Thomas A. Runkler

Data Mining

Methoden und Algorithmen intelligenter Datenanalyse

STUDIUM

VIEWEG+
TEUBNER

Thomas A. Runkler

Data Mining

Computational Intelligence

herausgegeben von
Wolfgang Bibel, Rudolf Kruse und Bernhard Nebel

Aus den Kinderschuhen der „Künstlichen Intelligenz“ entwachsen bietet die Reihe breitgefächertes Wissen von den Grundlagen bis in die Anwendung, herausgegeben von namhaften Vertretern ihres Faches.

Computational Intelligence hat das weitgesteckte Ziel, das Verständnis und die Realisierung intelligenten Verhaltens voranzutreiben. Die Bücher der Reihe behandeln Themen aus den Gebieten wie z. B. Künstliche Intelligenz, Softcomputing, Robotik, Neuro- und Kognitionswissenschaften. Es geht sowohl um die Grundlagen (in Verbindung mit Mathematik, Informatik, Ingenieurs- und Wirtschaftswissenschaften, Biologie und Psychologie) wie auch um Anwendungen (z. B. Hardware, Software, Webtechnologie, Marketing, Vertrieb, Entscheidungsfindung). Hierzu bietet die Reihe Lehrbücher, Handbücher und solche Werke, die maßgebliche Themengebiete kompetent, umfassend und aktuell repräsentieren.

Unter anderem sind erschienen:

Methoden wissensbasierter Systeme
von Christoph Beierle und Gabriele Kern-Isberner

Neuro-Fuzzy-Systeme
von Detlef Nauck, Christian Borgelt, Frank Klawonn und Rudolf Kruse

Evolutionäre Algorithmen
von Ingrid Gerdes, Frank Klawonn und Rudolf Kruse

Grundkurs Spracherkennung
von Stephan Euler

Quantum Computing verstehen
von Matthias Homeister

www.viewegteubner.de

Thomas A. Runkler

Data Mining

Methoden und Algorithmen
intelligenter Datenanalyse

Mit 72 Abbildungen und 7 Tabellen

STUDIUM

VIEWEG+
TEUBNER

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
<http://dnb.d-nb.de> abrufbar.

Höchste inhaltliche und technische Qualität unserer Produkte ist unser Ziel. Bei der Produktion und Auslieferung unserer Bücher wollen wir die Umwelt schonen: Dieses Buch ist auf säurefreiem und chlorfrei gebleichtem Papier gedruckt. Die Einschweißfolie besteht aus Polyäthylen und damit aus organischen Grundstoffen, die weder bei der Herstellung noch bei der Verbrennung Schadstoffe freisetzen.

1. Auflage 2010

Lektorat: Christel Roß | Walburga Himmel

Vieweg+Teubner ist Teil der Fachverlagsgruppe Springer Science+Business Media.
www.viewegteubner.de

Umschlaggestaltung: KünkelLopka Medienentwicklung, Heidelberg
Technische Redaktion: FROMM MediaDesign, Selters/Ts.
Druck und buchbinderische Verarbeitung: Krips b.v., Meppel
Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.
Printed in the Netherlands

ISBN 978-3-8348-0858-5

Vorwort

Die Information in der Welt verdoppelt sich etwa alle 20 Monate. Zu den wichtigsten Datenquellen gehören das Internet, industrielle und geschäftliche Prozesse, Bilderfassungssysteme und die Biomedizin. Diese Quellen liefern große Mengen numerischer Daten, Bilddaten, Textdaten und strukturierter Daten. Neben der Erfassung und Speicherung stellen die Aufbereitung, Analyse und Nutzung dieser Daten heute die größten Herausforderungen dar. Ziel ist es, aus den großen Datenmengen die relevanten Informationen, also das „Wissen“, zu extrahieren. Neben konventionellen statistischen Verfahren wie Korrelation und Regression werden hierzu Methoden aus den Bereichen Signaltheorie, Mustererkennung, Clusteranalyse, Neuroinformatik, Fuzzy-Systeme, evolutionäre Algorithmen, Schwarmintelligenz und maschinelles Lernen angewandt. Für die Analyse nichtnumerischer Daten sind relationale und strukturbasierte Algorithmen notwendig. Diese Datenanalysemethoden für numerische und nichtnumerische Daten werden unter dem Sammelbegriff „Data Mining“ zusammengefasst. Data Mining ist ein Prozess, der auch die Datenvorverarbeitung, Filterung und Visualisierung umfasst.

