Jürgen Cleve, Uwe Lämmel
Data Mining

Jürgen Cleve, Uwe Lämmel

Data Mining

DE GRUYTER
OLDENBOURG

Lektorat: Dr. Gerhard Pappert
Herstellung: Tina Bonertz
Grafik: Irina Apetrei

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

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Rosenheimer Straße 143, 81671 München, Deutschland
www.degruyter.com/oldenbourg
Ein Unternehmen von De Gruyter
Gedruckt in Deutschland
Dieses Papier ist alterungsbeständig nach DIN/ISO 9706.

ISBN 978-3-486-71391-6
eISBN 978-3-486-72034-1

Vorwort

Nicht nur täglich oder stündlich, sondern in jeder Sekunde werden Tausende oder Millionen von Daten gemessen, erhoben und gespeichert. Das Sammeln von Daten gehört zu unserem Alltag. Warum werden Daten gesammelt? Zum einen werden Daten gesammelt, weil es möglich ist und das Speichern kaum oder kein Geld kostet. Zum anderen wird selbstverständlich auch ein Nutzen erwartet, oft wird diese Frage aber erst nach dem Sammeln gestellt und beantwortet.

Das Datensammeln geschieht sowohl in unserem beruflichen als auch im privaten Umfeld: Unternehmensdaten, Finanztransaktionsdaten, Daten über den Kauf beziehungsweise Verkauf von Waren, Daten zur Steuerung von Fahrzeugen oder Prozessen bis hin zu Daten, die unseren Gesundheitszustand beschreiben. Unser Leben wird begleitet von einer Vielzahl von Daten.

Die Anzahl der Unternehmen, die sich der Erhebung von Daten und deren Auswertung verschrieben haben, steigt ständig. Daten und die daraus abgeleitete Information sind mittlerweile zu einer Handelsware und auch zu einem Produktionsfaktor geworden.

In diesem Buch behandeln wir Techniken, mit deren Hilfe solche Datenmengen ausgewertet werden können. Unser Ziel ist nicht, eine vollständige Übersicht über alle Verfahren zu geben. Vielmehr geht es in diesem Buch um eine Einführung in die Prozesse und Techniken der Mustererkennung in strukturierten Daten, die dem Teilgebiet des Data Mining zugeordnet werden. Diese Algorithmen und Basistechniken werden in vielen Anwendungen eingesetzt. Statistische Verfahren zur Datenauswertung werden nur am Rande erwähnt. Das Gebiet der Textanalyse, des sogenannten Text Minings wird von uns nicht betrachtet.

Was können Sie erwarten? Das Buch gibt eine Einführung in das Gebiet des Data Minings.

- Zunächst geben wir einen Überblick über Data Mining und diskutieren einige Grundbegriffe und Vorgehensweisen.
- Anschließend werden Anwendungsklassen beschrieben, um die Einsatzmöglichkeiten des Data Minings erkennen zu können.
- Die Möglichkeiten, die Modelle zu repräsentieren, werden vorgestellt.
- Die Verfahren zur Analyse der Daten – das eigentliche Data Mining – nehmen natürlich den größten Raum in diesem Buch ein.
- Die Daten müssen vorbereitet werden, um eine Analyse zu ermöglichen beziehungsweise die Qualität der Daten zu verbessern.

Die Qualität der vorliegenden Daten kann die Resultate stark beeinflussen. Folglich ist die Datenvorverarbeitung für den Erfolg einer Datenanalyse sehr wichtig. Der Datenvorbereitung ist deshalb ein separates Kapitel gewidmet.

- Es schließt sich die Bewertung der Resultate an. Sind sie neu, sind sie wirklich relevant?
  In diesem Kapitel gehen wir auch auf Möglichkeiten der Visualisierung ein. Das Ergebnis einer Datenanalyse muss geeignet dargestellt werden. Die adäquate Visualisierung der Resultate ist für die Akzeptanz einer Datenanalyse wichtig.

- Anhand eines größeren Beispiels wird die Vorgehensweise im Data Mining illustriert. Die vorher eingeführten Analysetechniken werden eingesetzt.

Den Block zum eigentlichen Data Mining haben wir etwas anders aufgebaut. Meistens findet man in Lehrbüchern Abschnitte, die sich beispielsweise ausschließlich mit Klassifizierung befassen. Dort werden dann alle Verfahren behandelt, die für eine Klassifizierung geeignet sind. Dies hat sich aus unserer Sicht als ungeschickt herausgestellt, da zum Beispiel Entscheidungsbäume sowohl für die Klassifizierung als auch die numerische Vorhersage geeignet sind. Folglich haben wir die Aufteilung etwas anders vorgenommen. Zunächst werden im Kapitel 3 Anwendungsklassen vorgestellt. Welche typischen Anwendungsgebiete gibt es für Data Mining?

Im Kapitel 4 wird dann auf geeignete Darstellungsmöglichkeiten für Data-Mining-Modelle eingegangen. Wie kann ein Klassifikationsmodell dargestellt werden? Wie kann man eine Cluster-Aufteilung repräsentieren?

In den anschließenden Kapiteln werden dann die Verfahren – den jeweiligen Anwendungsklassen zugeordnet – behandelt. Man wird also zum Thema Klassifikation in mehreren Kapiteln etwas finden.

Die Datenvorbereitung (Kapitel 8) haben wir, obwohl diese Phase natürlich in einem Data-Mining-Prozess die erste – und meistens auch eine für den Erfolg sehr wichtige – Etappe ist, weiter hinten platziert, da dies für den Leser unserer Meinung nach einfacher zu verstehen ist, wenn man die Data-Mining-Algorithmen bereits kennt. Erst nach dem Verstehen der Data-Mining-Verfahren kann man nachvollziehen, wieso und wie man bestimmte Daten vorverarbeiten muss.

Im Kapitel 9 betrachten wir einige Techniken zur Bewertung der Resultate, die durch Data Mining erzielt wurden.

Dieses Buch ist ein Lehrbuch. Data Mining ist mittlerweile in fast allen Curricula von Studiengängen mit einem Informatikbezug enthalten. Anliegen dieses Buchs ist es, eine Einführung in das interessante Gebiet des Data Minings zu geben. Wir haben bewusst bei einigen Verfahren auf die Darstellung der zugrunde liegenden mathematischen Details verzichtet. Ebenso haben wir uns auf grundlegende Algorithmen konzentriert.

Data Mining ist ein Gebiet, welches Erfahrung verlangt. Ein Projekt, in dem blind Data-Mining-Werkzeug eingesetzt werden, wird sehr selten erfolgreich sein. Das Verständnis für den jeweiligen Gegenstandsbereich, die vorliegenden Daten, aber eben auch für die Data-Mining-Verfahren ist eine notwendige Voraussetzung für ein erfolgreiches Projekt.

Die Beispiele haben wir zu großen Teilen in KNIME [KNI], WEKA [WEK] und JAVANNS [JAV] implementiert. Das Buch enthält viele Screenshots von KNIME-Workflows. Wir empfehlen, die Beispiele mittels der Werkzeuge nachzuvollziehen. Hierzu sollte man sich vorab die kurzen Einführungen in diese Werkzeuge nicht nur durchlesen, sondern die Systeme auch installieren und ausprobieren.

Es gibt mittlerweile eine Reihe von Wettbewerben, bei denen reale Probleme zu lösen sind. Die Autoren haben mit ihren Studenten mehrfach am Data Mining Cup [DMC] teilgenommen, der seit einer Reihe von Jahren unter der Leitung der Chemnitzer Firma `prudsys` durchgeführt wird. Ferner gibt es viele Plattformen, auf denen echte Probleme zu lösen sind, beispielsweise die Plattform `www.kaggle.com`.

Wer an Zusatzinformationen zum Thema Data Mining interessiert ist, findet viele gute Seiten im WWW. Informationen zu unserem Buch finden Sie unter:

`www.wi.hs-wismar.de/dm-buch`

An dieser Stelle möchten wir uns ausdrücklich bei der Firma prudsys sowie den Entwicklern von KNIME und WEKA bedanken, die uns die Verwendung von Beispielen aus ihrem Umfeld gestattet haben. Die Zusammenarbeit mit dem Oldenbourg-Verlag, insbesondere mit Herrn Breimeier und Herrn Dr. Gerhard Pappert, war sehr angenehm und konstruktiv. Ebenso danken wir Tobias Aagard und Roman Knolle, die mit vielen kritischen Hinweisen zum Gelingen des Buchs beigetragen haben.

