierten Datenreduktionsverfahren, aufgeschlüsselt nach verschiedenen Kategorisierungsperspektiven pro Methode und den dazugehörigen Kriterien.

3.1.4 Durchführung des Suchprozesses

Durch den Prozess der systematischen Literaturrecherche wurden verschiedene Literaturquellen identifiziert und analysiert, die für die Erreichung des Rechercheziels relevant sind. Eine schematische Darstellung des Suchprozesses ist in Abbildung 3.1 dargestellt, der basierend auf dem formulierten Vorgehen erfolgte.

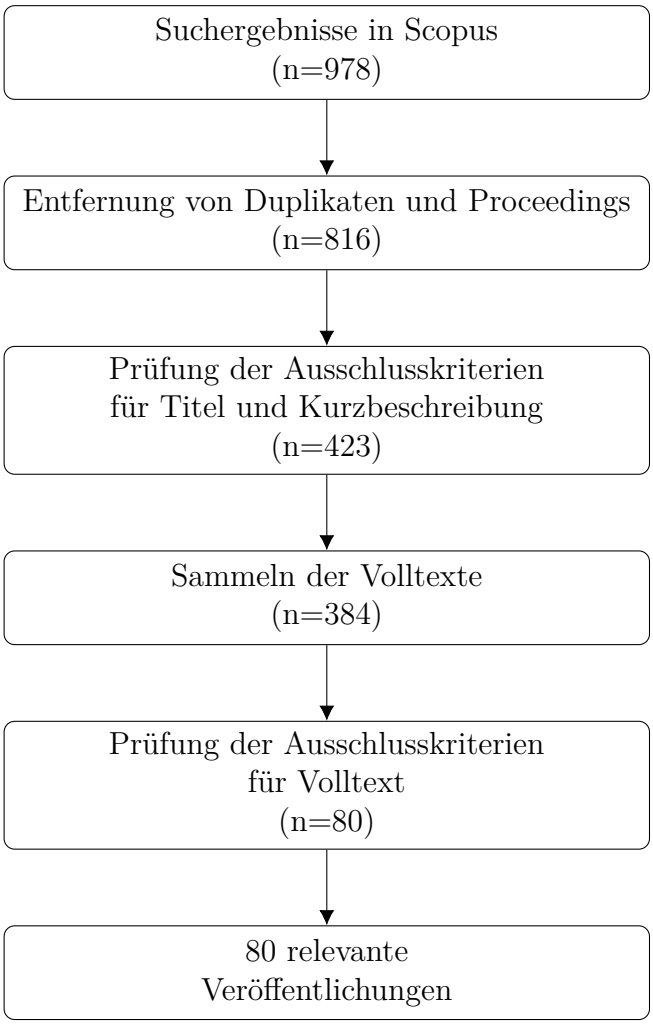


Abbildung 3.1: Darstellung des Suchprozesses

Die Synthese erfolgt in zwei Abschnitten. Zunächst werden aggregierte Kriterien für die Auswahl von Datenreduktionsverfahren dargelegt. Anschließend werden diese im Kontext der Wissensentdeckung in Datenbanken betrachtet und Kriterienausprägungen festgelegt. Im zweiten Abschnitt der Synthese werden anschließend die gebildeten Kategorisierungsperspektiven zu Datenreduktionsverfahren dargelegt und anhand der Kriterien bewertet. Die Dokumentation und Zuordnungen der Literatur befinden sich im Anhang und ist zu diesem Zwecke in drei Teile gegliedert. Anhang A.1 zeigt die Publikationsdetails zu den gesamten, untersuchten Veröffentlichungen. Anhang A.2 ordnet den Publikationen die thematisierten Kriterien zu. Anhang A.3 bietet eine Darstellung der identifizierten Datenreduktionsverfahren zu den jeweiligen Publikationen je Methode der Datenreduktion (vgl. Abschnitt 3.2).

3.2 Kriterien zur Auswahl von Datenreduktionsverfahren

Dieser Abschnitt präsentiert die Ergebnisse der Recherche bezüglich möglicher Kriterien zur Auswahl von Datenreduktionsverfahren. Identifizierte Kriterien werden zunächst

dargelegt und anschließend im Kontext der Wissensentdeckung in Datenbanken (vgl. Abschnitt 2.2.2) betrachtet. Danach erfolgt die Festlegung von Kriterienausprägungen anhand derer eine Zuordnung und Bewertung der Verfahren ermöglicht werden soll.

3.2.1 Identifizierte Kriterien zur Auswahl von Datenreduktionsverfahren

Durch die systematische Literaturrecherche identifizierte Auswahlkriterien wurden in *Input-* und *Output-* sowie *Prozedurale*-Kriterien eingeteilt und werden im Folgenden dargelegt. Die nachfolgenden Aussagen stützen sich hierbei auf den Anhang A.2 zur Kriterien-Auswertungstabelle der systematischen Literaturrecherche.

Input-Kriterien

Die identifizierten Input-Kriterien werden nachstehend in die Kategorien *formal*, *qualitativ* und *intrinsisch* klassifiziert. Die formalen Kriterien umfassen Aspekte, deren Nichterfüllung die Durchführbarkeit der Reduktionsverfahren einschränken oder nicht gegeben ist. Qualitative Kriterien des Inputs geben Aufschluss über die Eignung verschiedener Verfahren im Umgang mit qualitativen Mängeln des Inputs und deren Fähigkeit, trotzdem aussagekräftige Modelle zu generieren. Intrinsische Kriterien hingegen beziehen sich auf die Qualität der resultierenden Modellierung, allerdings nicht im Hinblick auf qualitative Merkmale des Datensatzes, sondern auf dessen inhärente Zusammenhänge. Während formale Datensatzeigenschaften für Anwender in der Regel unkompliziert zu bestimmen sind, erfordern qualitative und intrinsische Eigenschaften oft umfassendere Analysen. Die konsolidierten Input-Kriterien sind in Abbildung 3.2 visualisiert.

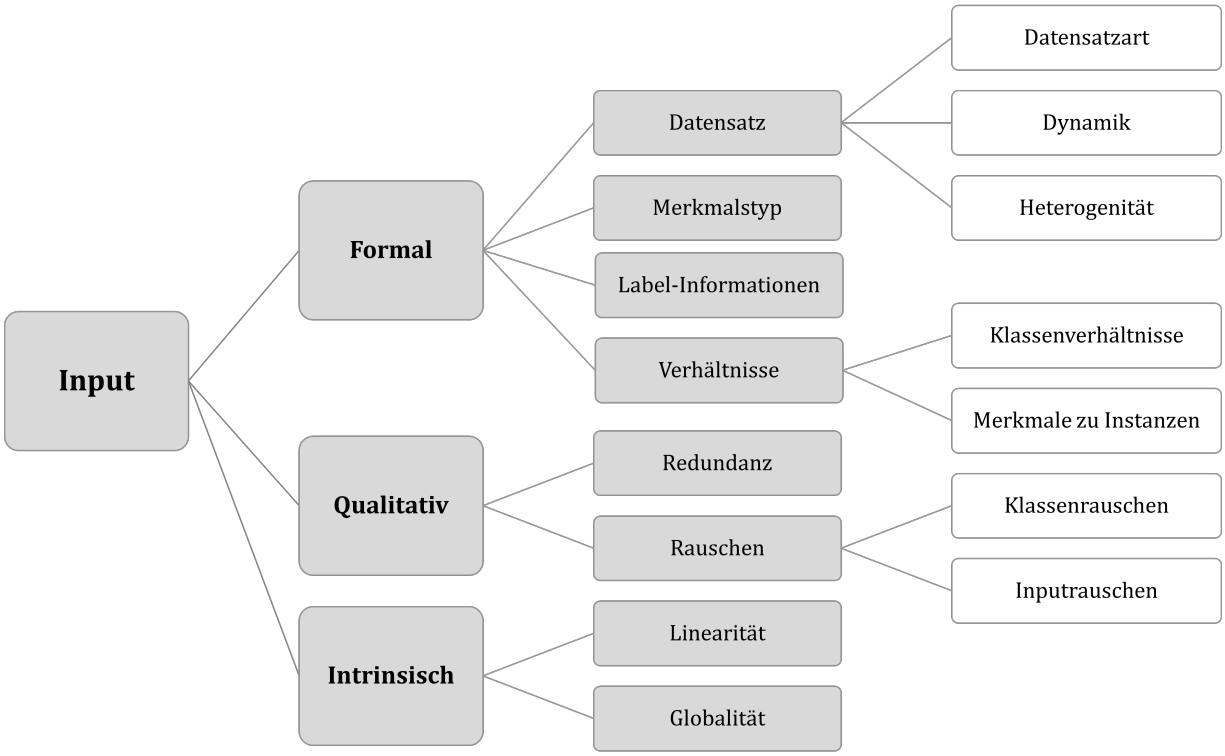


Abbildung 3.2: Input-Kriterien zur Auswahl von Datenreduktionsverfahren

Die systematische Auswertung der Forschungsliteratur zeigte, dass diverse Datenreduktionsverfahren spezifisch für die Anforderungen verschiedener Datensatzarten entwickelt wurden. Der Großteil der rezensierten Arbeiten konzentriert sich auf Reduktionsalgorithmen, die primär für die Verarbeitung einfacher Datenmatrizen (vgl. Abschnitt 2.1.2) ausgelegt sind. Vereinzelte Veröffentlichungen thematisieren, beziehungsweise präsentieren, innovative und spezialisierte Ansätze für den Umgang mit komplexeren Datenstrukturen, wie etwa Graphen und Netzwerken (vgl. Abschnitt 2.1.2). Diese Ansätze sind so gestaltet, dass sie die expliziten Verbindungen zwischen den Datenpunkten berücksichtigen, was den Erhalt von Schlüsselinformationen trotz der Reduktion der Daten gewährleistet. Über die Kategorisierung nach Datensatzarten hinaus wurden innovative Verfahren identifiziert, die für die Reduktion von heterogenen Daten und Datenströmen konzipiert sind. Exemplarisch seien hierfür die Verfahren aus der Methode der Merkmalsauswahl *Adaptive Multi-View Feature Selection* und das *Unsupervised Feature Selection for Multi-View Data* für heterogene Daten sowie *Online Streaming Feature Selection* für Datenströme genannt.

Die Recherche ergab, dass neben der Art des Datensatzes auch die spezifischen Merkmalstypen bei der Auswahl von Datenreduktionsverfahren entscheidend sind. Im Gegensatz zu den in Abschnitt 2.1.2 vorgestellten Kategorisierung in nominal, ordinal und metrisch wird in den analysierten Studien meist eine Unterscheidung in kategorische und numerische Daten bevorzugt. Die Präferenz für bestimmte Merkmalstypen basiert dabei auf theoretischen Grundlagen oder spezifischen Berechnungen der Verfahren. Je nach zugrunde liegender Theorie oder den angewandten Berechnungen erweisen sich Verfahren als geeignet für bestimmte Typen von Merkmalen (vgl. Abschnitt 2.1.2). Allerdings wurde diese Unterscheidung nur in wenigen untersuchten Studien thematisiert, nämlich in drei Veröffentlichungen zur Merkmalsauswahl. Obwohl in Summe 42 Veröffentlichungen zur

Instanzenreduktion sowie der Dimensionreduktion untersucht wurden, wird in keiner dieser explizit auf die Eignung der Verfahren zu bestimmten Merkmalstypen eingegangen. Diese Verfahren werden zwar anhand verschiedener Berechnungen beschrieben, die auf die Eignung unterschiedlicher Merkmalstypen hinweisen (vgl. Abschnitt 2.1.2), jedoch wird nicht weiter auf daraus resultierende Merkmalstypen eingegangen. Daher sind möglichen Implikationen zu Datenreduktionsverfahren hinsichtlich ihrer Eignung für verschiedene Merkmalstypen größtenteils lediglich durch Ableitungen aufgrund theoretischer Konzepte möglich. Als Bestätigung dieser Implikationen können Parallelen aufgrund der verwendeten Datensätze in den experimentellen Analysen der Studien betrachtet werden, bei denen beispielsweise ausschließlich numerische Merkmale verwendet wurden.

Der Umgang mit verrauschten Daten, sowohl in Bezug auf den Input als auch auf Label-Informationen, wird auch in den untersuchten Studien diskutiert, sowohl für Verfahren der Instanzenreduktion als auch der Merkmalsreduktion. Eine Studie hebt beispielsweise hervor, dass der *ReliefF*-Algorithmus, ein Verfahren zur Merkmalsauswahl, robuste Leistungen bei der Verarbeitung von verrauschten Daten aufweist. In den analysierten Studien wird häufig der Erhalt der Modellergebnisse für die Güte der Datenreduktion herangezogen. Im Gegensatz dazu zeigt der *InfoGain*-Algorithmus signifikante Einschränkungen bei Datensätzen mit hohem Rauschanteil. Zudem lassen sich spezielle Verfahren identifizieren, die explizit für den Umgang verrauschter Daten konzipiert wurden, wie etwa *Robust PCA* (vgl. Abschnitt 2.3.1).

Ein weiteres diskutiertes Kriterium bezüglich des Inputs, das sich aus den untersuchten Veröffentlichungen ergibt, betrifft die Verhältnisse im Datensatz, sowohl hinsichtlich der Klassenverhältnisse als auch der Verhältnisse von Merkmalen zu Instanzen. Zum Teil liegen in den analysierten Studien Datensätze vor, in denen die Anzahl der Merkmale die Anzahl der Instanzen deutlich übersteigt, wie beispielsweise in Studien der Genforschung oder der Text-Kategorisierung. In diesen Studien erreichen die Merkmalsanzahlen oft sechs- oder sogar siebenstellige Werte, während die Anzahl der Instanzen im Bereich von fünf- bis teilweise zweistelligen Zahlen liegt. Verschiedene Merkmalsreduktionsverfahren werden auf diese Datensätze angewendet, um allgemeingültige Aussagen zur Güte der Reduktion zu treffen. Eine Studie hebt beispielsweise die Güte des eben genannten Merkmalsauswahlverfahrens *InfoGain* hervor, insbesondere in Fällen, in denen das Verhältnis von Merkmalen zu Instanzen sehr hoch ist. Der Algorithmus *ReliefF* zeigt eine verbesserte Reduktionsgüte in Szenarien, in denen die Anzahl der Instanzen die der Merkmale übersteigt. Erweist sich jedoch als weniger effektiv als *InfoGain*, wenn die Anzahl der Merkmale die der Instanzen übertrifft. Bei einer aggregierten Betrachtung der Studien, die diese Verhältnisse untersuchten, lässt sich jedoch feststellen, dass stets auf die individuellen, verwendeten Datensätze und Data-Mining-Algorithmen hingewiesen wird und allgemeingültige Aussagen nicht oder nur eingeschränkt möglich sind.

Zudem ließ sich im Zusammenhang mit dem Input feststellen, dass oft eine Transformation der Daten in bestimmte Formate notwendig ist, bevor spezifische Datenreduktionsverfahren sinnvoll angewendet werden können. Ein Beispiel hierfür ist die Notwendigkeit, Daten für die Anwendung der Hauptkomponentenanalyse zu normalisieren (siehe Abschnitt 2.2.2). Dies wird im weiteren jedoch nicht als Kriterium betrachtet, da sofern Merkmale in numerischer Form vorliegen, diese Transformationen kein Hindernis für die Ausführbarkeit darstellen.

Prozedurale Kriterien

Datenreduktionsverfahren wurden häufig unter weiteren Gesichtspunkten analysiert die im Folgenden zu prozeduralen Kriterien zusammengefasst werden. Diese sind in Abbildung 3.3 dargestellt.

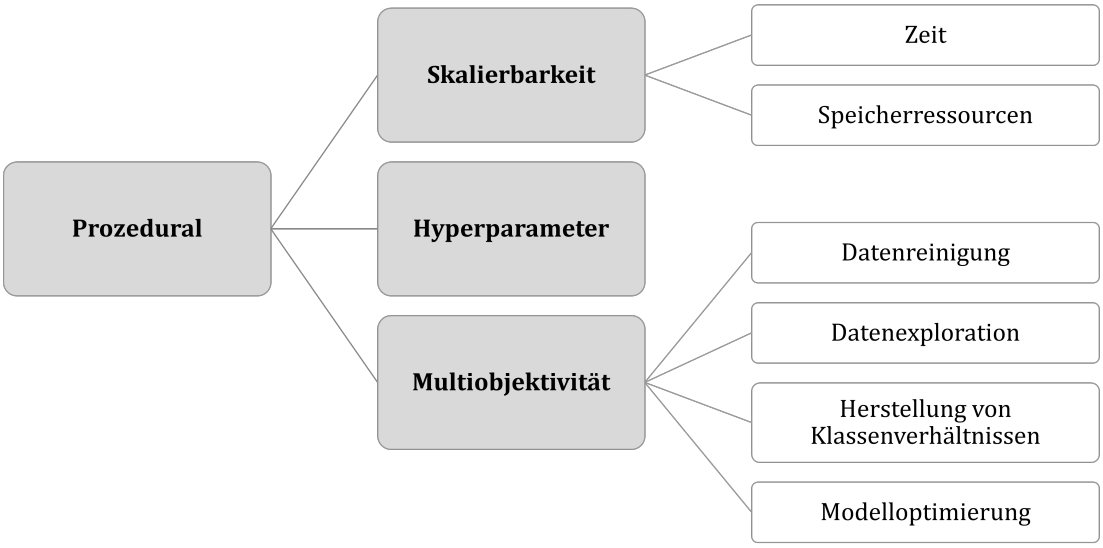


Abbildung 3.3: Prozedurale Kriterien zur Auswahl von Datenreduktionsverfahren

In einer Vielzahl der untersuchten Studien wurde das zeitliche Verhalten der Datenreduktionsverfahren thematisiert. Dies wurde dabei meist anhand der Zeitkomplexität beurteilt. Wenn das Verhalten der Verfahren bei sehr großem Input untersucht wurde, wurde hierfür die asymptotische Komplexität herangezogen. In dieser werden kleinere Terme in den Berechnungen der Verfahren vernachlässigt und nur dominierende Teile betrachtet. Entsprechend konnte auch die räumliche Komplexität der Verfahren in Bezug auf ihre Speicheranforderungen identifiziert werden. Diese wurde jedoch weitaus weniger diskutiert wurde als die zeitliche Komplexität. Zudem sei angemerkt, dass selbst bei dem Versuch, einen pragmatischen Zugang zu Datenreduktionsverfahren zu ermöglichen um deren Auswahl und Anwendung zu erleichtern, oft ein tiefgreifendes technisches Verständnis notwendig ist. Dies gilt insbesondere aufgrund der Konfiguration von Hyperparametern, die einen signifikanten Einfluss auf das Ergebnis haben können. Diese Herausforderung wird besonders bei Dimensionsreduktionsverfahren deutlich, bei denen die Mehrheit der Ansätze, insbesondere diejenigen, die auf Mannigfaltigkeitsannahmen basieren, eine oder mehrere konfigurierbare Parameter erfordern.

Des Weiteren kann die Reduktion der Daten beispielsweise durch das Herstellen von Klassengleichgewichten oder dem Bereinigen der Daten erfolgen (vgl. Abschnitt 2.2.2). Daher wird es als sinnvoll angesehen Multiobjektivitäten verschiedener Datenreduktionsverfahren als prozedurales Kriterium anzuführen. Als Beispiel wird hier das identifizierte Instanzenauswahlverfahren *Drop3* genannt, das eine Datenreduktion durch das Filtern stark verrauschter Daten erzielt. Insbesondere bei Befolgung von Vorgehensmodellen der Wissensentdeckung in Datenbanken, wo eine Datenbereinigung bereits vor der Datenreduktion

erfolgt, könnte die Wahl eines solchen Verfahrens nicht zielführend sein (vgl. Abschnitt 2.2.1).

Output-Kriterien

Neben den dargestellten Kriterien zum Input und der Prozedur galt es anhand der untersuchten Studien zudem Differenzierungen hinsichtlich des resultierenden Outputs der Datenreduktionsverfahren vorzunehmen. Identifizierte Kriterien im Zusammenhang des Outputs sind in Abbildung 3.4 dargestellt.

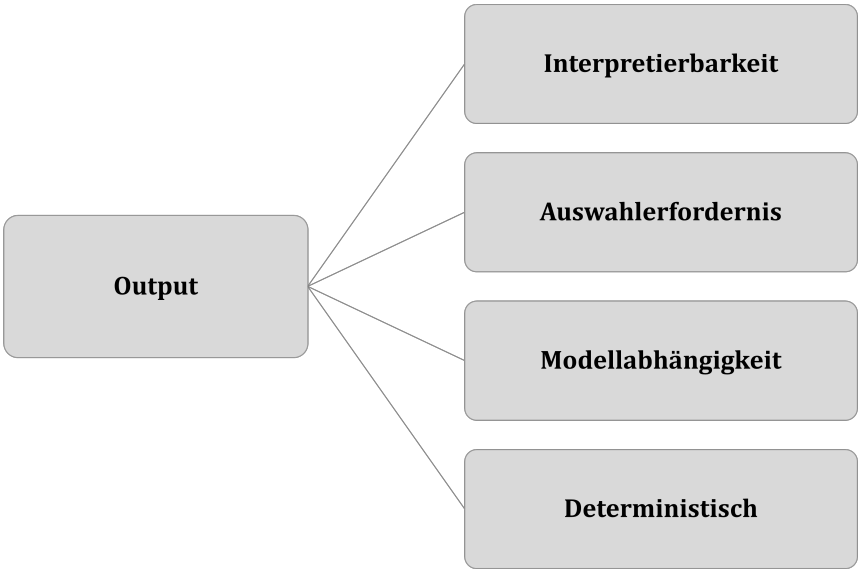


Abbildung 3.4: Output-Kriterien zur Auswahl von Datenreduktionsverfahren

Das erste Kriterium bezüglich des Outputs betrifft die Interpretierbarkeit. Wie in Abschnitt 3.2.3 beschrieben, basieren Verfahren der Merkmalsauswahl oder des Samplings darauf Merkmale oder Instanzen aus dem Datensatz zu wählen, wobei der resultierende reduzierte Datensatz die ursprüngliche Bedeutung beibehält. Der Output dieser Verfahren ist daher interpretierbar. Im Gegensatz dazu transformieren Verfahren der Dimensionsreduktion die Daten in einen anderen Merkmalsraum, wodurch der Output nicht oder nur schwer zu interpretieren ist (vgl. Abschnitt 2.3.1). Jedoch konnte im Kontext der Recherche festgestellt werden, dass selbst bei Verfahren der Dimensionsreduktion, die Merkmale in einen anderen Raum transformieren, eine gewisse Interpretation möglich ist. Beispielsweise ermöglichen verschiedene Visualisierungsmethoden eine gewisse Interpretation des Outputs. So konnte in den untersuchten Studien festgestellt werden, dass beispielsweise im Zusammenhang einer Datenreduktion mittels der *Hauptkomponentenanalyse* eine Interpretation des Outputs ermöglicht werden kann. Andererseits gibt es Verfahren, deren Output vollständig uninterpretierbar ist, wie es bei *Autoencodern* der Fall ist.

