

Data Mining – Methoden und Anwendungen

Johann Petrak

Österreichisches Forschungsinstitut für AI

Schotteng. 3, A-1010 Wien

E-mail: johann@ai.univie.ac.at

Technical Report TR-97-15¹

Data Mining oder *Knowledge Discovery in Databases* ist ein Forschungsgebiet, das in den letzten Jahren aufgrund seines großen Anwendungspotentials auf großes Interesse gestoßen ist. Dieser Report führt in das Gebiet ein und gibt einen Überblick über Motivation, Methoden und aktuelle Forschungsrichtungen und präsentiert einige typische Anwendungen von Data Mining.

¹Dieser Report ist eine gemeinsame Publikation von ÖFAI und CD-Labor und ist auch als Report CD-TR 96/107 des CD-Labors erhältlich.

Weitere Informationen zum Thema Data Mining sind verfügbar unter <http://www.ai.univie.ac.at/oefai/ml/kdd/>

Inhaltsverzeichnis

1	Einleitung	1
2	Wozu Data Mining?	1
2.1	Daten	1
2.2	Data Warehousing	3
3	Was ist Data Mining?	4
3.1	Data Mining – ein Prozess	4
3.2	Mustergewinnung aus Daten	8
3.3	Methoden und Techniken	10
4	Arten von Mustern	10
4.1	Klassifikationsregeln	11
4.2	Charakteristische Regeln	12
4.3	Regressionsregeln	13
4.4	Cluster	13
4.5	Abhängigkeitsmuster	13
4.5.1	Assoziationsregeln	13
4.5.2	Determinationen und funktionale Abhängigkeiten	14
4.6	Verbindungsmuster (Links)	14
4.7	Zeitliche Muster	14
4.8	Abweichungen	15
4.9	Formeln und Mathematische Gesetzmäßigkeiten	15
5	Exemplarische Anwendungen	15
5.1	SKICAT: Katalogisierung von Himmelsobjekten	15
5.2	SCOUT: Analyse von Basketballspielen	16
5.3	TASA: Fehleranalyse in Telekommunikationsnetzen	17
5.4	FAIS: Aufdeckung von Geldwäsche	17
5.5	Evaluierung von persönlichen Darlehen	18
5.6	Qualitätsanalyse von Flußwasser	18
5.7	Analyse von Daten aus dem Gesundheitswesen	18
6	Data Mining am ÖFAI	19

1 Einleitung

In den letzten Jahren ist das Forschungsgebiet *Data Mining*, das im Englischen oft auch synonym mit *Knowledge Discovery in Databases* bezeichnet wird, auf großes Interesse sowohl bei Forschern als auch in der Industrie gestoßen. Dieses Interesse hat mehrere Ursachen: zum einen werden immer mehr Anwendungsbereiche für Computer erschlossen, und damit fallen immer größere Datenmengen an, die mit herkömmlichen Methoden kaum mehr analysiert und interpretiert werden können. Zum anderen stehen nunmehr seit einigen Jahren Techniken aus dem Bereich des Maschinellen Lernens zur Verfügung und warten auf ihre praktische Anwendung.

Wie es für ein so junges Gebiet wie diesem nicht überraschend ist, herrscht unter den Wissenschaftlern und Entwicklern keineswegs Einigkeit darüber, welche Techniken und Wissensbereiche ihm überhaupt zuzuordnen sind. Auch die verwendeten Begriffe werden oft uneinheitlich oder ungenau verwendet. Das hängt auch damit zusammen, dass sich Forscher aus unterschiedlichen Wissenszweigen für das Gebiet interessieren und bei ihren Arbeiten Ergebnisse, die für das jeweilige persönliche Interessensgebiet wichtig sind, in den Vordergrund rücken.

Eine zusätzliche Schwierigkeit entsteht bei der Verwendung von deutschen Begriffen. Für viele Begriffe gibt es keine, oder nur recht holprige deutsche Übersetzungen. In diesem Report werden daher die meisten Begriffe unübersetzt gelassen, oder deutsche Übersetzungen nur in Klammern angegeben. Der Großteil der relevanten Literatur ist ohnehin praktisch ausschließlich in Englisch abgefasst.

Dieser Report gibt einen Überblick über das heterogene Gebiet Data Mining. Der Report ist folgendermaßen gegliedert: In Kapitel 2 werden die Ausgangslage und die Motivation für Data Mining-Techniken dargestellt. Kapitel 3 erläutert die Ziele des Data Mining, erklärt grundlegende Begriffe und definiert Data Mining als komplexen Prozess der die Einbindung menschlicher Experten genauso erfordert wie die Anwendung unterschiedlicher Methoden aus verschiedenen Forschungsbereichen. Kapitel 4 gibt anhand der unterschiedlichen Arten von Mustern einen Überblick über die wichtigsten im Data Mining Prozess angewandten Methoden. Schließlich werden in Kapitel 5 einige ausgewählte Beispiele für Data Mining-Anwendungen vorgestellt.

2 Wozu Data Mining?

2.1 Daten

Die Anzahl und Größe der weltweit routinemäßig anfallenden Datensammlungen und Datenbanken nimmt ständig zu. Man schätzt, dass sich die weltweit vorhandene Informationsmenge alle 20 Monate verdoppelt – bei Datenbanken ist die Rate wahrscheinlich noch höher (Frawley, Piatetsky-Shapiro, & Matheus 1991).

Viele dieser Datenbanken speichern auch enorme Datenmengen, mit Millionen oder gar Milliarden von Datensätzen. Die Daten werden ursprünglich meist für andere Zwecke als die Verwendung in Data Mining-Systemen gesammelt und routinemäßig archiviert.

Oft genannte Beispiele für solche Datenbanken sind:

Supermarkt-Scannerdaten: In den meisten Supermärkten werden die verkauften Waren mittlerweile an einer mit einem optischen Barcode-Scanner ausgerüsteten Kassa erfasst. Die Kassa erstellt nicht nur den Kassenzettel für den Kunden, sondern speichert auch sämtliche Verkaufsdaten in einer Datenbank. Aufgrund dieser Datenbank ist stets feststellbar, welche Waren verkauft wurden, welche Waren daher für das Lager nachgeliefert werden müssen und andere Informationen mehr. Zusätzlich könnten die Daten aber Informationen enthalten, für deren Auswertung sie ursprünglich gar nicht gesammelt wurden: Man kann den Effekt von Werbekampagnen oder bestimmten Präsentationsmethoden auf das Konsumverhalten feststellen, oder Kaufgewohnheiten der Käufer analysieren und das Angebot darauf abstimmen. Tatsächlich ist diese „Warenkorbanalyse“ ein wichtiges Anwendungsgebiet des Data Mining. Ein Beispiel für eine derartige Datenbank ist die der amerikanischen Firma Wal-Mart, die täglich mehr als 20 Millionen Transaktionen enthält (Babcock 1994).

Prozessdaten: Aus Echtzeitsystemen fallen oft routinemäßig große Mengen von Prozessdaten an. Daten aus der industriellen Fertigung könnten z.B. verwendet werden, um Regelmäßigkeiten bei der Produktion von Ausschuss zu finden, oder die Daten aus der Prozessüberwachung in Kraftwerken könnten wertvolle Informationen über Muster bei Störfällen enthalten.

Wissenschaftliche Daten: Auch im Bereich der wissenschaftlichen Forschung fallen immer größere Datenmengen an. Im *Human Genome Database Project* (HGP) wurden bereits Gigabytes von Daten über den menschlichen genetischen Code gesammelt, und in Zukunft wird noch weitaus mehr dazukommen (Fasman, Cuticchia, & Kingsbury 1994). Das für Ende der 90er Jahre geplante *Earth Observing System* (EOS) der NASA, bestehend aus einer Anzahl von Satelliten und Messinstrumenten, wird voraussichtlich 50 Gigabyte Fernerkundungsdaten pro Stunde liefern (Way & Smith 1991).

Es ließen sich noch eine Vielzahl weiterer Beispiele angeben (Fayyad, Piatetsky-Shapiro, & Smyth 1995). Allen ist gemeinsam, dass die Auswertung der Datenbanken mit herkömmlichen Methoden (manuelle Abfragen, klassische statistische Methoden u.a.) nicht machbar ist, oder nicht die gewünschten Ergebnisse liefert.

2.2 Data Warehousing

Insbesondere bei kommerziellen Anwendungen hat sich mit der zunehmenden Menge anfallender Daten oft die immer größere Diskrepanz zwischen der Menge der *Daten* und der Menge der *interessanten* und *nutzbaren* Informationen gezeigt.

