Dr. Fabio Valdés

Modul 63117

Data Mining

Lehrveranstaltung

Data Mining – Konzepte und Techniken

LESEPROBE

Fakultät für
Mathematik und
Informatik

FernUniversität in Hagen

Vorwort

Liebe Fernstudentin, lieber Fernstudent,

wir begrüßen Sie herzlich zum Modul 63117 „Data Mining – Konzepte und Techniken“ und hoffen, dass Sie den vorliegenden Lehrtext motiviert und mit Erfolg bearbeiten.

Bei Data Mining handelt es sich nicht etwa, wie die direkte Übersetzung andeutet, um die Ausgrabung verschollener Daten. Besser passt im Deutschen der etwas sperrige Begriff „Wissensentdeckung in Datenmengen“. Die Zielsetzung von Data Mining besteht darin, bestimmte Informationen, Strukturen, Muster oder Anomalien in großen und sehr großen Mengen von Daten zu identifizieren. Dies sollte auf möglichst zweckmäßige Weise und im Idealfall automatisiert geschehen. Data Mining ist ein besonders heterogenes Wissenschaftsfeld, das u. a. die Bereiche Statistik, maschinelles Lernen, Mustererkennung, künstliche Intelligenz und wissensbasierte Systeme umfasst.

Die Bedeutung von Data Mining hat in den vergangenen Jahrzehnten deutlich zugenommen. Das ist einerseits darauf zurückzuführen, dass immer mehr Wirtschaftszweige Daten als entscheidende Ressourcen betrachten und bestrebt sind, diese optimal einzusetzen und Informationen bzw. Wissen aus ihnen zu gewinnen. Dies lässt sich nicht nur im klassischen IT-Bereich und etwa in hochtechnisierten Fertigungsbetrieben feststellen, sondern auch im Einzelhandel, bei Verkehrsunternehmen oder bei Behörden. Gleichzeitig werden weltweit derart viele geschäftliche, private und wissenschaftliche Daten erzeugt – seien es Bewegungsdaten von Personen oder Fahrzeugen, Bezahlvorgänge bzw. Kontobewegungen, Interaktionen in sogenannten sozialen Netzwerken, Nutzungsdaten von Sprachassistenten und Haushaltsgeräten, Server-Protokolldaten, Browserverläufe oder Suchmaschinenhistorien –, dass zielführende und effiziente Methoden der Datenanalyse grundlegend für deren sinnvolle Nutzung sind.

Dieser stets weiter wachsende Überfluss an Daten steht auch im Zusammenhang damit, dass sowohl Speichermedien als auch datenerzeugende Sensoren deutlich günstiger geworden sind. Während letztere bei Forschung und Entwicklung unverzichtbar geworden sind und zudem massenhaft in Smartphones, Smartwatches und anderen Geräten verbaut werden, ist beispielsweise der Preis pro Gigabyte Festplattenkapazität in den vergangenen 40 Jahren von einigen Hunderttausend Dollar auf etwa 0.01 US-Dollar gesunken [McC22, Leg23]. Als Konsequenz dieses Preisverfalls ist die massenhafte Erhebung bzw. Erzeugung von Daten sowie deren dauerhafte Speicherung mit immer geringeren Kosten möglich. Deutlich aufwendiger ist es dagegen, erfolgreiche Strategien für deren effektive und effiziente Nutzung zu entwickeln.

Laut einer Studie der International Data Corporation [BR22] wird die im Jahr 2026 weltweit erzeugte Datenmenge eine Größe von 221 Exabytes ($221 \cdot 10^{18}$ Bytes bzw. 221 Millionen Terabytes) erreichen. Damit entstehen jährlich knapp dreimal so viele Daten wie im Jahr 2021 und fast doppelt so viele wie 2023. Zahlreiche aktuelle Statistiken zur weltweiten Datennutzung und -erzeugung werden u. a. in [Ray23] zusammengefasst.

Gliederung der Lehrveranstaltung

Im Rahmen dieses Lehrtextes bieten wir einen Überblick zu Data Mining und stellen zentrale Aspekte und Methoden genauer vor. Nach der Einführung (Kapitel 1) werden in Kapitel 2 Attributtypen und statistische Größen sowie Datenvisualisierung und Ähnlichkeits- bzw. Abstandsmaße behandelt. Kapitel 3 beschäftigt sich mit verschiedenen Techniken zur Vorverarbeitung von Daten, die die Anwendung von Data-Mining-Methoden effizienter gestalten oder überhaupt erst ermöglichen. In Kapitel 4 beleuchten wir grundlegende Konzepte zur Bestimmung von Mustern und Korrelationen, bevor in Kapitel 5 das Thema Klassifikation (überwachtes Lernen) vorgestellt wird. Als dritter klassischer Bestandteil des Data-Mining-Prozesses folgt Kapitel 6 mit der Clusteranalyse (unüberwachtes Lernen). Kapitel 7 befasst sich mit Data Mining auf komplexeren Strukturen wie Datenströmen, Textdaten, Zeitreihen, mehrdimensionalen Daten und Webdaten. Die Lehrveranstaltung endet mit praktischen Anwendungen in dem Data-Mining-Tool Weka (Kapitel 8).

Das Lehrmaterial wurde folgendermaßen in Lektionen aufgeteilt:

Lektion	Kapitel	Inhalt
1	1, 2	Einführung, Datencharakterisierung
2	3	Vorverarbeitung
3	4	Mustersuche
4	5	Klassifikation
5	6	Clusteranalyse
6	7	Analyse komplexer Strukturen I
7	7, 8	Analyse komplexer Strukturen II, Weka

Voraussetzungen

Diese Lehrveranstaltung geht davon aus, dass Sie grundlegende Kenntnisse in den Bereichen Statistik und Datenbanken besitzen. Formale Voraussetzungen bestehen jedoch nicht.

Übungen

Zum Verständnis und zur Verinnerlichung der in dieser Lehrveranstaltung vorgestellten Inhalte empfehlen wir die Bearbeitung der Einsende- und Selbsttestaufgaben.

Klausur

Das Modul „Data Mining – Konzepte und Techniken“ wird in Form einer Klausur geprüft.

Weitere Informationen

Organisatorische und sonstige Informationen zur Lehrveranstaltung finden Sie in einem gesonderten Anschreiben, welches Sie zusammen mit dieser ersten Lektion per Post erhalten haben bzw. über Moodle abrufen können.

Stoffeingrenzung

Prüfungsrelevant sind alle Inhalte der Lehrveranstaltung mit Ausnahme der folgenden Kapitel und Abschnitte (jeweils inklusive aller Unterabschnitte):

- 4.6
- 4.7
- 6.3.4
- 7.1.3
- 7.1.4
- 7.2.4
- 7.3.3
- 7.3.4
- 8

Literatur

Auf relevante und hilfreiche Literatur wird in den Literaturhinweisen am Ende jedes Kapitels hingewiesen. Das Literaturverzeichnis für die gesamte Lehrveranstaltung befindet sich am Ende der ersten Lektion (also hinter Kapitel 2).

Der Autor

Dr. Fabio Valdés. Bis 2011 Studium der Mathematik an der Technischen Universität Dortmund mit Abschluss Diplom. Anschließend tätig als wissenschaftlicher Mitarbeiter an der FernUniversität in Hagen, Lehrgebiet Datenbanksysteme für neue Anwendungen, danach Softwaretechnik und Theorie der Programmierung. Seit 2022 Dozent bzw. Lehrkraft für besondere Aufgaben im Zentralbereich der Fakultät [Fer23].

Inhaltsverzeichnis

1 Einführung

We are drowning in information, but starving for knowledge.
John Naisbitt (Autor, Trend- und Zukunftsforscher; * 1929)

Das Thema dieser Lehrveranstaltung ist Data Mining, ins Deutsche grob übersetzbar mit „Wissensentdeckung in Datenmengen/-banken“. Im Rahmen dieser Einleitung werden wir auf die rasant gewachsene Bedeutung von Data Mining eingehen, den Ablauf des Wissensentdeckungsprozesses skizzieren und ein Anwendungsbeispiel vorstellen.

Mittlerweile existieren unzählige Bereiche der Wirtschaft, des privaten Lebens und der Wissenschaft, in denen man aus sehr großen Datenmengen Informationen und Wissen generiert. Vor der Jahrtausendwende noch undenkbar, sind die weltweit wertvollsten und erfolgreichsten Unternehmen heute im Kerngeschäft Meister der Datenanalyse, beispielsweise basierend auf einer werbegestützten Internetsuchmaschine mit Ergebnispriorisierung, einem weltweiten Marktplatz für alle denkbaren Produkte, der Käufe und Kaufinteressen analysiert oder einem sogenannten sozialen Netzwerk mit auf die einzelne Nutzerin und den einzelnen Nutzer zugeschnittener Informationsflut. Auch abseits von Großkonzernen oder der klassischen IT-Branche ist es heute in vielen Bereichen hilfreich oder notwendig, große Mengen von Daten zu analysieren. Ob wir an Verkaufstransaktionen von Supermarktketten und Onlinehändlern denken, den weltweiten Wertpapierhandel, an Elektrizitätsnetze, den internationalen Flugverkehr oder an Telekommunikationsunternehmen mit Mobilfunknetzen und Transatlantikkabeln, an Messergebnisse unzähliger Sensoren in Wissenschaft und Entwicklung, Streams bzw. Downloads von Filmen, Serien und Sportereignissen, GPS-Positionsdaten von Mobiltelefonen, Wearables (z. B. Smartwatches, Datenbrillen) und Fahrzeugen, Aufzeichungen von Sprachassistenten – der Überfluss an erzeugten und gespeicherten Daten als Effekt gesellschaftlicher und technischer Veränderung ist offensichtlich. Das Erfordernis, daraus wertvolle Informationen abzuleiten, hat den Bereich Data Mining ins Leben gerufen.

