Freuden und Fallen des Data Mining

von

Alex Flemming

Alex Flemming
Südstraße 4
D-59439 Holzwickede

e-mail: flemming@vonformat.com

Übersicht

1. Vorwort – von Barry de Ville

When I first met Alex in December of 1993 the term ***Data Mining*** had not come into fashion yet. I was in the United Kingdom to talk to a group of European developers in order to promote an advanced version of KnowledgeSEEKER, a product that used decision tree approaches to automatically extract trends, patterns and relationships from collections of data. Decision trees are a type of software that use a variety of statistical and artificial intelligence techniques to automate the process of extracting knowledge from data. They have become one of the most popular data mining techniques … primarily because of their combined ease of use and effectiveness versus almost any other method.

As the inventor of KnowledgeSEEKER I was most certainly committed to this growing field of data mining. I'm not sure that any of us would have predicted that Alex too would become a solid convert to this emerging discipline but, as you'll see in the treatment that he has offered, he is not only a convert but a very accomplished one at that. Alex was, perhaps, an unlikely convert to the field of data mining. Although he clearly had a great deal of curiosity -- a hallmark of an early adopter -- he was also exceptionally applied in his thinking and very practical in terms of the possibilities offered by this new technology – almost to the point of being conservative. In short, it was not clear that Alex would have the desire for the kind of sometimes somewhat wild-eyed evangelism that was in front of him as the growing area of data mining slowly matured and found useful and rewarding applications over the decade following our initial meeting.

It was refreshing to see the interest that Alex had in this new technology: he was not just interested in a mathematical explanation of how methods work but wanted to go further into the use of the algorithms, their drawbacks and how they could and should be applied to achieve the best results. He was constantly focused on how to use the tools for business advantage. He had a mind for detail as well as the "big picture" view and so constantly probed into seemingly minor or even circumstantial areas in order to feed his notions of how to translate technology into business opportunity.

Alex brought an interesting background into his examination of new technology. Although he has been based in Germany for over 25 years he originally comes from London, England. So his experience incorporates lessons learned by a deep experience and exposure to multiple cultures. His academic background lies in historical studies; however, when he graduated in 1975 he became involved in languages. After his graduation he emigrated to Germany and eventually became involved in marketing and management with MacMillan. His exposure to the business world at MacMillan led him into work in 1987 as a self-employed business consultant. In this capacity he specialized in integrated business automation in both Unix and Novell networks for small and medium-sized companies as well as for the PC, client-

server and mainframe surroundings. His management experience extends to a variety of business areas (banking, insurance, engineering, telecommunications, retail, logistics, wholesale) in both public and private sectors.

Data mining is a methodology that employs analytical methods, business processes and procedures, and related software tools to make sense of the volumes of data that enterprises -- big and small -- collect and store on a daily basis. So rather than become swamped in data, the enterprise can in fact profit from the data so as to use the knowledge that is contained in data to design, develop and deliver enterprise programs more effectively to their customers and stakeholders. Data has always accumulated and now, with data mining technologies and methodologies, we have a means of profiting from it.

But data mining serves another, increasingly more import purpose: since the algorithms run in an automatic fashion it is possible to carry out complex analyses of data in a fraction of the time that a skilled data analyst would take. While this has always been a desirable characteristic in favour of data mining today it is a necessary pre-condition for success in the world of electronic commerce. These factors have had a major influence on the growth and acceptance of data mining over the past decade. It seems that data mining is poised for even greater growth in the future. A recent issue of ***Time*** magazine voted data mining as one of the 10 hottest jobs for the long term future.

Since the field of data mining is barely ten years old there is still a lot of mystery surrounding what data mining actually is and how it can be exploited for business advantage. It is at this point that Alex's contribution in this volume will prove its worth. Alex has preserved all of his native enthusiasm, his ever-deepening understanding of the business market place and his by now quite seasoned understanding of data mining tools, techniques and applications in treatment he offers in this volume. He presents a very practical view of how data mining can be used to extract profit from data.

Alex has a penchant for digging into details to find out how things work. At the same time he takes a broad view on business processes and market forces generally. And he knits this knowledge and viewpoint together in a way that is both easy to follow yet has enough detail and practical content to provide a great survival guide in the new world of data mining. He addresses the business user as well as the decision makers who need to know what as well as an insight into how and why. The specialist who needs to translate specialized experience into a business environment (with all the political difficulties involved) will also reap rewards from Alex's work here.

This is a timely and valuable document. I suspect that, given Alex's curiosity, his interest in creating value, and his gifts as a communicator it will not be the last. I expect that both the author and the reader will find this introduction useful and informative and I wish them well in their data mining ventures now and in the future.

June 15, 2000 Barry de Ville

(Es folgt die deutsche Übersetzung von Alex Flemming)

Als ich Alex im Dezember 1993 zum ersten Mal kennenlernte, war der Begriff ***Data Mining*** noch nicht in Mode gekommen. Ich befand mich im Vereinigten Königreich, um mit einer Gruppe europäischer Entwickler zusprechen, wo wir eine vorläufige Version von KnowledgeSEEKER vorstellen wollten. Da werden Entscheidungsbaumansätze verwendet, um automatisch Muster, Trends und Verbindungen aus Datenmengen zu extrahieren. Entscheidungsbäume sind einer Software, die eine Reihe statistischer und künstlicher Intelligenz Techniken verwenden, um den Vorgang der Wissensextrahierung aus Daten zu automatisieren. Sie sind unter den populärsten Data Mining Techniken zu finden ... hauptsächlich weil sie im Vergleich zu fast jeder anderen Methode eine einfache Bedienung mit hoher Effektivität verbinden.

Als Erfinder von KnowledgeSEEKER war ich sicher diesem wachsenden Feld des Data Minings verschrieben. Ich bin mir jedoch keineswegs sicher, dass einer von uns hätte vorhersagen können, dass Alex zu einem festen Überzeugten dieser heraufkommenden Disziplin werden würde, aber – wie Sie in seinem hier angebotenen Werk sehen werden – ist er nicht nur ein Überzeugter, sondern dazu auch ein äußerst Fähiger. Alex war vielleicht ein ungewöhnlicher Bekehrter für das Data Mining. Obwohl er ein deutliches Maß an Neugierde mitbringt – ein Zeichen der frühen Überläufer – seine angewandte Denkweise stellt Außergewöhnliches dar und sah die Möglichkeiten dieser neuen Technologie stets aus praktischer Sicht – fast bis zum konservativen Blickwinkel. Kurzum war es nicht klar, dass Alex die Lust für die manchmal großäugige Missionarsarbeit haben würde, die ihm bevorstand, während der wachsende Bereich des Data Mining langsam heranreifte und im Jahrzehnt nach unserem ursprünglichen Kennenlernen sich nützliche sowie lohnende Applikationen fanden.

Das Interesse, dass Alex für diese neue Technologie aufbrachte, war erfrischend: Er interessierte sich nicht nur für die mathematischen Erklärungen zur Funktionsweise der Methoden, sondern vielmehr wollte er in die praktische Anwendung der Algorithmen, ihre Nachteile und wie sie am besten angewandt werden konnten und sollten, um die besten Ergebnisse zu erzielen. Er war stets darauf bedacht, diese Werkzeuge zum geschäftlichen Vorteil einzusetzen. Er bedachte sowohl die Einzelheiten als auch verlor er nicht den Blick für das Gesamtbild und bohrte somit ständig in scheinbar kleinen oder gar zufälligen Gebieten, um seinen Vorstellungen Nahrung zu geben, Technologie in geschäftlichen Möglichkeiten umsetzen zu können.

Alex brachte einen interessanten Hintergrund in seine Prüfung der neuen Technologie mit hinein. Obwohl er seit über 25 Jahren in Deutschland sein Zuhause hat, stammt er aus London, England. Sein Erfahrungsschatz beherbergt die Lektionen der tiefgreifenden Erfahrungen eines Menschen, der mehrere Kulturen ausgesetzt wurde. Sein akademischer Hintergrund liegt in der Geschichte, jedoch nach seinem Abschluß im Jahre 1975 beschäftigte er sich mit Sprachen. Direkt nach dem Universitätsabschluß emigrierte er nach Deutschland und arbeitete nach und nach im Marketing und Management bei MacMillan. Die Erfahrungen der Geschäftswelt bei MacMillan führten 1987 zum Beginn seiner Arbeit als selbständiger Berater. In dieser Eigenschaft spezialisierte er sich in der integrierten Geschäftsautomation sowohl im UNIX- als auch im NOVELL-Bereich für kleine und mittlere Firmen. Er beschäftigte sich ebenfalls mit dem PC, Client-Server sowie Mainframe Umgebungen. Seine Management Erfahrung bezieht er aus einer Reihe Geschäftsfelder (Bankwesen, Assekuranz, Ingenieurwesen, Telekommunikation, Einzelhandel, Logistik, Großhandel) im öffentlichen sowie im privaten Sektor.

Data mining ist eine Methodologie, die analytische Methoden, Geschäftsprozesse und Prozeduren und artverwandte Software Werkzeuge einsetzt, um einen Sinn aus den Bergen von Daten zu ziehen, die Gesellschaften – groß und klein – täglich sammeln und speichern. Anstelle der Erstickung in Daten, kann die Gesellschaft in der Tat aus diesen Daten einen Profit schlagen, in dem sie das Wissen verwendet, das in den Daten enthalten ist, um Gesellschaftsaktionen zu gestalten, zu entwickeln und ihren Kunden und Anteilseigner effektiver beliefern zu können. Daten haben sich angehäuft und jetzt stehen uns mit Data Mining Techniken und Methodologien die Mittel zur Verfügung, einen Gewinn daraus zu schlagen.

Aber Data Mining dient einen anderen, zunehmend wichtigeren Zweck: Da Algorithmen in einer automatischen Art laufen, ist es möglich, komplexe Datenanalysen in einem Bruchteil der Zeit durchzuführen, die ein Fachanalytiker brauchen würde. Da diese wünschenswerte Eigenschaft immer für Data Mining gesprochen hat, ist sie zu einer notwendigen Vorbedingung des Erfolgs in der heutigen Welt des elektronischen Handels geworden. Diese Faktoren haben einen großen Einfluß auf die Verbreitung und auf die Anerkennung von Data Mining während des letzten Jahrzehnts ausgeübt. Es scheint, als stehe Data Mining vor einer noch größeren Entwicklung in der Zukunft. Eine kürzlich erschienene Ausgabe von ***Time*** Magazine erhob Data Mining zu einem der 10 heißesten Berufszweige der langfristigen Zukunft.

Da der Bereich des Data Mining kaum 10 Jahre alt ist, hüllt sich ihre tatsächliche Bedeutung noch im Geheimnis und die Ausnutzung zum geschäftlichen Vorteil steht noch nicht fest. Gerade bei diesem Punkt wird sich der Wert von Alex Beitrag in diesem Werk beweisen. Alex hat seinen ganzen ihm eigenen Enthusiasmus erhalten, seine immer tiefer werdende Erfahrung des Marktplatzes des Business und sein inzwischen ausgereiftes Verständnis für Data Mining Werkzeuge, Techniken und Applikationen in der hier vorliegenden Arbeit. Er präsentiert eine sehr praktische Ansicht darüber, wie man mit Data Mining einen Gewinn aus Daten extrahieren kann.

Alex hat eine Vorliebe dafür, in den Details zu graben, um herauszufinden, wie Dinge funktionieren. Gleichzeitig betrachtet er Geschäftsprozesse und Marktkräfte im Allgemeinen aus der Großansicht heraus. Und er bindet dieses Wissen und Ansicht in einer Art zusammen, die sowohl leicht verständlich als auch genügend Detail und praktischen Inhalt aufbietet, um einen großen Überlebensführer in der neuen Welt des Data Mining darzustellen. Er spricht sowohl den Business User als auch die Entscheider an, die das Was als auch einen Einblick in das Wie und das Warum benötigen. Der Spezialist, der Expertenerfahrung in eine Business Umgebung (mit allen damit verbundenen politischen Schwierigkeiten) umsetzen muß, wird ebenfalls aus dieser Arbeit seinen Vorteil zeihen können.

Dies ist ein zeitgemäßes und wertvolles Dokument. Wegen Alex Neugierde, seiner Hingabe zur Wertschöpfung und seines Kommunikationstalentes vermute ich, dass dies nicht sein letztes Werk sein wird. Ich gehe davon aus, dass sowohl der Autor als auch der Leser dieses Vorwort als nützlich und informativ empfinden und ich wünsche ihnen bei ihren jetzigen und zukünftigen Data Mining Unternehmungen alles Gute.

15. Juni 2000 Barry de Ville

Danksagung

Bei dieser Gelegenheit möchten wir uns bei Nigel Bishop, Ken Ono und Mamdouh Refaat von ANGOSS, ganz besonders bei Prof. Dr. Dr. Peter J.A. Reusch von der Fachhochschule Dortmund, Prof. Hans Peter Möller von der RWTH Aachen, Frau Tülay Aksu und vor allem Barry de Ville sowie vielen anderen für Ihre Hilfe, Unterstützung, Motivation und Impulse danken. Vor allem gilt der Dank meiner Familie. Ohne ihr Verstöndnis wäre dieses Werk nie entstanden.

Alex Flemming

August 2000

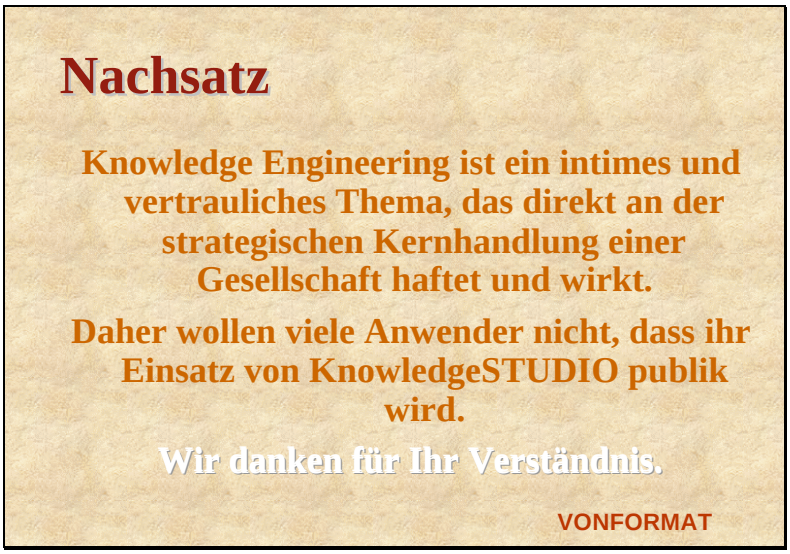
2. Einleitung

Dieses Werk entstand aus einem Kommunikationsnotstand. Auf dem Markt oder im Internet gab es aus dem deutschsprachigen Raum kaum umfangreiche Schriften zur praktischen Anwendung des Data Mining. Die vorhandenen Werke konzentrierten sich auf akademische und theoretische Belange des Data Mining und zielten vornehmlich auf die Spezialisten aus den mathematisch-statistischen Disziplinen.

Der Business Anwender, der die verschlossenen Werte aus seinen Datenbeständen zu Tage fördern wollte, fand hier kaum einen Zugang.

Dieser Notstand lag zum Teil darin, dass Data Mining eine neue Disziplin in der akademischen sowie in der praktischen Welt darstellte. Der Bedarf an Hilfe, Anregungen und Anleitungen für die Praxisarbeit mit Data Mining wurde durch den sprunghaften Anstieg an Interesse für das Thema angetrieben; denn einige begannen im Data Mining die mögliche Quelle eines zukünftigen geschäftlichen Vorteils zu erkennen.

Gleichzeitig aber erzeugte die intime Natur vieler Datenbestände, Dateninhalte sowie Projektergebnisse weitere Hemmnisse, die inzwischen entstandenen und vage formulierten Bedürfnisse befriedigen zu können. Hier nehmen wir einen Ausschnitt aus einer offiziellen Erklärung der Firma VONFORMAT.



Diese Arbeit basiert auf realen Erfahrungen aus der Praxis. Echter Fälle werden dargestellt, ohne jedoch einen direkten Namen zu erwähnen. Bei Bedarf werden Pseudonyme verwendet, aus denen die Branche erkennbar ist, um die Einordnung des jeweiligen Falls verständlicher machen zu können.

Die aufbereiteten Szenarien sollen einen konstruktiven Umgang mit Data Mining in der Praxis erleichtern sowie einige Impulse für Fachleute geben.

Ferner hoffen wir, etliche Anregungen für eigene Projekte geben zu können, denn eine große Portion Phantasie wird benötigt, um die Visionen einiger potentieller Ergebnisse dieser Arbeitsweise vorab erblicken zu können.

Die Konzentration auf die Projektarbeit hat einen einfachen Hintergrund. Data Mining ist wirklich ein Prozeß und kein Produkt. Daher geht es uns um den Einsatz von Data Mining und nicht um akademische Unterschiede von Algorithmen, Techniken, Methoden oder gar Technologien. Hier geht es schlicht darum, wie ein Mehrwert aus den unterschiedlichsten Datenbeständen gewonnen werden kann.

3. Was ist Data Mining?

Mit dem Wachstum des Faches Data Mining steigt die Anzahl von Definitionen und Erklärungen. Wir möchten versuchen, eine gewisse Erleichterung im Verständnis des Themas durch die Einführung unterschiedlicher Betrachtungsweisen zu propagieren.

Um unseren praktischen Ansatz aufrechterhalten zu können, möchten wir den Fokus auf die theoretische oder akademische Betrachtungsweise anderen Autoren überlassen, und uns auf die für die Praxis erforderlichen Ansicht konzentrieren.

Wir beginnen mit der fachlichen Ansicht, schreiten dann zur technischen und zum Abschluß folgt die operative Betrachtungsweise.

Fachlich gesehen, geht es beim „Data Mining“ darum Informationen zu erschließen oder um Daten-Modelle für die Vorhersage zu erstellen.

Die zu gewinnenden Informationen werden als „beschreibendes“ Ziel verstanden. Hierbei besteht die Aufgabe darin, das Verständnis für eine Datenmenge durch die Aufdeckung von Mustern und Korrelationen – schlicht Zusammenhänge - zu gewinnen.

Man sucht nach handfesten Erklärungen und gar Beweisen für irgendein aufgetretenes Phänomen. Es kann sich um die Reaktion von Menschen auf eine bestimmte medizinische oder gesellschaftliche Behandlung oder gar um die Verhaltensmuster bei der Kundschaft eines bestimmten Konzerns handeln.

Als Beispiel nehmen wir eine Marktanalyse, die im Bereich Marketing versucht festzustellen, warum eine bestimmte Kundengruppe eine erhöhte Kündigungsrate aufweist. Entsprechende Ergebnisse würden die Grundlage einer Gegensteuerung bilden.

In der Vorgehensweise sucht der Data Miner nach Anomalien – Ausreißern -, die von sonstigen vorhandenen Mustern abweichen oder er konzentriert sich auf die Inhalte der Muster. Interessant wird es, wenn man nach Anomalien und Mustern gleichermaßen Ausschau hält.