Wismar, Januar 2014 Jürgen Cleve und Uwe Lämmel

Inhaltsverzeichnis

1 Einführung

Data you don't need is never lost.
Ander's first negative Principle of Computers

1.1 Auswertung von Massendaten

Die Menge an verfügbaren Daten verdoppelt sich in immer kürzeren Abständen. Jedes Unternehmen, jede Institution sammelt freiwillig oder aufgrund rechtlicher Bestimmungen Unmengen an Daten. Banken speichern die Transaktionsdaten; Firmen speichern nicht nur Daten über ihre Kunden, beispielsweise über deren Kaufverhalten; Wetterinstitute sammeln Wetterdaten.

Denken Sie einmal nach: Wissen Sie, wer welche Daten über Sie speichert? Wenn Sie berufstätig sind, wie viele Daten werden von Ihrem Unternehmen gesammelt?

Durch leistungsfähige Computer sind wir nun nicht nur in der Lage, diese Daten zu speichern, sondern wir können diese Datenmengen auch analysieren und somit interessante Informationen generieren.

Mit der Veröffentlichung von Datenmengen ist man in den letzten Jahren deutlich vorsichtiger geworden, daher zunächst einige Beispiele aus dem Jahr 2000 [Run00]. Die anfallenden Datenmengen sind seitdem sicher auf ein Vielfaches gestiegen.

Industrielle Prozessdaten
Zur Analyse der Altpapieraufbereitung in einer Papierfabrik stehen an jeder der 8 DeInking-Zellen jeweils 54 Sensoren zur Verfügung: 3.800.000 Messwerte pro Tag.

Umsatzdaten
WalMart führte eine Warenkorb-Analyse durch, 20 Millionen Transaktionen pro Tag erforderten eine Datenbank mit einer Kapazität von 24 Terabyte.

Genom-Daten
In vielen Genom-Projekten wird versucht, aus den Genomen Informationen zu extrahieren. Das menschliche Genom enthält 60.000–80.000 Gene, das sind etwa 3 Milliarden DNA-Basen.

Bilder
Die NASA nimmt mit ihrem *Earth Observing System* Oberflächenbilder der Erde auf: 50 GByte pro Sekunde.

Textinformationen
Das World Wide Web enthält eine rapide wachsende Anzahl von Seiten und damit verbunden Daten und Informationen.

Die Frage, die sich natürlich sofort ergibt: Was machen wir mit den ganzen Daten, die wir täglich sammeln? Zur Auswertung von umfangreichen Daten reicht „Draufschauen“ nicht mehr aus. Auch die Statistik stößt häufig an ihre Grenzen. Vielmehr benötigt man weitere *Datenanalyse-Techniken*. Man sucht in den Daten nach Mustern, nach Zusammenhängen, um so beispielsweise Vorhersagen für ein bestimmtes Kundenverhalten treffen zu können.

Diese Suche nach Mustern oder Zusammenhängen in den Daten ist Gegenstand des *Data Minings*. Während man im Bergbau, zum Beispiel beim Coal Mining, die Kohle sucht und abbaut, will man im Data Mining nicht die Daten „abbauen“, sondern man sucht nach den Schätzen, die in den Daten verborgen sind.

Data Mining sucht nach unbekannten Mustern und Abhängigkeiten in den gegebenen Daten. Eines dieser Suchziele ist es, Objekte in Klassen einzuteilen, die vorher bekannt oder auch unbekannt sein können.

Folgende Anwendungsbeispiele verdeutlichen die praktische Relevanz:

- Kredit oder kein Kredit?
  Aus alten Kreditdaten werden Regeln als Entscheidungshilfe für die Bewertung der *Kreditwürdigkeit* eines Kunden abgeleitet.
- Typen von Reisenden:
  Man generiert Muster von typischen *Reisenden*, um auf den jeweiligen Kundentypen zugeschnittene Angebote zusammenzustellen.
- Windeln und Bier:
  Es wird das *Kaufverhalten* von Kunden analysiert, um logische Abhängigkeiten zwischen Produkten zu finden. Wer Windeln kauft, nimmt häufig auch Bier.
- Gen-Analyse:
  Es werden die Gene von *Diabetes-Kranken* mit dem Ziel untersucht, die für die Krankheit vermutlich verantwortlichen oder mitverantwortlichen Gene zu identifizieren.

Data Mining befasst sich mit der Analyse von Massendaten. Das Ziel des Data Minings ist es, aus Massendaten – wie beispielsweise Kunden- oder Unternehmensdaten, Unternehmenskennzahlen und Prozessdaten – nützliches Wissen zu extrahieren. Gelingt dies, so kann daraus ein entscheidender Wettbewerbsvorteil für das Unternehmen im Markt entstehen. Data Mining lässt sich somit in die Gebiete *Wissensmanagement* – es wird Wissen aus Daten extrahiert – und *Business Intelligence* – es werden die verschiedensten Daten aus unterschiedlichen Bereichen analysiert – einordnen.

1.2 Data Mining und Business Intelligence

Schaut man sich das Lexikon der Wirtschaftsinformatik an [KBG+12] und hier den Beitrag zum Data Mining [Cha12], so stellt man fest, dass das Data Mining als eine *Herangehensweise analytischer Informationssysteme, die wiederum dem Business Intelligence untergeordnet sind*, eingeordnet ist.

Die Begriffshierarchie aus dem Lexikon der Wirtschaftsinformatik:

Informations-, Daten- und Wissensmanagement
⇓
Business Intelligence
⇓
Analytische Informationssysteme, Methoden der
⇓
Data Mining

Wir schließen uns dieser Sichtweise an und sehen in Data Mining eine Sammlung von Techniken, Methoden und Algorithmen für die Analyse von Daten, die somit auch Grundtechniken für neuere und komplexere Ansätze, wie das *Business Intelligence* oder auch *Big Data* darstellen.

Business Intelligence (BI) ist ein relativ neuer Begriff. Ausgangspunkt ist die Beobachtung, dass in Zeiten der Globalisierung und des Internets ein effektiver und effizienter Umgang mit dem in einem Unternehmen verfügbaren Wissen nicht nur ein Wettbewerbsvorteil, sondern für das Überleben wichtig ist.

Unter Business Intelligence werden heute Techniken und Architekturen für eine effiziente Verwaltung des Unternehmenswissens zusammengefasst, natürlich einschließlich verschiedener Auswertungsmöglichkeiten. Die Aufgaben von Business Intelligence sind somit:

- Wissensgewinnung,
- Wissensverwaltung und
- Wissensverarbeitung.

Business Intelligence hat – aus Informatik-Sicht – viele Querbezüge zum Informations- und Wissensmanagement, zu Datenbanken und Data Warehouses, zur Künstlichen Intelligenz, sowie natürlich auch zum Data Mining (einschließlich OLAP – Online Analytical Processing, Statistik).

Eine allgemein akzeptierte Definition des Begriffs *Business Intelligence* gibt es bis heute nicht. Hinweise zur Entstehungsgeschichte des Begriffes kann man wieder dem Lexikon der Wirtschaftsinformatik entnehmen [KBG+12]. Man findet diese in [Hum12] unter dem entsprechenden Stichwort.

Unter Business Intelligence im engeren Sinn versteht man die Kernapplikationen, die eine Entscheidungsfindung direkt unterstützen. Hierzu zählt man beispielsweise das

Online Analytical Processing (OLAP), die Management Information Systems (MIS) sowie Executive Information Systems (EIS).

Ein etwas weiterer BI-Begriff stellt die Analysen in den Vordergrund. Folglich gehören hierzu alle Anwendungen, bei denen der Nutzer Analysen durchführt oder vorbereitet. Neben den oben genannten Anwendungen zählt man nun auch beispielsweise das Data Mining, das Reporting sowie das analytische Customer Relationship Management dazu.

Business Intelligence im weiten Verständnis umfasst schließlich alle Anwendungen, die im Entscheidungsprozess benutzt werden, also beispielsweise auch Präsentationssysteme sowie die Datenspeicherung und -verwaltung.

Schwerpunkt dieses Buchs sind die Techniken zur Wissensextraktion mittels Data Mining. Wir betrachten folglich nur einen kleinen Ausschnitt aus dem BI-Spektrum.

Für weiterführende Informationen sei auf [TSDK11] verwiesen.

Der Zusammenhang zwischen Data Mining und *Data Warehouses* ist offensichtlich. Data Warehouses haben den Anspruch, integrierte Daten für die Unterstützung von Managemententscheidungen bereitzuhalten, und sollten folgende Eigenschaften aufweisen (vgl. [Pet05, S. 40 ff.]):

- Es sollte sich am Nutzer, dem Entscheidungsträger oder Manager orientieren und so insbesondere den Informationsbedarf des Managements bedienen.
- Es umfasst alle entscheidungsrelevanten Daten in einer konsistenten Form.
- Ein Data Warehouse ist nur die „Sammelstelle“ für Daten aus externen Quellen. Eine Aktualisierung der Daten erfolgt normalerweise nur in fest definierten Abständen. Der Zugriff auf die Daten erfolgt dann im Data Warehouse nur noch lesend.
- Die Daten müssen zeitabhängig verwaltet werden, so dass Trends erkannt werden können.
- Die Daten werden nicht 1:1 aus den Quellen übernommen, sondern bereits kumuliert oder gefiltert. In einem Data Warehouse ist somit Redundanz möglich.