1.1 Historische Entwicklung

In den sechziger Jahren des vergangenen Jahrhunderts begann die Entwicklung leistungsstarker Datenbanksysteme, die die bis dahin vorherrschende Dateiverarbeitung allmählich ablösten. Relationale Datenbanken in Kombination mit Modellierungssprachen und -werkzeugen sowie Indexen für effizientere Zugriffe wurden in den 1970ern etabliert und boten Benutzeroberflächen, Anfragesprachen wie SQL sowie Anfrageoptimierung. Die Einführung der Echtzeit-Transaktionsverarbeitung (OLTP; Online Transaction Processing) als Gegenstück zur Stapelverarbeitung ermöglichte Transaktionssicherheit bei parallelen Anfragen und führte zu einer weiten Verbreitung relationaler Datenbanken.

Neben fortgeschrittenen Datenmodellen (etwa objektorientiert oder objektrelational) wurden in den achtziger Jahren spezialisierte Datenbanksysteme entwickelt, beispielsweise für Multimediaanwendungen sowie zur Verarbeitung geometrischer oder zeitlich

veränderlicher Objekte. Begünstigt durch den zügigen Fortschritt der Hardwaretechnologie und sinkende Preise für Computer und Speichermedien nahm die Verbreitung von Datenbanken weiter zu. Gegen Ende der 1980er Jahre kamen Data Warehouses (in der deutschsprachigen Literatur häufig als „Datenlager“ bezeichnet) auf. Dabei handelt es sich um zentrale Datenbanken, die Daten aus heterogenen Quellen zusammenführen und für Analysezwecke vorbereiten, z. B. durch Datenbereinigung, Integration oder OLAP (Online Analytical Processing), welches komplexe mehrdimensionale Analysen mit hohem Datenaufkommen beinhaltet.

Das letzte Jahrzehnt des 20. Jahrhunderts war von der Einführung des Internets gekennzeichnet. Im Zusammenhang damit stieg die Bedeutung heterogener und internetbasierter Datenbanksysteme (z. B. XML-Datenbanken).

Nach dem Jahr 2000 erkannten mehr und mehr Unternehmen die Chancen der globalen Vernetzung, während parallel Privatpersonen zu Internetnutzerinnen und -nutzern und damit zu potentieller Kundschaft und Datenerzeugenden wurden. Für die Verarbeitung großer Datenmengen entwickelten IT-Konzerne für ihre jeweiligen Bedürfnisse maßgeschneiderte Datenbanksysteme bzw. Frameworks wie MapReduce, Google File System, Bigtable, Spanner, F1 (Google), Dynamo (Amazon) oder Hive (Facebook).

Die massenhafte globale Verbreitung von technischen Geräten, die unabhängig vom Aufenthaltsort permanenten Internetzugang ermöglichen, und damit zusammenhängend die ständige Erreichbarkeit und Interaktionsfähigkeit, etwa hinsichtlich Aktivitäten in sogenannten sozialen Netzwerken, sowie die Automatisierung und Vernetzung in nahezu allen Wirtschaftsbereichen führen ab dem zweiten Jahrzehnt des 21. Jahrhunderts zu einer weiteren Steigerung der weltweit erzeugten Datenmenge. Dieser Effekt wird in den letzten Jahren durch Sprachassistenten, vernetzte Haushaltsgeräte (z. B. Kühlschränke, Beleuchtungssysteme, Temperatursteuerung) sowie Fahrzeuge mit zahlreichen Assistenzsystemen weiter verstärkt.

Offenbar sind in vielen Bereichen mächtige Werkzeuge erforderlich, um aus der riesigen und weiterhin streng monoton wachsenden Menge von Daten Informationen bzw. Wissen abzuleiten. Data Mining bietet eine Fülle von Möglichkeiten, um die Qualität von Entscheidungen zu optimieren und sie anschließend zu überprüfen, so dass große Datenmengen als wertvolle Ressource nutzbar werden.

1.2 Prozess der Wissensentdeckung

Data Mining findet nur selten isoliert statt. Der gesamte Prozess der Wissensentdeckung, dessen Hauptbestandteil Data-Mining-Methoden sind, der aber auch vor- und nachbereitende Phasen umfasst, lässt sich in folgende vier Schritte aufteilen[1]:

(i) **Datencharakterisierung**: Überblick über die Art und Ausprägung der Daten durch statistische Maße sowie Visualisierung

(ii) **Vorverarbeitung**: Integration verschiedener Quellen, Beseitigung von Inkonsistenzen sowie Ausreißern, Reduktion der zu verarbeitenden Datenmenge, Umwandlung in ein passendes Analyseformat

1. In der Literatur sind verschiedene Aufteilungen zu finden. Manche Schritte werden aufgeteilt bzw. zusammengefasst, zudem werden häufig Zielspezifikation oder Erstellung eines Projektplans als erste Phase bzw. Wissenspräsentation oder Implementierung in der Praxis als letzter Schritt genannt.

(iii) **Data Mining**: eigentlicher Analyseschritt; mit Hilfe bestimmter Methoden wie Mustersuche, Klassifikation oder Clusteranalyse werden Zusammenhänge oder Muster erkannt

(iv) **Evaluation**: Auswahl der interessanten Analyseergebnisse auf Basis entsprechender Maße, Kontrolle der erreichten Ziele

Streng genommen ist Data Mining also lediglich einer von mehreren Schritten der Wissensentdeckung. Allerdings ist es sowohl in der Wirtschaft als auch in der Wissenschaft üblich, Data Mining als den kompletten Prozess der Wissens- und Musterentdeckung in großen Datenmengen zu betrachten, was wir in diesem Lehrtext ebenfalls übernehmen.

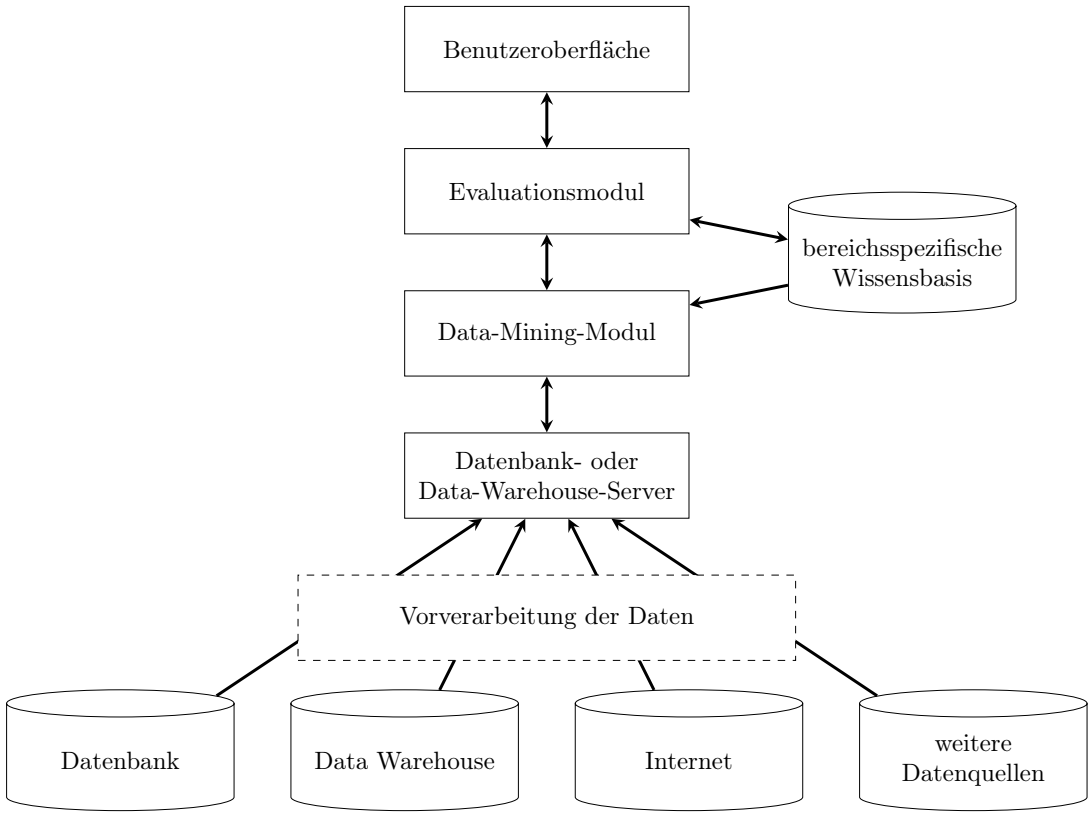

Abbildung 1.2.1: *Architektur eines typischen Data-Mining-Systems*

Entsprechend kann die Architektur eines Data-Mining-Gesamtsystems wie in Abbildung 1.2.1 dargestellt werden. Dabei führt der Server, abhängig von der Anfrage des Benutzers oder der Benutzerin, die ersten drei Schritte der obigen Auflistung durch. Die eigentliche Analyse der Daten findet im Data-Mining-Modul statt, etwa eine Klassifikation oder eine Ausreißeranalyse. Deren Ergebnisse werden vom Evaluationsmodul anhand bestimmter Relevanzmaße (vgl. Kapitel 4) und Grenzwerte überprüft und dem oder der Benutzenden ggf. in der Oberfläche angezeigt. Die Wissensbasis, stets im Austausch mit der Evaluationseinheit und einflussnehmend auf das Data-Mining-Modul, kann u. a.

