Grundbegriffe des Data Mining aufbereitet für eine Datenbank-Vorlesung

Von

Christian Ulrich

Inhaltsverzeichnis:

Abbildungsverzeichnis

1 Einleitung

In den letzten Jahren vollzog sich in der Unternehmensphilosophie ein Wandel von der althergebrachten Optimierung von Teilbereichen weg, hin zu Geschäftsprozessoptimierung. Diese berücksichtigt eine ganzheitliche Sicht des Unternehmens. In Anbetracht dieses Wandels rückt der Kunde und sein individueller Wert immer mehr in den Fokus. Um Wettbewerbsvorteile zu erhalten, muss man das Verhalten der Kunden analysieren, damit man daraus lernen und den Kunden enger an das eigene Unternehmen binden kann.
Die Informationen, die zu diesem Zweck gesammelt werden, erhöhen die ohnehin schon große Datenmenge, die ein Unternehmen täglich speichert, enorm. Dies sind zum Beispiel Informationen wie Namen, Preise, Adressen, Kaufinformationen oder Kommunikationsdaten.
Täglich speichern allein wissenschaftliche Organisationen ungefähr ein Terabyte an Daten. Und dies ist nur die Spitze des Eisbergs. Und es ist bekannt, dass die Wissenschaft nicht die meisten Daten speichert, sondern die Unternehmen. Die Mengen an gesammelten Daten sind so ungeheuer groß, dass sie von Menschen allein nicht mehr vernünftig verarbeitet werden können. Ebenso kann der Mensch kaum Zusammenhänge in diesen Datenmengen erkennen. Gerade aber solche Zusammenhänge können tiefe Einsichten in das Kundenverhalten vermitteln.
So ist es ohne spezielle Algorithmen fast nicht möglich, Fragen zu beantworten, die nur durch Zusammenhänge von verschiedenen Daten beantwortet werden können. Beispiele dafür wären:

Welche Güter biete ich einem bestimmten Kunden an?
Wie groß ist die Wahrscheinlichkeit, dass ein Kunde seinen Vertrag kündigt?
Was wird der Trend in der nächsten Saison?
Welche Waren werden von welchen Kunden typischerweise zusammen eingekauft?

Eine Analyse aus Kanada (Alpar, P, Data Mining im praktischen Einsatz, 2000, 8) zeigte zum Beispiel auf, dass Babywindeln und Bier häufig zusammen gekauft wurden. Die gegebene Erklärung lautet, dass junge Väter bei ihrem Biereinkauf oft den Auftrag erhalten, auch Windeln zu besorgen. Dieses Wissen wurde mit Hilfe des Data Mining erhoben. Das Data Mining fasst eine Reihe von unterschiedlichen Verfahren und Methoden zusammen, die es erlauben, datenverarbeitungsgestützte Algorithmen einzusetzen, welche selbständig große Datenbestände auf Zusammenhänge hin analysieren und dabei Korrelationen in diesen Datenbeständen entdecken, die bislang keine Berücksichtigung fanden (Krahl, Data Mining, 1998, 11). Solche Zusammenhänge wurden früher von speziell ausgebildeten Statistikern untersucht. Sie versuchten, Fragen durch herkömmliche statistische Verfahren die gesuchten Informationen aus den Datenbeständen zu extrahieren. Dies gelang nicht immer und dauerte meist mehrere Monate.
Der Einsatz von Data Mining beschleunigt und vereinfacht diesen Prozess und stellt so schneller neue entscheidungsrelevante Informationen zur Verfügung, mit denen sich die Wettbewerbsfähigkeit von Unternehmen stark verbessern lässt.
Allerdings ist Data Mining auch kein Allheilmittel, das heißt, es lassen sich nur Zusammenhänge finden, die auch in den erhobenen Daten enthalten sind.

Zielsetzung

Diese Arbeit soll die Grundlagen des Data Mining so erklären, dass sie für Studenten der Informatik und Wirtschaftsinformatik gleichermaßen verständlich sind. Die meisten Bücher zum Thema sind eher formal-mathematisch gestaltet und beinhalten selten verständliche Beispiele. So soll es das Ziel der Arbeit sein, den Studenten des Fachs Datenbanken die wichtigsten Grundlagen und einige ausgesuchte Algorithmen näher zubringen, ohne dass sie dazu einen Doktortitel in Statistik brauchen. Die Algorithmen sollen dann durch einfache, nachvollziehbare Beispiele eingeübt werden, damit dadurch auf dieser Grundlage das Lernziel überprüfbar ist. Der Schwerpunkt wird dabei auf den Verfahren „Entscheidungsbäume“ und „genetische Algorithmen“ liegen, die jeweils mit den grundlegenden Algorithmen aufgezeigt werden. Weiterhin sollen die Studenten die Wichtigkeit der Vorauswahl der Daten und die Einschränkungen des Data Mining kennen lernen.

Aufbau

In Kapitel 2 soll zuerst Allgemeines über das Data Mining erklärt werden. Hierbei halte ich mich im Groben an folgende Bücher: (Alpar, P: Data Mining im praktischen Einsatz, Braunschweig, 2000), (Chamoni, P, Gluchowski, P: Analytische Informationssysteme – Data Warehouse, On-Line Analytical Processing, Data Mining, Berlin, Heidelberg, 1998), (Ester, M, Sander, J: Knowledge Discovery in Databases, Techniken und Anwendungen, Berlin, 2000), (Fayyad, U.M.: Advances in knowledge discovery and data mining, Menlo-Park, Calif., 1996). Unter anderem werden, nach einer allgemeinen Einleitung über den Hintergrund des Data Mining, die Begriffe Induktion/Deduktion erklärt. Ebenso enthält dieses Kapitel eine Beschreibung der bei der Datenauswahl zu berücksichtigenden Kriterien und eine Erklärung der 3 Phasen des Data Mining.

In Kapitel 3 wird zuerst ein allgemeiner Überblick über verschiedene Data-Mining-Algorithmen gegeben, und danach folgt ein konkretes Beispiel. Hierbei beziehe ich mich auf: (Weigand, D.: Lernen mit Entscheidungsbäumen, Elektronische Publikation, URL am 23.02.01: http//www2.informatik.uni-erlangen.de/IMMD-II/Lehre/WS98_99/Machine_Learning/
Vortraege/Entscheidungsbaeume/Entscheidungsbaeume.pdf), (Krahl,Daniela: Data Mining – Einsatz in der Praxis, Bonn, 1998), (Chamoni, P, Gluchowski, P: Analytische Informationssysteme – Data Warehouse, On-Line Analytical Processing, Data Mining, Berlin, Heidelberg, 1998), (Alpar, P: Data Mining im praktischen Einsatz, Braunschweig, 2000)

Kapitel 4 zeigt eine praktische Anwendung des Data Mining. Hier beziehe ich mich auf: (Alpar, P: Data Mining im praktischen Einsatz, Braunschweig, 2000)

Am Schluss der Arbeit wird im Fazit (Kapitel5) das Data Mining kritisch beurteilt.

In einem Abschlussparagraphen zu jedem Kapitel wird kurz auf den pädagogischen Hintergrund der gewählten Erklärungen und Darstellungen eingegangen. Zu jedem Kapitel gibt es zusätzlich Folien in englischer Sprache, die bei einem Vortrag verwendet werden können.

Der Anhang enthält Übungsaufgaben und die dazugehörigen Lösungen.

2 Grundlagen des Data Mining

2.1 Einleitung

„Data Mining lässt sich mit „Schürfen oder Graben in Daten“ übersetzen, wobei das Ziel, nach dem gegraben wird, Informationen beziehungsweise Wissen ist. Wissen entspricht heute dem Gold, nach dem früher gegraben wurde, denn Unternehmen können daraus Umsätze und Gewinne generieren. (Alpar, P.: Data Mining im praktischen Einsatz, 2000, 3)

Der Begriff “Data Mining” kann als Synonym für Begriffe wie “Datenmustererkennung”, “Database exploration”, oder auch „Knowledge Discovery in Databases“ (KDD) aufgefasst werden (Alpar, P.: Data Mining im praktischen Einsatz, 2000, 3).

Die Ursprünge des Data Mining liegen in der Statistik, in der Datenbeziehungen analysiert werden, und in der Forschung zu Datenbankmanagementsystemen, wo man sich mit der Behandlung großer Datenbestände beschäftigt. In beiden Fällen dachte man dabei hauptsächlich an Algorithmen und Computerprogramme, mit denen die Beziehungen zwischen den betrachteten Daten, die Datenmuster, ermittelt werden konnten. Nach Fayyad (Fayyad; Advances in knowledge discovery and data mining, 1996, 6) ist Data Mining daher folgendermaßen definiert: „Data Mining ist die Anwendung spezifischer Algorithmen zur Extraktion von Mustern aus Daten.“

In der Statistik geht man meistens so vor, dass zuerst Hypothesen über Datenzusammenhänge aufgestellt werden, die dann mit Hilfe der Daten und Algorithmen entweder bestätigt oder verworfen werden. In den achtziger Jahren begannen Forscher aus dem Bereich der künstlichen Intelligenz, Algorithmen zu entwickeln, die umgekehrt vorgingen. Aus Daten sollten Hypothesen berechnet werden, die neu und interessant sind. Dieser so automatisierten Hypothesenfindung muss eine Überprüfung und Interpretation folgen, bevor Handlungsalternativen ausgearbeitet werden können. Bevor mit irgendwelchen Daten gerechnet wird, müssen die relevanten Objekte oder Merkmalsträger sowie ihre Merkmale

ausgewählt werden. Die Berechnungen stellen also nur einen Schritt im gesamten Prozess der Erkennung von Datenmustern dar. Dieser Prozess kann folgendermaßen definiert werden: „Wissensentdeckung in Datenbanken ist der nicht-triviale Prozess der Identifizierung valider, neuer, potentiell nützlicher und schließlich verständlicher Muster in Daten“(Fayyad, Advances in knowledge discovery and data mining. 1996, 6).

„Nicht-trivial“ bedeutet, dass ein Such- oder Schlussfolgerungsalgorithmus zur Anwendung kommt, womit man Data Mining von reinen Datenbankabfragen oder einfachen statistischen Auswertungen unterscheidet.
Die Forderung nach Validität besagt, dass die Gültigkeit der Datenmuster über die zugrunde liegenden Daten hinaus überprüft werden muss. Bei einem großen Datenbestand ist es sinnvoll, die Gültigkeit der in einer Stichprobe gefundenen Muster in anderen Stichproben zu überprüfen.
Die Forderungen nach neuen, potentiell nützlichen und verständlichen Mustern sind pragmatischer Natur und unmittelbar verständlich. Wenn eine Analyse von Kreditkartentransaktionen ergeben würde, dass das Hauptmerkmal von Kunden von Damenboutiquen das Geschlecht ist, wäre das ein verständliches, aber kaum ein neues und potentiell nützliches Datenmuster.
Abschließend lässt sich sagen, dass drei Gebiete hauptsächlich zum Data-Mining-Prozess beitragen. Die Statistik liefert Methoden zur Datenexploration, -auswahl und –transformation, zur Mustererkennung inklusive Validierung und zur Beschreibung und Visualisierung der Ergebnisse. Die Datenbankforschung stellt Methoden und Werkzeuge zur Verfügung, um die untersuchten Daten effizient zu speichern, wiederzugewinnen und auf Plausibilität und Integrität zu prüfen. Die künstliche Intelligenz liefert hauptsächlich weitere Verfahren für das eigentliche Data Mining, zum Beispiel genetische Algorithmen, künstliche neuronale Netze oder maschinelles Lernen.

2.2 Der Data-Mining-Prozess

Wie bereits erwähnt, sollte, bevor der Data-Mining-Prozess gestartet wird, Klarheit über die Ziele bestehen, die damit verfolgt werden.
Der Data-Mining-Prozess kann in drei Schritte zerlegt werden. Der erste Schritt, die Vorbereitung, besteht aus der Auswahl, der Vorbereitung und der Transformation der Daten. Der zweite Schritt ist das eigentliche Data Mining und der dritte Schritt beinhaltet die Interpretation und Evaluation der Daten. Der erste Schritt ist hochkomplex und oft gar nicht algorithmisierbar, daher nimmt er 75-85% der Gesamtanstrengungen in Anspruch.

