

Buchbesprechung, erschienen in: Information – Wissenschaft und Praxis 2001, vol. 52 (7). S. 427f.

Witten, Ian; Frank, Eibe (2001): Data Mining: Praktische Werkzeuge und Techniken für das maschinelle Lernen. Hanser Verlag München Wien. 386 Seiten.
Webseite der WEKA Software: <http://www.cs.waikato.ac.nz/ml/weka>

Maschinelles Lernen gewinnt in Zeiten stetig wachsender Datenmengen zunehmend an Bedeutung. Seit einigen Jahren bündelt sich die Forschung und Methoden, die weitgehend aus der Künstlichen Intelligenz kommen, unter dem Schlagwort Data Mining. Zunehmend werden Data Mining Verfahren in der Bibliometrie oder dem Information Retrieval eingesetzt. Deshalb sollte Data Mining auch in der informations- und bibliothekswissenschaftlichen Ausbildung eine Rolle spielen und sei es nur, um Einsatzmöglichkeiten, Risiken und Ergebnisse vernünftig bewerten zu können.

Eine gute Basis für einen Kurs bietet das Lehrbuch von Witten & Frank, das aus dem Englischen ins Deutsche übersetzt wurde. Das Software Paket WEKA (Waikato Environment for Knowledge Analysis), das die Autoren gratis im Internet zur Verfügung stellen, bietet eine wichtige Ergänzung des Buches und eines entsprechenden Kurses. Es bietet eine Implementierung der besprochenen Algorithmen in JAVA und damit eine Übungsmöglichkeit auch ohne JAVA Kenntnisse.

Der Aufbau des Buches orientiert sich nicht an den Algorithmen sondern verwendet die verschiedenen Aufgaben im Verlauf des Data Mining Prozesses als Grundlage für die Gliederung. Dabei kommen die Autoren auf einzelne Verfahren mehrfach unter verschiedenen Gesichtspunkten zurück.

Der erste Kapitel führt motivierend und anwendungsorientiert in die Thematik ein. Das zweite Kapitel behandelt die Eingabe der zu verarbeitenden Konzepte und greift dabei bereits das Datenformat für das WEKA Software-Paket auf, das ausführlich in Kapitel 8 erläutert wird. Kapitel 3 stellt das Format der Ausgabe von Data Mining Verfahren vor und nennt Entscheidungsbäume, verschiedene Regelmengen und Cluster. Die algorithmische Erzeugung dieser Wissensstrukturen aus den Rohdaten bildet den Kern des Data Mining. In Kapitel 4 beginnt die Vorstellung der Verfahren des maschinellen Lernens. Neben Vorarbeiten liegt der Schwerpunkt auf der Erzeugung von Regeln und linearen und statistischen Modellen. Die Evaluierung von Ergebnissen durch das Bilden von Testmengen und die numerische Abschätzung von Fehlerraten behandelt Kapitel 5. Auch Kostenfunktionen und die Bewertung von Clustern kommen zur Sprache. Das längste Kapitel 6 bespricht die Verfahren des maschinellen Lernens aus Kapitel 4 weiter und führt komplexere Algorithmen ein. Der Schwerpunkt liegt wieder auf Entscheidungsbäumen, Regeln zur Klassifikation, numerischer Vorhersage und Clustering. Zur Sprache kommen aktuelle, in der Forschung eingesetzte Algorithmen wie z.B. die mächtigen Support Vector Maschinen, C4.5 Entscheidungsbäume und Clustering durch Erwartungsmaximierung.

Insgesamt liegt der Schwerpunkt des Buches auf symbolischen Verfahren. Gerade das Fehlen von neuronalen Netzen ist bedauerlich, da sie ein weiteres wertvolles Werkzeug darstellen. Als Grund nennen die Autoren, dass die Ergebnisse neuronaler Netze nicht interpretierbar sind. Dies ist ein alter Streitpunkt zwischen symbolischen und sogenannten sub-symbolischen Lernverfahren, trotzdem finden der Backpropagation Algorithmus und zunehmend Kohonens Selbstorganisierende Karten (SOM) häufig Anwendung. Und auch ein Regelsystem oder ein Entscheidungsbaum auf Basis eines großen Datenbestandes sind aufgrund ihres Umfangs in der Praxis schwer im Detail nachvollziehbar.