Ein weiteres relevantes Kriterium betrifft die Art des Outputs, der je nach angewendetem Verfahren unterschiedlich ausfallen kann. Der Output kann beispielsweise als reduzierter Datensatz vorliegen oder in Form von Gewichtungen oder Rankings. Dies bedeutet, dass die Anwender selbst festlegen müssen, wie viele der ursprünglichen Daten beibehalten werden sollen und somit die Reduktionsrate, also das Verhältnis des Outputs zum

Input, bestimmen müssen. Generell lässt sich aus den Studien ableiten, dass die Anzahl der beibehaltenen Merkmale und Instanzen von den Eigenschaften der Daten abhängt. In Studien zur Dimensionsreduktion werden oft sogenannte Ellenbogenplots verwendet. Diese Plots zeigen die Eigenwerte oder den Anteil der Varianz, den jede Achse (zum Beispiel eine Hauptkomponente) erklärt. In der Merkmalsauswahl werden gelegentlich heuristische Verfahren angewendet, um diejenigen Merkmale zu identifizieren, die die beste Klassifizierungs- oder Clusterleistung bieten.

Des Weiteren lassen sich Unterscheidungen des Outputs hinsichtlich der Modellabhängigkeit treffen (vgl. Abschnitt 2.2.3). In den untersuchten Studien werden diesen Verfahren oft sowohl Vorteile als auch Nachteile zugeschrieben. Einerseits ermöglichen sie oft sehr gute Ergebnisse bezüglich der Modellierung, andererseits kann der Nachteil darin liegen, dass der Output möglicherweise nicht für das Training anderer Modelle geeignet ist. Ein weiterer Nachteil besteht oft in dem hohen rechnerischen Aufwand dieser Verfahren. Deterministische Verfahren gewährleisten, dass bei wiederholter Anwendung des Verfahrens auf denselben Datensatz stets derselbe Output erzielt wird, sofern die Eingabeparameter und der Ausgangszustand des Datensatzes gleich bleiben.

Die identifizierten Kriterien zur Auswahl von Datenreduktionsverfahren werden nun im Kontext der Wissensentdeckung in Datenbanken betrachtet. Dies ermöglicht es Verfahren gezielter in den Kontext der Wissensentdeckung in Datenbanken einordnen zu können.

3.2.2 Kontextualisierung der identifizierten Auswahlkriterien

In Abschnitt 3.2.1 wurden die Ergebnisse der systematischen Literaturrecherche zu allgemeinen Auswahlkriterien für Datenreduktionsverfahren dargelegt. Um Datenreduktionsverfahren nun differenziert anhand der vorgestellten Kriterien in Prozesse der Wissensentdeckung in Datenbanken betrachten und einordnen zu können, werden diese nun kontextualisiert. Um eine umfassende Perspektive über den Einsatz von Datenreduktionsverfahren innerhalb der Wissensentdeckung in Datenbanken zu ermöglichen, wird auf die Festlegung eines spezifischen Vorgehensmodells verzichtet. Stattdessen erfolgt die Betrachtung zur Kontextualisierung der Kriterien abstrahiert anhand der vorgestellten Modelle des KDD-Prozesses, des CRISP-DMs sowie des generischen Vorgehensmodells (vgl. Abschnitt 2.2.1).

Wie in den Abschnitten 2.2.1 und 2.2.2 beschrieben, ist es in der Praxis der Wissensentdeckung gängig, Datensätze im Zuge der Datenselektion und -integration in einen konsolidierten, statischen Datensatz zu überführen. Folglich sind die in Abschnitt 3.2.1 vorgestellten Kriterien bezüglich der Dynamik von Datensätzen im Kontext der Wissensentdeckung nicht zwingend von Relevanz und werden in den weiteren Erörterungen nicht berücksichtigt. Weiterhin gilt es als grundlegend die Ausführbarkeit der Datenreduktionsverfahren zu gewährleisten, je nach selektierter Datengrundlage. Dies korrespondiert mit den dargestellten formalen Input-Kriterien zur Datensatzart, dem Merkmalstyp sowie Label-Informationen, welche im Folgenden weiterhin Betrachtung finden.

Die qualitative Betrachtung der zu reduzierenden Daten offenbart, abhängig vom gewählten Vorgehensmodell, Unterscheidungen für die Anwendung von Datenreduktionsverfahren. Im KDD-Prozess wird die Datenreduktion ausschließlich auf bereits bereinigte Daten angewendet, da diese nach der Datenvorverarbeitung erfolgt (vgl. Abschnitt 2.2.1). Im

CRISP-DM ist diese Reihenfolge nicht in der Art vorgegeben, da in diesem die Datenreduktion Teil der Datenvorbereitung ist. Somit kann die Datenreduktion im CRISP-DM auch auf nicht vorverarbeitete Daten angewendet werden, ausgehend von der ersten Iteration. Ein Verfahren, wie in Abschnitt 3.2.1 zur Multiobjektivität bezüglich der Datenbereinigung beschrieben, wäre demnach nicht zielführend, sofern das Verfahren die Reduktion der Daten durch das Entfernen von Ausreißer erzielt. Wurden Daten jedoch noch nicht bereinigt könnte ein solches Verfahren zur Zeitersparnis in Wissensentdeckungsprojekten führen, da dadurch sowohl eine Bereinigung als auch eine Reduktion der Daten erfolgt.

Ein weiteres bedeutendes Kriterium, dass sich aus den Rahmenbedingungen in Wissensentdeckungsprojekten ergibt, insbesondere unter Berücksichtigung des CRISP-DMs, welches eine industrielle Ausrichtung aufweist (vgl. Abschnitt 2.2.1), stellt die Zeit dar. Hierbei ist insbesondere die Skalierbarkeit von Datenreduktionsverfahren auf sehr große Datensätze von zentraler Bedeutung. Die Reduktion der Daten kann, wie in Abschnitt 2.3.1 dargelegt, zahlreiche positive Auswirkungen auf das Data Mining haben kann, beispielsweise dem Entgegenwirken des Fluchs der Dimension, die Erhöhung der Interpretierbarkeit der Ergebnisse des Data Minnigs, sowie sogar eine Steigerung der Güte des Data Mining Ergebnis bleibt es entscheidend, dass der zeitliche Aufwand der Datenreduktionsverfahren in bestimmten Szenarien die der Data Mining Algorithmen nicht übersteigt (vgl. Abschnitt 2.3). Dies ist insbesondere relevant, wenn Datensätze sehr groß sind und die Ausführbarkeit des Data Mining Algorithmus dadurch unpraktikabel wird. Zu den zeitlichen Rahmenbedingungen von Wissensentdeckungsprojekten können zudem weitere Kriterien hinzugezählt werden, wie die Notwendigkeit einer möglicherweise mehrmaligen Konfiguration von Hyperparametern der Datenreduktionsverfahren, oder die Notwendigkeit der Analyse des Outputs zur Bestimmung der zu bewahrenden Merkmale oder Instanzen des Datensatzes, was mit dem Kriterium der Auswahlerfordernis einhergeht. Obwohl diese Kriterien als Flexibilität der Verfahren eingestuft werden können, sollte der damit einhergehende zeitliche Aufwand ebenfalls berücksichtigt werden.

Wie in Abschnitt 2.2.1 beschrieben ist zudem Vorgehensmodell-übergreifend die Interpretation und Evaluierung der Ergebnisse des Data Minings von wesentlicher Bedeutung. Wie in Abschnitt 2.3.1 beschrieben ist es in verschiedenen Domänen und Anwendungen zentral, dass die semantische Integrität der Daten während der Vorverarbeitung und der Datenreduktion gewahrt bleibt, um Data Mining Ergebnisse auf Merkmale zurückführen zu können. Das Kriterium der Interpretierbarkeit des Outputs von Datenreduktionsverfahren gilt es daher bereits vor dem Data Mining, beziehungsweise der Modellierung zu beachten.

Abbildung 3.5 zeigt die für weitere Betrachtungen festgelegten, kontextualisierten Auswahlkriterien zu Datenreduktionsverfahren. Die Kriterien werden hierbei in den Kontext der generischen Phasen der Wissensentdeckung in Datenbanken (vgl. Abschnitt 2.2.1) gesetzt und wie dargelegt anhand Interdependenzen zu anderen Phasen, sowie Rahmenbedingungen von Wissensentdeckungsprozessen dargestellt.

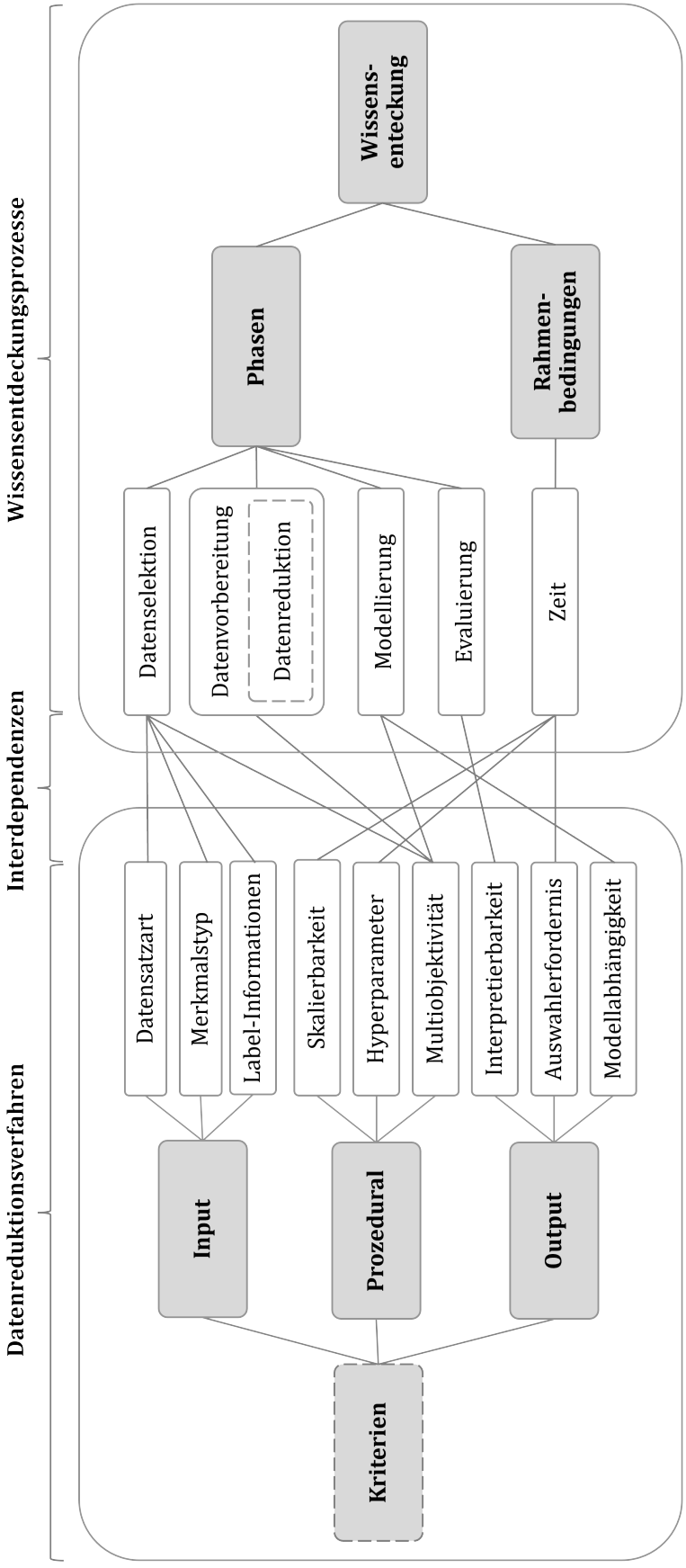


Abbildung 3.5: Kontextualisierte Auswahlkriterien zu Datenreduktionsverfahren

Bevor die Kategorisierung und Bewertung der identifizierten Verfahren anhand der kontextualisierten Kriterien erfolgt, werden zunächst unter Bezugnahme der Rechercheergebnisse Ausprägungen zu den Kriterien festgelegt.

3.2.3 Festlegung der Kriterienausprägungen

In Abschnitt 3.2.2 wurden die entwickelten Auswahlkriterien zu Datenreduktionsverfahren in den Kontext der Wissensentdeckung in Datenbanken gesetzt. Darauf aufbauend erfolgt nun die Festlegung der Kriterienausprägungen, welche die anschließende Einordnung und Bewertung der Datenreduktionsverfahren ermöglichen.

Die Anwendung von Datenreduktionsverfahren wurde in den untersuchten Studien nicht bezüglich der in Abschnitt 2.1.2 vorgestellten Klassifikation von Datensätzen in nicht-abhängigkeitsorientiert und abhängigkeitsorientiert, weiterhin in implizite und explizite Abhängigkeiten festgelegt. Es zeigte sich, dass Datensätze mit impliziter Abhängigkeit, häufig mit denselben Verfahren reduziert wurden wie nicht-abhängigkeitsorientierte Datensätze, wenn auch vereinzelt spezielle Verfahren behandelt wurden, deren Eignung sich auf den speziellen Umgang von implizit-abhängigen Datensätzen wie Zeitreihen konzentrieren (vgl. Abschnitt 2.1.2). Unterscheidungen manifestieren sich bei der Gegenüberstellung von Datensätzen mit nicht-abhängigkeitsorientierter Natur und solchen mit impliziter Abhängigkeit, im Vergleich zu Datensätzen mit expliziter Abhängigkeit. Im letzteren Fall kamen spezifische Verfahren zum Einsatz, um die expliziten Abhängigkeiten zu berücksichtigen. Aus diesem Grund erfolgt für die weitere Klassifizierung der Verfahren eine Unterteilung hinsichtlich einfacher, also nicht-abhängigkeitsorintierter und implizit abhängigkeitsorientierter Datensätze sowie komplexer Datensätze, die explizite Abhängigkeiten aufzeigen.

Bezüglich des Merkmalstyps wurden Studie identifiziert, die vereinzelte Merkmalsauswahlverfahren im Hinblick auf nominale und ordinale Merkmale differenzieren, diese jedoch nicht vertiefen. In den Studien wird vorrangig zwischen kategorischen und numerischen, respektive qualitativen und quantitativen Merkmalen unterschieden. Die Mehrzahl der Studien konzentriert sich auf die Anwendung von Datenreduktionsverfahren auf quantitative Merkmale. Trotzdem wurden Verfahren identifiziert, die qualitativen Daten reduzieren. Diese Methoden nutzen beispielsweise Häufigkeitsverteilungen oder Klassenunterscheidungen bezüglich des Informationsgehalts und sind sowohl für nominale als auch für ordinale Daten geeignet, wie in Abschnitt 2.2.2 dargelegt. Daher wird eine weitergehende Unterscheidung der Merkmalstypen in kategorische und numerische vorgenommen. Hierbei werden numerische Merkmale bei Möglichkeit weiter in kontinuierliche und diskrete unterteilt. Dies geschieht mit dem Ziel, die gemeinsamen Eigenschaften von kategorischen und numerisch-diskreten Merkmalen, hinsichtlich ihrer diskreten Natur zu nutzen und damit die Anwendbarkeit von Datenreduktionsverfahren auf diese Merkmalstypen zu erweitern, sofern möglich.

Bei der Analyse der Veröffentlichungen hat sich gezeigt, dass die Verfügbarkeit und Art der Label-Informationen in Datensätzen ausschlaggebend für die Anwendung bestimmter Verfahren ist. Es wurde festgestellt, dass in Szenarien, die beschreibende Analysen oder unüberwachtes Lernen erfordern, keine Label-Informationen vorhanden sind, was eine eigenständige Kategorie begründet. Im Bereich der Klassifikation müssen vorhandene Label-Informationen teilweise weiter in binäre und Multiklassen-Labels aufgeschlüsselt werden,

da identifizierte Datenreduktionsverfahren nicht immer in der Lage sind beide Aufgaben zu bewältigen. Die Mehrzahl der behandelten Verfahren ist jedoch für Mulitklassen-Aufgaben geeignet.

Die Festlegung der Ausprägungen der zeitlichen Skalierbarkeit wird anhand der Zeitkomplexität der Algorithmen vorgenommen, die durch Analyse der relevanten Studien ermittelt werden konnte. Verfahren mit geringer Skalierbarkeit zeichnen sich durch eine Zeitkomplexität aus, die mit der Anzahl der Merkmale oder Datensätze kubisch oder stärker anwächst. Methoden mit moderater Skalierbarkeit zeigen eine quadratische Zunahme der Zeitkomplexität. Algorithmen mit hoher Skalierbarkeit hingegen weisen einen linearen oder logarithmisch-linearen Anstieg der Laufzeit in Bezug auf die Anzahl der Merkmale oder Instanzen auf. Diese Ausprägungen der Skalierbarkeit ermöglichen Rückschlüsse für Anwender, die sich aus der Größe des zu reduzierenden Datensatzes ergeben, um diesen in angemessener Zeit zu verarbeiten. Eine mögliche, äquivalente Einteilung könnte in sehr große, große und normalgroße Datensätze erfolgen.

Des Weiteren werden die Ausprägungen bezüglich der Hyperparameter in die Kategorien *ja* für die Notwendigkeit von Einstellungen der Hyperparameter und *nein*, falls dies nicht der Fall ist, unterteilt. Obwohl im Rahmen der Recherche deutliche Unterschiede erkennbar sind, wie zum Beispiel bei Verfahren mit einem bis zu sieben Parametern, erfolgt die Einteilung dennoch nur in diese zwei Ausprägungen, um einen pragmatischen Ansatz zu verfolgen und den Anwendern eine einfachere Übersicht der Verfahren zu ermöglichen. Diese Argumentation wird auch für die Interpretierbarkeit des Outputs verfolgt. Hinsichtlich des Output-Kriteriums der Modellabhängigkeit ergeben sich jeweils nur die Ausprägungen *ja* und *nein*.

Wie bereits in Abschnitt 3.2.3 dargelegt, gibt es Verfahren, die die Datenreduktion lediglich durch das Entfernen von verrauschten Daten oder Ausreißern erreichen. Wenn ein Datensatz bereits in Bezug auf Rauschen oder Ausreißer bereinigt wurde, ist die Anwendung solcher Verfahren nicht zielführend. Ähnliches gilt für Methoden, die die Reduktion ausschließlich durch das Herstellen eines Klassengleichgewichts erreichen, falls dieses Gleichgewicht bereits in der Vorverarbeitung hergestellt wurde oder von Anfang an bestand. Als Ausprägungen der Kriterien der Multiobjektivität wird daher das Entfernen von Rauschen und Ausreißern als auch das Herstellen (Datenbereiniung) von Klassengleichgewichten festgelegt. Es sei angemerkt, dass Verfahren die eine solche Kennzeichnung erhalten, die Reduktion primär durch Entfernung besagter Daten erreichen. Verfahren die als Nebenprodukt der Reduktion eine Rauschmilderung herbeiführen, erhalten diese nicht. Abbildung 3.6 fasst die beschriebenen Kriterienausprägungen übersichtlich zusammen.

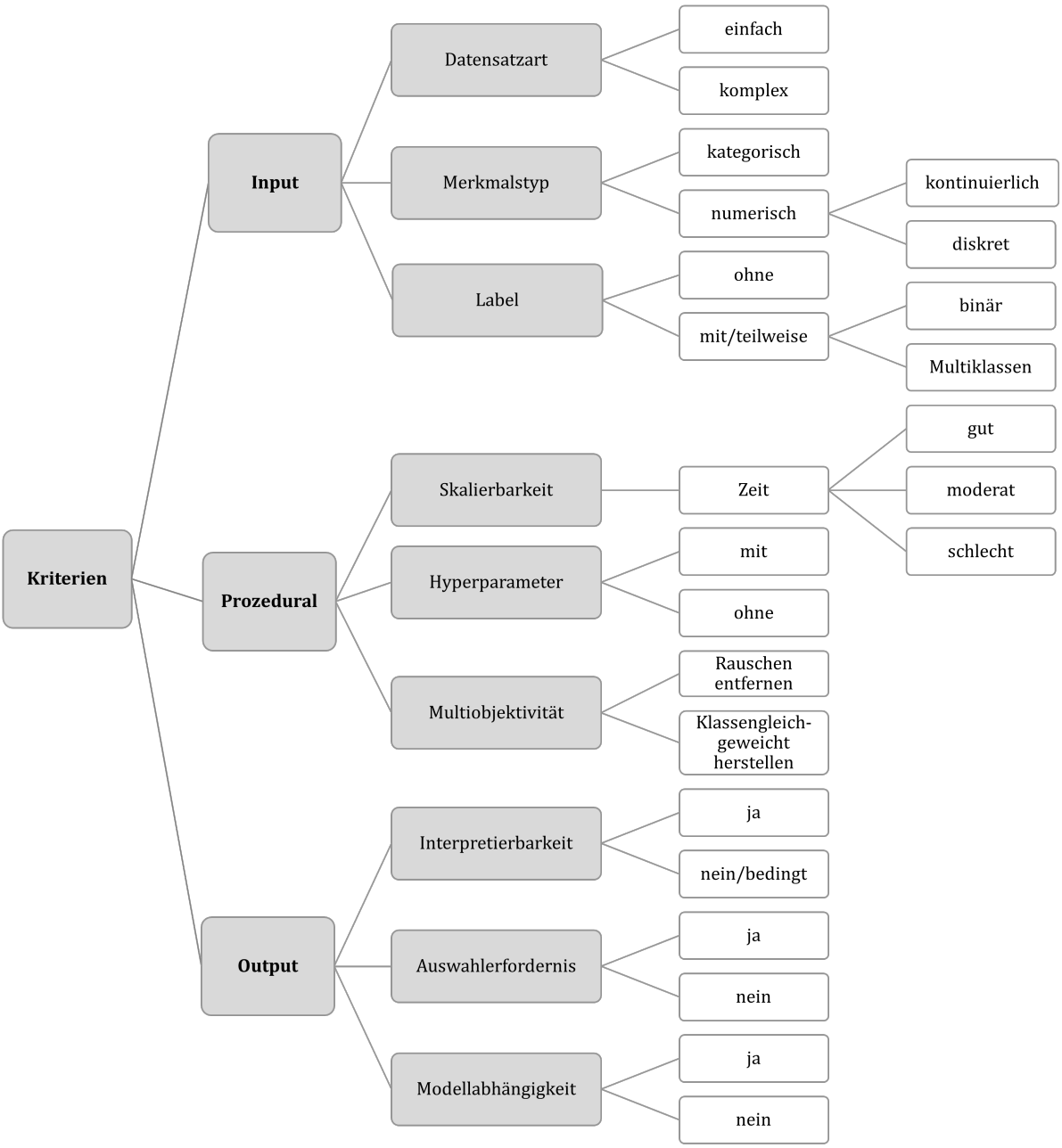


Abbildung 3.6: Festgelegte Kriterien-Ausprägungen zur Bewertung identifizierter Datenreduktionsverfahren

Nachdem die Ausprägungen der Kriterien festgelegt wurden, erfolgt nun die Bewertung der identifizierten Verfahren anhand dieser Kriterien.