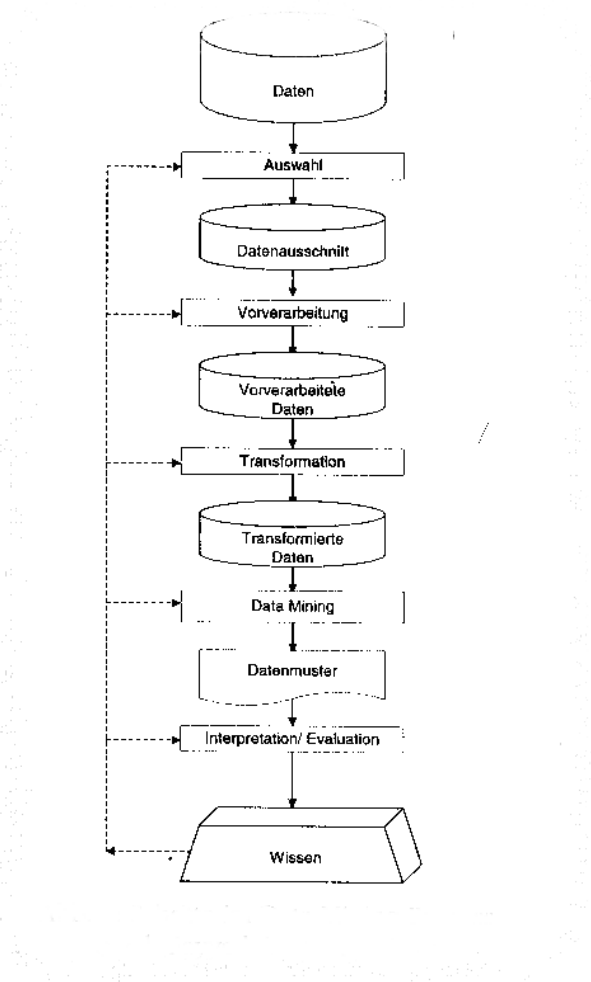


Abbildung 1: Schritte im Data Mining Prozess

Schritt 1: Vorbereitung

1) Auswahl der Daten

Im ersten Schritt geht es darum, ein Verständnis des Anwendungsbereichs und des bereits bekannten Anwendungswissens zu gewinnen. Darauf aufbauend wird das Ziel des Data Minings aus der Sicht des gegebenen Anwendungsbereichs definiert, denn das gewünschte Wissen soll ja bisher unbekannt und nützlich für den Anwendungsbereich sein. Es muss ferner festgelegt werden, in welchen Daten das Wissen gesucht werden soll und wie diese Daten zu beschaffen sind. Im einfachsten Fall kann man auf eine vorhandene Datenbank zurückgreifen und einen Teil davon als Grundlage für das Data Mining selektieren. Andernfalls müssen die Daten erst durch Messungen, Fragebögen oder ähnliche Methoden erhoben werden. Bei einem sehr großen Datenbestand reicht es oft aus, Data Mining in einer Stichprobe vorzunehmen. Damit die Stichprobe repräsentativ für den Gesamtbestand ist, müsste vor ihrer Ziehung eine Untersuchung der Verteilung der Werte der relevanten Datenfelder vorgenommen werden.
(Näheres zur Auswahl der Daten in Kapitel 2.4)

2) Vorverarbeitung

Ziel der Vorverarbeitung ist es, die benötigten Daten zu reinigen, also zu vervollständigen, konsistent zu machen und zu integrieren. Daten aus verschiedenen Quellen müssen integriert werden, da sie im Allgemeinen nach unterschiedlichen Konventionen gewonnen wurden. Verschiedene Abteilungen einer Firma benutzen zum Beispiel häufig verschiedene Namen für dieselben Attribute eines Objekts oder sammeln die Daten über unterschiedliche Zeiträume hinweg. In einer Abteilung könnte zum Beispiel der Umsatz tageweise gespeichert werden, während dieselbe Information in einer anderen Abteilung wochenweise gesammelt wird.

Inkonsistenzen in den Daten wie etwa verschiedene Werte desselben Attributs oder Schreibfehler für Namen treten häufig auf und müssen aufgelöst werden.
In realen Datenbanken fehlt meist ein signifikanter Teil aller Attributwerte: Es kann zum Beispiel ein Messfehler aufgetreten sein oder einige Fragen in einem Fragebogen wurden absichtlich nicht beantwortet. Je nach verwendetem Data-Mining-Algorithmus kann es notwendig sein, fehlende Attributwerte genauer zu spezifizieren, da diese Information für das Data Mining wichtig sein könnte. Man könnte zum Beispiel zwischen „Messung nicht durchgeführt“ und „Messfehler aufgetreten“ unterscheiden. In einer medizinischen Anwendung kann die Tatsache, dass ein bestimmter Test durchgeführt wurde, ausschlaggebend sein für die Klassifikation eines bestimmten Patienten.

3) Transformation

In diesem Schritt werden die vorverarbeiteten Daten in eine für das Ziel des Data Mining geeignete Form transformiert. Typische Transformationen sind die Attribut-Selektion und die Diskretisierung von Attributen, auf die im Folgenden näher eingegangen werden soll.
Im Allgemeinen sind nicht alle bekannten Attribute der Daten relevant für die Data-Mining-Aufgabe. Obwohl viele Algorithmen eine eigene Auswahl der relevanten Attribute vornehmen, kann eine zu große Anzahl von Attributen die Effizienz des Data Mining deutlich verschlechtern. Eine Attribut-Selektion ist also häufig in praktischen Anwendungen nötig. Wenn genügend Wissen über die Bedeutung der Attribute und über die gegebene Data-Mining-Aufgabe bekannt ist, kann dieses Wissen zu einer manuellen Attribut-Selektion genützt werden. Andernfalls muss eine automatische Attribut-Selektion durchgeführt werden. Ein optimaler Algorithmus, der alle Teilmengen betrachtet, ist hierfür zu aufwendig. Daher kommen dann heuristische Algorithmen zum Einsatz. Dazu verwendet man Clustering (Siehe Kapitel 3.1.2).
Manche Data-Mining-Algorithmen können keine kategorischen, sondern nur numerische Attribute verarbeiten, so dass eine Diskretisierung kategorischer Attribute erforderlich wird, das heißt eine Transformation in numerische Attribute. Einfache Verfahren teilen den Wertebereich eines Attributs in Intervalle gleicher Länge oder in Intervalle mit gleicher Häufigkeit von enthaltenen Attributwerten.

Ein kurzes Beispiel dazu: Ein Programm soll eine Heizung regulieren. Wenn es im Zimmer sehr kalt ist, soll geheizt werden, bis es angenehm warm ist. Wenn es zu heiß ist, soll die Heizung ausschalten, bis es wieder angenehm ist. Mit solchen Aussagen kann weder ein Computer noch ein Data-Mining-Algorithmus etwas anfangen. Deswegen werden jedem Ausdruck Werte zugewiesen. Zu kalt sei zum Beispiel alles kleiner 16 Grad.
Komplexere Verfahren berücksichtigen die unter Umständen bekannte Klassenzugehörigkeit der Daten und bilden Intervalle, so dass gewisse Maße wie der Informationsgewinn in Bezug auf die Klassenzugehörigkeit maximiert werden. In diesem Fall werden Attributwerte von Objekten derselben Klasse möglichst demselben Intervall zugeordnet.

Schritt 2: Data Mining

Data Mining ist die Anwendung effizienter Algorithmen, die die in einer Datenbank enthaltenen gültigen Muster finden. In diesem Schritt wird zuerst die relevante Data-Mining-Aufgabe identifiziert. Die wichtigsten Aufgaben werden im Folgenden kurz erläutert:

Entdecken von Ausreißern / Clustering
Ziel des Clustering ist die Aufteilung einer Datenbank in Gruppen (Cluster) von Objekten, so dass Objekte eines Clusters möglichst ähnlich, Objekte verschiedener Cluster möglichst unähnlich sind. Ausreißer sind Objekte, die zu keinem der gefundenen Cluster gehören.

Klassifikationstraining
Gegeben sind hier Trainingsobjekte mit Attributwerten, die bereits einer Klasse zugeordnet sind. Es soll eine Funktion gelernt werden, die unbekannte Objekte aufgrund ihrer Attributwerte einer der Klassen zuweist.
Z.B. Die automatische Flaschenerkennung in einem Getränkeautomaten. Sie erkennt, welcher Kategorie eine Flasche angehört und gibt dann den entsprechenden Pfandbetrag aus.

Generalisierung

Ziel der Generalisierung ist es, eine Menge von Daten möglichst kompakt zu beschreiben, indem die Attributwerte generalisiert und die Zahl der Datensätze reduziert wird. Das heißt, nur die wichtigen Werte werden übernommen.

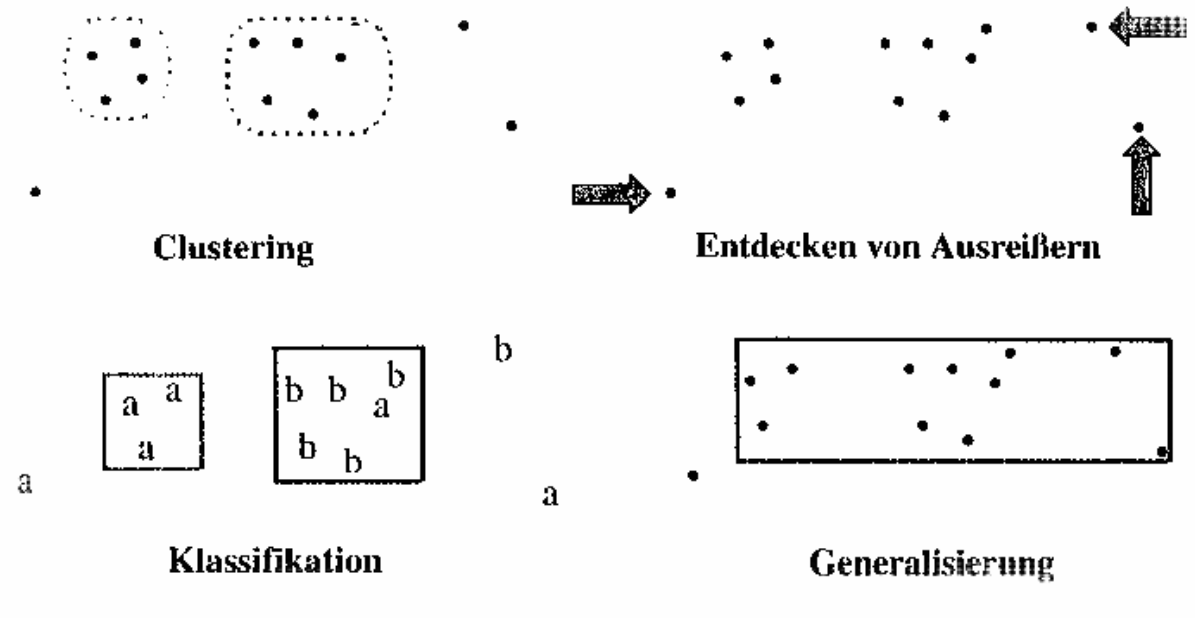


Abbildung 2: Die wichtigsten Data Mining Aufgaben

Aufgrund der Ziele der Anwendung und des Typs der Daten wird dann ein geeigneter Algorithmus ausgewählt.

Schritt 3: Interpretation / Evaluation

Im letzten Schritt werden die gefundenen Muster vom IT-System geeignet präsentiert und von Experten des Anwendungsbereichs in Bezug auf die definierten Ziele evaluiert. Falls die Ziele nach Einschätzung der Experten noch nicht erreicht sind, wird eine weitere Iteration des Data-Mining-Prozesses gestartet. Dieser neue Prozess kann bei einem beliebigen Schritt einsetzen, zum Beispiel beim Data Mining

oder der Vorverarbeitung. Sobald die Evaluation erfolgreich ist, wird das gefundene Wissen dokumentiert und in das bestehende IT-System integriert, zum Beispiel als Ausgangspunkt für zukünftige Data-Mining-Prozesse, die dann nur neues Wissen liefern. Grundlage der Evaluation ist eine geeignete Präsentation der gefundenen Muster durch das System. Eine solche Darstellung ist in vielen Anwendungsbereichen eine Herausforderung für sich, wenn sehr viele Muster gefunden wurden oder wenn die Daten sehr viele Attribute besitzen. Dieser Fall tritt oft beim Suchen von Assoziationsregeln auf. Häufig ist eine Visualisierung der gefundenen Muster für den Benutzer verständlicher als reine Textform. Gut dafür geeignet sind zum Beispiel Entscheidungsbäume.
Häufig ist es Ziel des Data Mining, mit Hilfe aus den vorhandenen Daten abgeleiteter Regeln Vorhersagen für die Zukunft zu treffen. Zentrale Aufgabe der Evaluation ist daher die Einschätzung der Vorhersagekraft, das heißt, man muss schätzen, wie gut die aus den vorhandenen Stichproben abgeleiteten Hypothesen sich auf zukünftige Daten verallgemeinern lassen. Die Stichprobendaten stellen eine Auswahl aus allen bisherigen und künftigen gesammelten Daten dar. Die Ergebnisse des Data Mining werden umso größere Vorhersagekraft besitzen, je größer die Stichprobe ist und je repräsentativer sie ist. Die Evaluation wird vereinfacht, wenn gewisse Annahmen über die Verteilung der Daten getroffen werden können, die eine Anwendung statistischer Tests erlauben. Sonst kann man aus den vorhandenen Daten eine Menge für die Stichprobe hernehmen und durch eine Überprüfung mit dem Rest die Hypothesen bestätigen.

2.3 Data Mining und Data Warehousing

Wie bereits erwähnt, nimmt der erste Schritt des Data Mining ungefähr 85% der Gesamtanstrengung in Anspruch. Dies lässt sich aber ein wenig reduzieren, wenn man auf eine konsistente, qualitativ hochwertige Datenbasis zurückgreifen kann. Eine Datenbasis ist eine strukturierte Sammlung von Daten, die dann zur Auswahl des für das Data Mining relevanten Datenbestandes benutzt werden kann.

Hierfür ist ein Data Warehouse besonders gut geeignet. Laut (Chamoni, Analytische Informationssysteme, 1998**,**13) wird unter einem Data Warehouse „... ein

unternehmensweites Konzept ..., dessen Ziel es ist, eine logisch zentrale, einheitliche und konsistente Datenbasis für die vielfältigen Anwendungen zur Unterstützung der analytischen Aufgaben von Fach– und Führungskräften aufzubauen...“ verstanden.

