Einführung in Datenanalyse und Data Mining mit intelligenten Technologien

Christian Borgelt
Otto-von-Guericke-Universität Magdeburg
Institut für Wissens- und Sprachverarbeitung
Universitätsplatz 2, D-39124 Magdeburg
Tel.: +49.391.67.12700, Fax: +49.391.67.12018
E-mail: christian.borgelt@cs.uni-magdeburg.de
WWW: http://fuzzy.cs.uni-magdeburg.de

We are drowning in information,
but starving for knowledge. John Naisbett

1 Einleitung

In jedem Unternehmen gibt es heute Systeme zur elektronischen Datenverarbeitung, sei es in der Produktion, im Vertrieb, in der Lagerhaltung oder im Personalwesen. Jedes dieser Systeme benötigt, um seine Aufgabe erfüllen zu können, Daten, die entweder noch in einfachen Dateien oder schon in modernen Datenbanksystemen abgelegt sind. Man könnte nun meinen, daß alles erreicht sei, wenn diese Daten in ausreichendem Umfang zur Verfügung stehen. Hat man z.B. eine gut geführte Kundenkartei, so „weiß“ man ja alles, was man über seine Kunden wissen muß, denn jede denkbare Einzelinformation kann jederzeit abgerufen werden.

Doch Daten allein genügen nicht. Man könnte sagen, daß man in einer Datenbank den Wald vor lauter Bäumen nicht sieht. Denn Einzelinformationen lassen sich aus einer Datenbank leicht abrufen, auch kann man einfache Aggregationen berechnen lassen (z.B. den durchschnittlichen Monatsumsatz im Raum Frankfurt im Jahre 1996), doch allgemeinere Muster, Strukturen, Regelmäßigkeiten bleiben unbemerkt. Gerade diese Muster können es jedoch sein, die sich z.B. zu einer Umsatzsteigerung ausnutzen lassen. Findet man z.B. in einem Supermarkt heraus, daß bestimmte Produkte oft zusammen gekauft werden, so kann die Verkaufszahl u.U. durch eine entsprechende Anordnung dieser Produkte in den Regalen des Marktes gesteigert werden.

In diesem Aufsatz versuche ich zunächst, den Unterschied zwischen „Daten“ und „Wissen“ zu fassen, um Begriffe zu gewinnen, mit denen sich deutlich machen läßt, warum das bloße Sammeln von Daten nicht ausreicht. Als Illustration führe ich ein Beispiel aus der Geschichte der Wissenschaft an. Ich gehe dann auf die im Zusammenhang mit Datenanalysen immer häufiger genannten Schlagworte „Knowledge Discovery in Databases“ (KDD) und „Data Mining“ (DM) ein und erläutere den KDD-Prozeß, in dem „Data Mining“ einen Schritt darstellt. Um die Ideen des Data Mining zu veranschaulichen, bespreche ich schließlich einige Beispiele aus der Vielzahl verfügbarer Data-Mining-Verfahren.

2 Daten und Wissen

Ich unterscheide in diesem Aufsatz zwischen Daten (data) und Wissen (knowledge). Aussagen wie „Kolumbus entdeckte Amerika im Jahre 1492.“ oder „Herr Meier fährt einen VW Golf.“ sind Daten. Dabei ist es irrelevant, ob ich das Jahr der Entdeckung Amerikas und den Typ des Wagens von Herrn Meier schon kenne oder nicht, ob ich diese Kenntnis im Moment benötige oder nicht, usw. Wesentlich ist, daß sich diese Aussagen auf Einzelfälle beziehen. Sie haben daher (wenn sie wahr sind) nur einen engen Gültigkeitsbereich und sind folglich nur sehr begrenzt nützlich.

Oft wird statt „Datum“ auch das Wort „Information“ verwendet. Dies hängt mit der Bedeutung zusammen, die dem Wort „Information“ in der Shannonschen Informationstheorie gegeben wird. Im Alltag benutzen wir jedoch „Information“ in anderer Weise. Nicht jedes Datum ist eine Information, dazu muß es in der vorliegenden Situation auch relevant sein. Um mögliche Mißverständnisse zu vermeiden, werde ich daher stets von „Daten“ sprechen.

Wissen besteht aus Aussagen wie „Alle Massen ziehen einander an.“ oder „Täglich um 17:04 Uhr fährt ein InterRegio von Magdeburg nach Braunschweig.“ Auch hier wird zunächst die Relevanz der Aussage vernachlässigt. Wesentlich ist, daß sich diese Aussagen nicht auf Einzelfälle beziehen, sondern

allgemeine Gesetze oder Regeln sind. Sie haben daher (wenn sie wahr sind) einen weiten Gültigkeitsbereich, und vor allem: Sie lassen Voraussagen zu und sind daher i.a. sehr nützlich.