Konzepthierarchien[2] aus dem jeweiligen Anwendungsbereich zu Hilfe nehmen. Auch die persönliche Einschätzung der Relevanz von Ergebnissen kann an dieser Stelle einfließen.

1.3 Anwendungsbeispiel

Im Folgenden führen wir ein Beispiel ein, auf das wir die im Laufe dieser Lehrveranstaltung vorgestellten Techniken anwenden werden. Wir betrachten den fiktiven Streamingdienst *Dreamstream*, dessen Kundschaft für eine monatliche Gebühr bestimmte Film-, Serien-, Sport- und sonstige Videopakete abonnieren kann, um sie auf einem Endgerät anzusehen. Die Unternehmensdaten sind in einer relationalen Datenbank organisiert, deren Relationenschemata in Tabelle 1.3.1 aufgelistet werden. Dabei ist jeweils das erste Attribut ein Identifikator bzw. Schlüssel.

Tabelle 1.3.1: *Relationenschemata für die relationale Datenbank der Firma Dreamstream*

Name	Schema
Kundschaft	(*KNr, Name, Geschlecht, Adresse, Alter, Beruf, Gehalt, Kreditwürdigkeit, ...*)
Produkt	(*ProdNr, Name, PaketNr, Genre, Länge, Größe, ...*)
Paket	(*PaketNr, Name, Standardpreis, ...*)
Personal	(*MNr, Name, Adresse, Alter, Tätigkeit, Abteilung, Gehalt, WurdeAbgemahnt, ...*)
Vertrag	(*VNr, KNr, PaketNr, Monatspreis, Abschlussdatum, SofortKündbar, ...*)

Dieses Datenbankschema bietet die Möglichkeit, Anfragen zu formulieren, mit denen man etwa alle Pakete, die von Rentnerinnen und Rentnern aus Nordrhein-Westfalen abonniert wurden, ermitteln oder den Speicherplatz, der von den Produkten des Filmpakets „Philip Marlowe“ belegt wird, berechnen kann. Des Weiteren lassen sich auch bestimmte Muster identifizieren und Vorhersagen treffen. Beispielsweise können aus Alter, Einkommen und Kreditwürdigkeit der Kundschaft Rückschlüsse auf das Risiko ausbleibender Zahlungen gezogen werden. Weiterhin besteht die Möglichkeit, Effekte von Rabatten zu untersuchen oder Gruppen der Kundschaft zu identifizieren, die mehrere Pakete abonnieren, ohne diese tatsächlich zu nutzen.

Auf Basis einer Transaktionsübersicht, aus der ersichtlich ist, welche Pakete zusammen bestellt wurden, kann eine Warenkorbanalyse durchgeführt werden. Findet man z. B. heraus, dass das eher spärlich vertriebene „Boris Karloff“-Paket, wenn überhaupt, gemeinsam mit dem besonders beliebten Paket „Buster Keaton“ abonniert wird, so kann man beide in Kombination anbieten, um den Absatz von „Boris Karloff“ zu steigern. Derartiges Wissen lässt sich mit Hilfe der Mustersuche ableiten.

1.4 Literaturhinweise

Zum Thema Data Mining wurden in den letzten 25 Jahren zahlreiche Bücher veröffentlicht. Exemplarisch erwähnt seien hier vor allem Han, Pei und Tong [HPT22] sowie von Aggarwal [Agg15], an denen sich diese Lehrveranstaltung im Wesentlichen orientiert. Weitere einschlägige Werke stammen von Hastie, Tibshirani und Friedman [HTF09] sowie von Tan, Steinbach, Karpatne und Kumar [TSKK19].

2. Hierbei handelt es sich um eine Folge von Abbildungen von einer speziellen zu einer allgemeineren Stufe, etwa 1-Euro-Stück → Münze → Bargeld → Zahlungsmittel.

Gut recherchierte und detaillierte Übersichten zur historischen Entwicklung der letzten 250 Jahre von Bayes bis Big Data findet man in den Artikeln von Foote [Foo21] und von Li [Li17].

Bei van der Aalst [Aal16] wird der Gesamtprozess der Wissensentdeckung in den Fokus genommen, unter Berücksichtigung geschäftlicher und organisatorischer Faktoren sowie softwaregestützter Methoden aus der Praxis.

Literaturverzeichnis

[AAD+96] Agarwal, S., Agrawal, R., Deshpande, P., Gupta, A., Naughton, J.F., Ramakrishnan, R. und Sarawagi, S. On the computation of multidimensional aggregates. In *Proc. Int. Conf. Very Large Data Bases (VLDB'96)*, S. 506–521. 1996.

[Aal16] van der Aalst, W.M.P. *Process Mining: Data Science in Action*. Springer, 2. Auflage, 2016.

[ABKS99] Ankerst, M., Breunig, M.M., Kriegel, H. und Sander, J. OPTICS: ordering points to identify the clustering structure. In *Proc. Int. Conf. Management of Data (SIGMOD'99)*, S. 49–60. 1999.

[AEEK99] Ankerst, M., Elsen, C., Ester, M. und Kriegel, H. Visual classification: An interactive approach to decision tree construction. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'99)*, S. 392–396. 1999.

[AGAV09] Amigó, E., Gonzalo, J., Artiles, J. und Verdejo, F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, 2009.

[Agg05] Aggarwal, C.C. On abnormality detection in spuriously populated data streams. In *Proc. Int. Conf. Data Mining (SDM'05)*, S. 80–91. 2005.

[Agg06] Aggarwal, C.C. On biased reservoir sampling in the presence of stream evolution. In *Proc. Int. Conf. Very Large Data Bases (VLDB'06)*, S. 607–618. 2006.

[Agg07] Aggarwal, C.C., Hrsg. *Data Streams: Models and Algorithms*, Band 31 von *Advances in Database Systems*. Springer, 2007.

[Agg09] Aggarwal, C.C. *Managing and Mining Uncertain Data*. Springer, 2009.

[Agg14] Aggarwal, C.C. A survey of stream classification algorithms. In *Data Classification: Algorithms and Applications*, S. 245–274. 2014.

[Agg15] Aggarwal, C.C. *Data Mining: The Textbook*. Springer, 2015.

[AGGR98] Agrawal, R., Gehrke, J., Gunopulos, D. und Raghavan, P. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. Int. Conf. Management of Data (SIGMOD'98)*, S. 94–105. 1998.

[AHK18] Andreß, H.J., Hagenaars, J.A. und Kühnel, S. *Analyse von Tabellen und kategorialen Daten: log-lineare Modelle, latente Klassenanalyse, logistische Regression und GSK-Ansatz*. Springer, 2. Auflage, 2018.

[AHS96] Arabie, P., Hubert, L.J. und de Soete, G. *Clustering and Classification*. World Scientific, 1996.

[AHWY03] Aggarwal, C.C., Han, J., Wang, J. und Yu, P.S. A framework for clustering evolving data streams. In *Proc. Int. Conf. Very Large Data Bases (VLDB'03)*, S. 81–92. 2003.

[AHWY04] Aggarwal, C.C., Han, J., Wang, J. und Yu, P.S. On demand classification of data streams. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'04)*, S. 503–508. 2004.

[AIS93] Agrawal, R., Imielinski, T. und Swami, A.N. Mining association rules between sets of items in large databases. In *Proc. Int. Conf. Management of Data (SIGMOD'93)*, S. 207–216. 1993.

[AKK96] Ankerst, M., Keim, D.A. und Kriegel, H.P. Circle segments : A technique for visually exploring large multidimensional data sets. In *Proc. Visualization '96, Hot Topic Session*. 1996.

[AL99] Aumann, Y. und Lindell, Y. A statistical theory for quantitative association rules. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'99)*, S. 261–270. 1999.