Die für den täglichen Betrieb vorgesehenen *On Line Transactional Processing* (OLTP)-Datenbanken sind für komplexe ad-hoc Abfragen und Zusammenfassungen unterschiedlicher Datenbestände nicht geeignet. Diese Aufgaben werden mit Hilfe von *Data Warehousing*-Technologien bewältigt.

Dabei werden die Daten aus den unterschiedlichen externen Quellen in einer zentrale Datenbank, dem *Data Warehouse* gesammelt und verwaltet. Die wichtigsten Aktivitäten sind dabei:

- Daten aus den verschiedenen Datenbanken für den täglichen Betrieb werden regelmäßig in das WH übernommen. Dabei werden die Daten restrukturiert und umkodiert und in eine abfrageorientierte Form gebracht.
- Durch Selektion, Projektion, Gruppierung usw. werden Überblicksdaten gebildet.
- Relevante Detail- und Summendaten werden in die benötigte Form – Charts, Spreadsheets, statistische Analysen, textuelle Reports etc. – gebracht.
- Alte Daten werden auf Massenspeicher archiviert und im WH nur mehr in zusammengefasster Form behalten.
- Die Anwender verwenden das WH, um komplexe Abfragen abzusetzen oder holen sich standardisierte Zusammenfassungen, Reports, oder Analysen.

Die Aufgaben, die eine WH-Datenbank zu erfüllen hat, unterscheiden sich also grundlegend von den Aufgaben der herkömmlichen Datenbanken:

	Data Warehouse	Operational Database
Zweck:	Analyse	täglicher Betrieb
Abfragen:	ad-hoc, über einen großen Teil der Datenbank	standardisiert, nur eine beschränkte Anzahl Datensätze sind involviert
Inhalt:	langlebig, statisch	kurzlebig, laufende Updates
Updates:	regelmäßige Massendates	Realtime
Verarbeitung:	datenorientiert	ereignisorientiert

Die auf die Verwaltung und Analyse von WH zugeschnittenen Datenbanksysteme werden oft *On Line Analytical Processing* (OLAP)-Systeme genannt, um sie von den herkömmlichen OLTP Systemen zu unterscheiden.

OLAP-Systeme erlauben es, die im Data Warehousing benötigten komplexen Datenbankabfragen einfach und effizient auszuführen. Damit ist es prinzipiell auch möglich, durch interaktive, explorative Abfragen komplexe, interessante und bisher unbekannte Zusammenhänge zu entdecken. Die Zielsetzung von *Data Mining Systemen* ist es, solche Zusammenhänge eigenständig zu finden, indem automatisch die richtigen Abfragen an das OLAP System abgesetzt werden. Allerdings existiert in der Praxis bei vielen Data Mining Anwendungen keine direkte Anbindung der Data Mining Algorithmen an ein Data Warehouse oder OLAP-System. Vielmehr müssen die Daten oft in einer Reihe von Vorbereitungsschritten zum Teil direkt aus archivierten OLTP-Datenbanken für die jeweiligen Data Mining-Algorithmen konvertiert werden (s. Kap. 3.1).

3 Was ist Data Mining?

Data Mining bezeichnet Techniken zum Finden von interessanten und nützlichen Mustern und Regeln („Wissen“) in großen Datenbanken. Der Begriff wird auch für die Anwendung dieser Techniken und für die Forschungsrichtung, die sich mit der Entwicklung dieser Techniken beschäftigt, verwendet. Eine allgemein verwendete exakte Definition des Begriffes gibt es nicht, in der in das Gebiet Data Mining einführenden Literatur finden sich oft recht unterschiedliche Definitionen.

Wir werden daher nicht versuchen „die richtige“ Definition des Begriffs anzugeben, sondern wesentliche Eigenschaften und Eigenheiten beschreiben.

3.1 Data Mining ist ein Prozess ...

Data Mining bezeichnet nicht eine einzelne Technik, sondern umfasst den gesamten Prozess von der Bereitstellung der Daten bis zur Anwendung der Erkenntnisse. Dies wird in den meisten einführenden Arbeiten immer wieder betont (Fayyad, Piatetsky-Shapiro, & Smyth 1995), allerdings wird in den Fachpublikationen nicht auf alle Schritte des Gesamtprozesses mit gleicher Genauigkeit eingegangen.

Die konkreten Einzelschritte hängen naturgemäß von der konkreten Situation ab. In den meisten Fällen lassen sich aber vier Phasen im Data Mining-Prozess erkennen, denen grob die folgenden Schritte zugeordnet werden können:

Planungsphase: Der wichtigste Schritt in dieser Phase ist das Festlegen der konkreten Aufgabenstellung: welche Ziele sollen durch das Data Mining erreicht werden? Welche Ergebnisse werden erwartet? Darüber hinaus ist es wichtig, dass die richtigen

Personen in den Data Mining Prozess eingebunden sind: Ohne Beteiligung von Experten des jeweiligen Sachgebiets ist ein sinnvolles Data Mining nur in den einfachsten Fällen möglich.

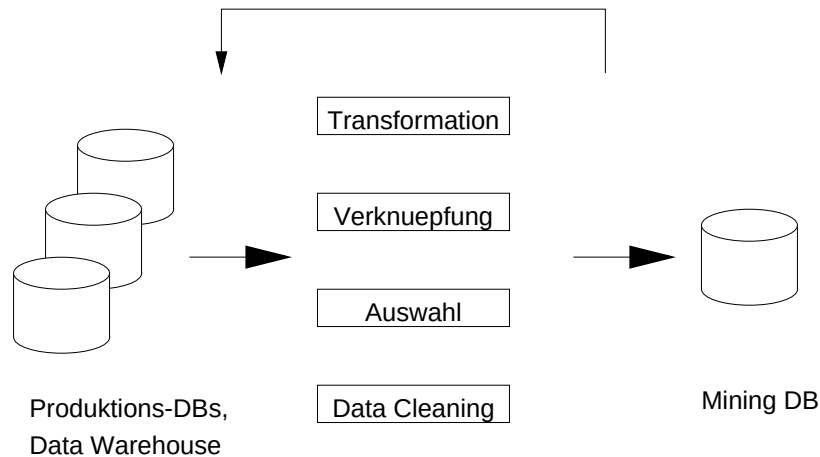


Abbildung 1: Vorbereitungsphase

Vorbereitungsphase: In dieser Phase (Abb. 1) werden die zu verwendenden Daten und Informationen vorbereitet. Ergebnis dieser Phase ist eine *Mining Base*, die die Ausgangsbasis für die weiteren Transformations- und Analyseschritte darstellt.

- **Datensammlung und Datenauswahl:** Es muss entschieden werden, welche Informationen für den Prozess verwendet werden sollen und in welcher Form diese Daten zur Verfügung stehen. Aus der großen Anzahl vorhandener Datenbanken, Tabellen, Datensätze und Attribute werden nur jene ausgewählt, die inhaltlich relevant erscheinen. Unter Umständen müssen einzelne Daten zusätzlich nach-erhoben werden. Das betrifft vor allem
 - Meta-Daten: Daten, die die verwendeten Daten näher beschreiben
 - Hintergrundwissen (optional): Regeln, die bekannte Zusammenhänge in den Daten beschreiben und dem Data Mining Algorithmus in maschinell lesbarer Form zur Verfügung stehen sollen.
- **Datenintegration:** Oft ist es sinnvoll, Datenbestände aus unterschiedlichen Quellen, Datenbanksystemen oder Anwendungen zusammenzuführen. Dabei sind z.B. Kompatibilitätsprobleme (unterschiedliche Dateiformate und Kodierungsarten) zu lösen oder Methoden zu finden, Datensätze aus verschiedenen Datenbanken, die sich auf den gleichen Fall beziehen, einander zuzuordnen.
- **Data Cleaning:** Entfernen oder Berichtigen offensichtlich falscher oder widersprüchlicher Daten.