Zwar werden, besonders im Alltag, auch Aussagen der Art „Kolumbus entdeckte Amerika im Jahre 1492.“ als Wissen bezeichnet, ich sehe aber von dieser Verwendung des Wortes „Wissen“ ab. Sammlungen von Daten über Einzelfälle sollen noch nicht als Wissen gelten.

Daten und Wissen können demnach durch folgende Eigenschaften gekennzeichnet werden:

Daten

- beziehen sich auf Einzelfälle (Zeitpunkte, Objekte, Personen, etc.)
- beschreiben individuelle Eigenschaften
- sind oft in großer Zahl/Menge vorhanden (Datenbanken)
- sind gewöhnlich leicht zu erfassen bzw. zu beschaffen (z.B. Scannerkassen im Supermarkt, Internet)
- lassen keine Vorhersagen zu

Wissen

- bezieht sich auf Fall*klassen* (Mengen von Zeitpunkten, Objekten, Personen, etc.)
- beschreibt allgemeine Muster, Strukturen, Gesetze, Prinzipien, etc.
- besteht aus möglichst wenigen Aussagen (dies ist eine Zielsetzung, s.u.)
- ist gewöhnlich schwer zu finden bzw. zu beschaffen (z.B. Suche nach Naturgesetzen, Ausbildung)
- läßt Vorhersagen zu

Aus diesen Kennzeichnungen läßt sich bereits ablesen, daß Wissen i.a. wesentlich nützlicher ist als reine Daten. Die Allgemeinheit der Aussagen und die Möglichkeit von Vorhersagen über die Eigenschaften bzw. das Verhalten neuer Fälle machen seine Überlegenheit aus.

Aber auch Wissen ist nicht gleich Wissen. Nicht jede allgemeine Aussage ist gleich wichtig, gleich gehaltvoll, kann mit gleichen Nutzen verwendet werden. Wissen muß daher bewertet werden. Die folgende Liste führt einige Bewertungskriterien auf, beansprucht jedoch nicht, vollständig zu sein.

Bewertungskriterien für Wissen

- Korrektheit (Wahrscheinlichkeit, Testerfolg)
- Allgemeinheit (Geltungsbereich, Geltungsbedingungen)
- Nutzbarkeit (Relevanz, Vorhersagekraft)
- Verständlichkeit (Einfachheit, Übersichtlichkeit)
- Neuheit (vorher unbekannt, unerwartet)

In der Wissenschaft stehen dabei Korrektheit, Allgemeinheit und Einfachheit (Sparsamkeit) im Vordergrund: Wissenschaft kann man (unter anderem) charakterisieren als die Suche nach einer Minimalbeschreibung der Welt. In der Wirtschaft wird man höheren Wert auf Nutzbarkeit, Verständlichkeit und Neuheit legen: Man versucht schließlich, einen Marktvorteil zu erlangen und so einen höheren Gewinn zu erzielen. In keinem der beiden Bereichen kann man es sich jedoch nicht leisten, die übrigen Kriterien zu vernachlässigen.

2.1 Tycho Brahe und Johannes Kepler

Tycho Brahe (1546–1601) war ein dänischer Adeliger und Astronom, der im Jahre 1576 mit der finanziellen Unterstützung von König Frederic II auf der Insel Ven, etwa 32 km nordöstlich von Kopenhagen, eine Sternwarte errichtete. Mit den besten Beobachtungsinstrumenten seiner Zeit (es gab damals noch keine Fernrohre, erst Galileo Galilei (1564–1642) und Johannes Kepler (siehe unten) setzten Fernrohre zur Himmelsbeobachtung ein) bestimmte er die Positionen der Sonne, des Mondes und der Planeten mit einer Genauigkeit von einer Bogenminute, was alle bis zu diesem Zeitpunkt durchgeführten Messungen weit übertraf. Sorgfältig zeichnete er die Bewegungen der Himmelskörper über mehrere Jahre hinweg auf.

Tycho Brahe sammelte Daten über unser Planetensystem. Sehr große Mengen von Daten. Aber er war nicht in der Lage, sie in einem einheitlichen Schema zusammenzufassen. Er konnte genau sagen, an welchem Ort des Himmels z.B. der Mars an einem bestimmten Tag des Jahres 1584 gestanden hatte, aber er war nicht in der Lage, die verschiedenen Positionen an verschiedenen Tagen so durch eine Theorie zueinander in Beziehung zu setzen, so daß sie mit seinen hochgenauen Beobachtungen “ubereinstimmten. Alle Hypothesen, die er versuchte, ließen sich nicht mit seinen Beobachtungen vereinen. Zwar entwickelte er das tychonische Planetensystem, nach dem sich Sonne und Mond um die Erde, alle anderen Planeten aber um die Sonne bewegen, doch bewährte sich dieses System nicht. Heute könnte man sagen: Tycho Brahe hatte ein „Data Mining“ oder „Knowledge Discovery“ Problem. Er verfügte über Daten, konnte aber das in ihnen enthaltene Wissen nicht herausziehen.