Dieses Buch gibt eine strukturierte Einführung in den Data-Mining-Prozess und die wichtigsten Data-Mining-Methoden. Die Gliederung orientiert sich am typischen schrittweisen Ablauf von Datenanalyse-Projekten. Das Buch richtet sich an Ingenieure, Informatiker und Mathematiker in Forschung und Lehre, an Studenten dieser Fachgebiete, aber auch an Praktiker, die mit großen Datenmengen arbeiten. Zum Verständnis werden lediglich grundlegende Mathematik-Kenntnisse vorausgesetzt.

Der Stoff dieses Buches basiert auf Vorlesungen über maschinelle Lernverfahren, Data Mining und Clusteranalyse, die der Autor seit vielen Jahren regelmäßig an der Fakultät für Informatik der Technischen Universität München hält. Das Buch enthält Ergebnisse aus zahlreichen Forschungs- und Entwicklungsprojekten im Learning Systems Department bei Siemens Corporate Technology in München.

Die Erstauflage dieses Buches erschien im Jahr 2000 im Vieweg-Verlag unter dem Titel „Information Mining“ (101). Für diese grundlegend überarbeitete Neuauflage wurden sämtliche Inhalte neu strukturiert und sorgfältig aktualisiert. Neu hinzugefügt wurden die Abschnitte Ähnlichkeitsmaße, Chi-Quadrat-Test, Merkmalsselektion, Zeitreihenprognose, naiver Bayes-Klassifikator, lineare Diskriminanzanalyse, Support-Vektor-Maschine und relationales Clustering.

Ich danke den Kollegen der Fakultät für Informatik der Technischen Universität München, insbesondere Herrn Prof. Dr. Dr. h.c. mult. Wilfried Brauer, Herrn Prof. Dr. Javier Esparza, Herrn Dr. Clemens Kirchmair, Herrn Dr. Volker Baier, Herrn Dipl.-Inform. Achim Rettinger und Frau Erika Leber für die Unterstützung bei der Planung und Realisierung der Vorlesungen. Mein besonderer Dank gilt meinem verstorbenen Kollegen Herrn Prof. Dr. Bernd Schürmann für die persönliche Förderung und die Unterstützung meiner Forschungs- und Lehrtätigkeiten. Für den langjährigen anregenden wissenschaftlichen Austausch danke ich Herrn Prof. Dr. James C. Bezdek, Herrn Prof. Dr. Eyke Hüllermeier, Herrn Prof. Dr. Uzay Kaymak, Herrn Prof. Dr. Jim Keller, Herrn Prof. Dr. Frank Klawonn, Herrn Prof. Dr. Rudolf Kruse und Herrn Prof. Dr. João M. Sousa. Für die gute Zusammenarbeit danke ich meinen Kollegen und ehemaligen Kollegen von Siemens Corporate Technology in München, insbesondere Herrn Dr. Ralph Grothmann, Herrn Prof. Dr. Hans Hellendoorn, Herrn Dr. Jürgen Hollatz, Herrn Prof. Dr. Rainer Palm, Herrn Dr. Martin Schlang, Herrn Volkmar Sterzing und Herrn Dr. Hans-Georg Zimmermann. Mein besonderer Dank gilt auch den zahlreichen Lesern, Studierenden und Rezensenten, die mich auf Ungereimtheiten, Fehler und Verbesserungsmöglichkeiten aufmerksam gemacht und damit maßgeblich zur Überarbeitung des Buches beigetragen haben. Außerdem danke ich den Herausgebern der Reihe, Herrn Prof. Dr. Wolfgang Bibel, Herrn Prof. Dr. Rudolf Kruse, Herrn Prof. Dr. Bernhard Nebel, dem Vieweg+Teubner-Verlag, insbesondere Frau Andrea Broßler, Frau Dr. Christel Anne Roß und Frau Sybille Thelen, sowie Frau Angela Fromm für die gute Zusammenarbeit bei der Konzipierung und Veröffentlichung dieses Buches. Nicht zuletzt danke ich meiner Familie, Anja, Marisa und Moritz, für die Unterstützung und Geduld während der vielen Stunden, die ich diesem Buch gewidmet habe.