Das Data Warehouse bildet die aus unterschiedlichen Quellen stammenden, für Auswertungszwecke benötigten Unternehmensdaten auf eine einheitliche, unternehmensweite und konsistente Datenbank ab. Für diesen zentralen Datenpool werden Informationen aus den operativen Systemen in bestimmten Zeitabständen übertragen. Die Daten werden dabei konsolidiert, indem redundante, inkonsistente und für die Data-Mining-Analysen nicht benötigte Daten herausgefiltert werden. Die Datenbasis kann bei diesem Vorgang auch neu strukturiert und themenorientiert zusammengefasst werden.
Durch die mit der Einführung eines Data Warehouse verbundene Entkopplung der Datenanalyse von den operativen IT-Systemen können diese entlastet werden. Dies ist, wie bereits gesagt, wichtig, da Data Mining-Auswertungen sehr rechenintensiv sind.
Die in einem Data Warehouse abgelegten Daten besitzen vier typische Merkmale:

1) Themenorientierung
2) Vereinheitlichung
3) Zeitorientierung
4) Beständigkeit

Themenorientierung bedeutet, dass die Informationseinheiten in einem Data Warehouse auf die inhaltlichen Kernbereiche der Organisation fokussiert sind. Dies ist ein Unterschied zu den üblichen applikations– bzw. prozessorientierten Konzepten der operativen IT–Anwendungen, die auf eine effiziente Abwicklung des Tagesgeschäftes ausgerichtet sind. Zum Beispiel sind Objekte wie ein spezifischer Kundenauftrag oder eine einzelne Produktionscharge kaum dazu geeignet, Entscheidungen zu unterstützen.
Im Data-Warehouse-Umfeld konzentriert man sich eher auf inhaltliche Themenschwerpunkte, wie zum Beispiel Kunden und Produkte. Operative Daten, die lediglich für die Prozessdurchführung wichtig sind und nicht der

Entscheidungsunterstützung dienen können, werden nicht ins Data Warehouse übernommen.

Ein zentrales Merkmal des Data-Warehouse-Konzepts ist, dass die Daten vereinheitlicht werden, bevor ihre Übernahme aus den operationalen Systemen erfolgt. Diese Vereinheitlichung kann in verschiedenen Formen auftreten und bezieht sich häufig auf Namensgebung, Skalierung und Kodierung. Das Ziel dieser Vereinheitlichung ist ein konsistenter Datenbestand, der sich stimmig und akzeptabel präsentiert, selbst wenn die Datenquellen große Heterogenität aufweisen.

Die Zeitorientierung der in einem Data Warehouse abgelegten Informationseinheiten zeigt sich auf folgende Weise. Mehrere zeitpunktsbezogene Daten werden zu zeitraumbezogenen Daten zusammengefasst. Beim Data Mining sind gerade die Analysen von Zeitreihen über längere und mittlere Zeiträume (Wochen-, Monats- oder Jahresbetrachtungen). Für solche Auswertungen reichen diese Informationen mit mäßiger Aktualität vollkommen aus.

Die beständige Bevorratung von Zeitreihendaten über lange Zeiträume hinweg erfordert durchdachte, anwendungsgerechte Sammelverfahren und optimierte Speichertechniken, um den Umfang des zu speichernden Datenmaterials und damit die Zeit, die für einzelne Auswertungen und Abfragen benötigt wird, in erträglichen Grenzen zu halten.

2.4 Beschränkungen der Data Mining Analyse

Die Daten für eine neue Data-Mining-Analyse können, wie in Kapitel 2.2 beschrieben, Altdaten sein; dass heißt, eine Datengrundlage existiert schon, eventuell in Form einer Datenbank oder eines Data Warehouse. Oder man erzeugt eine neue Datenbasis indem man einen neuen Fokus für das Sammeln von Daten festlegt, und speichert diese dann in einer Datenbank. Dabei ist aber Folgendes zu beachten:
Nicht alle Daten können in die Datenbasis übernommen werden, da sonst erstens die Performance leiden und zweitens die Datenbank viel zu groß werden würde. Deswegen werden, wie bereits erwähnt, die unwichtigen Daten nicht in die

Datenbasis übernommen, sondern nur die Daten, die laut Vorgabe für die Analyse wichtig sind. Also lässt sich feststellen, dass die Daten in der Datenbasis nur ein Modell, oder ein Abbild der Realität sind, aber nicht die Realität selbst. Das bedeutet, die Datenbasis ist im vergleich zur Realität selbst weder vollständig noch isomorph. Data Mining setzt also immer auf einer reduzierten und abstrahierten Realität auf.
Es lassen sich in der Analyse auch nur Beziehungen zwischen erhobenen Daten sichtbar machen.

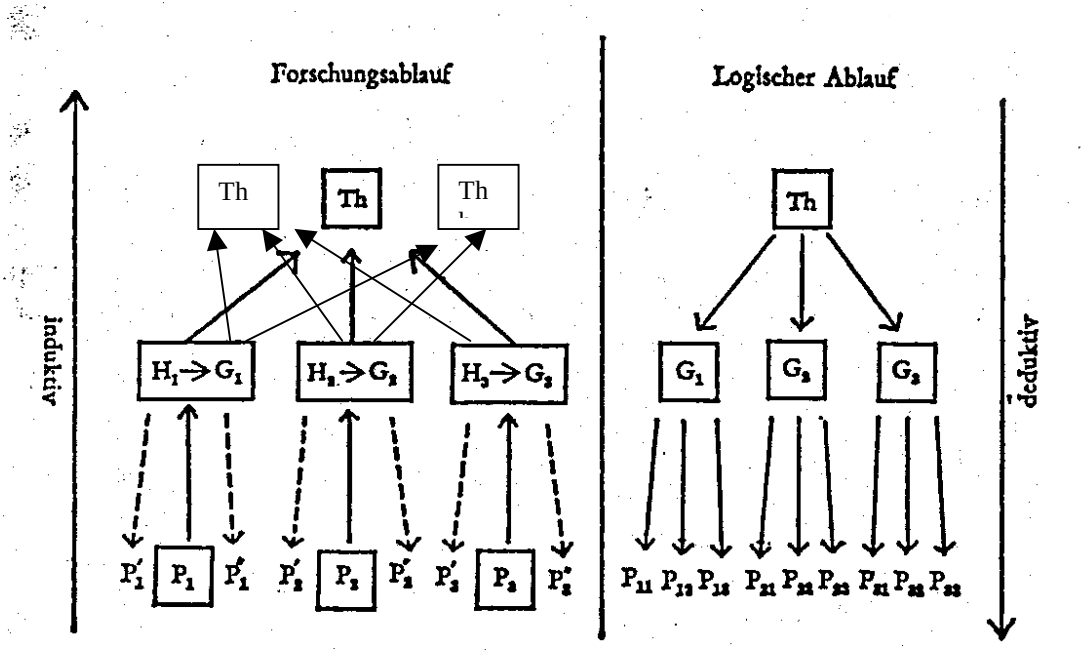
Abbildung 3: Realität und Modell

In Abbildung 3 sind die für die Datengrundlage ausgewählten Daten rot markiert. Nur Beziehungen zwischen diesen Daten lassen sich im Ergebnis der Data- Mining-Analyse aufzeigen. Die blau markierten Daten sind nicht in die Datenbasis übernommen worden. Zusammenhänge zwischen nicht übernommenen und nicht übernommenen und zwischen übernommenen und nicht übernommenen Daten können niemals im Ergebnis einer Data-Mining-Analyse entdeckt werden. Die Auswahl der Daten beeinflusst also das Ergebnis der Analyse und das Ziel der Data-Mining-Analyse beeinflusst die Auswahl der Daten.

2.5 Ziel der Data-Mining-Analyse

Das Ziel des Data Mining ist es, Theorien zu erstellen, mit deren Hilfe Aussagen über die Zukunft getroffen werden können. Zuerst geht man dabei induktiv vor. Dies bedeutet, dass mit Hilfe von ähnlichen Einzelbeobachtungen Theorien erzeugt werden. Hierzu benötigt man die verschiedenen Data Mining-Algorithmen. Man erstellt also aus Protokollaussagen (z.B. Verkaufszahlen, Umsätze) mit Hilfe von Algorithmen zuerst einmal Hypothesen. Aus diesen versucht nun der Mensch, Gesetze abzuleiten und dann schließlich die Gesetze in einer Theorie zusammenzufassen. Doch aus den Gesetzen lassen sich die verschiedensten Theorien ableiten.

Der für die Unternehmen interessantere Schritt ist der deduktive Ablauf. Hier versucht man mit Hilfe der erarbeiteten Theorien spezielle Vorhersagen zum Beilspiel über zukünftiges Kundenverhalten zu treffen.



Die Abkürzungen bedeuten: Th = Theorie, G = Gesetz, H = Hypothese, P = Protokollaussage.

Abbildung 4: Induktion - Deduktion

Das Ganze soll nun an einem kleinen Beispiel verdeutlicht werden:
Ein Supermarkt versucht, die Kaufgewohnheiten der Kunden mit Hilfe des Data Mining zu analysieren, damit für Produkte effektiver geworben werden kann und damit die Produkte besser gruppiert werden können. Dazu stehen dem Supermarkt Daten der Kunden wie unter anderem Geschlecht, gekaufte Produkte, Zeit des Einkaufs und Gesamtsumme zur Verfügung. Auf diese Daten wird ein Data Mining-Algorithmus angewandt, und es ergibt sich die Hypothese, dass von jungen Männern oft Bier und Windeln gekauft werden. Die Theorie wäre in diesem Fall, dass verheiratete Männer mit kleinen Kindern oft beim Bierholen von ihren Frauen gebeten werden, Windeln mitzubringen. Der nächste Schritt wäre dann, Bier neben die Windeln zu stellen, da beides von derselben Kundengruppe gekauft wird und so die

Verkaufszahlen für beide Artikel steigen könnten. Mit den Ergebnissen der Analyse ließen sich also unbekannte, interessante Zusammenhänge entdecken und damit Aussagen über zukünftiges Käuferverhalten machen.

2.6 Probleme bei der Analyse

Eines der Probleme bei der Sammlung quantitativer Daten über Data-Mining-Projekte besteht darin, die wirkliche Bedeutung der Daten zu verstehen. Leicht können Daten falsch interpretiert und falsche Schlussfolgerungen gezogen werden. Dies soll folgendes Beispiel veranschaulichen:
Ein Manager entschließt sich, die Anzahl der von Kunden gestellten Änderungsanträge zu überwachen, und geht davon aus, dass eine Beziehung zwischen diesen Anträgen und der Verwendbarkeit und Eignung des Produkts besteht. Je höher die Anzahl der Änderungsanträge ist, desto weniger entspricht die Software den Bedürfnissen des Kunden.

Die Verarbeitung von Änderungsanträgen und das Ändern der Software sind aufwändig. Daher entschließt sich die Organisation, ihren Ablauf zu ändern, um die Zufriedenheit bei den Kunden zu steigern und gleichzeitig die Änderungskosten zu senken. Hinter den Prozessänderungen steht die Absicht, die Produkte zu verbessern und weniger Änderungsanträge zu erhalten.

Also werden Prozessänderungen eingeleitet, die den Kunden mehr in den Prozess des Softwareentwurfs einbeziehen. Für alle Produkte wird eine Beta-Testphase eingeführt, und vom Kunden geforderte Änderungen werden in das ausgelieferte Produkt integriert. Neue Produktversionen, die mit diesem geänderten Prozess entwickelt werden, kommen zur Auslieferung. In einigen Fällen reduziert sich die Anzahl der Änderungsanträge, in anderen Fällen steigt sie. Der Manager ist verwirrt und kann die Auswirkungen der Prozessänderungen auf die Produktqualität nicht beurteilen.

Um zu verstehen, warum derartige Dinge passieren können, muss man verstehen, warum Änderungsanträge überhaupt erfolgen. Ein Grund dafür ist, dass die ausgelieferte Software nicht das tut, was der Kunde möchte. Eine weitere Möglichkeit

wäre, dass die Software sehr gut ist und auf breiter Ebene vielfach eingesetzt wird, manchmal sogar für Zwecke, für die sie ursprünglich nicht gedacht war. Da sie von so vielen Leuten verwendet wird, ist es nur natürlich, dass mehr Änderungsanträge gestellt werden.

Eine dritte Möglichkeit besteht darin, dass die Herstellerfirma der Software auf Änderungsanträge der Kunden schnell reagiert. Daher sind die Kunden mit der erhaltenen Unterstützung zufrieden. Sie stellen eine Menge Änderungsanträge, da sie wissen, dass diese Anträge ernst genommen werden. Ihre Vorschläge werden wahrscheinlich in künftige Versionen der Software integriert.

Die Anzahl der Änderungsanträge könnte sinken, da die Prozessänderungen Wirkung gezeigt haben und die Software dadurch besser verwendbar und geeigneter geworden ist. Andererseits könnte die Anzahl gesunken sein, da das Produkt gegenüber einem Konkurrenzprodukt an Marktanteil eingebüßt hat und es somit weniger Benutzer des Produkts gibt. Die Anzahl der Änderungsanträge könnte steigen, da es mehr Benutzer gibt, weil die Beta-Testphase die Benutzer davon überzeugt hat, dass der Hersteller bereit ist, Änderungen vorzunehmen, oder weil die Beta-Teststandorte nicht typisch für den häufigsten Gebrauch des Programms waren.

Zur Analyse der Daten des Änderungsantrags reicht es nicht, die Anzahl von Änderungsanträgen zu kennen. Wir müssen wissen, wer den Antrag gestellt hat, wie derjenige die Software eingesetzt und warum der Antrag gestellt wurde. Wir müssen außerdem wissen, ob externe Faktoren wie Verfahrensänderungen bei Änderungsanträge oder Marktänderungen vorliegen, die sich auswirken könnten. Mit diesen Informationen ist es dann möglich, herauszufinden, ob die Prozessänderungen wirksam zur Erhöhung der Produktqualität beigetragen haben.

Dies verdeutlicht, dass die Interpretation quantitativer Daten über ein Produkt oder einen Prozess ein ungewisser Vorgang ist. Vermessene Prozesse und Produkte lassen sich nicht isoliert von ihrer Umgebung betrachten, und Änderungen an dieser Umgebung könnten zur Ungültigkeit vieler Datenvergleiche führen. Quantitative Daten über menschliche Aktivitäten kann man nicht immer wörtlich nehmen. Die Gründe für den gemessenen Wert müssen möglicherweise untersucht werden.

2.7 Fazit

Beim Data Mining wird also versucht, ein reales, physisches Objekt in ein ideal mathematisches Objekt (gesäuberte Datenbasis) zu transformieren. Ein Data Mining-Algorithmus soll dann induktiv eine Theorie erstellen, mit deren Ergebnis die reale Welt beeinflusst werden soll.