[All01] Allison, P.D. *Missing Data*. SAGE Publications, 2001.

[AMO93] Ahuja, R.K., Magnanti, T.L. und Orlin, J.B. *Network Flows - Theory, Algorithms, and Applications*. Prentice Hall, 1993.

[AR14] Aggarwal, C.C. und Reddy, C.K., Hrsg. *Data Clustering: Algorithms and Applications*. CRC Press, 2014.

[AS94] Agrawal, R. und Srikant, R. Fast algorithms for mining association rules in large databases. In *Proc. Int. Conf. Very Large Data Bases (VLDB'94)*, S. 487–499. 1994.

[AS95] Agrawal, R. und Srikant, R. Mining sequential patterns. In *Proc. Int. Conf. Data Engineering (ICDE'95)*, S. 3–14. 1995.

[AT05] Adomavicius, G. und Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.*, 17(6):734–749, 2005.

[ATW+07] Aggarwal, C.C., Ta, N., Wang, J., Feng, J. und Zaki, M.J. Xproj: A framework for projected structural clustering of XML documents. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'07)*, S. 46–55. 2007.

[AV07] Arthur, D. und Vassilvitskii, S. k-means++: the advantages of careful seeding. In *Proc. SIAM Symposium on Discrete Algorithms (SODA'07)*, S. 1027–1035. 2007.

[AW10] Aggarwal, C.C. und Wang, H., Hrsg. *Managing and Mining Graph Data*, Band 40 von *Advances in Database Systems*. Springer, 2010.

[AY10] Aggarwal, C.C. und Yu, P.S. On clustering massive text and categorical data streams. *Knowledge and Information Systems*, 24(2):171–196, 2010.

[AZ12] Aggarwal, C.C. und Zhai, C., Hrsg. *Mining Text Data*. Springer, 2012.

[BB98] Bagga, A. und Baldwin, B. Entity-based cross-document coreferencing using the vector space model. In *Int. Conf. Computational Linguistics (COLING'98)*, S. 79–85. 1998.

[BBCC08] Bloch, M., Byron, L., Carter, S. und Cox, A. The ebb and flow of movies: Box office receipts 1986 – 2007. https://archive.nytimes.com/screenshots/www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.jpg, 2008. Abgerufen am 06.12.2023.

[BBD+02] Babcock, B., Babu, S., Datar, M., Motwani, R. und Widom, J. Models and issues in data stream systems. In *Proc. ACM Symp. Principles of Database Systems (PODS'02)*, S. 1–16. 2002.

[BCK08] Boriah, S., Chandola, V. und Kumar, V. Similarity measures for categorical data: A comparative evaluation. In *Proc. SIAM Int. Conf. Data Mining*, S. 243–254. 2008.

[BDF+97] Barbará, D., DuMouchel, W., Faloutsos, C., Haas, P.J., Hellerstein, J.M., Ioannidis, Y.E., Jagadish, H.V., Johnson, T., Ng, R.T., Poosala, V., Ross, K.A. und Sevcik, K.C. The new jersey data reduction report. *IEEE Data Eng. Bull.*, 20(4):3–45, 1997.

[Ber81] Bertin, J. *Graphics and Graphic Information Processing*. de Gruyter, 1981.

[BFOS84] Breiman, L., Friedman, J.H., Olshen, R.A. und Stone, C.J. *Classification and Regression Trees*. Wadsworth, 1984.

[BGMP03] Bonchi, F., Giannotti, F., Mazzanti, A. und Pedreschi, D. Exante: Anticipated data reduction in constrained pattern mining. In *Proc. Eur. Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD'03)*, S. 59–70. 2003.

[BGRS99] Beyer, K.S., Goldstein, J., Ramakrishnan, R. und Shaft, U. When is "nearest neighbor" meaningful? In *Proc. Int. Conf. Database Theory (ICDT'99)*, S. 217–235. 1999.

[BHR00] Bergroth, L., Hakonen, H. und Raita, T. A survey of longest common subsequence algorithms. In *Proc. Int. Symp. String Processing and Information Retrieval (SPIRE'00)*, S. 39–48. 2000.

[Bis07] Bishop, C.M. *Pattern recognition and machine learning*. Springer, 5. Auflage, 2007.

[BJRL15] Box, G.E.P., Jenkins, G.M., Reinsel, G.C. und Ljung, G.M. *Time series analysis : forecasting and control*. Wiley, 5. Auflage, 2015.

[Blo70] Bloom, B.H. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM*, 13(7):422–426, 1970.

[BMS97] Brin, S., Motwani, R. und Silverstein, C. Beyond market baskets: Generalizing association rules to correlations. In *Proc. Int. Conf. Management of Data (SIGMOD'97)*, S. 265–276. 1997.

[BMS+13] Buluç, A., Meyerhenke, H., Safro, I., Sanders, P. und Schulz, C. Recent advances in graph partitioning. *CoRR*, abs/1311.3144, 2013.

[BMUT97] Brin, S., Motwani, R., Ullman, J.D. und Tsur, S. Dynamic itemset counting and implication rules for market basket data. In *Proc. Int. Conf. Management of Data (SIGMOD'97)*, S. 255–264. 1997.

[BN92] Buntine, W.L. und Niblett, T. A further comparison of splitting rules for decision-tree induction. *Machine Learning*, 8:75–85, 1992.

[BP98] Brin, S. und Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[BR99] Baeza-Yates, R.A. und Ribeiro-Neto, B.A. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[BR22] Burgener, E. und Rydning, J. High data growth and modern applications drive new storage requirements in digitally transformed enterprises. *IDC White Paper*, 2022.

[Bre96] Breiman, L. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[Bre01] Breiman, L. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[BS98] Bunke, H. und Shearer, K. A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters*, 19(3-4):255–259, 1998.

[BSM+15] Bernard, J., Steiger, M., Mittelstädt, S., Thum, S., Keim, D.A. und Kohlhammer, J. A survey and task-based quality assessment of static 2d colormaps. In *Visualization and Data Analysis*, S. 93970M. 2015.

[BU95] Brodley, C.E. und Utgoff, P.E. Multivariate decision trees. *Machine Learning*, 19(1):45–77, 1995.

[Bun97] Bunke, H. On a relation between graph edit distance and maximum common subgraph. *Pattern Recognition Letters*, 18(8):689–694, 1997.

[BW08] Byron, L. und Wattenberg, M. Stacked graphs - geometry & aesthetics. *IEEE Trans. on Visualization and Computer Graphics*, 14(6):1245–1252, 2008.

[CBK12] Chandola, V., Banerjee, A. und Kumar, V. Anomaly detection for discrete sequences: A survey. *IEEE Trans. Knowl. Data Eng.*, 24(5):823–839, 2012.

[CFSV04] Cordella, L.P., Foggia, P., Sansone, C. und Vento, M. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(10):1367–1372, 2004.

[Cha02] Chakrabarti, S. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kaufmann, 2002.

[CHY96] Chen, M., Han, J. und Yu, P.S. Data mining: An overview from a database perspective. *IEEE Trans. Knowl. Data Eng.*, 8(6):866–883, 1996.

[CKL03] Chiu, B.Y., Keogh, E.J. und Lonardi, S. Probabilistic discovery of time series motifs. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, S. 493–498. 2003.

[CKPT92] Cutting, D.R., Karger, D.R., Pedersen, J.O. und Tukey, J.W. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proc. Int. Conf. Information Retrieval (SIGIR'92)*, S. 318–329. 1992.

[Cle93] Cleveland, W.S. *Visualizing Data*. AT&T Bell Laboratories, 1993.

[CM05] Cormode, G. und Muthukrishnan, S.M. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, 2005.

[CN89] Clark, P. und Niblett, T. The CN2 induction algorithm. *Machine Learning*, 3:261–283, 1989.

[Coh95] Cohen, W.W. Fast effective rule induction. In *Proc. Int. Conf. Machine Learning*, S. 115–123. 1995.

[CR17] Caldarola, E.G. und Rinaldi, A.M. Big data visualization tools: A survey - the new paradigms, methodologies and tools for large data sets visualization. In *Int. Conf. on Data Science, Technology and Applications*, S. 296–305. 2017.

[CX19] Chatfield, C. und Xing, H. *The Analysis of Time Series: An Introduction with R*. CRC Press, 7. Auflage, 2019.

[Dav17] Davies, J. Word cloud generator. https://www.jasondavies.com/wordcloud, 2017. Abgerufen am 06.12.2023.

[DDF+90] Deerwester, S.C., Dumais, S.T., Furnas, G.W., Landauer, T.K. und Harshman, R.A. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.*, 41(6):391–407, 1990.

[DE84] Day, W.H.E. und Edelsbrunner, H. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1):7–24, 1984.

[DH00] Domingos, P.M. und Hulten, G. Mining high-speed data streams. In *Proc. Int. Conf. Knowledge discovery and data mining (KDD'00)*, S. 71–80. 2000.

[Dhi01] Dhillon, I.S. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'01)*, S. 269–274. 2001.