München, im September 2009 Thomas Runkler

Inhaltsverzeichnis

Kapitel 1

Data-Mining-Prozess

Der Fokus dieses Buches ist die Analyse großer Datenmengen, zum Beispiel:

- Industrielle Prozessdaten: Zur Automatisierung industrieller Fertigungsprozesse werden zunehmend Prozessdaten erfasst und gespeichert. Die Informationstechnik fließt in alle Ebenen der Automatisierungspyramide ein und betrifft somit Signale von Sensoren und Aktuatoren in der Feldebene ebenso wie Steuerungs- und Regelungssignale in der Steuerungsebene, Bedienungs- und Beobachtungsdaten in der Prozessleitebene sowie Planungsdaten und Kennzahlen in der Betriebsleit- und Unternehmensebene. Auf allen diesen Ebenen lassen sich mit Data Mining Potenziale zur Prozessoptimierung und Steigerung der Wettbewerbsfähigkeit erschließen.

- Marketing: Modernes Marketing basiert auf der Nutzung umfangreicher Daten zu Kunden, Produkten und Absätzen. In der Warenkorbanalyse wird ausgewertet, welche Produkte in welchen Stückzahlen mit welchen anderen Produkten zusammen gekauft wurden, zum Beispiel im Supermarkt. Die Auswertung hat das Ziel, Produkte optimal zu platzieren und die Preisgestaltung zu optimieren. Ein anderes wichtiges Anwendungsgebiet ist die datenbasierte Kundensegmentierung, die für gezielte Kundenangebote und Werbemaßnahmen genutzt werden kann.

- Textdaten und strukturierte Daten: Neben den oben genannten numerischen Daten stehen zunehmend auch Textdaten und strukturierte Daten zur Analyse zur Verfügung. Zu den wichtigsten Quellen von Textdaten gehören konventionelle Textdokumente ebenso wie internet-basierte Dokumente wie E-Mail, World Wide Web und web-basierte Datenbanken (Deep Web). Data Mining hilft, die verfügbaren Informationen zu filtern, gewünschte Informationen zu finden und tiefergehende Analysen der verfügbaren Informationen durchzuführen. Zur Spezifikation und Speicherung von objektbezogenen Informationen werden außer unstrukturiertem Text zunehmend auch strukturierte Formate verwendet, in denen die Se-

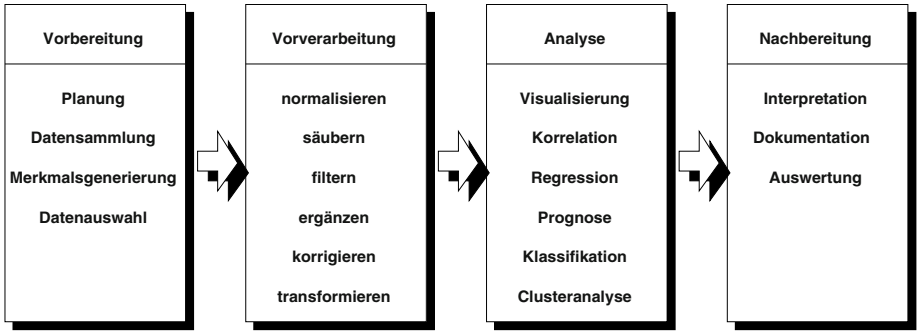

Abbildung 1.1: Der Data-Mining-Prozess

mantik von Objekten und deren Beziehungen untereinander spezifiziert werden können (Ontologien).

- Bilddaten: Durch die Verfügbarkeit von preisgünstigen elektronischen Kameras und satellitenbasierte Bildquellen (wie Google Earth) stehen große Mengen von Bilddaten für die Analyse zur Verfügung. Mit Data-Mining-Methoden werden Objekte gesucht und erkannt, Szenen analysiert und klassifiziert, und Bilddaten mit anderen Informationen in Beziehung gesetzt.

- Biomedizinische Daten: Ein wichtiges Anwendungsgebiet der biomedizinischen Informatik ist die Erbgutanalyse. Hierzu stehen zum Beispiel im Internet zahlreiche Datenbanken mit sequenzierte Genomdaten zu Verfügung. Zu den Zielen der Analyse gehören die Bestimmung von Genfunktionen und die Vorhersage von Strukturen.

Data Mining hat das Ziel, Wissen aus Daten zu extrahieren (11). Unter Wissen verstehen wir interessante Muster, und unter interessanten Mustern schließlich solche, die allgemein gültig sind, nicht trivial, neu, nützlich und verständlich (35). Inwieweit diese Eigenschaften auf die erkannten Muster zutreffen, muss von Anwendung zu Anwendung neu beurteilt werden, und zwar in der Regel gemeinsam mit Anwendungsexperten. Durch die Einbeziehung der Experten entsteht ein rückgekoppelter Prozess, der oft wiederholt durchlaufen wird, bis ein zufriedenstellendes Ergebnis erreicht ist.