3.3 Kategorisierung und Bewertung identifizierter Datenreduktionsverfahren

Nach Festlegung der Ausprägungen der Kriterien erfolgt nun die Kategorisierung und Bewertung der Datenreduktionsverfahren. Durch die Strukturierung der identifizierten Datenreduktionsverfahren aus technischen Perspektiven werden Implikationen für die Kriterien-Ausprägungen abgeleitet. Dabei werden repräsentative Verfahren aus den verschiedenen Kategorien vorgestellt und die Ergebnisse übersichtlich in tabellarischer Form zusammengefasst. Die Vorgehensweise orientiert sich an den in Abschnitt 2.3 dargelegten Methoden der Datenreduktion, wobei Verfahren der Merkmalsreduktion in Abschnitt 3.3.1 und die der Instanzenreduktion in Abschnitt 3.3.2 betrachtet werden. Sämtliche identifizierte Verfahren sind den korrespondierenden Quellen im Anhang A.3 zugeordnet.

3.3.1 Verfahren der Merkmalsreduktion

In diesem Abschnitt werden die Ergebnisse systematischen Literaturrecherche zu identifizierten und analysierten Datenreduktionsverfahren zur Merkmalsreduktion vorgestellt. Verfahren der Merkmalsreduktion wurden im Rahmen der systematischen Literaturrecherche in instgesamt 65 Veröffentlichungen thematisiert. In 21 dieser Studien wurden dabei sowohl Dimensionsreduktionsverfahren als auch Merkmalsauswahlverfahren behandelt.

Merkmalsauswahlsverfahren

Im Zuge der systematischen Literaturrecherche konnte festgestellt werden, dass Verfahren zur Merkmalsauswahl, neben der in Abschnitt 2.3.1 vorgestellten Einteilung in Filter, Wrapper und Embedded zudem anhand der Überwachung eingeteilt werden kann (vgl. Abschnitt 2.2.3). In diesem Zusammenhang lassen sich Verfahren der Merkmalsauswahl in überwacht, unüberwacht und semi-überwacht gliedern (vgl. Abschnitt 2.2.3). Abbildung 3.7 zeigt die aggregierten Kategorisierungsperspektiven zu Merkmalsauswahlsverfahren. Dieser ist die Kategorisierung nach der theorethischen Basis zu entnehmen, anhand derer Ableitungen zu Ausprägungen des Kriteriums des Merkmalstyps möglich sind, deren Darlegung nun folgt.

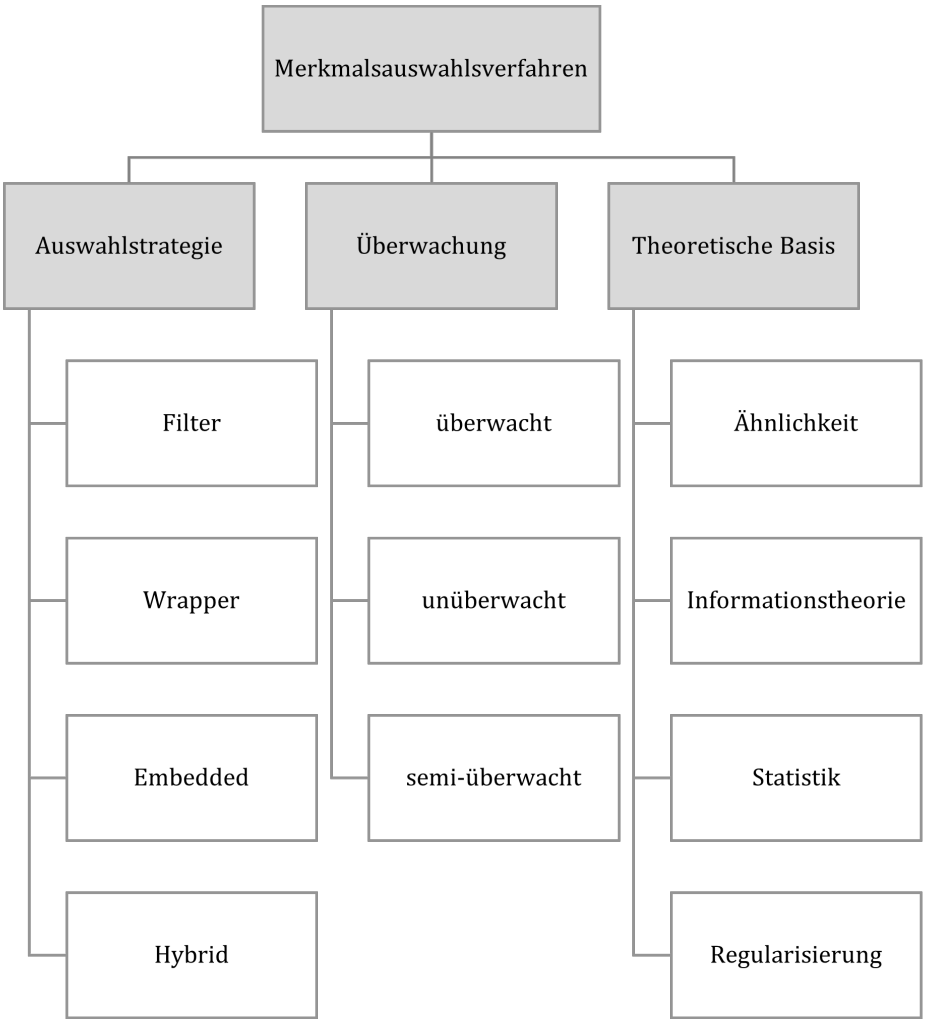


Abbildung 3.7: Kategorisierungsperspektiven identifizierter Merkmalsauswahlsverfahren

Verschiedene Algorithmen zur Merkmalsauswahl nutzen diverse Kriterien, um die Relevanz von Merkmalen zu definieren. Innerhalb dieser Algorithmen konnte eine Gruppe zusammengefasst werden, die die Wichtigkeit eines Merkmals anhand seiner Fähigkeit bewertet, die Ähnlichkeit von Daten zu bewahren. Bei der überwachten Merkmalsauswahl wird die Datenähnlichkeit aus den Labelinformationen abgeleitet. Bei nicht überwachten Merkmalsauswahlverfahren werden bei den identifizierten Verfahren Distanzmetriken genutzt, um die Datenähnlichkeit zu ermitteln. In diesem Zusammenhang lassen sich als in den Studien oft behandelte Verfahren wie *Laplacian Score*, *ReliefF*, *SPEC*, *Fisher Score* und *Trace Ratio* nennen. Ähnlichkeitsbasierte Algorithmen zur Merkmalsauswahl haben in den untersuchten Studien sowohl bei überwachten als auch bei nicht überwachten Lernproblemen sehr gute Ergebnisse erzielt. Zudem sind diese Methoden unabhängig von jeglichen Lernalgorithmen und die ausgewählten Merkmale eignen sich für viele nachfolgende Lernaufgaben. Als Nachteil wird in den Studien jedoch aufgeführt, dass die meisten der Verfahren die Redundanz von Merkmalen nicht bewältigen können, also dass sie wiederholt hoch korrelierte Merkmale während der Auswahlphase identifizieren. Aufgrund den zugrundeliegenden Berechnungen der sogenannten Affinitätsmatrix benötigen Verfahren dieser Kategorie numerische Werte als Input. Des Weiteren ist es bei den meisten Verfahren, die auf der Ähnlichkeit beruhen, notwendig die Merkmale zu skalieren (vgl. Abschnitt

2.2.2) um sicherzustellen, dass alle Merkmale in einem vergleichbaren Wertebereich liegen.

Weitere Verfahren lassen sich zu der Gruppe informationstheoretischer Verfahren zusammenfassen. Algorithmen dieser Familie nutzen verschiedene informationstheoretische Filterkriterien, um die Wichtigkeit von Merkmalen zu messen. Die Relevanz der Merkmale wird dabei üblicherweise durch seine Korrelation mit Klassenlabels gemessen, daher werden die meisten Algorithmen in dieser Familie auf eine überwachte Weise durchgeführt. Konzepte wie die *gegenseitige Information*, die *bedingte Entropie* und die *Informationsgewinnung*, werden für diese Verfahren genutzt, um die Beziehung zwischen Merkmalen und Klassenlabels quantitativ zu bewerten. Die gegenseitige Information misst beispielsweise, wie viel Information das Vorhandensein eines Merkmals über die Klassenzugehörigkeit eines Datensatzes offenbart. Durch die Minimierung der Redundanz zwischen Merkmalen und gleichzeitige Maximierung ihrer Relevanz für die Klassenlabels streben informationstheoretische Ansätze danach, eine Merkmalsmenge zu identifizieren, die die prädiktive Leistung von Lernmodellen optimiert. Oft behandelte Verfahren dieser Gruppe sind *minimum Redundancy Maximum Relevance* oder *Joint Mutual Information*. In den Studien wird im Zusammenhang dieser informationstheorethischen Verfahren beschrieben, dass diese aufgrund von Klassenvergleichen nur auf diskrete Variablen angewendet werden können. Bei kontinuierlichen Merkmalswerten sind vorab Techniken zur Diskretisierung erforderlich (vgl. Abschnitt 2.2.2). Im Gegensatz zu ähnlichkeitsbasierten Algorithmen zur Merkmalsauswahl können diese jedoch Merkmalsredundanz adressieren. Ähnlich wie bei ähnlichkeitsbasierten Methoden lässt sich verallgemeinern, dass diese Kategorie von jeglichen Lernalgorithmen unabhängig sind.

Die dritte Kategorie von Verfahren beruhen auf der Regularisierung und gehören zur eingebetteten Suchstrategie (vgl. Aschnitt 2.3.1). Fast alle indentifizierten Verfahren dieser Gruppe entwickeln ein Lernmodell unter der Überwachung von Klassenlabels (vgl. Abschnitt 2.2.3). Sie zielen darauf ab, die Anpassungsfehler zusammen mit verschiedenen Regularisierungstermen zu minimieren. Der Regularisierer zwingt viele Merkmalskoeffizienten dazu, klein oder genau null zu sein, wodurch die entsprechenden Merkmale einfach eliminiert werden können. Von der überwachten Seite aus nutzen diese Verfahren Klassenlabelinformationen, um die Auswahl der Merkmale zu leiten. Ein oft behandeltes Verfahren ist *Lasso* (Least Absolute Shrinkage and Selection Operator), die eine sogenannte L1-Regularisierung einsetzt um die Merkmalsauswahl vorzunehmen. Jedoch konnten auch nicht überwachte Verfahren dieser Kategorie identifiziert werden, wie das *Multi-Cluster Feature Selection*. Der Vorteil dieser Regularisierungsverfahren besteht darin, dass sie die Merkmalsauswahl in einen typischen Lernalgorithmus, wie beispielsweise der linearen Regression oder Support Vektor Maschinen, integrieren. Dadurch können sie oft eine sehr gute Leistung für den zugrunde liegenden Lernalgorithmus erzielen. Dennoch besteht der Nachteile dieser Methoden darin, dass sie durch die Merkmalsauswahl direkt einen bestimmten Lernalgorithmus optimieren, erzielen die ausgewählten Merkmale nicht notwendigerweise eine gute Leistung unter Anwendung anderer Data Mining Algorithmen. Da diese Verfahren zur Kategorie der eingebetteten Methoden gehören sind diese wie bereits in Abschnitt 2.3.1 dargelegt in der Regel rechenintensiver als Filterverfahren.

Eine weitere Kategorie von Algorithmen zur Merkmalsauswahl basiert auf verschiedenen statistischen Maßzahlen. Verfahren dieser Gruppe zeigen eine gute Skalierbarkeit auf. Es sei angemerkt, dass in den untersuchten Studien zur Merkmalsauswahl teilweise Verfahren

dieser Kategorie für eine schnelle, vorläufige Reduktion der Merkmale angewendet werden, um anschließend ausgefeiltere Verfahren, die jedoch höhere Zeitkomplexitäten aufweisen, anzuwenden. Ähnlich wie bei ähnlichkeitsbasierten Merkmalsauswahlverfahren bewerten diese die Wichtigkeit von Merkmalen individuell und können daher die Merkmalsredundanz nicht bewältigen. Algorithmen dieser Kategorie sind dabei nicht einem konkreten Merkmalstyp zuzuordnen sondern hängen von der jeweiligen statistischen Maßzahl ab (vgl. Abschnitt 2.1.2). Exemplarische Verfahren dieser Kategorie sind *Low-Variance*, *Chi-Quare* oder *Gini Index*.

Die meisten untersuchten Veröffentlichungen behandeln Verfahren der Merkmalsauswahl für einfache Datensätze basieren auf der starken Annahme, dass Merkmale voneinander unabhängig sind, während die inhärente Struktur der Merkmale ignoriert wird (vgl. Abschnitt 2.1.2). In vielen realen Anwendungen können Merkmale jedoch verschiedene Arten von Strukturen aufweisen, beispielsweise Baumstrukturen und Graphen (siehe Abschnitt 2.1.2). In solchen Fällen können Merkmalsauswahlalgorithmen, die Wissen über die Strukturinformation einbeziehen, relevantere Merkmale finden und somit nachfolgende Lernaufgaben verbessern (vgl. Abschnitt 2.1.2). Identifizierte Verfahren dieser Gruppe basieren auf dem Rahmenwerk der Regularisierung. Die Merkmalsstrukturen werden dabei a priori vorgegeben, was die automatische Inferenz der Strukturen aus den Daten für die Merkmalsauswahl zu einem herausfordernden Problem macht. Bekannte Vertreter von Verfahren dieser Kategorie sind das *Tree-Lasso* oder das *Graph-Lasso*.

In Tabelle 3.2 wird eine Auswahl der identifizierten Verfahren aus technischer Perspektive kategorisiert und anhand der vorgestellten Kriterien bewertet. Das Kriterium der Interpretierbarkeit des Outputs wird nicht explizit aufgeführt, da bei allen diesen Verfahren die ursprüngliche Form der Merkmale des Inputs beibehalten werden. Des Weiteren wurde auf Grund des Fokus auf die Datenreduktion lediglich Filter- und Embedded-Verfahren in die Bewertung aufgenommen, da Wrapper-Verfahren in den Studien aufgrund des hohen Rechenaufwands nicht auf große Datensätze angewendet werden.

Tabelle 3.2: Kategorisierung und Bewertung identifizierter Merkmalsauswahlsverfahren

Merkmalsauswahlsverfahren	Technisch			Input								Prozedural		Output	
	Strategie	Überwachung	Theoretische Basis	Merkmalsstruktur			Merkmalstyp			Label			hyperparameterfrei	Skalierbarkeit	Auswahlerfordernis	modellunabhängig
				einfach	komplex		kontinuierlich	diskret	kateg.	ohne	binär	mit				
					Graph	Baum	numerisch					Multiklassen				
Laplacian Score	Filter	nein	Ähnlichkeit	✓			✓	✓		✓					✓	✓
Trace Ratio	Filter	ja	Ähnlichkeit	✓			✓	✓			✓	✓			✓	✓
Fisher Score	Filter	ja	Ähnlichkeit	✓			✓	✓			✓	✓	✓	✓	✓	✓
ReliefF	Filter	ja	Ähnlichkeit	✓			✓	✓			✓	✓		○	✓	✓
Spectral Feature Selection	Filter	semi	Ähnlichkeit	✓			✓	✓		✓	✓	✓		○	✓	✓
Minimum Redundancy Maximum Relevance	Filter	ja	Informationstheorie	✓				✓	✓		✓	✓		○	✓	✓
Joint Mutual Information	Filter	ja	Informationstheorie	✓				✓	✓		✓	✓		○	✓	✓
Interaction Capping	Filter	ja	Informationstheorie	✓				✓	✓		✓	✓		○	✓	✓
Fast Correlation-Based Filter	Filter	ja	Informationstheorie	✓				✓	✓		✓	✓	✓	○		✓
Regularized Feature Selection	Embedded	ja	Regularisierung	✓			✓	✓			✓	✓			✓	
Least square loss	Embedded	ja	Regularisierung	✓			✓	✓			✓	✓	✓		✓	
Logistic loss	Embedded	ja	Regularisierung	✓			✓	✓				✓	✓		✓	
Multi-Cluster Feature Selection	Embedded	nein	Regularisierung	✓			✓	✓		✓					✓	
Low variance	Filter	nein	Stastistik	✓			✓	✓		✓			✓	✓		✓
Chi-square	Filter	ja	Stastistik	✓				✓	✓		✓	✓	✓	✓	✓	✓
Gini Index	Filter	ja	Stastistik	✓				✓	✓		✓	✓	✓	✓	✓	✓
T-score	Filter	ja	Stastistik	✓				✓	✓		✓		✓	✓	✓	✓
F-score	Filter	ja	Stastistik	✓				✓	✓		✓	✓	✓	✓	✓	✓
Graph Lasso	Embedded	ja	Regularisierung		✓		✓	✓			✓				✓	
Tree Lasso	Embedded	ja	Regularisierung			✓	✓	✓			✓				✓	

Nachdem in Tabelle 3.2 wesentliche Ergebnisse der Literaturrecherche zu Merkmalsauswahlverfahren übersichtlich zusammengefasst wurden erfolgt nun die Darlegung der Ergebnisse der Methode der Dimensionsreduktion.

Dimensionsreduktionsverfahren

Im Rahmen der Untersuchung von Verfahren zur Dimensionsreduktion wurden 25 wissenschaftliche Publikationen analysiert, die eine Vielzahl an Verfahren behandeln. Die durchgeführte Literaturrecherche ermöglichte die Bildung verschiedener alternativer Kategorisierungsperspektiven für Dimensionsreduktionsverfahren, wodurch Implikationen auf Ausprägungen verschiedener Kriterien möglich sind. Abbildung 3.8 visualisiert alternative Kategorisierungen der Dimensionsreduktionsverfahren, welche nachfolgend erörtert werden.

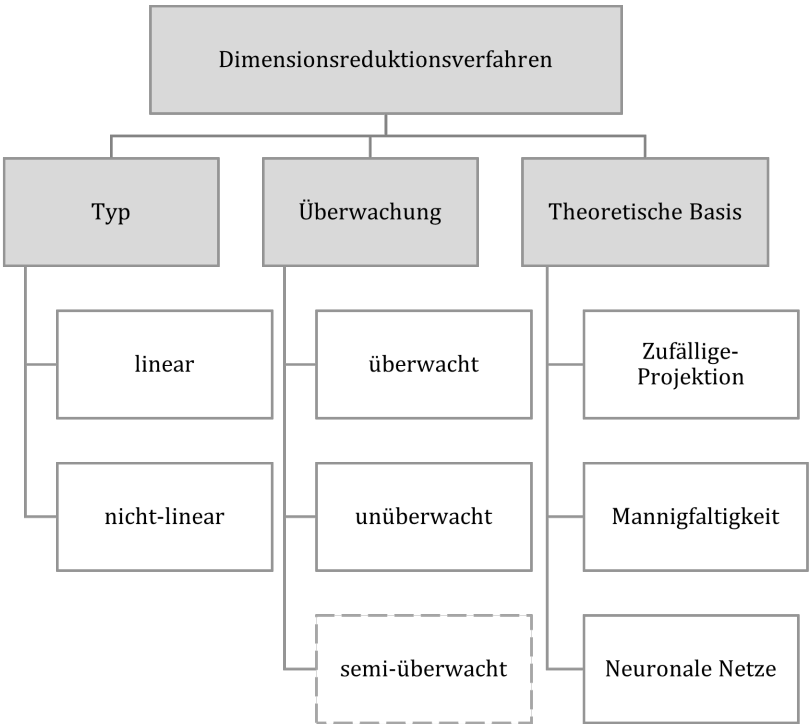


Abbildung 3.8: Kategorisierungsperspektiven identifizierter Dimensionsreduktionsverfahren

In den analysierten wissenschaftlichen Publikationen können Dimensionsreduktionsverfahren analog zu den Verfahren der Merkmalsauswahl nach ihrer Überwachung klassifiziert werden. Anhand der identifizierten Verfahren lässt sich jedoch formulieren, dass diese im Vergleich zu Merkmalsauswahlverfahren eher unüberwacht erfolgen.

Die Erste Kategorie zur Strategie wird im Folgenden als zufällige Projektion zusammengefasst. Diese Verfahren, darunter die in Abschnitt 3.3.1 vorgestellte Hauptkomponentenanalyse (Principal Component Analysis, PCA) basieren auf dem Prinzip, dass hochdimensionale Datenpunkte auf einen niedriger dimensionierten Raum auf eine zufällige weise projizieren, die sich auf Grund der Variablitätsrichtungen beziehungsweise datensatzcharakteristischen Merkmalsachsen beruhen. Die meisten Verfahren dieser Kategorie

sind primär für numerische Merkmale geeignet, da sie die Berechnung einer Kovarianzmatrix erfordern (vgl. Abschnitt 2.1.2). Identifizierte Erweiterungen wie die Korrespondenzanalyse (Correspondence Analysis, CA) ermöglichen jedoch die Verarbeitung kategorialer Daten. Die Korrespondenzanalyse nutzt eine Kontingenztafel um die Reduktion der Daten anhand gemeinsam auftretender Häufigkeiten der verschiedenen Kategorien zu leiten. Diese Kategorien können dabei nominaler aber auch ordinaler Natur sein, da lediglich die Häufigkeit der Kategorien betrachtet wird (vgl. Abschnitt 2.1.2). Auch numerische Merkmale diskreter Art könnten hierbei Anwendung finden. Die *Lineare Diskriminanzanalyse* nutzt klassenspezifische Informationen, um die Daten so zu projizieren, dass eine Maximierung der Klassentrennung erreicht wird, indem die Streuung zwischen den Klassen maximiert und die Streuung innerhalb der Klassen minimiert wird und stellt somit eines der in Unterzahl repräsentierten überwachten Verfahren der Dimensionsreduktion dar.