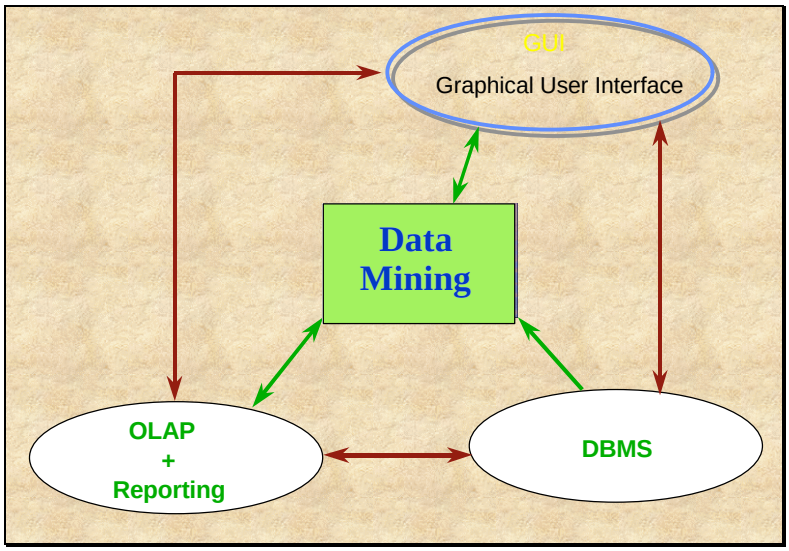
Obwohl der Wissensdurst nach diesem Schritt oft gestillt sein kann, erhebt sich möglicherweise die Frage, was mit dem neuen Wissen angefangen oder in längerer Frist angestellt werden kann.

An dieser Stelle beginnt man in Richtung Vorhersage zu denken. Hier werden Daten-Modelle aufgebaut, die in Anwenderapplikationen eingegliedert werden können.

In unserem Beispiel würde man versuchen, ein Alarm-System zu errichten, die im Vorfeld warnen kann, wenn die Vorbedingungen einer Kündigung aus unserer benannten Kundengruppe erfüllt sind und sich damit die Gefahr einer tatsächlichen Kündigung anbahnt.

Technisch gesehen, besteht „Data Mining“ aus der Anwendung von Software, Methoden und Techniken, die automatisch große Datenmengen untersuchen, um Muster, Verbindungen, Korrelationen und Anomalien - die unter den untersuchten Datenelementen vorkommen - besser verstehen zu können.

In dieser Hinsicht unterscheidet sich Data Mining grundsätzlich von Abfragen/Berichten und anderen vom Anwender gesteuerten und präparierten Werkzeuge, die präzise Eingaben vom Anwender benötigen, um den Vorgang einer Datenanalyse durchführen zu können.



Beziehung zwischen Anwenderoberfläche (Graphical User Interface - GUI), OLAP (OnLine Analytical Processing) und Datenbank (DataBase Management System)

Operativ gesehen, besteht „Data Mining“ aus der Untersuchung von Datenmengen aus der Sicht eines definierten Geschäftszieles, um zu erkennen,

(a) ob Wissen aus den identifizierten Mustern, Verbindungen und Korrelationen für das Geschäftsziel in relevanten Datenelementen gewonnen werden kann;

(b) ob und in welchem Umfang solches Wissen eine Entscheidungsbasis bildet; und
(c) welche Schritte oder Veränderungen in den Geschäftspraktiken eingeführt werden können, um das entscheidungsfähige Wissen aus dem Data Mining einzusetzen. Zum Beispiel gehören welche Erkenntnisse in die operative Geschäftspolitik und welche können mit den Ergebnissen anderer analytischer Werkzeuge, Firmenapplikationen oder operativen Systemen integriert werden.

Data Mining Initiativen bestehen aus der Anwendung von maschineller Intelligenz, durch die Untersuchung von großen Datenmengen unter Verwendung von modernen Techniken und menschlicher Intelligenz. Hierbei werden die Ergebnisse des Data Mining Vorgangs interpretiert und jene Resultate aus dem Verständnis des zu den Daten passenden geschäftlichen Umfeldes in dieses Umfeld eingegliedert.

Data Mining ist ein iterativer Vorgang – sowohl in der technischen wie auch in der organisatorischen Dimension – und, während der Anfangsphase, komplex, liefert aber einen beträchtlichen geschäftlichen Wert. Dieser geschäftliche Wert wird dann am größten, wenn Unternehmen Data Mining als strategisch wertvolle Kernkompetenz innerhalb ihrer Organisation ansehen.

Data Mining arbeitet parallel mit anderen analytischen Ansätzen und Techniken und bietet der Beschaffung sowie dem Management von Wissen in einem Unternehmen beträchtliche Erweiterungen.

Abfrage- und Berichtswerkzeuge liefern Antworten auf Fragen, die an eine Datenmenge gestellt werden. Um eine Antwort zu erhalten, müssen Sie jede Frage artikulieren und anschließend syntaktisch korrekt formulieren können. Die Antwort erscheint und die Bedeutung dieser Antwort müssen Sie selbst ermessen können. Weitergehende Fragen obliegen abermals der gleichen Technik.

Data Mining Technologie ist mächtiger und flexibler als solche Methoden, da die Artikulation und Formulierung der Fragen entfällt. Anstelle der Fragen stehen die Data Mining Algorithmen. Praktisch gesehen, können Sie sich auf die Interpretation der zusammenhängend entstandenen Antworten konzentrieren. Durch diese Entlastung keine Fragen – weder inhaltlich noch syntaktisch - ausdenken zu müssen, konzentriert sich Ihre Kraft auf die Interpretation der Ergebnisse. Hierdurch bewegen Sie sich deutlich näher am „Endprodukt Ergebnis“ als bei eher herkömmlichen Methoden und greifen nach neuen Erkenntnissen bzw. begreifen die Bedeutung dieser Erkenntnisse.

Die moderne IT-Welt bietet uns auch OLAP in diversen Variationen. OLAP steht für OnLine Analytical Processing und bedeutet die Versorgung einer größeren Menge Personen mit aktuellen (Online) analytischen Informationen. Diese Analysen werden auf der Basis bestehender Formeln erstellt. Zum Beispiel möchten man die Umsatzzahlen eines Konzerns in einem bestimmten Zeitraum betrachten. Das System

bietet die aktuellen Zahlen und teilt sie ein, nach Zeiträumen (Jahr, Quartal, Monat), nach geographischen Gebieten (Gesamt, Europa, Deutschland, Regionen, Bezirke) sowie nach Produkten (Gesamtkonzern, Ressort, Produktgruppen, Einzelprodukt, Elemente). Diese sogenannte Dimensionen können dann wiederum gegenüber Vorjahreszahlen, Planzahlen usw. gestellt werden. Diese Zahlen werden dann in einem schönen Bericht (neudeutsch: Report) angeboten sowie zusätzlich mit grafischen Darstellung wie Balkendiagramme und anderen Warnsignale wie rot für rückläufig gegenüber Vorjahr angereichert.

Entscheidend hierbei ist, dass die Aufbereitungsart der Zahlen meist vorgegeben ist und die vielen kleinen Formeln vorliegen. Zum Beispiel erhält das OLAP System einen aktuellen von DM 100.000. Diese Zahl wird auf ihre Bestandteile unterteilt und dementsprechend in die Quartale, Monate usw. eingeteilt. Ebenso könnte die erwartete Planzahl gegenüber der aktuellen Zahl gerechnet und die Abweichung als Zahl sowie als Prozentsatz im Falle eines Plus grün aufbereitet.

Der Input ist eine aktuelle Zahl mit Erklärungen der Bestandteile. Hieraus werden Informationen für die unterschiedlichen Blickwinkel eines Unternehmens ***anhand von vorhandenem Wissen*** errechnet.

Data Mining wurde dagegen geschaffen, um untersuchende Abfragen, Hypothesengenerierung und –prüfung sowie Wissensentdeckung zu betreiben. Es kann zum Beispiel als Entwicklungsvehikel benutzt werden, um die passendsten Berichte in einem Unternehmensumfeld während einer Planungsphase zu identifizieren.

Data Mining ist am sinnvollsten dort im Einsatz, wo der Ausgang von Fragen, Sorgen und Hypothesen sowie die betroffenen Faktoren von Beginn an unklar sind.

Hinweis für IT-Spezialisten

Darüber hinaus ist Data Mining bei der Errichtung von Data Warehouses und Data Marts ein wichtiges Werkzeug, denn erst durch den Data Mining Prozess, wird ein umfangreiches und flexibles Verständnis der notwendigen Schlüsseldimensionen und Zusammenhänge der Datenverbindungen ermöglicht, die zur Unterstützung der Data Mart oder Data Warehouse Datenstrukturen benötigt werden. Im Endergebnis entsteht meist ein Data Warehouse, das von geschäftlichen Belangen angetrieben wurde. Die technischen Feinheiten der fachlichen Data Warehouse Arbeit treten hierbei in den Hintergrund.

Für Institutionen, die gerade mit Data Mining anfangen, ist ein untersuchender Ansatz üblich. So wie die Erfahrung durch den Umgang mit Data Mining wächst, wird es operationalisiert und zur alltäglichen Norm. Die Automation von Data Mining Vorgänge sowie die Implementierung von Vorhersagemodellen in operativen

Systemen wie Call Center, Marketing Applikationen, Kreditzahlungen und Betrugsaufdeckung bietet den größten Gewinn. In der Hauptsache muß eine große Zahl von Mitarbeitern diese Modelle im Einsatz benutzen können, damit der Mehrwert für das Unternehmen maximiert wird.

Data Mining im Markt

Data Mining Technologie gibt es kommerziell im Markt seit Beginn der 1990er. Obwohl Data Mining Technologie bereits seit einiger Zeit im Markt verfügbar ist, bekam sie erst gegen Ende der 90er Jahre eine entsprechende Priorität in der Anwendung bei großen Unternehmen.

Noch heute werden Investitionen in Data Mining durch andere unternehmerische „Notwendigkeiten“ gehindert - Jahr 2000 Kompatibilität, weitere Infrastruktur wie Data Warehouse und Data Marts, die als Vorbedingungen für breiter angelegte Informations- und Entscheidungssysteme angesehen werden.

Data Warehouse (englisch: Datenlagerhaus) ist eine moderne Form der Datenhaltung und Datenbereitstellung , die eine große Menge Daten in komplexer, verbundener Form für eine große Personenzahl ermöglicht. Solche Datawarehäuser Können sich auf einem oder mehreren Computern erstrecken.

Die Auslagerung von Teilansichten solcher oder auch anderer Datenorganisationen auf einem einzelnen Computer stellen bestimmte Datenmengen einer ausgewählten Gruppe von Anwendern zur Verfügung. Diese ausgelagerten Datenmengen werden Datamart genannt. Analog zum englischen Data Warehouse (Datenlagerhaus) ist ein Datamart (Datensupermarkt) etwas kleiner, aber alles ist für den Endanwender (Verbraucher) leicht zugänglich.

Der Aufbau von einem Data Warehouse bindet meist immense Investitionen von Geld, Zeit und Personal. Die Planung, die Programmierung sowie der schrittweise Aufbau – es werden Datenbankstrukturen gebaut, sorgsam mit Daten gefüllt und nach und nach weiter gefüllt – der Datenhaltung ist oft ein mühsamer Prozess, denn sie wird neben der aktuell laufenden Datenhaltung gebaut. Der alltägliche Betrieb geht weiter. Dies kompliziert die Aufgabe um so mehr. Ferner benötigen Data Warehouses besonders kräftige Computer, um diese große Datenmengen unterbringen und treiben zu können. Leistungsfähigere, größere Computer sind entsprechend teuerer, müssen eingerichtet und in bestehenden Netzwerken eingebunden werden. Weiteres Personal wird zur Pflege benötigt. Zeit und Geld wird zusätzlich verbraucht und zusätzliches Personal gebunden.

Data Warehouse Projekte werden meist von der obersten IT-Management geleitet; da es sich hier in der Regel um eine Millioneninvestition handelt, werden solche Projekte in einigen Fällen sogar direkt unter der Leitung des Vorstandes gestellt.

Ein solches, wichtiges Projekt mit einem großen Etat und unter der höchsten Führung des Unternehmens stellt alle anderen IT-basierten Projekte in den Schatten. Hier wird die Kernverwaltung und Kommunikation des Unternehmens gebaut. Data Mining wird – unserer Meinung nach fälschlicherweise – im Vergleich hierzu als unwichtig angesehen. In der Tat herrscht die Meinung vor, dass alle verfügbaren Kräfte zuerst das Projekt Data Warehouse zu Ende führen sollen und anschließend kann über weitere Schritte nachgedacht werden.

Trotz aller Hindernisse und teilweise Schwierigkeiten im Verständnis von Data Mining befindet sich diese Disziplin bereits vielfältig im Markt. Unsere Erfahrung in bezug auf den aktuellen kommerziellen Einsatz ist wie folgt:

- Data Mining Technologie wird hauptsächlich bei den Unternehmen der "Global 2000" eingesetzt.
- Vielfach liegt der Einsatz im Finanz- und Telekommunikationssektor. Einzelne Unternehmen im Einzelhandel, Gesundheitswesen, und in den Bereichen Pharma, Technologie, Versorgung, Energie, Industrie und bestimmte Behörden setzen auf Data Mining.
- Unter den „Früheinsteigern“ befinden sich Analysten aus dem Heer der Unternehmensberater mit weltweiten oder nationalen Operationen. Diese Gruppe ist relativ klein, jedoch sie fordert Data Mining Lösungen
    - (a) mit hoher Leistung;
    - (b) mit Produktivitätssteigerung; und
    - (c) mit realistischen Preisen.

Data Mining definiert

Data Mining beinhaltet die disziplinierte und kreative Anwendung von Technologie und entsprechendem inhaltlichem Wissen und Verständnis der bearbeiteten Datenmenge. Es ist ein relativ neuer sich noch stark in der Entwicklung befindlichen Vorgang unter Einsatz von technologischen und organisatorischen Mitteln.

Es gibt zahlreiche Definitionen von „Data Mining“, die einzelne Aspekte der Disziplin aufzeigen. Hier sind einige:

„Das Extrahieren von einer bisher unbekannten, gültigen Entscheidungsbasis aus großen Datenbanken und die Anwendung dieser Informationen, um kritische Geschäftsentscheidungen zu fällen.“

IBM

Die obige Definition betont eher die Vorgehensweise des Data Mining; also neigt dazu „technology driven“ zu sein.

Die folgende Definition setzt den Brennpunkt eher bei der Feststellung des Wertes der Daten aus Wettbewerbssicht; also „business driven." Die zumindest implizierte Interdependenz der Zwecke „Wettbewerbsvorteil" und „Geschäftsziele" bietet eine Grundlage, die Data Mining ins reale unternehmerische Umfeld eingliedert.

„Den Wert Ihrer Daten zum Zweck des Wettbewerbsvorteils und der verbesserten operativen Leistung zu entfalten, durch die methodische Suche nach und Aufdeckung von einer Entscheidungsbasis, die identifizierte Geschäftsziele entspricht."

ANGOSS

„So wie ein Minenarbeiter im Bergwerk nach verborgenen Schätzen sucht, so werden beim Data Mining **aus dem 'Datenbergwerk'** *verborgene Informationen ans Tageslicht befördert. Data Mining-Verfahren sollen zu besseren Prognosen, differenzierteren Segmentierungen, Klassifizierungen und Bewertungen von Kundengruppen oder Märkten führen."*

Computerwoche

Data Mining ist „das Verfahren in großen Datenbeständen vorher **unbekannte, außergewöhnliche, unvorhersehbare und bedeutende Informationen** *zu identifizieren bzw. zu extrahieren."*

Meta Group

Data Mining ist „der Entdeckungsprozeß, der es ermöglicht, **mit Hilfe von Mustererkennungstechnologien**, *sowie* **statistischer und mathematischer Verfahren** *aus umfangreichen, abgespeicherten Datenmengen bedeutungsvolle neue Wechselbeziehungen, Muster und Trends* **herauszufiltern."**

Gartner Group

Fazit

Kurzum wird Data Mining in der Praxis dazu verwendet, bestehenden Datenbeständen einen Mehrwert zu entlocken. Dieser Mehrwert wiederum bildet einen Vorteil im strategischen oder alltäglichen Wettbewerb.

Dieser Wettbewerb kann

- gegen andere Konkurrenten im Markt (Marktwettbewerb),
- gegen die Zeit (Vorsprung),
- gegen den Mißbrauch (Aufdeckung) oder
- gegen das Unwissen (Transparenz) geführt sein.

Diverse Kombinationen sind vorstellbar. Ein typisches Beispiel des **Marktwettbewerbes** sind die Marketing Bestrebungen, das Kundenverhalten zu verstehen, um früher, näher und besser am Point-of-Sale zu sein.

Als Beispiel für den **Vorsprung** könnte man die Arbeit der Kriminalpolizei oder der Mediziner ansehen, die gegen die Uhr versuchen, durch Prognosen / Diagnosen zu einer sonst verborgenen Entscheidungsgrundlage zu gelangen.

Aufdeckung: Aufgedeckt werden können zum Beispiel Betrugsfälle bei einer Autoversicherung oder die Kreditprüfung bei einer Bank vorgenommen werden.

Das **Unwissen** ist oft die erste Hürde einer Untersuchung. Data Mining erzeugt eine durch die Darstellung der entscheidenden Zusammenhänge erklärende Transparenz, wodurch das Unwissen fundiert beseitigt werden kann.

Sogar bei dieser kurzen Beschreibung entstehen Überlappungen, jedoch sollte uns hier in der Hauptsache der jeweilige Fokus beschäftigen. Spätestens an dieser Stelle beginnt bei dem – vor allem neuen - Data Miner ein Konflikt zwischen dem absoluten Verständnis für eine Thematik und dem durch Data Mining aufgezeigten tendenziellen Ansatz aufzutreten.

Zusammenhänge erscheinen, die man auf Anhieb entweder nicht versteht oder spontan für unmöglich erklärt. Werte werden bis in die minimalsten Unterschiede auf die Goldwaage gelegt. Dabei wird der klare Blick von den wesentlichen Zusammenhängen und der Mustererkennung abgelenkt.

Bevor wir uns weiter der praktischen Projektarbeit nähern können, müssen wir diese kleine aber wesentliche Hürde nehmen.

4. Der Drang nach absoluten Werten

In unserem Alltag sind wir es gewohnt, mit absoluten präzisen Werten zu arbeiten. Dieser Funktionalitätstrieb erfordert wiederum eine wachsende Genauigkeit und Präzision. Schlicht in der heutigen, Technologie Gesellschaft geben wir uns mit unpräzisen oder unscharfen Antworten kaum zufrieden.

Zum Beispiel: Auf die Frage was etwas kostet, erwartet man in der Regel eine exakte Summe zu hören. Man gibt sich mit der Antwort: „Mehr als früher“ ungern zufrieden. Eine solche Antwort erfüllt nicht unsere Erwartungen.

Wir kennen alle die Frage, „Was ergibt 2 x 3“ und können alle rasch die präzise Antwort 6 liefern. Damit geben wir uns sofort zufrieden. Die Antwort: „Ungefähr das Doppelte von 3“ fällt aus dem erwarteten Rahmen und wir geben uns mit einer solchen Antwort nicht zufrieden.

Im Data Mining haben wir es zuerst hauptsächlich mit ungenauen Ergebnissen zu tun, die unsere Erwartungen auch nicht unbedingt erfüllen. Die ersten Ergebnisse einer Data Mining Untersuchung können sogar überraschenden Inhalts sein; dann müssen wir uns zuerst mit diesen unerwarteten Inhalten auseinandersetzen und ein Verständnis dafür suchen. Die Überraschung muß verarbeitet werden, bevor ein neues Verständnis für die neuen Inhalte gefunden werden kann.