In den meisten Anwendungen stehen zunächst nur Rohdaten zur Verfügung. Diese Rohdaten sind in der Regel ungeordnet, fehlerhaft, unvollständig, teilweise redundant oder unwichtig. Bevor also die eigentliche Datenanalyse durchgeführt werden kann, ist eine Vorbereitung und Vorverarbeitung der Daten notwendig. Dieser Ablauf ist in Abbildung 1.1 dargestellt. Häufig sind zu Beginn eines Projekts überhaupt noch keine nutzbaren Daten vorhanden, so dass zunächst die Datensammlung geplant und durchgeführt werden muss. Zur Datenvorbe-

reitung gehört weiter, dass aus den teilweise unübersichtlichen Daten geeignete Merkmale generiert und die überhaupt interessanten Daten ausgewählt werden. Nach der Datenvorbereitung steht ein Rohdatensatz zur Verfügung, der in der anschließenden Vorverarbeitung normalisiert, gesäubert, gefiltert oder transformiert wird; fehlende Einträge werden ergänzt und Fehler korrigiert. Der auf diese Weise erhaltene vorverarbeitete Datensatz kann anschließend mit verschiedenen Verfahren der Datenanalyse verarbeitet werden. Dabei spielen Visualisierungsverfahren eine große Rolle, aber auch reine Analysemethoden wie Korrelation, Regression, Prognose, Klassifikation und Clusteranalyse. Die Ergebnisse der Analyse können meist nicht unmittelbar verwendet werden, sondern müssen nachbereitet, also interpretiert, dokumentiert und ausgewertet werden. Die in jedem Verarbeitungsblock angegebenen Methoden stellen nur typische Beispiele dar, die keinen Anspruch auf Vollständigkeit erheben. Die tatsächlich benötigten Verarbeitungsmethoden sind von Anwendung zu Anwendung verschieden und stark von Art, Qualität und Quantität der Daten abhängig. Außerdem findet der Datenanalyseprozess in der Regel nicht streng sequentiell statt, wie in Abbildung 1.1 gezeigt, sondern springt je nach Ergebnis der einzelnen Schritte zu früheren Schritten zurück. Zum Beispiel werden in der Mustererkennung häufig neue fehlerhafte Daten erkannt, die eine erneute Datensäuberung notwendig machen.

Kapitel 2

Daten und Relationen

Zur Einführung wird in diesem Kapitel der in der Datenanalyse sehr häufig verwendete Iris-Datensatz vorgestellt. An diesem Beispiel werden einige grundlegende Begriffe und beispielhafte Fragestellungen erläutert. Daten können im Allgemeinen sehr unterschiedliche Charakteristika haben. Zum Beispiel können Daten numerisch oder nichtnumerisch sein, nichtnumerische Daten können Text oder strukturierte Information enthalten, numerische Daten können zu Vektoren oder Sequenzen zusammengefasst sein, sie können auf unterschiedlichen Skalen messbar sein, können Abtastwerte von Zeitsignalen sein und können quantisierte oder kontinuierliche Werte besitzen. Diese unterschiedlichen Charakteristika und ihre Konsequenzen für die Datenanalyse sind im Folgenden beschrieben.

2.1 Beispiel

Der Iris-Datensatz (3) wurde 1935 von dem amerikanischen Botaniker Edgar Anderson erstellt, der die geographischen Unterschiede der Schwertlilien (Iris) untersuchte. Die von Anderson erhobenen Daten über Schwertlilien auf der Gaspé-Halbinsel in Quebec (Kanada) wurden erstmals von Sir Ronald Aylmer Fisher 1936 als Beispiel für die Diskriminanzanalyse multivariater Daten verwendet (36) und entwickelten sich später zu einem der meistverwendeten Referenzdatensätze in Statistik und Datenanalyse. Der Iris-Datensatz umfasst Messdaten von 150 Irispflanzen, davon jeweils 50 Borsten-Schwertlilien (Iris Setosa), Virginia-Schwertlilien (Iris Virginica) und verschiedenfarbige Schwertlilien (Iris Versicolor). Der Datensatz enthält von jeder dieser 150 Pflanzen die Länge und Breite eines Kelchblatts sowie die Länge und Breite eines Blütenblatts (in Zentimetern). Den kompletten Datensatz zeigt Tabelle 2.1. In der Datenanalyse bezeichnen wir jede der 150 Irispflanzen als ein *Objekt*, jede der 3 Pflanzenarten als eine *Klasse* und jeden der 4 Messwerte als ein *Merkmal*. Typische Fragen, die uns bei der Analyse dieser Daten interessieren, sind:

- Welche dieser Daten enthalten möglicherweise Messfehler oder falsche Klassenzuordnungen?