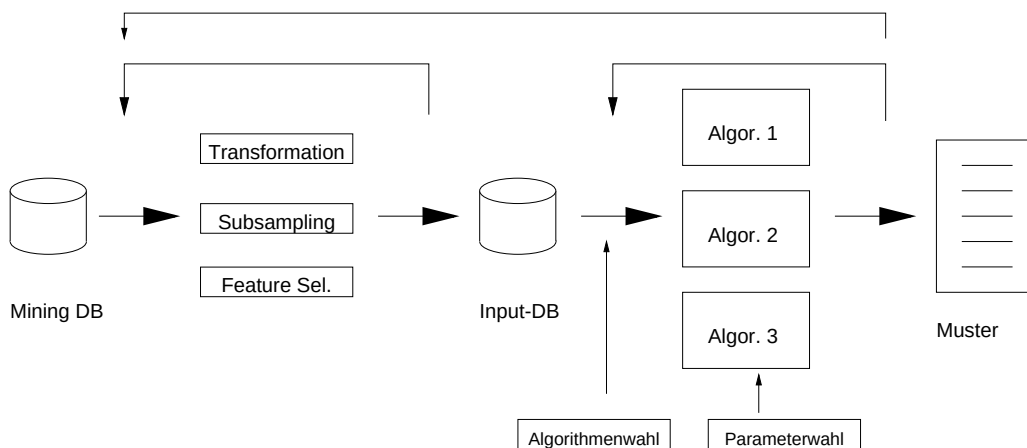


Abbildung 2: Miningphase

Mining Phase: In dieser Phase (Abb. 2) findet die eigentliche Suche nach interessanten Mustern statt. Die Erfahrungen, die im Laufe dieser Phase gewonnen werden, können es notwendig machen, Schritte aus der Vorbereitungsphase zu wiederholen, um z.B. zusätzliche Daten in die Mining Base zu inkludieren.

- *Subsampling:* Nur ein Teil der Datensätze der Mining Base wird verwendet. Das kann aus verschiedenen Gründen sinnvoll sein (Performanz, Analyse der Redundanz in den Daten etc.)
- *Feature selection:* Dabei wird versucht, möglichst nur die zum Finden der Muster relevanten Attribute aus der Mining Base auszuwählen. Dies kann händisch versucht werden, oder automatisch durch einen iterativen Suchprozess erfolgen.
- Auswahl der Analysesart: Festlegen, welche Art von Mustern gefunden werden sollen und wie diese Muster dargestellt werden sollen. Die verschiedenen Musterarten sind in Kap. 4 genauer beschrieben.
- Auswahl der Algorithmen: Entscheiden, wie gut die verschiedenen zur Verfügung stehenden Algorithmen zur Lösung des Problems geeignet sind.
- Transformation: In diesem Schritt wird die Kodierung der einzelnen Attribute verändert. Die Notwendigkeit für Transformationen kann sich entweder inhaltlich aus den Ergebnissen früherer Analyseschritte, oder aufgrund der Anforderungen des verwendeten Analyseprogramms ergeben. Die wichtigsten Arten von Transformationen sind:
 - Umkodierung: dabei wird nur die *Repräsentation* der Werte geändert, also z.B. symbolische Werte durch fortlaufende Nummern dargestellt.
 - Normalisierung: dient dazu, den Wertebereich eines Attributs auf einen Standardbereich abzubilden. Eine häufig verwendete Art der Normalisierung ist die Abbildung numerischer Werte auf den Bereich $[0 \dots 1]$.

- Diskretisierung: Dabei werden numerische Werte durch eine geringe Anzahl symbolisch repräsentierte Intervalle ersetzt. Das ist insbesondere dann notwendig, wenn der Data Mining-Algorithmus keine numerischen Werte verarbeiten kann.
- Auswahl der Parameter: viele Algorithmen erlauben die Angabe von Parametern, die das Verhalten des Algorithmus beeinflussen. Je nach Art der Daten und vorhergehender Analyseschritte müssen die Parameter geeignet gewählt werden, oder unter Umständen in einem eigenen iterativen Prozess gefunden werden.
- Analyse: Anwendung der Algorithmen auf die vorbereiteten Daten. Dieser Schritt ist der eigentliche *Kernschritt* des Data Mining Prozesses und wird daher manchmal als Data Mining im engeren Sinn bezeichnet² In diesem Schritt werden *interessante Muster* aus den Daten gewonnen (s. Kap 4).

Ergebnis dieser Phase ist meist eine große Anzahl komplexer Musterbeschreibungen, wie sie von den einzelnen Data Mining-Algorithmen erstellt werden.

Auswertungsphase: In dieser Phase wird versucht, die Ergebnisse aus der Mining Phase so aufzubereiten, dass sie auch für Nichtexperten verständlich sind.

- Interpretation: Dieser Schritt ist mindestens ebenso wichtig, wie das Finden der Muster im Analyseschritt. Es wird untersucht, ob und welche der gefundenen Muster interessant im Sinne der Aufgabenstellung sind und die statistische Signifikanz der gefundenen Muster abgeschätzt. Die *Visualisierung* der oft unübersichtlichen Ergebnisse des Analyseschritts spielt dabei eine wichtige Rolle. Wurden nicht die erwarteten Muster gefunden, muss die Ursache analysiert werden und zu einem früheren Schritt im Data Mining-Prozess zurückgekehrt werden (Verwendung anderer/anders kodierter Daten, anderer Algorithmen/Parameter etc.)
- Zusammenstellen der Ergebnisse in einer für den Anwender verständlichen Form
- Anwenden des gewonnenen Wissens: Insbesondere in kommerziellen Anwendungen wird das aus dem Data Mining Prozess gewonnenen Wissen oft zur Unterstützung eines Entscheidungsprozesses verwendet, beim Data-Mining Prozess handelt es sich daher um einen *Decision Support Process*.

²(Fayyad, Piatetsky-Shapiro, & Smyth 1995) nennen den Gesamtprozess *Knowledge Discovery in Data Bases (KDD)* und verstehen unter Data Mining nur die eigentliche Anwendung der Analysemethoden. Wir werden hier unter Data Mining jedoch weiterhin den Gesamtprozess verstehen.

Je nach Aufgabenstellung können die angegebenen Schritte mehrmals, auch in unterschiedlicher Reihenfolge durchlaufen werden. Data Mining ist also ein *iterativer* Prozess.

Die tatsächliche Planung und Ausführung des Prozesses ist kaum automatisierbar: die Entscheidungen, welche Schritte aufgrund der Ergebnisse früherer Aktivitäten durchzuführen sind, erfordern Wissen und Fingerspitzengefühl. Der Data Mining Prozess erfordert also die Interaktion mit einem Data Mining Experten, der die Ergebnisse richtig interpretieren und entsprechende Entscheidungen für den weiteren Prozess treffen kann.

In der Praxis steht also beim Data Mining der Mensch im Mittelpunkt, der Ziele und Vorgangsweise festlegt, mit einer oft sehr umfangreichen Datenbank interagiert und eine große Menge heterogener Werkzeuge einsetzt, um einzelne Teilprobleme zu lösen, bevor aufgrund seines Expertenwissens die Resultate zu entscheidungsrelevanten *Erkenntnissen* verdichtet werden können.

Die Vorgangsweise beim Data Mining wird manchmal im Gegensatz zum herkömmlichen *Verification Model* als *Discovery Model* bezeichnet:

- **Verification Model:** Hier werden Hypothesen und Fragen von einem Anwendungsexperten formuliert und dann mit Hilfe verschiedener Abfrage- und Analysetools anhand der Daten bestätigt oder verworfen. Dieses Modell liegt den meisten klassischen statistischen Analysemethoden zugrunde.
- **Discovery Model:** Hier werden interessante Hypothesen automatisch generiert und anhand der Daten überprüft. Der Algorithmus kann also neue und überraschende Hypothesen liefern.

In der Praxis verschwimmt diese Unterscheidung allerdings recht oft, da zwar einerseits nach neuen Hypothesen gesucht wird, andererseits Vorwissen und Ergebnisse aus verifizierenden Abfragen ein sinnvolles Hintergrundwissen für die Beschränkung der meist riesigen Anzahl in Frage kommender Hypothesen darstellen.

3.2 ...der Muster aus Daten gewinnt

Beim Data Mining ist man daran interessiert, *Wissen* aus den zur Verfügung stehenden Daten zu gewinnen. Aber was ist Wissen? Im Zusammenhang mit Data Mining versteht man unter Wissen im allgemeinen

Muster, die bestimmte zusätzliche *Eigenschaften* (s.u.) aufweisen und in einer meist formalen *Sprache* dargestellt werden.