Um diesen Anforderungen gerecht zu werden, sind – neben Komponenten zur effizienten und konsistenten Datenhaltung und für schnelle, flexible Zugriffe auf die Daten – natürlich auch Werkzeuge zur Datenanalyse notwendig. Deshalb verfügen viele Data-Warehouse-Systeme über Komponenten zur Datenanalyse oder zumindest über Schnittstellen zu externen Werkzeugen.

1.3 Ablauf einer Datenanalyse

Data Mining ist wie vorher beschrieben eingebettet in die analytischen Informationssysteme und kann allein oder integriert in Business Intelligence oder als Baustein eines Data Warehouses betrieben werden.

Wie läuft nun ein Data-Mining-Prozess selbst ab? Folgende Phasen können unterschieden werden:

Selektion – Auswahl der geeigneten Datenmengen:
Zunächst werden die verfügbaren Daten gesichtet und in eine Datenbank, meist sogar in *eine* Datentabelle übertragen.

Datenvorverarbeitung – Behandlung fehlender oder problembehafteter Daten:
In dieser Phase werden die Daten bereinigt. Fehler müssen beseitigt, fehlende oder widersprüchliche Werte korrigiert werden.

Transformation – Umwandlung in adäquate Datenformate:
Häufig ist in Abhängigkeit vom jeweils verwendeten Verfahren eine Transformation der Daten erforderlich, beispielsweise die Gruppierung von metrischen Werten in Intervalle.

Data Mining – Suche nach Mustern:
Hier geschieht das eigentliche Data Mining, die Entwicklung eines *Modells* wie die Erstellung eines Entscheidungsbaums.

Interpretation und Evaluation – Interpretation der Ergebnisse und Auswertung:
In der abschließenden Phase müssen die gefundenen Resultate geprüft werden. Sind sie neu, sind sie hilfreich?

In Abbildung 1.1 ist der Ablauf mit den hier vorgestellten Phasen des Data Minings dargestellt.

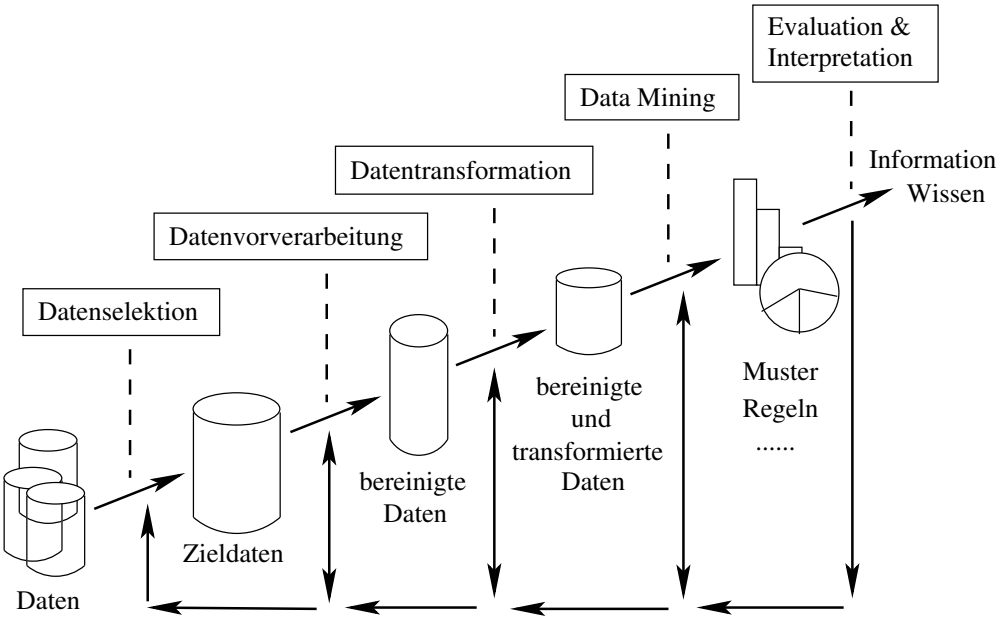

Abb. 1.1: *Ablauf eines Data-Mining-Prozesses [FPSS96]*

Häufig bezeichnet man den Gesamtprozess auch als *Knowledge Discovery in Databases (KDD)*.

Das CRISP-Data-Mining-Modell

Ein zweites Modell für Data-Mining-Prozesse ist das CRISP-Modell. Das CRISP-Modell wurde durch ein Konsortium, bestehend aus folgenden Firmen, entwickelt:

- NCR Corporation,
- Daimler AG,
- SPSS,
- Teradata und
- OHRA.

CRISP-DM steht für **Cr**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining. Ziel ist, einen Data-Mining-Prozess zu definieren und das berechnete Modell zu validieren.

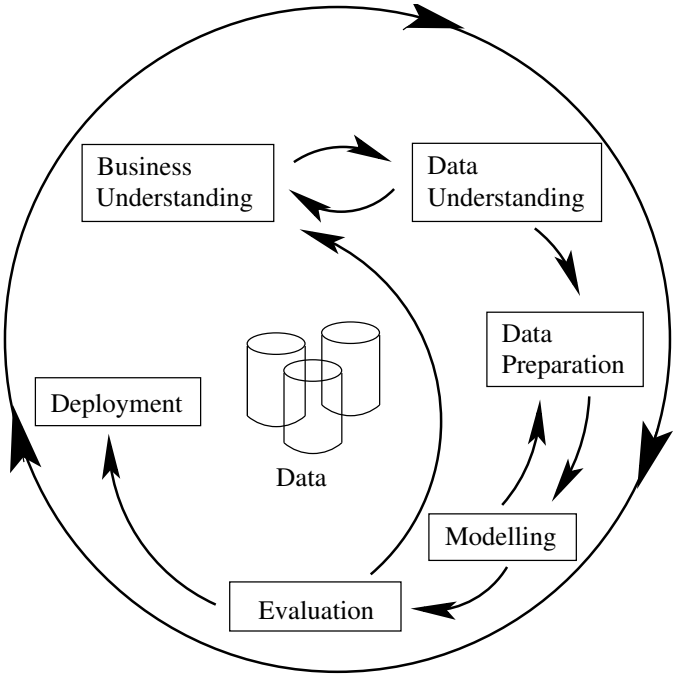

Abb. 1.2: *CRISP-Modell*

Das CRISP-Modell geht von einem Lebenszyklus in 6 Etappen aus (vgl. Abbildung 1.2):

1. **Verstehen der Aufgabe**
   Ohne ein grundsätzliches Verständnis des Fachgebiets, in dem eine Analyse stattfinden soll, ist ein gutes Resultat selten zu erzielen. In dieser Phase steht das allgemeine Verständnis der Aufgabe im Vordergrund. Es müssen die Ziele festgelegt werden: Ausgehend vom Ist-Zustand wird bestimmt, welche Ergebnisse im Rahmen des Data-Mining-Projekts erreicht werden sollen. Wie ist die Ausgangssituation? Welche Ressourcen stehen zur Verfügung?
   Weiterhin sind Erfolgskriterien zu definieren.
   Es müssen ebenso die Risiken betrachtet und möglichst quantifiziert werden. Risiken können finanzieller Natur sein. Beispielsweise kann sich das Beschaffen geeigneter Daten als aufwändiger als geplant herausstellen. Ebenso kann es passieren, dass bestimmte Daten aus rechtlichen oder firmenpolitischen Gründen nicht für das Projekt verfügbar sind.
   Natürlich müssen die Kosten geplant und der erwartete Nutzen abgeschätzt werden.
   Eine Datenanalyse, die noch nicht in die betrieblichen Prozesse integriert ist, besitzt immer Projektcharakter. Ein entsprechendes Projektmanagement ist erforderlich, insbesondere da hier Daten-Analysten, in der Regel Informatiker, mit den Anwendern aus anderen Fachgebieten zusammenarbeiten.
   Diese Phase schließt mit einem Projektplan ab.

2. **Verständnis der Daten**
   Die zweite Phase befasst sich mit den verfügbaren Daten. In Abhängigkeit vom jeweiligen Ziel der Analyse wird definiert, welche Daten benötigt werden. Parallel dazu wird untersucht, welche Daten verfügbar sind. Für ein erfolgreiches Projekt ist es erforderlich, die Daten und deren Bedeutung genau zu verstehen. Ein Data-Mining-Projekt, in dem man die verfügbaren Daten einfach nutzt, ohne ihre Semantik zu kennen, wird nicht erfolgreich sein.
   Die Daten sind also nicht nur zu sammeln, sondern sie sind auch zu beschreiben. Die Daten werden untersucht, nicht nur um sie zu verstehen, sondern auch um die Qualität der zur Verfügung stehenden Daten zu bestimmen. Erste statistische Untersuchungen, wie die Bestimmung statistischer Maßzahlen – beispielsweise Minima, Maxima, Mittelwerte sowie Korrelationskoeffizienten – geben Auskunft über die Daten.