[DHI12] Doan, A., Halevy, A. und Ives, Z. *Principles of Data Integration*. Morgan Kaufmann, 2012.

[DHS01] Duda, R.O., Hart, P.E. und Stork, D.G. *Pattern classification*. Wiley, 2. Auflage, 2001.

[DJ03] Dasu, T. und Johnson, T. *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, 2003.

[DK04] Deshpande, M. und Karypis, G. Item-based top-n recommendation algorithms. *ACM Trans. Inf. Syst.*, 22(1):143–177, 2004.

[DL97] Dash, M. und Liu, H. Feature selection for classification. *Intell. Data Anal.*, 1(1-4):131–156, 1997.

[DMM03] Dhillon, I.S., Mallela, S. und Modha, D.S. Information-theoretic coclustering. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, S. 89–98. 2003.

[Ega75] Egan, J.P. *Signal detection theory and ROC analysis*. Academic Press, 1975.

[EKSX96] Ester, M., Kriegel, H., Sander, J. und Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'96)*, S. 226–231. 1996.

[EKX95] Ester, M., Kriegel, H. und Xu, X. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. In *Proc. Int. Symp. Large Spatial Databases (SSD'95)*, S. 67–82. 1995.

[Elk97] Elkan, C. Boosting and naïve bayesian learning. Technical Report CS97-557, Dept. Computer Science and Engineering, Univ. Calif. at San Diego, 1997.

[EN15] Elmasri, R. und Navathe, S.B. *Fundamentals of Database Systems*. Addison-Wesley, 7. Auflage, 2015.

[Eng99] English, L.P. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. John Wiley & Sons, 1999.

[ET93] Efron, B. und Tibshirani, R.J. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[Fal85] Faloutsos, C. Access methods for text. *ACM Comput. Surv.*, 17(1):49–74, 1985.

[Fer23] FernUniversität in Hagen. Dr. Fabio Valdés. https://e.feu.de/valdes, 2023. Abgerufen am 06.12.2023.

[FH95] Fu, Y. und Han, J. Meta-rule-guided mining of association rules in relational databases. In *Proc. Int. Workshop Integration of Knowledge Discovery with Deductive and Object-Oriented Databases (KDOOD'95)*, S. 39–46. 1995.

[FHW16] Frank, E., Hall, M.A. und Witten, I.H. *The WEKA Workbench. Online Appendix for „Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, 4. Auflage, 2016.

[FI92] Fayyad, U.M. und Irani, K.B. The attribute selection problem in decision tree generation. In *Proc. Nat. Conf. Artificial Intelligence*, S. 104–110. 1992.

[Fis36] Fisher, R.A. Iris plants database. http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data, 1936. Abgerufen am 06.12.2023.

[Foo21] Foote, K.D. A brief history of data science. https://www.dataversity.net/brief-history-data-science, 2021. Abgerufen am 06.12.2023.

[For10] Fortunato, S. Community detection in graphs. *Physics Reports*, 486(3):75 – 174, 2010.

[FP08] Friedman, J.H. und Popescu, B.E. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.

[FPP07] Freedman, D., Pisani, R. und Purves, R. *Statistics*. International student edition. W.W. Norton & Company, 2007.

[FPRS07] Fouss, F., Pirotte, A., Renders, J. und Saerens, M. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Trans. Knowl. Data Eng.*, 19(3):355–369, 2007.

[Fri77] Friedman, J.H. A recursive partitioning decision rule for nonparametric classification. *IEEE Trans. Comput.*, 26(4):404–408, 1977.

[Fry08] Fry, B. *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. O'Reilly Media, Incorporated, 2008.

[FS97] Freund, Y. und Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

[FWG02] Fayyad, U.M., Wierse, A. und Grinstein, G.G. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann, 2002.

[GADc97] Güvenir, H.A., Acar, B., Demiröz, G. und Çekin, A.H. A supervised machine learning algorithm for arrhythmia analysis. In *Computers in Cardiology 1997*, S. 433–436. 1997.

[GAM97] Güvenir, H.A., Acar, B. und Müderrisoğlu, H. Arrhythmia data set. https://archive.ics.uci.edu/ml/datasets/arrhythmia, 1997. Abgerufen am 06.12.2023.

[Gar84] Garvin, D. What does "product quality" really mean? *MIT Sloan Management Review*, 26:25–43, 1984.

[GDD02] Gozalbes, R., Doucet, J.P. und Derouin, F. Application of topological descriptors in QSAR and drug design: History and new trends. *Current Drug Targets-Infectious Disorders*, 2(1):93–102, 2002.

[GE03] Guyon, I. und Elisseeff, A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[GGAH14] Gupta, M., Gao, J., Aggarwal, C.C. und Han, J. Outlier detection for temporal data: A survey. *IEEE Trans. Knowl. Data Eng.*, 26(9):2250–2267, 2014.

[GHP+03] Giannella, C., Han, J., Pei, J., Yan, X. und Yu, P.S. Mining frequent patterns in data streams at multiple time granularities. *Next generation data mining*, 212:191–212, 2003.

[GM99] Gibbons, P.B. und Matias, Y. Synopsis data structures for massive data sets. *External memory algorithms*, 50:39–70, 1999.

[GMMO00] Guha, S., Mishra, N., Motwani, R. und O'Callaghan, L. Clustering data streams. In *Proc. Symp. Foundations of Computer Science (FOCS'00)*, S. 359–366. 2000.

[GMV94] Guyon, I., Matic, N. und Vapnik, V. Discovering informative patterns and data cleaning. In *Workshop Advances in Knowledge Discovery and Data Mining (KDD'94)*, S. 145–156. 1994.

[Gra15] Grandjean, M. Introduction à la visualisation de données : l'analyse de réseau en histoire. *Geschichte und Informatik*, S. 109–128, 2015.

[GS18] Gold, O. und Sharir, M. Dynamic time warping and geometric edit distance: Breaking the quadratic barrier. *ACM Trans. Algorithms*, 14(4), 2018.

[Gus97] Gusfield, D. *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology*. Cambridge University Press, 1997.

[GW94] Guthrie, L. und Walker, E.A. Document classification by machine: Theory and practice. In *Int. Conf. Computational Linguistics (COLING'94)*, S. 1059–1063. 1994.

[HAK00] Hinneburg, A., Aggarwal, C.C. und Keim, D.A. What is the nearest neighbor in high dimensional spaces? In *Proc. Int. Conf. Very Large Data Bases (VLDB'00)*, S. 506–515. 2000.

[Hav02] Haveliwala, T.H. Topic-sensitive pagerank. In *Proc. Int. World Wide Web Conf. (WWW'02)*, S. 517–526. 2002.

[HBV01] Halkidi, M., Batistakis, Y. und Vazirgiannis, M. On clustering validation techniques. *J. Intell. Inf. Syst.*, 17(2-3):107–145, 2001.

[HCC+07] Han, J., Cai, Y.D., Chen, Y., Dong, G., Pei, J., Wah, B.W. und Wang, J. Multi-dimensional analysis of data streams using stream cubes. In C.C. Aggarwal, Hrsg, *Data Streams: Models and Algorithms*, Kapitel 6, S. 103–125. Springer, 2007.

[HDY99] Han, J., Dong, G. und Yin, Y. Efficient mining of partial periodic patterns in time series database. In *Proc. Int. Conf. Data Engineering (ICDE'99)*, S. 106–115. 1999.

[HEK09] Hartung, J., Elpelt, B. und Klösener, K.H. *Statistik: Lehr- und Handbuch der angewandten Statistik*. Walter de Gruyter, 15. Auflage, 2009.

[HF94] Han, J. und Fu, Y. Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases. In *AAAI Workshop Knowledge Discovery in Databases*, S. 157–168. 1994.

[HF95] Han, J. und Fu, Y. Discovery of multiple-level association rules from large databases. In *Proc. Int. Conf. Very Large Data Bases (VLDB'95)*, S. 420–431. 1995.

[HFH+09] Hall, M.A., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. und Witten, I.H. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[HG07] Hinneburg, A. und Gabriel, H. DENCLUE 2.0: Fast clustering based on kernel density estimation. In *Proc. Int. Conf. Intelligent Data Analysis (IDA'07)*, S. 70–80. 2007.

[HK98] Hinneburg, A. und Keim, D.A. An efficient approach to clustering in large multimedia databases with noise. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, S. 58–65. 1998.

[HK00] Han, E. und Karypis, G. Centroid-based document classification: Analysis and experimental results. In *Proc. Eur. Symp. Principles of Data Mining and Knowledge Discovery (PKDD'00)*, S. 424–431. 2000.

[Hol18] Holtz, Y. Parallel coordinates plot. https://www.data-to-viz.com/graph/parallel.html, 2018. Abgerufen am 06.12.2023.

[HP07] Hua, M. und Pei, J. Cleaning disguised missing data: A heuristic approach. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'07)*, S. 950–958. 2007.

[HPT22] Han, J., Pei, J. und Tong, H. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 4. Auflage, 2022.