Beim Data Mining müssen wir lernen, mit tendenziellen Aussagen zu leben, die unserem Streben nach Genauigkeit nicht – oder aus dem Projektablauf heraus genauer gesagt noch nicht – erfüllen können. Der geübte und innovative Analytiker oder der Marktforscher hat es hier meist einfacher, dieses Hindernis locker zu nehmen, da der Umgang mit Trends und Tendenzen zu seinem Alltag gehören. Es ist im Bereich Marketing nicht ungewöhnlich, den Werbeetat als Prozentsatz vom Umsatz auszudrücken.

Als zusätzliche Schwierigkeitselement wissen wir, dass solche neuen Erkenntnisse meist nicht einzeln, sondern im Konzert, parallel und mehrschichtig erscheinen und somit die intellektuelle Erfassung der neuen Erkenntnisse erschweren.

Hinzukommen weitere Erschwernisse wie Zweifel, Verschlossenheit und Vorurteile. Etwas Mut oder Freude, Wagnisse einzugehen, gehört auch dazu. Denn die Beherrschung der Zweifel, die Offenheit und die Freiheit sich von Vorurteilen zu lösen sowie der Mut neue Wege zu gehen, ermöglichen erst den Aufbau einer neuen Erkenntnis.

Erst nachdem man eine Erkenntnis auf- und ausgebaut hat, erreicht man allmählich eine Beherrschung von Data Mining Aussagen aus den eigenen Datenbeständen. (Zum Thema Beherrschungsgrad von Wissen bei einem Data Mining Projekt sehen

Sie Kapitel 8 „Stand des Wissens.“) Stufenweise beginnt man „spielerisch“ mit den eigenen Daten umzugehen und entwickelt eine Offenheit für neue Erkenntnisse.

Beschreibende Erkenntnisse aus einem Data Mining Prozess gekoppelt mit dem vorhandenen, erfahrenen Verständnis für die Inhalte der eigenen Datenmengen ermöglichen erst eine Interpretation der Data Mining Resultate. Diese überlegte Interpretation gibt einen Hinweis auf den Sinn oder Unsinn der Ergebnisse.

Im Zusammenhang mit der „fachlichen“ Interpretation der Data Mining Ergebnisse führt man im Data Mining eine weitere Prüfung durch, die den nutzbaren Wert der Ergebnisse untersucht. Erst nach dieser sogenannten Verifizierung beginnt man, endlich von Werten zu sprechen, die man als „absolut“ betrachten kann.

Um den allgemeinen Ablauf eines Data Mining Projektes besser nachvollziehen und die Entwicklung der Erkenntnisse anhand eines Beispiels verstehen zu können, verfolgen wir das Beispiel einer Autoversicherung, die versucht, die Daten ihrer Kfz-Schäden zu untersuchen.

5. Ein kleines Projektbeispiel

Die vorliegenden Daten aus dem Bereich der Kfz-Schäden zeigen deutlich auf, dass es einen signifikanten Zusammenhang zwischen dem gesellschaftlichen Status eines Versicherten und seinem Verhalten im Schadensfall gibt.

Ferner sehen wir die gleiche Datengrundlage aus der Sicht des Betruges und stellen fest, dass der Aufbau eines Datenmodells, als Basis der sofortigen Überprüfung einer Betrugswahrscheinlichkeit im Falle der Schadensmeldung möglich und sinnvoll ist. Die Erstellung einer einfach zu bedienenden Anwender Applikation für die Sachbearbeiter des Schadenbereichs kann vorgenommen werden.

In dem obigen Szenario ist die sofortige, schnelle und einfache Überprüfung jeder Schadensmeldung auf Betrugswahrscheinlichkeit bereits bei der Aufnahme der Schadensmeldung durch die Sachbearbeiter gegeben.

Die potentiellen Ersparnisse gehen in die Millionen.

Die folgenden Seiten zeigen den Weg zu diesen Erkenntnissen auf und bilden ein Beispiel für den ersten Schritt des **Data-Mining-Challenge**, wo Ansätze aufgezeigt werden können.

Ein Data-Mining-Challenge wurde entwickelt, um konkret aufzuzeigen, welche potentiellen Ergebnisse durch ein vollständiges Data Mining Projekt erzielt werden können. Ein Data-Mining-Challenge ist in der Zeit (wenige Wochen) überschaubar als auch von der Investition (wenige Tausend Mark) leicht zu entscheiden. Ein Data-Mining-Challenge liefert konkrete, hochrechenbare Ergebnisse, die für die Entscheider und für die betroffenen Teilnehmer aus einem gewohnten Umfeld stammen. Dadurch wird der Einsatz von Data Mining Techniken für interessierte Anwender berechenbar und die Freigabe von größeren Summen für vollständige Projekte vertretbar und für die Entscheider verständlich.

Vorgabe und Zielsetzung

Die Findung von sinnvollen Vorgaben und Zielsetzungen fußt auf unterschiedlichen Grundkriterien. Eine Managementvorgabe könnte sein:-

„Steigt die Betrugswahrscheinlichkeit bei der Kfz-Schadensabwicklung an und wenn ja, bitte um Vorschläge zur Bekämpfung?“

Wenn eine entsprechend hohe Betrugswahrscheinlichkeit bei der Kfz-Schadensabwicklung festgestellt wird, wie in der Frage des Managements vermutet

wird, könnte diese Vorgabe in ein Data Mining Projekt mit der folgenden Zielsetzung münden:-

Betrugswahrscheinlichkeit bei der Kfz-Schadensabwicklung durch Ermittlung der Einflußgrößen analysieren sowie ein Vorhersagemodell aufbauen. Eine ständige, schnelle und leicht zu bedienende Prüfung soll bereits direkt bei der Schadensmeldung durch ein Prognostik-Werkzeug erfolgen.

Der Sinn dieser Zielsetzung nimmt die Managementvorgabe auf und wandelt sie in eine Untersuchung von bleibendem Wert um, die in der Lage ist, die Schäden - zumindest um betrugswahrscheinliche Fälle - zu mindern. Über die Summe der betrugsverdächtigen Schadensmeldungen wird in der Tat meist geschwiegen. Erfahrungsgemäß werden jedoch bei solchen Data Mining Projekten bereits im ersten Jahr 10 - 20% der betrugsverdächtigen Schadensfälle sofort bei der Schadensmeldung identifiziert und daher eingespart.

Beim Data-Mining-Challenge beträgt die Dauer der Untersuchung 2 -3 Wochen. Die darauffolgende eingehende Untersuchung wird mit 2 - 3 Monate veranschlagt. Hiernach sollte die Grundlage für das Prognostik-Werkzeug gelegt sein. Das Prognostik-Werkzeug sollte innerhalb von weiteren 2-3 Monaten zumindest in Testbetrieb genommen werden können. Wie schnell sich ein solches Projekt rentiert, kann jeder anhand des Berichtes vom Data-Mining-Challenge für sein eigenes Unternehmen errechnen. Zusätzlich zur direkten Hochrechnung der Rentabilität eines Data Mining Projektes kommen weitere wertvolle Nebeneffekte.

Nebeneffekte sind hier die Zusammenhänge, die während der Untersuchung aufgedeckt werden und die möglichen Analyseansätze in anderen Richtungen aufzeigen. Entsprechende aufgezeigte Themen können in diesem Beispiel sein:-

Vertriebs-Controlling, Churn-Management, Wettbewerbskontrolle, Kalkulation sowie Zielgruppenermittlung, Portfoliobereinigung uvm. sogar Prämienermittlung.

Vorgehensweise und Ergebnisse

Die Beschaffung und Evaluierung der Daten stellt die schwierigste Aufgabe dar. Für diese Beispielanalyse haben wir eine Datenmenge aus den USA erhalten, die entsprechend neutralisiert (von spezifischen, persönlichen Angaben bereinigt) ist.

Die Satzstruktur sieht wie folgt aus:

Monat
Monatswoche
Wochentag
Hersteller
Unfallgebiet

Wochentag/Meldung
Monat/Meldung
Monatswoche/Meldung
Geschlecht
Status
Alter
Schuld
Police
PKW-Kategorie
PKW-Preis
Betrug
Policenummer
Vertreternummer
Freiheitsrabatt
Fahrerklasse
Tage: Police-Unfall
Tage: Police-Meldung
bisherige Meldungen
PKW-Alter
Versichertenalter
Polizeibericht
Unfallzeugen
Agenturart
Anzahl Zusätze
Adressenänderung/Meldung
Anzahl Autos
Jahr
Grundpolice

Das Feld „Alter“ wurde wegen Doppelung mit „Versichertenalter,“ sowie das Feld „Jahr“ aus dem Grund des Überfluß - alle Daten waren eh aus dem gleichen Jahr - ignoriert.

Die Datenmenge bestand aus 15.450 Sätzen.

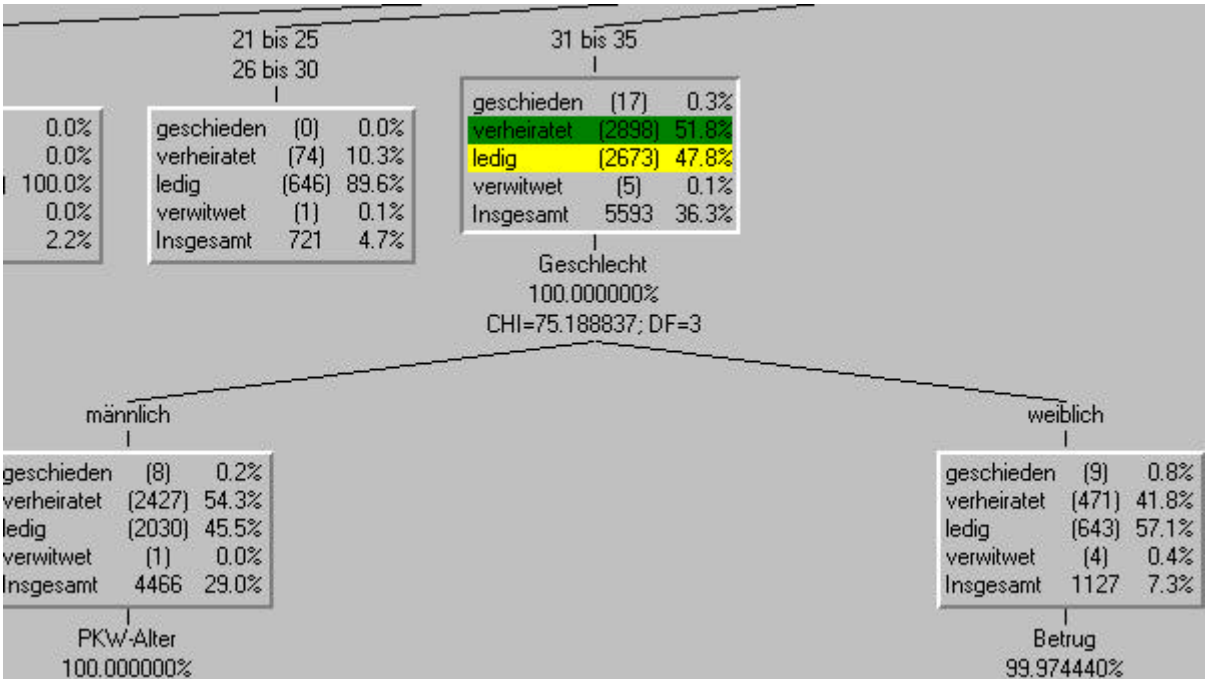
Zum Auftakt der Analyse wurde das Feld „Status“ als Ausgangspunkt der Untersuchung als „abhängige Variable“ gewählt. Der Grund hierfür lag darin, dass Gespräche mit Schadenssachbearbeitern bzw. Schadensreglern bei Kfz-Versicherungen, den Familienstatus als wichtiger Faktor bei Betrugsfällen hervorbrachten. Diese Annahme wollten wir als wichtiger Faktor bestätigen oder als wertlose Vermutung darstellen.
Irgendeinen Anfang muß gemacht werden, und wie so oft, hat uns das Feld „Status“ als anfänglicher Ausblickwinkel zwar noch nicht die gewünschten, alles erklärende Erkenntnisse beschert, aber einen sinnvollen und nützlichen Einblick gewährt. Der Einblick war in sofern nützlich, als wir dadurch weitere, sinnvolle und wertvollere

Blickwinkel gefunden haben, die uns eher zu einer vollständigen Erklärung und damit zur Erfüllung der Aufgabenzielsetzung führten.

Hiernach sind wir im Stil des Knowledge Discovery vorgegangen. Bei dieser Arbeitsweise sucht unser Werkzeug – in diesem Falle KnowledgeSEEKER, ein führendes Werkzeug der Entscheidungsbaum Technik – eigenständig nach statistisch bedeutungsvollen Zusammenhängen. Das Werkzeug erhält die Information welches Datenfeld als Ausgangspunkt (abhängige Variable) dienen soll und welche Datenfelder zur Untersuchung gehören sollen. Danach wird dem Werkzeug den gewünschte Algorithmus mit entsprechenden Varianzen zugeteilt. Wir haben uns mit den Standardeinstellungen des Werkzeuges zufrieden gegeben. Experten werden selbstbewußter vorgehen und weitere der gebotenen Möglichkeiten ausprobieren und im Ergebnis vergleichen.

Infolge der ersten Ergebnisse und Interpretationen haben wir teilweise Hypothesen gebildet und sie verfolgt. Als erste signifikante Ebene entdeckten wir das Alter der Versicherten. Es ergab sich eine Ballung in den Altersgruppen 31-35 Jahre sowie 36-40 Jahre.

Die Altersgruppe über 40 Jahre haben wir fortan aus der Untersuchung herausgelassen, weil hier keine eindeutigen Erkenntnisse auftauchten. Folglich haben wir uns ausschließlich auf die Altersgruppen 31-35 und 36-40 Jahren konzentriert.

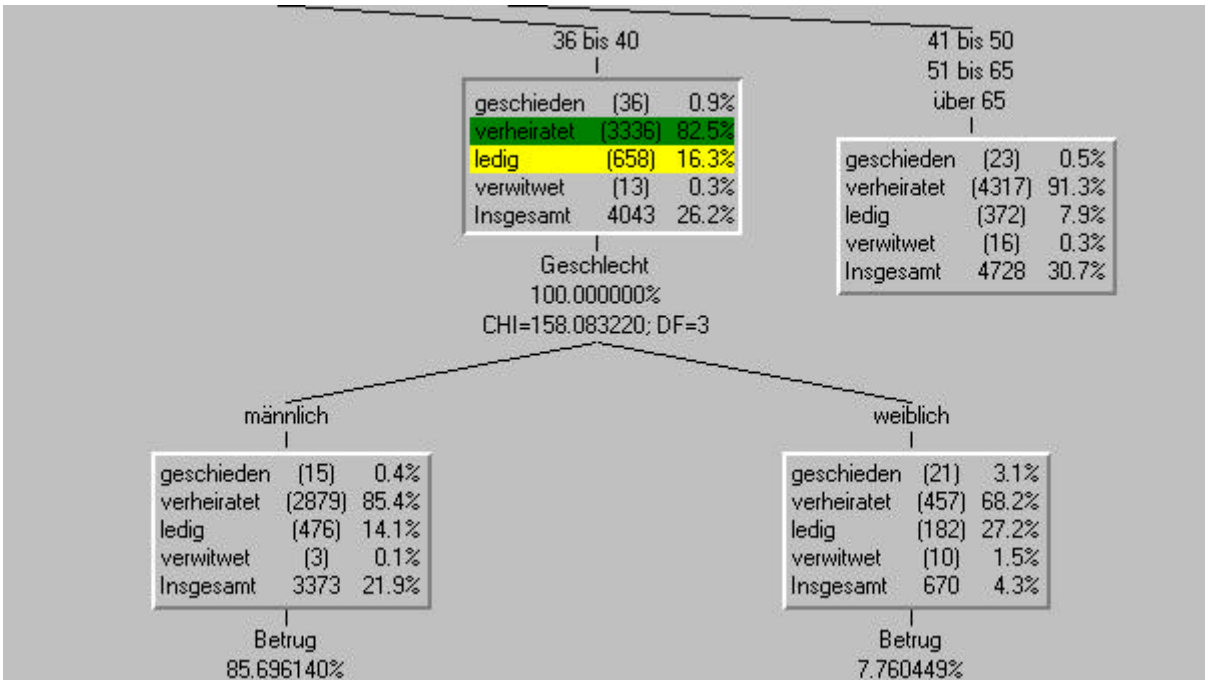


Die Grafik des Entscheidungsbaumes zeigt die Altersgruppe 31-35 Jahre, die nominell und prozentual in den vier „Status“ Kategorien aufgebaut sind. Die von ihrem

Anteil klar deutlichen Kategorien „verheiratet“ und „ledig“ sind zur Verdeutlichung an einer Stelle mit einem grünen bzw. gelben Balken hinterlegt.

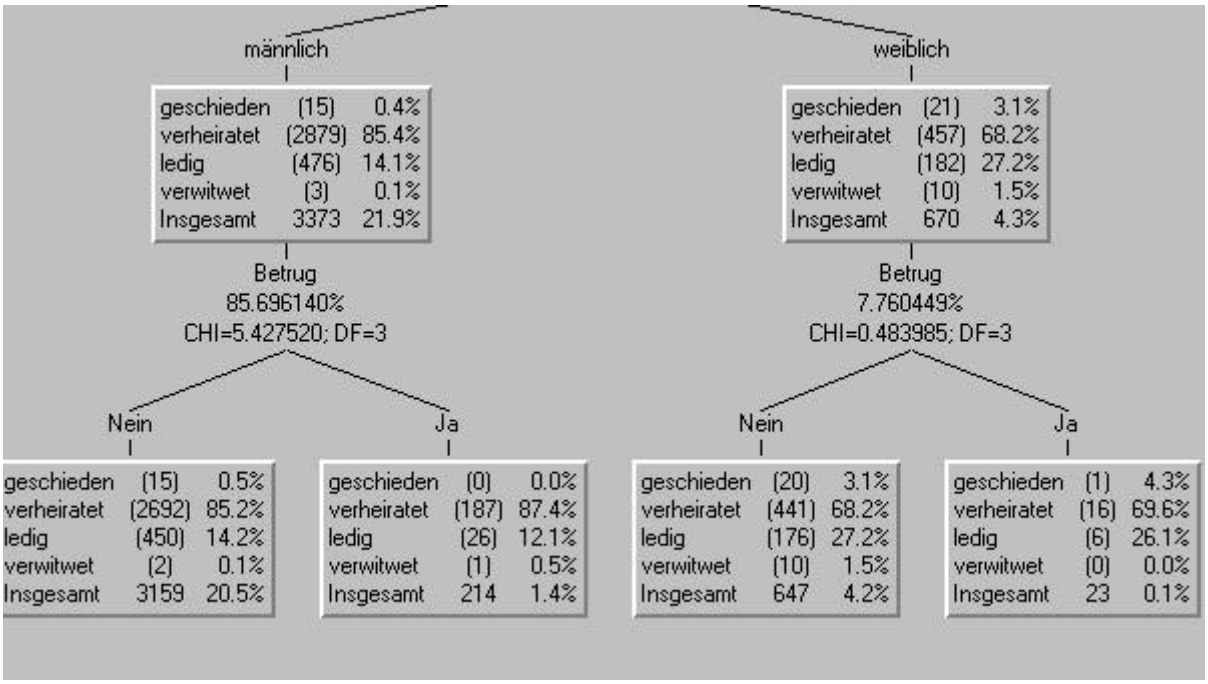
Die Altersgruppe 31-35 Jahre bildet mit 36,3% rund ein Drittel der Gesamtmenge. Die Gruppe wird anschließend nach Geschlecht geteilt. Hieraus sieht man eine leichte Mehrheit von verheirateten Männern aus ihrer Gruppe gegenüber leicht mehr ledige Frauen in ihrer Gruppe. Das Verhältnis der Geschlechter zueinander in dieser Altersgruppe ist etwa. 4 : 1 Männer.