Johannes Kepler (1571–1630) war ein deutscher

Astronom und Gehilfe von Tycho Brahe. Er vertrat das kopernikanische Planetensystem und versuchte sein Leben lang, die Gesetzmäßigkeiten zu finden, die die Bewegungen der Planeten bestimmen. Er suchte nach einer mathematischen Beschreibung, was für seine Zeit ein geradezu radikaler Ansatz war. Sein Ausgangspunkt waren die von Tycho Brahe gesammelten Daten. Nach vielen Versuchen und langen Rechnungen gelang es Johannes Kepler schließlich, die Daten Brahes in drei Gesetzen, den bekannten Keplerschen Gesetzen, zusammenzufassen. Nachdem er 1604 erkannt hatte, daß die Marsbahn eine Ellipse ist, veröffentlichte er die ersten beiden Gesetze 1609 in „Astronomia Nova“, das dritte zehn Jahre später in seinem Hauptwerk „Harmonica Mundi“.

1. Alle Planeten bewegen sich auf Ellipsen, in deren einem Brennpunkt die Sonne steht.
2. Eine von der Sonne zum Planeten gezogene Linie überstreicht in gleichen Zeiten gleiche Flächen.
3. Die Quadrate der Umlaufzeiten zweier Planeten verhalten sich zueinander wie die Kuben der großen Hauptachsen ihrer Umlaufbahnen.

Tycho Brahe hatte eine große Menge von Daten gesammelt, Johannes Kepler endeckte die Gesetze, durch die sie erklärt werden können. Er fand die in den Daten versteckte Information, das verborgene Wissen. Dadurch wurde Kepler zu einem der berühmtesten „Data Miner“ der Geschichte.

Heute sind die Arbeiten Tycho Brahes fast vergessen. Seine Kataloge haben nur noch historischen Wert. Kein Lehrbuch der Astronomie enthält Auszüge aus seinen Messungen. Seine Beobachtungen sind reine Daten und haben damit einen entscheidenden Nachteil: Sie lassen keine Vorhersagen zu. Keplers Gesetze werden dagegen in allen Astronomie- und Physiklehrbüchern behandelt, denn sie geben die Prinzipien an, nach denen sich sowohl Planeten als auch Kometen bewegen. Sie fassen alle Messungen Brahes in drei einfachen Aussagen zusammen. Außerdem lassen sie Vorhersagen zu: Kennt man die Position eines Planeten zu einem bestimmten Zeitpunkt, so kann man mit Hilfe der Keplerschen Gesetze seine Bahn vorausberechnen.

3 Knowledge Discovery und Data Mining

Wie hat Johannes Kepler seine Gesetze gefunden? Wie ist es ihm gelungen, aus den langen Tabellen und umfangreichen Katalogen von Tycho Brahe jene einfachen Gesetzen herauszudestillieren, die die Astronomie revolutionierten? Wir wissen nicht viel darüber. Viele Fehlversuche und langwierige Rechnungen werden notwendig gewesen sein. Hohes mathematisches Talent, ausdauernde Arbeit und ein nicht zu unterschätzendes Maß an Glück ermöglichten wohl schließlich den Erfolg. Eine universelle Methode zur Entdeckung physikalischer oder astronomischer Gesetze hat Kepler nicht gekannt.

Auch heute kennen wir keine solche allgemeine Methode. Nach wie vor ist Wissen sehr viel schwerer zu erlangen als Daten, von denen wir in der heutigen „Informationsgesellschaft“ geradezu überschwemmt werden. Es bedarf heute nicht einmal mehr der jahrelangen Fleißarbeit eines Tycho Brahe, um Daten zu erheben. Automatische Meßgeräte, Scanner, digitale Kameras und Computer haben diese Aufgabe übernommen. Die Fortschritte in der Datenbanktechnologie ermöglichen die Speicherung immer größerer Datenmengen. Wir ertrinken in einem Meer von Informationen, aber wir hungern nach Wissen (John Naisbett).

Wenn es schon einen Mann wie Johannes Kepler mehrere Jahre kostete, die nach heutigem Maßstab winzigen Datenbestände Tycho Brahes auszuwerten (wobei er sich sogar noch auf die Marsdaten beschränkte), wie sollen wir heute mit den zur Verfügung stehenden Datenmengen fertig werden? Manuelle Analysen sind schon lange nicht mehr durchführbar. Einfache Hilfsmittel, wie z.B. die Darstellung von Daten in Diagrammen, stoßen an ihre Grenzen. Will man nicht einfach vor der Datenflut kapitulieren, so ist man gezwungen, nach intelligenten rechnergestützten Verfahren Ausschau zu halten, mit denen sich die Datenanalyse wenigstens teilweise automatisieren läßt. Diese Verfahren sind es, die sich hinter den Schlagworten „Knowledge Discovery in Databases“ und „Data Mining“ verbergen. Zwar können sie einen Johannes Kepler noch lange nicht ersetzen, aber vielleicht wäre er, unterstützt von diesen Verfahren, schneller zum Ziel gelangt.

