CARL HANSER VERLAG

Ian H. Witten, Eibe Frank

Data Mining

Praktische Werkzeuge und Techniken für das maschinelle Lernen

3-446-21533-6

www.hanser.de

Vorwort

Die Konvergenz von Informatik und Telekommunikation hat eine Gesellschaft geschaffen, die von Informationen lebt. Allerdings sind die meisten Informationen nur in Rohform vorhanden: als Daten. Definiert man *Daten* als aufgezeichnete Fakten, so sind *Informationen* die Menge von Mustern oder Erwartungen, die hinter diesen Daten stecken. Datenbanken bergen riesige Informationsmengen – potenziell wichtige Informationen, die aber noch nicht entdeckt oder ausformuliert wurden. Unsere Mission ist es, sie zu Tage zu fördern.

Data Mining ist die Gewinnung impliziter, bislang unbekannter und potenziell nützlicher Informationen aus Daten. Zu diesem Zweck gilt es, Computerprogramme zu entwickeln, die Datenbanken automatisch durchforsten und dabei nach Regelmäßigkeiten oder Mustern suchen. Starke Muster werden sich, wenn sie erst einmal gefunden sind, mit großer Wahrscheinlichkeit verallgemeinern und für genaue Vorhersagen über künftige Daten nutzen lassen. Natürlich werden dabei Probleme auftreten. Viele Muster werden sich als banal oder uninteressant erweisen. Andere werden Störfeuer sein, die auf zufällige Koinzidenzen in der aktuell verwendeten Datenmenge zurückzuführen sind. Und echte Daten sind nicht perfekt: Manche Teile sind verfälscht, andere fehlen ganz. Was immer auch entdeckt wird, es wird ungenau sein: Es wird Ausnahmen zu jeder Regel geben und Fälle, die von keiner einzigen Regel abgedeckt werden. Algorithmen müssen deshalb robust genug sein, um mit unperfekten Daten fertig zu werden und Regelmäßigkeiten zu extrahieren, die ungenau, aber nützlich sind.

Die technische Basis des Data Mining ist das maschinelle Lernen. Es wird genutzt, um Informationen aus rohen Daten in Datenbanken zu gewinnen – Informationen, die in einer verständlichen Form ausgedrückt sind und für vielfältige Zwecke genutzt werden können. Wir haben es also mit einem Abstraktionsprozess zu tun: Aus den Daten, einschließlich aller darin enthaltenen Schönheitsfehler, wird die zugrunde liegende Struktur abgeleitet. In diesem Buch geht es um die Werkzeuge und Techniken des maschinellen Lernens, die beim Data Mining eingesetzt werden, um strukturelle Muster in Daten zu finden und zu beschreiben.

Wie jede hoffnungsvolle neue Technologie, die ein starkes kommerzielles Interesse hervorruft, wird Data Mining in der Fachpresse und gelegentlich auch in der Tagespresse mit überhöhten Erwartungen befrachtet. Überzogenen Berichten zufolge sollen sich alle möglichen Geheimnisse aufklären lassen, indem man auf die Datenflut Lernalgorithmen ansetzt. Dabei hat maschinelles Lernen nicht das Ge-

ringste mit Hexerei, Alchimie oder verborgenen Mächten zu tun. Es ist einfach ein definierter Korpus eingängiger und praktikabler Techniken, mit denen sich nützliche Informationen aus rohen Daten ziehen lassen. Dieses Buch beschreibt diese Techniken und zeigt, wie sie funktionieren.