Das siebente Kapitel behandelt weiterführende Themen der Ein- und Ausgabe. Bei der Eingabe bedeutet dies Komplexitätsreduktion durch die vorteilhafte Auswahl von Attributmengen, Diskretisierung von numerischen Attributen und Datensäuberung in Entscheidungsbäumen. Bei der Ausgabe besprechen die Autoren die Kombination mehrerer Data Mining Algorithmen zu Committee Machines, auch ein in der Forschung intensiv diskutiertes Thema. Kapitel 8 behandelt das WEKA Software-Paket das u.a. viele der besprochenen Algorithmen als JAVA Code enthält. Dieses Kapitel liegt der Software als PDF

Datei bei, so dass die Software auch unabhängig vom Buch genutzt werden kann. Für den Einsatz von WEKA bestehen drei Möglichkeiten:

- ?? Benutzung des Systems als Anwender mittels Benutzungsoberfläche
- ?? Einbinden von WEKA Routinen in eigene JAVA Programme
- ?? Realisierung eigener Lernverfahren durch das Ausnutzen der offenen Strukturen von WEKA

Das WEKA Paket ist eine ausgezeichnete Ressource für Lehre und Forschung, die den Vergleich von mehreren Algorithmen erleichtert. Sie unterstützt den gesamten Data Mining Prozess von der Vorbereitung bis zur Evaluierung, wobei die Stärke auf der Vielfalt der realisierten Verfahren liegt. Das Medium Software ist natürlich dynamischer als das statische Medium Buch. WEKA wird regelmäßig gepflegt und ergänzt und hat inzwischen den im Buch geschilderten Umfang überschritten. Die WEKA Version 3.1 umfasste z.B. auch neuronale Netze, wesentlich mehr Kombinationsverfahren als im Buch vorkommen und v.a. zwei komfortablere Benutzungsschnittstellen:

- ?? Ein verbesserter Kommandozeilen-Interpreter mit Eingabe-Wiederholung und scrollbarer Ausgabe
- ?? Eine graphische, fensterbasierte Benutzungsoberfläche

Im großen und ganzen läuft WEKA sehr stabil und ist ein wertvolles Werkzeug. Im Detail finden sich bei komplexer Software natürlich Probleme. Die relativ neue Implementierung von neuronalen Netzen ist instabil und die entstehenden Modelle sind nicht sehr flexibel. Allerdings liegt bereits eine neuere Version vor. Die Klasse für Support Vector Maschinen heisst nicht SVM, was die übliche Abkürzung wäre, sondern SMO nach einem der Basisalgorithmen. Auf die Lehre wirkt sich problematisch aus, dass die Ausgabe der Modelle teilweise zu kryptisch und nicht ausreichend dokumentiert ist. Dabei stören auch die Inkonsistenzen zwischen dem Buch und dem sich entwickelnden Software-Paket.

Kapitel 9 schließt den Rahmen des Buches mit einem Ausblick der u.a. Visualisierung und Text Mining enthält.

Obwohl der zirkuläre Aufbau von den meisten Lehrbüchern zu diesem Thema abweicht und sich am Prozeß des Data Mining orientiert, führt er nicht zu erheblichen Verbesserungen. Die zwei Kapitel zu Algorithmen hätten in mehrere Kapitel zu einzelnen Verfahren aufgeteilt werden können. Dadurch wären auch geeignetere Häppchen für die Sitzung einer Vorlesung entstanden. In einem Kurs sollte der Lehrende konkrete Algorithmen auch vor dem vierten Kapitel einführen oder zumindest ein Verfahren beispielhaft besprechen, sonst können die Studierenden lange nicht mit Übungen beginnen.

Das Buch und die Software sind mit geringen mathematischen Vorkenntnissen lesbar bzw. benutzbar. Im Buch sind mathematisch etwas weiterführende Abschnitte mit grauen Randbalken gekennzeichnet. Manchem motivierten Studierenden geht die Formalisierung und der Detaillierungsgrad aber sicher auch dort nicht weit genug, denn einige Algorithmen werden nicht bis zur letzten Konsequenz formal erläutert. Dafür schließt jedes Kapitel mit einem Ausblick auf weiterführende Literatur.

Einige wenige unvollständige Sätze haben sich in der Übersetzung des ursprünglich englischen Buches von 2000 eingeschlichen. Problematischer ist die Bemühung, den lockeren Stil des Originals ins Deutsche zu übernehmen. Etwas verwirrend ist der hintere Klappentext. Java wirkt dort wie ein spezielles Verfahren, das sozusagen die Grundlage des Teigs für bestimmte innovative Methoden ist. Dagegen ist es nicht weiter als der Topf, in dem sowohl alte als auch neue Algorithmen „gegart“ werden.

Insgesamt bleibt das Buch aber empfehlenswert. Zielgruppe sind wie bereits erwähnt Studierende in nicht formal orientierten Studiengängen aber auch Anwender, die im Selbststudium schnell moderne Verfahren des maschinellen Lernens einsetzen möchten.

(Thomas Mandl, Universität Hildesheim, mandl@rz.uni-hildesheim.de)