Im Falle der Altersguppe 36-40 Jahre entsteht ein anderes Bild. Der Anteil an der Gesamtmenge ist etwas kleiner (26,2%) jedoch ist der Anteil Frauen deutlich geringer – das Verhältnis beträgt hier 5:1 Männer. Der Anteil verheiratete Frauen ist mit 68,2% ebenfalls deutlich höher.



Die Männer sind zu 85% verheiratet, was zum größten Teil demoskopisch zu erklären ist. Die Abweichung zwischen den Werten der verheirateten Männern und den verheirateten Frauen erklärt sich dadurch, daß Autoversicherungen oft im Namen des Mannes abgeschlossen werden und die Frau als zusätzliche versicherte Person gilt. Daher erscheint die Ehefrau nicht direkt als Versicherungsnehmerin.

Im allgemeinen wird in der gesamten Altersgruppe 31-40 am unteren Bildrand der beiden bisherigen Grafiken „Betrug“ als Folgekriterium erkennbar, was in der folgenden Grafik verdeutlicht wird.

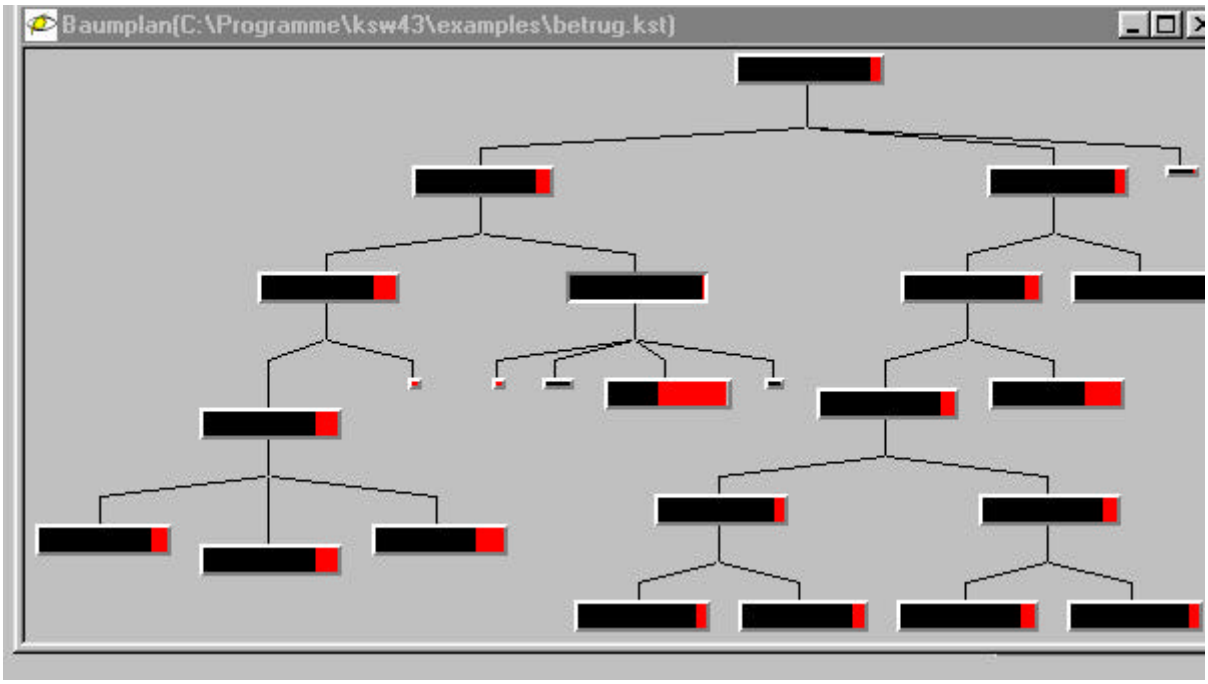


Obwohl erkennbar, finden wir an dieser Stelle noch keinen signifikanten Aufschluß darüber, welche Einflüsse beim Betrug eine Rolle spielen. Wir wissen nun, dass der Betrug bereits als viertes Kriterium nach Status, (in diesem Bild) Altersgruppe 36-40 Jahre, Geschlecht aufgetaucht ist.

In der jüngeren Altersgruppe 31-35 erleben wir (nicht im Bild), dass bei den Frauen eher die „Verheirateten“ zum Betrug neigen, aber bei den Männern kommen erst andere Eigenschaften hervor und der Betrug taucht vorläufig gar nicht auf.

Anhand diese Ansätze entscheiden wir uns, die „abhängige Variable“ zu wechseln und damit den Betrachtungswinkel unserer Analyse vom Feld „Status“ weg und hin zum „Betrug“ zu verlagern. Diese Fälle sind eindeutig als historische Schadenfälle erkennbar.

Die folgende Grafik zeigt einen schematischen Zusammenhang von zwei möglichen Verkettungen an, die einen Betrug erkennen lassen. Dieses Schema möchten wir etwas auflösen.



In dieser Grafik sehen wir die Übersicht eines Entscheidungsbaumes. Die abhängige Variable ist „Betrug“ und den Kategorien „Ja“ und „Nein“ sind die Farben rot bzw. schwarz zugeordnet worden. Dort wo die Farbe rot zunimmt, wächst der Betrugsanteil.

Die linke Verästelung stellt den Betrug aus der Sicht der Grundpolice = Vollkasko dar. Von den insgesamt 923 Betrugsfällen werden 452 also knapp die Hälfte in diesem Bereich „Grundpolice = Vollkasko“ gefunden. 435 Betrugsfälle also etwa die andere Hälfte werden in der rechten Verästelung „Grundpolice = Unfall“ gefunden.

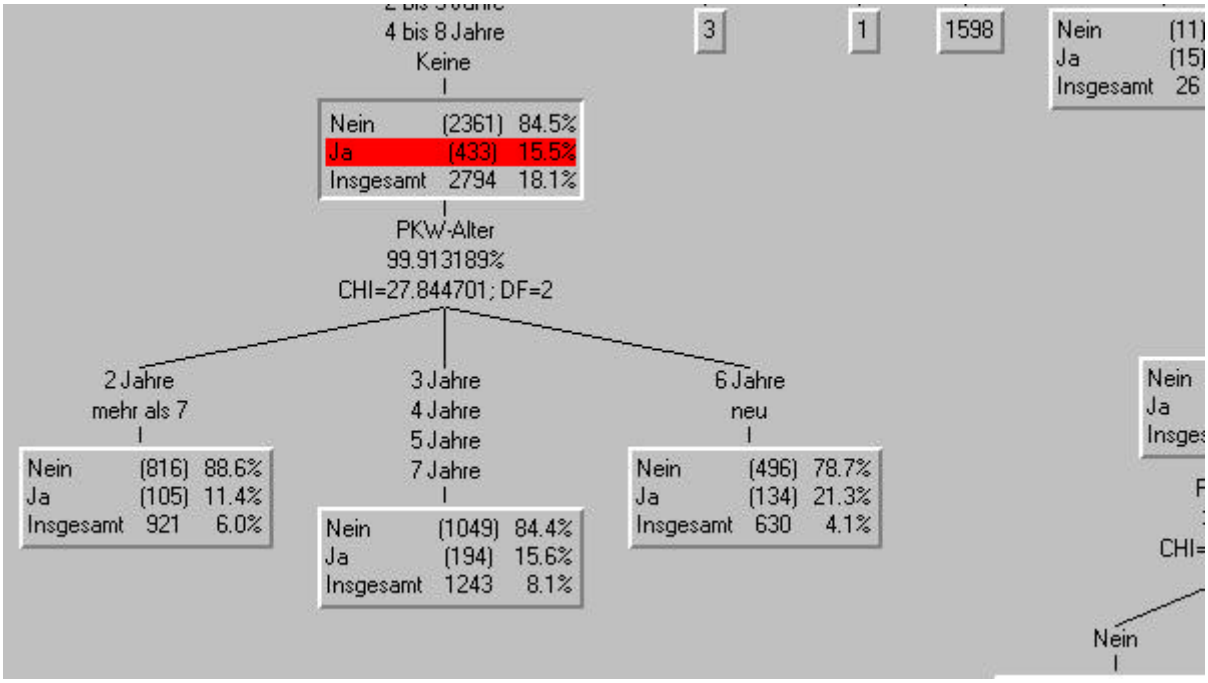
Erstes Fazit:

Im Betrugsfall ist die Grundpolice fast ausschließlich Vollkasko oder Unfall.

Die Verzweigung Vollkasko löst sich weiter auf, in dem der Unfallgegner als wichtiger Indikator ausscheidet. Also ist es hauptsächlich der Versicherter selber, der Schuld am Unfall hat. Hiernach spielt die Dauer der letzten Adressenänderung bis zur Meldung des Schadens eine signifikante Rolle. Alle Variationen außer „unter 6 Monate“ spielen eine Rolle. An dieser Stelle sind noch 433 der ursprünglichen 452 Fälle in diesem Sektor betroffen, die sich in der folgenden Grafik je nach Alter des Fahrzeuges weiter aufsplitten.

Wir sehen die „Betrugsrate“ (roter Balken) bereits bei 15,5% aber im Falle der Neuwagen oder 6 Jahre alten Autos, die vom Werkzeug statistisch gruppiert wurden, einen Wert von 21,3%, obwohl die höchste Anzahl (194 Schäden) der vom Betrug betroffenen Fälle im PKW-Altersbereich der 3, 4, 5 und 7 Jahre alten Fahrzeuge liegt.

Wir schließen daraus, dass die ältesten Autos (über 7 Jahre alt) am wenigsten von Betrügern genutzt werden. Dies ergibt Sinn, denn dort sind die zu erzielenden finanziellen Werte am geringsten. Eine nähere Untersuchung der Daten ergab, dass die 2 Jahre alten Autos kaum vorkamen.



Umgekehrt war die höchsten Raten bei den Neuwagen entdeckt, wobei hier die 6 Jahre alten Autos in der Tat kaum repräsentiert waren. Soweit war dies alles leicht erklärbar. Aber der mittlere Block der Grafik stellt die zahlenmäßig größte Gruppe dar. Hierfür liegt keine Erklärung so offen auf der Hand.

An dieser Stelle ist die Gruppierung der Fahrzeuge nach ihrem Alter zwar interessant, leider gibt die Verteilung der Betrugsfälle aber hier keinen Aufschluß. Es müßte eine weitergehende Analyse der Datenmenge aus der Sicht des PKW-Alters bzw. eine Bereinigung der Datenmenge um einige „ausgeschlossene“ Kriterien erfolgen, um hier Klarheit zu schaffen.

Wir wenden uns nun der anderen Verzweigung „Grundpolice = Unfall“ zu.

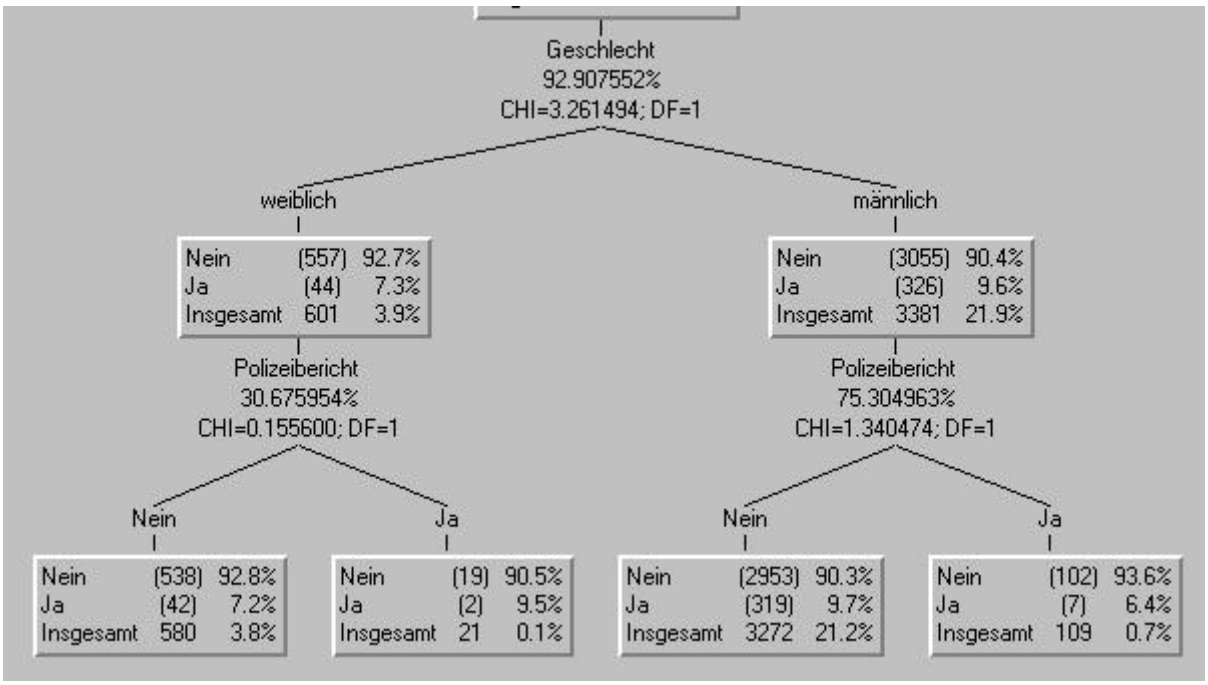
Von unseren 435 Betrugsfällen werden 20 bei der nächsten Ebene - Schuld - eliminiert. In lediglich 20 dieser Fälle hat der Unfallgegner Schuld – Ähnlichkeit mit der Verästelung „Grundpolice = Vollkasko.“ Also diese Datenmenge scheint zu sagen, dass wenn es Betrugsverdacht gibt, betrifft es den eigenen Versicherten und nicht die Unfallgegner. Auch dies ist logisch und sinnvoll, denn es ist wohl kaum möglich eine fremde Versicherung um Geld zu prellen. Dazu muß man erst dort versichert sein, um am Spiel des Betruges teilnehmen zu können.

In der nächsten Ebene „PKW-Kategorie“ scheiden weitere 45 Fälle aus, die zur Kategorie Sportwagen gehören. Es bleiben 370 in Betracht.

An dieser Stelle schmilzt einer unserer Vorurteile dahin. Wir hatten geglaubt, Sportwagen stellen die Hauptquelle von Kfz-Schadensbetrügerei dar. Hier mussten wir einsehen, dass Glauben nicht Wissen ist. Also bleiben uns die „Limousine und Andere“ als PKW-Kategorie, die wir nach Geschlecht aufteilten.

Die folgende Grafik zeigt die Aufteilung nach Geschlecht sowie die Folgesplittung nach Polizeibericht. An dieser Stelle haben wir eine erste deutliche Teil-Erklärung für die Grundlagen des Betrugsfalles.

Immerhin betrifft diese Erklärung einen beträchtlichen Teil unserer gesamten Datenmenge, nämlich rund 25%. Hierbei können wir schon behaupten, dass die hier noch präzise herauszustellenden aber gefundenen Bedingungen im Bereich Grundpolice = „Unfall“ in mindestens 25% aller Betrugsfälle vorkommen.



Insbesondere wird die Bedeutung des Polizeiberichts in der Grafik ersichtlich. Man kann fast behaupten, ob männlich oder weiblich, wenn kein Polizeibericht vorhanden, ist Betrug ein Dauergast.

Diese gefundene Bedingung möchten wir an einem Ausschnitt verdeutlichen.

Polizeibericht , Nein
Geschlecht, männlich
PKW-Kategorie, Limousine Andere

Schuld, Versicherter
Grundpolice, Unfall

Wenn die obigen Bedingungen zutreffen, werden bereits 34.56% aller Betrugsfälle eine Erklärung finden. 923 Betrugsfälle gibt es insgesamt. Wir haben mit der obigen Verkettung 319 Fälle erklärt. 319 / 923 ergibt 34.56%.

Fazit: Etwa jeder dritter Betrugsfall kann mit der obigen Kette von Bedingungen aufgedeckt werden.

Dieses einfache Datenmodell wird qualitativ deutlich erhöht, wenn wir die Vorbedingung Unfallversicherung festlegen. Von 435 Fällen haben wir 319 erklärt. 319 / 435 ergibt 73.34% oder 3 von 4 Betrugsfällen im Bereich Grundpolice = Unfall sind mit der obigen Bedingungskette gefunden.

Neue, weitere Daten sollten selektiert werden, um den hier entstandenen Regelsatz zu prüfen. Der hier benutzte Ausschnitt der Datenmenge beinhaltet die Datensätze, die genau unserer Regel entsprechen.

Der folgende Regelsatz betrifft die rechte Verästelung „Grundpolice = Unfall“ und würde, wenn man ihn als Abfrage an eine ähnlich strukturierte Datenmenge passend formulieren und anwenden würde, alle entsprechende Datensätze aus der Sicht „Betrug = Ja/Nein“ selektieren.

Regelsatz:

RULE_1 IF
Grundpolice = Unfall
THEN
Betrug = Nein 92.7%
Betrug = Ja 7.3%

RULE_2 IF
Schuld = Versicherter
Grundpolice = Unfall
THEN
Betrug = Nein 90.0%
Betrug = Ja 10.0%

RULE_3 IF
PKW-Kategorie = Limousine or Andere
Schuld = Versicherter
Grundpolice = Unfall
THEN
Betrug = Nein 90.7%
Betrug = Ja 9.3%

RULE_4 IF
Geschlecht = weiblich
PKW-Kategorie = Limousine or Andere
Schuld = Versicherter
Grundpolice = Unfall
THEN
Betrug = Nein 92.7%
Betrug = Ja 7.3%

RULE_5 IF
Polizeibericht = Nein
Geschlecht = weiblich
PKW-Kategorie = Limousine or Andere
Schuld = Versicherter
Grundpolice = Unfall
THEN
Betrug = Nein 92.8%
Betrug = Ja 7.2%

RULE_6 IF
Polizeibericht = Ja
Geschlecht = weiblich
PKW-Kategorie = Limousine or Andere
Schuld = Versicherter
Grundpolice = Unfall
THEN
Betrug = Nein 90.5%
Betrug = Ja 9.5%

RULE_7 IF
Geschlecht = männlich
PKW-Kategorie = Limousine or Andere
Schuld = Versicherter
Grundpolice = Unfall
THEN
Betrug = Nein 90.4%
Betrug = Ja 9.6%

RULE_8 IF
Polizeibericht = Nein
Geschlecht = männlich
PKW-Kategorie = Limousine or Andere

Schuld = Versicherter
Grundpolice = Unfall
THEN
Betrug = Nein 90.3%
Betrug = Ja 9.7%

RULE_9 IF
Polizeibericht = Ja
Geschlecht = männlich
PKW-Kategorie = Limousine or Andere
Schuld = Versicherter
Grundpolice = Unfall
THEN
Betrug = Nein 93.6%
Betrug = Ja 6.4%

Natürlich suchen wir den „endgültigen“ Regelsatz, der immer ein „gutes“ Ergebnis liefert. Hier soll uns dieser Regelsatz lediglich als Beispiel für einen gefundenen Ansatz zur Weiterverarbeitung in folgenden Analysen dienen. Die weitere Verfeinerung und Prüfung solcher Regelsätze erfolgt in einem weiteren Schritt. Unsere Aufgabe ist hier Ansätze aufzuzeigen.