Die häufigsten *Arten* von Mustern sind:

- regelhafte Beziehungen zwischen Eigenschaften von Objekten

- Beziehungen zwischen Objekten
- zeitliche Verläufe
- räumliche Muster
- Gruppierungen ähnlicher Objekte
- mathematische Gesetzmäßigkeiten
- Abweichungen von statistischen Verteilungen
- Ausnahmen, auffällige Objekte

Im Kapitel 4 werden die wichtigsten Arten von Mustern im Detail erklärt und Techniken vorgestellt, wie diese Muster in Datenbanken gefunden werden können.

Von den gefundenen Mustern wird oft verlangt, dass sie mehrere der folgenden Eigenschaften erfüllen:

- *understandable* (verständlich): das Wissen ist in einer leicht verständlichen Sprache formuliert oder wird in graphischer Form dargestellt. Die Menge der gelieferten Information ist klein genug, um vom Anwender noch erfasst werden zu können. Verständlichkeit ist schwer objektivierbar, man kann versuchen, sie mit einer heuristischen *simplicity measure function* abzuschätzen. Dabei sollte sowohl die Form (Syntax) als auch der Inhalt (Semantik) des dargestellten Wissens berücksichtigt werden.
- *valid* (gültig): das gefundene Wissen soll nicht nur auf die vorhandenen, sondern auch auf zukünftige Daten mit einer gewissen Sicherheit (*certainty*) zutreffen. Zur Abschätzung der Sicherheit wird eine *certainty measure function* verwendet.
- *novel* (neu): das gelieferte Wissen sollte neu, vielleicht sogar überraschend sein.
- *useful* (nützlich): das gefundene Wissen soll für die konkrete Aufgabenstellung verwendbar und relevant sein.
- *non-trivial* (nicht-trivial): Eine Reihe von Ergebnissen, die konventionelle Datenbanksysteme liefern können, erfüllen alle bisherigen Bedingungen. Der durchschnittliche Umsatz aller Filialen in Ostösterreich ist verständlich, gültig, neu und nützlich. Die Art der Berechnung ist allerdings trivial. Bei der Suche nach Wissen im Sinne des Data Mining wird vom Algorithmus verlangt, autonom komplexe Zusammenhänge zu untersuchen und nur interessante als Wissen zu präsentieren. Ein Data Mining System könnte also z.B. eigenständig in einer Datenbank entdecken, dass der Mittelwert fuer eine bestimmte Gruppe von Objekten signifikant vom Gesamtmittelwert abweicht.

- *interesting* (interessant): Ein Maß der Interessantheit könnte man z.B. aus den oben erwähnten Funktionen für certainty und simplicity (und vielleicht auch Funktionen, die novelty und usefulness abschätzen) ermitteln.

Muster werden in einer formalen Sprache ausgedrückt. Im Data Mining sollen die Ausdrücke dieser Sprache auch für Menschen leicht verständlich sein oder zumindest in eine Darstellung übersetzt werden können, die möglichst leicht verständlich ist. Oft verwendete Darstellungsformen sind z.B. Regeln, Entscheidungsbäume oder PROLOG-Programme. Oft sind graphische Darstellungsformen leichter zu erfassen, die *Visualisierung* von Ergebnissen spielt im Data Mining daher eine wichtige Rolle.

3.3 ... mittels unterschiedlicher Methoden und Techniken

Viele der Methoden zum Finden von Mustern in Datenbanken kommen ursprünglich aus dem Bereich des Maschinellen Lernens, aber auch statistische Methoden und interaktive Analyse mittels Visualisierungsmethoden werden verwendet. Die Methoden lassen sich grob nach der Art der zu findenden Muster einteilen; für manche Musterarten gibt es eine Vielzahl zur Verfügung stehender Methoden und Algorithmen. Der Data Mining-Experte kennt die unterschiedlichen Eigenschaften der verschiedenen Methoden und kann daher entscheiden, welche Methode in welcher Situation einzusetzen ist, oder wie Unterschiede in den Ergebnissen zwischen verschiedenen Methoden zu bewerten sind.

Ein Überblick über im Data Mining häufig eingesetzte Methoden und Algorithmen wird in Kap. 4 anhand der verschiedenen Arten von Mustern gegeben.

4 Arten von Mustern

Beim Data Mining ist man daran interessiert, *interessante Muster* in Daten zu entdecken. Welche Art von Mustern interessant sind, kann sich dabei jeweils aus der konkreten Aufgabenstellung ergeben. Die meisten Arten von Mustern, an denen man für gewöhnlich interessiert ist, lassen sich aber einer der folgenden Gruppen zuordnen:

- Regeln, die Zusammenhänge zwischen den verschiedenen Attributen eines Objekts beschreiben (Klassifikationsregeln, charakteristische Regeln, Regressionsregeln, Assoziationsregeln, Determinationen)
- Muster, die Verknüpfungen zwischen Objekten beschreiben (Verbindungsmuster)
- häufig wiederkehrende Abfolgen (Sequences)

- Gruppierungen ähnlicher Objekte (Cluster)
- auffällig unterschiedliche Attribute bei Objekten, die einer bestimmten Gruppe angehören (Abweichungen)

Diese Gruppen werden im Folgenden genauer beschrieben und mit einfachen Beispielen illustriert.

4.1 Klassifikationsregeln

Das Lernen von Klassifikationsregeln ist die im maschinellen Lernen am häufigsten studierte Aufgabe. Eine Klassifikationsregel versucht, den Wert einer abhängigen Zielvariable (die Klasse) aus den Werten von bekannten Variablen vorauszusagen. Ein – natürlich stark vereinfachtes – Beispiel wäre³:

```

WENN KreditType = A UND
    BesitztHaus = ja UND
    Angestellter = ja UND
    Betrag < 500.000
    Dauer > 6
DANN: Kreditwuerdigkeit = hoch [2341/4]

```

Diese Regel besagt:

„Wenn jemand, der ein Haus besitzt und Angestellter ist, einen Kredit der Type A mit Laufzeit 6 Jahre und einem Betrag kleiner 500.000 beantragt, ist seine Kreditwürdigkeit hoch einzustufen.“

Meist ist man daran interessiert, Mengen von Regeln zu finden, die für alle vorkommenden Klassen Bedingungen angeben und einen möglichst großen Anteil der vorkommenden Objekte abdecken. Dabei ist man an möglichst einfachen, kompakten Regeln interessiert.

Klassifikationsregeln werden immer dann benötigt, wenn fehlende Informationen aus bekannten Eigenschaften von Objekten abgeleitet werden sollen.

Klassifikationsregeln werden typischerweise aus einer Menge von vorklassifizierten Trainingsbeispielen gelernt. Dazu stehen eine Vielzahl verschiedener Algorithmen zur Verfügung. Die drei im Data Mining wichtigsten Ansätze sind:

Lernen von Decision Trees: Ein Decision Tree (Entscheidungsbaum) ist eine Hierarchie von Tests, deren Ausgang gemeinsam die Klassenzugehörigkeit eines Objekts bestimmt. Das bekannteste Verfahren, ID3 (Quinlan 1983), beruht auf

³Alle Beispiele dienen nur zur Illustration, die angegebenen Muster entsprechen daher nicht der Realität

der Idee, Tests die möglichst viel Information über die Klassenzugehörigkeit des Objekts gewinnen, zuerst durchzuführen, also nahe der Wurzel des Baums anzusiedeln. Die bekanntesten Vertreter dieser Familie sind C4.5 (Quinlan 1993) und CART (Breiman, Friedman, Olshen, & Stone 1984).

Regellernen mittels Covering: Das Prinzip das dem Covering zugrundeliegt, besteht darin, eine einzelne Regel zu finden, die möglichst viele Beispiele einer Klasse abdeckt. Diese Beispiele werden dann entfernt. Aus den restlichen Beispielen wird eine neue Regel gelernt, solange bis alle Beispiele durch eine Regel abgedeckt werden. Die prototypischen Algorithmen für diese Klasse von Algorithmen sind AQ (Michalski 1969) und CN2 (Clark & Niblett 1989).

Attribut-orientierte Induktion: Dieses Verfahren findet Regeln, indem identische Datensätze in einer Datenbank, die durch das Generalisieren von Attributwerten entstehen verschmolzen werden (Cai, Cercone, & Han 1991). Als wesentliche Hintergrundinformation werden *Abstraktionshierarchien* für die vorkommenden Werte benötigt. Diese Methode wird auch im kommerziellen Programm DB-MINER (Han, Fu, Wang, Chiang, Gong, Koperski, Li, Lu, Rajan, Stefanovic, Xia, & Zaiane 1996) verwendet.