Mannigfaltigketisbasierte Verfahren gründen auf dem Prinzip, Daten nicht zufällig aufgrund von Variablitäten in den Daten zu reduzieren, sonder Daten in einen Raum reduzierter Dimension zu projizieren, um so essenziellen nachbarschaftlichen Beziehungen zwischen den Datenpunkten zu bewahren. Die Algorithmen benötigen eine Nachbarschaftsrepräsentation der Datenpunkte in der anschließend eine Abbildung dieser Repräsentation in einem Raum geringerer Dimension erfolgt. Hierbei wird angestrebt Distanzen oder Ähnlichkeiten die im hochdimensionalen Raum existieren, möglichst genau im reduzierten Raum abzubilden. Ein zentraler Aspekt dieser Techniken ist ihre Anpassungsfähigkeit an unterschiedliche Merkmalstypen, die durch die Auswahl verschiedener Distanzmetriken ermöglicht wird. In Studien wurde beispielsweise die Euklidische- oder Manhattan-Distanz für numerische Daten, die Hamming-Distanz für kategoriale Daten oder die Gower-Distanz für gemischte Datensätze aus numerischen und kategorialen Merkmalen verwendet. Diese Verfahren ermöglichen auch die Analyse komplexer Datensätze zu Graphen, indem beispielsweise Distanzmetriken wie der kürzeste Pfad verwendet werden. Dennoch benötigen die meisten Verfahren dieser Gruppe die Einstellung mindestens einem, meist jedoch mehrerer Parameter.

Eine weitere Verfahrensgruppe, die zusammengefasst werden kann ist die auf neuronalen Netzen basiert. Neuronale Netz basierte Dimensionsreduktionsverfahren verwenden tiefgreifende Lernarchitekturen, um eine Transformation der Daten von einem hochdimensionalen Raum in einen Raum reduzierter Dimension zu ermöglichen. Diese Transformation erfolgt durch das Training des Netzwerks, das darin besteht, eine Reihe von Gewichtungen zu lernen, welche die Daten in einer Weise abbilden, dass wesentliche Informationen erhalten bleiben, während redundante oder weniger informative Aspekte minimiert werden. Ein exemplarisches identifiziertes Verfahren dieser Kategorie ist der *Autoencoder*. Ein bedeutender Vorteil dieser Verfahren ist ihre hohe Anpassungsfähigkeit, die es ermöglicht, verschiedene Arten von Datenstrukturen und -komplexitäten zu behandeln. Durch die Anpassung der Netzwerkarchitektur, wie die Anzahl der Schichten und die Aktivierungsfunktionen, können diese Modelle für spezifische Anforderungen maßgeschneidert werden. Dennoch erfordert die Optimierung dieser Modelle oft eine sorgfältige Einstellung mehrerer Hyperparameter, wie Lernrate, Anzahl der Epochen und die Größe der versteckten Schichten, um optimale Ergebnisse zu erzielen.

Tabelle 3.3 bietet eine übersichtliche Zusammenfassung der Ergebnisse zu Verfahren der Dimensionsreduktion. In Klammern gesetzten Haken bezüglich des Merkmalstyps oder des Datensatzes deuten auf die potenzielle Eignung der Verfahren hin, welche jedoch von der

spezifischen Wahl der Distanzmetrik abhängig ist.

Tabelle 3.3: Kategorisierung und Bewertung identifizierter Dimensionsreduktionsverfahren

Dimensionsreduktionsverfahren	Technisch			Input					Prozedural		Output
				Merkmalsstruktur		Merkmalstyp		Label			
	nicht-linear	Überwachung	Theoretische Basis	einfach	komplex	numerisch	kategorisch	ohne	mit	hyperparameterfrei	Skalierbarkeit	Auswahlerfordernis
Principal Component Analysis		nein	zufällige Projektion	✓		✓		✓		✓	○	✓
Kernel Principal Component Analysis	✓	nein	zufällige Projektion	✓		✓		✓				✓
Linear Discriminant Analysis		ja	zufällige Projektion	✓		✓			✓		○	
Partial Least Squares		ja	zufällige Projektion	✓		✓			✓		○	✓
Neighborhood Components Analysis	✓	ja	zufällige Projektion	✓		✓			✓		○	
Correspondence Analysis		nein	zufällige Projektion	✓			✓	✓		✓	○	
Multiple Correspondence Analysis		nein	zufällige Projektion	✓			✓	✓		✓	○	
Classical Multidimensional Scaling		nein	zufällige Projektion	✓		✓		✓		✓	○	✓
Isometric Mapping	✓	nein	Mannigfaltigkeit	✓	(✓)	✓	(✓)	✓				✓
Diffusion Mapping	✓	nein	Mannigfaltigkeit	✓	(✓)	✓	(✓)	✓				✓
t-Distributed Stochastic Neighbor Embedding	✓	nein	Mannigfaltigkeit	✓	(✓)	✓	(✓)	✓			○	✓
Local Linear Embedding	✓	nein	Mannigfaltigkeit	✓	(✓)	✓	(✓)	✓			○	✓
Laplacian Eigenmaps	✓	nein	Mannigfaltigkeit	✓	(✓)	✓	(✓)	✓			○	✓
Maximum Variance Unfolding	✓	nein	Mannigfaltigkeit	✓	(✓)	✓	(✓)	✓				✓
Non-metric Multidimensional Scaling		nein	Mannigfaltigkeit	✓	(✓)	✓	(✓)	✓				✓
Bottleneck Neural Network	✓	ja	Neuronale Netze	✓	✓	✓	✓		✓		○	✓
Autoencoders	✓	ja	Neuronale Netze	✓	✓	✓			✓		○	✓

Neben einer Bevorzugung von Merkmalsauswahlverfahren gegenüber Dimensionsreduktionsverfahren zur Bewahrung der semantischen Integrität der Daten lässt sich abschließend zu den Methoden der Merkmalsreduktion aus den Studienergebnissen zusammenfassen, dass Dimensionsreduktionsverfahren tendenziell einen höheren zeitlichen Aufwand für die Skalierbarkeit erfordern und komplexere Parameterkonfigurationen benötigen als Merkmalsauswahlverfahren. Keine der identifizierten Verfahren der Dimensionsreduktion konnte anhand der vorgenommen Klassifikation als gut skalierbar bewertet werden.

3.3.2 Verfahren der Instanzenreduktion

Nach der Darstellung der Ergebnisse zu Verfahren der Merkmalsreduktion in Abschnitt 3.3.4 folgt nun die Präsentation der Ergebnisse der Instanzenreduktion. Es lässt sich festhalten, dass Verfahren der Merkmalsreduktion in der wissenschaftlichen Literatur deutlich häufiger thematisiert werden als Verfahren der Instanzenreduktion. Im Rahmen der Untersuchung zu Verfahren der Instanzenreduktion wurden insgesamt 17 Publikationen analysiert, deren Ergebnisse im nachfolgenden erörtert werden.

Instanzenauswahlverfahren

Zu Verfahren der Instanzenauswahl wurden insgesamt 11 Veröffentlichungen untersucht, in denen 33 Verfahren identifiziert werden konnten. Es wurden, neben der in Abschnitt 2.3.2 vorgestellten Kategorisierung in Filter- und Wrapper-Methoden, weitere Kategorisierungsperspektiven aggregiert. Abbildung 3.9 zeigt die gebildeten Kategorien die im folgenden beschrieben werden.

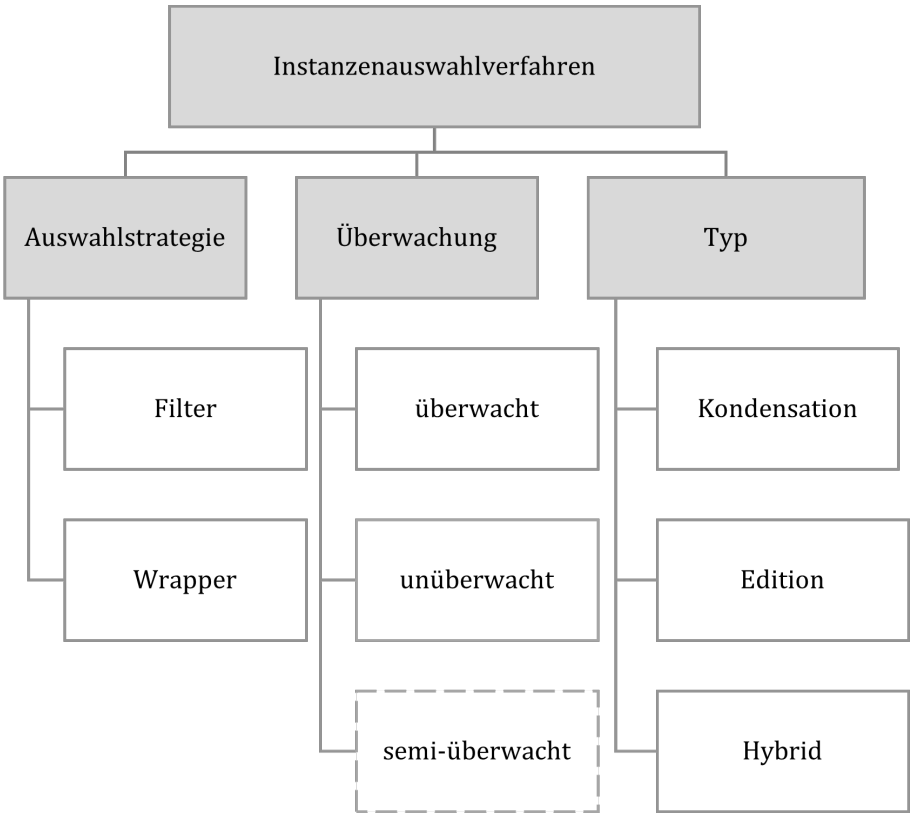


Abbildung 3.9: Kategorisierungsperspektiven identifizierter Instanzensauswahlsverfahren

Im Bereich der Instanzenauswahl lassen sich zwei Kategorien differenzieren sowie eine ergänzende, die die ersten beiden kombiniert: Kondensationsverfahren und editierende Verfahren sowie hybride Verfahren. Editierende Verfahren zielen darauf ab, den Datensatz zu reduzieren, indem sie ihn von Rauschen und Ausreißern bereinigen. Ein Beispiel eines solchen Verfahrens ist der *Edited Nearest Neighbor*. Dieser verbessert die Datenqualität, indem er Instanzen aus dem Trainingsdatensatz entfernt, deren Klassifikation durch die KNN-Regel nicht bestätigt wird.

Im Gegensatz dazu stehen die kondensierenden Verfahren, deren zentraler Algorithmus der *Condensed Nearest Neighbor* ist. Dieser reduziert den Trainingsdatensatz, indem er nur solche Instanzen behält, die für die korrekte Klassifikation der übrigen Instanzen notwendig sind. Das Ziel ist es, einen kompakteren Datensatz zu schaffen, der dennoch eine hohe Klassifikationsgenauigkeit gewährleistet. Trotz der hohen Effektivität des Verfahrens muss jedoch angemerkt werden, dass dieser Algorithmus aufgrund seiner kubischen Zeitkomplexität nicht effizient auf große Datensätze angewendet werden kann. Diese Einschränkung erfordert eine sorgfältige Abwägung des Einsatzes, insbesondere in umfangreichen oder zeitkritischen Anwendungsszenarien.

Verfahren die eine Reduktion der Daten sowohl durch die Edition als auch die Kondensation erreichen, werden als hybride Verfahren zusammengefasst. Eines dieser Verfahren ist Drop3. *Drop3* ist Teil einer größeren Familie von Algorithmen, die durch das systematische Entfernen von Instanzen aus einem Set gekennzeichnet ist, wobei anschließend überprüft wird, ob sich die Klassifikationsgenauigkeit verringert. *Drop3* beginnt zunächst mit der Anwendung des ENN-Algorithmus, um Rauschinstanzen zu eliminieren, und schreitet dann

zur weiteren Reduktion des Datensatzes vor. Die überwiegende Mehrheit der identifizierten Verfahren basiert auf Algorithmen für Klassifikationsprobleme, insbesondere auf dem k-Nearest Neighbors (kNN)-Algorithmus. Aufgrund der Variablität bezüglich der gewählten Distanzmetrik können von diesem unterschiedliche Merkmalstypen verarbeitet werden.

Neben diesen überwachten Verfahren konnte beispielsweise mit dem *Local Density-Based Instance Selection* auch ein unüberwachtes Instanzenauswahlverfahren identifiziert werden. Dieser wählt Instanzen aufgrund ihrer lokalen Dichte aus. Dabei werden Instanzen in dicht besiedelten Regionen des Merkmalsraums bevorzugt, um charakteristische Datenpunkte beizubehalten. Auch dieser ist je nach Distanzfunktion für verschiedene Merkmalstypen geeignet.

Tabelle 3.4 stellt repräsentative, identifizierte Instanzenauswahlverfahren dar. Die Abkürzung *Ra* in der Spalte zur Multiobjektivität stellt dabei dar, dass die Reduktion in erster Linie durch die Entfernung von Rauschen erfolgt.

Tabelle 3.4: Kategorisierung und Bewertung identifizierter Instanzenauswahlverfahren

Instanzenauswahlsverfahren	Technisch			Input				Prozedural			Output	
	Strategie	Überwachung	Typ	Datensatz einfach	Merkmalstyp		Label	hyperparameterfrei	Multiobjektivität	Skalierbarkeit	Auswahlerfordernis	modellunabhängig
					numerisch	kategorisch	ohne	mit				
Condensed Nearest Neighbor	Wrapper	ja	Kondensation	✓	✓	(✓)		✓				✓
Fast Condensed Nearest Neighbor	Wrapper	ja	Kondensation	✓	✓	(✓)		✓		○		✓
Generalized Condensed Nearest Neighbor	Wrapper	ja	Kondensation	✓	✓	(✓)		✓				✓
Edited Nearest Neighbor	Filter	ja	Edition	✓	✓	(✓)		✓	Ra	○		✓
All-K-Nearest Neighbor	Filter	ja	Edition	✓	✓	(✓)		✓	Ra	○		✓
Drop 3	Filter	ja	Edition	✓	✓	(✓)		✓	Ra			✓
Iterative Case Filtering	Filter	ja	Edition	✓	✓	(✓)		✓	Ra	○		
Local Set Border Selector	Wrapper	ja	Hybrid	✓	✓	(✓)		✓	✓	○		✓
Modified Selective Subset Selection	Wrapper	ja	Hybrid	✓	✓	(✓)		✓		○		✓
Local Density-Based Instance Selection	Filter	nein	Hybrid	✓	✓		✓			○	✓	
Clustering-Based Instance Selection	Filter	nein	Hybrid	✓	✓	(✓)	✓			○		
Class Conditional Instance Selection	Wrapper	ja	Hybrid	✓	✓	(✓)		✓				✓

Nachdem die Ergebnisse zu Instanzenauswahlverfahren dargelegt wurden folgt nun die Darlegung zu Samplingverfahren.

Samplingverfahren

Im Rahmen der systematischen Literaturrecherche wurden sechs Veröffentlichungen zu Samplingverfahren analysiert. Alternative Kategorisierungen zu Samplingverfahren, neben der in Abschnitt 2.3.2 in wahrscheinlichkeitsorientiert und nicht wahrscheinlichkeitsorientiert, sind Abbildung 3.10 dargestellt.

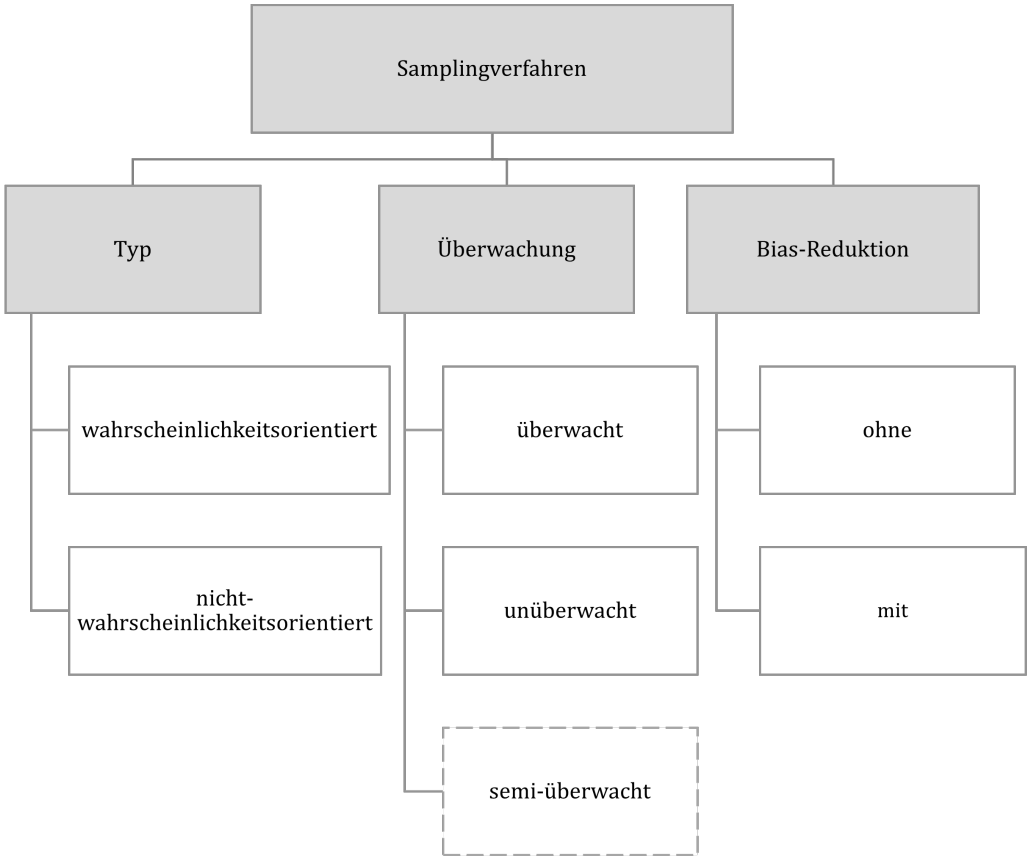


Abbildung 3.10: Kategorisierungsperspektiven identifizierter Samplingverfahren

Alternative Kategorisierungen der Samplingverfahren ließe sich dahingehend vornehmen, ob eine Bias-Reduktion angestrebt wird. Im vorangegangenen Abschnitt 2.3.1 wurde das *Random Sampling* eingeführt, welches zufällige Instanzen aus einem Datensatz auswählt. Dies führt zu Ungleichverteilungen im resultierenden reduzierten Datensatz, wenn Ungleichgewichte in den Daten vorhanden sind. Unterschiedliche identifizierte Samplingverfahren setzen sich von diesem simplen Stichprobenzug ab, indem sie während des Samplings eine Bias-Reduktion anstreben, um eine repräsentative Stichprobe verschiedener Populationen zu gewährleisten.

Im Bereich des überwachten Lernens zielt das *Random Under Sampling* darauf ab das Gleichgewicht zwischen den verschiedenen Klassen im Datensatz herzustellen und berücksichtigt dabei die Multidimensionalität des Klassenausgleichs. Des Weiteren wurden

spezifische Samplingverfahren identifiziert, die eine Datenreduktion ermöglichen, während bestimmte Merkmalsverteilungen erhalten bleiben. Im Gegensatz zur rein zufälligen Samplingmethode, bei der die Merkmalstypen des Datensatzes keine Rolle spielen, setzen die Samplingverfahren zur Bias-Reduktion mindestens ein Merkmal ein, um die Reduktion zu leiten. Im Fall des *Random Under Sampling* erfolgt dies beispielsweise anhand der Klassenlabels. Andere Verfahren zur Bias-Reduktion variieren je nach angestrebter Zielfunktion der Merkmalsverteilung. Im Unterschied zu den bereits vorgestellten Verfahren der Instanzen- und Merkmalsreduktion muss der gesamte Datensatz nicht in dieser Form vorliegen, sondern lediglich das Merkmal, anhand dessen die gewünschte Verteilung erzielt werden soll.

Ähnlich lässt sich diese Argumentation auf das in Abschnitt 2.2.2 vorgestellte *stratifizierte Sampling* anwenden. Diese Methode unterteilt den Datensatz so, dass eine gleichmäßige Repräsentativität aus verschiedenen Schichten gewährleistet ist, die mindestens ein Merkmal in Form von Kategorien aufweisen mussen. Es zeigt sich daher, dass die Auswahl von Sampling-Techniken eng mit dem gewünschten Auftreten verschiedener Populationen im Datensatz verbunden sind. Dies impliziert jedoch, dass zunächst eine Analyse anhand von Verteilungen im Datensatz durchgeführt werden muss. Gleichzeitig verdeutlicht dies den individuellen Charakter der verschiedenen Sampling-Techniken, der eng mit dem angestrebten Ziel verbunden ist. Demnach ließe sich die zufällige Stichprobe auch von der zweckmäßigen unterscheiden. Zudem lässt sich formulieren, dass sofern Zeitkomplexitäten aus den Studien extrahiert werden konnten, Samplingverfahren meist als gut einzustufen sind.

Die einfache Zufallsauswahl ist geeignet für allgemeine explorative Datenanalysen, bei denen keine spezifischen Verteilungen oder Muster innerhalb des Datensatzes erwartet werden. Diese Methode ist ideal, um eine gleichmäßige und unvoreingenommene Stichprobe aus der gesamten Datenmenge zu gewinnen, was sie zur bevorzugten Wahl für grundlegende statistische Analysen macht. Die einfache Zufallsauswahl kann daher gut als Ausgangspunkt für viele Data-Mining-Projekte, da sie eine unvoreingenommene Stichprobe der gesamten Datenpopulation bietet und somit grundlegende Einblicke ohne vorausgesetzte Hypothesen ermöglicht. Dieses Verfahren ist besonders effektiv für homogene Datensätze, in denen die Relevanz jedes Datenpunkts gleich ist und keine signifikanten Untergruppen oder Cluster existieren. Die Implementierung der einfachen Zufallsauswahl ist unkompliziert und gewährleistet, dass jeder Datensatz die gleiche Wahrscheinlichkeit hat, in die Stichprobe aufgenommen zu werden. Dies ist von besonderer Bedeutung, wenn keine vorherigen Informationen oder Vermutungen über die Daten vorliegen. Die einfache Zufallsauswahl zeichnet sich durch eine niedrige zeitliche Komplexität aus, die primär von der Größe des Datensatzes und der Effizienz der implementierten Zufallszahlengeneratoren abhängt. Generell ist diese Methode aufgrund ihrer schnellen Durchführbarkeit und der geringen technischen Anforderungen gut geeignet für leicht zugängliche Datensätze, die rasch verarbeitet werden können.