3. **Datenvorbereitung**
   Nach dem Verstehen der Aufgabe, der betrieblichen Hintergründe und dem Verstehen der Daten gilt es nun, den eigentlichen Data-Mining-Schritt vorzubereiten. Zunächst werden die aus der Sicht der Aufgabe relevanten Daten selektiert und in eine konsistente Datentabelle überführt. Die Daten müssen gesäubert werden, fehlerhafte und inkonsistente Daten werden korrigiert. Es ist zu überlegen, wie man mit fehlenden Daten umgeht. Gegebenenfalls werden neue Attribute eingeführt oder auch Attribute zusammengeführt. Die Daten werden geeignet transformiert, damit sie durch die Data-Mining-Verfahren verarbeitet werden können. Die Datenvorverarbeitung kann über Erfolg oder Misserfolg einer Datenanalyse mitentscheiden. Deshalb stellen wir die Möglichkeiten der Datenvorverarbeitung in einem separaten Kapitel (Kapitel 8) vor.

4. **Data Mining (Modellbildung)**
   In dieser Phase geschieht die eigentliche Datenanalyse: Es wird ein Modell erstellt, welches beschreibt, wie die Daten einzuordnen oder zu behandeln sind. Modelle können Entscheidungsbäume, verschiedene Formen von Regeln oder Beschreibungen von Clustern sein.
   In Abhängigkeit von der jeweiligen Aufgabe wird ein adäquates Verfahren ausgewählt. Die durchzuführenden Experimente werden konzipiert und konfiguriert, Parameter für das Verfahren werden gesetzt. Dann werden die Experimente durchgeführt und gegebenenfalls mit modifizierten Parametern wiederholt, um das Modell zu verfeinern und zu verbessern.

5. **Evaluation**
   Die erzielten Modelle und Resultate müssen geprüft und bewertet werden. Die Ergebnisse werden an den in Phase 1 festgelegten Erfolgskriterien gemessen: Wird zum Beispiel der erwartete wirtschaftliche Nutzen durch die Ergebnisse erzielt? Eine Fehleranalyse kann Möglichkeiten für neue Experimente aufzeigen, und es wird zur Phase 3 – Datenvorbereitung – oder direkt zur Data-Mining-Phase zurückgekehrt.

6. **Einsatz im und Konsequenzen für das Unternehmen**
   In der letzten Phase gilt es, den Einsatz der erzielten Resultate vorzubereiten. Dies ist eine kritische Phase für viele Data-Mining-Projekte. Ohne ein gut geplantes Monitoring und eine ausreichende Motivation und Unterstützung kann die erfolgreiche Umsetzung scheitern.
   In der Regel ist ein Data-Mining-Projekt nur *eine* Entwicklungsphase, das Ergebnis ist dann in den Regelbetrieb zu übernehmen und in die laufenden Prozesse zu integrieren. Übliche Forderungen an einen Projektabschluss gelten natürlich auch für Data-Mining-Projekte.

Das CRISP-Modell unterscheidet sich von dem in Abbildung 1.1 auf Seite 5 dargestellten Ablauf nach Fayyad. Die Punkte 1–2 und 6 des CRISP-Modells sind im Fayyad-Modell nicht explizit aufgeführt. Der Schritt 3 des CRISP-Modells enthält die ersten drei Phasen des Fayyad-Modells. Das Fayyad-Modell konzentriert sich auf die eigentliche Datenbereitstellung und die Datenanalyse, während das CRISP-Modell die Sicht der Industrie auf Data-Mining-Projekte widerspiegelt.

Wir werden uns am in Abbildung 1.1 dargestellten Fayyad-Modell orientieren, da die Phasen 1, 2 und 6 des CRISP-Modells sehr stark projektabhängig sind und folglich nicht Bestandteil einer Einführung in das Data Mining sein können.

Wir werden zwischen den Begriffen *Data Mining* und *Knowledge Discovery in Databases (KDD)* nicht unterscheiden und diese synonym verwenden, bevorzugen jedoch den Begriff *Data Mining*.

Es gibt weitere Modelle für den Ablauf eines Data-Mining-Projekts. Das SEMMA-Vorgehensmodell wird von der Firma SAS Institute Inc. in Zusammenhang mit ihrem Produkt Enterprise Miner verwendet, siehe [SAS13]. SEMMA steht für *Sample, Explore, Modify, Model and Assess*. Ein SEMMA-Prozess besteht aus den folgenden Schritten:

1. Die für die Analyse relevanten Daten werden gesammelt (Sample).

2. Die Daten werden – vor der eigentlichen Data-Mining-Modellbildung – untersucht. Das Ziel ist, die Datenqualität zu prüfen sowie ein Datenverständnis zu erreichen. Auch erste Visualisierungen werden vorgenommen (Explore).

3. Die Daten werden modifiziert, um die Datenqualität zu verbessern. Sie werden in ein für das gewählte Verfahren adäquate Form transformiert (Modify).

4. Nun erfolgt die eigentliche Analyse und Modellbildung (Model).

5. Die Resultate werden evaluiert (Assess).

Im Folgenden wenden wir uns nun detaillierter den Teilphasen des Fayyad-Modells (Abbildung 1.1 auf Seite 5) zu, da wir uns an diesem Vorgehensmodell orientieren.

Datenselektion

In der ersten Phase des KDD-Prozesses sind die Daten, die für die vom Anwender angeforderte Analyse benötigt werden oder für eine Analyse geeignet erscheinen, zu bestimmen und aus den gegebenen Datenquellen zu exportieren. Neben dem Basisdatenbestand können auch externe Daten für die Analyse herangezogen werden. So bieten beispielsweise Adressbroker Informationen an, mit denen potentielle Kunden oder Interessenten besser erkannt werden können. In der Phase der Datenselektion wird geprüft, welche Daten nötig und verfügbar sind, um das gesetzte Ziel zu erreichen.

Können die selektierten Daten aufgrund technischer oder rechtlicher Restriktionen nicht in einen Zieldatenbestand überführt werden, ist die Datenselektion entsprechend zu überdenken und erneut durchzuführen. Technische Restriktionen, welche die Überführung in einen Zieldatenbestand verhindern, sind zum Beispiel Kapazitäts- und Datentyp-Beschränkungen des Zielsystems oder fehlende Zugriffsrechte des Anwenders. Eine Möglichkeit, diese Probleme – zumindest zum Teil – zu umgehen, ist die Beschränkung der Auswahl auf eine repräsentative Teildatenmenge des Gesamtdatenbestands.

Datenvorverarbeitung

Da die Zieldaten aus den Datenquellen lediglich extrahiert werden, ist im Rahmen der Datenvorverarbeitung die Qualität des Zieldatenbestands zu untersuchen und – sofern nötig – durch den Einsatz geeigneter Verfahren zu verbessern. Aufgrund technischer oder menschlicher Fehler können die Daten operativer Systeme *fehlerhafte Elemente* enthalten. Man rechnet damit, dass bis zu 5% der Felder eines realen Datenbestands falsche Angaben aufweisen. Die Kenntnis der Schwächen der Analysedaten ist für die Qualität der Untersuchungsergebnisse wichtig. Die Anwender der Analysewerkzeuge müssen auf die Zuverlässigkeit und Korrektheit der Daten vertrauen können. Fehlerhafte Daten verfälschen möglicherweise die Resultate, ohne dass der Anwender von diesen Mängeln Kenntnis erlangt.

Fehlende Daten verhindern eventuell die Berechnung von Kennzahlen wie den Umsatz einer Firma. Die zunehmende Durchführung (teil-)automatisierter Datenanalysen hat eine erhöhte Anfälligkeit gegenüber Datenmängeln zur Folge, der durch geeignete Mechanismen zur Erkennung und Beseitigung solcher Schwächen zu begegnen ist. Eine

häufige, leicht zu identifizierende Fehlerart besteht in *fehlenden Werten*. Zur Behandlung von fehlenden Werten stehen unterschiedliche Techniken zur Verfügung, die im Abschnitt 8.2.2 diskutiert werden.

Eine weitere potentielle Fehlerart wird durch *Ausreißer* hervorgerufen. Dabei handelt es sich um Wertausprägungen, die stark vom Niveau der übrigen Werte abweichen. Bei diesen Ausprägungen kann es sich um korrekt erfasste Daten handeln, die damit Eingang in die Analyse finden oder aber um falsche Angaben, die nicht berücksichtigt werden dürfen und daher aus dem Datenbestand zu löschen sind. Die Erkenntnisse, die der Benutzer eines Data-Mining-Systems in dieser Phase über den Datenbestand gewinnt, können Hinweise auf die Verbesserung der Datenqualität der operativen Systeme geben.

