Christian Gottermeier

Data Mining

Modellierung und Durchführung ausgewählter Fallstudien mit dem SAS Enterprise Miner

Diplomarbeit

Diplom.de

Bibliografische Information der Deutschen Nationalbibliothek:

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

ISBN: 9783832472177

Christian Gottermeier

Data Mining

Modellierung und Durchführung ausgewählter Fallstudien mit dem SAS Enterprise Miner

Diplom.de

Christian Gottermeier

Data Mining

Modellierung und Durchführung ausgewählter Fallstudien mit dem SAS Enterprise Miner

Diplomarbeit
Ruprecht-Karls-Universität Heidelberg
Fachbereich Wirtschafts- und Sozialwissenschaften
Abgabe Januar 2003

Diplom.de

Diplomica GmbH
Hermannstal 119k
22119 Hamburg

Fon: 040 / 655 99 20
Fax: 040 / 655 99 222

agentur@diplom.de
www.diplom.de

ID 7217

ID 7217
Gottermeier, Christian: Data Mining - Modellierung und Durchführung ausgewählter Fallstudien mit dem SAS Enterprise Miner
Hamburg: Diplomica GmbH, 2003
Zugl.: Ruprecht-Karls-Universität Heidelberg, Universität, Diplomarbeit, 2003

Diplomica GmbH
http://www.diplom.de, Hamburg 2003
Printed in Germany

I N H A L T S V E R Z E I C H N I S

1. Einführung

Entscheidungen sind ein Akt des menschlichen Verhaltens, bei denen eine Festlegung für eine unter mehreren Möglichkeiten stattfindet. Da bei diesen Handlungen die Berufung auf Traditionen oder Autoritäten oftmals nicht möglich ist, wurde schon früh auf verschiedenste Hilfsmittel zurückgegriffen. So ließ sich Julius Cäsar von einem Würfelergebnis leiten, General Wallenstein von einem Astrologen beraten oder es wurden Prognosen mit Hilfe von Glaskugeln, Spielkarten oder dem Stand der Sterne getroffen.

Unter wirtschaftlichen Gesichtspunkten sind Entscheidungen eine rationale Wahl zwischen mehreren Möglichkeiten, wobei der Entscheidungsprozess als tragendes Element der ökonomischen Tätigkeit herausgestellt wird. Gerade in diesem Umfeld wird die Entscheidungsfindung – nun allerdings wissenschaftlich fundiert und mit weitreichenden Konsequenzen – durch folgende Verfahren unterstützt: Analysemethoden wie Benchmarking, Lebenszyklus- oder Erfahrungskurvenkonzept und Prognoseverfahren wie die Delphi-Methode oder die Szenario-Technik. Allerdings sind die meisten dieser Verfahren i.d.R. auf spezielle Problemstellungen ausgerichtet. Ganzheitliche Lösungsansätze werden seit den 60er Jahren zur Unterstützung des Managements bereitgestellt. Mit Hilfe von Informationssystemen soll die Entscheidungsfindung verbessert werden. Häufig wechselnde Schlagworte wie z.B. Management Information System (MIS) oder Decision Support System (DSS) konnten allerdings noch keine durchschlagenden Erfolge erzielen. Seit Mitte der 90er Jahre wurden mit neuen konzeptionellen Ansätzen, die meist unter dem Oberbegriff „Business Intelligence“ zusammengefasst werden, erfolgsversprechende Lösungen zum Aufbau entscheidungsorientierter Informationssysteme (EIS) etabliert. EIS setzen sich dabei aus Werkzeugen zur Selektion und Speicherung entscheidungsrelevanter Informationen (Data Warehouse) sowie zur entscheidungsunterstützenden Modellierung (OLAP-Tools) zusammen. Eine konsequente Umsetzung des Data Warehouse Gedanken führt zu immensen Datensammlungen, die, um die Archivierung nicht zum Selbstzweck werden zu lassen, dann auch ausgewertet werden sollen. An dieser Stelle setzt Data Mining an.

In Kapitel 2 werden die Grundzüge des Data Mining dargestellt, eine Verbindung zu Data Warehouse und OLAP gezogen und die Einsatzgebiete skizziert, in denen sich Data Mining durchgesetzt hat. In Kapitel 3 wird der erste wichtige Schritt, der vor der eigentlichen Modellierung stattfinden sollte, das Pre-Processing, erläutert. Die Modelle und die damit verbundenen Methodiken der Data Mining-Verfahren werden in Kapitel 4 vorgestellt. Stets wird eine Verbindung zum SAS® Enterprise Miner™ gesucht und so eine

Anpassung der dort verankerten Möglichkeiten an die Theorie vorgenommen. Die Vorgehensweise der Modellbewertung und die dafür existierenden Kriterien werden in Kapitel 5 dargestellt. Die praktische Umsetzung der Data Mining-Modelle wird anhand verschiedener Fallstudien im sechsten Kapitel gezeigt. Dafür werden die von der SAS® Institute Inc. erstellten Fälle bearbeitet. Diese Daten sind stark idealisiert, d.h. sofort analysierbar und deshalb sehr gut geeignet, um die einzelnen Schritte Pre-Processing, Modellierung der einzelnen Verfahren und Modellbewertung durchzuführen.