Für Untersuchungen, in denen Populationen in klar definierten Schichten, beispielsweise nach Alter oder Einkommen, analysiert werden, ist die stratifizierte Stichprobenziehung besonders geeignet. Diese Methode gewährleistet repräsentative Stichproben innerhalb jeder definierten Schicht und ist essentiell, wenn unterschiedliche demografische Gruppen miteinander verglichen oder spezifische Bevölkerungsteile detailliert analysiert werden sollen. Stratifizierte Stichprobenziehung kommt besonders dann zum Einsatz, wenn Datensät-

ze offensichtliche Schichtungen oder Kategorien aufweisen, die häufig bei demografischen Daten wie Alter, Geschlecht und Einkommen der Fall sind. Durch das Ziehen von Stichproben aus jeder Schicht proportional zu ihrer Größe in der Gesamtpopulation können Verzerrungen minimiert und die Genauigkeit der Analyse in jeder Gruppe maximiert werden. Die Zeitkomplexität der stratifizierten Stichprobenziehung ist allerdings höher als die der einfachen Zufallsauswahl, da sie eine vorherige Segmentierung der Daten in verschiedene Schichten oder Gruppen erfordert. Nach der Identifikation und Definition der Schichten müssen separate Stichproben für jede Gruppe gezogen werden. Diese zusätzlichen Schritte steigern sowohl die Komplexität als auch die benötigte Zeit für die Durchführung, besonders wenn die Kriterien für die Schichtung komplex sind oder wenn umfangreiche und vielschichtige Datensätze verarbeitet werden müssen.

Die Entscheidung, ob Instanzenauswahlverfahren oder Samplingverfahren für die Instanzenreduktion angewendet werden sollte, hängt wie erörtert, von mehreren Faktoren ab. Während Instanzenauswahlverfahren die Datenqualität durch die Eliminierung von Rauschen und die Auswahl hochrelevanter Datenpunkte priorisieren, konzentrieren sich Samplingverfahren auf die reine Reduktion der Datenquantität, oft ohne die Qualität der ausgewählten Daten zu berücksichtigen. Instanzenauswahlverfahren erfordern in der Regel mehr Rechenaufwand, da eine gründliche Analyse der Daten notwendig ist, um die besten Instanzen für die Auswahl zu identifizieren. Sampling hingegen ist in der Anwendung einfacher und weniger rechenintensiv, leistet jedoch in de Regel keine Steigerung der Qualität der Daten. Bei sehr großen Datensätzen erscheint anhand der untersuchten Studien Sampling aus zeitlicher Sicht bevorzugt werden zu sollen. Letztlich sollte die Wahl zwischen diesen beiden Methoden auf einer sorgfältigen Bewertung der spezifischen Anforderungen der Analyse oder des Modells basieren. Beide Techniken bieten wertvolle Strategien zur Datenreduktion, wobei ihre Eignung von den Zielen und Bedingungen des jeweiligen Projekts abhängt. In jedem Fall können die dargelegten Ergebnisse für die Auswahl der Verfahren für verschiedene Kontexte herangezogen werden. Aus der systematischen Literaturrecherche ließe sich jedoch empfehlen Samplingverfahren im Kontext der Wissensentdeckung in Datenbanken in anfänglichen Phasen der Datenselektion zu wählen, bevor weitere Analysen angestrebt werden.

Die vorgestellten Kategorisierungen und Bewertungen der Datenreduktionsverfahren anhand der dargelegten Kriterien bieten eine Übersicht, die zur Auswahl geeigneter Verfahren herangezogen werden kann. Um diese Auswahl weiter zu präzisieren, werden im Folgenden zusätzliche Empfehlungen und qualitative Zusammenfassungen der analysierten Studien dargelegt. Diese weiterführenden Informationen sollen dabei helfen, die Entscheidungsfindung für spezifische Anwendungskontexte zu verfeinern.

Die Entscheidung, ob eine Reduktion von Instanzen oder Merkmalen vorgenommen werden sollte, lässt sich auf Basis der untersuchten Studien nicht pauschal treffen und hängt wesentlich von den spezifischen Gegebenheiten des Datensatzes, sowie den Analysezielen ab. Die analysierten Studien bieten in diesem Kontext lediglich allgemeine Orientierungspunkte. Es zeigt sich, dass Verfahren zur Merkmalsreduktion über einen breiten Bereich von niedrigen zweistelligen bis hin zu siebenstelligen Merkmalsdimensionen Anwendung finden können. Pauschalaussagen, etwa im Verhältnis von eins zu zehn, die zur Entscheidung zwischen Merkmals- oder Instanzenreduktion herangezogen werden könnten, erweisen sich als irreführend. Laut den analysierten Studien wäre es nicht empfehlenswert, bei einem Datensatz mit 100 Merkmalen und 1000 Instanzen eine Instanzenreduktion ohne vorherige

Merkmalsreduktion vorzunehmen. Falls allgemeine Empfehlungen, ob eine Merkmals- oder Instanzenreduktion zum Ziele der Datenreduktion bevorzugt werden sollte, dann tendieren diese dazu, dass eine Merkmalsreduktion bereits ab dem mittleren zweistelligen Bereich sinnvoll sein kann. Die Reduktion der Instanzen sollte mindestens in das Verhältnis eins zu zehn resultieren, sofern bei der Reduktion der Instanzen kein Modell trainiert, beziehungsweise optimiert wird.

Zudem legt die qualitative Aggregation der Studienergebnisse nahe, dass es sinnvoll ist Verfahren der Merkmals- und Instanzenreduktion entsprechend der Größe des Datensatzes und der Komplexität der jeweiligen Verfahren auszuwählen. Wenn der Merkmalsraum klein ist, ist es ratsam Verfahren mit höheren Zeitkomplexitäten zu wählen, da diese in der Regel ausgereifter sind und zu besseren Ergebnissen in der anschließenden Analyseaufgabe führen, ohne in einer zeitlich unpraktikablen Reduktion der Daten zu resultieren. Eine vorsichtige Einteilung der Größenordnungen, basierend auf den untersuchten Studien, könnte folgendermaßen aussehen: Ein normaler Merkmalsraum umfasst bis zu 10^2 Merkmale, große Merkmalsräume liegen im Bereich von 10^2 bis 10^4, und sehr große Merkmalsräume beginnen ab 10^4 Merkmalen. Die Skalierbarkeit von Merkmalsreduktionsverfahren, wie sie in Abschnitt 3.3.1 erfolgte, sollte entsprechend diesen Einstufungen betrachtet werden und geeignete Verfahren aus diesem Bereich bevorzugt werden. Anwender sollten sich zunächst fragen, ob die Interpretation der Merkmale für die weiteren analytischen Betrachtungen von Bedeutung ist. Wenn dies nicht der Fall ist, legen die untersuchten Studien eine signifikantere Reduktion der Merkmale durch Dimensionsreduktionsverfahren nahe. Wenn ein spezifischer Data-Mining-Algorithmus optimiert werden soll und die Ergebnisse nicht mittels mehrerer Data-Mining-Algorithmen evaluiert werden sollen und zudem der Merkmalsraum in einem normalen Bereich liegt, ist es ratsam, bevorzugt eingebettete Merkmalsauswahlverfahren einzusetzen. Für weitere Betrachtungen können wie formuliert die Kriterien sowie Bewertungstabellen herangezogen werden. Größenordnungen zum Instanzenraum ließen sich grob äquivalent einteilen, jedoch jeweils um den Faktor zehn bis 100 in größere Richtung verschoben.

Nachdem nun Datenreduktionsverfahren analysiert, strukturiert und bewertet wurden, erfolgt im anschließenden Kapitel die Demonstration der erarbeiteten Ergebnisse. Dies geschieht durch die exemplarischen Anwendung verschiedener Datenreduktionsverfahren. Hierzu wird ein Fallbeispiel vorgestellt und Verfahren anhand den formulierten Kriterien und Empfehlungen ausgewählt und angewendet. Dies soll nicht nur die praktische Relevanz und Effektivität der zuvor diskutierten Verfahren demonstrieren, sondern auch deren Stärken und mögliche Limitationen in Anwendungsszenarien aufzeigen.

4 Exemplarische Anwendung verschiedener Datenreduktionsverfahren

Nachdem im vorangegangenen Kapitel Datenreduktionsverfahren identifiziert und analysiert wurden, werden nun die Ergebnisse anhand einer exemplarischen Anwendung verschiedener Datenreduktionsverfahren evaluiert. Dazu wird in Abschnitt 4.1 ein Fallbeispiel aus dem industriellen Bereich vorgestellt. In Abschnitt 4.2 werden anhand der Kategorisierung und Bewertungen der Datenreduktionsverfahren aus dem Abschnitt 3.3 Verfahren ausgewählt. In Abschnitt 4.3 folgt sodann die Durchführung des Fallbeispiels. Die Ergebnisse werden abschließend im Gesamtkontext der Arbeit diskutiert und ein Fazit gezogen.

4.1 Vorstellung des Fallbeispiels

Der SECOM-Datensatz, verfügbar auf der UCI Machine Learning Repository Plattform, dient als exemplarisches Fallbeispiel für die Anwendung von Datenreduktionsverfahren und stammt aus der Domäne der modernen Halbleiterfertigung. Dieser Datensatz umfasst eine Vielzahl von Signalen, die kontinuierlich während des Produktionsprozesses von Sensoren erfasst werden. Die Daten enthalten sowohl nützliche Informationen als auch irrelevante Daten und Störgeräusche. Jede Instanz repräsentiert eine einzelne Produktionsentität mit entsprechend gemessenen Merkmalen. Die verfügbaren Daten sind zweigeteilt. Zum einen besteht der Datensatz aus einer Matrix mit 591 Merkmale und 1567 Instanzen. Zum anderen enthält der Datensatz eine zugehörige Label-Datei. Die zugehörigen Labels geben ein einfaches Bestehen und Nicht-Bestehen der internen Produktionslinientests wieder. Ein Wert von -1 entspricht einem Bestehen des Tests, während ein Wert von 1 einem Nicht-Bestehen entspricht. Wie bei Datensätzen, die aus realen Produktionsumgebungen stammen, sind auch hier fehlende Werte präsent. Die Intensität der fehlenden Werte variiert über die verschiedenen Merkmale, was die Notwendigkeit von sorgfältigen Vorverarbeitungsschritten unterstreicht, um die Integrität und Verwertbarkeit der Daten zu gewährleisten. Die Datendarstellung in Rohdatentextdateien, bei denen jede Zeile einer Instanz entspricht und die Merkmale durch Leerzeichen getrennt sind, verdeutlicht die Notwendigkeit, die Daten sorgfältig zu laden und zu interpretieren. Die Platzhalter für fehlende Werte sind mit *NaN* gekennzeichnet. Die Anwendung von Datenreduktionsverfahren zielt in diesem Fallbeispiel darauf ab, die Anzahl der betrachteten Merkmale zu reduzieren, indem unwesentliche oder redundante Signale entfernt werden. Dies soll es Ingenieuren ermöglichen, diejenigen Signale zu identifizieren, die maßgeblich für Abweichungen in der Produktionsausbeute verantwortlich sind, um daraus zuverlässige Vorhersagemodelle zu trainieren.

4.2 Auswahl der Datenreduktionsverfahren

Nachdem in Abschnitt 4.1 das Fallbeispiel vorgestellt wurde, erfolgt nun die Auswahl der Datenreduktionsverfahren anhand der in Abschnitt 3.3 präsentierten Ergebnisse. Die Anzahl der Instanzen liegt mit etwas 1500 im normalen Bereich. Das Verhältnis von Instanzen zu Merkmalen beträgt etwa eins zu drei. Basierend auf den Erkenntnissen in Abschnitt 3.3.3 und der formulierten Aufgabenstellung, ist in diesem Fall eine Merkmalsreduktion dringend zu empfehlen. Da aus den Beschreibungen in Abschnitt 4.1 hervorgeht, dass es notwendig ist Ergebnisse auf einzelne kritische Signale zurückführen zu können, sollte nach den Darlegungen in Abschnitt 3.3 die Merkmalsauswahl der Dimensionsreduktion bevorzugt werden. Bei weiterer Prüfung der Kriterien des Abschnitts 3.1 lässt sich bei Betrachtung des Datensatzes formulieren, dass es sich um einen einfachen Datensatz, mit lediglich numerisch-kontinuierlichen Werten handelt. Zudem sind diskrete Label-Informationen binärer Art vorgegeben, womit es sich um eine überwachte Klassifikationsaufgabe handelt (vgl. Abschnitt 2.2.3). Nach Empfehlungen des Abschnitts 3.3.1 sollten demnach bei Möglichkeit auch überwachte Merkmalsauswahlsverfahren verwendet werden, da diese eine gezielte, aufgabenorientierte Reduktion der Daten erreichen. Weitere Forderungen zur Kriterienerfüllung gehen aus der Aufgabenstellung nicht hervor und sind somit frei wählbar. Bei Betrachtung der in Tabelle 3.2 zusammengefassten Ergebnisse kommen hierfür zunächst mehrere überwachte Merkmalsauswahlsverfahren für einfache Datensätze mit der Fähigkeit des Umgangs numerisch-kontinuierlicher Werte in Frage. Anhand den Empfehlungen in Abschnitt 3.3.2 sollten jedoch wenn möglich zur Datensatzgröße passende Verfahren gewählt werden. Die Größe des Merkmalsraums liegt mit knappen 600 Merkmalen zwar nicht im sehr großen, aber im großen Bereich. Daher kommen die moderat skalierbaren Verfahren in Frage. Durch diese Eingrenzung reduzieren sich die Verfahren der Merkmalsauswahl auf den *ReliefF*-Algorithmus. Der Tabelle 3.2 ist somit zudem vorab zu entnehmen, dass dieser Algorithmus eine Auswahlerfordernis aufzeigt, also die Anzahl der verbleibenden Merkmale für die weitere Modellierung selbst gewählt werden muss. Zudem enthält der Algorithmus Hyperparameter zur Initialisierung. Des Weiteren lässt sich der Tabelle entnehmen, dass das Verfahren modellunabhängig operiert und der Output somit für das Training und der Evaluation verschiedener Data Minnig Algorithmen sinnvoll ist.

Im gegebenen Kontext erfolgt die Selektion eines zusätzlichen Verfahrens zur Dimensionsreduktion, um die Bedeutung der Interpretierbarkeit der Ergebnisse zu verdeutlichen. In Analogie zur Auswahl des Merkmalsauswahlverfahrens *ReliefF* wird ein für numerisch-kontinuierliche Daten adäquates Dimensionsreduktionsverfahren ausgewählt, welches die Anforderungen der Datensatzgröße in Bezug auf die Größe erfüllt und überwacht durchgeführt wird. Unter Betrachtung der in Tabelle 3.3 präsentierten Resultate erweist sich die *Lineare Diskriminanzanalyse* (LDA) als geeignetes Verfahren. Ferner lässt sich der Tabelle entnehmen, dass keine Konfiguration von Hyperparametern erforderlich ist und der Output keine weiteren Auswahlprozesse erfordert.

Nachdem Verfahren anhand der Erkenntnisse des Abschnitts 3.3.1 ausgewählt wurde erfolgt nun die Durchführung des Fallbeispiels, in dem die in der Theorie formulierten Kriterien sowie Kriterien-Ausprägungen in der Praxis demonstriert werden.

4.3 Durchführung des Fallbeispiels

Zunächst werden die Spezifikationen und die genutzte Softwareumgebung dargelegt, um eine transparente und reproduzierbare Forschungsgrundlage zu gewährleisten. Die Implementierung und Ausführung der Datenreduktionsverfahren erfolgten in einer Python-Umgebung mit folgender Bibliotheken:

- Python 3.12.1
- pandas 1.5.3
- NumPy 1.21.4
- scikit-learn 1.0.2
- skrebate 0.62

Bevor nun die Datenreduktionsverfahren angewendet werden erfolgt zunächst die Datenvorverarbeitung.

4.3.1 Datenvorverarbeitung

Vor der Anwendung der Datenreduktionsverfahren ist es erforderlich die Daten entsprechend ihrer vielen fehlenden Werte vorzubereiten. Die Anzahl der fehlenden Werte variiert stark innerhalb des Datensatzes, was bedeutet, dass das Löschen von Instanzen mit fehlenden Werten aufgrund des bereits geringen Verhältnisses von Merkmalen zu Instanzen und der hohen Anzahl von fehlenden Werten nicht zielführend ist. Das Löschen aller Merkmale mit fehlenden Werten ist aufgrund der Verteilung der fehlenden Werte über verschiedene Merkmale nicht sinnvoll. Daher wird ein pragmatischer Ansatz verfolgt, der darauf abzielt, Merkmale beizubehalten die weniger als 50 Prozent fehlenden Werten haben und diese durch eine einfache Mittelwertimputation aufzufüllen und alle Merkmale mit mehr als 50 Prozent fehlenden Werten zu löschen. Zur Imputation wird die Funktion *SimpleImmputer* verwendet mit der Strategie *mean*. Auf diese Weise werden die Merkmale auf 562 reduziert.

Wie in Tabelle 3.2 dargelegt, handelt es sich beim *ReliefF*-Algorithmus um ein ähnlichkeitsbasiertes Merkmalsauswahlsverfahren. Dementsprechend ist es nach den Darlegungen in Abschnitt 3.3.1 notwendig die Merkmale zu normalisieren (vgl. Abschnitt 2.2.2). Ebenso ist dies für die Anwendung der *LDA* notwendig (vgl. Abschnitt 3.3.1). Hierfür wird die Funktion *MinMaxScaler* verwendet. Algorithmus 1 stellt die Python-Implementierung zur Datenvorverarbeitung dar. Dieser beinhaltet zudem das Importieren, der für die Vorverarbeitung der Daten notwendige Bibliotheken, als auch das Laden der Daten selbst.

Algorithm 1 Datenvorbverarbeitung des SECOM-Datensatzes

```
Importieren der Bibliotheken:
import pandas as pd
from sklearn.impute import SimpleImputer
Daten laden:
data = pd.read_csv('secom.data', header=None, delim_whitespace=True)
labels = pd.read_csv('secom_labels.data', header=None, usecols=[0],
delim_whitespace=True)
Berechnung des Prozentsatzes fehlender Werte je Spalte:
missing_percentage = data.isnull().mean() * 100
Entfernen von Spalten mit mehr als 50% fehlenden Werten:
columns_to_drop = missing_percentage[missing_percentage > 50].index
data_cleaned = data.drop(columns=columns_to_drop)
Anwendung der Mittelwert-Imputation auf verbleibende Spalten:
imputer = SimpleImputer(strategy='mean')
data_imputed = imputer.fit_transform(data_cleaned)
Umwandlung der imputierten Daten zurück in ein DataFrame:
data_imputed_df = pd.DataFrame(data_imputed, columns=data_cleaned.columns)
Daten normalisieren:
scaler = MinMaxScaler()
data_normalized = scaler.fit_transform(data_imputed_df)
```

Da der Fokus auf der Anwendung der Datenreduktionsverfahren liegt, wird von weiteren Vorverarbeitungsschritten, wie der zur Rauschbehandlung, abgesehen und es folgt die Anwendung der Datenreduktionsverfahren.

4.3.2 Anwendung der Datenreduktionsverfahren

Im folgenden wird zunächst der *ReliefF* angewendet und anhand der formulierten Kriterien sowie Ausprägungen geprüft. Anschließend folgt die Anwendung der *LDA*.

ReliefF

Nachdem der Datensatz nun vorverarbeitet wurde erfolgt nun die Anwendung des *ReliefF*-Algorithmus. Wie bereits in Abschnitt 4.2 beschrieben und in Tabelle 3.2 dargestellt, benötigt der Algorithmus die Einstellung eines Hyperparameters. Der Algorithmus arbeitet, indem er für jedes Merkmal einen Score berechnet, der auf dem Grad basiert, zu dem das Merkmal dazu beiträgt ähnliche Beispiele näher und unähnliche Beispiele weiter voneinander zu trennen. Der zu konfigurierende Hyperparamter ist daher die Anzahl der nächsten Nachbarn (vgl. Abschnitt 2.2.3). Da die optimale Konfiguration von Hyperparamtern nicht Gegenstand dieser Arbeit ist, wird diese daher auf zehn gesetzt, da dies anhand der untersuchten Literatur eine gängige Spezifikation von nächsten Nachbarn bei Datensätzen ähnlicher Größe ist.

Nach Anwendung des Verfahrens resultiert der Output, wie Tabelle 3.2 zu entnehmen, mit

einer Auswahlerfordernis. Dementsprechend erhält man keinen reduzierten Datensatz, sondern mit Wichtigkeitsscore bewertete Merkmale. Im Falle des *ReliefF*-Algorithmus basiert die Berechnung des Wichtigkeitsscores auf der Differenz zwischen den durchschnittlichen Distanzen zu den nächsten Nachbarn derselben Klasse und den nächsten Nachbarn der gegnerischen Klasse. Ein positives Score-Ergebnis zeigt an, dass das Merkmal effektiv zwischen den Klassen unterscheidet. Ein negatives Score-Ergebnis deutet hingegen darauf hin, dass das Merkmal weniger effektiv differenziert oder sogar negativ zur Klassenunterscheidung beiträgt. Wie in Abschnitt 3.2.1 beschrieben werden hierbei verschiedene Methoden angewendet, um geeignete Merkmale beizubehalten. Wie beschrieben, ist ein gängiges Vorgehen die Scores aufsteigend beziehungsweise absteigend zu visualisieren und nach einem *Ellenbogen* im Plot zu suchen. Die Implementierung dieses Vorgehens nach der Datenvorverarbeitung ist Algorithmus 2 zu entnehmen. Die nach Wichtigkeit sortierten Merkmale sind Abbildung 4.1 zu entnehmen.