[HPY00] Han, J., Pei, J. und Yin, Y. Mining frequent patterns without candidate generation. In *Proc. Int. Conf. Management of Data (SIGMOD'00)*, S. 1–12. 2000.

[HRU96] Harinarayan, V., Rajaraman, A. und Ullman, J.D. Implementing data cubes efficiently. In *Proc. Int. Conf. Management of Data (SIGMOD'96)*, S. 205–216. 1996.

[HS54] Hopkins, B. und Skellam, J.G. A new method for determining the type of distribution of plant individuals. *Annals of Botany*, 18(2):213–227, April 1954.

[HSD01] Hulten, G., Spencer, L. und Domingos, P.M. Mining time-changing data streams. In *Proc. Int. Conf. Knowledge discovery and data mining (KDD'01)*, S. 97–106. 2001.

[HTF09] Hastie, T., Tibshirani, R. und Friedman, J.H. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2. Auflage, 2009.

[Hua98] Huang, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.*, 2(3):283–304, 1998.

[IWM00] Inokuchi, A., Washio, T. und Motoda, H. An apriori-based algorithm for mining frequent substructures from graph data. In *Proc. Eur. Symp. Principles of Data Mining and Knowledge Discovery (PKDD'00)*, S. 13–23. 2000.

[Jai10] Jain, A.K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[Jam85] James, M. *Classification Algorithms*. Wiley-Interscience, 1985.

[JD88] Jain, A.K. und Dubes, R.C. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[JL96] John, G.H. und Langley, P. Static versus dynamic sampling for data mining. In *Int. Conf. on Knowledge Discovery and Data Mining*, S. 367–370. 1996.

[Jl01] Jensen, A. und la Cour-Harbo, A. *Ripples in Mathematics - The Discrete Wavelet Transform*. Springer, 2001.

[Jol02] Jolliffe, I. *Principal Component Analysis*. Springer, 2. Auflage, 2002.

[JW02] Jeh, G. und Widom, J. Simrank: a measure of structural-context similarity. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'02)*, S. 538–543. 2002.

[JW07] Johnson, R.A. und Wichern, D.W. *Applied Multivariate Statistical Analysis*. Pearson, 6. Auflage, 2007.

[KCPM01] Keogh, E., Chakrabarti, K., Pazzani, M. und Mehrotra, S. Dimensionality reduction for fast similarity search in large time series databases. *J. Knowledge and Information Systems*, 3:263–286, 2001.

[Kei97] Keim, D.A. Visual techniques for exploring databases. In *Int. Conf. on Knowledge Discovery in Databases (KDD'97)*. 1997.

[Ker92] Kerber, R. ChiMerge: Discretization of numeric attributes. In *Nat. Conf. on Artificial Intelligence*, S. 123–128. 1992.

[KH97] Kononenko, I. und Hong, S.J. Attribute selection for modelling. *Future Generation Comp. Syst.*, 13(2-3):181–195, 1997.

[KHC97] Kamber, M., Han, J. und Chiang, J. Metarule-guided mining of multi-dimensional association rules using data cubes. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'97)*, S. 207–210. 1997.

[KHK99] Karypis, G., Han, E. und Kumar, V. Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer*, 32(8):68–75, 1999.

[KJ18] Kuhn, M. und Johnson, K. *Applied predictive modeling*. Springer, 2. Auflage, 2018.

[KK01] Kuramochi, M. und Karypis, G. Frequent subgraph discovery. In *Proc. Int. Conf. Data Mining (ICDM'01)*, S. 313–320. 2001.

[Kle99] Kleinberg, J.M. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[KLR+98] Kennedy, R.L., Lee, Y., van Roy, B., Reed, C.D. und Lippman, R.P. *Solving Data Mining Problems Through Pattern Recognition*. Prentice Hall, 1998.

[KLR04] Keogh, E.J., Lonardi, S. und Ratanamahatana, C. Towards parameter-free data mining. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'04)*, S. 206–215. 2004.

[KM94] Kivinen, J. und Mannila, H. The power of sampling in knowledge discovery. In *ACM Principles of Database Systems*, S. 77–85. 1994.

[KMR+94] Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H. und Verkamo, A.I. Finding interesting rules from large sets of discovered association rules. In *Proc. Int. Conf. Information and Knowledge Management (CIKM'94)*, S. 401–407. 1994.

[KNN05] Kutner, M.H., Nachtsheim, C.J. und Neter, J. *Applied Linear Statistical Models*. McGraw-Hill, 5. Auflage, 2005.

[Koh95] Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI'95)*, Band 2, S. 1137–1145. 1995.

[Kon04] Konstan, J.A. Introduction to recommender systems: Algorithms and evaluation. *ACM Trans. Inf. Syst.*, 22(1):1–4, 2004.

[KR05] Kaufman, L. und Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley, 2005.

[KRM05] Knauer, U., Reulke, R. und Meffert, B. Fahrzeugdetektion und -erkennung mittels mehrdimensionaler Farbhistogrammanalyse. In *Workshop Farbbildverarbeitung*, S. 93–100. 2005.

[Kru64] Kruskal, J.B. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.

[KTGA13] Kotsakos, D., Trajcevski, G., Gunopulos, D. und Aggarwal, C.C. Time-series data clustering. In *Data Clustering: Algorithms and Applications*, S. 357–380. 2013.

[LCH+09] Lo, D., Cheng, H., Han, J., Khoo, S. und Sun, C. Classification of software behaviors for failure detection: A discriminative pattern mining approach. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'09)*, S. 557–566. 2009.

[Leg23] Legitimate Data Company. Disk prices (US). https://diskprices.com, 2023. Abgerufen am 06.12.2023.

[Len02] Lenzerini, M. Data integration: A theoretical perspective. In *ACM Principles of Database Systems*, S. 233–246. 2002.

[LHTD02] Liu, H., Hussain, F., Tan, C.L. und Dash, M. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6(4):393–423, 2002.

[Li17] Li, R. History of data mining. https://hackerbits.com/data/history-of-data-mining, 2017. Abgerufen am 06.12.2023.

[Liu11] Liu, B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Data-Centric Systems and Applications. Springer, 2. Auflage, 2011.

[LJ90] Lawson, R.G. und Jurs, P.C. New index for clustering tendency and its application to chemical problems. *J. Chemical Information and Computer Sciences*, 30(1):36–41, 1990.

[LKCH03] Lee, Y., Kim, W., Cai, Y.D. und Han, J. Comine: Efficient mining of correlated patterns. In *Proc. Int. Conf. Data Mining (ICDM'03)*, S. 581–584. 2003.

[LLMZ04] Li, Z., Lu, S., Myagmar, S. und Zhou, Y. CP-Miner: A tool for finding copy-paste and related bugs in operating system code. In *Symp. Operating System Design and Implementation (OSDI'04)*, S. 289–302. 2004.

[Llo82] Lloyd, S.P. Least squares quantization in PCM. *IEEE Trans. Inf. Theor.*, 28(2):129–137, 1982. (ursprüngliche Version: Technical Report, Bell Labs, 1957).

[LLS00] Lim, T., Loh, W. und Shih, Y. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, 40(3):203–228, 2000.

[LR02] Little, R.J. und Rubin, D.B. *Statistical Analysis with Missing Data*. Wiley, 2. Auflage, 2002.

[LRU14] Leskovec, J., Rajaraman, A. und Ullman, J.D. *Mining of Massive Datasets*. Cambridge University Press, 2. Auflage, 2014.

[LSW97] Lent, B., Swami, A.N. und Widom, J. Clustering association rules. In *Proc. Int. Conf. Data Engineering (ICDE'97)*, S. 220–231. 1997.

[Lux07] von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[LWT+10] Lim, C.P., Wang, S.L., Tan, K.S., Navarro, J.C. und Jain, L.C. Use of the circle segments visualization technique for neural network feature selection and analysis. *Neurocomputing*, 73:613–621, 2010.

[Mac67] MacQueen, J.B. Some methods for classification and analysis of multivariate observations. In *Proc. Berkeley Symp. Math. Statist. Prob.*, Band 1, S. 281–297. 1967.

[MAE05] Metwally, A., Agrawal, D. und El Abbadi, A. Efficient computation of frequent and top-k elements in data streams. In *Proc. Int. Conf. Database Theory (ICDT'05)*, S. 398–412. 2005.

[Mar14] Marsland, S. *Machine Learning - An Algorithmic Perspective*. CRC Press, 2. Auflage, 2014.

[MB98] Messmer, B.T. und Bunke, H. A new algorithm for error-tolerant subgraph isomorphism detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(5):493–504, 1998.

[MC12] Murtagh, F. und Contreras, P. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[McC22] McCallum, J.C. Disk drive prices 1955+. https://jcmit.net/diskprice.htm, 2022. Abgerufen am 06.12.2023.

[Mei03] Meilă, M. Comparing clusterings by the variation of information. In *Computational Learning Theory and Kernel Machines (COLT'03)*, S. 173–187. 2003.

[Mei05] Meilă, M. Comparing clusterings: an axiomatic view. In *Int. Conf. Machine Learning (ICML'05)*, S. 577–584. 2005.