Datentransformation

Die im Unternehmen verfügbaren Rohdatenbestände erweisen sich häufig in ihrer Ursprungsform als nicht für Data-Mining-Analysen geeignet. In der Phase der Datentransformation wird der analyserelevante Zieldatenbestand in ein Datenbankschema transformiert, das von dem verwendeten Data-Mining-System verarbeitet werden kann. Dabei können neue Attribute oder Datensätze generiert beziehungsweise vorhandene Attribute transformiert werden. Dieser Schritt ist nötig, da Analyseverfahren spezifische Anforderungen an die Datenstruktur der Eingangsdaten stellen. Ziel der Transformation ist insbesondere die Gewährleistung invarianter Datendarstellungsformen (beispielsweise durch Übersetzung textueller Informationen in eindeutige Schlüssel oder Codierungen) sowie die Einschränkung von Wertebereichen zur Verringerung der Anzahl zu betrachtender Ausprägungen (Dimensionsreduktion). Letzteres kann durch Verallgemeinerung von Attributwerten auf eine höhere Aggregationsstufe, zum Beispiel durch Nutzung von Taxonomien oder durch Bildung von Wertintervallen geschehen, wodurch sich die Granularität der Daten ändert.

Data Mining

Liegen geeignete Datenbestände in akzeptabler Qualität vor, können die Analysen durchgeführt werden. In dieser Phase erfolgt die Verfahrensauswahl und deren Einsatz zur Identifikation von Mustern auf der Basis des vorbereiteten Datenbestandes. In einem ersten Schritt wird zunächst entschieden, welche grundlegende Data-Mining-Aufgabe (beispielsweise Klassifizierung oder Cluster-Bildung) vorliegt. Daran schließt sich die Auswahl eines geeigneten Data-Mining-Verfahrens an. Nach der Auswahl eines für die konkrete Problemstellung geeigneten Verfahrens wird dieses konfiguriert. Diese Parametrisierung bezieht sich auf die Vorgabe bestimmter methodenspezifischer Werte, wie zum Beispiel die Festlegung minimaler relativer Häufigkeiten für einen Interessantheitsfilter, die Auswahl der bei der Musterbildung oder -beschreibung zu berücksichtigenden Attribute oder die Einstellung von Gewichtungsfaktoren für einzelne Eingabevariablen. Wenn eine zufriedenstellende Konfiguration gefunden wurde, kann mit der Suche nach interessanten Mustern in den Daten begonnen werden. Die Analyse-Verfahren erzeugen ein Modell, welches dann als Grundlage für die Bewertung dieser oder anderer Daten dient.

Evaluation und Interpretation

In dieser Phase des KDD-Prozesses werden die entdeckten Muster und Beziehungen bewertet und interpretiert. Diese Muster sollen den Anforderungen der *Gültigkeit, Neuartigkeit, Nützlichkeit* und *Verständlichkeit* genügen, um neues Wissen zu repräsentieren und einer Interpretation zugänglich zu sein. Letztere ist Voraussetzung für die Umsetzung der gewonnenen Erkenntnisse im Rahmen konkreter Handlungsmaßnahmen. Bei weitem nicht alle der aufgedeckten Muster erfüllen diese Kriterien. Die Analyseverfahren fördern häufig viele Regelmäßigkeiten zutage, die irrelevant, trivial, bedeutungslos oder bereits bekannt waren, aus denen dem Unternehmen folglich kein Nutzen erwachsen kann, oder die nicht nachvollziehbar sind. Die Bewertung von Mustern kann anhand des Kriteriums der Interessantheit vollzogen werden. Folgende Dimensionen der *Interessantheit* sind sinnvoll:

- Die *Validität* (Gültigkeit) eines Musters ist ein objektives Maß dafür, mit welcher Sicherheit das gefundene Modell (beispielsweise ein Muster oder eine Assoziationsregel) auch in Bezug auf neue Daten gültig ist.
- Das Kriterium der *Neuartigkeit* erfasst, inwieweit ein Muster das bisherige Wissen ergänzt oder im Widerspruch zu diesem steht.
- Das Kriterium der *Nützlichkeit* eines Musters erfasst den praktischen Nutzen für den Anwender.
- Die *Verständlichkeit* misst, wie gut eine Aussage von einem Anwender verstanden werden kann.

Die korrekte Interpretation von Data-Mining-Ergebnissen erfordert ein hohes Maß an Domänenkenntnissen. Die Interpretation dient dazu, das Domänenwissen des Anwenders effektiv zu verändern. Im Idealfall wird ein Team von Experten aus unterschiedlichen Bereichen gebildet, um sicherzustellen, dass die Bewertung korrekt ist und die gewonnenen Informationen bestmöglich genutzt werden können. Die Interpretationsphase lässt sich durch geeignete Präsentationswerkzeuge sowie durch die Verfügbarkeit zusätzlicher Informationen über die Anwendungsdomäne unterstützen. Typischerweise erfolgt in dieser Phase ein Rücksprung in eine der vorherigen Phasen. So ist meist eine Anpassung der Parameter oder die Auswahl einer anderen Data-Mining-Technik nötig. Es kann auch erforderlich sein, zur Datenselektionsphase zurückzukehren, wenn festgestellt wird, dass sich die gewünschten Ergebnisse nicht mit der benutzten Datenbasis erreichen lassen.

1.4 Interdisziplinarität

Data Mining ist keine abgeschlossene Nischentechnik, die es erst seit einigen Jahren gibt. Im Gegenteil: Data Mining nutzt bewährte Techniken aus vielen Forschungsgebieten und fügt diesen neue Ansätze hinzu. Letztendlich basieren alle Analyse-Verfahren des Data Minings auf der Mathematik. Insbesondere die Statistik steuert eine Reihe eigener Ansätze für die Datenanalyse bei, wird aber auch für die Datenvorverarbeitung

eingesetzt. Statistik ist zudem Grundlage einiger Verfahren, wie zum Beispiel *Naive Bayes*.

Erst das Gebiet der Datenbanken ermöglicht die Verwaltung großer Datenmengen, und dies wird durch den synonymen Begriff für das Data Mining sogar explizit deutlich: *KDD – Knowledge Discovery in Databases*.

Die Künstliche Intelligenz als die Wissenschaft der Wissensverarbeitung stellt insbesondere Techniken für die Darstellung der Analyseergebnisse bereit: die Repräsentation von Wissen als logische Formeln und insbesondere als Regeln.

Eine andere Form der Ergebnisdarstellung als Grundlage einer Nutzung oder Bewertung ist die graphische Darstellung, die Visualisierung. Computer-Graphik ist somit eine weitere Disziplin, die im engen Kontakt zum Data Mining steht. Abbildung 1.3 illustriert diese Interdisziplinarität des Data Minings.

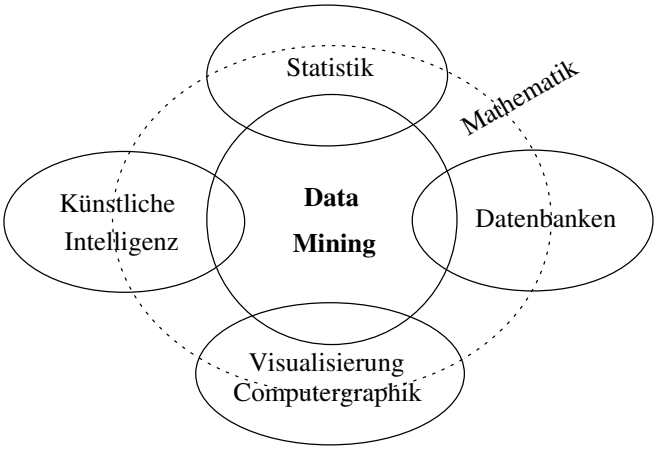

Abb. 1.3: *Interdisziplinarität*

Datenbanken und Data Warehouses

Datenbanken bilden in vielen Fällen die Grundlage des Data Minings. Häufig wird in bereits existierenden Datenbeständen nach neu zu entdeckenden Zusammenhängen oder Auffälligkeiten gesucht.

Ein *Data Warehouse* setzt sich in der Regel aus mehreren Datenbanken zusammen und enthält unter anderen auch die Daten, die zu analysieren sind. Nach Gluchowski [Glu12] sind die Merkmale eines Data Warehouse die Themenorientierung, die Vereinheitlichung, die Zeitorientierung sowie die Beständigkeit. Technisch ist die Vereinheitlichung hervorzuheben: Daten aus unterschiedlichen Quellen und mit möglicherweise verschiedenen Skalierungen oder Maßeinheiten werden korrekt zusammengeführt. Zudem werden alle Daten zeitbehaftet gespeichert, so dass Zeitreihen entstehen, die für die Auswertung genutzt werden können. Die sogenannte Beständigkeit besteht darin, dass

ein Data Warehouse beständig wächst, die mit ihrem Zeitstempel versehenen Daten werden akkumuliert.

