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