Ferner haben wir hiermit ein Beispiel eines gefundenen Regelsatzes, der uns die berühmte Nadel im Heuhaufen zeigt. Vorher hatten wir gar nicht gewußt, welche Frage wir an unsere Datenmenge hätten stellen sollen. Jetzt wissen wir die Antworten auf Fragen, die wir gar nicht formulieren mussten. Nachträglich können wir die Fragen oder Regelsätze leicht in verschiedener Form (Generisch wie oben, SQL, SAS, Java usw.) ausgeben, um direkt mit unserer Datenmenge „sprechen“ zu können.

Zusammenfassung

Die Suche nach den Zusammenhängen im Betrugsfall bei einer Kfz-Schadensmeldung hat eine Verkettung von Bedingungen und Folgebedingungen aufgezeigt, die in der Signifikanz dargestellt werden.

Hiermit haben wir die Grundlage eines Datenmodells gefunden, das zur Ermittlung der Betrugswahrscheinlichkeit dienen kann. Jede einzelne, in der Tat aktuell anfallende Schadensmeldung kann nach ihren Kriterien erfaßt und „über unser Datenmodell gefahren“ werden. Die Wahrscheinlichkeit des Betruges kann dann in Sekunden prognostiziert werden.

Das ganze wird in einer Anwender Applikation verankert. Im Hintergrund liegt das Datenmodell, das jederzeit verfeinert und aktualisiert werden kann. Dieses Datenmodell wird von Fachleuten kontrolliert und als „Schablone“ in der Applikation

verschlossen, die von Sachbearbeitern der Schadensabteilung bedient werden kann. Innerhalb von Sekunden kann die Betrugswahrscheinlichkeit jeder Schadensmeldung von einem Sachbearbeiter der Schadensabteilung – nicht von Spezialisten der EDV oder der Statistik – als Prognose ermittelt werden.

Der Schadenssachbearbeiter wird je nach Ergebnis der Prognose verfahren und die Gesellschaft wird einen beträchtlichen Teil der bisherigen Schadenssumme nicht mehr auszahlen; denn eine hohe Betrugswahrscheinlichkeit wird andere Akteure einschalten, die solche Fälle behandeln. Im Endeffekt werden die meisten solcher „verdächtigen“ Schadensmeldungen zurückgezogen, aber auf jeden Fall wird die Auszahlung bis auf weiteres zurückbehalten und die Verluste durch Betrug minimiert.

Vorschlag

Data-Mining-Challenge ist eine Aufforderung Ihre Daten in dieser Hinsicht aufzubereiten. In einer Phase von 2 - 3 Wochen können Sie einen solchen Bericht über Ihre Situation erhalten.

Die Kosten von mindestens DM 10.000,- netto (Sonderwünsche wie Einsatzort, Einsatzdauer usw. können die Maßnahme verteuern) geben Ihnen einen praktischen Aufschluß über mögliche Gewinne solcher Projekte in Ihrem Unternehmen.

Unser Beispiel zeigt auf, welche Gewinne zu erzielen sind, wenn ein meßbarer Input erbracht wird. Sie messen in Ertrag oder neudeutsch „Return-on-Investment“ und müssen sich nicht mit der akademisch-wissenschaftlichen Belastung des virtuellen Faches Data Mining auseinandersetzen, um seine Wirkung zu messen und verstehen zu können.

Weiterführende Analysen auf der Basis zusätzlicher „lernfähiger“ Techniken kommen in einer fortführenden Phase zum Einsatz.

Nun möchten wir sehen, welche Strukturen wir in der Arbeit eines Data Mining Projektes zur leichteren Orientierung einbauen können.

6. Das Data Mining Projekt-Rad

Ein Data Mining Projekt besteht aus deutlich verschiedenen Schritten, die unterschiedliche Kenner auf ihrer Weise kategorisiert haben. Wir möchten uns bei keiner Bewertung dieser diversen Vorgehensweisen aufhalten, sondern lediglich aussagen, dass der folgende Phasenaufbau eines Data Mining Projektablaufes aus vielen Methoden sowie Projekten gewachsen ist und sich in der Praxis – insbesondere bei der Übertragung auf neue Data Miner – als leicht verständliches und in den unterschiedlichen Umgebungen anwendbares Mittel zum Zweck der klaren und geordneten Data Mining Projekten erwiesen hat.

Die verschiedenen Schritte eines Data Mining Projektes sind untereinander verbunden. Die Wege sind keine Einbahnstraßen, sondern führen ineinander, weiter, sowie auch teilweise zueinander zurück: Denn dieses Projekt-Rad ist nicht statisch, sondern befindet sich meist in Fluß. Ferner ist es mehrdimensional zu verstehen. Jeder einzelne Schritt des Projektes liefert mögliche Erkenntnisse oder Teileinblicke ins zentrale digitale Nervensystem einer jeden Institution.

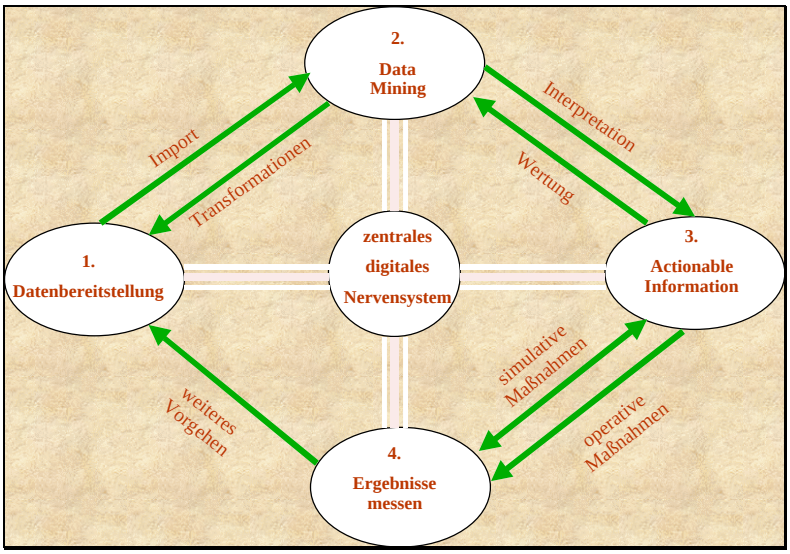
Anhand des folgenden Diagramms werden wir die verschiedenen Schritte eines Projektes sehen und einige Rückkoppelungen festmachen, die dazu verwendet werden, notwendige Zwischenschritte zu bewältigen, die wiederum zur Steigerung der Qualität der Data Mining führen.

Bevor wir loslegen können, müssen einige datenbezogene Vorbedingungen geklärt werden.

Vorab muß man wissen, dass Data Mining im Vergleich zu eher herkömmlichen Methoden absolut saubere Daten verlangt. Data Mining wird nach dem Prinzip „Mist ´rein, Mist ´raus“ keine Qualität in den Ergebnissen liefern, wenn Sie nicht für beste Eingabequalität gesorgt haben. Wie dies erreicht wird und welche Schritte dabei zu beachten sind, sehen wir gleich in der Zusammenfassung im Abschnitt unten „i) Datenbereitstellung,“ die wir als nächste wichtigste Aufgabe verfolgen.

Im folgenden Diagramm sehen wir unten im Überblick eine schematische Darstellung des Data Mining Projektrades, die vier Hauptstufen ähnlich den primären Windrichtungen des Kompasses aufzeigt. Die Vor- und Rückkoppelungen sind mit grünen Pfeilen angezeigt mit allgegenwärtigen breiten Verbindungen zum zentralen digitalen Nervensystem der Institution.

Dieses Schema wird uns immer wieder begegnen und teilweise gute Dienste als Orientierungshilfe ähnlich einer Landkarte leisten.



Das Rad des Data Mining Projektes

1. Datenbereitstellung

Zur Datenbereitstellung zählen wir alle Schritte von Daten besorgen über Daten säubern bis hin zur Datentransformationen. Desweiteren betrachten wir die Nutzung des Data Mining bei verschiedenen Schritten in der Datenbereitstellung.

„No Data, no Data Mining."

Als erste und gleichzeitig schwierigste Schritte müssen Daten besorgt werden. Die üblichen Quellen sind die eigenen Datenbanken – Kundendaten, Auftragsdaten, technische Daten usw. Dabei entdeckt man sofort, dass die IT-Welt nicht unmittelbar auf Data Mining eingestellt ist.

Es gibt eigentlich keine fertige Data Mining Datenmenge. Man muß sich die Daten zusammensuchen. Oder es gibt keine geeigneten Daten und man muß sie ganz oder teilweise erst erstellen. Hierbei sind Erhebungen wie Umfragen, die Personendaten einer Teilnahme an einem Rätselwettbewerb und ähnliches häufig als Beschaffungsmittel anzutreffen.

Daher ist dieser erste Schritt gleichzeitig der schwierigste und der wesentlichste, denn welche Daten soll man nehmen?

Wenn man mit „Data Mining Laien" zusammenarbeitet, die gleichzeitig entscheiden sollen, mit welchen Daten zu arbeiten ist, dauert diese Phase lange. Die Entscheidung und die Taten schieben sie vor sich her, da Laien nicht wissen gar wissen können, was sie an Daten bieten oder nehmen sollen.

Meist schwimmt auch etwas die Angst mit, dass sie sich blamieren könnten, wenn sie nicht das Optimum (sic!) an Daten liefern. Dabei können Sie aus der Sicht des Data Mining wahrscheinlich nur Daten-Rohmaterial liefern.

Es heißt also: Hauptsache liefern und Ansätze aufdecken. Zur Illustration hier ein kleines Beispiel aus den Anfängen unserer eigenen Webseite.

Wir haben die Uhrzeit, das Datum und die besuchte Stelle unserer Webseite sowie den Server (Quelle) eines Hits festgehalten. Diese wenigen Daten haben wir zu nützlichen Grunddaten ausbauen können.

Der Monat, der Quartal, der Wochentag und die Zugriffsstunde wurden aus dem vorhandenen Datenrohmaterial heraus errechnet und anschließend per Datentransformation schnell als weitere Informationen (zusätzliche Spalte in der Tabelle) hinzu gebaut.

Hier sehen Sie in welchem Zustand die Daten ursprünglich angefallen sind.

```
02:14,03.09.98,1,diskuss,208.209.42.37
22:25,03.09.98,1,diskuss,ics1f.k.srv.t-online.de
08:24,04.09.98,1,diskuss,pc19f1d94.dip.t-online.de
12:43,04.09.98,1,diskuss,192.168.100.5
16:10,04.09.98,1,diskuss,proxy.emea.compaq.com
20:24,08.09.98,1,diskuss,204.92.243.201
12:45,10.09.98,6,diskuss,gw.taschen.com
14:47,10.09.98,1,diskuss,pacific.ndh.net
17:36,11.09.98,1,diskuss,195.24.94.5
15:00,14.09.98,1,diskuss,ics2f.e.srv.t-online.de
15:27,15.09.98,1,diskuss,gaowww.gao.de
16:35,19.09.98,1,diskuss,proxy3.cityweb.de
20:25,21.09.98,1,diskuss,ics2f.ac.srv.t-online.de
23:48,23.09.98,1,diskuss,idefix.whu-koblenz.de
11:26,29.09.98,1,diskuss,pc19f7a3a.dip.t-online.de
17:40,29.09.98,1,diskuss,195.24.90.240
12:45,01.10.98,1,diskuss,195.52.235.150
18:37,01.10.98,1,diskuss,212.53.208.94
00:58,03.10.98,1,diskuss,ics1f.k.srv.t-online.de
18:23,04.10.98,1,diskuss,funnel47.btx.dtag.de
18:55,19.10.98,3,diskuss,gw.taschen.com
08:01,21.10.98,1,diskuss,pc19f1ca8.dip.t-online.de
13:16,21.10.98,1,diskuss,inehou-pxy04.compaq.com
22:14,21.10.98,1,diskuss,ics1f.ac.srv.t-online.de
12:29,25.10.98,1,diskuss,ics2f.ac.srv.t-online.de
00:39,26.10.98,1,diskuss,ics1f.k.srv.t-online.de
18:08,27.10.98,1,diskuss,12.3.94.35
19:01,27.10.98,1,diskuss,195.24.96.1
18:27,31.10.98,1,diskuss,ics1f.ac.srv.t-online.de
12:23,05.11.98,1,diskuss,ics1f.do.srv.t-online.de
11:47,06.11.98,1,diskuss,lx42.lur.rwth-aachen.de
```

In dieser Form sind die Daten für den Data Miner unbrauchbar. Die ersten 14 Zeichen (inkl. Trennzeichen) bilden einen Zeitstempel. Zuerst steht die Uhrzeit in Stunde und Minuten gefolgt vom Datum. Die einzelne Zahl zeigt an, welche Anzahl Zugriffe erfolgt ist. Die Rubrik „diskuss“ ist ein Bereich des Website und zum Schluß sehen wir den Server des Website „Besuchers.“

Mit etwas datentechnischem Fleiß und ein wenig Kreativität kann man solche Zahlen zu einer Data Mining Nützlichkeit und Aussagekraft wie folgt umformen.

```
"Uhrzeit","Std","Datum","Wochentag","Monat","Quartal","Jahr","Zugriffe","Ort","Adresse"
"19:47",19,"03.04.99","Sa",4,2,1999,1,"diskuss","pc19f808c.dip.t-online.de"
"10:29",10,"10.04.99","Sa",4,2,1999,1,"diskuss","ics1f.ms.srv.t-online.de"
"06:01",6,"17.04.99","Sa",4,2,1999,1,"diskuss","slip129-37-153-164.on.ca.ibm.net"
"16:23",16,"25.04.99","So",4,2,1999,1,"diskuss","ics1f.bn.srv.t-online.de"
"20:50",20,"25.04.99","So",4,2,1999,1,"Knowledge","196.26.208.194"
```

Das geübte Auge findet bereits einige merkliche Erweiterungen. Aus der Uhrzeit haben wir die Stunde „Std“ herauskristallisiert. Aus dem Datum sind die neuen Spalten Wochentag, Monat, Quartal und Jahr entstanden. Die Anzahl der Zugriffe sowie der Websitebereich „Ort“ sind hier unverändert aufgeführt. Zum Schluß der Tabelle finden wir die Webadresse des Servers unseres Besuchers.

Die folgende Excel Tabelle zeigt die obigen Werte als Ausschnitt einer fertigen Tabelle, die man weiter verarbeiten könnte oder direkt als Import für ein Data Mining Tool verwenden kann. Als Beispiel für weitere Erweiterungen wäre die Aufteilung der Wochentage in Arbeitstagen und Wochenenden/Feiertagen. Die Stunden könnte man ebenfalls aufteilen, zum Beispiel in Perioden wie vormittags (8-13 Uhr), nachmittags (13-17 Uhr), abends (17-22 Uhr) und nachts (22-06 Uhr). Welche Aufteilungen man letztendlich wählt, hängt zum größten Teil von der Aufgabe ab. Solange man die Aufgabe nicht fest definiert hat, dass diese zusätzlichen Werte ausgeschlossen werden können, sollte man sie möglichst einschließen.

Uhrzeit	Std	Datum	Wochentag	Monat	Quartal	Jahr	Zugriffe	Ort	Adresse
19:47	19	03.04.99	Sa	4	2	1999	1	diskuss	pc19f808c.dip.t-online.de
10:29	10	10.04.99	Sa	4	2	1999	1	diskuss	ics1f.ms.srv.t-online.de
06:01	6	17.04.99	Sa	4	2	1999	1	diskuss	slip129-37-153-164.on.ca.ibm.net
16:23	16	25.04.99	So	4	2	1999	1	diskuss	ics1f.bn.srv.t-online.de
20:50	20	25.04.99	So	4	2	1999	1	Knowledge	196.26.208.194

Zum leichteren Verständnis haben wir die Titel der neuen Spalten in roter Farbe gekennzeichnet.

Festzuhalten bleibt, dass diese Grundlage aus drei ziemlich einfachen Basisinformationen „gebaut“ wurde, Zeit- und Datumsstempel sowie Seitenziel und Adresse des Besuchers.

Die Adresse des Besucherservers gibt ebenfalls einiges her. Die vierte Zeile ist ein T-Online Kunde aus Bonn, die zweite Zeile ist auch ein T-Online Kunde aus Münster, die dritte Zeile ist ein bekannter Partner aus dem IBM Net und die letzte Zeile eine IP-Adresse, die wir zufällig kannten. Zur langfristigen Automatisierung dieses Prozesses könnte man die vollständige Identifikation solcher Adressen in einer Datenbank festhalten und per Programm entsprechend aus diesem kryptischen Zustand in eine leicht verständliche Form wie Kundennummer oder Kundenname in eine weitere Spalte umstellen.

Leider ist es jedoch so, dass die Aussagequalität einer Datenmenge für einen Data Mining Neuling im voraus schlecht zu erahnen ist. Woher soll ein Data Mining Anfänger wissen können, mit welchen Informationen man wie vorgehen kann?

Daher wird aus der Praxis empfohlen, die rohen Daten sofort mit einem Data Mining Tool (z.B. bei strukturierten Daten empfiehlt sich der Entscheidungsbaum wegen der Transparenz und Nachvollziehbarkeit seiner Ergebnisse) zu untersuchen.

Diese Erstuntersuchung führt meist zu einem ersten Fokus und Verständnis für die möglichen Inhalte. Dies öffnet wiederum die Wege zu den möglichen Erweiterungspotentiale.

Bei der Erstuntersuchung findet man bereits im Bereich des **Profiling** (erste Annäherung an die Daten mit allen Werten wie Anzahl der Kategorien, Minimum- und Maximumwerte, Standardabweichung usw.) einen Überblick über entsprechender Datenqualität.

#	Feldname	Datentyp	# von Kategorien	# fehlende Werte	Minimum	Maximum	Standardabweichung
1	Uhrzeit	String	312	0	00:00	23:48	312
2	Std	Zahl	23	0	0	23.0	5.77
3	Datum	String	170	0	01.02.99	31.12.98	170
4	Wochentag	String	7	0	Di	So	7
5	Monat	Zahl	9	0	1.0	12.0	3.02
6	Quartal	Zahl	4	0	1.0	4.0	0.97
7	Jahr	Zahl	2	0	1998.0	1999.0	0.41
8	Zugriffe	Zahl	4	0	1.0	6.0	0.39
9	Ort	String	7	0	AAG	diskuss	7
10	Adresse	String	227	0	12.3.94.35	ws26.zsw.uni-ulm.de	227

In der Aufstellung sieht man zum Beispiel die eindeutigen Werte der Datenmenge.

Wenn man ein Feld „Monat“ in der Datenmenge hat, die alle Monate als Namen aufführen soll, und der Wert 15 unter dem Feld „Monate“ vorkommt, stimmt etwas mit

den Daten nicht. Es sollte 1 bis 12 Werte (je nachdem wieviele unserer 12 Monate überhaupt vorkommen) geben . Wenn es mehr als 12 könnte man sogar 13 akzeptieren, wenn die Software eine Art Datenmülleimer wie „???“ anwenden kann. Mehr als 13 ist nicht akzeptabel. Der Grund könnte in diesem Beispiel jedoch sein, dass die jeweiligen Monate teilweise als Zahlen von 1 bis 12 und teilweise als textliche Monatsbezeichnungen vorkommen. Wir haben im obigen Beispiel lediglich 9 Monate. D.h. nicht alle 12 Monate kommen vor.