Neben Datenbanken oder einem Data Warehouse können natürlich auch Textdateien oder WWW-Seiten Basis eines Data-Mining-Prozesses sein.

Expertensysteme

Expertensysteme – besser *wissensbasierte Systeme* versuchen, einen oder mehrere qualifizierte menschliche Experten bei der Problemlösung in einem abgegrenzten Anwendungsbereich zu simulieren. Sie enthalten große Wissensmengen über ein eng begrenztes Spezialgebiet. Sie berücksichtigen auch Faustregeln, mit denen Erfahrungen aus den Teilgebieten für spezielle Probleme nutzbar gemacht werden sollen. Gelingt es, in einem Data-Mining-Prozess Wissen aus den Daten zu extrahieren, so kann dieses Wissen dann – zum Beispiel in Form von Regeln – in einem Expertensystem repräsentiert und angewendet werden.

Maschinelles Lernen

Der Begriff *Lernen* umfasst viele komplexe Aspekte. Nicht jeder davon kann auf einem Rechner nachgebildet werden. Beim *Maschinellen Lernen* (engl. Machine Learning) versucht man, computerbasierte Lernverfahren verfügbar zu machen, so dass das Programm aus Eingabeinformationen *Wissen* generieren kann.
Bei maschinellen Lernsystemen ist – wie auch in der menschlichen Psychologie – die einfachste Lernstrategie das Auswendiglernen. Dabei wird das präsentierte Wissen einfach in einer Liste oder Datenbank abgespeichert. Eine ebenso einfache Form des Maschinenlernens ist das unmittelbare Einprogrammieren des Wissens in den Sourcecode eines entsprechenden Programms.

Dies ist jedoch nicht das, was in der Künstlichen Intelligenz mit *Maschinellem Lernen* gemeint ist. Hier wird mehr ein Verständnis von Zusammenhängen und Hintergründen (beispielsweise das Erkennen von Mustern oder Abhängigkeiten) angestrebt, um beispielsweise Muster oder Abhängigkeiten erkennen zu können. Beim Induktiven Lernen wird unter anderem versucht, aus Beispielen zu verallgemeinern und so neues Wissen zu erzeugen.

Statistik

Ohne *Statistik* ist Data Mining nicht denkbar: Seien es die statistischen Maßzahlen, die helfen, die Daten zu verstehen, oder die statistischen Verfahren zum Aufdecken von Zusammenhängen.

Nicht immer ist es möglich oder sinnvoll, ein maschinelles Data-Mining-Verfahren zu entwickeln und anzuwenden. Manchmal bringen auch schon statistische Lösungen einen ausreichenden Erfolg, falls beispielsweise ein Zusammenhang zwischen zwei Merkmalen durch eine Korrelationsanalyse gefunden wird.

Des Weiteren können statistische Verfahren dabei helfen zu erkennen, ob Data Mining überhaupt zu einem gewünschten Ergebnis führen kann.

Visualisierung

Eine gute *Visualisierung* ist für den Erfolg eines Data-Mining-Projekts unerlässlich. Da Data Mining meistens zur Entscheidungsfindung oder -unterstützung eingesetzt wird und Entscheidungen nicht immer von den Personen getroffen werden, die direkt am Prozess des Data Minings beteiligt sind, müssen die Resultate des Data Minings veranschaulicht werden. Nur wenn es gelingt, gefundenes Wissen anschaulich und nachvollziehbar darzustellen, wird man Vertrauen in die Ergebnisse erzeugen und eine Akzeptanz der Resultate erreichen.

Man kann Visualisierung aber nicht nur zur Darstellung der Resultate einsetzen, sondern auch beim eigentlichen Data Mining. Häufig erkennt man durch eine geschickte Darstellung der Daten erste Zusammenhänge zwischen den Attributen. Man denke hier an Cluster-Bildung oder an besonders einflussreiche Attribute bei einer Klassifizierung.

Visualisierung stellt somit nicht nur die entwickelten Modelle graphisch dar, sondern kann auch als eigene Data-Mining-Technik in der Datenanalyse eingesetzt werden. Da die Ergebnisdarstellung maßgeblich über den Projekterfolg entscheidet, wird der Visualisierung ein eigener Abschnitt gewidmet, siehe Abschnitt 9.6.

1.5 Erfolgreiche Beispiele

Ähnlich wie die Verwandtschaftsbeziehungen in fast jeder Einführung in die logische Programmiersprache PROLOG zu finden sind, wird in Data-Mining-Lehrbüchern sehr oft das Wetter-Golf-Spiel-Ja-Nein-Beispiel (siehe Abschnitt A.3) eingesetzt. Auch hier in diesem Buch werden Sie immer wieder auf dieses Beispiel treffen, unter anderem in den Abschnitten 1.6.3 oder 5.2. Gibt es nun auch Beispiele aus der Realität, die zeigen, dass Data Mining erfolgreich ist?

Aus den Anfangsjahren des Data Mining sind erfolgreiche Beispiele überliefert, die mittlerweile zu Klassikern wurden:

- Die amerikanische Handelskette Wall Mart soll herausgefunden haben, dass an bestimmten Tagen Windeln besonders häufig zusammen mit Bier verkauft wurden. Obwohl dieses Beispiel von vielen zitiert wird, gibt es immer wieder Diskussionen, ob dies belegt ist oder in das Reich der Legenden gehört.[1]
- Die Bonitätsprüfung oder die Prüfung der Kreditwürdigkeit von Bankkunden ist eine schon „alte“ Anwendung und wird ebenso häufig angeführt [Han98].
- Die personalisierte Mailing-Aktion als Marketing-Strategie findet sich ebenso in vielen Data-Mining-Einführungen: Aufgrund vorhandener Daten wird ein Modell für die Klassifikation in die Klassen *Anschreiben* sowie *Nichtanschreiben* entwickelt und so die Werbe-Information nur an potenzielle Kunden gesendet.

[1] Unter `www.kdnuggets.com/news/2000/n13/23i.html` wird berichtet, dass dieses Beispiel von Tom Blishok (einem Einzelhandelsberater) ca. 1992 erfunden wurde. Es soll wohl nie eine wirkliche Analyse gegeben haben.

Jede Aufzählung von Beispielen aus der realen Welt ist zum Zeitpunkt des Aufschreibens bereits veraltet. Es ist wohl besser, sich der Frage zu stellen: Wie finde ich erfolgreiche, aktuelle Beispiele? In den Lehrbüchern finden sich in der Regel die Verweise auf die gerade erwähnten Anwendungsfälle. In der Zwischenzeit ist Data Mining den Kinderschuhen entwachsen und hat Eingang in viele Anwendungsfelder – vom Finanzbereich bis zur Medizin, von der Kundenanalyse bis zum E-Learning – gefunden.

Es gibt eine Reihe von Büchern, die erfolgreiche Data-Mining-Anwendungen dokumentieren. Eine gute Auswahl an praktischen Anwendungen findet man in [Gab10]. Diese gehen von räumlichen Analysen in geographischen Systemen, über Anwendungen in der Chemie und Bioinformatik bis zu erfolgreichen Analysen in der Astronomie. In [HKMW01] wird auf eine Vielzahl von erfolgreichen Anwendungen im Marketing eingegangen. Eine typische Anwendung, die Kundensegmentierung – das Finden von Gruppen von ähnlichen Kunden – wird in diesem Buch an mehreren Beispielen vorgestellt. Die dort vorgestellten Projekte stammen aus dem Automobilbereich und der Kundenspezifischen Ansprache. Erfolgreiche Projekte aus der zweiten großen Anwendungsklasse – der Klassifikation – werden ebenso vorgestellt, beispielsweise die Bonitäts- und Kreditwürdigkeitsprüfung von Kunden. Auch Cross selling, wie es von vielen Online-Plattformen genutzt wird, ist dort mit einem erfolgreichen Projekt vertreten. Ähnliche Beispiele – aus dem Bereich E-Business und Finanzen – findet man auch in [SPM+08].

Selbst im E-Learning werden Data-Mining-Techniken auf vielfältige Art eingesetzt. In [RV06] und [RVPB11] wird eine Vielzahl von Möglichkeiten vorgestellt, wie Data Mining zur Verbesserung der Lehre eingesetzt wird. Dies geht vom Erkennen von typischen Lerner-Mustern, die eine Nutzer-angepasste Präsentation von Inhalten (sogenannte adaptive Story-Boards) ermöglicht, bis zur automatischen Erkennung von Problemen im Lernprozess.