Wir interpretieren maschinelles Lernen als den Erwerb struktureller Beschreibungen aus Beispielen. Die gefundenen Beschreibungsarten können zur Vorhersage, Erklärung und als Verständnishilfe eingesetzt werden. Manche Data-Mining-Anwendungen konzentrieren sich auf die Vorhersage: die Voraussage, was in neuen Situationen passieren wird, anhand von Daten, die beschreiben, was in der Vergangenheit geschah, wobei häufig die Klassenzuordnung neuer Beispiele erraten wird. Genauso stark, wenn nicht sogar stärker, sind wir jedoch an Anwendungen interessiert, deren „Lern"-Ergebnis die Beschreibung einer Struktur ist, die zur Klassifizierung von Beispielen eingesetzt werden kann. Diese strukturelle Beschreibung unterstützt nicht nur die Vorhersage, sondern auch das Erklären und Verstehen. Unserer Erfahrung nach sind bei den meisten Data-Mining-Anwendungen die Erkenntnisse, die der Anwender gewinnt, von größtem Interesse; und tatsächlich gehören solche Erkenntnisse zu den Hauptvorteilen des maschinellen Lernens gegenüber der klassischen statistischen Modellierung.

Das Buch erklärt maschinelle Lernmethoden unterschiedlichster Art. Einige davon werden aus rein didaktischen Motiven beschrieben: einfache Verfahren, die dafür entworfen sind, die Funktionsweise grundlegender Ideen anschaulich zu erklären. Andere sind durch und durch praxisnah: reale Systeme, die in heutigen Anwendungen eingesetzt werden. Viele der beschriebenen Lernmethoden sind hochaktuell und wurden erst in den letzten Jahren entwickelt.

Zur Illustration der in diesem Buch vorgestellten Konzepte wurde eine umfassende, in Java geschriebene Software-Utility geschaffen. Sie heißt Waikato Environment for Knowledge Analysis oder kurz Weka[1] und ist als Quellcode im World Wide Web unter www.mkp.com/datamining oder www.cs.waikato.ac.nz/ml/weka erhältlich. Weka ist eine vollständige Implementierung nahezu aller in diesem Buch beschriebenen Techniken auf Industrieniveau. Neben Code zu Anschauungszwecken umfasst es voll funktionsfähige Implementierungen maschineller Lernmethoden. Es bietet klare, sparsame Implementierungen der einfachsten Techniken, die darauf abzielen, das Verständnis der beteiligten Mechanismen zu erleichtern. Es stellt darüber hinaus eine Workbench mit funktionierenden Implementierungen auf dem neuesten Stand der Technik für viele beliebte Lernverfahren bereit, die für praktische Data-Mining-Anwendungen oder Forschungszwecke genutzt werden können. Schließlich enthält es ein Framework in der Form einer Java-Klassenbibliothek, das Anwendungen, die eingebettetes maschinelles Lernen nutzen, ebenso unterstützt wie die Implementierung neuer Lernverfahren.

[1] Der *weka* ist ein flugunfähiger, besonders neugieriger Vogel, den es nur auf den neuseeländischen Inseln gibt.

Ziel dieses Buchs ist es, die Werkzeuge und Techniken für maschinelles Lernen vorzustellen, die beim Data Mining eingesetzt werden. Nachdem Sie es gelesen haben, werden sie diese Techniken kennen und verstehen und ihre Stärke und Praktikabilität schätzen. Wenn Sie mit eigenen Daten experimentieren möchten, so können Sie das mit Hilfe der Weka-Software tun.

Das Buch schlägt eine Brücke zwischen dem stark praxisbezogenen Ansatz, den wirtschaftsnahe Bücher mit Fallstudien über Datamining verfolgen, und der eher theoretischen, grundlagenorientierten Darstellung, wie sie für aktuelle Fachbücher über maschinelles Lernen typisch ist. Zwischen diesen Ansätzen klafft eine relativ breite Kluft. Um Techniken des maschinellen Lernens produktiv einsetzen zu können, müssen Sie ihre Arbeitsweise verstehen, denn maschinelles Lernen ist keine Technologie, die auch bei blinder Anwendung noch gute Ergebnisse liefert. Unterschiedliche Probleme erfordern unterschiedliche Techniken. Weil Data Mining jedoch eine sehr junge Disziplin ist, ist nie klar, welche Techniken in einer gegebenen Situation die geeigneten sind: Deshalb ist es notwendig, die Bandbreite in Frage kommender Lösungen zu kennen. Dieses Buch behandelt denn auch ein extrem breites Spektrum von Techniken. Das ist möglich, weil wir im Gegensatz zu vielen wirtschaftsnahen Büchern keiner bestimmten kommerziellen Software das Wort reden. Dieses Buch enthält sehr viele Beispiele. Die Datenmengen dienen jedoch Anschauungszwecken und sind so klein, dass Sie die beschriebenen Abläufe detailliert verfolgen können. Reale Datenmengen wären dafür viel zu groß (und werden von den Firmen unweigerlich vertraulich behandelt). Unsere Datenmengen sind nicht dafür ausgewählt, tatsächliche umfassende Probleme aus der Praxis zu illustrieren. Sie sollen Ihnen vielmehr helfen zu verstehen, was die verschiedenen Techniken leisten, wie sie funktionieren und wo ihre Anwendungsmöglichkeiten liegen.