Oft werden die Begriffe „Knowledge Discovery“ und „Data Mining“ synonym gebraucht, ich werde sie hier jedoch unterscheiden. Ich verstehe unter „Knowledge Discovery in Databases“ einen Prozeß, der mehrere Schritte umfaßt und der Entdeckung von gültigen, nutzbaren, verständlichen, unbekannten und unerwarteten Zusammenhängen in Daten dient. Ein Schritt in diesem Prozeß ist „Data Mining“. In ihm werden bestimmte Modellierungs- und Entdeckungstechniken auf bereits vorverarbeitete Daten angewandt. Die Ergebnisse dieses Schritts werden in einer weiteren Phase visualisiert, interpretiert, bewertet und in Werkzeugen (Programmen) nutzbar gemacht.

3.1 Der KDD-Prozeß

In diesem Aufsatz wird der KDD-Prozeß in zwei Vor- und fünf Hauptstufen gegliedert, doch ist diese Gliederung keineswegs verbindlich. Ein einheitliches, allgemein anerkanntes Schema liegt bislang noch nicht vor.

Vorstufen

- Bestimmung des Nutzenpotentials
- Anforderungs- und Durchführbarkeitsanalyse

Hauptstufen

- Sichtung des Datenbestandes, Datenauswahl, ggf. Datenerhebung
- Vorverarbeitung (60-80% des Aufwandes)
  - Vereinheitlichung und Transformation der Datenformate
  - Säuberung (Behandlung von Fehlern, Fehlstellen und Ausreißern)
  - Reduktion / Fokussierung (Stichproben, Attributauswahl, Prototypenbildung)
- **Data Mining** (mit verschiedenen Verfahren)
- Visualisierung (auch parallel zu Vorverarbeitung, Data Mining und Interpretation)
- Interpretation, Prüfung und Bewertung der Ergebnisse
- Anwendung und Dokumentation

In den Vorstufen soll vor allem geklärt werden, ob die Hauptstufen des KDD-Prozesses überhaupt durchlaufen werden sollten. Denn nur wenn der potentielle Nutzen hoch genug, die Kosten einer Durchführung nicht zu hoch und die Anforderungen durch Data-Mining-Verfahren erfüllbar sind, kann mit einem Nutzen gerechnet werden.

In den Hauptstufen werden zunächst die Daten, die auf verborgenes Wissen hin untersucht werden sollen, ausgewählt und in eine Form gebracht, in der Data-Mining-Verfahren angewendet werden können. Dieser Vorverarbeitungsschritt ist gewöhnlich der aufwendigste des ganzen Prozesses. Abhängig von der in der Anforderungsanalyse festgestellten Data-Mining-Aufgabe (siehe unten) werden anschließend Entdeckungstechniken eingesetzt, deren Ergebnisse zur Prüfung und Interpretation visualisiert werden können. Da sich das gewünschte Ergebnis nur selten schon nach dem ersten Versuch ergibt, müssen einige Schritte der Vorverarbeitung (z.B. die Attributauswahl) und die Anwendung der Data-Mining-Verfahren ggf. mehrfach durchlaufen werden. Spätestens hier zeigt sich, daß KDD kein völlig automatisierter, sondern ein interaktiver Prozeß ist. Der Benutzer prüft und bewertet die erzielten Ergebnisse und nimmt, wenn nötig, Anpassungen am Ablauf des KDD-Prozesses vor.

3.2 Data-Mining-Aufgaben

Im Laufe der Zeit haben sich typische Aufgaben herauskristallisiert, die Data-Mining-Verfahren lösen können sollten. Zu diesen gehören vor allem die folgenden, die ich, neben ihrer Bezeichnung, durch eine typische Fragestellung zu charakterisieren versucht habe.

- Klassifikation (classification)
  Ist dieser Kunde kreditwürdig?
- Konzeptbeschreibung (concept description)
  Welche Eigenschaften haben reparaturanfällige Fahrzeuge?
- Segmentierung (segmentation, clustering)
  Was für Kundengruppen habe ich?
- Prognose (prediction, trend analysis)
  Wie wird sich der Dollarkurs entwickeln?
- Abhängigkeitsanalyse
  (dependency/association analysis)
  Welche Produkte werden zusammen gekauft?
- Abweichungsanalyse (deviation analysis)
  Gibt es jahreszeitliche Umsatzschwankungen?

Am häufigsten sind Klassifikations- und Prognoseprobleme, da ihre Lösung unmittelbare Auswirkungen auf den Umsatz und den Gewinn eines Unternehmens haben kann. In letzter Zeit werden aber auch Abhängigkeitsanalysen immer öfter benötigt, z.B. wenn Verbundkäufe in Supermärkten untersucht werden (Warenkorbanalyse).