Es ist in den letzten Jahren der Trend zu beobachten, dass über erfolgreiche Data-Mining-Anwendungen nur noch im akademischen Umfeld detailliert berichtet wird. Dies hat mehrere Gründe. Einerseits ist meistens der wissenschaftliche Neuwert nicht mehr gegeben. Andererseits sind in jeder Data-Mining-Anwendung auch viele Daten enthalten, aus denen erfolgskritisches Wissen abgeleitet wurde. Dieses geben Firmen natürlich ungern preis.

Erfolgreiche Data-Mining-Anwendungen sind für eine andere Gruppe wiederum ein gutes Marketing-Argument: Die Hersteller von Data-Mining-Software beziehungsweise die branchenübergreifenden IT-Service-Unternehmen auf dem Gebiet der Datenanalyse sind auf erfolgreiche Projekte angewiesen, um wieder neue Kunden gewinnen zu können.

Auf den Seiten der von uns in diesem Buch eingesetzten, frei verfügbaren Werkzeuge sind nur auf den Seiten der Knime-Software einige Anregungen zu finden. Diese tragen aber eher den Charakter von Einführungsbeispielen. Kommerzielle Anbieter von Data-Mining-Software werben dagegen mit ihren Referenzen: Referenz-Kunden oder Referenz-Projekte. Die Firma *Easy.Data.Mining* aus München stellt unter der Überschrift „Data-Mining-Beispiele und Fallstudien aus der Praxis“ [Eas] ihre Projekte vor.

Die Liste der Referenzkunden, die den *Rapid Miner* einsetzen, ist lang und umfasst viele Bereiche von der Elektronik-, Luft- und Automobilbranche über Handel und Marktforschungsunternehmen bis hin zu Banken und Versicherungen, der Pharma- und Biotech-

nologiebranche oder der IT-Branche selbst. Konkrete Anwendungsbeispiele werden zwar nicht aufgeführt, die aufgezählten Branchen und Unternehmen geben aber Hinweise auf den Einsatz von Data-Mining-Lösungen.

Das System *SPSS*[2] wird von IBM eingesetzt, um Data-Mining-Lösungen im Bereich des sogenannten Predictive Analytics zu entwickeln. Einige spektakuläre Anwendungsfälle sind während der Entstehung dieses Buches als Video auf YouTube zu sehen (gewesen)[3]:

- Die Datenanalyse durch die amerikanische Polizei führt dazu, dass man vorhersagen kann, wo und wann demnächst ein Verbrechen stattfinden wird. Diese Prognose erfolgt sicherlich nur mit einer gewissen Wahrscheinlichkeit[4].
- Der Zusammenhang zwischen dem Wetter und dem Keks-Verkauf in deutschen Bäckereien ist nicht nur unterhaltsam, sondern betriebswirtschaftlich relevant.

Nicht zuletzt können wir auch das Archiv des Data Mining Cups [DMC] durchsehen. Die Daten für die Aufgaben werden von Unternehmen bereitgestellt, die Aufgaben sind somit praxisrelevant.

In den Jahren 2000 und 2001 ging es um die bereits erwähnten Mailingaktionen. Lohnt es sich, einen Kunden anzuschreiben oder nicht? Beide Wettbewerbe wurden – nach Aussagen der Veranstalter – zur großen Zufriedenheit der hinter der Aufgabe stehenden Firmen abgeschlossen. Im Kapitel 10 greifen wir das Thema aus dem Jahre 2002 auf, die Mailing-Aktion eines Energieversorgers.

In mehreren Aufgaben des Data Mining Cups wird das Kundenverhalten analysiert, und es werden Vorhersagen getroffen, sei es für den Einsatz von Gutscheinen oder Rabatt-Coupons, für die Verkaufszahlen von Büchern oder das Verhalten der Kunden in einer Lotterie.

Erfolgreiche Data-Mining-Anwendungen sind in das operative Geschäft vieler Anwendungsgebiete eingebunden und werden tagtäglich genutzt. Suchen Sie selbst – beispielsweise im World Wide Web – und lassen Sie sich von interessanten Anwendungen des Data Minings überraschen.

1.6 Werkzeuge

Für die Lösung der in diesem Buch behandelten Aufgaben kann spezielle Data-Mining-Software eingesetzt werden.

Unabhängig davon sind Kenntnisse in der *Tabellenkalkulation* hilfreich. Mit einem Tabellenkalkulationsprogramm können Daten einer ersten Analyse unterzogen werden; es lassen sich Abhängigkeiten zwischen Attributen entdecken, oder die Ergebnisse eines Data-Mining-Modells können analysiert beziehungsweise nachgearbeitet werden.

[2] `www.ibm.com/de/de/`, Unterpunkt Lösungen, 03.09.2013

[3] IBM: Advertisements on YouTube:
`http://www.youtube.com/playlist?list=PLAD6EEA3C161A84F1`, 02.09.2013

[4] Siehe auch „Der Spiegel“ 2013/20.

Ein *Text-Editor*, der zudem die Arbeit mit Makros ermöglicht, ist ein weiteres nützliches Werkzeug in der Vor- sowie Nachbereitung einer Datenanalyse.

Nicht zuletzt sei darauf verwiesen, dass hin und wieder die eigene Entwicklung kleiner Programme für die Datenvorverarbeitung sowie die Analyseauswertung notwendig werden kann. Dazu kann eine beliebige Programmiersprache herangezogen werden. Einige Systeme, wie beispielsweise KNIME, ermöglichen den Einbau eigener kleiner Java-Programme (Java Snippets) in den Analyseprozess.

An dieser Stelle geben wir eine kurze Einführung in Data-Mining-Software, die für die Lösung der im Buch behandelten Aufgaben von uns eingesetzt werden:

- Der *Konstanz Information Miner* (KNIME) ist ein System zur Beschreibung ganzer Data-Mining-Prozesse, welcher eine Vielzahl Algorithmen für die verschiedenen Analyse- Phasen bereithält.
  Siehe `http://www.knime.org/`.
- Das *Waikato Environment for Knowledge Analysis* (WEKA) stellt eine Reihe von in Java implementierten Algorithmen bereit, die sowohl interaktiv als auch im Kommandozeilen-Modus ausgeführt werden können.
  Siehe `http://www.cs.waikato.ac.nz/ml/weka/`.
  Die WEKA-Algorithmen können in KNIME als KNIME-Erweiterung integriert werden.
  Siehe hierzu: `http://www.knime.com/downloads/extensions`.
- Speziell auf die Arbeit mit neuronalen Netzen ist der JAVANNS ausgerichtet. JAVANNS ist ein in Java implementiertes Nutzer-Interface für den *Stuttgarter Neuronale Netze Simulator* (SNNS).
  `http://www.ra.cs.uni-tuebingen.de/software/JavaNNS/welcome_e.html`.

Es gibt viele weitere, leistungsfähige Data-Mining-Tools, beispielsweise den *Rapid Miner* (`http://rapid-i.com/`) und den *IBM SPSS Modeler*. Ebenso sind in vielen Data-Warehouse- und Datenbank-Systemen Data-Mining-Komponenten integriert. In diesem Buch beschränken wir uns aber auf die drei genannten Systeme. Diese sind zum Erlernen und zum Experimentieren sehr gut geeignet.

1.6.1 KNIME

KNIME ist ein Data-Mining-Tool, welches ursprünglich an der Universität Konstanz entwickelt wurde. Die Abkürzung KNIME steht für *Konstanz Information Miner*. KNIME läuft unter allen Betriebssystemen, erforderlich ist JAVA.

KNIME ist in verschiedenen Versionen verfügbar. Wir verwenden im Buch den *KNIME Desktop*, der auf der KNIME-Seite heruntergeladen werden kann (`www.knime.org`).

KNIME zeichnet sich durch eine sehr einfache Drag&Drop-Bedienung sowie durch eine große Anzahl an verfügbaren Algorithmen und Methoden aus. Der Aufbau ist modular und wird ständig um neue Komponenten erweitert, die leicht intern über das Programm geladen werden können. Neben eigenen Algorithmen lässt sich die Software so um eine Vielzahl weiterer Inhalte erweitern. So sind auch nahezu alle WEKA-Verfahren für

KNIME umgesetzt worden. KNIME ist in seiner Leistungsfähigkeit durchaus vergleichbar mit einer Vielzahl von kommerziellen Data-Mining-Programmen. Der komplette Mining-Prozess vom Datenimport über Datenvorverarbeitung und Datenanalyse bis hin zur Darstellung der Ergebnisse lässt sich mit den bereitgestellten Methoden bewerkstelligen.

Für einige Beispiele ist die Integration der WEKA-Verfahren in das KNIME-System erforderlich. WEKA-Verfahren können als KNIME-Extension eingebunden werden.