[MES95] Malerba, D., Esposito, F. und Semeraro, G. A further comparison of simplification methods for decision-tree induction. In *Learning from Data - Int. Workshop Artificial Intelligence and Statistics, AISTATS*, S. 365–374. 1995.

[Min89] Mingers, J. An empirical comparison of pruning methods for decision tree induction. *Machine Learning*, 4:227–243, 1989.

[Mit97] Mitchell, T.M. *Machine learning*. McGraw-Hill, 1997.

[Mit10] Mitsa, T. *Temporal Data Mining*. Chapman & Hall/CRC, 2010.

[MKZ+09] Mueen, A., Keogh, E.J., Zhu, Q., Cash, S. und Westover, M.B. Exact discovery of time series motifs. In *Proc. Int. Conf. Data Mining (SDM'09)*, S. 473–484. 2009.

[MM95] Major, J.A. und Mangano, J.J. Selecting among rules induced from a hurricane database. *J. Intell. Inf. Syst.*, 4(1):39–52, 1995.

[MM02] Manku, G.S. und Motwani, R. Approximate frequency counts over data streams. In *Proc. Int. Conf. Very Large Data Bases (VLDB'02*, S. 346–357. 2002.

[MMBR14] Moorthy, T.N., Mostapa, A.M.B., Boominathan, R. und Raman, N. Stature estimation from footprint measurements in indian tamils by regression analysis. *Egyptian Journal of Forensic Sciences*, 4(1):7–16, 2014.

[MMHL86] Michalski, R.S., Mozetic, I., Hong, J. und Lavrac, N. The multi-purpose incremental learning system AQ15 and its testing application to three medical domains. In *Proc. Nat. Conf. Artificial Intelligence*, S. 1041–1047. 1986.

[MN98] McCallum, A. und Nigam, K. A comparison of event models for naive bayes text classification. In *AAAI Workshop on Learning for Text Categorization*, Band 752, S. 41–48. 1998.

[MR13] Mooney, C. und Roddick, J.F. Sequential pattern mining - approaches and algorithms. *ACM Comput. Surv.*, 45(2):19:1–19:39, 2013.

[MRS08] Manning, C.D., Raghavan, P. und Schütze, H. *Introduction to information retrieval*. Cambridge University Press, 2008.

[Mur98] Murthy, S.K. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Min. Knowl. Discov.*, 2(4):345–389, 1998.

[Mut05] Muthukrishnan, S.M. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2):117–236, 2005.

[MXC+07] Mei, Q., Xin, D., Cheng, H., Han, J. und Zhai, C. Semantic annotation of frequent patterns. *ACM Trans. Knowledge Discovery from Data (TKDD'07)*, 1(3):11, 2007.

[Nat23] National Oceanic and Atmospheric Administration. Climate at a glance. https://www.ncdc.noaa.gov/cag/global/time-series/globe/land_ocean/ann/12/1950-2020, 2023. Abgerufen am 06.12.2023.

[NH94] Ng, R.T. und Han, J. Efficient and effective clustering methods for spatial data mining. In *Proc. Int. Conf. Very Large Data Bases (VLDB'94)*, S. 144–155. 1994.

[NLHP98] Ng, R.T., Lakshmanan, L.V.S., Han, J. und Pang, A. Exploratory mining and pruning optimizations of constrained association rules. In *Proc. Int. Conf. Management of Data (SIGMOD'98)*, S. 13–24. 1998.

[Ols03] Olson, J.E. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, 2003.

[Omi03] Omiecinski, E. Alternative interest measures for mining associations in databases. *IEEE Trans. Knowl. Data Eng.*, 15(1):57–69, 2003.

[OMM+02] O'Callaghan, L., Meyerson, A., Motwani, R., Mishra, N. und Guha, S. Streaming-data algorithms for high-quality clustering. In *Proc. Int. Conf. Data Engineering (ICDE'02)*, S. 685–694. 2002.

[Pal13] Palpanas, T. Real-time data analytics in sensor networks. In *Managing and Mining Sensor Data*, S. 173–210. 2013.

[PBTL99] Pasquier, N., Bastide, Y., Taouil, R. und Lakhal, L. Discovering frequent closed itemsets for association rules. In *Proc. Int. Conf. Database Theory (ICDT'99)*, S. 398–416. 1999.

[PCT+03] Pan, F., Cong, G., Tung, A.K.H., Yang, J. und Zaki, M.J. Carpenter: finding closed patterns in long biological datasets. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, S. 637–642. 2003.

[PCY95] Park, J.S., Chen, M. und Yu, P.S. An effective hash based algorithm for mining association rules. In *Proc. Int. Conf. Management of Data (SIGMOD'95)*, S. 175–186. 1995.

[PD11] Peck, R. und Devore, J.L. *Statistics: The Exploration & Analysis of Data*. Brooks/Cole, Cengage Learning, 7. Auflage, 2011.

[PHL01] Pei, J., Han, J. und Lakshmanan, L.V.S. Mining frequent item sets with convertible constraints. In *Proc. Int. Conf. Data Engineering (ICDE'01)*, S. 433–442. 2001.

[PHL04] Parsons, L., Haque, E. und Liu, H. Subspace clustering for high dimensional data: a review. *SIGKDD Explorations*, 6(1):90–105, 2004.

[PHM00] Pei, J., Han, J. und Mao, R. CLOSET: An efficient algorithm for mining frequent closed itemsets. In *Proc. Int. Conf. Management of Data (SIGMOD'00)*, S. 21–30. 2000.

[PHM+01] Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U. und Hsu, M. PrefixSpan: Mining sequential patterns by prefix-projected growth. In *Proc. Int. Conf. Data Engineering (ICDE'01)*, S. 215–224. 2001.

[PHM+04] Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U. und Hsu, M. Mining sequential patterns by pattern-growth: The PrefixSpan approach. *IEEE Trans. Knowledge and Data Engineering*, 16(11):1424–1440, 2004.

[Pia91a] Piatetsky-Shapiro, G. Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, S. 229–248. AAAI/MIT Press, 1991.

[Pia91b] Piatetsky-Shapiro, G. Notes of AAAI workshop knowledge discovery in databases. In *Proceedings of AAAI*, Band 91. 1991.

[PKLL02] Patel, P., Keogh, E.J., Lin, J. und Lonardi, S. Mining motifs in massive time series databases. In *Proc. Int. Conf. Data Mining (ICDM'02)*, S. 370–377. 2002.

[PTVF07] Press, W.H., Teukolsky, S.A., Vetterling, W.T. und Flannery, B.P. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 3. Auflage, 2007.

[QC93] Quinlan, J.R. und Cameron-Jones, R.M. FOIL: A midterm report. In *Proc. European Conf. Machine Learning (ECML'93)*, S. 3–20. 1993.

[QR89] Quinlan, J.R. und Rivest, R.L. Inferring decision trees using the minimum description length principle. *Inf. Comput.*, 80(3):227–248, 1989.

[Qui86] Quinlan, J.R. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[Qui87] Quinlan, J.R. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3):221–234, 1987.

[Qui88] Quinlan, J.R. An empirical comparison of genetic and decision-tree classifiers. In *Proc. Int. Conf. Machine Learning*, S. 135–141. 1988.

[Qui89] Quinlan, J.R. Unknown attribute values in induction. In *Int. Machine Learning Workshop*, S. 164–168. 1989.

[Qui90] Quinlan, J.R. Learning logical definitions from relations. *Machine Learning*, 5:239–266, 1990.

[Qui93] Quinlan, J.R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[Qui96] Quinlan, J.R. Bagging, boosting, and C4.5. In *Proc. Nat. Conf. Artificial Intelligence (AAAI'96)*, Band 1, S. 725–730. 1996.

[Ray23] Rayaprolu, A. How much data is created every day in 2023? https://techjury.net/blog/how-much-data-is-created-every-day, 2023. Abgerufen am 06.12.2023.

[Red92] Redman, T.C. *Data Quality: Management and Technology*. Bantam Books, 1992.

[Rij79] van Rijsbergen, C.J. *Information Retrieval*. Butterworth, 2. Auflage, 1979.

[RN21] Russell, S.J. und Norvig, P. *Artificial Intelligence: A Modern Approach*. Pearson Education, 4. Auflage, 2021.

[Ros17] Rose, F. CLIQUE: Clustering in quest. https://list01.biologie.ens.fr/wws/d_read/machine_learning/SubspaceClustering/CLIQUE_algorithm_grid-based_subspace_clustering.pdf, 2017. Abgerufen am 06.12.2023.

[RS05] Rennie, J.D.M. und Srebro, N. Fast maximum margin matrix factorization for collaborative prediction. In *Proc. Int. Conf. Machine Learning (ICML'05)*, S. 713–719. 2005.

[Rul22] Rulequest Research. Data mining tools See5 and C5.0. http://www.rulequest.com/see5-info.html, 2022. Abgerufen am 06.12.2023.