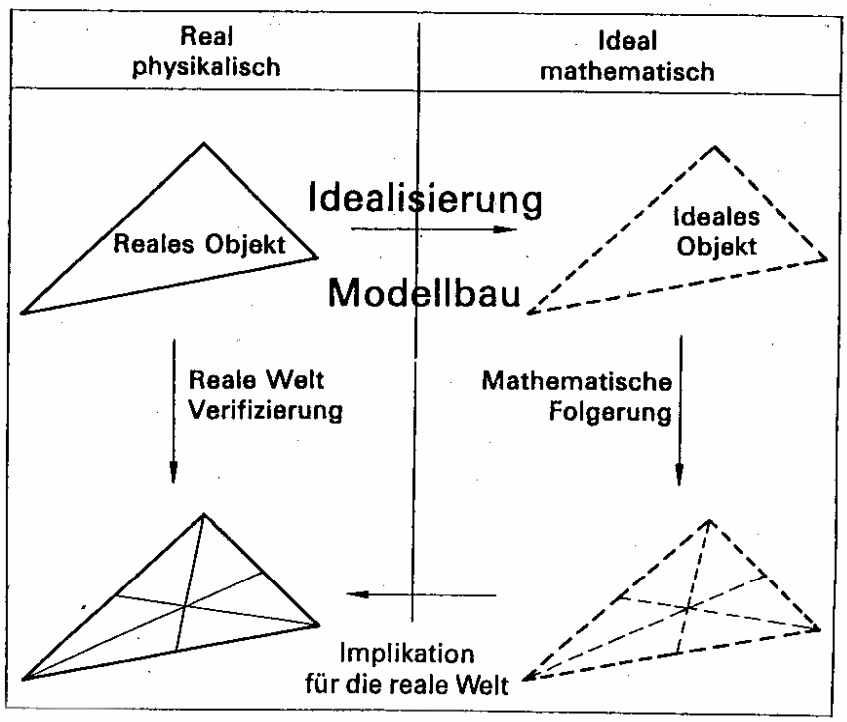


Abbildung 5: Übersicht des Data-Mining-Prozess

2.8 Anmerkungen

In der Einleitung zu Kapitel 2 wird die Frage geklärt, aus welchen Zweigen der Wissenschaft sich Data Mining entwickelt hat. Zusätzlich werden in einer Definition des Data Mining ausführlich die Ziele erklärt, damit eine Verständnisbasis gebildet werden kann. Diese Definition soll den Studenten eine Idee geben, wozu das Data Mining gut ist, bevor sie sich die abstrakten Definitionen anhören müssen.
Im nächsten Teil werden die einzelnen Schritte des Data Mining mit Hilfe der Abbildung 1 erklärt, die den Zusammenhang zwischen den einzelnen Schritten verdeutlichen soll. Dies wird auch als Folie vorliegen. Zum Schritt Data Mining werden einzelne Beispiele genannt, um die Möglichkeiten einer Data-Mining-Analyse aufzuzeigen. Hierzu wird ebenfalls eine Folie aufgelegt.
Im Punkt „ Auswahl der Daten“ wird die Möglichkeit, ein Data Warehouse als Datenbasis zu verwenden, erläutert. Es wird kurz auf die Vorteile der Merkmale der der in einem Data Warehouse gespeicherten Daten eingegangen. Dies soll den Studenten die Wichtigkeit einer einheitlichen, konsistenten Datenbank zeigen.
Der Punkt „Beschränkungen der Data-Mining-Analyse“ zeigt wichtige Grundlagen beim Zusammenstellen einer Datenbasis und die Zusammenhänge zwischen Auswahl der Daten und Ziel der Analyse. Diese Zusammenhänge sind wichtige Grundlagen, die für das Verständnis des Data Mining von größter Wichtigkeit sind.
In Punkt 2.5 „Die Ziele des Data Mining“ werden abschließend die Begriffe Induktion und Deduktion erklärt, also die Schritte, wie man von anfänglichen Daten zu einer Theorie und dann zu Aussagen über die Zukunft kommt. Dies wird mittels eines kleinen Beispiels verdeutlicht, zu dem dann noch eine Skizze der Abbildung 4 an der Tafel angefertigt wird.
Abschließend wird noch ein Fazit gezogen, in dem noch einmal der ganze Data-Mining-Prozess erklärt wird. Auch hierfür wird eine Skizze der Abbildung 5 an die Tafel gezeichnet. Die Anwendung von verschiedenen Medien, wie Folien, Tafelbild und Sprache, soll den Lernprozess vertiefen. Und eine mehrmalige Wiederholung der Informationen dient demselben Zweck.

3 Data-Mining-Methoden

3.1.1 Anmerkung

Kapitel 3 beginnt mit einer kurzen Beschreibung der Data-Mining-Aufgaben. Danach werden die Methoden des Data Minings besprochen. Zuerst wird hier zwischen unüberwachtem und überwachtem Lernen unterschieden. Dann wird gezeigt, welche Methoden zum überwachten und welche zum unüberwachten Lernen gehören. Die wichtigsten werden danach kurz beschrieben. Es wird kurz auf Entscheidungsbäume eingegangen, die unter 3.2 genauer behandelt werden. Bei Clustering wird zwischen agglomerativen und divisiven Verfahren unterschieden. Dann werden noch kurz neuronale Netze und Assoziationsregeln erklärt. In 3.2.1 werden dann die Eigenschaften von Entscheidungsbäumen näher betrachtet. 3.2.2 beschreibt kurz den Ablauf von ID3 und in 3.2.3 wird dieses Vorgehen an einem Beispiel verdeutlicht. Anmerkung für die Übung: der log2 wird mit dem Taschenrechner folgendermaßen berechnet: $\log_2 X = \ln X / \ln 2$.

3.1.2 Überblick über die Data-Mining-Methoden

Die Methoden des Data Mining werden in der Literatur nach verschiedenen Kriterien klassifiziert. Da einzelne Methoden für verschiedene Fragestellungen genutzt werden können, werden auch unterschiedlich viele Klassifikationsebenen genutzt. Je nach Autor und Buch kann es eine bis 3 Dimensionen geben. Uns genügen für eine Übersicht allerdings schon zwei.

Die **erste Dimension** wird als „**Aufgabe**“ bezeichnet. Sie ergibt sich aus dem konkreten unternehmerischen Anlass für Data Mining.
Data Mining eignet sich, wie bereits erwähnt, für folgende Aufgaben

- Klassifikation
- Generalisierung
- Entdecken von Ausreißern

Diese Aufgaben wurden schon kurz in Kapitel 2.2 besprochen.

Die **zweite Dimension** wird als „**Methode**“ bezeichnet.
Die Methoden lassen sich in zwei verschiedene Kategorien einteilen, nämlich das überwachte und das unüberwachte Lernen.

Beim **überwachten Lernen** liegt eine bestimmte Menge an Trainingsbeispielen vor, für die eine richtige Klassifikation bekannt ist. Das Verfahren besteht dann aus zwei Schritten: Zunächst muss das lernende System anhand der Trainingsbeispiele eine Klassifikation erlernen und diese im zweiten Schritt auf die Gesamtmenge der Objekte anwenden (Krahl, Data Mining 1998, 62).

Beim **unüberwachten Lernen** liegen diese Trainingsbeispeile nicht vor und das System muss ohne Vorgaben interessante Zusammenhänge in den Daten erkennen. Dabei kommen zwei unterschiedliche Sichtweisen in Betracht: Segmentierung und Assoziationen (Krahl, Data Mining 1998, 78f).

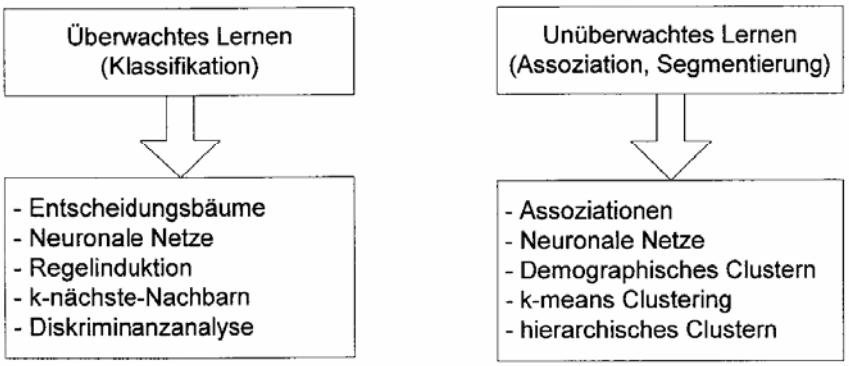


Abbildung 6: Übersicht Data-Mining-Methoden

Bei der **Segmentierung** ist das Ziel eine Zuordnung eines jeden Datensatzes zu einem Segment. Es wird nach einer globalen Strukturierung der Daten gesucht. Zu den wichtigsten Verfahren zählt hier die Clusteranalyse.

Bei der **Assoziationsanalyse** wird versucht, Beziehungen zwischen den Datensätzen zu finden, um daraus eigenständige Assoziationsregeln zu formulieren, die häufig auftretende versteckte Regeln oder Muster in Datenbeständen beschreiben.

Es wäre etwas viel, auf alle Methoden detailliert einzugehen, deswegen soll im Anschluss nur eine kurze Übersicht über die wichtigsten Methoden gegeben werden.

Im folgenden Absatz werden nun einige Beispiele für überwachtes Lernen gegeben. Bei Methoden der **Regelinduktion** oder **Entscheidungsbäumen**, die am weitesten im Bereich des Maschinellen Lernens entwickelt wurden, werden Objekte, deren Klassenzuordnung bekannt ist, sukzessive mit Hilfe einzelner Merkmale in Gruppen aufgeteilt, die in sich homogen, aber voneinander möglichst unterschiedlich sind. Am Ende des Verfahrens steht ein Baum, aus dessen Verzweigungskriterien Regeln gebildet werden können, die dann auf nicht zugeordnete Objekte angewendet werden können. Entscheidungsbäume werden hauptsächlich zur Klassifikation angewandt (siehe auch 3.2 für mehr Details zu Entscheidungsbäumen). Traditionelle statistische Verfahren wie Diskriminanzanalyse, k-nächste-Nachbarn oder logistische Regression werden ebenfalls zum Klassifizieren angewandt.

Die folgenden Absätze sollen eine kurze Übersicht über einige Methoden des unüberwachten Lernens geben. Bei **neuronalen Netzen** werden so genannte künstliche Neuronen in Schichten angeordnet, in denen alle Neuronen einer Schicht mit allen Neuronen der beiden Nachbarschichten verbunden sind. Die erste Schicht, die die zu verarbeitenden Daten aufnimmt, wird als Input- und die letzte, die das Ergebnis liefert, als Outputschicht bezeichnet.
Künstliche Neuronen sind Prozessoren, die bestimmte Eigenschaften der Signalerzeugung natürlicher Neuronen simulieren. Sie empfangen über Verbindungen Signale anderer Neuronen, wobei der Quotient aus der gesendeten und der empfangenen Signalstärke als Verbindungsstärke oder “Gewicht" bezeichnet wird. Die Fähigkeiten neuronaler Netze beruhen auf der Möglichkeit, diese Gewichte zu verändern. Passiert die Gewichtsveränderung in einer sinnvollen Weise, kann von „Lernen“ gesprochen werden. Die gewichteten Inputs werden dann im Neuron aufsummiert. Schließlich wird durch eine Ausgabe- oder Übertragungsfunktion aus dieser Summe bestimmt, wie stark das Signal ist, welches den anderen Neuronen mitgeteilt wird. Dem Netz werden so lange Beispiele (Daten) präsentiert, bis es die erwünschten Ergebnisse zeigt, d.h. bis die Gewichte optimal sind. Wenn die Ergebnisse der Beispiele bekannt sind, dann spricht man von überwachtem Lernen.

Es gibt auch neuronale Netze für unüberwachtes Lernen, die als Selbstorganisierende Netze bezeichnet werden.

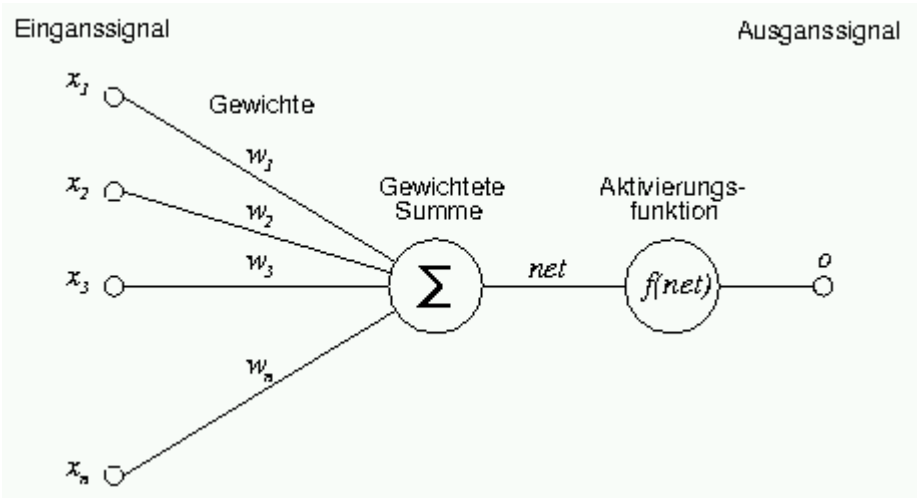


Abbildung 7:Schematische Darstellung eines Neurons

Aus den Eingangssignalen xi wird die gewichtete Summe „net“ gebildet. Aus dieser Summe wird dann mit der Aktivierungsfunktion f(net) der Ausgangswert o berechnet.

