Vorwort

Vier Jahre sind seit dem Erscheinen der 2. Auflage unseres Buches vergangen. Vier Jahre, in denen sich das Gebiet der Analyse von Massendaten rapide entwickelt hat. Schlagworte wie *Big Data*, *Predicitve Analytics*, *Data Science* oder *Machine Learning* hören oder lesen wir fast täglich.

Daten werden in vielen Bereichen gesammelt: In der Forschung, um neue Zusammenhänge zu entdecken oder vorhandene Verfahren zu verbessern, Unternehmen sammeln Daten, um durch individualisiertes Marketing höhere Verkaufszahlen zu erreichen oder um eine Verringerung von Ausfallzeiten von Maschinen oder Anlagen zu erzielen.

Mit Hilfe massenweise gesammelter Daten wurden und werden die Verfahren zur Text- oder Schrifterkennung, zum Sprachverstehen oder auch zur Bilderkennung verbessert, so dass immer leistungsfähigere intelligente Systeme entstehen.

Daten werden aus verschiedenen Gründen gesammelt: Erstens ist das Sammeln und Speichern schlicht technisch und finanziell möglich. Zweitens gibt es rechtliche Vorgaben, dass Daten gespeichert werden müssen, zumindest für eine bestimmte Zeit. Drittens wird selbstverständlich – und hier nähern wir uns nun dem Data Mining – auch ein Zweck mit der Datensammlung verfolgt, somit ein Nutzen erwartet.

Welche Daten werden gesammelt? Alle, soweit das Sammeln nicht rechtlich eingeschränkt ist: technische Daten, betriebswirtschaftliche Daten und auch private Daten.

In diesem Buch behandeln wir Techniken, mit deren Hilfe solche Datenmengen ausgewertet werden können. Wir stellen dabei die Prozesse und Techniken für die Analyse strukturierter Daten in den Mittelpunkt, die gemeinhin dem Data Mining zugeordnet werden. Statistische Verfahren zur Datenauswertung werden nur am Rande erwähnt. Die Gebiete der Textanalyse, des sogenannten Text Minings, und auch des Web Minings werden nur skizziert.

Das vorliegende Buch gibt eine Einführung in das Gebiet des Data Minings für strukturierte Daten:

- Zunächst geben wir einen Überblick über Data Mining und diskutieren Grundbegriffe und Vorgehensweisen.
- Anschließend werden Anwendungsklassen beschrieben, um die Einsatzmöglichkeiten des Data Minings erkennen zu können.
- Aus den Daten werden Modelle abgeleitet: Wir stellen dar, wie diese Modelle repräsentiert beziehungsweise gespeichert werden können.
- Die Verfahren zur Analyse der Daten – das Data Mining im engeren Sinne – nehmen natürlich den größten Raum in diesem Buch ein.

https://doi.org/10.1515/9783110676273-202

- Die Daten müssen vorbereitet werden, um eine Analyse zu ermöglichen beziehungsweise die Qualität der Daten zu verbessern.
  Die Qualität der Daten kann die Resultate stark beeinflussen. Daher ist die Datenvorverarbeitung für den Erfolg einer Datenanalyse wichtig. Der Datenvorbereitung ist deshalb ein separates Kapitel gewidmet.
- Es schließt sich die Bewertung der Analyse-Ergebnisse an: Sind sie neu, sind sie wirklich relevant?
  Dieses Kapitel betrachtet auch mögliche Visualisierungen der Ergebnisse als eine Form der Bewertung. Eine geeignete graphische Darstellung hilft sowohl die Qualität der Ergebnisse einzuschätzen als auch Vertrauen in ein Modell aufzubauen.
- Zum Abschluss spielen wir Data Mining an einem konkreten Beispiel durch. Wir illustrieren somit die Vorgehensweise im Data Mining und setzen die im Buch vorgestellten Analysetechniken ein.

Anders als in vielen anderen Büchern trennen wir zwischen den Modellen, zum Beispiel Entscheidungsbaum oder künstliche neuronale Netze, und den Data-Mining-Verfahren, die diese verwenden, zum Beispiel die Generierung eines Entscheidungsbaumes für eine Klassifikation. Damit wird deutlicher, dass einige Modelle für verschiedene Aufgaben eingesetzt werden können.

