Knowledge Discovery in Databases

Knowledge Discovery in Databases: Techniken und Anwendungen. Martin Ester und Jörg Sander. Springer-Verlag, 2000. 281 Seiten, Brosch. ISBN: 3-540-67328-8. Preis 69.- DM. http://www.dbs.informatik.uni-muenchen.de/buecher/kdd.html

Das Aufspüren von bisher unbekanntem Wissen ist wohl in jeder Branche ein verlockendes Ziel — im Finanzwesen genauso wie im Wissensmanagement. Zur Einführung in das Gebiet "Knowledge Discovery in Databases" (KDD) bietet sich dem interessierten Neuling nun das neue Buch von Ester und Sander an. Es entstand aus Seminaren und Vorlesungen an der Ludwig-Maximilians Universität in München und richtet sich an Studenten der Informatik sowie Praktiker mit fundiertem Fachwissen.

Obwohl es nur knappe 300 Seiten zählt lässt der Titel auf eine tiefgehende Betrachtung des KDD schließen. Die Kapitel sind dabei immer nach dem gleichen Schema aufgebaut. Eine kurze Einführung in die entsprechende Technik erläutert die grundlegende Idee zu deren Realisierung einige Algorithmen in Pseudokode vorgestellt werden. Die Beschreibung der Anwendungsgebiete sowie spezieller Eigenschaften wie Zeitkomplexität und Ergebnisqualität runden die Betrachtung der Algorithmen ab. Abbildungen und Beispiele unterstützen dabei meist das Verständnis. Leider wurden nicht alle Algorithmen mit guten Beispielen untermauert. An einigen Stellen ist eigene Denkarbeit gefragt wodurch der Lesefluss ins stocken gerät. Beendet werden die Kapitel immer von einer Zusammenfassung und einem eigenen Literaturverzeichnis.

Inhaltlich beginnt das Buch mit einer *Einleitung* in das Gebiet KDD. Hier werden die verschiedenen Ursprünge — Statistik, Datenbanken und Maschinelles Lernen — sowie die Teilschritte des KDD beschrieben. Dabei erwähnen die Autoren, dass im folgenden Text nur der Teilschritt "Data Mining" betrachtet wird. Eine kurze Betrachtung der Anwendungsgebiete des KDD sowie eine Übersicht der nachfolgenden Kapitel beenden das Kapitel.

Die elementaren *Grundlagen* der Datenbanksysteme und Statistik sind Mittelpunkt des zweiten Kapitels. Neben einer kurzen aber guten Beschreibung relationaler Datenbanksysteme und -sprachen werden einige Indexstrukturen vorgestellt. Eine ebenso gute Auffrischung der statistischen Grundlagen wie Korrelation oder Wahrscheinlichkeitsrechnung beschließt das Kapitel. Für eine tiefergehende Einführung in die beiden Gebiete benötigt ein Anfänger natürlich weitere Literatur.

Das erste Kapitel über die Techniken des DM beschäftigt sich mit dem *Clustering*. Dabei sollen Objekte in mehrdimensionalen Räumen zu Gruppen (Clustern) zusammengefasst werden. Neben einer ausführlichen Einführung in das einfache und hierarchische Clustering wird hier bspw. auch das Problem nicht-kugelförmiger Cluster betrachtet.

Die *Klassifikation* neuer Objekte sowie die Ermittlung von Klassifikationswissen ist Thema des vierten Kapitels. Hier werden drei Verfahren der Klassifikation vorgestellt. Bayes-Klassifikatoren verwenden Wahrscheinlichkeiten zur Bestimmung der Klassenzugehörigkeit. Nächste-Nachbar Klassifikatoren hingegen berechnen den Abstand zum Mittelpunkt einer Klasse und Entscheidungsbaum-Klassifikatoren extrahieren aus den Trainingsdaten Entscheidungskriterien für einen Baum. Insbesondere beim letzten Verfahren wird das Klassifikationswissen explizit abgebildet. Ein weiteres Thema ist die Optimierung der Algorithmen für große Datenmengen.

Assoziationsregeln (AR) stehen im Mittelpunkt des darauffolgenden Kapitels. Prominentestes Beispiel ist wohl die Warenkorbanalyse, bei der Zusammenhänge beim Einkauf verschiedener Produkte ermittelt werden. Neben einfachen und hierarchischen AR ist auch die Betrachtung quantitativer AR Teil des Kapitels. Während hierarchische AR zur Entdeckung von Zusammenhängen zwischen abstrakten Warengruppen dienen, beziehen qualitative AR genauere Wertebereiche der Attribute ein.

Die *Generalisierung* ist die letzte vorgestellte Technik des DM. Es sollen viele einzelne Daten zu wenigen Informationen verarbeitet werden um eine bessere Übersicht zu erhalten. Als Beispiel für eine manuelle Generalisierung dient hierbei der Data Cube aus dem Bereich der Data Warehouses. Zur Beschreibung der automatischen Generalisierung wird die attributorientierte Induktion verwendet. Maßnahmen bei Updates der Basisdaten beenden das Kapitel.

Das siebte Kapitel behandelt besondere *Datentypen und Anwendungen* des DM. Zeitlich abhängige Daten ermöglichen mittels des sog. *Temporal Data Mining* Trends oder Regeln zu entdecken. Im *Spatial Data Mining* finden Daten besondere Betrachtung, welche zueinander abhängige Attribute besitzen — bspw. räumliche Dimensionen. Abschließend wird die Wissensentdeckung in großen unstrukturierten Textmengen — dem *Text- und Web-Mining* untersucht.

Das Buch endet mit einer Betrachtung *weiterer Verfahren* aus dem Gebiet DM. Neben *Induktiver Logik-Programmierung* und *Genetischen Algorithmen* zur Ermittlung von Regeln werden auch *Neuronale Netze* zur Klassifikation betrachtet.

Zusammenfassend lässt sich sagen, dass dieses Lehrbuch für Einsteiger in das Gebiet "Data Mining" sehr gut geeignet ist. Anders als der Titel verspricht, gehen die Autoren aber nicht auf den gesammten Prozess des KDD, sondern auf die Techniken des DM ein. Andere Teilgebiete des KDD wie beispielsweise die Vorverarbeitung oder Visualisierung von Daten werden nicht betrachtet.

Zum Verständnis dieses Lehrbuchs sind Grundkenntnisse aus den Gebieten Statistik und Datenbanktechnik notwendig. Wissen über KDD oder DM wird nicht benötigt. Zur Zielgruppe gehören daher insbesondere Studenten der Informatik und Mathematik im Hauptstudium. Wissenschaftler und Praktiker mit mathematischen Grundkenntnissen werden ebenfalls auf ihre Kosten kommen. Um einen speziellen Algorithmus zu implementieren und an ein Problem anzupassen wird aber weiterführende Literatur benötigt. Für Manager und Berater ist dieses Buch nicht zu empfehlen, da es zu sehr die Techniken des DM fokusiert und Aufbau, Management und Wartung von KDD- oder DM-Systemen nicht betrachtet.

Kontaktadresse:
Dipl.-Inform. Jörg Rech
AG Software Engineering
Universität Kaiserslautern
67653 Kaiserslautern
email: rech@informatik.uni-kl.de