Dieses Buch zielt auf den technisch interessierten, allgemeinen Leser ab, der sich für die Prinzipien und Techniken interessiert, auf die praktische Data-Mining-Anwendungen zurzeit aufsetzen. Es ist darüber hinaus für IT-Spezialisten von Interesse, die sich in diese neue Technologie des Data Mining einarbeiten möchten, und für alle, die sich ein detailliertes technisches Hintergrundwissen über maschinelles Lernen aneignen möchten. Es ist für ein gemischtes Publikum aus Anwendern von Informationssystemen, Programmierern, Beratern, Entwicklern, IT-Managern, Spezifikationsspezialisten, Patentanwälten, interessierten Laien sowie Studenten und Professoren geschrieben, die ein leicht zu lesendes Buch mit vielen Abbildungen brauchen, das die wichtigsten Techniken des maschinellen Lernens, ihre Leistungsfähigkeit, Einsatzmöglichkeiten und Arbeitsweise beschreibt. Es ist praxisorientiert – fast schon ein „How-to"-Buch – und enthält Algorithmen, Code und Implementierungen. Alle, die Data Mining in der Praxis einsetzen, werden von den beschriebenen Techniken direkt profitieren. Das Buch wendet sich auch an Leser, die wissen möchten, was hinter der Begeisterung über maschinelles Lernen steckt und nach einem praktischen, nicht therorielastigen,

unprätentiösen Ansatz suchen. Wir haben darauf verzichtet, theoretische oder mathematische Spezialkenntnisse vorauszusetzen. Einzige Ausnahme sind einige wenige Abschnitte, die mit einem hellgrauen Balken am Rand markiert sind. Diese Passagen enthalten Material für stärker technisch oder theoretisch ausgerichtete Leser und können ohne weiteres übersprungen werden.

Das Buch ist in Schichten organisiert, die die Konzepte des Data Mining sowohl Lesern erschließen, die sich ein Grundlagenwissen aneignen möchten, als auch Lesern, die eine vertiefte Behandlung und Detailinformationen über die behandelten Techniken wünschen. Wir sind der Meinung, dass Anwender des maschinellen Lernens eine Vorstellung davon haben müssen, wie die eingesetzten Algorithmen funktionieren. Es zeigt sich immer wieder, dass Datenmodelle nur so gut sind wie die Person, die sie interpretiert, und dass diese Person etwas über die Erzeugung der Modelle wissen muss, um Stärken und Grenzen der Technologie richtig einschätzen zu können. Nicht alle Anwender benötigen jedoch ein vertieftes Detailwissen über die einzelnen Algorithmen.