Die **Clusteranalyse** ist ein statistisches Verfahren, das in sehr vielen Varianten vorkommt. Die wichtigsten sind die agglomerative und die divisive Methode. „**Agglomerativ**“ bedeutet, dass die Gruppierung (Bildung von Clustern) durch fortschreitende Zusammenfassung von Objekten erfolgt, während bei der **divisiven** Methode zu Anfang alle Objekte in einer Gruppe (in einem Cluster) zusammengefasst sind und durch das Verfahren in Untergruppen unterteilt werden. Dabei werden die so entstehenden Gruppierungen in einem Dendrogramm dargestellt (Chamoni, Ausgewählte Verfahren des Data Mining, 1998, 307).

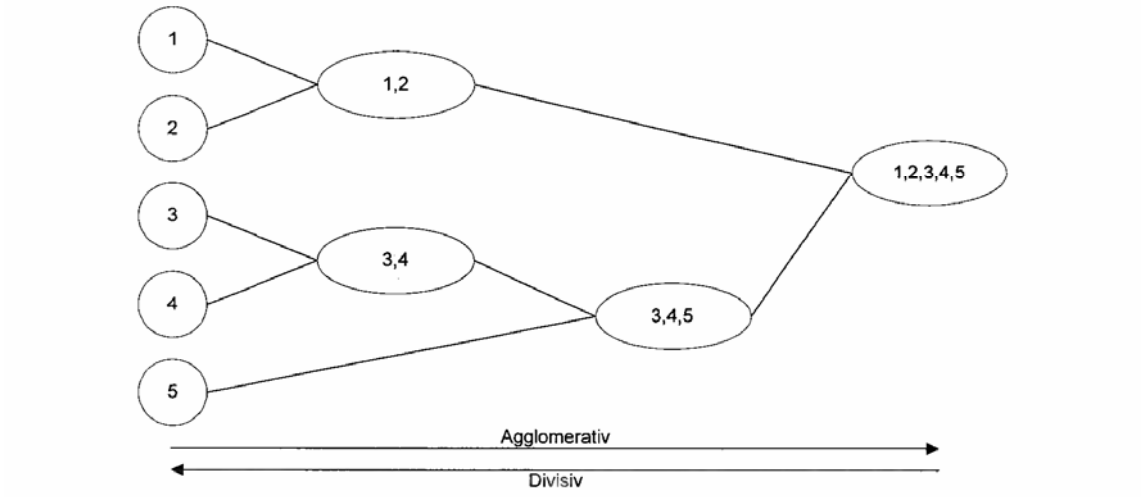


Abbildung 8: Dendrogramm mit agglomerativer und divisiver Clusterung

Agglomerative Verfahren beginnen mit einelementigen Gruppen, die während des Verfahrens zusammengefasst werden. Dabei werden bei jedem Verfahrensschritt zwei oder mehr Gruppen bestimmt, die eine neue Gruppe bilden. Dieses Verfahren bricht spätestens dann ab, wenn sich alle Elemente in einer Gruppe befinden.

Divisive Verfahren gehen umgekehrt vor. Zu Anfang befinden sich alle Objekte in einer Gruppe. Bei jedem Verfahrensschritt wird die Gruppe weiter in Untergruppen unterteilt. Dieses Verfahren bricht spätestens dann ab, wenn alle Gruppen einelementig sind.
Die Kernaufgabe der Algorithmen besteht darin zu entscheiden, welche Gruppen gebildet werden sollen, sei es durch Zusammenfassung oder sei es durch Untergruppenbildung. Voraussetzung für alle Verfahren ist eine so genannte Verschiedenheitsmatrix, aus der man ablesen kann, worin die Verschiedenheit zweier Objekte besteht.

3.2.1 Entscheidungsbäume

Ein Entscheidungsbaum ist eine Darstellungsform für eine Klassifikationsregel, anhand derer Objekte in Klassen eingeteilt werden können. Dabei gestaltet man den Baum so, dass in jedem Knoten ein Attribut abgefragt und eine Entscheidung getroffen wird, bis man ein Blatt erreicht. Ein Blatt ist ein Knoten, an dem keine weitere Verzweigung durchgeführt wird. Hier kann die Klassifikation abgelesen werden. Daher werden solche Bäume auch Klassifikationsbäume genannt. Grundlage für den Aufbau eines Entscheidungsbaumes sind Trainingsbeispieldatensätze, deren Klassenzugehörigkeit bekannt ist. Mit deren Hilfe wird von dem **ID3-Algorithmus** (Induction of Decision Trees) ein Klassifikator mit einem diskreten Wert hervorgebracht. Ein Klassifikator entscheidet, welcher Klasse ein neuer Datensatz zuzuordnen ist. Die Trainingsbeispieldatensätze bestehen aus mehreren Attributen, wobei ein Attribut das Zielattribut darstellt. Das Zielattribut ist die Größe, die klassifiziert werden soll. Jedes Blatt des Entscheidungsbaumes wird anhand des Zielattributs klassifiziert, erhält also einen möglichen Wert des

Zielattributes. Das Zielattribut darf bei der Entscheidungsauswahl nicht berücksichtigt werden, da es ja das Ergebnis dieser Entscheidungen darstellt. Es taucht daher nie selbst als Knoten im Baum auf. Zum Beispiel taucht das Zielattribut „Kündigung“ aus dem Beispiel des Kapitels 3.2.3 nicht als Blatt des Baumes auf. Siehe Abbildung 10.

Die Entscheidungsbäume werden top-down aufgebaut, indem ausgehend von der Wurzel des Baumes Unterteilungen anhand der Attributwerte vorgenommen werden. Dabei variieren die Verfahren darin, nach welchem Kriterium diese Unterteilung vorgenommen wird. „Häufig implementierte Baumtypen sind sogennante CARTs (Classification And Regression Trees) und CHAIDs (Chi-Square Automatic Interaction Detectors)” (Krahl, Data Mining 1998, 70). Daneben findet auch der ID3-Algorithmus Verwendung.

Der Ablauf bei all diesen Verfahren ist immer der gleiche bis auf das Attribut-Auswahlverfahren, wodurch die Unterteilung des Baumes gesteuert wird.
Dabei trennen die CART-Bäume nach dem Informationsgehalt. Die CHAID-Bäume verwenden den CHI-Quadrat-Unabhängigkeitstest, um den Attributswert auszuwählen, der von dem Zielattribut die größte Abhängigkeit hat.
Nun soll der ID3-Algorithmus näher erläutert werden, um eines der verschiedenen Attribut-Auswahlverfahren verständlich zu machen.

3.2.2 ID3-Methode

Das ID3-Verfahren startet mit einem elementaren Baum, der nur aus der Wurzel besteht, die alle Trainingsbeispiele enthält. Die Trainingsbeispiele sollen nun gemäß ihrer Attributwerte aufgeteilt werden. Auf jeder Ebene muss für jeden Knoten entschieden werden, welches Attribut ihm zugewiesen wird. Dies ist das zentrale Auswahlproblem von ID3. Welches Attribut passt am besten zu den an einem Knoten vorhandenen Trainingsbeispielen, also welches Attribut klassifiziert diese Beispiele am Besten an einem Knoten? Ist diese Frage entschieden, so wird für jeden möglichen Wert des ausgewählten Attributs eine Kante mit Kindknoten erzeugt und der Algorithmus startet mit den Kindknoten erneut, wobei einmal ausgewählte Attribute im selben Pfad nicht mehr vorkommen dürfen.

Der Auswahlalgorithmus von ID3 basiert auf einem Maß für den Informationsgehalt eines Attributs, genannt **Information Gain**. Dieses Maß stützt sich auf die Größe **Entropie**, ein Maß für Unordnung. Ein niedriger Wert für die Entropie des Baums entspricht einem hohen Informationsgehalt des Baums, ein hoher Wert einem kleinen Informationsgehalt. Der größte Informationsgehalt wird bei einem Entropiewert von 0 erreicht. Also muss die Entropie des Baums möglichst gering sein. Um dieses Auswahlproblem zu lösen, wird in jedem Schritt die Entropie aller Attribute errechnet. Dann wird anhand des Attributs mit der niedrigsten Entropie der Baum geteilt. Dadurch nimmt der Informationsgehalt des Baumes von der Wurzel zu den Blättern hin ab. Die Formeln für Entropie und Informationsgehalt werden im nächsten Kapitel genannt.

Ein Blatt im bisher aufgestellten Baum kann nicht weiter aufgeteilt werden, wenn es sich um einen so genannten **homogenen Knoten** handelt. Darunter versteht man ein Blatt, dessen Beispieldatensätze identische Attributwerte haben. Das Verfahren ist beendet, wenn alle Blätter des Baumes homogene Knoten sind.

Das Verfahren lässt nominale und numerische Attributwerte zu. Numerische Attributwerte sollten dabei allerdings anhand von Schwellenwerten in Klassen eingeteilt werden, damit man nicht zu viele Attributausprägungen und damit zu weit verzweigte Bäume erhält.

3.2.3 Durchführung der ID3-Methode

Der Vertriebsvorstand eines Unternehmens ist daran interessiert, welche Kundenprofile besonders kündigungsgefährdet sind. Zu diesem Zweck wurden die Daten aller Kunden in Abbildung 9 zusammengefasst.

Das Attribut **Kundennummer** beinhaltet die vom System vergebene Nummer, durch die die Kunden unterschieden werden.

Das Attribut **Vertragsdauer** beinhaltet anstatt Monatsangaben drei Klassen. Dies sind die Klassen „niedrig“, „mittel“ und „hoch“. Einer niedrigen Vertragsdauer entsprechen Verträge von 1 bis 6 Monaten, einer mittleren Vertragsdauer Verträge von 7 bis 18 Monaten und einer hohen Vertragsdauer die Verträge ab 19 Monaten.

Das Attribut **Berufsstatus** enthält die vier selbsterklärenden Attribute „Nicht-Erwerbstätige“, „Beamte“, „Angestellte“ und „Selbständige“.

Unter **Kündigung** versteht man die Unterscheidung in Kunden, die ihren Vertrag bereits gekündigt haben oder nicht. Das Attribut „Kündigung“ stellt das Zielattribut für die anschließende Klassifikation dar.

Kundennummer	Vertragsdauer	Berufsstatus	Kündigung
1	Mittel	Nicht-Erwerbstätige	Nein
2	Niedrig	Beamte	Ja
3	Niedrig	Angestellte	Ja
4	Mittel	Selbständige	Nein
5	Mittel	Angestellte	Nein
6	Hoch	nicht-Erwerbstätige	Nein
7	Mittel	Angestellte	Nein
8	Hoch	Beamte	Nein
9	Hoch	Selbständige	Ja
10	Mittel	nicht-Erwerbstätige	Nein

Abbildung 9: Eingangsdaten für die ID3-Analyse

Zu Anfang befinden sich alle Kunden in der Wurzel des Entscheidungsbaumes. Daher wird zunächst gemäß dem in Kapitel 3.2.2 vorgestellten Verfahren die Entropie der Wurzel des Baumes ausgerechnet. Diese berechnet sich nach folgender Formel:

$$\text{Entropie } (S,Z) = \sum_{z \in Z} - p(z) \log_2 p(z)$$

S ist die gesamte Stichprobe, auf die der Algorithmus angewandt wird.
Z steht für das Zielattribut nach dem klassifiziert wird, in unserem Beispiel also für die Kündigung.
„**z**“ steht für die möglichen Attributwerte des Zielattributs Z, in dem Beispiel sind dies „Ja“ und „Nein“.
„**p**“ ist die Wahrscheinlichkeit dafür, dass ein z vorliegt, also die Anzahl aller Datensätze mit z geteilt durch die Mächtigkeit der Stichprobe.

Im konkreten Fall ist die relative Häufigkeit der „Vertragskündigenden“ Kunden 0,3 und die der „nicht-vertragskündigenden“ Kunden 0,7, da es insgesamt 10 Kunden sind. Als Entropie ergibt sich somit:

$$\text{Entropie (Kunden 1-10, Kündigung)} = -0{,}3 * \log_2 0{,}3 - 0{,}7 * \log_2 0{,}7 = 0{,}881$$

Nun erfolgt eine Entscheidung aufgrund des ID3-Algorithmus, ob die erste Unterteilung anhand des Attributs „Vertragsdauer“ oder „Berufsstatus“ erfolgen soll.
Dies geschieht anhand des Gütemaßes „InfoGain“, welches den Informationsgehalt eines Attributs angibt. Dieses Gütemaß berechnet man nach folgender Formel:

$$\text{InfoGain}(S,Z,A) = \text{Entropie}(S,Z) - \sum_{v \in A} \frac{\#S_v}{\#S} * \text{Entropie}(S_v,Z)$$

Hier steht **A** für das Attribut, dessen InfoGain berechnet werden soll.
„**v**“ ist ein Platzhalter für alle möglichen Attributwerte von A.
$\mathbf{S}_v$ ist die Untermenge der Stichprobe S, deren Elemente beim Attribut A den Wert v aufweisen.
„**#**“ bedeutet die Mächtigkeit der Mengen S und S_v.

Man ermittelt für den jeweiligen Attributwert (z.B. niedrige, mittlere und hohe Vertragsdauer) die Entropie und gewichtet ihn mit der relativen Häufigkeit des Attributs. Die Einzelergebnisse werden aufsummiert und anschließend von der

Entropie der gesamten Stichprobe subtrahiert. Der Attributwert mit dem größten ermittelten InfoGain wird somit ausgewählt. Im Beispiel ergeben sich dadurch folgende Werte:

1. Attribut Vertragsdauer:

InfoGain(Kunden1-10, Kündigung, Vertragsdauer) =

Entropie (Kunden 1-10, Kündigung) - $\sum_{vEVertragsdauer} \frac{\#Sv}{\#S}$ * Entropie (S_v,Kündigung) =

Die Entropie der Kunden 1 bis 10 und der Vertragsdauer wurde bereits berechnet. Deswegen muss nur noch die folgende Formel ausgewertet werden.