Algorithm 2 Merkmalsauswahl mit ReliefF und Visualisierung der Merkmalswichtigkeit

```
1: ReliefF instanziieren und anpassen:
2: relief = ReliefF(n_neighbors=10)        ▷ Setze die Anzahl der Nachbarn auf 10
3: relief.fit(data_normalized, labels.values.ravel())
4: Feature-Wichtigkeiten und Rankings abrufen:
5: feature_scores = relief.feature_importances_
6: sorted_scores = np.sort(feature_scores)[::-1]
7: Visualisierung der Merkmals-Wichtigkeiten:
8: plt.figure(figsize=(15, 6))
9: plt.bar(range(len(sorted_scores)), sorted_scores, color='dodgerblue')
10: plt.xlabel('Rang der Merkmalswichtigkeit')
11: plt.ylabel('Wichtigkeitsscore')
12: plt.show()
```

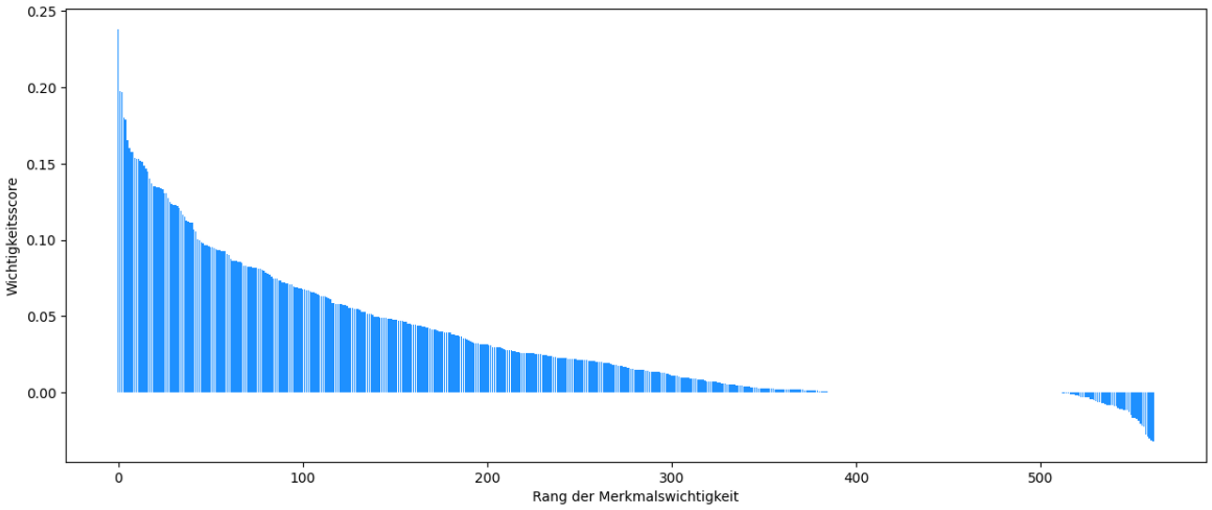


Abbildung 4.1: Merkmalsrang nach Wichtigkeitsscore nach Anwendung von ReliefF

Inwiefern der Datensatz nun reduziert wird hängt somit von der Entscheidung ab, wie viele der ursprünglichen Merkmale beibehaltet werden. Die Anzahl der Instanzen bleibt unverändert. Die semantische Integrität der Merkmale bleibt erhalten. Die Ausprägung

des Kriteriums der Auswahlerfordernis, das in diesem Verfahren als *ja* eingestuft ist, kann wie in Abschnitt 3.2.1 beschrieben, sowohl als Vorteil als auch zeitlicher Nachteil ausgelegt werden. Zum einen haben Anwender nun die Freiheit zu entscheiden wie viele der Merkmale für das anschließende Data Mining beibehaltet werden sollen und so gegebenenfalls hochwertigere Data Mining Modelle zu erstellen, zum anderen hat das zeitliche Auswirkungen auf das Projekt, da sich dadurch weitere Analysen der Daten ergeben, sofern mehrere Merkmalskombinationen verwendet werden. Exemplarisch könnten sich Anwender nun Merkmale mit einem Wichtigkeitsscore über 0.15 extrahieren, um die signifikantesten Merkmale im Zusammenhang mit der Klassifikationsaufgabe zu identifizieren. Die Ergebnisse sind in Tabelle 4.1 und Abbildung 4.2 dargestellt.

Tabelle 4.1: Wichtigkeitsscores der Merkmale über 0,15 nach Anwendung von ReliefF

Merkmal	Score	Merkmal	Score
55	0.152814	126	0.178851
58	0.180027	175	0.151228
59	0.238079	255	0.157276
64	0.197454	305	0.151585
65	0.196501	385	0.165312
76	0.157850	435	0.153152
78	0.159818	515	0.153692

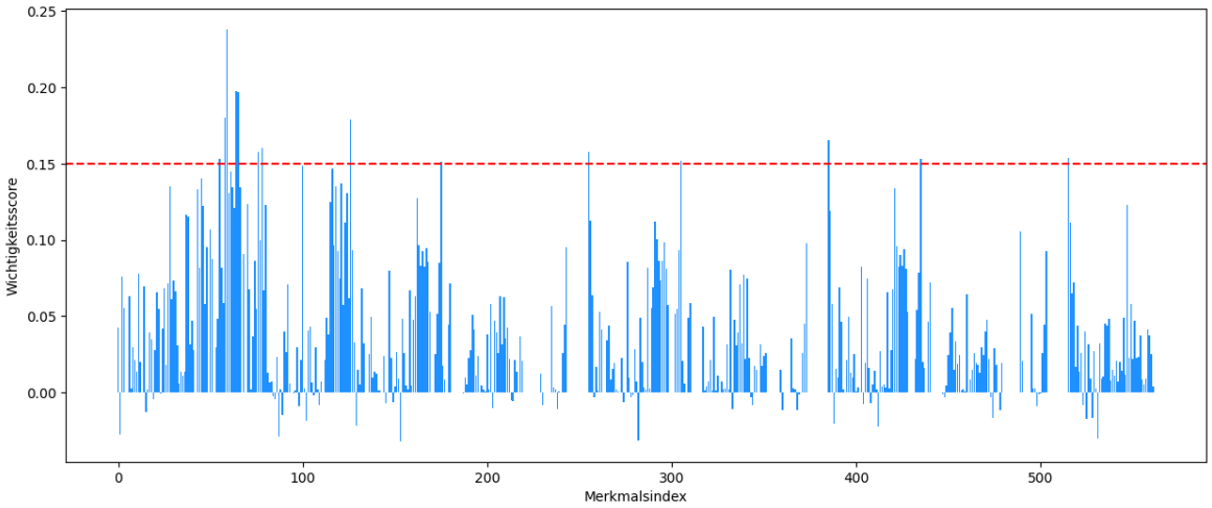


Abbildung 4.2: Merkmalsindex nach Wichtigkeitsscore nach Anwendung von ReliefF

Im praktischen Einsatz des Algorithmus ließen sich somit alle, im Zusammenhang des Fallbeispiels untersuchten, in Abschnitt 3.3.1 theoretisch formulierten Kriterien sowie Ausprägungen demonstrieren und bestätigen. Darüber hinaus ermöglichte die Anwendung eine Darlegung der Auswirkungen dieser Kriterien, wie sie in Abschnitt 3.2.3 erörtert wurden. Anschließend erfolgt nun die exemplarische Anwendung der *LDA*

LDA

Nach Implementierung des *ReliefF*-Algorithmus wird nun die *LDA* angewendet, um das Kriterium der Interpretation der resultierenden Merkmale sowie der nicht vorliegenden Auswahlerfordernis darzulegen. Gemäß den Ausführungen in Abschnitt 3.3.1 vollzieht die *LDA* eine Reduktion der Daten, indem eine lineare Transformation angewandt wird, welche die Daten auf einen Raum geringerer Dimension projiziert. Dies erfolgt unter Maximierung der Klassenseparation und Minimierung der Varianz innerhalb der Klassen. Die Anzahl der dabei entstandenen Dimensionen entsprechen dabei der Anzahl der Klassen minus eins. Im gegebenen Fallbeispiel resultiert bei zwei Klassen eine eindimensionale Darstellung des Datensatzes. Die Anzahl der Merkmale des Datensatzes reduziert sich demnach auf ein einziges Merkmal. Die Anzahl der Instanzen bleibt konstant. Abbildung 4.3 zeigt die Instanzen des Datensatzes, die auf die erste lineare Diskriminanzkomponente (LDA-1) projiziert wurden. Die x-Achse repräsentiert die Werte der ersten Diskriminantenkomponente, die eine Linearkombination der ursprünglichen Merkmale darstellt und so konstruiert ist, dass sie die Varianz zwischen den Klassen maximiert. Die Farbskala auf der rechten Seite repräsentiert die Klassenzugehörigkeit, wobei jede Farbe einer Klasse entspricht.

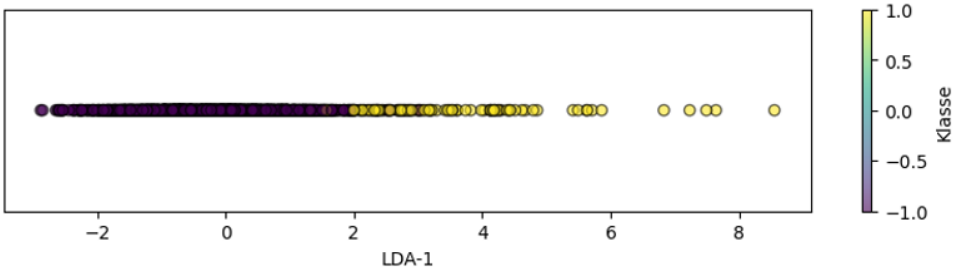


Abbildung 4.3: Darstellung der Diskriminanzkomponente nach Anwendung der LDA

Der Output ist somit im Vergleich zum *ReliefF*-Algorithmus, wie in Abschnitt 3.3.3 erörtert, nicht unmittelbar interpretierbar, da dieser nicht auf einzelne Merkmale zurückzuführen ist. Demnach ist eine direkte Rückschluss auf die Relevanz individueller Merkmale nicht gegeben. Für das gegebene Fallbeispiel, die eine Rückführung der Data-Mining-Ergebnisse auf einzelne Merkmale erfordert, ist die Anwendung des Verfahrens nicht direkt zielführend, was wiederum die Notwendigkeit des in Abschnitt 3.2.1 dargelegten Kriterium der Interpretierbarkeit unterstreicht.

Falls eine Rückführung auf die einzelnen Merkmale nicht erforderlich ist, leistet die *LDA* jedoch gemäß Abbildung 4.3 eine gute Trennung der Klassen. Um die Empfehlungen aus Abschnitt 3.3.3 zu evaluieren, welche die Bevorzugung von überwachten Reduktionsverfahren in überwachten Lernaufgaben nahelegen, wird nun exemplarisch das unüberwachte Dimensionsreduktionsverfahren *PCA* (vgl. Abschnitt 3.3.1) angewendet. Bei Bewahrung einer in der Literatur gängigen Empfehlung von 95 Prozent der Varianz, reduziert sich hierbei der Datensatz auf elf Hauptkomponenten, welche als reduzierter Datensatz in elf Merkmalen resultiert. Auf die reduzierten Datensätze werden nun fünf gängige Klassifikationsmodelle trainiert und hinsichtlich der Klassifikationsgenauigkeit evaluiert. Diese misst, wie in Abschnitt 2.3.2 beschrieben, den Prozentsatz der korrekt klassifizierten Instanzen in einem Klassifizierungsproblem und quantifiziert somit die Fähigkeit des Modells zur richtigen Vorhersage der Klassenlabels. Der Datensatz wird hierzu im Verhältnis von drei

zu eins aufgeteilt. Drei viertel des Datensatzes werden für das Training genutzt und ein viertel zum Testen. Die Ergebnisse sind in Abbildung 4.4 dargestellt.

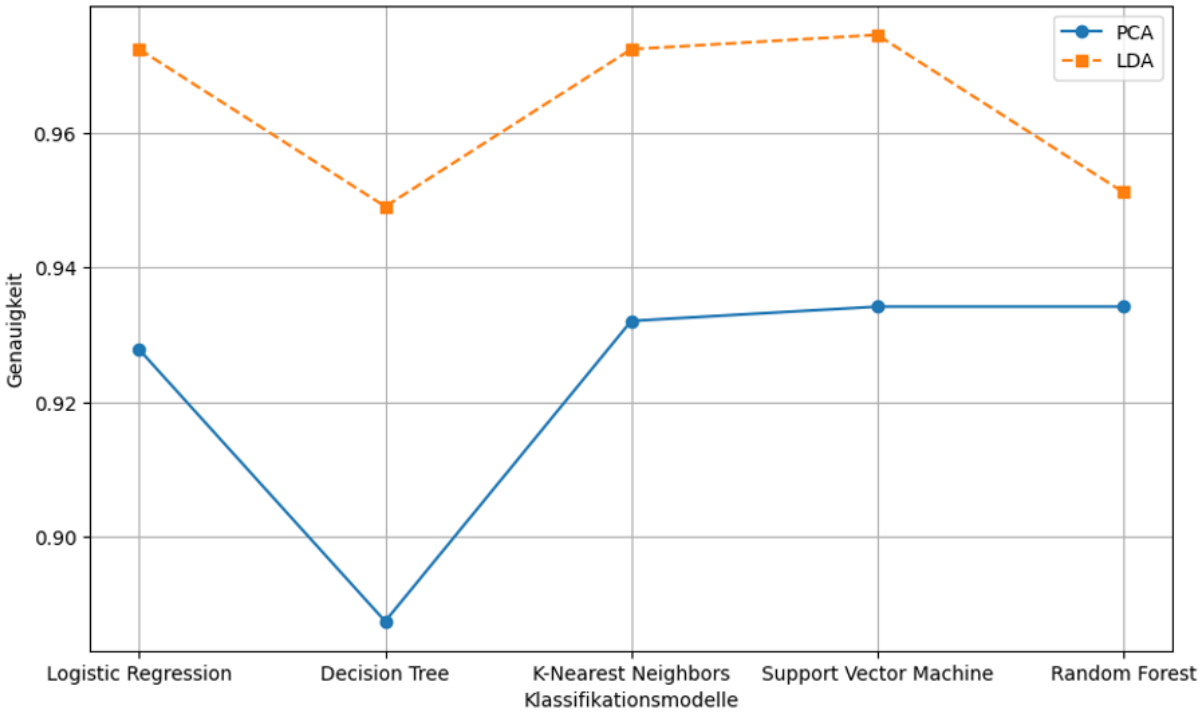


Abbildung 4.4: Gegenüberstellung der Klassifikationsgenauigkeit unter Anwendung von PCA und LDA über verschiedene Klassifikationsmodelle

In der Grafik wird ersichtlich, dass durch Reduktion der Daten mittels *LDA* eine konsistent höhere Klassifikationsgenauigkeit als bei einer Reduktion der Daten mittels *PCA* über verschiedene Klassifikationsmodelle erzielt wird. Mit einer Klassifikationsgenauigkeit von über 94 Prozent über mehrere Modelle hinweg, zeigt die *LDA* somit trotz der signifikaten Reduktion der Daten eine hohe Genauigkeit. Die Ergebnisse bestätigen die Empfehlung, dass überwachte Reduktionsverfahren für überwachte Aufgaben zu bevorzugen sind, da sie eine gezieltere Reduktion der Daten im Zusammenhang der Data-Minnig-Aufgabe ermöglichen. Es muss jedoch angemerkt werden, dass diese Schlussfolgerungen aufgrund der Beschränkung auf einen einzelnen Datensatz nicht generalisiert werden können. Dennoch kann dies als Indikator der Empfehlungen dienen. Zudem demonstriert dies, dass trotz einer signifikanten Reduktion der Daten mittels *LDA* auf lediglich ein Merkmal eine hohe Klassifikationsgenauigkeit erzielt werden kann. Anwender sollten demgemäß, wie in den Empfehlungen formuliert, zunächst prüfen, ob die Rückführbarkeit auf einzelne Merkmale von Bedeutung ist. Ist dies nicht der Fall, kann eine starke Reduktion der Daten durch Dimensionsreduktion erreicht und dennoch gute Klassifikationsgenauigkeiten erzielt werden. Durch das dargestellte Beispiel konnten somit ebenfalls die untersuchten Kriterienausprägungen und Empfehlungen bestätigt und demonstriert werden.

4.4 Diskussion und Fazit

Die Entwicklungen dieser Arbeit zielen darauf ab, eine Lücke in der Literatur zur Datenvorverarbeitung für das Data Mining und der Wissensentdeckung in Datenbanken zu schließen. Es wurde dargelegt, dass die Anzahl an Datenreduktionsverfahren stetig zunimmt und unübersichtlicher wird. Anwender stehen zunehmend vor der Herausforderung, sich in der umfangreichen Literatur der Verfahren zurechtzufinden. Zudem steigt die Komplexität der Verfahren, wodurch neben der Menge an Verfahren, auch die technische Komplexität Anwender vor die Herausforderung stellt, Verfahren zu identifizieren, die für verschiedene Kontexte des Data Minings anwendbar, beziehungsweise sinnvoll anwendbar sind. Klassische, in der Literatur dokumentierte Kategorisierungen zu Datenreduktionsverfahren geben Aufschluss über verschiedene technische Aspekte der Verfahren, ermöglichen Anwendern jedoch keinen praktikablen Zugang zu den Verfahren. Aufgrund der interdisziplinären Natur des Data Minings und der Wissensentdeckung in Datenbanken, sowie der Vielzahl an Techniken, die in diesem Rahmen angewendet werden, ergeben sich schiere Anzahlen an technischen Herangehensweisen und Konzepten, deren übergreifendes Verständnis selbst versierte Experten vor Herausforderungen stellt.

Um ein Fazit zu der vorliegenden Arbeit ziehen zu können, wird diese anhand ihrer Zielerreichung diskutiert. Hierfür galt es zunächst als grundlegend, Datenreduktionsverfahren zu identifizieren, zu strukturieren und aus einer rein praktikablen Perspektive zu differenzieren. Zu diesem Zweck wurde eine systematische Literaturrecherche durchgeführt und Veröffentlichungen untersucht, die sich mit Datenreduktionsverfahren auseinandersetzten. Daraus resultierte der erste wesentliche Beitrag dieser Arbeit. Die Identifikation und Aggregation von Kriterien in Gruppen des Inputs, Outputs und der Prozedur hilft dabei, die Komplexität der Datenreduktionsverfahren zu reduzieren und macht die Unterschiede zwischen den Verfahren aus praktikabler Sicht transparenter. Dadurch wird deutlich, welche Anforderungen und Überlegungen jeweils in Bezug zur Anwendung von Datenreduktionsverfahren beachtet werden sollten. Durch die darauf aufbauende Kontextualisierung der Kriterien zu Datenreduktionsverfahren in Wissensentdeckungsprozesse konnten zudem Interdependenzen zu anderen Phasen der Wissensentdeckung sowie zeitlichen Rahmenbedingungen aufgedeckt werden, sodass die Auswirkungen kontextspezifisch verschärft und eine gezieltere Eignung von Verfahren ermöglicht.

Anschließend erfolgte die Entwicklung von Kriterienausprägungen, um eine solide Grundlage für die Einordnung und Bewertung von Datenreduktionsverfahren zu schaffen. Es ist jedoch anzumerken, dass hierbei Kritikpunkte zu erwähnen sind. Insbesondere die Ausprägungen bezüglich der Datensatzarten erscheinen recht undifferenziert. In Abschnitt 2.1.2 wurden verschiedene Unterscheidungen zu Datensätzen vorgenommen, jedoch ließ sich lediglich eine methodenübergreifende, generalisierte Einteilung in einfache und komplexe Datensätze vornehmen. Trotz dieser Herausforderungen ermöglicht die in dieser Arbeit vorgenommene Einteilung dennoch eine differenzierte Betrachtung von Datensätzen, um die Anwendbarkeit von Datenreduktionsverfahren sicherzustellen.

Anschließend wurden identifizierte Datenreduktionsverfahren anhand der Kriterien eingeordnet und bewertet. Als positiv hervorzuheben ist ebenfalls, dass diese Perspektiven das Verständnis der Transition aus der technischen Perspektive fördern und es den Anwendern ermöglichen, vorab Rückschlüsse auf gewisse Kriterienausprägungen zu ziehen. Auch wenn diese Perspektiven oft nur vereinzelte Implikationen auf Ausprägungen zulassen und letzt-

endlich die eindeutige Einordnung auf Verfahrensebene stattfindet, fördern diese dennoch das Verständnis. Die vorgenommene, übersichtliche Einordnung und Bewertung der Verfahren erlaubt es schließlich Anwendern, Verfahren für verschiedene Kontexte auszuwählen. Da nicht für jede Kombination der Kriterienausprägungen Verfahren eingeordnet wurden, was als Kritikpunkt angeführt werden könnte. Dennoch kann das Ziel, die Analyse und Einordnung der Verfahren vorzunehmen, insgesamt als erfolgreich betrachtet werden. Der Kategorisierungsrahmen hinsichtlich der Kriterien bietet einen vielversprechenden Ansatz für zukünftige Forschung und Entwicklung, da er flexibel genug ist, um für verschiedene Verfahren erweitert zu werden.

Abschließend wurde eine exemplarische Anwendung von Datenreduktionsverfahren durchgeführt, um die Ergebnisse zu evaluieren und das Verständnis für ihre Anwendung anhand eines konkreten Fallbeispiels zu vertiefen. Die formulierten Ergebnisse des vorangegangenen Kapitels spielten dabei eine wesentliche Rolle, indem sie maßgeblich zum Verständnis und zur Auswahl geeigneter Datenreduktionsverfahren beitrugen. Durch das gewählte Fallbeispiel und die angewendeten Verfahren konnte erfolgreich die Bestätigung aller für die Verfahren zugeordneten Kriterienausprägungen erreicht werden.

Schließlich lässt sich das Fazit ziehen, dass die vorliegende Arbeit einen Beitrag zum Verständnis und dem praktikablen Umgang mit Datenreduktionsverfahren als Vorbereitung für das Data Mining im Kontext der Wissensentdeckung in Datenbanken leistet. Die entwickelten Kriterien vermitteln Anwendern ein Verständnis für zu berücksichtigende Eigenschaften der Datenreduktionsverfahren. Die vorgenommene Einordnung und Bewertung bietet eine übersichtliche Zusammenfassung der Ergebnisse die für eine Auswahl von Verfahren herangezogen werden kann und ist zudem flexibel, sodass weitere Verfahren in diese eingeordnet und bewertet werden können. Jedoch sind erweiterte Untersuchungen zur Differenzierung hinsichtlich der Ausprägungen verschiedener Kriterien sowie Verfahren sinnvoll.

5 Zusammenfassung und Ausblick

Im Verlauf dieser Arbeit wurden Datenreduktionsverfahren als Vorbereitung für Data Mining im Kontext der Wissensentdeckung in Datenbanken systematisch analysiert um diese anhand verschiedener Kriterien einordnen und bewerten zu können und so Anwendern ein Übersicht zur Eignung verschiedenen Methoden und Verfahren für unterschiedliche Kontexte zur Verfügung zu stellen.

