Raphael FOCKEL, Paderborn

Data Mining und Knowledge Discovery in Databases

Data Mining (engl.: „Daten schürfen“) als Kern des Knowledge Discovery in Databases („Wissensentdeckung in Datenbanken“) kennzeichnet die Identifikation von Mustern, Zusammenhängen und Trends in Datenbeständen. Es handelt sich um eine eigenständige Disziplin, die sich insbesondere aufgrund des Phänomens der Datenflut (Informationsflut) entwickelt hat. Im täglichen Arbeitsumfeld sammeln sich fortlaufend große Mengen an strukturierten und unstrukturierten Daten, in denen häufig interessante und wertvolle Informationen versteckt sind, die jedoch mit bloßem Auge kaum noch zu erkennen sind. Beim Data Mining wird nun versucht mithilfe bestimmter mathematisch-statistischer Verfahren in akzeptabler Rechenzeit valide, neuartige, potenziell nützliche und verständliche Muster in diesen Daten zu suchen und verwertbar zu machen.

Im Rahmen eines Projektes an der Universität Paderborn versuchen wir uns in zweierlei Hinsicht der Wissenschaftsdisziplin des Data Mining zu nähern. Einerseits unter einer *Forschungsperspektive*. Hier geht es um die kritische Analyse von Einsatzpotenzialen des Data Mining bei der Entscheidungsunterstützung in Lehre und Forschung & Entwicklung unter besonderer Berücksichtigung des Datenschutzes, z.B. bei der Evaluation von Lehrveranstaltungen (Scientific Data Mining, Educational Data Mining). Hierbei handelt es sich um einen Bereich, der in der Schnittstelle zwischen Statistik, Wirtschaftsinformatik und Mathematikhochschuldidaktik steht. Andererseits versuchen wir in einer *didaktischen Perspektive* herauszufinden, inwiefern das Themengebiet „Data-Mining-Methoden“ im Rahmen von Lehrveranstaltungen zur „Angewandten Mathematik“ einen Mehrwert bieten kann. Im Folgenden sollen zunächst typische Anwendungsgebiete und Ausprägungen des Data Mining vorgestellt werden, anschließend sollen Data-Mining-Methoden in der Lehre thematisiert werden.

Anwendungsgebiete des Data Mining

Data Mining hat sich als eigenständige Wissenschaftsdisziplin etabliert, die in der Schnittstelle mit benachbarten Disziplinen wie Statistik (insbesondere multivariate Datenanalyse), Maschinelles Lernen, Künstlicher Intelligenz und Visualisierungstechniken anzusiedeln ist. Der Einsatz der Methoden in der Praxis ist dabei in den vergangenen Jahren sprunghaft angestiegen. In naher Zukunft wird es vermutlich kaum noch Bereiche in Wissenschaft, Technik, Medizin, Handel, Banken, Verwaltung usw. geben, die sich nicht Data Mining bedienen werden. Klassische Beispiele für die Anwendung des Data Mining (die durchaus auch unter einer didaktischen Per-

spektive interessant sein können) finden sich bei der *Segmentierung* von Kunden durch Unternehmen aufgrund ähnlicher Eigenschaften, um neue Erkenntnisse über die Kunden herauszufinden. Solche grundlegenden Data-Mining-Methoden zur Segmentierung finden sich vielfach auch bei Internet-Suchmaschinen wie Google oder in Bilderportalen wie flickr.de (Clustern von Bildern bei der Bildersuche). Ein weiteres Anwendungsbeispiel findet sich bei *Assoziationsanalysen*. So kann ein Handelsunternehmen aufgrund der Einkäufe der Kunden mithilfe von Assoziationsregeln herausfinden, welche Produkte häufig zusammen, d.h. im Verbund, gekauft wurden. Eine so gefundene Regel kann dabei z.B. wie folgt aussehen: „In 80% aller Einkäufe, bei denen Wein und Tomaten gekauft wurden, wurden auch Spaghetti gekauft". Aufgrund solcher Informationen können Händler ihre Waren im Geschäft sinnvoll anordnen oder dementsprechend ihre Werbung ausrichten. Typisch sind solche Assoziationsanalysen mittlerweile in vielen Internetshops. So findet sich beispielsweise bei amazon.de bei Bücherangeboten die Produktempfehlung „Kunden, die diesen Artikel gekauft haben, kauften auch: …", die auf Data-Mining-Analysen beruht.