[RVSZ+17] Ramos, R., Valdez-Salas, B., Zlatev, R., Schorr Wiener, M. und Bastidas Rull, J.M. The discrete wavelet transform and its application for noise removal in localized corrosion measurements. *International Journal of Corrosion*, 2017.

[RY98] Ristad, E.S. und Yianilos, P.N. Learning string-edit distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(5):522–532, 1998.

[SA95] Srikant, R. und Agrawal, R. Mining generalized association rules. In *Proc. Int. Conf. Very Large Data Bases (VLDB'95)*, S. 407–419. 1995.

[SA96] Srikant, R. und Agrawal, R. Mining sequential patterns: Generalizations and performance improvements. In *Int. Conf. Extending Database Technology (EDBT'96)*, S. 3–17. 1996.

[Sch90] Schapire, R.E. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

[SCZ98] Sheikholeslami, G., Chatterjee, S. und Zhang, A. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In *Proc. Int. Conf. Very Large Data Bases (VLDB'98)*, S. 428–439. 1998.

[SD16] Saini, S. und Dewan, L. Application of discrete wavelet transform for analysis of genomic sequences of mycobacterium tuberculosis. *SpringerPlus*, 5(1):64, 2016.

[SDS96] Stollnitz, E.J., Derose, T.D. und Salesin, D.H. *Wavelets for Computer Graphics: Theory and Applications*. Morgan Kaufmann, 1996.

[SE10] Seni, G. und Elder, J.F. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool, 2010.

[Seb02] Sebastiani, F. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[SKKR01] Sarwar, B.M., Karypis, G., Konstan, J.A. und Riedl, J. Item-based collaborative filtering recommendation algorithms. In *Proc. Int. World Wide Web Conf. (WWW'01)*, S. 285–295. 2001.

[SM84] Salton, G. und McGill, M. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1984.

[SON95] Savasere, A., Omiecinski, E. und Navathe, S.B. An efficient algorithm for mining association rules in large databases. In *Proc. Int. Conf. Very Large Data Bases (VLDB'95)*, S. 432–444. 1995.

[SS94] Sarawagi, S. und Stonebraker, M. Efficient organization of large multidimensional arrays. In *Proc. Int. Conf. Data Engineering (ICDE'94)*, S. 328–336. 1994.

[SS17] Shumway, R.H. und Stoffer, D.S. *Time Series Analysis and Its Applications: With R Examples*. Springer, 4. Auflage, 2017.

[Sta23] Statista. Weltweite und europäische Kunststoffproduktion in den Jahren von 1950 bis 2022. https://de.statista.com/statistik/daten/studie/167099/umfrage/weltproduktion-von-kunststoff-seit-1950, 2023. Abgerufen am 06.12.2023.

[Sto74] Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. Royal Statistical Society*, 36:111–147, 1974.

[SZLY08] Shang, H., Zhang, Y., Lin, X. und Yu, J.X. Taming verification hardness: an efficient algorithm for testing subgraph isomorphism. *Proc. Int. Conf. Very Large Data Bases (VLDB'08)*, 1(1):364–375, 2008.

[TKS02] Tan, P., Kumar, V. und Srivastava, J. Selecting the right interestingness measure for association patterns. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'02)*, S. 32–41. 2002.

[Toi96] Toivonen, H. Sampling large databases for association rules. In *Proc. Int. Conf. Very Large Data Bases (VLDB'96)*, S. 134–145. 1996.

[TSKK19] Tan, P.N., Steinbach, M., Karpatne, A. und Kumar, V. *Introduction to Data Mining*. Pearson, 2. Auflage, 2019.

[Tuf90] Tufte, E.R. *Envisioning Information*. Graphics Press, 1990.

[Tuf97] Tufte, E.R. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, 2. Auflage, 1997.

[Tuf01] Tufte, E.R. *The Visual Display of Quantitative Information*. Graphics Press, 2. Auflage, 2001.

[Ull76] Ullmann, J.R. An algorithm for subgraph isomorphism. *J. ACM*, 23(1):31–42, 1976.

[Uni22] University of Waikato. Weka 3 - data mining with open source machine learning software in java. https://www.cs.waikato.ac.nz/ml/weka, 2022. Abgerufen am 06.12.2023.

[VC06] Vuk, M. und Curk, T. ROC curve, lift chart and calibration plot. *Metodološki zvezki*, 3(1):89–108, 2006.

[Vit85] Vitter, J.S. Random sampling with a reservoir. *ACM Trans. Mathematical Software*, 11(1):37–57, 1985.

[WB98] Westphal, C.R. und Blaxton, T. *Data mining solutions: methods and tools for solving real-world problems*. Wiley, 1998.

[WCH10] Wu, T., Chen, Y. und Han, J. Re-examination of interestingness measures in pattern mining: a unified framework. *Data Min. Knowl. Discov.*, 21(3):371–397, 2010.

[WFHP16] Witten, I.H., Frank, E., Hall, M.A. und Pal, C.J. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 4. Auflage, 2016.

[WFYH03] Wang, H., Fan, W., Yu, P.S. und Han, J. Mining concept-drifting data streams using ensemble classifiers. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, S. 226–235. 2003.

[WHH00] Wang, K., He, Y. und Han, J. Mining frequent itemsets using support constraints. In *Proc. Int. Conf. Very Large Data Bases (VLDB'00)*, S. 43–52. 2000.

[WHLT05] Wang, J., Han, J., Lu, Y. und Tzvetkov, P. TFP: An efficient algorithm for mining top-k frequent closed itemsets. *IEEE Trans. Knowl. Data Eng.*, 17(5):652–664, 2005.

[WHP03] Wang, J., Han, J. und Pei, J. CLOSET+: Searching for the best strategies for mining frequent closed itemsets. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, S. 236–245. 2003.

[Win14] Winner, L. Left footprint length and height in indian adult male tamils. http://users.stat.ufl.edu/~winner/data/india_foot_height.dat, 2014. Abgerufen am 06.12.2023.

[WK91] Weiss, S.M. und Kulikowski, C.A. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, 1991.

[WYM97] Wang, W., Yang, J. und Muntz, R.R. STING: A statistical information grid approach to spatial data mining. In *Proc. Int. Conf. Very Large Data Bases (VLDB'97)*, S. 186–195. 1997.

[WZZ+07] Wang, J., Zhang, Y., Zhou, L., Karypis, G. und Aggarwal, C.C. Discriminating subsequence discovery for sequence clustering. In *Proc. SIAM Int. Conf. Data Mining*, S. 605–610. 2007.

[XCYH06] Xin, D., Cheng, H., Yan, X. und Han, J. Extracting redundancy-aware top-k patterns. In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'06)*, S. 444–453. 2006.

[YF00] Yi, B. und Faloutsos, C. Fast time sequence indexing for arbitrary lp norms. In *Proc. Int. Conf. Very Large Data Bases (VLDB'00)*, S. 385–394. 2000.

[YH02] Yan, X. und Han, J. gSpan: Graph-based substructure pattern mining. In *Proc. Int. Conf. Data Mining (ICDM'02)*, S. 721–724. 2002.

[YJF98] Yi, B., Jagadish, H.V. und Faloutsos, C. Efficient retrieval of similar time sequences under time warping. In *Proc. Int. Conf. Data Engineering (ICDE'98)*, S. 201–208. 1998.

[YK08] Yildirim, H. und Krishnamoorthy, M.S. A random walk method for alleviating the sparsity problem in collaborative filtering. In *Proc. Conf. Recommender Systems (RecSys'08)*, S. 131–138. 2008.

[YW03] Yang, J. und Wang, W. CLUSEQ: efficient and effective sequence clustering. In *Proc. Int. Conf. Data Engineering (ICDE'03)*, S. 101–112. 2003.

[YYH04] Yan, X., Yu, P.S. und Han, J. Graph indexing: A frequent structure-based approach. In *Proc. Int. Conf. Management of Data (SIGMOD'04)*, S. 335–346. 2004.

[Zak00] Zaki, M.J. Scalable algorithms for association mining. *IEEE Trans. Knowl. Data Eng.*, 12(3):372–390, 2000.

[Zak01] Zaki, M.J. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1/2):31–60, 2001.

[ZRL96] Zhang, T., Ramakrishnan, R. und Livny, M. BIRCH: An efficient data clustering method for very large databases. In *Proc. Int. Conf. Management of Data (SIGMOD'96)*, S. 103–114. 1996.

[ZRL97] Zhang, T., Ramakrishnan, R. und Livny, M. BIRCH: A new data clustering algorithm and its applications. *Data Min. Knowl. Discov.*, 1(2):141–182, 1997.

[ZWFM06] Zhang, S., Wang, W., Ford, J. und Makedon, F. Learning from incomplete ratings using non-negative matrix factorization. In *Proc. SIAM Int. Conf. Data Mining*, S. 549–553. 2006.

[ZYHY07] Zhu, F., Yan, X., Han, J. und Yu, P.S. gprune: A constraint pushing framework for graph pattern mining. In *Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, (PAKDD'07)*, S. 388–400. 2007.