Im gleichen Beispiel kommen 23 unterschiedliche Stundenzahlen, die vollen 7 Wochentage, alle 4 Quartale und zwei verschiedene Jahreszahlen vor.

Weitere Anzeichen für fehlerhafte Daten könnten zum Beispiel vier Geschlechter sein. Wir erwarten zwei und könnten mit drei Werten inkl. Datenmülleimer leben. Mehr gibt es nicht. Kurz nachschauen und dann sieht man, ob sich was in den Daten verdreht hat, oder ob es mehrfache Darstellungen gibt, wie bei diesem Beispiel „männlich“, „weiblich“, „m“, „w“, „female“, „male“ in einer Mischform.

Ähnlich muß man vorgehen, wenn 10 Wochentage vorkommen. Da stimmt was nicht. Gibt es ebenfalls 6-stellige Postleitzahlen aus Deutschland, stimmt was nicht. Als weiteres Beispiel: Gibt es Umsätze mit einem Minus-Vorzeichen, obwohl keine Gutschriften in dieser Spalte erscheinen dürften, stimmt wieder was nicht.

Gibt es vielleicht fehlende Werte und wenn ja in welchem Umfang. Eine große Anzahl fehlender Werte in entscheidenden Feldern, könnte die ganze Datenmenge unbrauchbar machen.

Nach der Aufdeckung solcher Fehler muß man zurück zur Datenmenge, um die Korrektur dieser Fehler vorzunehmen, die man Datensäuberungen nennt. Alle solche Anzeichen zeugen von Unsauberkeit innerhalb der Dateninhalte und müssen bearbeitet, d.h. repariert, ergänzt und/oder erneut geladen werden.

Als nächstes werden die Daten wieder vom Data Mining Tool aufgenommen und erneut einem solchen Profiling unterzogen. Hierbei werden kleinere Fehler entdeckt und notiert, während man beginnt sich die Zusammenhänge zu merken.

Kleinere Blessuren wie „0“ und „1“ durch „Nein“ und „Ja“ zu ersetzen kann unter Umständen durch einen Mapping im Data Mining Tool vorgenommen werden. Bei großen Datenmengen empfiehlt es sich grundsätzlich an der Datenquelle - zur Entlastung des Systems – solche Transformationen vorzunehmen.

Der Vorgang der Datenaufbereitung kann sich mehrfach wiederholen. Nachdem längst erste Data Mining Untersuchungen durchgeführt wurden, kann es immer noch vorkommen, dass sich Datenerweiterungen oder –bereicherungen anbieten. Denn oft führen erst fortgeschrittene Data Mining Erkenntnisse dazu, wie man die Qualität seiner Ergebnisse durch besondere Datenerweiterungen steigern kann.

Dieser Abschnitt wird eigentlich nie endgültig abgeschlossen, jedoch er wird irgendwann mit dem Schwerpunkt der Arbeit in den nächsten Haupt-Arbeitsschritt des Projektrades übergehen.

2. Data Mining

Nach der oft zeitraubenden und umfangreichen Fleißarbeit der Datenbeschaffung und –vorbereitung folgt nun der Schritt des eigentlichen Data Mining, der meist als der spannendste Teil des ganzen Projektes angesehen wird. Hier werden erste umfangreiche Zusammenhänge oder zum ersten Mal deutliche Anomalien in der zu untersuchenden Datenmenge erkannt.

Die genaue Gestaltung des Arbeitsschrittes Data Mining ist direkt von der Aufgabenstellung und den dafür geeigneten Methodologien abhängig, jedoch bietet die folgende Reihenfolge für die meisten Fälle einen zuverlässigen Ansatz. Eine eingehende
Beschreibung dieser Funktionen finden wir im Kapitel 11 „Data Mining Verfahren im Vergleich.

- **Profiling** – um sich der neuen Datenmenge zu nähern und somit erste Zusammenhänge und Verteilungen zu erkennen. Profiling besteht aus der Berechnung und Darstellung von statistischen Werten wie Standardabweichung, Minimum- und Maximumwerte usw. wie oben aufgeführt. Ferner gehören hierzu die grafische Darstellung der Daten u.a. aus der Sicht eines bestimmten Variablen in zum Beispiel Balken- oder Tortengrafiken. In einer solchen speziellen Ansicht kann man die Datenverteilung leicht erkennen. Zum Beispiel möchten man (um beim obigen Beispiel zu bleiben) sehen, wieviele Besucher man an einem bestimmten Wochentag im Vergleich zu den anderen Wochentagen hat. Kurzum: Man bekommt hierbei einen Einblick in die Verteilung der Inhalte aus bestimmter Sicht. Dies ist noch kein eigentliches Data Mining, sondern lediglich eine bestimmte Ansicht der Datenmenge.

- **Entscheidungsbaum** und/oder **Clustering** – um Transparenz zu erzeugen. Beim Beginn des Data Mining Prozesses sucht man meist nach deutlichen Zusammenhängen und Korrelationen. Durch die Erkenntnisse versucht man einen Fokus zu schaffen, bestimmte Vorgänge zu erklären oder deutliche Segmentierungen zu erreichen. Kurzum: Diese Funktionen liefern deutliche mathematisch statistische Gruppierungen, die man auch als „Verkettung von Faktoren nach Wichtigkeit geordnet“ verstehen kann.
    - Die Ermittlung der signifikanten (wichtigen) Variablen bildet eine Kette von Bedingungen, die einen deutlichen Fokus und damit klare Verhältnisse liefert. Ferner dient dieses Ergebnis als Vorbau zur Konfiguration eines Neuronalen Netzwerkes.

  - Die Regelsätze erklären jeden Entscheidungsbaum. Im Prinzip sind dies die Fragen, die man an die Datenmenge hätte stellen müssen, um die im Entscheidungsbaum erzeugten Ergebnisse direkt zu erzeugen. Sich jedoch ohne Entscheidungsbaum diese präzise und komplexe Fragestellung einfallen lassen zu können, ist im Vorfeld zeitraubend, wenn nicht nahezu unmöglich. Den Entscheidungsbaum aufzubauen und anschließend Regelsätze zu generieren, ist dagegen kinderleicht.
  - die Hebelwirkungen, zeigen alle Einflüsse auf eine bestimmte Kategorie einer abhängigen Variablen auf. Solche Informationen erklären die Einflüsse bei einem Entscheidungsbaum in einem einzigen Bericht. Ebenso werden die Bestandteile eines bestimmten Clusters (oder Segmentes) klar aufbereitet.

- **Neuronale Netzwerke** – bieten diverse komplexe mathematische Gruppierungsmethoden an, die auf spezielle Weise Zusammenhänge bilden, die sich gut zur Ermittlung von Vorhersagen eignen. Diese Ermittlung erfolgt auf der Basis einer mathematisch errechneten Prognose. Wenn man vorher dafür Sorge trägt, dass die verwendeten Variablen deutliche Indikatoren der gestellten Zielaufgabe sind, können erstaunlich genaue Prognosen ermittelt werden. Praktische Endergebnisse kennt man zum Beispiel im deutschen Fernsehen bei den Hochrechnungen diverser Institute an Wahlabenden.

- **Scoring/Validierung** – Die Ergebnisse von neuronaler Arbeit ist auf Anhieb schwer verständlich. Um die Resultate eines neuronalen Netzes verständlicher machen zu können, haben wir die Validierung. Darüber hinaus, kann man vor allem große Datenmengen mit der Wertigkeit eines neuronalen Vorhersagemodells versehen – dies nennt man Scoring (oder auch Ranking). Das Scoring ist im Prinzip die Eingliederung eines neuen Ordnungskriteriums in die Datenmenge. Zum Beispiel könnte man eine Kundendatenbank mit dem Kriterium versehen, welche Kunde am wahrscheinlichsten einen Kauf tätigen wird. Somit könnten Verkaufsanstrengungen auf diese per Scoring ermittelte Gruppe konzentriert werden. Setzt natürlich voraus, dass ein solches neuronales Modell bereits aufgebaut wurde.

- **Lift Charts** – um unterschiedliche Entscheidungsbäume nach den unterschiedlichen Kategorien oder die Ergebnisse unterschiedlich konfigurierter neuronaler Netzwerke grafisch darzustellen und vom Nutzwert miteinander vergleichen zu können.

- **Genauigkeitsgrafiken** - um vor allem bei abhängigen Variablen mit fortlaufenden Werten verschiedene Entscheidungsbäume oder die

Ergebnisse unterschiedlich konfigurierter neuronaler Netzwerke grafisch darzustellen und miteinander vergleichen zu können.

- Im Falle von **nicht strukturierten Daten** (Textstrings oder Pixel-Grafiken evtl. sogar von Fotos erstellt) gibt es eine Reihe von geeigneten Werkzeugen, die zuerst eine Ordnung und dann eine Strukturierung der Daten erzeugen können.
  - **Textstrings** können mit Extrawerkzeugen, wie Connex, nach Häufigkeiten bestimmtet Strings untersucht, die anschließend bei der Überführung in eine strukturierte Form prägend wirken können. Zum Beispiel könnte der Verfassungsschutz oder die Nachrichtenspezialisten bei der Bundeswehr Telefon- oder Funksignale aufgefangen haben. Die Signale möchte man nach bestimmten Kriterien untersuchen. Vorab muß man die Häufigkeit bestimmter Textstrings (oder Begriffe) ermitteln. Diese Ergebnisse werden in eine Struktur wie in eine Tabelle eingeordnet. Wenn man hiernach die Palette der Data Mining Werkzeuge wie oben anwendet, hat man schnellstens eine Einblick in die Bedeutung der aufgefangenen Signale entdeckt.
  - **Pixelgrafiken** müssen mit speziell dafür geeigneten Werkzeuge des Data Mining bearbeitet werden. Eine sinnvolle Umwandlungen in strukturierte Form erscheint von der Sache her sinnlos. Spezielle Systeme werden hier angewendet, die zum Beispiel im Bereich der Bilderkennung bei Sicherheitssystemen eingesetzt werden.

Hier wäre der erste und grundsätzliche Arbeitsschritt des Data Mining beendet. Dieser Schritt bildet einen untersuchenden Ansatz, der bestehende Situationen erklären kann und gleichzeitig die Grundlage für weiterführende Schritte bietet. Die möglichen Folgeschritte stellen die deutliche und für manche schwer nachvollziehbare Macht des Data Mining dar. Leider können viele kommerziell vertriebene Data Mining Produkte diese „Zauberei“ des Data Mining nicht unterstützen. Hierzu gehören im allgemeinen zwei Folgeschritte:-

- Aufbau eines Datenmodells in Form eines neuronalen Netzes (oder einfacher in einem Entscheidungsbaum) als Vorhersagemodell.

- Einbau eines Datenmodells in eine Data Mining getriebene Anwender Applikation, die es Sachbearbeitern ermöglichen, die Macht eines Data Mining Modells zu nutzen, Mehrwerte erzeugen zu können, ohne irgendeine Ahnung von Data Mining zu haben. Dies könnte die präzise und umfangreiche Prüfung von Kreditanträgen durch Kundenberatern einer Bank sein.

Diese vorgeschlagene Reihenfolge des Data Mining kann jederzeit abgebrochen oder verzweigt oder unterbrochen werden. Zum Beispiel könnte ein Entscheidungsbaum einen Regelsatz ergeben. Dieser Regelsatz könnte man als SQL-Abfrage ausgeben

und separat speichern. Diese SQL-Abfrage könnte als zusätzliche Query in eine bestehende SQL basierte Anwender Applikation eingebaut werden.

Als weiteres Beispiel könnte es sich um ein SIAS (Suck-it-and-see)-Projekt handeln. SIAS-Projekte dienen der Bewältigung der Frage, ob überhaupt was „Interessantes“ in einer Datenmenge zu entdecken ist. (Übrigens bilden diese SIAS-Projekte die einzige Projektart, die in der Zielsetzung / Aufgabenstellung keine Meßlatte des Erfolges beinhalten.)

Auf jeden Fall muß der Schritt des Data Mining mit handfesten Erkenntnissen in Form von nachvollziehbaren Ergebnissen abschließen. Diese Ergebnisse können in grafischer Form, Regeln in natürlichen Sprache (z.B. deutsch), Regelsätze in Programmiersprache (Java, SQL usw.), Leverage Reports, Lift Charts usw. sowie Entscheidungsbäume vorliegen.

Aus diesem entdeckten vielfältigen Wissen muß in der nächsten Phase des Data Mining Projekt-Rades eine Entscheidungsbasis gebaut werden, die in englischer Sprache bildlich „actionable information“ genannt wird.

3. Actionable Information

Dieser schöne lebendige Begriff aus dem Englischen spricht von Informationen auf deren Basis agiert werden kann. Diese Phase wird in der deutschen Sprache als Herstellung einer Entscheidungsbasis verstanden.

Aus den diversen Erkenntnissen, die man während der bisherigen Data Mining Phasen gewonnen hat, muß bei diesem Arbeitsschritt eine bewertete Zusammenstellung erfolgen.

Meist ergibt sich diese Entscheidungsbasis aus der qualifizierten Interpretation der Erkenntnisse durch die „Kenner“ der Dateninhalte. Hier gilt es nun, diese Erkenntnisse in eine verständliche Form zu bringen, woraus entsprechende Vorschläge und Empfehlungen für Aktionen hervorgehen.

Diese Entscheidungsbasis wiederum muß oft in eine "Vorstandsform" (möglichst eine Seite) zusammengefaßt werden. An dieser Stelle sind weniger die möglichen Entscheidungen und ihre jeweiligen Erklärungen bedeutsam, sondern reine Management Überlegungen von Bedeutung. Fragen wie, was kostet, was bringt es oder welches Risiko wird eingegangen, dürften zum Fragenkomplex gehören.

Nun gilt es beratend tätig zu werden, denn die Entscheidung zur weiteren Vorgehensweise kann operativ oder simulativ ausfallen. Mit „operativ“ meinen wir, dass die Entscheidung sofort in den praktischen Einsatz gehen soll. Mit „simulativ“ meinen wir eine mit dem Computer künstlich hergestellte Situation, die dazu dienen kann, vorab eine Aktion zu testen. Zum Beispiel kann man eine Marketingaktion

anhand von Datenmengen im Computer auf ihre mögliche Erfolgsrate „prüfen." Der simulative Einsatz ist klassischerweise bei Marketing Kampagnen eine große Hilfe. Hier lassen sich unterschiedliche Kampagnen auf dem Computer in Form einer vorhersagenden Aktion künstlich – und damit vergleichsweise für wenig Geld und schnell – erzeugen.

Ob man sich für die operative oder vorerst für die simulative Lösung entscheidet, es ist Aufgabe dieser Phase für die Bereitsstellung der entsprechend detaillierten Entscheidungsbasis zu sorgen. Bei einer Marketingaktion kann dies zum Beispiel lediglich die präzise Mitteilung der effektivsten Zielgruppe bedeuten oder bei einer Cross-Selling Aktion, die Erstellung, Eingliederung und Einrichtung entsprechender Software mit Datenmodellen umfassen.

Mit anderen Worten findet eine Entscheidung statt, heißt es nun mit den vom Data Mining Projekt gefundenen Erkenntnissen im geschäftlichen Umfeld endlich tätig zu werden.

Auch an dieser Stelle kann es sein, dass die Ergebnisse aus den unterschiedlichen „politischen" Gründen eines Unternehmens nicht zum Einsatz kommen und der Prozeß abgebrochen wird. Ferner könnte es passieren, dass man eine andere Gruppe - eventuell ein externes Team - darum bittet, eine „unabhängige" Prüfung der gefundenen Entscheidungsbasis vorzunehmen.

Das Ziel dieser Tätigkeit kann nur sein, bestimmte Vorteile für Ihre eigene Institution zu erreichen. Ob und inwieweit diese Ziele wirklich Vorteile schaffen, wird an der abschließenden Phase gemessen.

4. Ergebnisse messen

Nach Einsatz diverser Erkenntnisse – ob operativ oder simulativ – möchte jeder wissen, was die Maßnahme gebracht hat. Der Data Miner sollte grundsätzlich darauf bestehen, dass dieser letzte Schritt seines Prozesses durchgeführt wird. Denn eine Data Mining Maßnahme ohne bewerteten Abschluß ist wertlos. Data Mining ermöglicht Mehrwerterzeugung. Data Mining ist damit als eine der wenigen Computer gestützten Disziplinen, die eine solche Behauptung aufstellen und halten können. Daher sollten Sie auch in Erfahrung bringen, welche Höhe dieser erzeugte Mehrwert in der Tat auch ausmacht. Denn wenn Data Mining Projekte im Ergebnis gemessen werden und als Erfolg ausfallen, wird der weitere Einsatz von Data Mining Projekten meist wenig in den Weg gestellt werden.

Der Hauptvorteil liegt jedoch in der meist strategischen Natur der üblichen Data Mining Projekte. Ein Teil der Zielsetzung muß mit einem quantitativen Bestandteil formuliert sein. Dies bietet die Möglichkeit der Messung am Ende des Projektes. Die Kosten/Nutzen Diskussion verfällt somit zu einer Aufsummierung des „return-on-

investment“ (ROI) oder ein Vergleich von aktuellen Kosten zu ehemaligen Kosten oder gar Marktanteil heute gegenüber vor zwei Jahren usw.

An dieser Stelle am Abschluß des ganzen Data Mining Projektes wird endgültig in absoluten Zahlen der Mehrwert des kompletten Vorgangs plastisch und arithmetisch ausgerechnet und deutlich gemacht. Zum ersten Mal wird für jeden noch so unbeteiligten Zuschauer deutlich, „warum und wozu das ganze.“

Gleichzeitig werden notwendige oder sinnvolle Folgeprozesse aufgezeigt und können recht einfach nach gewinnbringenden Prioritäten eingestuft werden.

Aus der Sicht der IT-Leute wird der Kampf um Etats endlich durch eine neue Dimension erleichtert. Mit Data Mining verfügt man über eine Sparte, die nicht nur Geld kostet, wie Hardware oder andere Software, sondern – vorausgesetzt Data Mining wird ordentlich betrieben – meßbare Geldvorteile erzeugen läßt.

Kurzum**: Mit Data Mining wird Geld gemacht. Eine durchaus neue Erfahrung für die meisten IT-Verantwortlichen**.

7. Vertraulichkeitserklärung

Bevor Daten zur Verfügung gestellt werden, sollte die sensible Natur der Daten als auch vor allem der potentiellen Ergebnisse bedacht werden.

Data Mining ist eine äußerst intime Angelegenheit, daher ist es beinahe ein gesichertes Muß, dass sich beide Seiten vor Beginn der Arbeiten – egal welche Größenordnung von Projekt avisiert ist -, durch eine Absicherung aller Eventualitäten schützen.

Solche Maßnahmen werden meist Vertraulichkeitserklärung oder volkstümlich „Schindludererklärung“ genannt. In der Anlage 1 finden Sie Beispiele für solche Absicherungen.

Viele Konzerne haben heute bereits viel Erfahrung mit der Intimität ihrer Datenbestände und verfügen über entsprechende Vordrücke, die aus ihrer Rechtsabteilung stammen und den Segen des Vorstandes genießen.