Tabelle 2.1: Iris-Datensatz (3)

Setosa				Versicolor				Virginica			
Kelchblatt		Blütenblatt		Kelchblatt		Blütenblatt		Kelchblatt		Blütenblatt	
Länge	Breite	Länge	Breite	Länge	Breite	Länge	Breite	Länge	Breite	Länge	Breite
5,1	3,5	1,4	0,2	7	3,2	4,7	1,4	6,3	3,3	6	2,5
4,9	3	1,4	0,2	6,4	3,2	4,5	1,5	5,8	2,7	5,1	1,9
4,7	3,2	1,3	0,2	6,9	3,1	4,9	1,5	7,1	3	5,9	2,1
4,6	3,1	1,5	0,2	5,5	2,3	4	1,3	6,3	2,9	5,6	1,8
5	3,6	1,4	0,2	6,5	2,8	4,6	1,5	6,5	3	5,8	2,2
5,4	3,9	1,7	0,4	5,7	2,8	4,5	1,3	7,6	3	6,6	2,1
4,6	3,4	1,4	0,3	6,3	3,3	4,7	1,6	4,9	2,5	4,5	1,7
5	3,4	1,5	0,2	4,9	2,4	3,3	1	7,3	2,9	6,3	1,8
4,4	2,9	1,4	0,2	6,6	2,9	4,6	1,3	6,7	2,5	5,8	1,8
4,9	3,1	1,5	0,1	5,2	2,7	3,9	1,4	7,2	3,6	6,1	2,5
5,4	3,7	1,5	0,2	5	2	3,5	1	6,5	3,2	5,1	2
4,8	3,4	1,6	0,2	5,9	3	4,2	1,5	6,4	2,7	5,3	1,9
4,8	3	1,4	0,1	6	2,2	4	1	6,8	3	5,5	2,1
4,3	3	1,1	0,1	6,1	2,9	4,7	1,4	5,7	2,5	5	2
5,8	4	1,2	0,2	5,6	2,9	3,6	1,3	5,8	2,8	5,1	2,4
5,7	4,4	1,5	0,4	6,7	3,1	4,4	1,4	6,4	3,2	5,3	2,3
5,4	3,9	1,3	0,4	5,6	3	4,5	1,5	6,5	3	5,5	1,8
5,1	3,5	1,4	0,3	5,8	2,7	4,1	1	7,7	3,8	6,7	2,2
5,7	3,8	1,7	0,3	6,2	2,2	4,5	1,5	7,7	2,6	6,9	2,3
5,1	3,8	1,5	0,3	5,6	2,5	3,9	1,1	6	2,2	5	1,5
5,4	3,4	1,7	0,2	5,9	3,2	4,8	1,8	6,9	3,2	5,7	2,3
5,1	3,7	1,5	0,4	6,1	2,8	4	1,3	5,6	2,8	4,9	2
4,6	3,6	1	0,2	6,3	2,5	4,9	1,5	7,7	2,8	6,7	2
5,1	3,3	1,7	0,5	6,1	2,8	4,7	1,2	6,3	2,7	4,9	1,8
4,8	3,4	1,9	0,2	6,4	2,9	4,3	1,3	6,7	3,3	5,7	2,1
5	3	1,6	0,2	6,6	3	4,4	1,4	7,2	3,2	6	1,8
5	3,4	1,6	0,4	6,8	2,8	4,8	1,4	6,2	2,8	4,8	1,8
5,2	3,5	1,5	0,2	6,7	3	5	1,7	6,1	3	4,9	1,8
5,2	3,4	1,4	0,2	6	2,9	4,5	1,5	6,4	2,8	5,6	2,1
4,7	3,2	1,6	0,2	5,7	2,6	3,5	1	7,2	3	5,8	1,6
4,8	3,1	1,6	0,2	5,5	2,4	3,8	1,1	7,4	2,8	6,1	1,9
5,4	3,4	1,5	0,4	5,5	2,4	3,7	1	7,9	3,8	6,4	2
5,2	4,1	1,5	0,1	5,8	2,7	3,9	1,2	6,4	2,8	5,6	2,2
5,5	4,2	1,4	0,2	6	2,7	5,1	1,6	6,3	2,8	5,1	1,5
4,9	3,1	1,5	0,2	5,4	3	4,5	1,5	6,1	2,6	5,6	1,4
5	3,2	1,2	0,2	6	3,4	4,5	1,6	7,7	3	6,1	2,3
5,5	3,5	1,3	0,2	6,7	3,1	4,7	1,5	6,3	3,4	5,6	2,4
4,9	3,6	1,4	0,1	6,3	2,3	4,4	1,3	6,4	3,1	5,5	1,8
4,4	3	1,3	0,2	5,6	3	4,1	1,3	6	3	4,8	1,8
5,1	3,4	1,5	0,2	5,5	2,5	4	1,3	6,9	3,1	5,4	2,1
5	3,5	1,3	0,3	5,5	2,6	4,4	1,2	6,7	3,1	5,6	2,4
4,5	2,3	1,3	0,3	6,1	3	4,6	1,4	6,9	3,1	5,1	2,3
4,4	3,2	1,3	0,2	5,8	2,6	4	1,2	5,8	2,7	5,1	1,9
5	3,5	1,6	0,6	5	2,3	3,3	1	6,8	3,2	5,9	2,3
5,1	3,8	1,9	0,4	5,6	2,7	4,2	1,3	6,7	3,3	5,7	2,5
4,8	3	1,4	0,3	5,7	3	4,2	1,2	6,7	3	5,2	2,3
5,1	3,8	1,6	0,2	5,7	2,9	4,2	1,3	6,3	2,5	5	1,9
4,6	3,2	1,4	0,2	6,2	2,9	4,3	1,3	6,5	3	5,2	2
5,3	3,7	1,5	0,2	5,1	2,5	3	1,1	6,2	3,4	5,4	2,3
5	3,3	1,4	0,2	5,7	2,8	4,1	1,3	5,9	3	5,1	1,8