Während die meisten Lernalgorithmen nur Tests auf einzelne Werte von Attributen berücksichtigen können, hat sich in letzter Zeit im Gebiet des *Inductive Logic Programming* (Muggleton 1992; De Raedt 1995) eine Familie von Algorithmen entwickelt, die auch Relationen und Abhängigkeiten zwischen den Attributen als Tests in Klassifikationsregeln zulassen können. Die beiden wichtigsten Vertreter sind FOIL (Quinlan 1990) and PROGOL (Muggleton 1995).

4.2 Charakteristische Regeln

Charakteristische Regeln beschreiben möglichst kompakt, welche Eigenschaften allen zu einer Klasse gehörenden Objekte gemeinsam sind (aber im Unterschied zu Klassifikationsregeln möglicherweise auch den nicht zur Klasse gehörenden Objekten). Ein einfaches Beispiel für eine charakteristische Regel könnte etwa folgendermaßen aussehen:

WENN Kreditwürdigkeit = hoch
DANN KreditType = A,B UND
Angestellter = ja

„Personen mit hoher Kreditwürdigkeit sind typischerweise Angestellte und beantragen einen Kredit der Type A oder B“

Charakteristische Regeln können verwendet werden, um „typische“ Eigenschaften von Objekten einer bestimmten Klasse zu beschreiben. Eine gängige Methode zum Finden charakteristischer Regeln ist Attributorientierte Induktion (s. vorheriger Abschnitt).

4.3 Regressionsregeln

Mit Regressionsregeln wird im Unterschied zu Klassifikationsregeln nicht die Zugehörigkeit eines Objekts zu einer von n Klassen bestimmt, sondern ein numerischer Wert für dieses Objekt vorhergesagt. Das ist dann notwendig, wenn der Wert eines numerischen Attributs aus anderen Attributen abgeleitet werden soll, und eine Diskretisierung des numerischen Attributes nicht sinnvoll ist.

Eine an das Lernen von Decision Trees angelehnte Methode ist die Induktion von *Regression Trees* (Breiman, Friedman, Olshen, & Stone 1984; Kramer 1996b)

4.4 Cluster

Während beim Lernen von Klassifikationsregeln die Einteilung der Objekte in Klassen schon vorgegeben ist, muss beim Lernen von *Clusters* diese Einteilung selbständig ermittelt werden. Beim Clustering werden Objekte zu Gruppen zusammengefasst, also eine selbstständige Klassifizierung vorgenommen.

Die Bildung der Gruppen kann mit Hilfe unterschiedlicher Strategien erfolgen. Klassische statistische Verfahren basieren auf den Abständen zwischen multidimensionalen Datenpunkten, (MacQueen 1967; Jarvis & Patrick 1973), neuere Algorithmen verwenden wahrscheinlichkeitsbasierte (Fisher 1987; Gennari, Langley, & Fisher 1989), informationstheoretische oder Bayes'sche Methoden (Cheeseman & Stutz 1995). Während die statistischen Verfahren sich auf das Finden der Gruppen beschränken, ist man bei den Data Mining Methoden zusätzlich daran interessiert eine verständliche Beschreibung der Gruppen, etwa in Form von Klassifikationsregeln, zu erhalten.

Ein einfaches Beispiel für eine in einer Kundendatenbank gefundene Gruppe könnte etwa folgendermaßen aussehen:

Gruppe-020: (N=103)
Bestellhäufigkeit = selten
durchschn-Bestellvol < 10000
Alter > 45

4.5 Abhängigkeitsmuster

4.5.1 Assoziationsregeln

Assoziationsregeln (Agrawal & Srikant 1994; Agrawal, Imielinski, & Swami 1993; Mannila, Toivonen, & Verkamo 1993; Srikant & Agrawal 1995) beschreiben, welche Gruppen von Objekten, Eigenschaften, etc. häufig gemeinsam auftreten. Die am häufigsten genannte Anwendung ist die Analyse von Warenkorbdaten, also Daten, die beschreiben, welche Produkte ein Kunde bei einem Kauf gemeinsam erworben hat. Eine Assoziationsregel könnte dann folgendermaßen aussehen:

Die Regel wäre dann zu interpretieren als:

„In 34% Prozent der Transaktionen, in denen Milch, Brot und Butter gekauft wurde, wurde auch Kaffee gekauft. Insgesamt wurden Milch, Brot, Butter und Kaffee in 3% aller Transaktionen gemeinsam gekauft.“

4.5.2 Determinationen und funktionale Abhängigkeiten

Mit diesen Begriffen beschreibt man Abhängigkeiten zwischen Mengen von Attributen. Eine funktionale Abhängigkeit liegt vor, wenn der Wert eines Attributs bei allen Objekten eindeutig durch die Werte derselben Menge von Attributen bestimmt wird. Von Determinationen spricht man, wenn der Wert zwar nicht immer, aber sehr oft von anderen Attributen bestimmt wird. Das Finden von funktionalen Abhängigkeiten (Mannila & Rähä 1994; Kivinen & Mannila 1994) und Determinations (Pfahring & Kramer 1995) kann helfen, kausale Zusammenhänge in einer Datenbank aufzuspüren.

4.6 Verbindungsmuster (Links)

Bei den bisher besprochenen Mustern ist man an Zusammenhängen zwischen den Attributen einzelner Objekte interessiert. Eine weitere interessante Aufgabenstellung ist es, Regelmäßigkeiten zwischen verschiedenen Objekten, und unter Umständen verschiedenen Datenbanken zu finden.

Eine Anwendung, die sich in diese Art von Problemstellung einordnen lässt, ist das Entdecken von Geldtransaktionen, bei denen der Verdacht auf Geldwäsche bestehen könnte. Eine derartige Anwendung existiert bei der US Behörde „Financial Crime Enforcement Network“ (FinCEN) mit dem System FAIS (FinCEN AI System) (Senator, Goldberg, Wooton, Cottini, Khan, Klinger, Llamas, Marrone, & Wong 1995a).

4.7 Zeitliche Muster

Assoziationsregeln beschreiben Muster innerhalb einer einzigen Transaktion. Oft ist aber auch die *zeitliche Abfolge* von solchen Kaufmustern interessant. Solche Muster werden *frequent Sequences* genannt. Ein Beispiel:

$\langle \text{(VRC)} [\leq 5] (\text{VidCass}, \text{Batt}) \rangle [0.13]$

„In 13% aller Abfolgen von Käufen pro Kunde werden zuerst ein Videorekorder und dann innerhalb von fünf Tagen sowohl Videokassetten als auch Batterien gekauft.“

4.8 Abweichungen

Abweichungen sind Muster, bei denen eine Teilmenge von Objekten einer Datenbank andere Eigenschaften hat, als aufgrund der Gesamtmenge eigentlich statistisch zu erwarten wäre.

Im folgenden Beispiel hat der Algorithmus in einer Angestellten Datenbank das mittlere Einkommen für die Gruppen „jünger als 32“ und „32 oder älter“ untersucht und signifikante Abweichungen für drei Gruppen gefunden:

```
In population empl1, compare AGE < 32 vs. AGE ≥ 32
Means of CURRENT-SALARY larger in first pupulation for:
FEMALES                                (11115 vs. 9613)
EDUCATION < 12                         (10282 vs. 9623)
EDUCATION 12–15, WORK EXPERIENCE 3–8  (11678 vs. 9552)
```

Das bekannteste System, das derartige Muster findet, ist EXPLORA (Klösgen 1995a; Klösgen 1995b).

4.9 Formeln und Mathematische Gesetzmäßigkeiten

Bei der Analyse von numerischen Daten, etwa aus der Auswertung von wissenschaftlichen Experimenten, ist man oft daran interessiert, den Zusammenhang zwischen Variablen in mathematischer Form herauszufinden. Die Techniken die sich mit dieser Aufgabenstellung befassen, werden auch als *Scientific Discovery* bezeichnet.

5 Exemplarische Anwendungen

Seit Beginn der 90-er Jahre, als man begann, KDD als ein eigenes Forschungsgebiet anzusehen, wurden Anwendungen in den verschiedensten Gebieten entwickelt. In den folgenden Abschnitten werden wir einige Anwendungen diskutieren. Bei der Auswahl wurde insbesondere auf die Vielfalt in den Anwendungsgebieten und der Art der zu entdeckenden Muster Wert gelegt.