Die unterschiedlichen Absicherungen reichen von halben Gesetzesvorlagen bis hin zu einem formlosen Schreiben, wobei alle entsprechenden Ausschlußhandlungen und Verpflichtungen aufgeführt werden.

Zum Schutz beider Parteien – MINER und KUNDE – sollte erst nach dem Abschluß einer schriftlichen Vertraulichkeitserklärung die Arbeit aufgenommen werden.

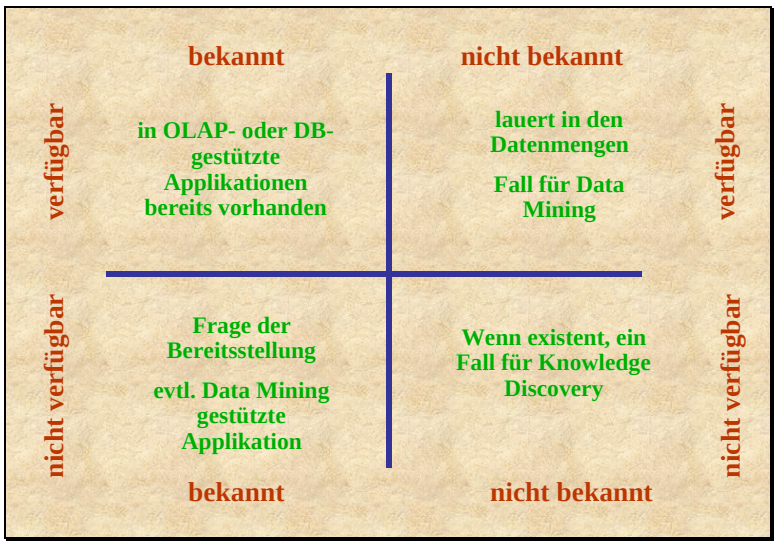
8. Stand des Wissens

Nach dem Abschluß der Vertraulichkeitserklärung sowie nach den ersten Gedanken zum Projektinhalt und eventuell ersten Einblicken in die verfügbaren Datenmengen sollten man sich über den aktuellen Stand des zum Thema des Projektes vorhandenen Wissens Klarheit verschaffen.

Insbesondere bei Hauptprojekten, wo entweder ein neuer Bereich untersucht wird oder das Projekt einen übergeordneten, strategischen Charakter besitzt, kann diese Überlegung erste große Gewinne sowie eine unbezahlbare Orientierung oder Meßlatte darstellen.

Auch der Data Miner befindet sich in einer Position der Verantwortlichkeit, die er nicht unterschätzen sollte. Das heißt, er muß mit der Umgebung und in der Situation leben und arbeiten, die er vorfindet. Es bringt dem auftraggebenden Unternehmen nichts, wenn der Data Miner die wunderbarsten Dinge entdeckt und anpreist, um unmittelbar zu erfahren, dass dies alles bereits bekannt ist und schlimmer noch, sich bereits im Einsatz befindet. Der schlimmste Albtraum des Data Miners besteht darin zu erfahren, dass seine Erkenntnisse und Entdeckungen bereits vor kurzer Zeit aus unterschiedlichen – beispielsweise firmenpolitischen – Gründen verworfen wurden.

Das kostet viel Zeit und Geld und wirkt äußerst peinlich. Leider ist diese Situation nicht immer zu vermeiden. In manchen Unternehmen wird trotz Knowledge Management (ein System zur Verwaltung und Bereitstellung des im Unternehmens verfügbaren Wissen und der nachweislichen Fähigkeiten, die Mitarbeiter vorweisen können) etwas „undurchsichtig“ mit Wissen umgegangen.



Stand des Wissens

Manchmal wird aus fast perversen Sicherheits- oder Machtüberlegungen heraus, erst nach Abschluß eines Projektes damit geprahlt, dass diese oder jene Erkenntnis bereits vorher erkannt wurde. Solchen Menschen ist wohl kaum zu helfen.

Um bereits im Vorfeld eines Projektes eine geordnete, geplante Vorgehensweise zu gewährleisten, sollte man versuchen, Schwerpunkte zur Orientierung festlegen. Der obige Matrix kann Ihnen dabei nützlich sein.

Das Diagramm „Stand des Wissens“ stellt die Verfügbarkeit des Wissens dem Bekanntheitsgrad des Wissens gegenüber. Es gibt vier mögliche Betrachtungsvarianten, die wir näher einzeln nacheinander betrachten wollen.

i) Wissen bekannt und verfügbar

In zweierlei Hinsicht bildet dieser Quadrant den größten Stolperstein für den Data Miner.

Erstens steht ihm dieses bekannte Wissen zur Verfügung und er muß entscheiden, ob er dieses Wissen als wertvoll und richtungsweisend anzunehmen möchte. In der Tat steht er unter Zeit- und/oder Leistungsdruck und bestehende Annahmen in einem Unternehmen werden zu oft einfach akzeptiert. Dies erweist sich später bei der Projektarbeit als Fehler. Während eines Projektes werden plötzlich seltsame Entdeckungen macht, die man auf Anhieb nicht erklären kann. Zum Schluß wird erkannt, dass die ursprünglich akzeptierten Grundannahmen nicht mehr stimmten. Es kommt vor, dass bekanntes und verfügbares Wissen in einem Unternehmen lange nicht mehr überprüft wurde.

Die Annahmen in einem Unternehmen oder gar einer ganzen Branche werden ohne Prüfung einfach weitergeführt und erhalten einen Status als beinahe feststehende Wahrheit. Zum Beispiel ist uns ein Fall bekannt, wobei ein Versicherungsunternehmen im Kfz-Bereich seine wahrscheinlichen Schadenshöhen gegenüber der Schadenshäufigkeit berechnete. Daraufhin wurden die vom Versicherten zu zahlenden Prämien errechnet. Die grundsätzliche Berechnung dieser entscheidenden Eckwerte lag inzwischen fast 10 Jahre zurück. In der Zwischenzeit hatte eine massive Verschiebung der Kundenstruktur stattgefunden. Die Folge wäre vielleicht eine absolute Katastrophe geworden, denn jene Versicherungssparte wäre Pleite gegangen und gemerkt hätte man es kaum jemand. Schließlich nahm man an, dass die Eckwerte stimmten, und was nicht sein darf, wird nicht passieren.

Also grundsätzlich alle Eckwerte und Annahmen sollten vorher nachgeprüft werden. Das hier angesprochene Wissen soll als bekannt und lediglich als vermutet gelten. Alle Schritte müssen auf einem gesicherten Grund erfolgen.

Zweitens könnte der Data Miner dieses bekannte verfügbare Wissen mißachten, ein enormes Projekt durchführen, das in sich ein Spitzenvorgang bildet, um spätestens in

der Phase 3 des Projektrades „actionable information“ zu erfahren, dass sich dieses Wissen bereits im Einsatz befindet.

Hier ist Vorsicht geboten. Bei einem Projekt, dass nicht auf der Basis von komplett neuen Daten erfolgen soll, bringt es viel, wenn man sich erkundigt, welches Wissen bereits vorherrscht.

Wenn bereits Wissensperlen genannt werden, kann man eine Überprüfung der Richtigkeit dieser vermeintlichen Annahmen vornehmen. Anschließend wird der Ausbau des vorhandenen Wissens behutsam betrieben.

ii) Wissen bekannt aber nicht verfügbar

Eine weitere Kategorie des Matrix „Stand des Wissen“ ist, dass das Wissen zwar bekannt ist, aber nicht verfügbar. Die Gründe hierfür können vielfältig sein.

Wenn das Wissen nicht verfügbar ist, kann es viele Erklärungen hierfür geben. Es kann sein, dass das bekannte Wissen deswegen nie verfügbar gemacht wurde, weil entsprechende Regelsätze, die das bekannte Wissen im Alltag nutzbar gemacht hätten, nie erstellt wurden. Es können „firmen- oder personalpolitische“ Gründe vorherrschen, die einen Einsatz des bekannten Wissens verhindert haben. Vielleicht fehlten die passenden Tools oder die entsprechenden Fähigkeiten, um das Wissen nutzbar zu machen. Oder die entscheidenden Menschen des Unternehmens haben den Wert des bekannten Wissens nicht erkannt oder erkennen wollen.

In jedem Fall ist diese Rubrik eine Frage der Bereitstellung und der Data Miner kann seine Fähigkeiten und Tools verwenden, um – bei Bedarf – die Verfügbarkeit dieses bekannten Wissens herbeizuführen.

iii) Wissen unbekannt aber verfügbar

Dieser Bereich unseres Matrix „Stand des Wissens“ bildet das dankbarste sowie das momentan häufigste Betätigungsfeld des Data Miners.

Wenn Wissen verfügbar ist, jedoch noch nicht erkannt wurde, liegt ein Tätigkeitsfeld vor, auf dem der Data Miner seinen Wert deutlich zeigen kann. Unter Verwendung seiner Werkzeuge und Methoden kann er dieses unbekannte Wissen aufdecken, verfügbar machen und dabei helfen, es begreiflich zu machen.

iv) Wissen unbekannt und nicht verfügbar

Hier findet der Vollblut Data Miner die Aufgabe, die ihn wirklich herausfordert.

Wenn das Wissen nicht verfügbar und gänzlich unbekannt ist, muß er mit der Schaffung seiner Datengrundlage sowie mit dem Aufbau seiner Bewertungsbasis anfangen. Dies muß diese ohne hilfreiche Zielsetzung geschehen, denn er befindet sich in der Rolle des Entdeckers, der im voraus nicht genau wissen kann, worauf er eigentlich stoßen wird.

Die Kreativität des Data Miners sowie die entstehende Innovation in der Interpretation bestreiten neue Wege der Erkenntnis für die einsetzende Institution. Spätestens an dieser Stelle beginnt die Diskussion ob nun Data Mining eine Wissenschaft oder eine Kunst ist. Hier finden wir jedoch die vielleicht purste Form des Data Mining.

Fazit

Es ist bestimmt hilfreich, die Aufdeckung von nutzbarem, mehrwerthaltigem Wissen sowie seine Bereitstellung als Spiel zu betrachten. Diese Einstellung bietet den notwendigen Abstand, was wiederum hoffentlich die hilfreiche, intellektuelle Lockerheit freimacht, die zur Erkennung neue Wissensperlen erforderlich ist.

Die Ergebnisse des Spiels stehen zum Schluß so plastisch dar, wie das Fußball-Ergebnis eines Spiels, das man selber nicht gesehen hat. Man riskiert gar nichts, wenn man manchmal recht kontroverse und unorthodoxe Mittel und Wege bestreitet, die ein Sachkenner vielleicht von Beginn an verurteilen würde.

Witz und Mut zum Experiment sind gefragt. Da es sich meist um wenigen Minuten oder gar Stunden handelt, sollte man gelegentlich „verrückt“ anmutende Vorhaben einfach ausprobieren. Oder glauben Sie, dass Kolumbus und seines gleichen oder unsere Raumfahrer ihre Leistung innerhalb der geltenden Normen erreichten. Nein. Alle Menschen der Geschichte, die für neue Erkenntnisse und neue Wege gesorgt haben, mussten sich außerhalb der üblichen Wege und der als Norm geltenden Grenzen bewegen. Man muß gelegentlich andere Betrachtungsweise wagen und deswegen kann es sein, dass man als Data Miner schnell als unbequem oder gar verrückt angesehen wird. Lassen Sie sich nicht entmutigen. Sie verfügen schließlich über genügend Werkzeuge und Methoden, kühn anmutende Behauptung überprüfen gar beweisen zu können.

9. Data Mining Challenge

Nachdem man die diversen Erkenntnisse über den möglichen Ablauf eines neuen Data Mining Projektes verstanden und verinnerlicht hat, treten bereits die ersten Schwierigkeiten auf, die den Beginn des Projektes hindern. Zum Beispiel: Es gibt keine Daten oder die Vorteile des Data Mining sind bei manchen entscheidenden Mitarbeitern der auftraggebenden Institution nicht eingehend verstanden worden.

Am häufigsten tritt folgendes Problem auf. Die entscheidende Mitarbeiter nehmen an eine ausführliche Demonstration des Data Mining teil. Die Vorteile eines Data Mining Projektes werden mit Fremddaten illustriert. Der durchschnittliche Teilnehmer einer solchen Veranstaltung versteht während der Demonstration zwar die Wirkung des Data Mining auf die vorgeführten Datenmenge, kann sich jedoch gar nicht vorstellen, wie sich das vorgeführte in sein eigenes Umfeld mit seinen gewohnten Dateninhalten übertragen lässt. Erst wenn das Verständnis bei den entscheidenden Mitarbeitern für die Möglichkeiten des Data Minings in deren gewohnten Umfeld erreicht wurde, können praktische und sinnvolle Vorstellungen sowie Hochrechnungen über einen möglichen Einsatz des Data Mining vorgenommen werden.

Leider zeigt die Erfahrung, dass etwa 90% aller Menschen diesen Sprung von der Demonstration auf der Basis fremder Daten bis zum eigenen Umfeld nicht schaffen. Für den Data Miner bildet diese Situation eine Blockade zum Projektbeginn. Um diese Blockade zu durchbrechen, muß eine Möglichkeit gefunden und für die entscheidenden Mitarbeiter in einem Unternehmen verständlich gemacht werden, die ihre gewohnten Inhalte und ihr bekanntes Umfeld in ein Data Mining Projekt abbildet.

Bei VONFORMAT hat man einen Weg entwickelt, den man mutig „Data Mining Challenge“ nennt. Data-Mining-Challenge baut eine WIN-WIN Situation auf , wo sich sowohl das interessierte Unternehmen als auch der lieferwillige Data Miner in die Pflicht nehmen.

Das interessierte Unternehmen bietet seine eigenen Daten und eine geringe Summe (ab DM 10.000,-). Auf der Basis dieser Datenmenge liefert der Data Miner innerhalb drei Wochen einen Bericht ab. In diesem Bericht ist ein Mini-Projekt abgebildet. Jeder Entscheider kann inhaltlich seine eigene Umgebung wieder entdecken und erhält für ihn sachlich nachvollziehbare Ergebnisse, die den möglichen Gewinn (Return on Investment) eines größeren Projektes hochrechnen lassen.

Im Kapitel 5 sahen Sie ein Beispiel für ein Data Mining Challenge Bericht. Das Szenario ist die Schadensabteilung einer Versicherung. Hier sollen Schadensdaten auf mögliche Muster untersucht werden.

Das Ziel eines Data-Mining-Challenge ist nicht, eine vollständige Untersuchung vorzulegen, sondern dem auftraggebenden Unternehmen einen Einblick in die

möglichen Werte seiner Datenmenge zu geben. Gleichzeitig wird für Data Mining Fremdlinge aus der Sicht ihrer eigenen Daten, einen Anfangsszenario gebaut, das ihnen den Wert eines Data Mining Projektes daran aufzuzeigen, welche Erkenntnisse man überhaupt gewinnen kann.

Für den Data Mining Neuling das Ganze im bekannten Umfeld aufzubereiten, macht das Verständnis der möglichen Vorteile beinahe zur Selbsterklärung, da er sich in seinen Dateninhalten wieder finden kann. Eine Live-Vorführung der Ergebnisse direkt vor der Übergabe des Berichtes erzeugt eine zusätzliche Spannung. Vom Ausgangspunkt der erhaltenen Datenmenge, über die diversen Zwischenschritte bis hin zum vorläufigen Endergebnis kann - unter Verwendung der Data Mining Werkzeuge – der Weg deutlich illustriert werden und gleichzeitig Fragen behandelt werden.

„Seeing is believing“ sagt man englisch; von der Bedeutung her etwa, was man selber gesehen hat, kann man glauben. Es hat sich über Jahre bestätigt, dass ein Vorgehen – wie oben beschrieben – für den Start als auch für den weiteren Ablauf der Data Mining Arbeit von grundsätzlicher Bedeutung ist, die entscheidenden Mitarbeiter für solche Maßnahmen zu gewinnen.

10. Pilotprojekt

Ein Data Mining Challenge ist eine einfache Möglichkeit praktische Testresultate für wenig Geld und mit geringem Aufwand zu erstellen, auf deren Basis größere Projekte in der Gewinnträchtigkeit hochgerechnet werden können. Einige Auftraggeber übergehen den ersten Schritt und steigen sofort bei dem zweiten Schritt bei einem Pilotprojekt ein.

Die Dauer des Data Mining Challenge ist von Beginn bis zur Ablieferung des Berichtes maximal 3 Wochen. Der Schritt des Pilotprojektes dauert ab 4 Wochen bis zu 3 Monaten.

Das Pilotprojekt ist bereits ein kleines Data Mining Projekt, wobei meist „Training on the Job“ dazu gehört. Also findet bereits im Pilotprojekt ein Knowledge-Transfer statt. Ferner wird oft ein „Proof of Concept“ an diese Stelle gefahren. Hier einiges zum Thema „Proof of Concept.“

i) Proof of Concept

Obwohl die eher traditionelle Literatur ein „Proof of Concept“ zu Beginn vorschreibt, wo der Wert dieser Data Mining Technologie für das jeweilige Unternehmen geprüft wird, scheitern solche Unterfangen meist an den Kosten und der zusätzlichen Dauer der Arbeiten.

Wo es um Zeit und Geld geht, sind Unternehmen immer bemüht, möglichst bald und zu einem „geringerem“ Preis zu einer Entscheidungsbasis zu kommen. Es gibt immer weniger Unternehmen die eine stattliche Summe ausgeben wollen, um erfahren zu können, ob eine ihnen unbekannte Methodik gewinnbringend weiterhelfen kann.

Den Unternehmen ist es lieber, solche Testuntersuchungen mit dem eigenen hausinternen ausführlichen Test eines Produktes oder eines Vorganges zu koppeln oder am liebsten mit einem nützlichen – wenn auch virtuellen - Gewinn zu paaren.

Der praktische Weg der Unternehmen in Deutschland liegt oft in der Beschäftigung eines Studenten mit einer Diplom- oder Doktorarbeit zu dem Thema. Hierbei sind schon einige maßgebliche und lesenswerte Arbeiten erstellt worden, die zu oft dem Vertraulichkeitsschutz des jeweiligen Unternehmens unterliegen.

Alles jedoch in ein Pilotprojekt aufzunehmen, kostet weniger, kommt wesentlich schneller zum Ergebnis und wird von externen Profis geliefert sowie von hauseigenen Datenkenner mitgestaltet. Der automatische Wissenstransfer bildet einen praktischen Einblick in die Materie, der anschließend von hohem nutzbaren Wert wird und eine sichere Basis zur weiteren Entscheidung bildet.

ii) Vorbedingungen eines Pilotprojektes

Bevor ein Pilotprojekt beginnen kann, soll der Zeitplan, die Zuständigkeit, die Teilnahme und das Ziel des Projektes festgelegt werden.

Der Zeitplan – d.h. der Liefertermin – soll dafür sorgen, dass ein Eindruck über die Laufzeiten in Verbindung mit zu erzielenden Ergebnissen gewonnen werden kann. Bei Data Mining Projekten ist der Zeitfaktor eine der wesentlichen Pluspunkte wenn man die Qualität der Ergebnisse messen möchte.

Die Zuständigkeit innerhalb des Unternehmens geklärt und festgelegt zu haben ist stets eine große Hilfe, denn ein eventuelles Kompetenzgerangel kann dem Fortschritt des Projektes hinderlich sein. Hier empfiehlt es sich eine im Unternehmen möglichst hochgestellte Person als „Schirmherr“ zu gewinnen.