- Welcher Fehler wird durch die Rundung der Daten auf 0,1 Zentimeter verursacht?
- Wie stark ist der Zusammenhang zwischen Länge und Breite einer Blüte?
- Zwischen welchen Messwerten besteht der am stärksten ausgeprägte Zusammenhang?
- Keine der Pflanzen in diesem Datensatz hat eine Kelchbreite von 1,8 Zentimetern. Welche Kelchlänge würden wir bei einer solchen Pflanze erwarten?
- Zu welcher Pflanzenart gehört vermutlich eine Iris mit einer Kelchbreite von 1,8 Zentimetern?
- Sind in den drei betrachteten Arten auch Unterarten enthalten, die aus den Daten ersichtlich werden?

In diesem Buch werden wir zahlreiche Methoden kennenlernen, solche und ähnliche Fragestellungen zu beantworten. Zum Verständnis dieser Methoden ist es notwendig, zunächst einige grundlegende Eigenschaften von Daten und deren Relationen zueinander formal zu definieren.

2.2 Maßskalen

Daten können auf unterschiedlichen Skalen messbar sein (67). Tabelle 2.2 zeigt die wichtigsten Maßskalen, die dazugehörigen definierten Operationen, Beispiele sowie die zugehörigen statistischen Maße (46). Die erste Zeile in Tabelle 2.2 zeigt nominal skalierte Daten. Diese sind Daten, für die lediglich die Gleichheit oder Ungleichheit definiert ist. Beispiele sind „reine“ Objektdaten wie Patientennamen oder Produktbezeichnungen. Nominal skalierte Daten können statistisch mit dem Modus beschrieben werden, der dasjenige Objekt bezeichnet, das im Datensatz am häufigsten vorkommt. Die zweite Zeile in Tabelle 2.2 zeigt ordinal skalierte Daten. Für diese ist neben Gleichheit und Ungleichheit auch eine Ordnungsrelation $>$ bzw. $<$ definiert, Beispiel: Schulnoten. Ordinal skalierte Daten lassen sich statistisch mit dem Median beschreiben, der dasjenige Objekt bezeichnet, für das im Datensatz (ungefähr) ebenso viele kleinere wie größere Objekte existieren. Die Berechnung eines Durchschnitts von Schulnoten ist also

Tabelle 2.2: Übersicht über verschiedene Maßskalen

Skala	Operation		Beispiel	stat. Maß
nominal	$=$	$\neq$	Müller, Meier, Schulz	Modus
ordinal	$>$	$<$	sehr gut, gut, befriedigend	Median
Intervall	$+$	$-$	$20°C$, 1999 n. Chr.	Mittelwert
proportional	$\cdot$	$/$	$273°K$, 45 min	Mittelwert