5.1 SKICAT: Katalogisierung von Himmelsobjekten

Anwendungsgebiet: Astronomie

Methodik: Lernen von Decision Trees

Literatur: (Fayyad, Weir, & Djorgovski 1993; Fayyad, Djorgovski, & Weir 1996)

Das *California Institute of Technology (CalTech)* führt eine Bestandsaufnahme von Himmelsobjekten durch. Dazu wurden 3000 photographische Platten des Palomar Himmelsobservatoriums digitalisiert und für jedes Objekt auf diesen Platten zirka 40 Attribute ermittelt. Für ungefähr 1700 der insgesamt mehr als 10^8 Objekte wurden Aufnahmen mit höherer Auflösung gemacht, sodaß es für eine Astronomen möglich war, diese Objekte in die Kategorien Sterne, Galaxien, und andere Objekte einzuteilen. Ein Lernalgorithmus, basierend auf der Induktion von Decision Trees, lernte Klassifikationsregeln, die mit einer Genauigkeit von 94% in der Lage waren, aus den 40 Attributen vorherzusagen, ob das vorliegende Objekt ein Stern, eine Galaxie, oder ein anderes Himmelsobjekt ist. Dazu wurden verbesserte Methoden zur Diskretisierung von Daten (Fayyad & Irani 1992; Fayyad & Irani 1993), zur Generierung von möglichen Tests (Cheng, Fayyad, Irani, & Qian 1988), und zur Auswahl der Tests im Decision Tree (Fayyad & Irani 1990) entwickelt. SKICAT kann die Klassifizierungsaufgabe viel schneller als menschliche Astronomen durchführen und ist überdies noch in der Lage, korrekte Voraussagen für schwer erkennbare Objekte zu treffen, was Astronomen nur unter Hinzuziehung von Aufnahmen mit höherer Bildauflösung möglich wäre. Im Dezember 1995 wurde bekanntgegeben, daß mit Hilfe von SKICAT 16 neue Quasare entdeckt worden sind.

5.2 SCOUT: Analyse von Basketballspielen

Anwendungsgebiet: Sport

Methodik: Entdecken von Assoziationsregeln

Literatur:

Für viele Mannschaftssportarten lassen sich unzählige Statistiken aufstellen. In Basketball werden zum Beispiel zusätzlich zum Resultat die Anzahl der Schußversuche, die Anzahl der 3-Punkt Würfe, die Anzahl der Ballverluste und vieles mehr oft auch für jeden Spieler jeder Mannschaft routinemäßig gesammelt (zumindest in den USA). Diese Statistiken beinhalten eine Fülle von Informationen, deren Auswertung dem Trainer wertvolle Hinweise auf die Stärken und Schwächen der eigenen Mannschaft, des nächsten Gegners bzw. einzelner Spieler geben kann, jedoch ist es nicht einfach, die entscheidenden Informationen aus der großen Zahl von Statistiken herauszufiltern. Im IBM T.J. Watson Research Center wurde daher eine Data Mining-Anwendung erstellt, die Basketballtrainern helfen soll, die Unmengen von Daten, die nach jedem Spiel anfallen, zu organisieren und zu interpretieren. SCOUT erlaubt den Trainern, komplexe Anfragen an die Datenbank zu stellen, schlägt selbsttätig neue Fragen vor, und findet interessante Muster, die relevant zur Beantwortung der gestellten Fragen sind, die es dann auch automatisch durch entsprechende Video-Clips aus Spielen belegen kann. Das System wurde von einigen Trainern der Nordamerikanischen Basketball-Liga getestet, und die Genehmigung seines Einsatzes dem Verband empfohlen.

5.3 TASA: Fehleranalyse in Telekommunikationsnetzen

Anwendungsgebiet: Telekommunikation

Methodik: Entdecken von zeitlichen Mustern

Literatur: (Hätönen, Klemettinen, Mannila, Ronkainen, & Toivonen 1996)

An der Universität von Helsinki hat eine Forschungsgruppe ein System zur Analyse der zeitlichen Abfolge von Alarmen in einem Telekommunikationsnetz entwickelt, den *Telecommunication Network Alarm Sequence Analyzer (TASA)*. Ein Telekommunikationsnetz besteht aus einer Vielzahl von miteinander verbundenen Komponenten. Jede Komponente und ihre Subkomponenten können in gewissen Situationen Alarm schlagen, was 200 bis 10,000 mal am Tag passieren kann. TASA sucht in diesen Alarmen nach zeitlichen Mustern der Art "Wenn auf einen Alarm vom Typ A innerhalb von 5 Sekunden ein Alarm des Typs B folgt, so wird mit einer Wahrscheinlichkeit von 0.7 ein Alarm des Typs C folgen." Das Programm ist noch im experimentellen Einsatz, jedoch ist geplant, Teile des gefundenen Wissens in das Alarm Handling System zu integrieren.

5.4 FAIS: Aufdeckung von Geldwäsche

Anwendungsgebiet: Bekämpfung von Wirtschaftskriminalität

Methodik: Verbindungsanalyse

Literatur: (Senator, Goldberg, Wooton, Cottini, Khan, Klinger, Llamas, Marrone, & Wong 1995b)

Das *Financial Crimes Enforcement Network (FINCEN)* ist eine Agentur des Finanzministeriums der Vereinigten Staaten, deren Ziel es ist, Geldwäsche zu verhindern bzw. aufzudecken. In den USA sind alle Bargeldtransaktionen über mehr als 10,000\$ meldepflichtig und werden in einer Datenbank gespeichert. 200,000 solcher Transaktionen landen wöchentlich in dieser Datenbank. Das FINCEN AI System (FAIS) fügt zu den transaktionsbasierten Einträgen in der Datenbank noch konten- und personsbezogene Sichtweisen hinzu. Diese werden dann mit Hilfe eines regelbasierten Expertensystems analysiert, und verdächtige Konten, Personen oder Transaktionen werden dem Benutzer mitgeteilt. Dieser hat dann die Möglichkeit den Fluß der verdächtigen Mittel mit Hilfe eines übersichtlichen Benutzer-Interfaces zu verfolgen und zu entscheiden, ob eine Anklage eingeleitet werden soll bzw. weiter Ermittlungen durchgeführt werden sollen. Das Programm befindet sich seit März 1993 im Einsatz und hat seither mehr als eine Milliarde \$ an verdächtigen Mitteln aufgedeckt. Bis Dezember 1995 kam es zu einer Verurteilung. Mit Hilfe von FAIS sind Mensch und Computer in der Lage, eine Aufgabe durchzuführen, die keiner für sich allein lösen könnte.

5.5 Evaluierung von persönlichen Darlehen

Anwendungsgebiet: Bankwesen

Methodik: Lernen von Decision Trees

Literatur: (Feelders & Verkooijen 1995)

Bei dieser Fallstudie der Niederländischen Bank AMR AMRO wurde die Anwendbarkeit des Decision Tree Lernprogramms C4.5 zur Evaluierung persönlicher Darlehen untersucht. Das Programm ermittelte aus einer Datenbank von Darlehensbeschreibungen Regeln, die die Nichterfüllung der Darlehensrückzahlung voraussagen sollten. Diese Methode wurde mit der „klassischen“ statistischen Methode der Linearen Diskriminanzanalyse verglichen. Erste Ergebnisse zeigten, daß die gelernten Regeln ein etwas besseres Vorhersageverhalten zeigten, vor allem aber leichter verständlich waren als das statistische Modell.

5.6 Qualitätsanalyse von Flußwasser

Anwendungsgebiet: Biologie/Chemie

Methodik: Lernen von Klassifikationsregeln

Literatur: (Džeroski & Grbović 1995)

Die Wassergüte slovenischer Flüsse wird aufgrund von chemischen und biologischen Analysen festgelegt. Bei den biologischen Analysen wird die Wasserqualität aufgrund des Vorhandenseins oder Fehlens bestimmter Arten von Lebewesen (Bioindikatoren) abgeschätzt. Das Data Mining System dieser Anwendung wurde dazu eingesetzt, einerseits den Einfluß physikalischer und chemischer Eigenschaften des Flußwassers auf diese Bioindikatoren zu untersuchen, andererseits, die Wasserqualität direkt aus den Daten über Bioindikatoren bzw. physikalisch/chemischer Messungen abzuleiten. Dazu wird das Programm CN2 zum Finden von Klassifikationsregeln eingesetzt.