2. Data Mining

2.1 Definitionen und Erklärungen

Durch die Entwicklung der Informationstechnologie ist die Datenerhebung, -speicherung und -verwaltung stark ausgeprägt. Datenbanken der Größenordnung Gigabyte oder Terabyte sind weit verbreitet. Schätzungen zufolge verdoppeln sich die weltweit vorhandenen Informationen alle 20 Monate, in Datenbanken ist die Rate noch größer. Allerdings konnten die Möglichkeiten der manuellen Analysetechnik mit dieser Entwicklung nicht Schritt halten. Die Folge ist, dass die Datenmengen nicht zu aussagefähigen Informationen verdichtet werden. Der verschärfte Wettbewerbsdruck zwingt aber zur Nutzung aller Informationsquellen.

Der Begriff Data Mining bezeichnet eine relativ neue Forschungs- und Anwendungsrichtung, obwohl die Bestandteile schon lange existieren. Verfahren der klassischen statistischen Datenanalyse, Anwendungen aus der künstlichen Intelligenz (KI), der Mustererkennung und des maschinellen Lernens wurden in das sog. Knowledge Discovery in Databases (KDD) integriert. Wie der Name schon assoziiert, besteht auch eine starke Verbindung zur Datenbanktechnologie. KDD und Data Mining werden in dieser Arbeit und auch meistens in der Literatur simultan verwendet.

Die Aufgabe des Data Minings ist die Entwicklung, Implementierung und Ausführung von Datenanalysemethoden. Ein Fokus wird dabei auf sehr große Datensätze mit komplexen Strukturen gelegt. Die vier Hauptaufgaben sind:

a.) Vorhersage- und Klassifikationsmodelle,

b.) Segmentierungen oder Clusterung,

c.) Dimensionsreduktion und

d.) Assoziationsanalysen.

Vorhersage und Klassifikation unterscheiden sich hauptsächlich in den Skalierungsmaßen. Ist die abhängige Variable intervallskaliert, spricht man von Vorhersagemodellen. Klassifikationsmodelle dagegen besitzen einen binären, ordinalen oder nominalen Regressand. Man spricht von Klassifikation, weil die Klassenwahrscheinlichkeiten angegeben werden sollen. Vertreter dieser Modelle sind u.a. die Regressionsanalyse, das Entscheidungsbaumverfahren oder die künstlichen neuronalen Netze.

Die Besonderheit des Clusterung ist das Fehlen einer abhängigen Variablen. Als Modelle, die eine Klassenunterscheidung aufgrund der Homogenität bzw. Heterogenität der unabhängigen Variablen vornehmen, können beispielsweise das K-Means-Verfahren oder sog. selbstorganisierende Karten, die Kohonen-Netze, angeführt werden.

Die Dimensionsreduktion ist in komplexen Systemen unerlässlich, um eine Visualisierung der Problemstellung zu ermöglichen. Auch wenn eine Reduktion auf zwei bis drei Dimensionen, die eine grafische Anschauung ermöglicht, nicht gelingt, ist eine Variablenselektion das Instrument, um Modelle zu vereinfachen und damit die Fähigkeit zur Modellentwicklung zu erhöhen (vgl. auch Abschnitt 3.2).

Eine Assoziationsanalyse soll Beziehungen zwischen Variablen aufdecken. Als Verfahren werden Assoziationsregeln, Sequenzmuster oder Link-Analysen verwendet.

In den Vorhersage- und Klassifikationsmodellen liegt der Schwerpunkt dieser Arbeit, sowohl bei der Analyse der Modellierung als auch bei den Fallstudien, bei denen ausschließlich diese Modelle untersucht werden[1].

An dieser Stelle sollte auch auf die Schwierigkeiten eingegangen werden, eine allgemeingültige Definition oder eine einheitliche Terminologie zu finden. Entgegen der Gleichstellung der Begriffe KDD und Data Mining (siehe oben) existiert auch die Sichtweise, KDD als den Gesamtprozess der Analyse anzusehen und Data Mining als Synonym für die einzelnen eingesetzten Methoden zu bezeichnen. Fayyad[2], der das Thema Data Mining seit den Anfängen zu Beginn der 90er Jahre prägte, stellte folgenden Zusammenhang zwischen beiden Begriffen her:

"KDD is the non-trivial process of identifying valid, novel, potential useful and ultimately understandable pattern in data."

"Data Mining is a simple step in the KDD process that under acceptable computational efficience limitations enumerates structures (patterns or models) over the data."

Data Mining mit seinen Methoden aus den verschiedensten Gebieten ist eine interdisziplinäre Wissenschaft. Schon deshalb wird es je nach Schwerpunkt und wissenschaftlicher Herkunft verschiedene Sichtweisen, Ansätze und vor allem Notationen geben. Darüber hinaus finden sich nicht nur in der Literatur, sondern auch in der praktischen Umsetzung der verschiedenen Anbieter von Data Mining-Software unterschiedliche Ansätze, die Teilgebiete ausdrücklich ausgrenzen oder neue Methoden einführen. Der Ansatz dieser Arbeit besteht nicht darin, alle Methoden und Sichtweisen darzustellen, sondern eine Einführung in die Problematik zu bieten, aufgrund derer die nachher folgenden Fallstudien durchgeführt werden.

Die deutsche Data Mining-Übersetzung Datenmustererkennung nennt die Identifizierung von Mustern als Ziel des Data Mining-Prozesses, während der Begriff KDD eine

[1] Der Ablauf bei der Erstellung von Vorhersage- und Klassifikationsmodellen kann im Anhang in Kapitel A.1 nachvollzogen werden.

[2] Küppers (1999), S. 23.