Ausprägungen des Data Mining

Beim Data Mining gibt es mittlerweile eine Vielzahl an Ausprägungen. Beispielsweise kennzeichnet das *Web Mining* die Anwendung von Data-Mining-Methoden auf Internetdaten. So können Data-Mining-Verfahren eingesetzt werden, um das Nutzerverhalten von Webseitenbesuchern auf interessante Muster zu analysieren. Mithilfe des *Text Mining* können aus digital vorliegenden Texten neue und relevante Zusammenhänge entdeckt werden. Mit dem Aufkommen digitaler Medien durch Bilder-, Video- oder Audiodateien kommt auch dem *Multimedia Data Mining* eine immer größere Bedeutung zu. Hierzu gehört z.B. die Klassifizierung von Fotos, Sprache oder Musik. Im *Scientific Data Mining* geht es um die Anwendung von Data-Mining-Methoden an Datensätzen speziell aus Bereichen von Wissenschaft und Technik. Hieran anknüpfend hat sich in den letzten Jahren, insbesondere in der englischsprachigen Literatur der Wirtschaftsinformatik und Informatikdidaktik, eine Disziplin entwickelt, die sich mit dem Einsatz des Data Mining im Bildungsbereich beschäftigt. Beim sog. *Educational Data Mining* werden Daten auf Muster untersucht, insbesondere um Lernende und ihre jeweilige Lernumgebung besser verstehen zu können. Beispiele für das Educational Data Mining finden sich in der Prognose des Studienerfolgs und des Studienverhaltens, der Auswertung des Nutzerverhaltens in E-Learning-Kursen, beim Erkennen von schwer verständlichen Inhalten in Materialien, beim Erkennen von Hürden in Lernprozessen oder auch bei der Evaluation von Lehrveranstaltungen.

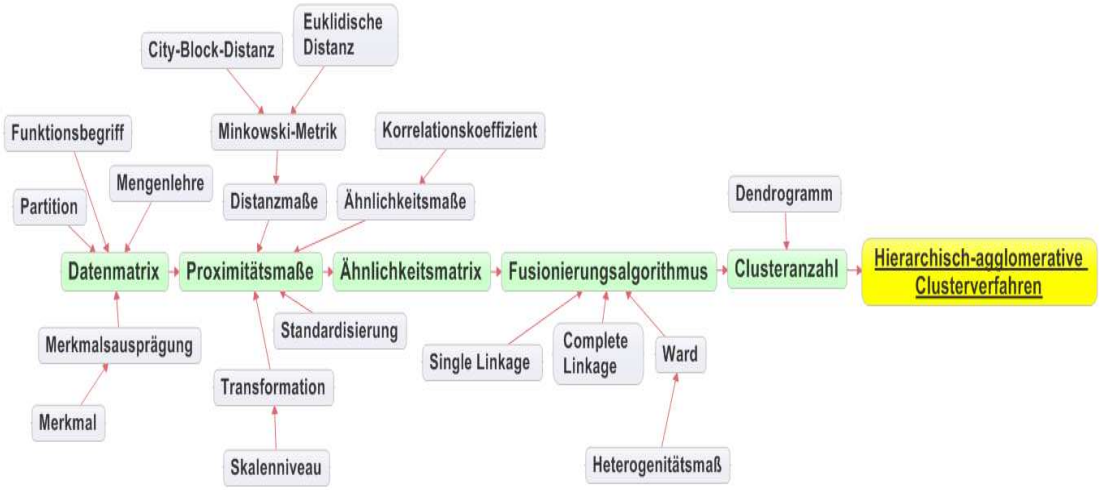
Data Mining in der Lehre

Neben der zuvor genannten Forschungsperspektive ist je nach Zielgruppe die Thematisierung grundlegender Methoden des *Data Mining in der universitären Mathematikausbildung* (auch im Bereich des schulischen Mathematikunterrichts) durchaus denkbar. Es existieren einige Data-Mining-Methoden, die zum Teil durch einen verblüffend einfachen Algorithmus gekennzeichnet sind und interessante Zusammenhänge zur Elementaren Statistik liefern können. Methoden wie Clusteranalysen, Naive-Bayes-Verfahren, Nearest-Neighbor-Verfahren, Assoziationsanalysen oder Entscheidungsbäume können in vereinfachter Form in der Lehre thematisiert werden. (Exemplarisch versucht der unten abgebildete stark vereinfachte Netzplan, wesentliche Zusammenhänge und Voraussetzungen aufzuzeigen, die grundlegend für das Verständnis einer hierarchisch-agglomerativen Clusteranalyse sind.) Es gibt dabei verschiedene Ziele, die man mit dem Einsatz des Themengebiets „Data-Mining-Methoden“ im Rahmen von Lehrveranstaltungen zur Angewandten Mathematik oder Statistik erreichen kann. Exemplarisch seien die folgenden Ziele genannt:

- *Interdisziplinäres Arbeiten:* Erkennen von Zusammenhängen z.B. zur Linearen Algebra, Elementarer Statistik, Informatik ,Wirtschaft, usw.
- *Entdeckendes Lernen* und Förderung von *Problemlösefähigkeit*. An idealtypischen Datensätzen können selbständig Muster entdeckt werden. Es bieten sich verschiedene Aufgaben (z.B. Segmentierungen, Prognosen, Assoziationen) und Lösungswege an, um Informationen aus Daten zu extrahieren.
- Kenntnisse zu webbezogenen *Anwendungsgebieten/Internettechnologien*. So kann an verschiedenen Problemstellungen thematisiert werden, wie Unternehmen wie z.B. Google, Yahoo, Amazon oder Vodafone Data-Mining-Methoden einsetzen.
- *Kritikfähigkeit* im Umgang mit Daten (Welche Daten kann man im Internet, z.B. in sozialen Netzwerken über sich preisgeben? Was können Unternehmen damit anfangen?).
- *Arbeiten mit Werkzeugen:* Statistik- und Data-Mining-Software (Potenziale bietet Software wie XLMiner, Statistica, SPSS, R Commander).

In den von uns durchgeführten und ausgewerteten Mathematikseminaren (gekennzeichnet durch Vorträge und Ausarbeitungen der Studierenden) zeigte sich, dass die verschiedenen Anwendungsgebiete und Methoden von den Studierenden zumeist als recht interessant bewertet wurden. Es zeigten sich vielschichtige Diskussionen, beispielsweise zu Methoden und Algorithmen oder es wurden Vermutungen über Prognoseergebnisse aufgestellt. Die zunächst recht komplex und fremdartig anmutenden Methoden und

Problemstellungen konnten im Wesentlichen von den Studierenden erarbeitet und erklärt werden, was häufig zu Erfolgserlebnissen führte. Es zeigte sich in den Seminaren jedoch auch, dass die Beziehungen zu Bereichen der Elementaren Mathematik und Statistik nicht immer von alleine hergestellt und erkannt wurden, z.B. Skalenniveaus bei Proximitätsmaßen. Ursache für solche Problemstellungen kann u.a. darin gefunden werden, dass die zu bearbeitende Literatur aus Gebieten der Informatik oder Wirtschaftswissenschaften einen anderen Adressatenkreis und andere Zielsetzungen hat. Fachmathematische Zusammenhänge, wie sie insbesondere für Mathematikseminare wichtig sind, werden dabei verständlicherweise weniger deutlich angesprochen. So konzentriert sich beispielsweise Literatur aus dem informatischen Bereich eher auf die Programmierung von Algorithmen, Literatur aus dem wirtschaftswissenschaftlichen Bereich eher auf Anwendungspotenziale. Bedarf liegt daher in einem adressatengerechten didaktischen Gesamtkonzept, das dementsprechend, je nach Zielgruppe, Data Mining eher aus einer fachmathematischen und mathematikdidaktischen Perspektive betrachtet und zusätzlich, so weit wie möglich, einen Praxisbezug zum späteren Berufsfeld deutlich werden lässt. Insbesondere aufgrund der rasanten Entwicklung von Internettechnologien, der zunehmenden Datenflut sowie der Marktentwicklung der Statistiksoftware kann eine frühzeitige Auseinandersetzung mit dem Themengebiet „Data Mining“ gleichermaßen gewinnbringend für Forschung und Lehre sein.



Literatur

Fahrmeir, L., Hamerle, A., Tutz, G. (Hrsg.) (1996). Multivariate statistische Verfahren, 2. Auflage; Berlin; New York: de Gruyter

Han, J., Kamber, M. (2006). Data mining. Concepts and Techniques. 2. Auflage, Amsterdam: Elsevier/Morgan Kaufmann

Otte, R.; Otte, V.; Kaiser, V. (2004). Data Mining für die industrielle Praxis, München, Wien: Carl Hanser Verlag

Romero, C., Ventura, S. (2007). Educational Data Mining. A Survey from 1995 to 2005. In: Expert Systems with Applications 33(1), S. 135-146