$\sum_{v \in Vertragsdauer} \frac{\#Sv}{\#S}$ * Entropie (S_v,Kündigung) =

Jetzt wird die Entropie der einzelnen Attributwerte von „Vertragsdauer“ berechnet, nämlich für die Attributwerte „niedrig“, „mittel“ und „hoch“.

Für „niedrig“ ergibt sich:

2/10 * Entropie ($S_{Vertragsdauerniedrig}$, Kündigung)

2 der 10 Beispiele haben eine niedrige Vertragsdauer. Die Attributwerte von „Kündigung“ sind in diesem Fall beide „Ja“, also identisch. Bei identischen Werten ist die Entropie 0, daher fällt dieser Term weg (Multiplikation mit 0!).

Für „mittel“ ergibt sich:

5/10 * Entropie($S_{Vertragsdauermittel}$, Kündigung)

5 der 10 Beispiele der Stichprobe haben eine mittlere Vertragsdauer. Die Attributwerte von „Kündigung“ sind in diesem Fall „nein“, daher ist die Entropie in diesem Fall wieder 0, und auch dieser Term fällt weg.

Für „hoch“ ergibt sich:

3/10 * Entropie($S_{Vertragsdauerhoch}$, Kündigung)=

3/10 * (- 1/3 * $\log_2$ 1/3 – 2/3 * $\log_2$ 2/3) = 0,275

3 der 10 Beispiele der Stichprobe haben eine hohe Vertragsdauer. Die Attributwerte von „Kündigung“ sind in diesem Fall zwei Mal „nein“ und einmal „ja“. Deswegen muss hier die Entropie berechnet werden.

Die Gesamtformel für den InfoGain lautet also:
InfoGain(Kunden 1-10, Kündigung, Vertragsdauer) = 0,881 – 0,275 = 0,606

2. Attribut Berufsstatus:

InfoGain(Kunden1-10, Kündigung, Berufsstatus)=
0,881- {#$S_{\text{nicht-erwerbstätig}}$/#S * Entropie($S_{\text{nicht-erwerbstätig}}$, Kündigung) +
#S_{beamte}/#S * Entropie(S_{beamte}, Kündigung) +
#$S_{\text{angestellte}}$/#S * Entropie($S_{\text{angestellte}}$, Kündigung) +
#$S_{\text{selbständige}}$/#S * Entropie($S_{\text{selbständige}}$, Kündigung)} =

$\frac{3}{10}$ * 0 +
$\frac{2}{10}$ * (-½ * $\log_2$ ½ - ½ * $\log_2$ ½) +
$\frac{3}{10}$ * (-⅓ * $\log_2$ ⅓ – $\frac{2}{3}$ * $\log_2 \frac{2}{3}$) +
$\frac{2}{10}$ * (-½ * $\log_2$ ½ - ½ * $\log_2$ ½) =

$\frac{2}{10}$ * 1 + 0,275 + $\frac{2}{10}$ * 1 = 0,675

InfoGain(Kunden 1-10, Kündigung, Berufsstatus) = 0,881 – 0,675 = 0,206

Der Wert für die Vertragsdauer (0,606) ist höher als der für den Berufsstatus (0,206). Also unterteilt man den Entscheidungsbaum zunächst nach der Vertragsdauer. Dazu wird für jede Ausprägung ein Ast gebildet. Man stellt fest, dass für die Ausprägung „niedrig“ und „mittel“ bereits homogene Knoten vorliegen (siehe Abbildung 10). Deshalb wird nur der Ast „hoch“ nach dem Berufsstatus weiter unterteilt. Die nun entstehenden Kindknoten sind ebenfalls alle homogen. „Nicht-Erwerbstätige“ und „Beamte“ mit „hoher Vertragsdauer“ gehören alle der Klasse an, die den Vertrag nicht gekündigt haben. „Selbständige“ mit „hoher Vertragsdauer“ gehören hingegen ausnahmslos der Klasse an, die den Vertrag bereits gekündigt haben. Dem Zweig „Angestellte“ kann keine Klasse zugeordnet werden, weil für die Ausprägung mit „hoher Vertragsdauer“ kein Trainingsbeispiel vorliegt. Abbildung 10 zeigt das Ergebnis:

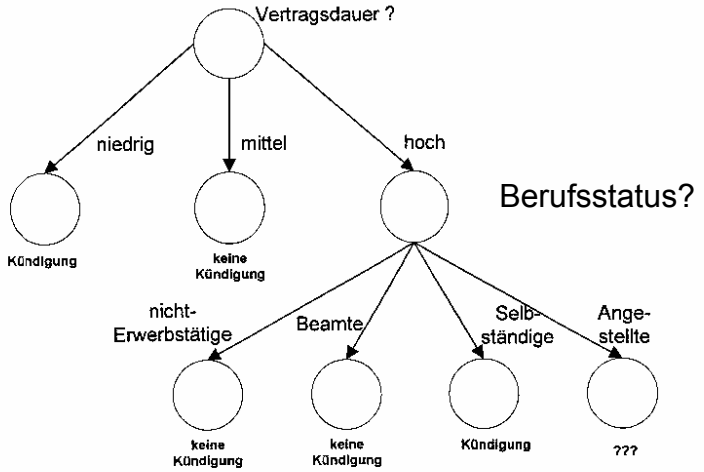


Abbildung 10: ID3-Entscheidungsbaum

Nun können die Kunden, die neu hinzukommen, bezüglich ihrer Kündigungsgefährdung klassifiziert werden. Folgende zwei Kunden sollen zugeordnet werden:

Kundennummer	Vertragsdauer in Monaten	Berufsstatus
11	mittel	Beamte
12	niedrig	Selbständige

Abbildung 11: Beispieldatensätze für die Klassifikation nach ID3

Mit Hilfe des in Abbildung 10 gezeigten Baumes können diese neuen Kunden leicht klassifiziert werden. Verfolgt man für den Kunden mit der Kundennummer 11 ausgehend von der Wurzel den Ast „mittel“, so ist man bereits bei einem Blatt des Baumes angelangt. Hier kann man direkt die Annahme ablesen, dass der Kunde nicht kündungsgefährdet ist. Das Ergebnis für den Kunden mit der Kundennummer 12 ist anders. Verfolgt man den Ast „niedrig“, kann man direkt ablesen, dass der Kunde kündungsgefährdet ist.

Aus dem Baum lassen sich insgesamt folgende Entscheidungsregeln ableiten:

- „Bei niedriger Vertragsdauer ist der Kunde gefährdet."
- „Bei mittlerer Vertragsdauer ist der Kunde nicht gefährdet."
- „Bei hoher Vertragsdauer und Berufsstatus Nicht-Erwerbstätige ist der Kunde nicht gefährdet."
- „Bei hoher Vertragsdauer und Berufsstatus Beamte ist der Kunde nicht gefährdet."
- „Bei hoher Vertragsdauer und Berufsstatus Selbständige ist der Kunde gefährdet."

Gerade die letzte Regel ist ein Beispiel für einen unerwarteten Zusammenhang. Doch gerade die Analyse solcher Kundenprofile kann Ansatzpunkte für wirksame Kündigungsbekämpfung liefern.

4 Fallbeispiel: Bonitätsprüfung

4.1 Vorwort

Ein Beispiel für die tatsächliche Anwendung von Data Mining bei einer Firma zu finden ist gar nicht einfach. Die meisten Firmen wollen natürlich nicht über die Daten, die sie dazu benutzen oder die Ergebnisse, die sie erhalten reden. Daher findet man meist nur oberflächliche Beispiele in denen grob auf die angewendeten Algorithmen ein und geben so gut wie keine firmenbezogenen Daten wieder. Aber trotzdem kann aus so einem Beispiel dennoch einiges interessantes gelernt werden.

4.2 Einleitung

Die Gewährung von Krediten spielt nicht nur für Bankinstitute sondern auch für Industrie- und Handelsunternehmen eine zunehmende Rolle, man denke etwa an Ratenzahlungen im Versandhandel oder bei Leasing-Gesellschaften. Durch den Einsatz von automatischen Verfahren zur Bonitätsprüfung ergeben sich zahlreiche Vorteile und Rationalisierungspotentiale. Zugleich eröffnen sich auch Anwendungsbereiche für Data-Mining-Verfahren.

Vor dem Einsatz entsprechender Verfahren haben Datenvorbereitung und Merkmalsauswahl große Bedeutung. In diesem Kapitel wird die Anwendung von Entscheidungsbaumverfahren auf eine Stichprobe aus dem Versandhandel dargestellt.

Bei jeder Kreditgewährung fallen die Leistung des Kreditgebers und die Gegenleistung des Schuldners – seine Rückzahlungen und Zinsen – zeitlich auseinander. Damit stellen die zukünftigen Zahlungen des Schuldners für den Kreditgeber unsichere Werte dar und beinhalten Risiken. Es besteht die Möglichkeit des teilweisen oder vollständigen Ausfalls von Zahlungen sowie ihres zeitlichen Verzugs. Aufgabe der Bonitätsprüfung ist die Beschaffung und Verarbeitung von

Informationen zur Bestimmung des so genannten Bonitätsrisikos. Dies ist der Wert für die Wahrscheinlichkeit einer Kreditrückzahlung.

Neben der traditionellen Kreditprüfung und Vergabeentscheidung durch Sachbearbeiter lassen sich verschiedene Ansätze unterscheiden. Diese reichen über den Einsatz statistischer Verfahren bis hin zu Ansätzen aus den Bereichen des induktiven Lernens und der künstlichen Intelligenz. Somit ist eine breite Palette von Data-Mining-Verfahren angesprochen. Als Ergebnis im Sinne des Data Mining können dabei solche Bonitätsmerkmale gelten, die zwischen guten und problembehafteten Kreditarrangements unterscheiden können.

Ein Großteil der eingesetzten Verfahren ist der Klassifikation zuzurechnen: Ausgehend von erfassten Attributen soll ein Kunde oder eine Firma einer vorgegebenen Bonitätsklasse zugeordnet werden. In der Vergangenheit abgewickelte Kreditfälle bilden dabei die Basis zur Konstruktion entsprechender Klassifikatoren, denn hier sind Informationen über die Attribute und Bonitätsklasse verfügbar. In diesem Beispiel werden aber nur Entscheidungsbaum-Klassifikatoren betrachtet.

Als wesentliche Ziele des Einsatzes von Credit-Scoring-Systemen und EDV-gestützten Verfahren gelten die Objektivierung der Kreditentscheidung und Standardisierung solcher Prozesse. Die daraus folgende Qualitätsverbesserung schützt den Kreditgeber vor zu hohen Kreditausfällen und den Kunden vor unberechtigten Kreditablehnungen. Des Weiteren gibt es Rationalisierungseffekte: Kreditanträge können schneller bearbeitet, eine Kostendeckung auch bei Kleinstkrediten erreicht werden.

4.3 Verwendeter Datenbasis

Die für dieses Beispiel verwendete Datenbasis enthält reale Kreditdaten aus dem Bereich der Bonitätsprüfung eines großen deutschen Versandhauses: Dort gehen täglich zwischen 50000 und 130000 Bestellungen bei einem Kundenstamm von 8 Millionen ein. Dabei müssen bis zu 8000 Neukunden pro Tag auf ihre Bonität geprüft werden. Eingesetzt werden verschiedene Scoring-Systeme, wobei zwischen

Application-Scoring, dies bezeichnet das Neukundengeschäft, und **Behaviour-Scoring** unterschieden wird. Mit statistischen Klassifikationsverfahren wie logistischer Regression und Diskriminanzanalyse, wurden 78% der Testdaten richtig klassifiziert. Der mit der Verbesserung der Klassifikationsgenauigkeit um 1-2% verbundene potentielle Anstieg des Gewinns wird mit einer Größenordnung von mehreren Hunderttausend bis Millionen DM beziffert.

Die Datenbasis ist der Bonitätsprüfung von bestehenden Kundenbeziehungen, dem Behaviour-Scoring, zuzuordnen. Die Selektion der Objektmenge (Stichprobenziehung) und Zusammenstellung der Datenbasis erfolgte durch Experten des Versandhauses. Die Datenbasis umfasst 5921 Kreditfälle, die in die Bonitätsklassen „gut“ und „schlecht“ eingeordnet sind. Es liegen 2982 gute und 2939 schlechte Kredite vor. Damit ist eine wesentliche Anforderung an die Stichprobe erfüllt, dass nämlich auch die schlechten Kredite in einer statistisch aussagekräftigen Anzahl vorkommen. Im Allgemeinen treten nämlich die guten Kredite in der Gesamtheit aller abgewickelten Kreditfälle wesentlich häufiger auf.

Die Datenbasis enthält 107 Attribute, alle mit ganzzahligen Attributwerten. Die Stichprobe bezog sich auf einen Zeitraum von 12 Monaten. Hier war vor allem die Entwicklung des Zahlungsverhaltens der Kunden wichtig. Daraus resultierte dann die Einstufung der Kunden in die Bonitätsklassen „gut“ oder „schlecht“. Die erfassten Attribute enthalten Angaben zu der Fälligkeitsstruktur, den Buchungsvorgängen, wie etwa Einzahlungen, Retouren und Annahmeverweigerungen, und verschiedenen Saldowerten. Beispiele sind die Kontodauer in Monaten, der Gesamtwert der Einzahlungen, die Anzahl der Retouren sowie die Anzahl der Verweigerungen im Beobachtungszeitraum. Eine Vielzahl der Merkmale repräsentiert dabei Prozentangaben und Maßzahlen. Daneben treten in der Datenbasis zwei Sondercodierungen –9999999 und –9999998 auf. Es handelt sich dabei nicht um tatsächlich erfasste Werte, sondern diese drücken aus, dass die Berechnung entsprechender Quotienten-Maßzahlen problembehaftet oder nicht möglich war (z.B. Division durch Null).