5.7 Analyse von Daten aus dem Gesundheitswesen

Anwendungsgebiet: Gesundheitswesen

Methodik: Finden von Abweichungen

Literatur: (Matheus, Piatetsky-Shapiro, & McNeill 1996)

Bei dieser Anwendung wird der „Key Findings Reporter“ (KEFIR), ein System, das nach „key findings“ in großen, veränderlichen Datenbanken sucht, dazu eingesetzt, Einrichtungen des Gesundheitswesens zu verbessern und Kosten zu senken. KEFIR erstellt selbsttätig Berichte die interessante Trends und Abweichungen enthalten, in der Form von Hypertext-Dokumenten. KEFIR ist in der Lage, die „Interessantheit“ der gefundenen Abweichungsmuster abzuschätzen und zur Reihung im Bericht zu verwenden.

6 Data Mining am ÖFAI

Das Österreichischen Forschungsinstitut für Artificial Intelligence (ÖFAI) arbeitet seit Jahren auf dem Gebiet des maschinellen Lernens. Die Abteilung für maschinelles Lernen verfügt über ausgedehnte Erfahrungen mit Möglichkeiten, Methoden und Anwendungen des maschinellen Lernens und Data Minings. Unsere Expertise beruht auf langjähriger intensiver und international stark beachteter Forschungstätigkeit. Als Folge dieser Aktivitäten ist das ÖFAI Mitglied in zwei Networks of Excellence der Europäischen Union im Bereich Machine Learning: dem *Network of Excellence in Machine Learning (MLNet)* und dem *Inductive Logic Programming Pan-European Scientific Network (ILPNet)*. Darüber hinaus arbeitet das ÖFAI regelmäßig mit Partnern aus Wirtschaft und Forschung an gemeinsamen Projekten.

Die folgende Aufstellung soll eine kurzen Auswahl aus den Aktivitäten der letzten Zeit im Bereich maschinelles Lernen und Data Mining am ÖFAI geben:

- **Klassifikation / kategorische Vorhersage**

Methoden für

- das Lernen von Entscheidungsbäumen (Kramer 1996a)
- das Lernen von Regeln
 - ▷ in propositionaler Form (Pfahring 1995d)
 - ▷ in relationaler Form (Fürnkranz 1994b)
- die automatische Konstruktion relevanter komplexer Deskriptoren während des Lernens (konstruktive Induktion) (Kramer 1994; Pfahring 1994)
- die automatische Identifizierung relevanter Variabler für Induktionsprobleme (*feature selection*) (Pfahring 1995b)
- die Diskretisierung numerischer Variabler (Pfahring 1995a)
- die informationstheoretische Bewertung / Evaluierung von Lernresultaten mit Hilfe des *Minimum Message Length Principle* (Pfahring 1995e)

- **numerische Vorhersage**

- Integration von Regelinduktion und *case-based learning* für die numerische Vorhersage (Widmer 1993)

- Algorithmen zur Konstruktion von symbolisch-numerischen Regressionsbäumen (*structural regression trees*) (Kramer 1996a)
- **Wissensgewinnung aus Daten**
 - Lernen logischer Klassifikations- und Vorhersageregeln aus relationalen Daten (relationales Lernen, *Inductive Logic Programming*) (Fürnkranz 1994a; Fürnkranz 1995)
 - Entdecken (quasi-)funktionaler Abhängigkeiten in Datenbanken (Pfahring & Kramer 1995; Kramer & Pfahring 1996)
 - Extraktion relevanter Fälle aus Datenbanken durch *case-based / instance-based learning* (Petrak 1995)
 - Methoden zur Integration von Vorwissen in den Lernprozess mit dem Ziel der Induktion besserer und verständlicherer Regeln:
 - ▷ *plausible explanation-based learning* basierend auf qualitativen Modellen des Anwendungsbereichs (Widmer 1994)
 - ▷ Nutzung anatomischen Wissens bei Anwendungen in der medizinischen Bildanalyse (Petta 1994)
 - praktische Anwendung dieser Methoden auf realistische Datenbanken (Trappl, Fürnkranz, & Petrak 1996)
- **On-line-Lernen**
Lernen (Klassifikation/Vorhersage) in dynamischen, veränderlichen Umgebungen:
 - durch dynamische Anpassung an ‘Concept Drift’ (Widmer & Kubat 1996)
 - durch automatische Erkennung von Kontexteinflüssen und Anpassung an veränderliche Kontexte (Widmer 1996; Widmer)
- **Optimierung**
mittels genetischer Algorithmen (Pfahring 1995c)

Danksagung

Die Forschung des Österreichischen Instituts für Artificial Intelligence (ÖFAI) auf dem Gebiet Data Mining wird teilweise vom Österreichischen Fonds zur Förderung der wissenschaftlichen Forschung (FWF) unterstützt. Das ÖFAI wird teilweise vom Bundesministerium für Wissenschaft, Verkehr und Kunst unterstützt.

- Agrawal, R., T. Imielinski, & A. Swami (1993). Mining association rules between sets of items in large databases. In: *Proceedings of the International Conference on Management of Data*.
- Agrawal, R. & R. Srikant (1994). Fast algorithms for mining association rules. In: *Proceedings of the 20th International Conference on Very Large Databases*.
- Babcock, C. (1994, Sep. 26). Parallel processing mines retail data. *Computer World* 6.
- Breiman, L., J. Friedman, R. Olshen, & C. Stone (1984). *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth & Brooks.
- Cai, Y., N. Cercone, & J. Han (1991). Attribute-oriented induction in relational databases. In: Piatetsky-Shapiro & Matheus (1991), Kapitel 12.
- Cheeseman, P. & J. Stutz (1995). Bayesian classification (Auto Class): Theory and results. In: U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, Kapitel 6. Menlo Park: AAAI Press.
- Cheng, J., U. M. Fayyad, K. B. Irani, & Z. Qian (1988). Improved decision trees: A generalized version of ID3. In: *Proceedings of the 5th International Conference on Machine Learning*, pp. 100–106.
- Clark, P. & T. Niblett (1989). The CN2 induction algorithm. *Machine Learning* 3(4), 261–283.
- De Raedt, L. (Ed.) (1995). *Advances in Inductive Logic Programming*, Volume 32 of *Frontiers in Artificial Intelligence and Applications*. IOS Press.
- Džeroski, S. & J. Grbović (1995). Knowledge discovery in a water quality database. In: Fayyad (1995), pp. 81–95.
- Fasman, K. H., A. J. Cuticchia, & D. T. Kingsbury (1994). The GDB (TM) human genome database anno 1994. *Nucl. Acid. R.* 22(17), 3462–3469.
- Fayyad, U. M. (Ed.) (1995). *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA. AAAI Press.
- Fayyad, U. M., S. G. Djorgovski, & N. Weir (1996, Summer). From digitized images to online catalogs — data mining a sky survey. *AI Magazine* 17(2), 51–66.
- Fayyad, U. M. & K. B. Irani (1990). What should be minimized in a decision tree? In: *Proceedings of the 8th National Conference on Artificial Intelligence (AAAI-90)*, Menlo Park, CA, pp. 749–754. AAAI Press.
- Fayyad, U. M. & K. B. Irani (1992). On the handling of continuous-valued attributes in decision tree generation. *Machine Learning* 8(2).
- Fayyad, U. M. & K. B. Irani (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In: R. Bajcsy (Ed.), *Proceedings of*