Zunächst werden im Kapitel 3 Anwendungsklassen vorgestellt. Welche typischen Anwendungsgebiete gibt es für das Data Mining? Anschließend gehen wir im Kapitel 4 auf die Data-Mining-Modelle und die Möglichkeiten ihrer Darstellung ein: Wie kann ein Klassifikationsmodell dargestellt werden? Wie kann man eine Cluster-Aufteilung repräsentieren? Die sich anschließenden Kapitel behandeln die Verfahren, zugeordnet zu den jeweiligen Anwendungsklassen.

Kapitel 8 thematisiert die Datenvorbereitung. Wenngleich diese Phase in einem Data-Mining-Prozess die erste und oft auch eine für den Erfolg ausschlaggebende Etappe ist, haben wir diese weiter hinten platziert. Kennt man die Data-Mining-Algorithmen schon, ist es einfacher zu verstehen, wieso und wie bestimmte Daten vorverarbeitet werden müssen.

Im Kapitel 9 betrachten wir einige Techniken zur Bewertung der Resultate, die durch Data Mining erzielt wurden.

Dieses Buch ist ein Lehrbuch. Data Mining ist mittlerweile in fast allen Curricula von Studiengängen mit einem Informatikbezug enthalten. Anliegen dieses Buchs ist es, eine Einführung in das interessante Gebiet des Data Minings zu geben. Wir haben bewusst bei einigen Verfahren auf die Darstellung der zugrunde liegenden mathematischen Details verzichtet. Ebenso haben wir uns auf grundlegende Algorithmen konzentriert.

Data Mining ist ein Gebiet, welches Erfahrung verlangt. Ein Projekt, in dem blind Data-Mining-Werkzeug eingesetzt werden, wird selten erfolgreich sein. Das Verständnis

für den jeweiligen Gegenstandsbereich, die vorliegenden Daten, aber eben auch für die Data-Mining-Verfahren ist eine notwendige Voraussetzung für ein erfolgreiches Datenanalyse-Projekt.

Mit diesem Buch erhalten Sie einen Einstieg in das Gebiet des Data Minings, so dass Sie Datenanalyse-Projekte strukturiert und zielgerichtet durchführen können. Sie können die Data-Mining-Verfahren hinsichtlich ihrer Anwendungsgebiete und ihrer Leistungsfähigkeit einschätzen und sind in der Lage, die Daten für diese Verfahren entsprechend aufzubereiten.

Die Beispiele haben wir zu großen Teilen in KNIME [KNIME] und WEKA [WEKA] implementiert. Das Buch enthält viele Screenshots von KNIME-Workflows. Wir empfehlen, die Beispiele mittels der Werkzeuge nachzuvollziehen. Installieren Sie diese Systeme, probieren Sie kleine Beispiele aus, und machen Sie sich so mit der Handhabung vertraut.

Wer an Informationen zum Thema Data Mining interessiert ist, findet viele gute Seiten im WWW. Es gibt mittlerweile eine Reihe von Wettbewerben, bei denen reale Probleme zu lösen sind. Die Autoren haben mit ihren Studenten mehrfach am Data Mining Cup [DMC] teilgenommen, der seit einer Reihe von Jahren unter der Leitung der Chemnitzer Firma PRUDSYS durchgeführt wird. Ferner gibt es viele Plattformen, auf denen echte Probleme zu lösen sind, beispielsweise die Plattform www.kaggle.com. Weitere Plattformen haben wir auf den WWW-Seiten zum Buch aufgeführt.

Nicht nur Informationen zum Buch, sondern auch Beispiele finden Sie unter: www.wi.hs-wismar.de/dm-buch

An dieser Stelle möchten wir uns bei der Firma PRUDSYS (www.prudsys.de) sowie den Entwicklern von KNIME und WEKA bedanken, die uns die Verwendung von Beispielen aus ihrem Umfeld gestattet haben. Bitte sehen Sie uns nach, dass die Screenshots der Systeme nicht alle aus den aktuellen Versionen stammen.

Auch in der 3. Auflage haben wir die Grundstruktur beibehalten. Wir haben etliche Änderungen, Aktualisierungen und Erweiterungen vorgenommen. An dieser Stelle möchten wir uns ausdrücklich für die Anregungen unserer Leser bedanken.

Die Zusammenarbeit mit dem De Gruyter-Verlag war angenehm und unkompliziert. Wir bedanken uns insbesondere bei Frau Schedensack (Verlag De Gruyter) und Frau Hausmann (Fa. Konvertus).

Wismar, August 2020 Jürgen Cleve und Uwe Lämmel