4.4 Datentransformation und Merkmalsauswahl

Aufgrund der vorliegenden Datenbasis und der großen Zahl von Attributen wurde im Vorfeld, vor dem eigentlichen Einsatz der Klassifikationsverfahren, viel Zeit für Datentransformationen und Merkmalsauswahl verwendet. Eine kleinere Zahl von Merkmalen kann sich dabei sehr positiv auf das Laufzeitverhalten entsprechender Algorithmen auswirken. Daneben wurde die Auswirkung von alternativen Transformationsverfahren auf die Klassifikationsergebnisse untersucht.

Im Zuge einer explorativen Datenanalyse wurden die wichtigsten statistischen Maßzahlen ermittelt und die Häufigkeitsverteilung der Attribute durch Histogramme dargestellt. Alle Attribute sind statistisch signifikant von der Normalverteilung verschieden. Eine Korrelationsanalyse zeigt, dass die Attribute mit dem Zielattribut (gut oder schlecht) nur den maximalen Betrag der Abweichung von 0,5 haben, einige Attribute untereinander aber hohe Korrelationen aufweisen. Der Autor geht hierbei nicht ein, ob auch Attributskombinationen überprüft wurden. Sollte dies nicht der Fall gewesen sein, so wurden wichtige Zusammenhänge außer Acht gelassen. Natürlich können auch Attributkombinationen eine wichtige Rolle spielen, da auch zwischen den Attributen Zusammenhänge bestehen können. Diese wären so einfach übersehen worden.

Als nächstes wurden verschiedene Verfahren auf die Datenbasis angewandt, und die Datenbasis A, B und C genannt.

Für die Konstruktion der Klassifikatoren stehen 103 Attribute zur Verfügung. Aus praktischen Gründen wurde daher eine Vorselektion der Attribute vorgenommen.
Eingesetzt wurden verschiedene Methoden und dimensionsreduzierende Verfahren der Statistik.
Mit einem Chi-Quadrat-Unabhängigkeitstest wurden die Zusammenhänge jedes Attributs mit der Bonitätsklasse untersucht. Der Chi-Quadrat-Unabhängigkeitstest konnte keinen Hinweis auf besondere signifikante Attribute geben.
Ebenso führte eine Untersuchung der Attribute mit dimensionsreduzierenden Verfahren (hier: Hauptkomponentenanalyse) zu nichts. Es konnten so keine Hinweise auf besonders wichtige und daher auszuwählende Attribute gefunden werden.

Für die Attributsauswahl wurde eine schrittweise Diskriminanzanalyse benutzt. Die schrittweise Diskriminanzanalyse wurde in diesem Fall als Hilfsmittel zur Vorauswahl der Attribute verwendet, was sich nicht zuletzt durch die guten Ergebnisse beim Einsatz von Entscheidungsbäumen auf den selektierten Attributen begründen lässt. Pro Datensatz werden dann ~ 30 Attribute vorselektiert.

4.5 Empirische Ergebnisse bei der Anwendung von Entscheidungsbaum-Klassifikatoren

Die Entscheidungsbäume wurden mit einem der ID3-Methode ähnlichen Verfahren erstellt.

Für die ersten Tests und zur Ermittlung der geeigneten Parameter wurde aus jedem der drei Datenbasen (A, B und C) eine geschichtete Zufallsstichprobe vom Gesamtumfang 1000 gezogen. Diese enthält jeweils 500 gute und 500 schlechte Kredite.

Einer der so erzeugten Bäume ist in Abbildung 12 dargestellt. Ausgangspunkt ist die Wurzel des Baumes, die 500 schlechte und 500 gute Kredite enthält. Als bestes Attribut für das Splitting wird Attribut CHAR555 herangezogen. Hier ist leider völlig unklar, um was für ein Attribut es sich handelt. Dadurch, dass vom Autor nur ein nichtverständliches Kürzel verwendet wird, geht einiges an Nachvollziehbarkeit und Verständnis verloren.
Die Ausgangsmenge wird in die beiden Teilmengen mit {CHAR555 < 3,5} und {CHAR555 >= 3,5} partitioniert; die Informationen zur Partitionierung sind jeweils an den Kanten des Baumes angegeben, darüber steht das zugehörige Verzweigungsmerkmal. Durch diese Aufteilung der Objektmenge gelangen 168 schlechte und 430 gute Kredite in den linken Nachfolgeknoten sowie 332 schlechte und 70 gute Kredite in den rechten Nachfolgeknoten. Betrachtet man die weitere Aufspaltung der 402 Kreditfälle im rechten Nachfolgeknoten (dem Knoten 2 der Ebene 2), so wird hier anhand des Attributs CHAR557 verzweigt. Es gelangen 294 schlechte und 70 gute Kredite in seinen rechten Nachfolgeknoten. Dagegen enthält der linke Nachfolger nur 38 schlechte Kredite; es handelt sich um einen homogenen Knoten, so dass hier eine weitere Aufteilung keinen Sinn macht.

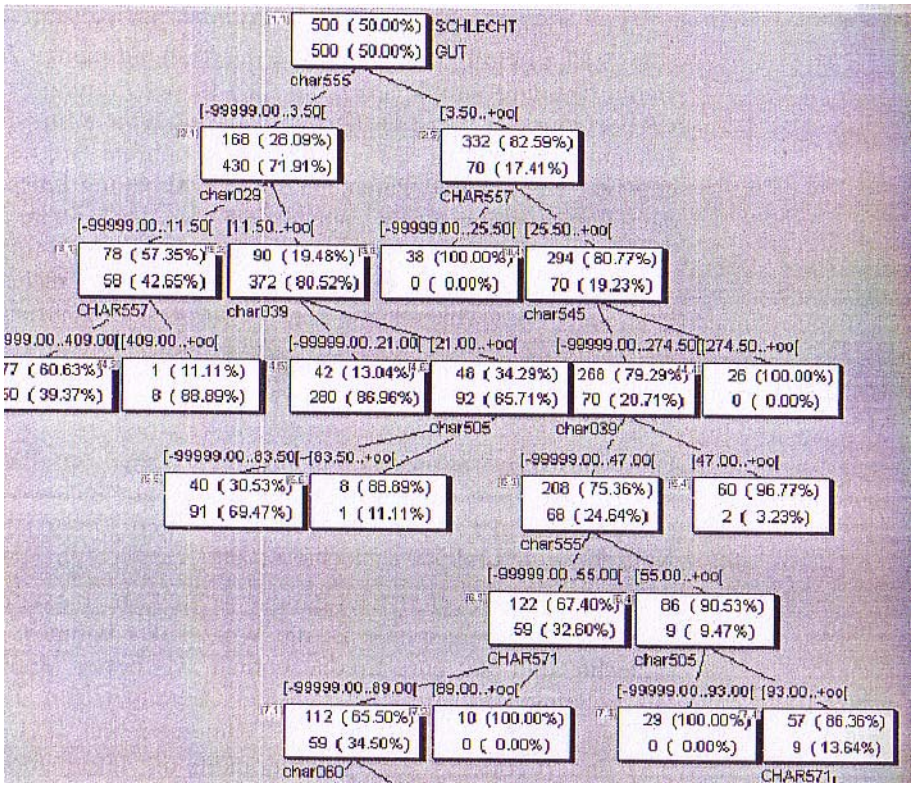


Abbildung 12: Beispiel eines Entscheidungsbaumes

Die hierbei erzielte Klassifikationsgenauigkeit liegt bei 79 - 80%. Die absolut schlechteste Klassifikation des Tests beträgt immerhin noch 75%.

Der konstruierte Baum ist relativ kompakt. Die aus dem Baum abgeleiteten Entscheidungsregeln enthalten selten mehr als sechs Bedingungen pro Regel. Auch die Anzahl der Regeln ist mit durchschnittlich 15 Regeln moderat. Sie bieten sich damit zum Entwurf eines einfachen Expertensystems an, das sehr schnell über die Bonität des Kunden urteilen kann, zum Beispiel während eines Bestellvorgangs über das Telefon oder das Internet. Zwei Beispiele für Entscheidungsregeln:

If CHAR007 = { -99999.00 .. 1,92 then Klasse = Schlecht)
Mit geschätzter Genauigkeit: 0,83
Regel gilt für 2245 Fälle
Klassifiziert als (Schlecht: 1867; Gut: 378)

If CHAR007 = [1,92 .. ∞) AND CHAR022 = [0,69 .. ∞)
then Klasse = Gut
Mit geschätzter Genauigkeit: 0,82
Regel gilt für 2400 Fälle
Klassifiziert als (Schlecht: 438; Gut: 1962)

4.6 Fazit

Verschiedene Entscheidungsbaumverfahren wurden auf einen realen Datensatz aus dem Bereich der Bonitätsprüfung angewandt. Dabei wurde eine durchschnittliche Klassifikationsgenauigkeit von 79% erzielt.

Ein Großteil der Zeit wurde auf die Datenvorbereitung, die Erprobung verschiedener Transformationsverfahren und die Auswahl der Attribute verwendet. Transformationen der Ursprungsdaten müssen mit Bedacht durchgeführt werden, da sie Zeit und Ressourcen in Anspruch nehmen und zudem einen Informationsverlust bedeuten. Ein weiteres Problem stellt die Auswahl der zu benutzenden Attribute dar. Klassische Auswahlverfahren, wie etwa die schrittweise Diskriminanzanalyse haben hier eher unterstützenden Charakter. Durch die Hinzunahme weiterer Attribute steigt zugleich die Rechenzeit deutlich an. Die erzielten Ergebnisse zeigen, dass beim vorliegen vieler Attribute die Attributssauswahl weniger kritisch als erwartet ist. Insbesondere scheint es auszureichen, mit rund 10% der insgesamt zur Verfügung stehenden über 100 Attributen zu arbeiten. Inwieweit dies noch unterboten werden kann und welche Attribute sich als besonders dirkriminanzstark erweisen, muss im Einzelfall untersucht werden.

Neben der Prognosefunktion liefern Entscheidungsbäume auch eine strukturierte und einfach zu interpretierende Darstellung der zur Klassifikation benutzten Attribute. Mit anderen Klassifikationsverfahren aus dem Bereich der Künstlichen Intelligenz, beispielsweise Multi-Layer-Perceptron-Netzen, lassen sich im Prinzip vergleichbare Ergebnisse erzielen. Für den Anwender stellt das Neuronale Netz jedoch lediglich eine Black Box dar, entsprechende Wirkungszusammenhänge und Attribute bleiben weitgehend unerkannt. Die vorliegenden Ergebnisse geben einen Einblick in die Leistungsfähigkeit von Entscheidungsbaum-Algorithmen bei realen Problemstellungen, wie sie die Bonitätsprüfung darstellt.

5 Fazit

Data Mining wird dazu benutzt, um unbekannte Zusammenhänge in einer für einen Menschen nicht zu bewältigenden Datenmenge zu finden. Doch die erhaltenen Ergebnisse müssen aus mehreren Gründen mit Vorsicht behandelt werden.
Zuerst einmal muss man sich fragen, auf welche Weise die für die Analyse benötigten Daten ermittelt werden. Jede Methode der Datenerhebung hat ihre speziellen Schwachstellen und Tücken. Ein großes Problem besteht zum Beispiel darin, Kunden zur Mitarbeit zur bewegen. Denn ohne ihre Einwillingung Kundendaten zu verwenden, verstößt gegen Datenschutzbestimmungen. Zudem ist es fraglich, ob all die Kundendaten verlässlich sind. So wird ein zufriedener Kunde gerne seine Meinung in eine Auswertung einfließen lassen, aber wie sieht es mit der genauso nötigen Kritik von unzufriedenen Kunden aus? Melden sich diese doch meistens nicht persönlich, sondern drücken ihre Unzufriedenheit durch einen Wechsel zu Konkurrenz aus.
Weiterhin ist auch die Merkmalsauswahl ein wichtiger Faktor - nur die Attribute können ausgewertet werden, die erfasst wurden. Daher stellt sich natürlich die Frage nach dem ob und wie der Auswahlgestaltung der in die Datenbasis zu übernehmenden Attribute. Nach welchen Kriterien kann ich ein Attribut als unwichtig ausschließen? Und was ist mit Merkmalskombinationen? Diese sollten schließlich nicht vernachlässigt werden, aber um alle möglichen Kombinationen auszuwerten, ist meist nicht genug Zeit und Rechenleistung vorhanden. Denn hier kommt man beispielsweise allein schon bei 50 Merkmalen sehr schnell in einen Bereich, in dem es einfach zu mühsam und unübersichtlich wird, alle Kombinationen zu berücksichtigen.
Hier ist daher die Anwendung der richtigen statistischen Bewertungsverfahren von großer Bedeutung. Aber welche sind nun in der Praxis die „richtigen“? Fehler, die durch eine Fehleinschätzung auftreten könnten, sollen ja vermieden werden. Die Kernfrage zur Lösung dieses Dilemmas lautet daher: Zu wie viel Prozent ist das Ergebnis der Vorhersage zutreffend?
Schlussendlich muss man sich auch darüber im klaren sein, dass die Data-Mining-Algorithmen stets nur auf ein Abbild der Realität anwendbar sind. Also kann ein Ergebnis einer Data-Mining-Analyse nie 100%ig zutreffend sein, wenn schließlich die Realität ins Spiel kommt, die immer allerlei Unwägbarkeiten mit sich bringt. Vielleicht,

wird die Technik eines Tages soweit sein, die Realität gänzlich im Computer abbilden zu können, doch solange Menschen diese Computer bedienen, wird das Data Mining die beste Möglichkeit bleiben, wichtige Zusammenhänge zu ermitteln.

Anhang

Übungsbeispiele

Aufgabe 1 Kündigung

Gesucht ist die Antwort auf die Frage, welche Kunden besonders kündigungsgefährdet sind. Das Zielattribut sei also „Kündigung“.