Das einführende Kapitel dieser Arbeit diente dazu, essenzielle technische und kontextuelle Grundlagen der Datenreduktion zu erläutern. Dies umfasste eine grundlegende Betrachtung von Datenbanken sowie eine differenzierte Darlegung von Datensätzen sowie ihrer Eigenschaften. Des Weiteren wurde ein grundlegendes Verständnis von Wissensentdeckungsprozessen in Datenbanken vermittelt, einschließlich der Phasen der Datenvorverarbeitung und des Data Minings. Die Grundlagen zur Datenreduktion wurden erörtert, wobei verschiedene Methoden und repräsentative Verfahren vorgestellt wurden, um ein einheitliches Verständnis für die weiteren Ausführungen zu schaffen.

Im Hauptteil erfolgte eine eingehende Analyse zu Datenreduktionsverfahren mittels einer systematischer Literaturrecherche. Hierfür wurden Studien zu Datenreduktionsverfahren analysiert und Kriterien für die Auswahl von Datenreduktionsverfahren entwickelt. Nachdem diese im Kontext der Wissensentdeckung in Datenbanken betrachtet wurden folgte die Festlegung von Kriterien-Ausprägungen. Anschließend wurden identifizierte Datenreduktionsverfahren der untersuchten Studien strukturiert und anhand den Kriterien bewertet. Die Ergebnisse wurden übersichtlich in tabellarischer Form zusammengefasst und können so als Unterstützung zur Auswahl von Datenreduktionsverfahren herangezogen werden. Die erarbeiteten Ergebnisse wurden durch eine exemplarische Anwendungen von Datenreduktionsverfahren auf ein konkretes Anwendungsbeispiel demonstriert, woraufhin eine eingehende Diskussion und ein abschließendes Fazit folgten.

Ein Ausblick auf mögliche zukünftige Forschungsarbeiten in diesem Bereich zeigt auf, dass erweiterte Untersuchungen zur Differenzierung der Kriterien-Ausprägungen sowie eine Einordnung und Bewertung weiterer Verfahren von Interesse sein könnten.

Literatur

Abdi, H. und L.J. Williams (2010). „Principal component analysis“. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4, S. 433–459.

Aggarwal, Charu C. (2013). *Outlier Analysis* -. Berlin Heidelberg: Springer Science und Business Media. ISBN: 978-1-461-46396-2.

– (2015). *Data Mining - The Textbook*. Berlin, Heidelberg: Springer. ISBN: 978-3-319-14142-8.

Agrawal, Rakesh, Tomasz Imieliński und Arun Swami (1993). „Mining association rules between sets of items in large databases“. In: *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*. ACM, S. 207–216.

Alpaydin, Ethem (2020). *Introduction to Machine Learning, fourth edition*. Cambridge: MIT Press. ISBN: 978-0-262-35806-4.

Anand, S.S., A.G. Büchner und Financial Times Management (1998). *Decision Support Using Data Mining*. Financial Times management briefings: Information technology. Financial Times Management. ISBN: 9780273632696.

Azevedo, Ana und Manuel Filipe Santos (2008). „KDD, SEMMA and CRISP-DM: A Parallel Overview“. In: *IADIS European Conference Data Mining*, S. 182–185.

Bengio, Yoshua, Aaron Courville und Pascal Vincent (2013). *Representation Learning: A Review and New Perspectives*. IEEE Transactions on Pattern Analysis und Machine Intelligence.

Berkhin, Pavel (2006). „A Survey of Clustering Data Mining Techniques“. In: *Grouping Multidimensional Data*, S. 25–71.

Bethlehem, Jelke (2009). *Applied Survey Methods - A Statistical Perspective*. New York: John Wiley und Sons. ISBN: 978-0-470-49498-1.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Berlin, Heidelberg: Springer. ISBN: 978-0-387-31073-2.

Bolón-Canedo, Verónica und Beatriz Remeseiro (2019). „Recent advances and emerging challenges in feature selection for machine learning: a survey“. In: *Neurocomputing* 335, S. 191–200.

Brachman, R. J. und T. Anand (1996). „The process of knowledge discovery in databases“. In: *Advances in Knowledge Discovery and Data Mining*. Hrsg. von U. M. Fayyad et al. In Fayyad, U. M. et al. (Eds.) AAAI Press / The MIT Press.

Cabena, Peter, Pablo Hadjinian, Rolf Stadler, Jaap Verhees und Alessandro Zanasi (1998). *Discovering Data Mining - From Concept to Implementation*. USA: Prentice-Hall, Inc. ISBN: 0137439806.

Cattell, Rick (2011). „Scalable SQL and NoSQL Data Stores“. In: *ACM SIGMOD Record* 39.4, S. 12–27.

Celko, Joe (2010). *Joe Celko's SQL for Smarties - Advanced SQL Programming*. Amsterdam: Elsevier. ISBN: 978-0-080-46004-8.

Chae, Bongsug (2014). „Insights from hashtag supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research". In: *International Journal of Production Economics* 165, S. 247–259.

Chandrashekar, Girish und Ferat Sahin (2014). „A survey on feature selection methods". In: *Computers and Electrical Engineering* 40.1, S. 16–28.

Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer und Rüdiger Wirth (2000). *CRISP-DM 1.0 - Step-by-step Data Mining Guide*. SPSS Inc.

Chaudhuri, S. und U. Dayal (1997). „An Overview of data warehousing and OLAP technology". In: *ACM Sigmod Record* 26.1, S. 65–74.

Chen, Hsinchun, Roger H.L. Chiang und Veda C. Storey (2012). „Business Intelligence and Analytics: From Big Data to Big Impact". In: *MIS Quarterly* 36.4, S. 1165–1188.

Chen, Peter Pin-Shan (März 1976). „The entity-relationship model—toward a unified view of data". In: *ACM Trans. Database Syst.* 1.1, S. 9–36. ISSN: 0362-5915.

Choudhary, Alok K., Jennifer A. Harding und Manoj Kumar Tiwari (2009). „Data mining in manufacturing: a review based on the kind of knowledge". In: *Journal of Intelligent Manufacturing* 20.5, S. 501–521.

Cios, Krzysztof, Anna Teresińska, S Konieczna, J Potocka und S Sharma (Juli 2000). „A knowledge discovery approach to diagnosing myocardial perfusion". In: *IEEE engineering in medicine and biology magazine : the quarterly magazine of the Engineering in Medicine and Biology Society* 19, S. 17–25.

Cios, Krzysztof J., Witold Pedrycz, Roman W. Swiniarski und Lukasz Kurgan (2008). *Data Mining - A Knowledge Discovery Approach*. Springer US. ISBN: 978-0-387-51317-1.

Cleve, Jürgen und Uwe Lämmel (2020). *Data Mining* -. Berlin: Walter de Gruyter GmbH und Co KG. ISBN: 978-3-110-67627-3.

Cochran, William Gemmell (1977). *Sampling Techniques* -. New York: Wiley. ISBN: 978-0-471-16240-7.

Codd, E. F. (1970). „A Relational Model of Data for Large Shared Data Banks". In.

Cukier, Kenneth und Viktor Mayer-Schoenberger (2013). „The Rise of Big Data: How It's Changing the Way We Think About the World". In: *Foreign Affairs* 92.3, S. 28–40. ISSN: 00157120.

Date, C. J. (2013). *An Introduction to Database Systems*.

– (2011). *SQL and Relational Theory - How to Write Accurate SQL Code*. Sebastopol: Ö'Reilly Media, Inc." ISBN: 978-1-449-31974-8.

Davenport, Thomas H und Jeanne Harris (2012). *Data Analytics for Corporate Debt Markets: Using Data for Investing, Trading, Capital Markets, and Portfolio Management*. FT Press Analytics.

De Backer, S., A. Naud und P. Scheunders (1998). „Non-linear dimensionality reduction techniques for unsupervised feature extraction“. In: *Pattern Recognition Letters* 19.8, S. 711–720. ISSN: 0167-8655.

Doan, AnHai, Alon Halevy und Zachary Ives (2012). *Principles of Data Integration*. Amsterdam: Elsevier. ISBN: 978-0-123-91479-8.

Efron, Bradley und R.J. Tibshirani (1994). *An Introduction to the Bootstrap* -. Boca Raton, Fla: CRC Press. ISBN: 978-0-412-04231-7.

Elmasri, Ramez und Shamkant B. Navathe (2020). *Fundamentals of Database Systems*. 8. Aufl. Pearson.

Esling, Philippe und Carlos Agon (2012). „Time-series data mining“. In: *ACM Computing Surveys (CSUR)* 45.1, S. 12.

Fahrmeir, Ludwig, Christian Heumann, Rita Künstler, Iris Pigeot und Gerhard Tutz (2016). *Statistik - Der Weg zur Datenanalyse*. 8. Aufl. Berlin Heidelberg New York: Springer-Verlag. ISBN: 978-3-662-50372-0.

Fayyad, Usama, Gregory Piatetsky-Shapiro und Padhraic Smyth (1996a). „Data Mining and Knowledge Discovery: Making Sense Out of Data“. In: *IEEE Expert* 11.5, S. 20–25.

Fayyad, Usama M, Gregory Piatetsky-Shapiro und Padhraic Smyth (1996b). „Advances in Knowledge Discovery and Data Mining“. In: *American Association for Artificial Intelligence*, S. 37–54.

Fukunaga, Keinosuke (1990). *Introduction to Statistical Pattern Recognition*. 2. Aufl. Academic Press.

Gao, Z., L. Xie, D. Tao und C. Zhang (2016). „Image Super-Resolution With Sparse Neighbor Embedding“. In: *IEEE Transactions on Image Processing* 21.7, S. 3194–3205.

García, Salvador, Joaquín Derrac, José Cano und Francisco Herrera (2012). „Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.3, S. 417–435.

García, Salvador, Julián Luengo und Francisco Herrera (2015). *Data Preprocessing in Data Mining*. Berlin, Heidelberg: Springer. ISBN: 978-3-319-10247-4.

Garcia-Laencina, Pedro J, Jose-Luis Sancho-Gomez und Anibal R Figueiras-Vidal (2010). „Pattern analysis and machine intelligence“. In: *IEEE Transactions on* 32.5, S. 925–938.

Guyon, Isabelle und André Elisseeff (2003). „An introduction to variable and feature selection“. In: *Journal of Machine Learning Research* 3, S. 1157–1182.

Haerder, Theo und Andreas Reuter (1983). „Principles of Transaction-Oriented Database Recovery“. In: *ACM Computing Surveys (CSUR)* 15.4, S. 287–317.

Haglin, David, Richard Roiger, Jon Hakkila und Timothy Giblin (Jan. 2005). „A tool for public analysis of scientific data“. In: *Data Science Journal* 4, S. 39–53.

Hair, Joseph F., Barry J. Babin, William C. Black und Rolph E. Anderson (2019). *Multivariate Data Analysis*. Cengage. ISBN: 978-1-473-75654-0.

Halevy, Alon, Anand Rajaraman und Joann Ordille (2006). „Data Integration: The Teenage Years“. In: *Proceedings of the 32nd international conference on Very large data bases*, S. 9–16.

Han, Jiawei, Micheline Kamber und Jian Pei (2011). *Data Mining: Concepts and Techniques*. Amsterdam: Elsevier. ISBN: 978-0-123-81480-7.

Harding, Jennifer A., Muhammad Shahbaz und Kusi Srinivas (2006). „Data Mining in Manufacturing: A Review“. In: *Journal of Manufacturing Science and Engineering* 128.4, S. 969–976.

Hastie, Trevor, Robert Tibshirani und Jerome Friedman (2009). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction, Second Edition*. Berlin Heidelberg: Springer Science und Business Media. ISBN: 978-0-387-84858-7.

He, X., D. Cai und P. Niyogi (2005). „Laplacian score for feature selection“. In: *In Advances in Neural Information Processing Systems*, S. 507–514.

Hernández, Mauricio A. und S. Stolfo (1998). „Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem“. In: *Data Mining and Knowledge Discovery* 2, S. 9–37.

Hinton, G.E. und S.T. Roweis (2002). „Stochastic Neighbor Embedding“. In: *Advances in Neural Information Processing Systems* 15.

Hipp, Jochen, Ulrich Güntzer und Gholamreza Nakhaeizadeh (2000). „Algorithms for association rule mining—a general survey and comparison“. In: *SIGKDD Explorations* 2.1, S. 58–64.

Hofmann, Thomas, Bernhard Schölkopf und Alexander J. Smola (2008). „Kernel methods in machine learning“. In: *The Annals of Statistics* 36.3, S. 1171–1220.

Jain, Anil K, M Narasimha Murty und Patrick J Flynn (1999). „Data clustering: A review“. In: *ACM computing surveys (CSUR)* 31.3, S. 264–323.

Jiang, B. und Y. Ren (2011). „ABiSEE: A binary search embedding algorithm for fast image retrieval“. In: *Journal of Visual Communication and Image Representation* 22.7, S. 610–621.

Jolliffe, I.T. (2006). *Principal Component Analysis*. Berlin Heidelberg: Springer Science und Business Media. ISBN: 978-0-387-22440-4.

Kelleher, John D. und Brendan Tierney (2018). *Data Science* -. Cambridge: MIT Press. ISBN: 978-0-262-53543-4.

Kemper, A. und A. Eickler (2015). *Datenbanksysteme - eine Einführung*.

Kim, J. und S. Choi (2021). „Deep Learning Approaches for Dimensionality Reduction: A Comprehensive Review“. In: *Expert Systems with Applications* 167, S. 114158.

Kitchenham, Barbara und Stuart Charters (Jan. 2007). „Guidelines for performing Systematic Literature Reviews in Software Engineering“. In: 2.

Kohavi, Ron und George H. John (1997). „Wrappers for feature subset selection“. In: *Artificial Intelligence*. Bd. 97. 1-2, S. 273–324.

Krcmar, Helmut (2015). *Informationsmanagement*. Berlin Heidelberg New York: Springer-Verlag. ISBN: 978-3-662-45863-1.

Kriegel, Hans-Peter, Peer Kröger, Jörg Sander und Arthur Zimek (2011). „Density-based Clustering“. In: *WIREs Data Mining and Knowledge Discovery*. Bd. 1. 3. Wiley Online Library, S. 231–240.

Kurgan, Lukasz A. und Petr Musilek (2006). „A Survey of Knowledge Discovery and Data Mining process models“. In: *The Knowledge Engineering Review* 21.1, S. 1–24.

Lee, A. und B. Kang (2018). „Advanced Techniques for Dimensionality Reduction and Data Visualization“. In: *Journal of Computing Sciences* 12.4, S. 450–460.

Lee, John A. und Michel Verleysen (2007). *Nonlinear Dimensionality Reduction*. Berlin Heidelberg: Springer Science und Business Media. ISBN: 978-0-387-39351-3.

Lenzerini, Maurizio (2002). „Data integration: A Theoretical Perspective“. In: *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, S. 233–246.

Leskovec, Jure, Anand Rajaraman und Jeffrey David Ullman (2020). *Mining of Massive Datasets*. Cambridge: Cambridge University Press. ISBN: 978-1-108-75131-5.

Little, Roderick J.A. und Donald B. Rubin (2019). „Statistical Analysis with Missing Data“. In.

Liu, Huan und Hiroshi Motoda (2002). *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers.

Liu, X. und Y. Wang (2022). „Efficient Feature Selection Methods for High-Dimensional Data: A Comprehensive Review“. In: *Knowledge-Based Systems* 232, S. 107599.

Lohr, Sharon L. (2009). *Sampling: Design and Analysis* -. Clifton Park, NY: Cengage Learning. ISBN: 978-0-495-10527-5.

Loshin, David (2013). *Big Data Analytics - From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph*. Amsterdam: Elsevier. ISBN: 012-4-186-645-.

Maaten, Laurens van der (2014). „Accelerating t-SNE using tree-based algorithms“. In: *Journal of Machine Learning Research* 15, S. 3221–3245.

Maaten, Laurens van der und Geoffrey Hinton (2008). „Visualizing data using t-SNE“. In: *Journal of Machine Learning Research* 9, S. 2579–2605.

Maimon, Oded und Lior Rokach (2010). *Data Mining and Knowledge Discovery Handbook*. Berlin Heidelberg: Springer Science, Business Media. ISBN: 978-0-387-09823-4.

Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh und Angela Hung Byers (Juni 2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. Techn. Ber. McKinsey Global Institute.

Martínez, A.M. und A.C. Kak (2001). „PCA versus LDA“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.2, S. 228–233.

Masaeli, M., G. Fung und J. G. Dy (2010). „From transformation-based dimensionality reduction to feature selection“. In: *In Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, S. 751–758.

Mika, Sebastian, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf und Klaus-Robert Mullers (1999). „Fisher discriminant analysis with kernels“. In: *IEEE Neural Networks for Signal Processing IX*, S. 41–48.

Mladenic, Dunja und Marko Grobelnik (1999). „Feature selection for Unbalanced Class Distribution and Naive Bayes“. In: *ICML*. Bd. 99, S. 258–267.

Murphy, Kevin P. (2012). *Machine Learning - A Probabilistic Perspective*. Cambridge: MIT Press. ISBN: 978-0-262-01802-9.

Nguyen, H. M., E. W. Cooper und K. Kamei (2014). „Borderline over-sampling for imbalanced data classification“. In: *International Journal of Knowledge Engineering and Soft Data Paradigms* 5.1, S. 3–29.

North, Klaus (2016). *Wissensorientierte Unternehmensführung - Wissensmanagement gestalten*. Berlin Heidelberg New York: Springer-Verlag. ISBN: 978-3-658-11643-9.

Peng, H., F. Long und C. Ding (2005). „Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8, S. 1226–1238.

Piatetsky-Shapiro, Gregory (1989). „Knowledge Discovery in Databases: Towards a Working Definition and an Answer to the Question "Why?"“ In: *Knowledge Discovery in Databases*, S. 1–20.

Pokorny, Jaroslav (2013). „NoSQL databases: a step to database scalability in web environment“. In: *International Journal of Web Information Systems* 9.1, S. 69–82.

Probst, Gilbert, Steffen Raub und Kai Romhardt (2012). *Wissen managen - Wie Unternehmen ihre wertvollste Ressource optimal nutzen*. 7. Aufl. Wiesbaden, Deutschland: Springer Gabler.

Pyle, Dorian (1999a). *Data Preparation for Data Mining*. San Francisco, Calif: Morgan Kaufmann. ISBN: 978-1-558-60529-9.

– (1999b). *Data Preparation for Data Mining*. San Francisco, Calif: Morgan Kaufmann. ISBN: 978-1-558-60529-9.

Rahm, Erhard und Philip A Bernstein (2001). „A survey of approaches to automatic schema matching“. In: *the VLDB Journal—The International Journal on Very Large Data Bases* 10.4, S. 334–350.

Reinartz, Thomas (2002). „A Unifying View on Instance Selection“. In: *Data Mining and Knowledge Discovery* 6, S. 191–210.

Robnik-Šikonja, M. und I. Kononenko (2003). „Theoretical and Empirical Analysis of ReliefF and RReliefF“. In: *Machine Learning* 53.1-2, S. 23–69.

Rodriguez, P. und L. Gutierrez (2019). „Feature Selection and Dimensionality Reduction Techniques for Machine Learning“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.7, S. 1785–1798.

Sadalage, Pramod J. und Martin Fowler (2012). *NoSQL Distilled - A Brief Guide to the Emerging World of Polyglot Persistence*. Amsterdam: Addison-Wesley. ISBN: 978-0-133-03612-1.

Särndal, Carl-Erik, Bengt Swensson und Jan Wretman (2003). *Model Assisted Survey Sampling* -. Berlin Heidelberg: Springer Science und Business Media. ISBN: 978-0-387-40620-6.

Schicker, Edwin (2017). *Datenbanken und SQL - Eine praxisorientierte Einführung mit Anwendungen in Oracle, SQL Server und MySQL*. Berlin Heidelberg New York: Springer-Verlag. ISBN: 978-3-658-16129-3.

Schölkopf, Bernhard, Alexander Smola und Klaus-Robert Müller (1998). „Nonlinear component analysis as a kernel eigenvalue problem“. In: *Neural Computation* 10.5, S. 1299–1319.

Schwab, Klaus (2017). *The Fourth Industrial Revolution* -. New York: Crown. ISBN: 978-1-524-75887-5.

Shekhar, Shashi und Sanjay Chawla (2015). *Spatial Databases - A Tour*. Prentice Hall.

Silberschatz, Abraham (2010). *Database System Concepts*. McGraw-Hill Publishing. ISBN: 978-0-077-41800-7.

Stevens, Stanley Smith (1946). „On the Theory of Scales of Measurement“. In: *Science* 103.2684, S. 677–680.

Strauch, Christof (2011). *NoSQL Databases*. http://www.christof-strauch.de/nosqldbs.pdf.

Sutton, Richard S. und Andrew G. Barto (2018). *Reinforcement Learning, second edition - An Introduction*. Cambridge: MIT Press. ISBN: 978-0-262-03924-6.

Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne und Vipin Kumar (2019). *Introduction to Data Mining*. München: Pearson. ISBN: 978-0-133-12890-1.

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

Thompson, Steven K. (2012). *Sampling* -. New York: John Wiley und Sons. ISBN: 978-1-118-16294-1.

Tiwari, Shashank (2011). *Professional NoSQL*. New York: John Wiley und Sons. ISBN: 978-1-118-16780-9.

Tukey, John Wilder (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley Publishing Company. ISBN: 978-0-201-07616-5.

Tyrer, Stephen und Bob Heyman (2016). „Sampling in epidemiological research: issues, hazards and pitfalls“. In: *BJPsych Bulletin* 40.2, S. 57–60.

Van Der Maaten, Laurens, Eric Postma und Jaap Van den Herik (2009). „Dimensionality reduction: a comparative review“. In: *Journal of Machine Learning Research* 10, S. 66–71.

Wilson, Dennis L., Tony R. Martinez und Robert Holte (2000). „Reduction Techniques for Instance-Based Learning Algorithms“. In: *Machine-mediated Learning* 38.3, S. 257–286.

Wirth, Rüdiger und Jochen Hipp (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

Witten, Ian H., Eibe Frank, Mark A. Hall und Christopher J. Pal (2016). *Data Mining - Practical Machine Learning Tools and Techniques*. 4. Aufl. San Francisco, Calif: Morgan Kaufmann. ISBN: 978-0-128-04357-8.

Xu, Rui und Don Wunsch (2005). *Clustering* -. New York: IEEE Press Series on Computational Intelligence.

Xu, Rui und Donald Wunsch (2015). „A comprehensive survey of clustering algorithms“. In: *Annals of Data Science* 2.2, S. 165–193.

Yang, P., Z. Zhang, B. B. Zhou und A. Y. Zomaya (2011). „A review of ensemble methods in bioinformatics“. In: *Current Bioinformatics* 5.4, S. 296–308.

Yu, L. und H. Liu (2003). „Feature selection for high-dimensional data: A fast correlation-based filter solution“. In: *In Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, S. 856–863.