Nach dem Start von KNIME öffnet sich das in Abbildung 1.4 dargestellte Fenster.

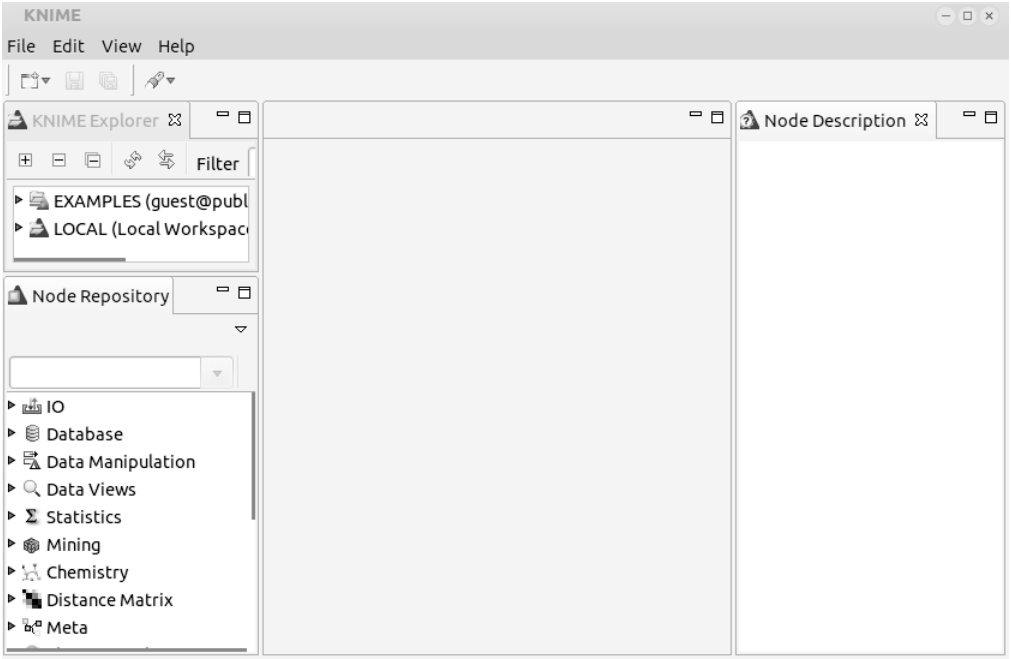

Abb. 1.4: KNIME – *Start-Fenster*

Die Oberfläche enthält folgende Komponenten:

1. Workflow-Fenster
   Das mittlere Fenster ist das Hauptfenster, in dem der Data-Mining-Prozess abgebildet wird.

2. Projektfenster
   Das Projektfenster *KNIME Explorer* dient der Verwaltung von Projekten.

3. Node Repository
   Fundament jeglicher KNIME-Anwendung ist die Sammlung der implementierten Algorithmen, aus denen der Nutzer einen Workflow zusammensetzt. Der Fundus an Algorithmen umfasst Komponenten

- für den Datenimport und -export aus Dateien oder Datenbanken,
- zur Datenvorverarbeitung und Datenmanipulation,
- für grundlegende statistische Analysen,
- für die eigentliche Datenanalyse (sowohl KNIME als auch WEKA) und
- zur Visualisierung.

4. Node Description
   In diesem Fenster erhält man eine detaillierte Beschreibung der KNIME-Knoten:
   - Beschreibung des Algorithmus
   - Benötigte Inputformatierungen der Daten
   - Art des erzeugten Outputs
   - Kurze Beschreibung der Konfigurationsmöglichkeiten

Nun kann man sich die entsprechenden Werkzeuge als Nodes (Knoten) aus dem Repository in das Hauptfenster ziehen und dort durch Pfeile verbinden. In Abbildung 1.5 auf der nächsten Seite ist ein Workflow dargestellt, der aus folgenden Komponenten besteht.

1. Filereader: Hier werden die Daten eingelesen.
2. Partitioning: Die Daten werden in Trainings- und Testmenge aufgeteilt (siehe Abschnitt 9.4).
3. Decision Tree Learner: Aus den Trainingsdaten wird ein Entscheidungsbaum erzeugt.
4. Decision Tree Predictor: Der erlernte Entscheidungsbaum wird auf die Testmenge angewendet.

Das Hauptfenster dient folgenden Aufgaben:

- Es ist das primäre Modellierungsfenster des Miningprozesses.
- Die Knoten können verwaltet und konfiguriert werden (Rechtsklick auf den jeweiligen Knoten).
- Die Knoten können miteinander verknüpft und so zu einem Workflow eines Data-Mining-Prozesses zusammengesetzt werden. Dabei signalisieren die Pfeile den Datenfluss.

Ist ein Workflow zusammengesetzt, muss man die Knoten konfigurieren. Dazu geht man auf den jeweiligen Knoten und aktiviert durch einen rechten Mausklick (oder durch einen Doppelklick mit der linken Maustaste) die entsprechende Auswahl. Beim *Filereader* kann man so das einzulesende File auswählen.

Wenn ein Knoten korrekt konfiguriert ist, wechselt die Ampel unter dem Knoten von rot auf gelb. Nun führt man den Knoten aus, wieder über den rechten Mausklick. Ist

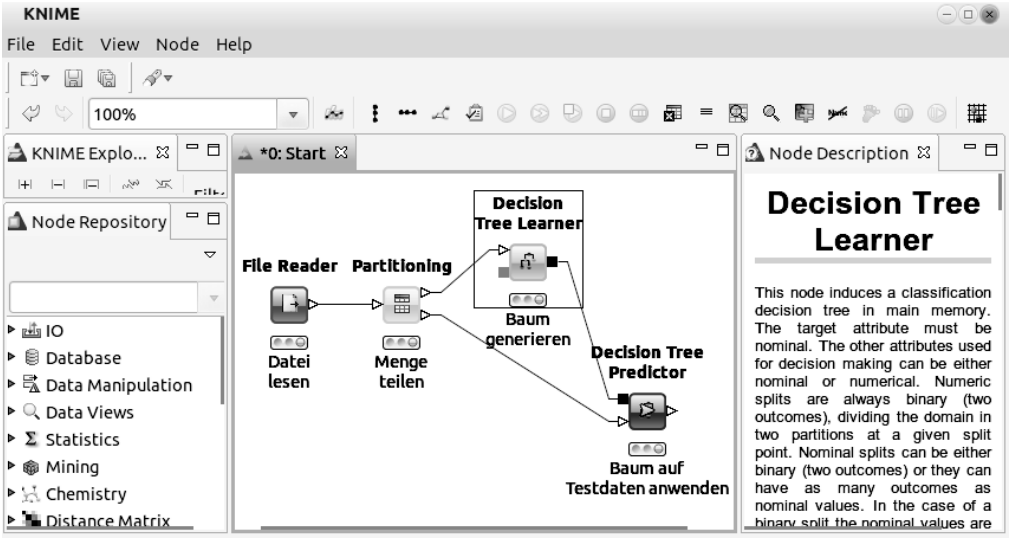

Abb. 1.5: KNIME – *Workflow*

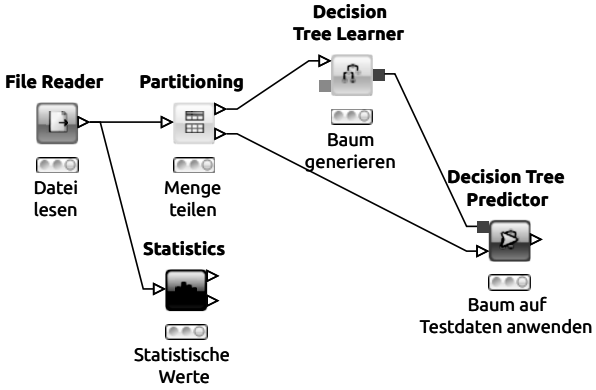

Abb. 1.6: KNIME – *Wetterbeispiel*

alles korrekt verlaufen, wird die Ampel grün. Dies setzt man bei den Folgeknoten fort. Zeigt sich unter dem Knoten ein Warndreieck, sollte man die hinter diesem Symbol platzierte Fehler- oder Warnmeldung lesen und darauf reagieren. Abbildung 1.6 zeigt den fertigen Workflow für das Wetter-Beispiel (siehe Abschnitt A.3).

Der in den Workflow integrierte Statistik-Knoten liefert die bereits angesprochenen statistischen Maßzahlen und fördert das Verständnis für die zu analysierenden Daten.

Betrachten wir nochmal die wichtigen Schritte im Data-Mining-Workflow.

Zunächst muss die Datei eingelesen werden. Um eine Datei einzulesen, muss in der Knotenauswahl unter *IO* im Abschnitt *Read* der *Filereader* ausgewählt und auf die