- the 13th International Joint Conference on Artificial Intelligence, San Mateo, CA, pp. 1022–1029. Morgan Kaufmann.
- Fayyad, U. M., G. Piatetsky-Shapiro, & P. Smyth (1995). From data mining to knowledge discovery: An overview. In: Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy (1996), Kapitel 1.
- Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.) (1996). *Advances in Knowledge Discovery and Data Mining*. Menlo Park: AAAI Press.
- Fayyad, U. M., N. Weir, & S. G. Djorgovski (1993). SKICAT: A machine learning system for automated cataloging of large scale sky surveys. In: *Proceedings of the 10th International Conference on Machine Learning*, San Mateo, CA, pp. 112–119. Morgan Kaufmann.
- Feelders, A. & W. Verkooijen (1995, Jan). Which method learns most from the data? - methodological issues in the analysis of comparative studies. In: *Proceedings of the fifth international workshop on AI and Statistics*, Fort Lauderdale, Florida, pp. 219–225.
- Fisher, D. (1987). Improving inference through conceptual clustering. In: *Proceedings of the 6th National Conference on Artificial Intelligence (AAAI-87)*, Los Altos, CA, pp. 461–465. Morgan Kaufmann.
- Frawley, W. J., G. Piatetsky-Shapiro, & C. J. Matheus (1991). Knowledge discovery in databases: An overview. In: Piatetsky-Shapiro & Matheus (1991), Kapitel 1, pp. 1–27.
- Fürnkranz, J. (1994a). *Efficient Pruning Methods for Relational Learning*. Dissertation, Technical University of Vienna, Vienna, Austria.
- Fürnkranz, J. (1994b). FOSSIL: A robust relational learner. In: F. Bergadano & L. de Raedt (Eds.), *Proceedings of the 7th European Conference on Machine Learning (ECML-94)*, Berlin, pp. 122–137. Springer Verlag.
- Fürnkranz, J. (1995). A tight integration of pruning and learning. In: N. Lavrac & S. Wrobel (Eds.), *Proceedings of the 8th European Conference on Machine Learning (ECML-95)*, Berlin. Springer Verlag.
- Gennari, J. H., P. Langley, & D. Fisher (1989). Models of incremental concept formation. *Artificial Intelligence* 40, 11–61.
- Han, J., Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu, A. Rajan, N. Stefanovic, B. Xia, & O. Zaiane (1996). DBMiner: A system for mining knowledge in large relational databases. In: Simoudis, Han, & Fayyad (1996), pp. 250–255.
- Hätönen, K., M. Klemettinen, H. Mannila, P. Ronkainen, & H. Toivonen (1996). Knowledge discovery from telecommunication network alarm databases. In: *Proceedings of the 12th International Conference on Data Engineering (ICDE-96)*.

- Jarvis, R. A. & E. A. Patrick (1973). Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers* C-22(11), 1025–1034.
- Kivinen, J. & H. Mannila (1994). The power of sampling in knowledge discovery. Technical Report Publication Series C, No. C-1993-66, University of Helsinki, Department of Computer Science, Helsinki, Finland.
- Klösgen, W. (1995a). Efficient discovery of interesting statements in databases. *Journal of Intelligent Information Systems* 4(1), 53–69.
- Klösgen, W. (1995b). *Explora - A support system for Discovery in Databases*. Sankt Augustin, Germany: Gesellschaft für Mathematik und Datenverarbeitung. Manual.
- Kramer, S. (1994). CN2-MCI: A two-step method for constructive induction. In: *Proceedings of the Workshop on Constructive Induction and Change of Representation*, New Brunswick, New Jersey. 11th International Conference on Machine Learning (ML-94/COLT-94).
- Kramer, S. (1996a). Structural regression trees. In: *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)*, Portland, Oregon. AAAI Press.
- Kramer, S. (1996b). Structural regression trees: First steps. In: R. Trappl (Ed.), *Cybernetics and Systems '96. Proceedings of the 13th Conference on Cybernetics and Systems Research*, Vienna. Austrian Society for Cybernetic Studies.
- Kramer, S. & B. Pfahringer (1996). Efficient search for strong partial determinations. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Menlo Park, CA. AAAI Press.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkley Symposium on Mathematics, Statistics and Probability*, pp. 281–296.
- Mannila, H. & K.-J. Räihä (1994). Algorithms for inferring functional dependencies from relations. *Data & Knowledge Engineering* 12, 83–99.
- Mannila, H., H. Toivonen, & A. I. Verkamo (1993). Improved methods for finding association rules. Technical Report Report C-1993-65, University of Helsinki Department of Computer Science.
- Matheus, C. J., G. Piatetsky-Shapiro, & D. McNeill (1996). Selecting and reporting what is interesting: The KEFIR application to healthcare data. In: Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy (1996), Kapitel 20, pp. 495–516.
- Michalski, R. S. (1969). On the quasi-minimal solution of the covering problem. In: *Proceedings of the 5th International Symposium on Information Processing (FCIP-69)*, Volume A3 (Switching Circuits), Bled, Yugoslavia, pp. 125–128.
- Muggleton, S. H. (Ed.) (1992). *Inductive Logic Programming*. London: Academic Press Ltd.

- Muggleton, S. H. (1995). Inverse entailment and Progol. *New Generation Computing* 13(3,4), 245–286. Special Issue on Inductive Logic Programming.
- Petrak, J. (1995). VIE-CBR2: An object-oriented case-based learning system. Master’s thesis, Department of Medical Cybernetics and Artificial Intelligence, University of Vienna.
- Petta, P. (1994). The value of background knowledge for ILP: A case study in the interpretation of Tl-201 myocardial SPECT polar maps. Technical Report TR-94-15, Austrian Research Institute for Artificial Intelligence.
- Pfahringer, B. (1994). Controlling constructive induction in CIPF: An MDL approach. In: *Proceedings of the 7th European Conference on Machine Learning (ECML-94)*, Berlin, pp. 242–256. Springer Verlag.
- Pfahringer, B. (1995a). Compression-based discretization of continuous attributes. In: A. Prieditis & S. Russell (Eds.), *Proceedings of the 12th International Conference on Machine Learning (ML95)*, San Francisco. Morgan Kaufmann. `file://ftp.ai.univie.ac.at/papers/oefai-tr-95-31.ps.Z` (Draft version).
- Pfahringer, B. (1995b). Compression-based feature subset selection. In: *Proceedings of the IJCAI-95 Workshop on Data Engineering for Inductive Learning*, Montreal, Canada.
- Pfahringer, B. (1995c). Evolving good tsp tours by means of heuristic repair and strong crowding. Technical Report TR-95-33, Austrian Research Institute for Artificial Intelligence, Vienna.
- Pfahringer, B. (1995d). A new MDL measure for robust rule induction. In: N. Lavrac & S. Wrobel (Eds.), *Proceedings of the 8th European Conference on Machine Learning (ECML-95)*, Berlin. Springer Verlag.
- Pfahringer, B. (1995e). *Practical Uses of the Minimum Description Length Principle in Inductive Learning*. Dissertation, Technical University of Vienna.
- Pfahringer, B. & S. Kramer (1995). Compression-based evaluation of partial determinations. In: *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, Montreal, Canada. `file://ftp.ai.univie.ac.at/papers/oefai-tr-95-31.ps.Z` (draft version).
- Piatetsky-Shapiro, G. & C. Matheus (Eds.) (1991). *Knowledge Discovery in Databases*. Menlo Park: AAAI Press/MIT Press.
- Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end games. In: R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning. An Artificial Intelligence Approach*, pp. 463–482. Palo Alto, CA: Tioga.
- Quinlan, J. R. (1990). Learning logical definitions from relations. *Machine Learning* 5, 239–266.

- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.
- Senator, T. E., H. G. Goldberg, J. Wooton, M. A. Cottini, A. F. U. Khan, C. D. Klinger, W. M. Llamas, M. P. Marrone, & R. W. H. Wong (1995a, Winter). The financial crimes enforcement network AI system (FAIS). *AI Magazine*, 21–39.
- Senator, T. E., H. G. Goldberg, J. Wooton, M. A. Cottini, A. F. U. Khan, C. D. Klinger, W. M. Llamas, M. P. Marrone, & R. W. H. Wong (1995b, Winter). The Financial Crimes Enforcement Network AI System (FAIS). *AI Magazine* 16(4), 21–39.
- Simoudis, E., J. Han, & U. Fayyad (Eds.) (1996). *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Menlo Park. AAAI Press.
- Srikant, R. & R. Agrawal (1995). Mining generalized association rules. In: *Proceedings of the 21st International Conference on Very Large Databases*, Zurich.
- Trappl, R., J. Fürnkranz, & J. Petrak (1996). Digging for peace: Using machine learning methods for assessing international conflict databases. In: *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI-96)*, Chichester, UK. Wiley.
- Way, J. & E. A. Smith (1991). The evolution of synthetic aperture radar systems and their progression to the EOS SAR. *IEEE Trans. on Geoscience and Remote Sensing* 29(6), 962–985.
- Widmer, G. Tracking changing contexts through meta-learning. *Machine Learning*. (to appear).
- Widmer, G. (1993). Combining knowledge-based and instance-based learning to exploit qualitative knowledge. *Informatica* (17), 371–385.
- Widmer, G. (1994). Learning with a qualitative domain theory by means of plausible explanations. In: R. Michalski & G. Tecuci (Eds.), *Machine Learning: A Multistrategy Approach*, Volume IV. San Mateo, CA: Morgan Kaufmann.
- Widmer, G. (1996). Recognition and exploitation of contextual clues via incremental meta-learning. In: *Proceedings of the 13th International Conference on Machine Learning*, San Francisco, CA. Morgan Kaufman.
- Widmer, G. & M. Kubat (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning* 23(1), 69–101.