Kapitel 1

Einleitung

“We are drowning in information but starving for knowledge.” (John Naisbett)

“Who[ever] has information fastest and uses it wins.” (Don Keough, früherer Präsident von Coca-Cola)

“Datamining will become more important, and companies will throw away nothing about their custumers because it will so be so valuable. If you’re not doing this you’re out of business.”

(Arno Penzias, Nobelpreisträger 1999)

Wir leben heute in einer Zeit des ungeheuren Wachsums digitaler Information. Schätzungen besagen, dass sich die verfügbare digitale Information alle 20 Monate und die schiere Speicherkapazität alle 9 Monate verdoppeln (Fayyad und Uthurusamy 2002). Einer der Gründe ist die exponentiell wachsende Verfügbarkeit von Rechen- und Speicherkapazität zu dramatisch fallenden Preisen – eine stetige Entwicklung, die schon von Moore (1965) quantifiziert wurde.

Die Verfügbarkeit von Information und die schnelle Umsetzung durch Extraktion von Wissen sind Schlüsselfaktoren in Entscheidungsprozessen – dies wurde von vielen auch als wettbewerbsrelevant erkannt.

Der Begriff *Datamining* ist in Analogie zum Bergbau, engl. *mining*, eingeführt worden. Die Kunst des Bergbaus besteht darin, wertvolle Erze und Mineralien in einer großen Menge von so genanntem „taubem“ Gestein

aufzuspüren, zu identifizieren, herauszuarbeiten und zu bergen. Dazu gehört, im Gestein zu navigieren, den reichen Flözen und Gängen zu folgen und das Gestein aufzuschießen, zu separieren, um letztlich die „Goldklumpen“ (*nuggets*) zu heben. In verschiedenen Bereichen treffen wir in unserer Zeit auf eine solche Situation an: riesige Datenhalden, die heute, dank winziger Speicherkosten, teilweise nie mehr gelöscht werden.

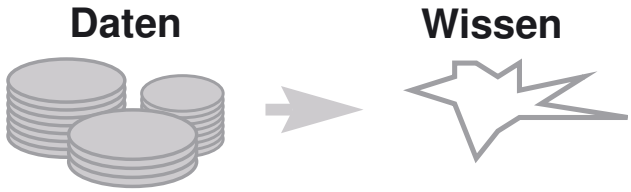

Abbildung 1.1: Die übergeordnete Aufgabe des *knowledge-discovery*-Prozesses ist: Wie gewinnt man interessantes Wissen aus einer möglicherweise erdrückenden Fülle von Daten?

So, wie die Aufgabe des Bergbaus darin besteht, „Gold“ aus dem großen Steingemenge zu bergen, so ist die Aufgabe des Dataminings, wertvolles Wissen in großen Datenmengen zu entdecken. Diese Aufgabe wird im Begriff *knowlegde discovery in databases* (**KDD**) noch stärker zum Ausdruck gebracht und zudem wird Bezug auf die typische Speicherform in Datenbanken genommen.

Nach der Definition von Fayyad et al. (1996) beschreibt ***Knowlegde discovery*** den *(i)* nicht-trivialen Prozess der Identifikation von *(ii)* validen, *(iii)* neuen, *(iv)* potentiell nützlichen und letztlich *(v)* verständlichen bzw. umsetzbaren *(vi)* Mustern und Regularitäten aus *(vii)* Datenbeständen.

Im Folgenden wird diese prägnante Beschreibung an verschiedenen Stellen weiter erläutert.

Die Nichttrivialität *(i)* bedeutet, dass *knowlegde discovery* in komplexen Domänen kein vollkommen automatisierbarer Vorgang ist. Neben teilautomatisierten Verfahren schließt er das Vorwissen und das Interpretationsvermögen von Experten der Anwendungsdomäne unverzichtbar mit ein. Daher ist die effiziente Integration des Menschen in den iterativen Prozess der Wissensgewinnung eine Schlüsselherausforderung. Dies impliziert Aspekte der Gestaltung von Benutzerschnittstellen (*human computer interface*, HCI) als auch der Einbindung der Experten in den KDD-Prozess.