4 Beispiele für Data-Mining-Verfahren

Data-Mining-Verfahren stammen aus den verschiedensten Bereichen. Klassische statistische Verfahren finden ebenso Verwendung wie Entwicklungen der künstlichen Intelligenz, des maschinellen Lernens oder des sogenannten „Soft Computing“. Data Mining ist also stark interdisziplinär.

In diesem Abschnitt werden an einigen Beispielen aus der Vielzahl der verfügbaren Data-Mining-Verfahren einige Kernideen, wenn auch nur sehr knapp, erläutert. Es ist klar, daß nicht alle Verfahren berücksichtigt werden können — ihre große Zahl verbietet dies.

4.1 Entscheidungsbäume

Entscheidungsbäume [5, 18, 19, 4] sind eine der bekanntesten Formen von Klassifikatoren, d.h. von Verfahren, die einen Fall anhand seiner Eigenschaften einer Klasse zuordnen. Die Entscheidung über

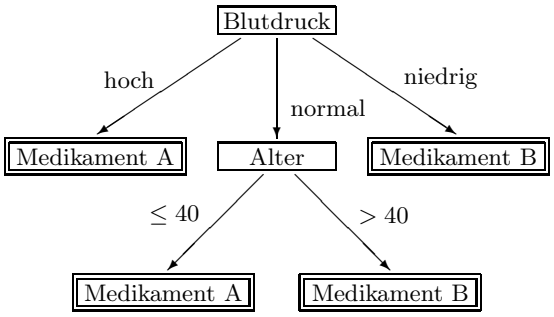

Abbildung 1: Ein Entscheidungsbaum zur Bestimmung eines Medikamentes. Zuerst wird der Blutdruck getestet. Ist der Wert hoch oder niedrig, kann das richtige Medikament sofort angegeben werden. Ist der Blutdruck normal, wird das Alter des Patienten geprüft, wodurch auch in diesem Fall das richtige Medikament bestimmt werden kann.

die Klassenzugehörigkeit eines Falles wird durch das Durchlaufen eines Baumes gefällt, wobei an jeder Verzweigung eine Eigenschaft des zu klassifizierenden Falles geprüft wird. Es wird dann derjenige Zweig verfolgt, der dem festgestellten Wert der Eigenschaft zugeordnet ist. Dies geschieht so lange, bis ein Blatt erreicht wird, das die Klasse angibt. Ein Beispiel, in dem ein Medikament aus Patientenmerkmalen bestimmt wird, zeigt Abbildung 1.

Liegt ein Datensatz bereits klassifizierter Fälle vor, so lassen sich Entscheidungsbäume leicht automatisch konstruieren. Entscheidungsbaumlernprogramme nehmen dazu eine schrittweise Unterteilung der Fälle nach bestimmten Kriterien vor, bis eine möglichst gute Klassifizierung erreicht ist. Ein Beispiel soll dies illustrieren. Gegeben sei der in Tabelle 1 gezeigte Datensatz, das wirksame Medikament soll vorhergesagt werden. Ein Entscheidungsbaumlernprogramm unterteilt die Fälle wie in Tabelle 2 gezeigt, was anhand von Blutdruck und Alter eine perfekte Vorhersage ermöglicht. Der zugehörige Entscheidungsbaum ist in Abbildung 1 dargestellt.

No	Geschecht	Alter	Blutdruck	Med.
1	männlich	20	normal	A
2	weiblich	73	normal	B
3	weiblich	37	hoch	A
4	männlich	33	niedrig	B
5	weiblich	48	hoch	A
6	männlich	29	normal	A
7	weiblich	52	normal	B
8	männlich	42	niedrig	B
9	männlich	61	normal	B
10	weiblich	30	normal	A
11	weiblich	26	niedrig	B
12	männlich	54	hoch	A

Tabelle 1: Patientendaten zusammen mit einem wirksamen Medikament (wirksam in bezug auf eine nicht näher spezifizierte Krankheit). Es fällt schwer, direkt aus dieser Tabelle eine Regel für die Vorhersage des richtigen Medikamentes abzulesen.

No	Blutdruck	Alter	Med.
3	hoch	37	A
5	hoch	48	A
12	hoch	54	A
1	normal	20	A
6	normal	29	A
10	normal	30	A
7	normal	52	B
9	normal	61	B
2	normal	73	B
11	niedrig	26	B
4	niedrig	33	B
8	niedrig	42	B

Tabelle 2: Durch ein Entscheidungsbaumlernprogramm vorgenommene Aufteilung der Fälle aus Tabelle 1. Blutdruck und Alter des Patienten bestimmen zusammen das wirksame Medikament.