Wir berücksichtigen diese Situation, indem wir die Methoden des maschinellen Lernens in aufeinander folgenden Detailstufen beschreiben. Der Leser lernt die grundlegenden Konzepte, die oberste Ebene, kennen, wenn er die drei ersten Kapitel liest. Kapitel 1 beschreibt anhand von Beispielen, was maschinelles Lernen ist und in welchen Bereichen es eingesetzt werden kann und stellt darüber hinaus reale praktische Anwendungen vor. Kapitel 2 und 3 behandeln die verschiedenen Ein- und Ausgabeformen – oder *Wissensrepräsentationen* –, die beim Data Mining eine Rolle spielen. Unterschiedliche Ausgabeformen geben unterschiedliche Arten von Algorithmen vor. Deshalb beschreibt Kapitel 4, die nächste Stufe, die grundlegenden Methoden des maschinellen Lernens in vereinfachter, leicht verständlicher Form. Die beteiligten Prinzipien werden hier anhand von unterschiedlichen Algorithmen vermittelt, ohne auf spitzfindige Details oder schwierige Implementierungsfragen einzugehen. Um Fortschritte in der Anwendung maschineller Lerntechniken auf konkrete Data-Mining-Probleme zu erzielen, ist es notwendig, das Erreichte messen zu können. Kapitel 5, das unabhängig von den anderen Kapiteln gelesen werden kann, befähigt den Leser, die durch maschinelles Lernen gewonnenen Ergebnisse zu evaluieren, und behandelt die manchmal komplexen Probleme der Leistungsevaluierung.

Auf der niedrigsten und detailliertesten Stufe legt Kapitel 6 schonungslos verzwickte Probleme bei der Implementierung verschiedener Algorithmen des maschinellen Lernens offen und zeigt die vielschichtigen Überlegungen, die notwendig sind, wenn die Algorithmen in der Praxis gut funktionieren sollen. Auch wenn viele Leser diesen Detailinformationen wenig werden abgewinnen können, repräsentieren sie doch die Ebene, auf der die vollständigen, funktionierenden, ausgetesteten Java-Implementierungen der maschinellen Lernsysteme geschrieben sind. Kapitel 7 diskutiert praktische Themen der Aufbereitung der Eingaben für das maschinelle Lernen – zum Beispiel die Auswahl und Diskretisierung von

Attributen – und behandelt einige anspruchsvollere Techniken zur Verfeinerung und Kombination der Ausgabe verschiedener Lerntechniken. Kapitel 8 beschreibt den Java-Code, der dieses Buch ergänzt. Sie können sich diesem Kapitel direkt von Kapitel 4 aus zuwenden, wenn Sie schnell mit der Analyse Ihrer Daten beginnen möchten, und sich nicht mit den technischen Details aufhalten wollen. Kapitel 9 schließlich gibt einen Ausblick auf die Zukunft.

Das Buch behandelt nicht alle Methoden des maschinellen Lernens. Insbesondere gehen wir nicht auf neuronale Netze ein, weil diese Technik eher Vorhersagen als strukturelle Beschreibungen erzeugt, und darüber hinaus gut in eben erst erschienenen Büchern über Data Mining beschrieben ist. Auch das so genannte Reinforcement Learning, das in praktischen Data-Mining-Anwendungen nur selten eine Rolle spielt, ist in diesem Buch kein Thema. Das Gleiche gilt für genetische Algorithmusansätze, die eigentlich nur eine Optimierungstechnik sind; für Bayessche Netzwerke, weil die Algorithmen, sie zu erlernen, für den praktischen Einsatz nicht robust genug sind; sowie relationales Lernen und induktive logische Programmierung, die in gängigen Data-Mining-Anwendungen selten genutzt werden.

Die Techniken des maschinellen Lernens, die dieses Buch ergänzen, wurden mit Java implementiert, weil Java als objektorientierte Programmiersprache ein einheitliches Interface zu den Lernverfahren und -methoden für die Vor- und Nachverarbeitung erlaubt. Wir haben Java den Vorzug gegenüber C++, Smalltalk oder anderen objektorientierten Sprachen gegeben, weil in Java geschriebene Programme auf fast jedem Computer ausgeführt werden können, ohne neu kompilieren, komplizierte Installationsprozeduren ausführen oder – noch schlimmer – den Code selbst ändern zu müssen. Ein Java-Programm wird in Byte-Code übersetzt, der auf jedem Computer ausgeführt werden kann, der mit einem geeigneten Interpreter ausgestattet ist. Dieser Interpreter heißt *Java-Virtual-Machine*. Java-Virtual-Machines – und natürlich auch Java-Compiler – sind für alle wichtigen Plattformen kostenlos erhältlich.