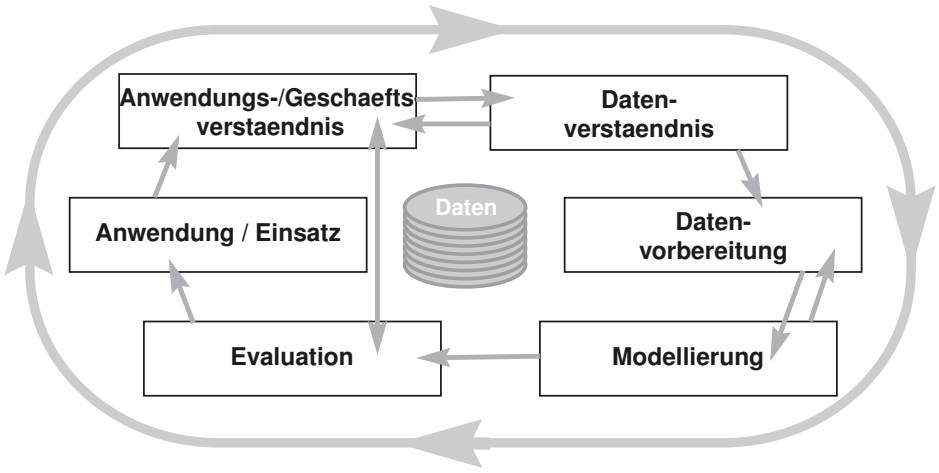

Abbildung 1.2: Der Datamining-Prozess nach dem CRISP-DM-Standard (*CRoss Industry Standard Process for Datamining*) durchläuft mehrere Teilschritte, wobei die Schrittfolge nicht als linear aufeinander folgend, sondern iterativ und situationsgerecht angepasst vorgesehen ist.

Vorgehensmodell CRISP-DM

Um den KDD-Prozess als solchen besser zu planen und zu standardisieren, wurden einige Vorgehensmodelle entwickelt, von denen eines der verbreitetsten hier kurz dargestellt werden soll. Der *CRoss Industry Standard Process for Datamining* (**CRISP-DM**) wurde von einem internationalen Industriekonsortium (NCR, Daimler Chrysler, SPSS, OHRA, Chapman et al. 2000) entwickelt. Wie in Abb. 1.2 illustriert, beschreibt CRISP-DM sechs Kernbereiche eines Datamining-Projekts, deren zeitliche Abfolge nicht in sechs streng sukzessiven Phasen vorgeschlagen wird, sondern im Gegenteil, es wird ein iterativer und stark vernetzter Ablauf angelegt:

- Im ersten Schritt eines Projekts gilt es, zunächst ein Verständnis für die relevanten Vorgänge im Geschäfts- bzw. Anwendungsbereich zu gewinnen (*business understanding*). Da das KDD-Ziel ist, *neues* Wissen (*iii*) zu entdecken, ist es bedeutsam eine gute Orientierung darüber zu haben, welche Bereiche bekannt und welche potentiell relevant sind. Das Ergebnis soll eine klare Festlegung der Anforderungen, Projektziele (aus Geschäfts- bzw. Anwendungssicht) und Rahmenbedingungen beinhalten.

  Das Fokussieren dieser Ziele und das Stellen relevanter Fragen sind die wichtigsten Planungsschritte im Datamining-Prozess. Gute Fragen sind solche, deren Antwort irgendwo in den Daten verborgen

liegt. Ihre Formulierung erfordert oft die Kooperation von verschiedenen Partnern: Man muss die Geschäfts- oder Wissensziele verstehen, die verfügbaren Daten interpretieren und die Datamining-Verfahren der nachfolgenden Phasen kennen. Mitarbeiter, die in allen drei Bereichen kompetent sind, gibt es zwar ganz selten, aber sie sind äußerst wertvoll;