4.2 Schlußfolgerungsnetze

Schlußfolgerungsnetze haben sich aus der Wahrscheinlichkeitstheorie als Bayessche Netze (nach Thomas Bayes, 1702–1761) [17] und Markovnetze (nach Andrej Andrejewitsch Markov, 1856–1922) [13] entwickelt, heute werden aber auch andere Unsicherheitskalküle als die Wahrscheinlichkeitstheorie zum Aufbau von Schlußfolgerungsnetzwerken herangezogen [8, 3]. Schlußfolgerungsnetze eignen sich besonders für die Darstellung von Abhängigkeiten mehrerer Merkmale. Welche Mermale voneinander abhängen wird durch ein Netzwerk (Hypergraph) dargestellt. Die Details der Abhängigkeiten werden z.B. durch Wahrscheinlichkeits- oder Possibilitätsverteilungen beschrieben, die den Verbindungen in diesem Netzwerk zugeordnet sind.

Durch das automatisierte Lernen von Schlußfolgerungsnetzen aus Daten [6, 9, 3] läßt sich eine Abhängigkeitsanalyse durchführen. In Zusammenarbeit mit der Daimler-Benz Forschungsabteilung Ulm konnte ich eine solche Analyse auf einer Mercedes-Benz-Fahrzeugdatenbank

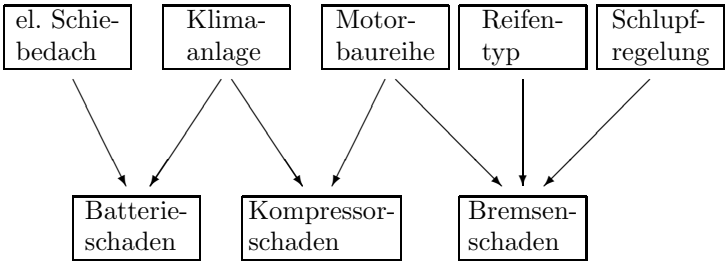

Abbildung 2: Ein Ausschnitt eines fiktiven zweischichtigen Schlußfolgerungsnetzes, das die Abhängigkeiten zwischen Schäden/Fehlern (untere Schicht) und Bauzustandsmerkmalen (obere Schicht) beschreibt. Übereinstimmungen mit tatsächlichen Abhängigkeiten sind rein zufällig.

(fiktive) Häufigkeit von Batterieschäden		Klimaanlage mit	Klimaanlage ohne
elektrisches Schiebedach	mit	9 %	3 %
	ohne	3 %	2 %

Tabelle 3: Ein fiktives Teilnetz, das die Abhängigkeit eines Batterieschadens vom Vorhandensein eines elektrischen Schiebedaches und einer Klimaanlage beschreibt.

durchführen, um nach Abhängigkeiten zwischen den Ausstattungsmerkmalen eines Fahrzeugs und Schäden bzw. Fehlern zu suchen. Ein fiktives Beispiel für das Ergebnis einer solchen Analyse (echte Ergebnisse unterliegen selbstverständlich der Geheimhaltung) ist in Abbildung 2 und Tabelle 3 gezeigt. Tabelle 3 würde darauf hindeuten, daß die Batterie der erhöhten Belastung durch die Sonderausstattungen Klimaanlage und elektr. Schiebedach nicht ganz gewachsen ist.

Lernprogramme für Schlußfolgerungsnetze haben viel mit Entscheidungsbaumlernprogrammen gemeinsam. Auch sie teilen die Daten nach einigen ihrer Merkmalen auf, jedoch etwas anders als Entscheidungsbäume. Durch diese Aufteilung versuchen sie, die Unterschiede in den Werthäufigkeiten anderer Merkmale zu maximieren.

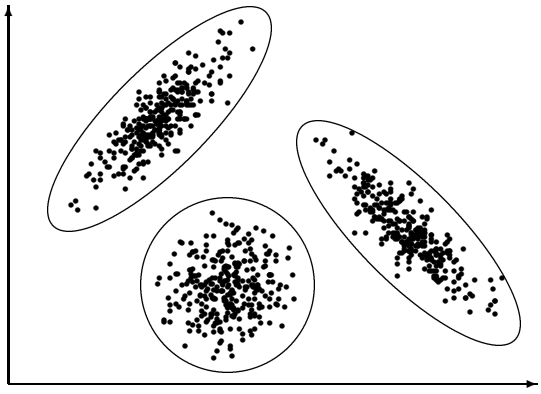

Abbildung 3: Eine Einteilung zweidimensionaler Daten in Cluster. Der Kreis und die Ellipsen deuten an, welche Punkte zu einer Gruppe zusammengefaßt wurden.

4.3 Clusteranalyse

Durch eine Clusteranalyse [2, 22, 21, 14] werden Fälle zu Gruppen zusammengefaßt. Das Ziel ist, daß zwei Fälle aus einer Gruppe sich möglichst ähnlich sind, und zwei Fälle aus zwei verschiedenen Gruppen möglichst unterschiedlich. Gelingt eine solche Einteilung in Gruppen, so kann man oft alle Fälle einer Gruppe wie einen Fall behandeln (Prototypenbildung), oder man kann versuchen, die die Gruppen unterscheidenden Merkmale zu bestimmen (Konzeptbeschreibung). Ein Beispiel für eine Clustereinteilung zeigt Abbildung 3.