Wie alle verbreiteten Programmiersprachen steht auch Java im Kreuzfeuer der Kritik. Obwohl hier nicht der Platz ist, näher auf diese Frage einzugehen, haben die Kritiker in verschiedenen Punkten durchaus Recht. Von allen derzeit verfügbaren Programmiersprachen, die allgemein anerkannt, standardisiert und umfassend dokumentiert sind, schien die Sprache Java für die Zwecke dieses Buches am besten geeignet zu sein. Ihr Hauptnachteil ist ihre Ausführungsgeschwindigkeit – beziehungsweise ihr Mangel daran. Die Ausführung eines Java-Programms ist um ein Mehrfaches langsamer als die Ausführung eines entsprechenden C-Programms, weil die Virtual Machine den Byte-Code vor der Ausführung in Maschinencode übersetzen muss. Nach unserer Erfahrung liegt der Unterschied bei einem Faktor drei bis fünf, wenn die Virtual Machine einen *Just-in-Time-Compiler* verwendet. Statt jeden Byte-Code einzeln zu übersetzen, übersetzt ein Just-in-Time-Compiler ganze Byte-Code-Fragmente in Maschinencode, sodass eine sig-

nifikante Geschwindigkeitssteigerung erreicht wird. Ist auch diese Möglichkeit für Ihre Anwendung zu langsam, so gibt es Compiler, die Java-Programme direkt in Maschinencode übersetzen und den Byte-Code-Schritt völlig umgehen. Natürlich kann dieser Code nicht auf anderen Plattformen ausgeführt werden, sodass einer der Hauptvorteile von Java verloren geht.

Dank

Den Dank zu schreiben ist immer das Schönste! Viele Menschen haben uns unterstützt, und wir freuen uns sehr über diese Gelegenheit, ihnen zu danken. Dieses Buch ist aus einem Forschungsprojekt über maschinelles Lernen am Fachbereich Informatik der University of Waikato in Neuseeland entstanden. Die wissenschaftlichen Mitarbeiter an diesem Projekt haben uns in unserem Vorhaben unermüdlich bestätigt und großzügig unterstützt: John Cleary, Sally Jo Cunningham, Matt Humphrey, Lyn Hunt, Bob McQueen, Lloyd Smith und insbesondere Geoff Holmes als Projektleiter und Inspirationsquelle. Alle, die am Projekt über maschinelles Lernen mitgearbeitet haben, haben unser Denken beeinflusst: Insbesondere möchten wir Steve Garner, Stuart Inglis und Craig Nevill-Manning erwähnen, die uns geholfen haben, das Projekt am Anfang, als der Erfolg noch ungewiss und die Dinge schwierig waren, auf den Weg zu bringen.

Das Weka-System, das die beschriebenen Konzepte illustriert, ist eine entscheidende Komponente dieses Buches. Es wurde von den Autoren konzipiert und von Eibe Frank zusammen mit Len Trigg und Mark Hall entworfen und implementiert. Viele Menschen im Labor für maschinelles Lernen in Waikato haben einen erheblichen Beitrag geleistet, ganz besonders Yong Wang mit seiner Implementierung von M5‘.

Abgeschottet von der Welt, wie wir in einer fernen (aber sehr schönen) Ecke der südlichen Hemisphäre leben, freuen wir uns über alle Besucher unseres Fachbereichs: Sie spielen eine entscheidende Rolle als Resonanzboden und helfen uns bei der Erweiterung unseres Horizonts. Wir möchten in diesem Zusammenhang insbesondere Rob Holte, Bernhard Pfahringer, Carl Gutwin und Russell Beale erwähnen, die alle mehrere Monate bei uns zu Gast waren; David Aha, der zwar nur für ein paar Tage, dafür aber in einer frühen und fragilen Phase des Projekts zu uns kam und uns mit seiner Begeisterung und Ermutigung anfeuerte; und Kai Ming Ting, der zwei Jahre lang mit uns an vielen der in Kapitel 7 dikutierten Themen arbeitete und uns half, im Mainstream des maschinellen Lernens mitzuschwimmen.