- der zweite Bereich fokussiert auf die vorhandenen bzw. noch zu beschaffenden Datenbestände. Dies beinhaltet die Sichtung, Dokumentation und überblicksartige Exploration der Daten. Die starke Interdependenz mit der Zielplanung wird durch die Pfeile in beide Richtungen ausgedrückt;
- die Datenvorbereitung schafft integrative Zugänge zu den Daten durch Zusammenführung der (möglicherweise weit verstreuten) Datensätze und deren Aufbereitung in geeignete Datenformate. Dies involviert auch die Beseitigung von Inkonsistenzen, die Behandlung fehlender Werte und die Erweiterung um abgeleitete, später hilfreiche Größen;
- die Modellbildung umfasst die Auswahl der verwendeten Algorithmen und Bewertungsverfahren, die meist iterative Gewinnung der Modelle und deren technische Bewertung;
- bei der Evaluierung steht, im Gegensatz zu vorher, die Beurteilung des gewonnenen Modells aus der Anwendungsperspektive im Vordergrund. Hierzu werden die gefundenen Muster adäquat präsentiert und vom Experten entsprechend der definierten Projektziele bewertet. Die Zwischenschritte werden überprüft und die nächsten Schritte geplant;
- wenn die Evaluierung erfolgreich ist, werden die Ergebnisse dokumentiert und ggf. ein Umsetzungsplan für das gefundene Wissen erstellt und ausgeführt. Dies kann zum Beispiel die Durchführung einer Werbekampagne (s.u.) sein oder die Integration einer Risikobewertung in ein klinisches Informationssystem (s. Kap. 9).

Anwendungsbeispiele für Datamining

Ein typisches und kommerziell sehr erfolgreiches Einsatzfeld für Datamining Verfahren ist der Bereich Marketing und Kundenbindung – (neudeutsch:) *Customer Relationship Management* (**CRM**).

Kampagnenplanung: Zur Optimierung der Marketingaktivität werden Kundendaten analysiert, um z.B. herauszufinden, welche typischen Kunden profitabel sind oder werden könnten. Aus der Schätzung der Empfänglichkeit für eine bestimmte Werbemaßnahme wird genau jenes Kundensegment bestimmt, in das in einer Marketingaktion investiert wird, oder umgekehrt, nicht investiert wird.

Churn management ist eine dazu verwandte Anwendung mit umgekehrtem Vorzeichen: Für eine Telefongesellschaft gilt es z.B., der Kündigung eines Dienstleistungsvertrages gezielt entgegenwirken zu können. Indem man die mögliche Unzufriedenheit des Kunden abzuschätzen versucht, kann die Kundenbindung durch konkrete Problembehebung oder gezielte Freundlichkeiten (Anruf, Brief, spezielle Angebote etc.) wieder verbessert werden. Die Datenbasis kann sich hier auf die Vertragsdetails und die gesamte Beziehungshistorie mit den Kunden ausweiten.

Webshops: Die Entwicklung des Internets hat die vollautomatische Rund-um-die-Uhr Onlinevermarktung von Produkten und Services ermöglicht. Man mag annehmen, dass die Vermarktungskosten deutlich geringer sind als im klassischen Geschäft. Dies ist durchaus nicht immer so, aber es ergeben sich durch Datamining-Techniken ganz neue Möglichkeiten, Kunden durch Zusatzservices zu binden. Man unterscheidet zwei grundsätzliche Arten von Käuferverhalten. Der *Sucher* weiß, was er will – er kann online durch gute Datenbanksuchoptionen und -strategien hocheffizient bedient werden. Der *Spontankäufer* dagegen schlendert und wartet, bis er etwas Reizvolles findet, das bei ihm den nötigen Kaufimpuls auslöst. Ein reales Ladengeschäft dekoriert sein Sortiment und präsentiert es dem umherschweifenden Blick des Kunden. Ein Online-Webshop kann dies nicht annähernd in der Blickweite nachbilden, aber er kann versuchen, gleich *etwas Passendes* zu präsentieren.

Mit Datamining-Verfahren kann die Beratungsleistung eines persönlich vertrauten Kundenbetreuers nachgebildet werden. Aufgrund der Erfahrung bei vergangenen Besuchen werden individuelle Kaufempfehlungen unterbreitet, verbunden mit einem breiten Angebot von Zusatzinformation und verzugsloser, freundlicher Bedienung. Dies erfordert nicht nur eine geschickte Dialoggestaltung, in der die relevanten Daten unaufdringlich erhoben werden, sondern auch die Kombination mit evtl. demographischen Daten (z.B. straßenbezogene Schätzungen von sozio-ökonomischen Daten, Kaufkraft und Präferenzen, natürlich unter Wahrung der Datenschutzregelungen) und der Einbeziehung von aktuellen Kauftrends, die in der