Ein häufig verwendetes Verfahren zur Clusteranalyse ist das hierarchische Clustern. Bei diesem Verfahren bildet man zunächst Ein-Fall-Cluster, die dann schrittweise vergrößert werden, indem man stets die beiden am dichtesten zusammenliegenden Cluster vereinigt. Wenn nur noch eine bestimmte vorgegebene Zahl von Clustern übrig ist, wird das Verfahren abgebrochen. Es ist klar, daß das Ergebnis des Verfahrens stark von dem verwendeten Abstandsmaß abhängt, da dieses ja bestimmt, welche Cluster vereinigt werden.

Eine Erweiterung der klassischen Clusteranalyse ist die Fuzzy-Clusteranalyse [10], die mit graduellen Zugehörigkeiten zu einem Cluster arbeitet. D.h., es ist erlaubt, daß ein Fall zu zwei Clustern gleichzeitig gehört, jedoch nur z.B. mit 50%. Graduelle und Mehrfachzugehörigkeiten bieten besonders in der Bildverarbeitung Vorteile. Im Gegensatz zum hierarchischen Clustern wird bei der Fuzzy-Clusteranalyse eine Funktion optimiert.

4.4 Neuronale Netze

Neuronale Netze [1, 20, 16] wurden ursprünglich den Gehirnstrukturen von Lebewesen nachempfunden. Ihre Entwicklung ging jedoch bald eigene We-

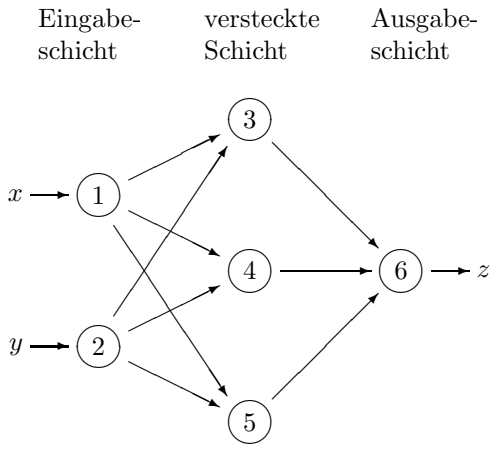

Abbildung 4: Ein einfaches neuronales Netz aus sechs Neuronen mit Eingabe-, Ausgabe- und versteckter Schicht.

ge, so daß man heute nur noch von einer entfernten Analogie zum biologischen Ursprung sprechen kann. Neuronale Netze sind aus kleinen Recheneinheiten, den sogenannten Neuronen, aufgebaut, die über gewichtete Verbindungen miteinander verschaltet werden. Jedes Neuron berechnet die gewichtete Summe seiner Eingänge und aus dieser über eine festzulegende Funktion einen Aktivierungsgrad, der die Ausgabe des Neurons darstellt. Die Ausgabe eines Neurons kann die Eingabe eines anderen Neurons sein. So lassen sich aus einfachen Einheiten komplexe Strukturen aufbauen, die z.B. in der Lage sind, beliebige Funktionen zu approximieren.

Neuronale Netze sind gewöhnlich in Schichten organisiert. Verbindungen gibt es dann nur zwischen Neuronen zweier benachbarter Schichten. Es gibt eine Eingabeschicht, eine Ausgabeschicht und gewöhnlich eine oder mehrere „versteckte“ Schichten (siehe Abbildung 4). Eine solche Struktur kann leicht auf die Ausgabe bestimmter Werte bei Anliegen bestimmter Eingangswerte „trainiert“ werden. Die am häufigstens verwendete Trainingsmethode ist die sogenannte Rückpropagation (back propagation), bei der der Fehler, d.i. die Abweichung des gelieferten vom gewünschten Ausgabewert, von den Ausgabeneuronen durch die versteckten Schichten in Richtung auf die Eingabeneuronen weitergegeben wird. Auf diesem Wege werden die Verbindungsgewichte so angepaßt, daß sich der Fehler verringert.

Neuronale Netze haben den großen Nachteil, daß sie zwar sehr mächtig sind (mit speziellen Netzstrukturen läßt sich sogar eine Clusteranalyse durchführen), ein trainiertes Netz jedoch eine „black box“ darstellt, von der man nicht verstehen kann, wie sie zu ihren Ergebnissen gelangt. Eine Alternative bieten Neuro-Fuzzy-Systeme [16], in denen man versucht, die Vorteile von neuronalen Netzen mit der Verständlichkeit von Fuzzy-Systemen zu kombinieren.

4.5 Weitere Ansätze

Es ist klar, daß in einem so kurzen Aufsatz wie diesem nicht alle bekannten Verfahren besprochen werden können. Einige weitere sollen aber wenigstens erwähnt werden.