Frühere Studenten in Waikato haben bei der Entwicklung des Projekts eine bedeutende Rolle gespielt. Jamie Littin arbeitete an der Erzeugung von Regeln mit Ausnahmen und relationalem Lernen. Brent Martin befasste sich mit instanzbasiertem Lernen und verschachtelten instanzbasierten Repräsentationen. Murray Fife brütete über relationalem Lernen. Nadeeka Madapathage erforschte, wie sich

Algorithmen des maschinellen Lernens mit funktionalen Sprachen ausdrücken lassen. Andere Studenten haben uns in vielfältiger Weise beeinflusst, insbesondere Gordon Paynter, der sich intensiv mit dem in Abschnitt 9.4 beschriebenen Verfahren zur Extrahierung von Schlüsselbegriffen auseinander setzte, und Zane Bray, der an dem im gleichen Abschnitt vorgestellten Text-Mining-Verfahren arbeitete. Steve Jones, Tony Smith und Malika Mahoui haben ebenfalls großartige und weitreichende Beiträge zu diesen und anderen Projekten des maschinellen Lernens geleistet.

Ian Witten möchte sich für den Beitrag seiner früheren Studenten in Calgary bedanken, besonders Brent Krawchuk, Dave Maulsby, Thong Phan und Tanja Mitrovic, die ihm ebenso wie die Fakultätsmitglieder Bruce MacDonald, Brian Gaines und David Hill in Calgary und John Andreae an der University of Canterbury halfen, seine frühen Ideen über maschinelles Lernen weiter zu entwickeln. Eibe Franks besonderer Dank gilt seinem früheren Betreuer an der Universität Karlsruhe, Klaus-Peter Huber (heute am SAS-Institut), der ihn mit der faszinierenden Idee lernender Maschinen infizierte. Bronwyn Webster hat uns in Waikato ausgezeichnet unterstützt.

Danken möchten wir auch für die Bemühungen der anonymen Gutachter, von denen uns vor allem einer durch sehr viele relevante und konstruktive Kommentare half, dieses Buch signifikant zu verbessern. Darüber hinaus möchten wir den Bibliothekaren des Repository of Machine Learning Databases an der University of California in Irvine danken, deren sorgsam zusammengetragene Datenmengen für unsere Forschungen von unersetzlichem Wert waren.

Unsere Forschungsarbeiten wurden von der New Zealand Foundation for Research, Science and Technology und der Royal Society of New Zealand Marsden Fund finanziert. Der Fachbereich Informatik an der University of Waikato hat uns in vielerlei Hinsicht großzügig unterstützt, und unser besonderer Dank gilt Mark Apperley für seine inspirierende Führung und herzliche Ermutigung. Ein Teil dieses Buches entstand, während beide Autoren zu Gast an der University of Calgary in Kanada waren, und wir danken dem dortigen Informatik-Fachbereich für seine Unterstützung – ebenso wie den geduldigen Studenten, die uns in unserer Vorlesung über maschinelles Lernen als Versuchskaninchen ausgeliefert waren, für ihre positive und hilfreiche Einstellung.

Vor allem – und am allermeisten – sind wir unseren Familien und Partnerinnen dankbar. Obwohl Pam, Anna und Nikki genau wussten, was es heißt, einen Autor im Haus zu haben („nicht noch einmal!“), ließen sie Ian das Buch schreiben. Julie war immer verständnisvoll, selbst wenn Eibe bis spät in die Nacht im Labor für maschinelles Lernen saß. Wir sechs kommen aus Kanada, England, Deutschland, Irland und Samoa: Neuseeland führte uns zusammen und war ein idealer, ja sogar idyllischer Schauplatz, um dieses Buch zu schreiben.