Zhou, Wengang, Houqiang Li und Qi Tian (2018). „Deep Learning for Content-Based Image Retrieval: A Comprehensive Study“. In: *Proceedings of the ACM on Multimedia Conference*.

Zhu, Xiaojin und Andrew B. Goldberg (2009). „Introduction to Semi-Supervised Learning“. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3.1, S. 1–130.

A Anhang

A.1 Literaturtabelle der systematischen Literaturrechere

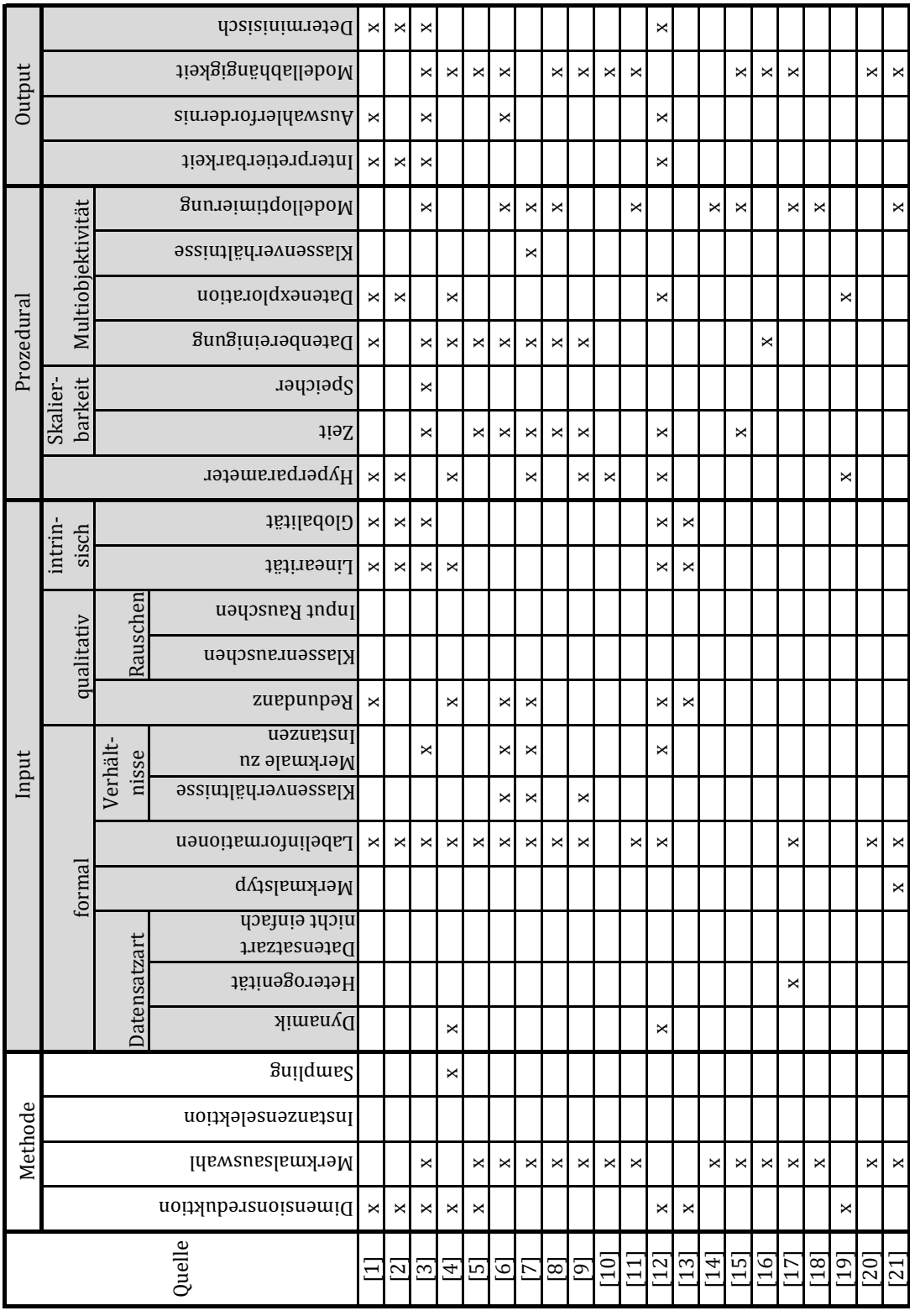
Index	Titel	Autor	Jahr
[1]	Overview and comparative study of dimensionality reduction techniques for high dimensional data	Ayeshaa et al.	2020
[2]	Analysis of Dimensionality Reduction Techniques on Big Data	Ray et al.	2020
[3]	Various dimension reduction techniques for high dimensional data analysis: a review	Papla et. Al	2021
[4]	Review of classical dimensionality reduction and sample selection methods for large-scale data processing	Xu et al.	2018
[5]	Feature selection methods on gene expression microarray data for cancer classification: A systematic review	Alhenawi et al.	2022
[6]	A Review of the Modification Strategies of the Nature Inspired Algorithms for Feature Selection Problem	Khurma et al.	2022
[7]	Enhanced whale optimization algorithm for medical feature selection: A COVID-19 case study	Nadimi-Shahraki et al.	2022
[8]	A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction	Pudjihartono et al.	2022
[9]	Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019)	Agrawal et al.	2021
[10]	Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance	Bagherzadeh et al.	2021
[11]	Supervised feature selection techniques in network intrusion detection: A critical review	Di Mauro et al.	2021
[12]	Toward a Quantitative Survey of Dimension Reduction Techniques	Espadato et al.	2021
[13]	A Review of Principal Component Analysis Algorithm for Dimensionality Reduction	Hasan et al.	2021
[14]	A novel hybrid whale optimization algorithm with flower pollination algorithm for feature selection: Case study Email spam detection	Mohammedzadeh et al.	2020
[15]	Review of swarm intelligence-based feature selection methods	Rostami et al.	2021
[16]	A Comprehensive Analysis of Nature-Inspired Meta-Heuristic Techniques for Feature Selection Problem	Sharma et al.	2020
[17]	Attack classification using feature selection techniques: a comparative study	Thakkar et al.	2020
[18]	Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria	Katrutsa et al.	2016
[19]	Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study	Allaoui et al.	2020
[20]	A survey on hybrid feature selection methods in microarray gene expression data for cancer classification	Almugren et al.	2019

[21]	Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset	Kasongo et al.	2020
[22]	Dragonfly algorithm: Theory, literature review, and application in feature selection	Mafarja et al.	2020
[23]	A review of machine learning methods of feature selection and classification for autism spectrum disorder	Rahman et al.	2020
[24]	A review of unsupervised feature selection methods	Solorio-Fernandes et al.	2019
[25]	Performance analysis of feature selection methods in software defect prediction: A search method approach	Balogun et al.	2019
[26]	A comparative study of feature selection techniques for classify student performance	Punlumjeak et al.	2015
[27]	A Review on Dimensionality Reduction Techniques	Jindal et al.	2017
[28]	A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction	Iqbal et al.	2016
[29]	Selection of appropriate statistical methods for data analysis	Mishra et al.	2019
[30]	A review of feature selection methods in medical applications	Remeseiro et al.	2020
[31]	Comparison of Feature Selection Methods and Machine Learning Classifiers for Radiomics Analysis in Glioma Grading	Sun et al.	2019
[32]	Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis	Sun et al.	2019
[33]	Feature selection methods and genomic big data: a systematic review	Tadist et al.	2019
[34]	Improved whale optimization algorithm for feature selection in Arabic sentiment analysis	Tubishat et al.	2018
[35]	A Review of Dimensionality Reduction Techniques for Efficient Computation	Velliangiri et al.	2019
[36]	A review of Feature Selection and its methods	Venkatesh et al.	2019
[37]	An overview of variable selection methods in multivariate analysis of near-infrared spectra	Yun et al.	2019
[38]	A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model	Zhang	2019
[39]	Intrusion detection with comparative analysis of supervised learning techniques and fisher score feature selection algorithm	Aksu et al.	2018
[40]	Swarm intelligence algorithms for feature selection: A review	Brezocnik et al.	2018
[41]	Discriminant Analysis-Based Dimension Reduction for Hyperspectral Image Classification: A Survey of the Most Recent Advances and an Experimental Comparison of Different Techniques	Li et al.	2018
[42]	Variable selection methods in multivariate statistical process control: A systematic literature review	Peres et al.	2017

[43]	A Deceptive Reviews Detection Method Based on Multidimensional Feature Construction and Ensemble Feature Selection	Li et al.	2023
[44]	A study of feature selection algorithms for predicting students academic performance	Zaffar et al.	2018
[45]	A large-scale study of the impact of feature selection techniques on defect classification models	Ghotra et al.	2017
[46]	Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier	Manek et al.	2015
[47]	An optimized feature selection technique based on incremental feature analysis for bio-metric gait data classification	Semwal et al.	2016
[48]	A Survey on semi-supervised feature selection methods	Sheikhpour et al.	2017
[49]	Dimension reduction techniques for the integrative analysis of multi-omics data	Menget al.	2016
[50]	Big Data Reduction Methods: A Survey	Rehman et al.	2016
[51]	Feature selection methods for big data bioinformatics: A survey from the search perspective	Wang et al.	2016
[52]	Advances in feature selection methods for hyperspectral image processing in food industry applications: A review	Dai et al.	2014
[53]	Similarity of feature selection methods: An empirical study across data intensive classification tasks	Dessi et al.	2015
[54]	A review of feature selection and feature extraction methods applied on microarray data	Hira et al.	2015
[55]	A review of feature selection methods with applications	Jovic et al.	2015
[56]	Analysis of instance selection algorithms on large datasets with Deep Convolutional Neural Networks	Albelwi et al.	2016
[57]	A Unified Analysis of Variational Inequality Methods: Variance Reduction, Sampling, Quantization, and Coordinate Descent	Beznosikov et al.	2022
[58]	Ensembles of instance selection methods: A comparative study	Blachnik	
[59]	A first analysis of meta-learned per-instance algorithm selection in scholarly recommender systems	Collins et al.	2019
[60]	A Comparative Survey of Instance Selection Methods applied to Non-Neural and Transformer-Based Text Classification	Cunha et al.	2023
[61]	Analysis of training sample selection strategies for regression-based quantitative landslide susceptibility mapping methods	Erener et al.	2017
[62]	Analysis of inconsistent source sampling in monte carlo weight-window variance reduction methods	Griesheimer et al.	2017
[63]	On Time-frequency Feature Selection Method for Neural Imaging Analysis with Small Sample Size	He et al.	2021

[64]	Three new instance selection methods based on local sets: A comparative study with several approaches from a bi-objective perspective	Leyva et al.	2013
[65]	Prototype reduction algorithms comparison in nearest neighbor classification for sensor data: Empirical study	Rosero-Montalvo et al.	2017
[66]	Instance selection and feature extraction using cuttlefish optimization algorithm and principal component analysis using decision tree	Suganthi et al.	2017
[67]	A study of prototype selection algorithms for nearest neighbour in class-imbalanced problems	Valero-Mas et al.	2017
[68]	An experimental study on rank methods for prototype selection	Valero-Mas et al.	2017
[69]	An analysis of dimension reduction methods applicable for out of sample problem in hyperspectral images	Yilmaz et al.	2019
[70]	Parameter Sensitivity Analysis for the Progressive Sampling-Based Bayesian Optimization Method for Automated Machine Learning Model Selection	2020	
[71]	Performance Analysis of Deep Autoencoder and NCA Dimensionality Reduction Techniques with KNN, ENN and SVM Classifiers	Siddique et al.	2019
[72]	Classification of inter-ictal and ictal EEGs using multi-basis MODWPT, dimensionality reduction algorithms and LS-SVM: A comparative study	Zhang et al.	2018
[73]	Comparison between Dimensionality Reduction Techniques for Pileup Detection in Digital Gamma Ray Spectroscopy	Abd eL Hamid et al.	2018
[74]	Bottleneck Channels Algorithm for Satellite Data Dimension Reduction: A Case Study for IASI	Pellet et al.	2019
[75]	A comprehensive study of eleven feature selection algorithms and their impact on text classification	Vora et al.	2017
[76]	Supervised Filter Feature Selection method for mixed data based on Spectral Feature Selection and Information-theory redundancy analysis	Solorio-Fernandes et al.	2020
[77]	Comparison of Mutual Information-based Feature Selection Method for Biological Omics Datasets	Huang	2021
[78]	Comparison of instance selection and construction methods with various classifiers	Blachnik et al.	2020
[79]	A density-based approach for instance selection	Carbonara et al.	2015
[80]	A modified genetic algorithm and weighted principal component analysis based feature selection and extraction strategy in agriculture	Shastry et al.	2021

A.2 Kriterien Auswertungstabelle der systematischen Literaturrecherche



[22]		x							x		x												x			x	
[23]	x	x			x		x		x	x													x		x		
[24]		x										x														x	
[25]		x							x								x						x			x	
[26]		x							x																x		
[27]	x								x									x								x	
[28]		x							x														x				
[29]		x						x																			
[30]		x								x		x	x	x		x		x							x	x	
[31]		x							x														x			x	
[32]	x	x											x	x	x			x		x	x						
[33]		x							x														x			x	
[34]		x							x								x						x		x	x	
[35]	x											x								x					x		
[36]		x																x									
[37]		x			x				x														x				
[38]		x							x														x		x	x	
[39]		x							x																x		
[40]		x							x														x			x	
[41]	x								x						x	x											
[42]	x	x							x																		
[43]		x		x					x	x													x			x	
[44]		x							x																		
[45]		x							x																x		
[46]		x						x	x	x																	
[47]		x							x														x		x	x	
[48]		x							x														x			x	
[49]	x														x	x	x										
[50]	x				x													x	x								
[51]		x								x			x	x								x	x			x	
[52]		x							x														x			x	
[53]		x							x		x														x		
[54]	x	x							x						x								x	x		x	
[55]	x	x			x				x														x			x	

[56]			x						x									x		x			x			x	
[57]				x																							x
[58]			x						x								x	x		x			x			x	
[59]			x		x				x											x			x			x	
[60]			x						x											x			x			x	
[61]				x																							x
[62]				x					x													x					x
[63]		x							x																		x
[64]			x						x	x	x		x	x			x			x			x			x	
[65]			x						x											x			x			x	
[66]	x		x						x									x					x			x	
[67]			x						x													x	x				
[68]			x										x	x						x					x		
[69]	x														x	x											
[70]				x													x						x			x	
[71]	x																			x	x			x		x	
[72]	x														x	x					x						
[73]	x													x	x		x	x									
[74]	x								x						x	x								x			
[75]		x							x	x																	
[76]		x							x									x							x		
[77]		x							x																x		
[78]			x						x								x			x			x			x	
[79]			x									x						x					x			x	
[80]	x																x								x		

A.3 Identifizierte Datenreduktionsverfahren der systematischen Literaturrecherche

Merkmals-auswahl	Laplacian Score	Trace Ratio	Fisher Score	Relieff	Spectral Feature Selection	Minimum Redundancy Maximum Relevance (mRMR)	Information gain (IG)	Mutual information (MI)	Joint Mutual Information	Joint Mutual Information Maximization (JMIM)	Fast Correlation-Based Filter	Least square loss (L1)	Multi-Cluster Feature Selection	Low variance / Variance Threshold	Chi-square	Gini Index	T-score (Student's T-statistic)	F-score (ANOVA F-statistic)	Graph Lasso	Tree Lasso	Recursive Feature Elimination (RFE)	Genetic Algorithms	Least Absolute Shrinkage and Selection Operator (LASSO)	Correlation Coefficients	Ridge Regression	Consistency-Based Subset Evaluation (CSE)	Elastic Net	Pearson Correlation	Backward Elimination	Fuzzy Rough-Based Feature Selection	XGBoost-based	Stepwise Selection	Gain Ratio	Orthogonal Centroid Algorithm (OC)	Incremental OC (IOC)	Orthogonal Centroid Feature Selection (OCFS)	Sparse Leaning Models	Sequential Foward Selection	Sequential Backward Selection	Greedy Randomized Adaptive Search Procedures (GRASP)	Time-Frequency Feature Consideration	Particle Swarm Optimization (PSO)	Ant Colony Optimization (ACO)	Nature-Inspired Algorithms (NIAs)	Binary Whale Optimization Algorithm (BWOA)	Enhanced Whale Optimization Algorithm (E-WOA)	Bee Colony Optimization	Whale Optimization Algorithm (WOA)	Flower Pollination Algorithm (FPA)	Grey Wolf Optimizer (GWO)	Binary Dragonfly Algorithm (BDA)	Dragonfly Algorithm (DA)	Online Streaming Feature Selection (OSFS)
[3]														x						x																														
[4]						x																													x	x	x												
[5]																		x			x	x	x																										
[6]																																												x					
[7]																																																
[8]							x								x						x	x	x																										
[9]																						x																				x	x				x		
[10]								x						x				x					x																										
[11]							x								x						x		x	x																									
[14]																																														x	x		
[15]																																										x	x				x		
[16]																						x																				x	x				x		
[17]							x								x						x																											
[18]								x																x																								
[20]	x					x	x	x										x																														
[21]																																x																	
[22]																																															x	x	
[23]																					x	x																											
[24]					x							x	x	x							x			x																									
[35]							x																																									
[26]						x	x															x																										
[28]																						x																										
[29]															x		x	x																														
[30]							x								x						x		x	x	x																							
[31]				x		x		x					x		x		x			x								x																				
[32]																					x	x										x															
[33]							x								x						x	x	x	x																							
[34]							x																																						x			
[36]		x					x						x		x					x	x	x		x																							
[37]																					x	x																									
[38]																												x																			
[39]			x																																												
[40]																																											x	x			x		
[42]																						x	x																									
[43]							x								x																x																
[44]							x								x						x	x	x	x	x																							
[45]			x	x			x	x							x	x			x		x	x	x	x					x																	
[46]																x																															
[47]																			x																												
[48]	x				x															x																x										
[51]					x															x																x										
[52]																						x																								
[53]							x	x							x						x	x	x	x	x													x	x	x							
[54]							x								x						x	x	x	x	x																						
[55]							x	x										x	x		x	x	x					x																			x
[63]			x																						x																						
[75]	x		x	x		x	x					x				x																															
[76]					x																																											
[77]						x				x	x																																				

Dimensions-reduktion	Principal Component Analysis (PCA)	Kernel Principal Component Analysis (KPCA)	Sparse PCA	Incremental PCA	Robust PCA	Randomized PCA	Functional PCA	Noise-Adjusted PCA (NAPCA)	Weighted PCA	Self-Organizing Maps (SOM)	Singular Value Decomposition (SVD)	Linear Discriminant Analysis (LDA)	Independent Component Analysis (ICA)	t-Distributed Stochastic Neighbor Embedding (t-SNE)	Locally Linear Embedding (LLE)	(Classical) Multidimensional Scaling (MDS)	Isometric Mapping (Isomap)	Projection Pursuit	Learning Vector Quantization	Locality Preserving Projections (LPP)	Uniform Manifold Approximation and Projection (UMAP)	Autoencoders	Deep Autoencoders	Sparse Graph-based Discriminant Analysis (SGDA)	Spectral-Spatial Discriminant Analysis (SSDA)	Canonical Correlation Analysis (CCA)	Partial Least Squares (PLS)	Locality Preserving Projections (LPP)	Neighborhood Preserving Embedding (NPE)	Orthogonal Locality Preserving Projections (OLPP)	Orthogonal Neighborhood Preserving Embedding (ONPE)	Nystrom Method	Neighborhood Components Analysis (NCA)	Correspondence Analysis	Diffusion Mapping (DM)	Laplacian Eigenmaps (LE)	Maximum Variance Unfolding (MVU)	Bottleneck Neural Network
[1]	x	x								x	x	x	x	x	x	x	x	x	x																	
[2]	x	x									x	x	x																							
[3]	x										x	x	x	x	x	x	x					x														
[4]	x	x									x	x	x							x																
[5]	x																																				
[12]	x											x	x	x		x					x															
[13]	x		x	x	x	x	x																														
[19]																					x															
[23]	x																																				
[27]	x	x									x	x	x	x		x					x															
[32]	x																																				
[35]	x	x									x	x		x	x		x					x														
[41]		x										x								x				x	x											
[42]	x																																				
[49]	x													x												x	x									
[50]	x																																				
[54]	x											x																									
[55]	x											x	x																								
[66]	x																																				
[69]																												x	x	x	x	x					
[71]																							x										x				
[72]	x	x											x		x		x																		x		
[73]	x	x														x	x																		x		
[74]	x							x																													x
[80]									x		x																						x				

Instanzenselektion	Random Mutation Hill Climbing (RMHC)	Proposed Fast RMHC	Bagging (Bootstrap Aggregating)	Feature Bagging	AdaBoost (Adaptive Boosting)	Additive Noise	Local Set Border Selector (LSBS)	Condensed Nearest Neighbor (CNN)	Local Set-based Smoother (LSS)	Edited Nearest Neighbor (ENN)	Instance-Based Learning 3	Drop 3	Iterative Case Filtering (ICF)	Local Density-Based Instance Selection	Prototyp Selection by Clustering	Genetic Algorithmus for Instance Selection	Random Forest-based Instance Selection	Learning Vactor Quantization	Relative neigbor graph editing (RNGE)	Hit miss network editing (HMN-E)	Class conditional instance selection (CCIS)	All-KNN	Nearest Centroid Neighbor Edition (NCNEdit)	Edited Normalized Radial Basis Function (ENRBF)	Reduced NN	Modified Condensed NN	Fast Condensed NN	Cuttlefish Optimization Algorithm for Instance Selection	EFCNN	DROP	Modified selective subset selection (MSS)	Generalized CNN (GCNN)	CHC
[56]	x	x						x	x	x																						
[58]			x	x	x	x		x		x	x	x	x						x	x	x	x										
[59]																	x															
[60]							x	x	x																							
[64]								x		x		x	x		x						x		x	x								
[65]	x							x		x		x				x						x			x	x	x					
[66]																												x				
[67]								x		x		x															x		x			x
[68]								x		x			x														x			x		x
[78]								x		x		x	x						x	x										x		
[79]								x		x	x	x	x	x											x				x		x	

Sampling	Random Sampling	Non-selective Nearby Sampling (NNS)	Progressive Sampling-Based Bayesian Optimization (PSBO)	Importance Sampling	Consistent Adjoint-driven - Importance Sampling (CADIS) Method	Proportional Random Sampling (PRS)	Random Under-Sampling	Selective Nearby Sampling (SNS)	Mini-batch Sampling
[4]	x								
[43]							x		
[57]				x					x
[61]		x				x		x	
[62]					x				
[70]			x						