Aufgabe1:			
Kündigungswahrscheinlichkeit für Kunden?			
Kundennummer	Vertragsdauer	Berufsstatus	Kündigung
1	Mittel	Nicht-Erwerbstätig	Nein
2	Niedrig	Beamte	Ja
3	Niedrig	Angestellte	Ja
4	Mittel	Selbständige	Nein
5	Mittel	Angestellte	Nein
6	Hoch	Nicht-Erwerbstätig	Nein
7	Mittel	Angestellte	Nein
8	Hoch	Beamte	Nein
9	Hoch	Selbständige	Ja
10	Mittel	Nicht-Erwerbstätig	Nein

Entropie (Kunden 1-10, Kündigung) =
-0,3 * log2 0,3 – 0,7 * log2 0,7 = 0,881;

InfoGain(Kunden1-10, Kündigung, Vertragsdauer) =
Entropie (Kunden 1-10, Kündigung)
- (2/10 * Entropie (SVertragsdauerniedrig, Kündigung)
+ 5/10 * Entropie(SVertragsdauermittel, Kündigung)
+ 3/10 * Entropie(SVertragsdauerhoch, Kündigung) =
0,881 – (2/10 * 0 + 5/10 * 0 + 3/10 * (- 1/3 * log2 1/3 – 2/3 * log2 2/3) =
0,881 – 0,275 = 0,606

InfoGain(Kunden1-10, Kündigung, Berufsstatus) =

Entropie (Kunden 1-10, Kündigung)

- (3/10 * Entropie(Snicht-erwerbstätig, Kündigung)

+ 2/10 * Entropie(Sbeamte, Kündigung)

+ 3/10 * Entropie(Sangestellte, Kündigung)

+ 2/10 * Entropie(Sselbständige, Kündigung)) =

0,881 –

(3/10 * 0 + 2/10 * (-½ * log2 ½ - ½ * log2 ½)

+ 3/10 * (-⅓ * log2 ⅓ – 2/3 * log2 2/3)

+ 2/10 * (-½ * log2 ½ - ½ * log2 ½) =

0,881 – (2/10 * 1 + 0,275 + 2/10 * 1) =

0,881 – 0,675 = 0,206

Alle Knoten sind nun homogen, und somit ist das Verfahren beendet.
Der Baum ist in Abbildung 10 abgebildet.

Aufgabe 2 Tennisspielen

Gesucht ist die Antwort auf die Frage, ob am Zieltag Tennisspielen möglich war. Das Zielattribut sei „Tennis spielen?“.

Aufgabe2:					
Ist am Vortag Tennis spielen möglich gewesen?					
Nummer	Aussicht	Temperatur	Luftfeuchtigkeit	Wind	Tennis spielen?
1	sonnig	heiß	hoch	nein	nein
2	sonnig	heiß	hoch	ja	nein
3	bewölkt	heiß	hoch	nein	ja
4	regen	mild	hoch	nein	ja
5	regen	kalt	normal	nein	ja
6	regen	kalt	normal	ja	nein
7	bewölkt	kalt	normal	ja	ja
8	sonnig	mild	hoch	nein	nein
9	sonnig	kalt	normal	nein	ja
10	regen	mild	normal	nein	ja
11	sonnig	mild	normal	ja	ja
12	bewölkt	mild	hoch	ja	ja
13	bewölkt	heiß	normal	nein	ja
14	regen	mild	hoch	ja	nein

Entropie(Nummer 1-14, Tennis spielen?)=

- 9/14 * log2 9/14 – 5/14 log2 5/14 = - (-0,4098 – 0,5305) = 0,9403

InfoGain(Nummer 1-14, Tennis spielen, Aussicht) =

Entropie(Nummer 1-14, Tennis spielen?)

– (5/14 * Entropie(Aussichtsonnig, Tennis spielen)

+ 4/14 * Entropie(Aussichtbewölkt, Tennis spielen)

+ 5/14 * Entropie (Aussichtregen, Tennis spielen)) =

0,9403 – (5/14 * (- 3/5 * log2 3/5 – 2/5 * log2 2/5) + 4/14 * 0

+ 5/14 * ((- 3/5 * log2 3/5 – 2/5 * log2 2/5)) =

0,9403 – 0,6935 = 0,247

InfoGain(Nummer 1-14, Tennis spielen, Temperatur) =

Entropie(Nummer 1-14, Tennis spielen?)

–(4/14 * Entropie(Temperaturheiß, Tennis spielen)

+ 6/14 * Entropie(Temperaturmild, Tennis spielen)

+ 4/14 * Entropie (Temperaturkalt, Tennis spielen)) =

0,9403 – (4/14 * (- ½ log2 ½ - ½ log2 ½) + 6/14 * (- 2/3 log2 2/3 – 1/3 log2 1/3)

+ 4/14 * (- ¾ log2 ¾ - ¼ log2 ¼) =

0,9403 – (4/14 * 1 + 6/14 * (0,3899 + 0,5283) + 4/14 * (0,3112 + 0,5) =

0,9403 – 0,9109 = 0,029

InfoGain(Nummer 1-14, Tennis spielen, Luftfeuchtigkeit) =

Entropie(Nummer 1-14, Tennis spielen?)

–(7/14 * Entropie(Luftfeuchtigkeithoch, Tennis spielen)

+ 7/14 * Entropie(Luftfeuchtigkeitnormal, Tennis spielen))=

0,9403 – (½ * (- 4/7 log2 4/7 – 3/7 log2 3/7) + ½ (- 1/7 log2 1/7 – 6/7 log2 6/7) =

0,9403 – (1/2 * (0,4613 + 0,5238) + ½ * (0,4010 + 0,1906) =

0,9403 – (0,4925 + 0,2958) = 0,152

InfoGain(Nummer 1-14, Tennis spielen, Wind) =

Entropie(Nummer 1-14, Tennis spielen?)
– (8/14 * Entropie(Windnein, Tennis spielen)
+ 6/14 * Entropie(Windja, Tennis spielen)) =
0,9403 – (8/14 * (- 2/8 log2 2/8 – 6/8 log2 6/8)
+ 6/14 * (- 2/6 log2 2/6 – 4/6 log2 4/6)) =
0,9403 – (8/14 * (½ + 0,3112) + 6/14 * (0,3899 + 0,5283) =
0,9403 – 0,8570 = 0,0833

➔ InfoGain für Aussicht ist am größten, also wird im 1. Knoten Aussicht ausgewählt

Die Werte bei „bewölkt“ sind schon homogen, aber die anderen Knoten müssen weiter berechnet werden.

Knoten „sonnig“:
InfoGain(Nummer 1-14, Tennis spielen, Temperatur) =
Entropie(Nummer 1-14, Tennis spielen?)
– (2/5* Entropie(Temperaturheiß, Tennis spielen)
+ 2/5 * Entropie(Temperaturmild, Tennis spielen)
+ 1/5 * Entropie (Temperaturkalt, Tennis spielen)) =
0,9403 – (2/5 * 0 + 2/5 (-1/2 log2 ½ - ½ log2 ½) + 1/5 * 0) =
0.9403 – 2/5 = 0,5403

InfoGain(Nummer 1-14, Tennis spielen, Luftfeuchtigkeit) =
Entropie(Nummer 1-14, Tennis spielen?)
–(3/5 * Entropie(Luftfeuchtigkeithoch, Tennis spielen)
+ 2/5 * Entropie(Luftfeuchtigkeitnormal, Tennis spielen))=
0,9043 – (3/5 * 0 + 2/5 *0) = 0,9043

InfoGain(Nummer 1-14, Tennis spielen, Wind) =
Entropie(Nummer 1-14, Tennis spielen?)
–(3/5 * Entropie(Windnein, Tennis spielen) + 2/5 * Entropie(Windja, Tennis spielen))=
0,9403 – (3/5 * (- 2/3 log2 2/3 – 1/3 log2 1/3) + 2/5 * (-1/2 log2 ½ - ½ log2 ½) =
0,9403 – (3/5 * (0,3899 + 0,5283) + 2/5 * 1) = 0,9403 – 0,9509 = - 0,0106

Der größte InfoGain ist hier bei „Luftfeuchtigkeit“.
Nachdem nach diesem Attribut der Baum geteilt wurde, entstehen 2 homogene Knoten.

Jetzt folgt der rechte Knoten „regen“

InfoGain(Nummer 1-14, Tennis spielen, Temperatur) =
Entropie(Nummer 1-14, Tennis spielen?)
– (0/5* Entropie(Temperaturheiß, Tennis spielen)
+ 3/5 * Entropie(Temperaturmild, Tennis spielen)
+ 2/5 * Entropie (Temperaturkalt, Tennis spielen)) =
0,9043 – (3/5 * (- 2/3 log2 2/3 – 1/3 log2 1/3) + 2/5 (-1/2 log2 ½ - ½ log2 ½)=
=-0,0106

InfoGain(Nummer 1-14, Tennis spielen, Luftfeuchtigkeit) =
Entropie(Nummer 1-14, Tennis spielen?)
– (2/5 * Entropie(Luftfeuchtigkeithoch, Tennis spielen)
+ 3/5 * Entropie(Luftfeuchtigkeitnormal, Tennis spielen))=
0,9403 – (2/5 * (-1/2 log2 ½ - ½ log2 ½) + 3/5 * (- 2/3 log2 2/3 – 1/3 log2 1/3)=
=-0,0106

InfoGain(Nummer 1-14, Tennis spielen, Wind) =
Entropie(Nummer 1-14, Tennis spielen?)
–(3/5 * Entropie(Windnein, Tennis spielen) + 2/5 * Entropie(Windja, Tennis spielen))=
0,9403 – (3/5 * 0 + 2/5 * 0) = 0,9403

Es wird also nach Wind geteilt, und es liegen 2 homogene Knoten vor.
Der Baum wird in Abbildung 13 gezeigt. Damit sind alle Knoten homogen, und somit ist das Verfahren beendet.

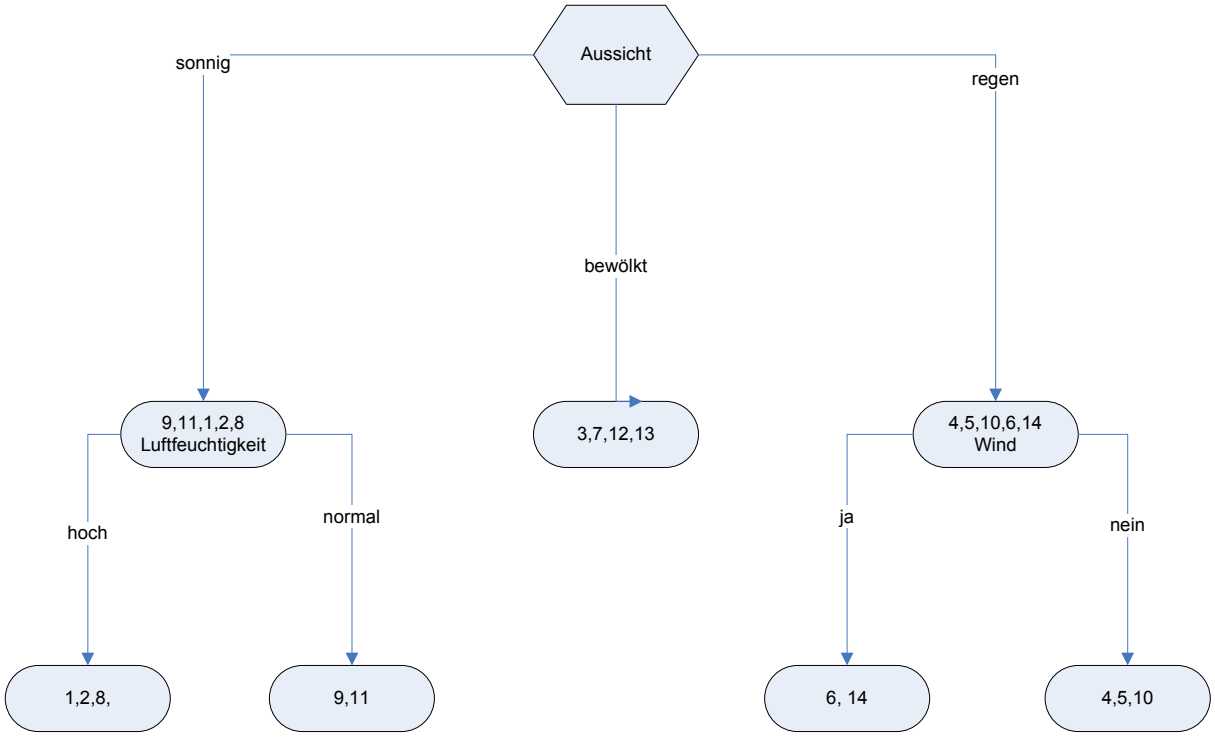


Abbildung 13: Baum Aufgabe2

Erklärung

Hiermit erkläre ich, Christian Ulrich, dass ich diese Arbeit selbständig verfasst habe, dass ich habe diese Arbeit nicht anderweitig für Prüfungszwecke vorgelegt habe, dass ich keine Hilfsmittel als die angegebenen Quellen und Hilfsmittel benutzt habe sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

Literaturverzeichnis

Alpar, P:

Data Mining im praktischen Einsatz, Braunschweig, 2000

Chamoni, P, Gluchowski, P:

Analytische Informationssysteme – Data Warehouse, On-Line Analytical Processing, Data Mining, Berlin, Heidelberg, 1998

Ester, M, Sander, J:

Knowledge Discovery in Databases, Techniken und Anwendungen, Berlin, 2000

Fayyad, U.M.:

Advances in knowledge discovery and data mining, Menlo-Park, Calif., 1996

Krahl,Daniela:

Data Mining – Einsatz in der Praxis, Bonn, 1.Auflage 1998

Weigand, D.:

Lernen mit Entscheidungsbäumen. Elektronische Publikation, URL am 23.02.01: http//www2.informatik.uni-erlangen.de/IMMD-II/Lehre/WS98_99/Machine_Learning/Vortraege/Entscheidungsbaeume/Entscheidungsbaeume.pdf