- klassische statistische Verfahren
  - Diskriminanzanalyse
  - Regressionsanalyse
  - Hauptkomponentenanalyse
  - Faktorenanalyse
  - Zeitreihenanalyse
  - k-nearest neighbour
- Verfahren des maschinellen Lernens
  - induktive logische Programmierung
  - conceptual clustering
  - instance based learning
- evolutionäre/genetische Algorithmen

Von diesen Verfahren verdienen sicherlich die statistischen die größte Aufmerksamkeit, denn sie haben eine gesicherte mathematische Grundlage — im Gegensatz zu den meisten anderen Verfahren (z.B. Entscheidungsbaumlernern), die nur Heuristiken sind.

5 Zusammenfassung

Daten allein sind noch kein Wissen, aber in Daten kann wichtiges Wissen verborgen sein. Dieses Wissen zu finden und nutzbar zu machen ist jedoch nicht leicht. Manuelle Analysen sind bei den heute zu verarbeitenden Datenmengen undurchführbar, so daß der Mensch durch intelligente Computerprogramme unterstützt werden muß. Zwar können diese Programme den Menschen noch lange nicht ersetzen, noch muß er ihren Einsatz steuern, aber sie können ihm eine wertvolle Hilfestellung leisten. Dennoch ist Vorsicht geboten. Schwere Fehler können auftreten, wenn komplexe Methoden nicht richtig angewandt oder vom Anwender nicht hinreichend verstanden werden.

Literatur

[1] J.A. Anderson. *An Introduction to Neural Networks*. MIT Press, Cambridge, MA, 1995

[2] H.H. Bock. *Automatische Klassifikation (Cluster-Analyse)*. Vandenhoek & Ruprecht, Göttingen, 1974

[3] C. Borgelt and R. Kruse. Evaluation Measures for Learning Probabilistic and Possibilistic Networks. *Proc. 6th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE'97)*, Vol. 2:pp. 1034–1038, Barcelona, Spain, 1997

[4] C. Borgelt und R. Kruse. Attributauswahlmaße für die Induktion von Entscheidungsbäumen: Ein Überblick. In: G. Nakhaeizadeh, ed. *Data Mining: Theoretische Aspekte und Anwendungen* pp. 77-98, Physica-Verlag, Heidelberg, 1998

[5] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*, Wadsworth International, Belmont, CA, 1984

[6] G.F. Cooper and E. Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning* 9:309–347, Kluwer, 1992

[7] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds. *Advances in Knowledge Discovery and Data Mining*. AAAI Press / MIT Press, Cambridge, MA, 1996

[8] J. Gebhardt and R. Kruse. POSSINFER — A Software Tool for Possibilistic Inference. In: D. Dubois, H. Prade, and R. Yager, eds. *Fuzzy Set Methods in Information Engineering: A Guided Tour of Applications*, Wiley, 1995

[9] D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning* 20:197–243, Kluwer, 1995

[10] F. Höppner, F. Klawonn und R. Kruse. *Fuzzy-Clusteranalyse: Verfahren für die Bilderkennung*. Vieweg, Wiesbaden, 1997

[11] R. Kruse and D. Meyer. *Statistics with Vague Data*. Reidel, Dordrecht, 1987

[12] R. Kruse, J. Gebhardt, and F. Klawonn. *Fuzzy-Systeme, 2. erweiterte Auflage*. Teubner, Stuttgart, 1995.

[13] S.L. Lauritzen and D.J. Spiegelhalter. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society, Series B*, 2(50):157–224, 1988

[14] H.-J. Mucha. Clusteranalyse mit Mikrocomputern. Akademie-Verlag, Berlin, 1992

[15] G. Nakhaeizadeh. *Data Mining: Theoretische Aspekte und Anwendungen*. Physica-Verlag, Heidelberg, 1998

[16] D. Nauck, F. Klawonn und R. Kruse. *Neuronale Netze und Fuzzy-Systeme, 2. erweiterte Auflage*. Vieweg, Wiesbaden, 1996.

[17] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (2nd edition)*. Morgan Kaufman, New York 1992

[18] J.R. Quinlan. Induction of Decision Trees. *Machine Learning* 1:81–106, 1986

[19] J.R. Quinlan. *C4.5: Programs for Machine Learning*, Morgan Kaufman, 1993

[20] R. Rojas. *Theorie der Neuronalen Netze: Eine systematische Einführung*. Springer, Berlin, 1993.

[21] H. Späth. Cluster-Formation und Analyse. Oldenbourg, München, 1983

[22] D. Steinhausen und K. Langer. *Clusteranalyse: Einführung in Methoden und Verfahren der automatischen Klassifikation*. Walter de Gruyter, Berlin, 1977

[23] S.M. Weiss and N. Indurkhya. *Predictive Data Mining*. Morgan Kaufman, San Francisco, CA, 1997