Bereits bei der Datenbeschaffung, kann etwas „Rang“ äußerst hilfreich sein: Denn ein Datenbankadministrator nimmt sich ohne Auftrag selten freiwillig viel Zeit, um entsprechende Daten für ein Data Mining Projekt zur Verfügung zu stellen. Denn seine Zeit ist knapp bemessen.

Ferner müssen wir gerade mit Data Mining Projekten oft ungewöhnliche Wege bestreiten (es kann sich um eine unkonventionelle Datenzusammenstellung handeln, die in Vorschriften nicht vorgesehen sind oder gar um vertrauliche Daten), da kann etwas ranghohe Unterstützung ungeahnte Kompetenzschwierigkeit auf dem kleinen Dienstweg lösen helfen.

Bei den Teilnehmern sind Datenkenner, Analytiker, spätere oder gestandene Data Miner, Datenmanipulatoren sowie einen zuständigen Projektleiter seitens des auftraggebenden Unternehmens. Es sind nicht alle permanent anwesend im Projekt, aber ihre Dienste sowie entsprechende Plattformen sind fast eine Selbstverständlichkeit.

Sinnvolle Beispiele für Projektziele sind Untersuchungen von Datenmengen, die noch nicht vorgenommen werden konnten, weil die Aufgabenstellung bisher für herkömmliche Untersuchungsmittel zu ungenau oder „unscharf“ formuliert war. Entscheidend ist lediglich, dass die Meßlatte festgelegt wird. Es muß unbedingt in der Aufgabenstellung formuliert sein, wie der Erfolg zu messen ist. Soll die Maßeinheit als „Marktanteil gegenüber dem Vorjahr“, oder als Umsatz, Absatz, Kundenanzahl oder gar „Gewinn gegenüber Finanzplan“ gesehen werden? Obwohl die Möglichkeiten vielfältig sind, muß die Maßeinheit des Erfolgs im Vorfeld entschieden und festgelegt werden.

Ansonsten bietet die Nachuntersuchung einer bereits abgeschlossenen Analyse eine sinnvolle Vergleichsmöglichkeit für jedes Pilotprojekt. In jenem Fall kann man sich

zum Abschluß auf die Vergleiche der Ergebnisse und des jeweilig notwendigen Aufwandes konzentrieren.

Darüber hinaus, sollte das Data Mining versuchen, zumindest Ansätze eines Vorhersagemodells aufzeigen; denn solche Modelle sind zumindest mit herkömmlichen Mittel der Analytik überhaupt nicht herzustellen.

Neben der inhaltlichen Planung eines solchen Projektes sollte für externe Mitarbeiter auch hier die Vertraulichkeitserklärung nicht vergessen werden (Siehe Anlage I).

iii) Nachgestellte Untersuchung

Die Nachstellung einer bereits abgeschlossenen Untersuchung birgt einige Gefahren, aber gleichzeitig eine Menge nette Überraschungen in sich. Die Gefahren liegen in der Planung.

Vorab muß unbedingt klar gestellt werden, dass die an der hauptsächlichen Arbeit der nachgestellten Untersuchung **NICHT** bereits vorher an der ursprünglichen Untersuchung beteiligt waren. Dies gewährleistet eine vorurteilslose Untersuchung.

Ferner muß unbedingt ein Kenner der Dateninhalte zur Verfügung stehen, der bei der Interpretation der entstehenden Ergebnisse unterstützen soll. Dies dient der prompten Entscheidung über Sinn und Unsinn sowie Wertigkeit diverser Zwischenresultate.

Nichtzuletzt sollten die Vergleichskriterien vorher festgelegt werden. Meistens werden Zeitdauer, Ergebnisse und Mehrwert in den Ring geworfen.

Fallen die Ergebnisse gleich aus, wird der Vergleich der Zeitdauer in Kosten umgerechnet. Eventuelle hierbei entstandene Mehrwerte können als zusätzlicher Gewinn angesehen werden.

Deutlich ungleiche Ergebnisse (sind allerdings selten) müssen mit den Fachleuten erörtert werden. Abweichungen gibt es öfter in den zusammenhängenden Details einer Ergebniskette. In diesen Abweichungen liegen oft die entscheidenden Impulse einer Erklärung verborgen.

Gegenüber herkömmliche Analyse Methoden wie regelbasierte Systeme oder Abfrage Systematiken gewinnt üblicherweise das Data Mining immer , und zwar mit einem teilweise gewaltigen Vorteil in der Zeitdauer sowie beim notwendigen Aufwand. Oft ist mit einem Mehrwert in Form von Vorhersagemodelle sowie anderen Folgeerkenntnissen zu rechnen.

Vergleiche von unterschiedlichen Data Mining Verfahren können recht diffus ausfallen.

Manche Vergleiche sind rein akademischer Natur ohne praktischen Nutzen. Die Unterschiede zwischen mancher neuronaler Verfahren können zum Beispiel absolut akademisch verlaufen.

In der offenen Diskussion werden hier vehemente Debatten von allen Beteiligten geliefert. Jeder wird bemüht sein, die optimalen Eigenschaften seiner favorisierten Methode herauszustellen. Diese Debatten können bereits im frühen Stadium die fachliche Kompetenz sowie die gesamte Arbeitsqualität des Gegenüber komplett in Frage stellen. Der Wert solcher Gespräche ist meist gering und für den Data Mining Neuling absolut verwirrend und wertlos. Zumal unter vier Augen beim Bier von den Streithähnen zugegeben wird, dass die Unterschiede von vielen Verfahren für den alltäglichen Data Mining Einsatz kaum von praktischer Bedeutung sind.

Ähnlich verhält es sich beim Einsatz neuronaler Systeme oder nicht neuronaler Methoden. Manchmal kann man schnelle und plausible Ergebnisse mit einem Entscheidungsbaum erzeugen, die mit einem neuronalen System große Umstände, sowie Zeit und immense Systemkraft bedeutet.

Es hängt in der Tat von der genauen Aufgabe und der speziellen Situation ab, welches Verfahren sinnvoll ist. (Hierzu sehen Sie unten 7 v) „Welches Verfahren für welche Aufgabe?“ und Kapitel 8 „Data Mining Verfahren im Vergleich.“)

iv) Ein optimales Data Mining Tool gibt es nicht

Ein Optimum in der Palette der Data Mining Tools gibt es nicht. Der viel gepriesene und oft vergeblich gesuchte „Alleskönner“ gibt es im Data Mining gar nicht.

Hinzu kommen die vielen teils verwirrenden Überlappungen der Anwendbarkeit unterschiedlicher Verfahren, insbesondere im Vergleich der regelbasierten Systeme mit algorithmisch gesteuerten, „intelligenteren“ Verfahren.

Entscheidend für den Einsatz einer bestimmten Werkzeugart ist allein die jeweils zu bewältigende Aufgabe. Da Data Mining eine noch relativ junge Disziplin ist, werden oft die Erfahrungen der bisherigen Pionierarbeiten herangezogen.

Im Einzelhandel wurden die ersten Erfolge mit „Association Rules“ erzielt, die uns mittlerweile zur „Basketanalyse“ geführt haben sind. Dies ist eine Trial-and-Error Methode. Sie baut auf eigenen postulierten Hypothesen auf, die im Test eventuell zu kleinen Erfolgen führen, aus denen wiederum entsprechende Erkenntnisse gezogen werden. Hierauf werden dann weitere Schritte dieser Art entwickelt.

Eine Menge Zeit, einige Mühen und das Basiswissen über die Vorgänge mit Kunden im Einzelhandel werden als Input zusammengebracht. Idee werden aufgestellt und anschließend getestet. Im Test werden die Ideen entweder bestätigt oder sie stellen

sich als unbrauchbar heraus. Nach und nach versucht man gesicherte Erkenntnisse aufzubauen.

Der an den Entscheidungsbaum gewohnte Mensch wird mit seinem bevorzugten Algorithmus zu ähnliche Erkenntnissen in einem Bruchteil der Zeit kommen und dabei vielleicht auf weitere Effekte stoßen. Diese Zusatzeffekte können schnell die Grenzen des Basiswissens eines Einzelhändlers sprengen und völlig neue Wege eröffnen.

Wenn man diese neuen Wege gar nicht will, sondern lieber auf dem gewohnten Pfad bleiben möchte, ist der zuletzt beschriebene Weg überflüßig: Seine Qualität jedoch unbestritten wertvoll.

Der neuronale Spezialist wird bestimmt weitere Möglichkeiten oder andere Algorithmen in den Ring beordern. Summa summarum bleibt die Zielstellung, das Aufgabenumfeld sowie die Vorlieben und Erfahrungen der mitwirkenden Personen für die Auswahl entscheidend.

Als wichtigster Faktor herrscht die verfügbare Zeit. Je eher der Liefertermin gesetzt ist, je weniger werden Methoden gewählt werden können, die ein unbestimmtes Quantum an Prüfungsschritten benötigen, um endgültige Ergebnisse bieten zu können.

Nachdem wir uns dem „human factor“ zugewendet haben, stellt sich nun die Frage, welches Verfahren eigentlich für welche Aufgabe geeignet ist.

v) Welches Verfahren für welche Aufgabe?

An dieser Stelle möchten wir einige Verfahren aus der Sicht der Eigenschaften und Einsatzgebiete besprechen. Im folgenden Kapitel werden die meisten Verfahren aufgeführt, erläutert und in einer Tabelle nach der Wirksamkeit in der Projektarbeit gegenüber gestellt.

Wenn man eine Datenmenge zum ersten Mal untersuchen will, steht man vor der Frage wie man am besten beginnen soll. Die Antwort dieser Frage möchten wir von der Kenntnis der Inhalte der Datenmenge abhängig machen.

Wenn man ein Feld in der Datenmenge (Data Miner sprechen von Variablen) als aussagekräftig kennt - dies könnte sinnvollerweise Kundengruppen, Umsatzkategorien, Produktgruppen, Altersgruppen, Verkaufsregionen usw. sein -, hat man damit einen potentiellen Ausgangspunkt einer Analyse gefunden. Data Miner sprechen von einer abhängigen Variable, die meist in einem direkten Zusammenhang mit der Zielvorstellung der Projektaufgabe steht.

Mit einer abhängigen Variable eröffnen sich gezielte Möglichkeiten sofort mit der Transparenz von Profiling-Verfahren oder Entscheidungsbaumtechniken zu arbeiten.

Ein CHAID-artiger Entscheidungsbaum für den Anfang bietet eine Menge flexiblen und schnellen Einblick in die Zusammenhänge der Datenmenge.

Gleichzeitig werden weitere signifikante (wichtige) Variablen ermittelt, die zur Weitergabe an andere Systeme (z.B. neuronale) dienen können. Die Ausgabe dieser Zusammenhänge in Form von Regelsätzen, Kreuztabellen, Grafiken usw. ist ebenfalls möglich (Siehe Kapitel 8 „Data Mining Verfahren im Vergleich:“

Hat man keine Vorstellung welche Variable als „abhängige“ dienen kann, muß man sich anderweitig der Datenmenge nähern. Eine Cluster-Analyse bietet einen schönen Anfang, da hier keine abhängige Variable benötigt wird. Dadurch hat man die Möglichkeit, Einflußgrößen und Zusammenhänge ohne irgendeine Prejudizierung vornehmen zu können.

Eine Clusteranalyse (Gruppierung/Segmentierung) kann das Ziel der Untersuchung liefern und / oder sie bietet die Übergabe an weitere Methodologien.

Unterschiedliche Formen von neuronalen Netzwerken sollte man zuerst als „Blackbox“ betrachten, die eine Vorhersage liefern kann. Diese Vorhersage kann man als mathematische Simulation betrachten, die deutlich von den eingegebenen Werten abhängt. Um möglichst „brauchbare“ Werte für die Eingabe in ein neuronales System bereitzustellen, sollte man vorab eine Untersuchung mit einem der oben genannten Methoden beginnen. Die dort erzielten Ergebnisse bilden eine gesicherte Basis für die Eingabe in die neuronale „Blackbox.“

Die Engländer sprechen von „horses for courses“ (das passende Pferd für die jeweilige Rennbahn). Auch beim Data Mining brauchen Sie das passende Werkzeug für die jeweilige Aufgabe.

Näheres zu den jeweiligen Verfahren finden Sie im folgenden Kapitel.

11. Data Mining Verfahren im Vergleich

Data Mining ist eine junge Disziplin und wird permanent mit neuen Verfahren, Wirksamkeiten und Anwendbarkeiten für den Anwender verbessert. Daher kann dieser Vergleich lediglich eine zeitlich beschränkte Aufnahme darstellen. Er konzentriert sich auf die mögliche Wirkung des Verfahrens in der Praxis.

Nach dieser Tabelle betrachten wir die Verfahren im einzelnen.

Vergleichstabelle der Data Mining Verfahren

Verfahren	Zeitdauer	Flexibilität	Umfang	Power	Input	Integration	Praxis-Faktor
Cluster	2	2	3	2	2	3	3
K-Means	2	4	3	2	2	3	3
Expectation-Maximization	2	3	3	2	2	3	2
Kohonen	2	3	3	2	2	3	2
Entscheidungsbaum	1	1	1	1	2	1	1
Kass	1	1	1	1	1	1	1
CHAID/XAID	1	1	1	1	1	1	1
Entropie	1	1	1	1	2	2	2
C4.5/5.0	1	1	1	1	2	1	1
ID3	1	1	1	1	1	1	1
Neuronale Netzwerke	3	3	2	3	1	2	2
Multi-Layered Perceptron	2	1	2	3	1	2	2
Radial Basis Function	3	2	2	2	1	2	2
Lineare Regression	2	2	2	2	1	2	3
Logistische Regression	2	2	2	2	1	2	3
Probablistic	3	2	2	2	1	2	2
Regelbasierte Systeme	5	5	3	1	1	2	4
Associated Rules	5	5	4	1	1	1	4
Basket Analysis	5	5	3	1	1	2	5
Künstliche Intelligenz	5	5	3	1	1	3	3

Die Spalten dieser Tabelle haben folgende Bedeutung:-

Verfahren

Die Verfahren sind hier in dunkelrote Farbe nach Gruppierungen eingeordnet. Das jeweilige Verfahren ist in grauer Farbe gesondert aufgeführt.

Zeitdauer

Mit der Zeitdauer ist sowohl die Dauer des Vorganges als auch die Überschaubarkeit des zeitlichen Aufwandes gekennzeichnet. Werte reichen von 1 – 6 entsprechend den Schulnoten.

Flexibilität

Die Flexibilität des Verfahrens zeigt die Anpassungsfähigkeit des Verfahrens an verschiedene Anforderungen des Bedieners an. Ein starres System wird für eine Anwendung ohne Konfigurationsmöglichkeit oder sonstige Adaption des Bedieners mit 6 gewertet. Werte reichen von 1 – 6 entsprechend den Schulnoten.

Umfang

Mit dem Umfang stellen wir die Anwendungsmöglichkeiten nach verschiedensten Szenarien dar. Je höher der Wert, je höher die Anwendungsvielfalt. Werte reichen von 1 – 6 entsprechend den Schulnoten.

Power

Mit der Spalte „Power“ möchten wir die Leistungskraft des Verfahrens zum Ausdruck bringen, die nach der möglichen maximalen Größe der Datenmenge gemessen wird. Werte reichen von 1 – 6 entsprechend den Schulnoten.

Input

Als Input verstehen wir die Komplexität der Datenmengen, die das System verarbeiten kann. Solange die Datenmengen strukturiert sind, können zum Beispiel Entscheidungsbäume alle möglichen Formate verarbeiten. Manche neuronale Systeme können sogar unstrukturierte Datenmengen verarbeiten und in Strukturen bringen, Werte reichen von 1 – 6 entsprechend den Schulnoten.

Integration

Nicht jedes Verfahren eignet sich zur Zusammenarbeit mit anderen Methoden. Werte reichen von 1 – 6 entsprechend den Schulnoten.

Praxis-Faktor

Aus dem Praxis-Faktor lesen wir die praktische Einsatzstufe des Verfahrens, auch unter Berücksichtigung der Hardware-Anforderungen sowie alle vorangegangenen Werte. Werte reichen von 1 – 6 entsprechend den Schulnoten.

Im Stile eines Data Miners, der seine Daten gerne kategorisiert, haben wir uns erlaubt, alle Data Mining Verfahren in vier Kategorien einzuteilen. In der

Vergleichstabelle sind sie in dunkelroter Farbe gekennzeichnet und werden neben den dazugehörigen einzelnen Verfahren ebenfalls bewertet.

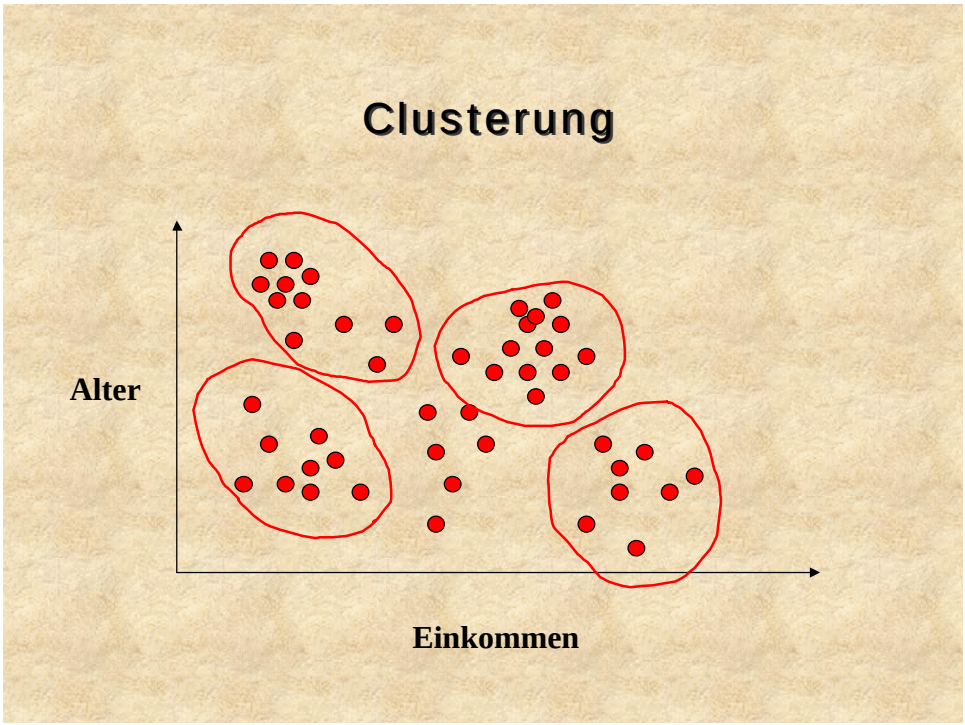
Hierbei haben wir uns bemüht, ein passendes Einsatzszenario für die jeweiligen Verfahren aufzubauen. Leider herrscht bei den einzelnen Verfahren manchmal keine deutliche Klarheit, denn die entstehenden Überlappungen führen oft zu Verwirrungen. Auch hier versuchen wir die Verwirrung deutlich aufzuklären, obwohl es nicht immer gelingen kann.

Cluster

Jedes Cluster-Verfahren gruppiert die dazugehörigen Bestandteile so gut wie möglich nach den Anforderungen der jeweiligen Konfiguration. 5 Cluster heißt 5 Gruppen. Alle Mitglieder der Datenmenge sollen nun in diese 5 Räume eingeteilt werden.

Die Zugehörigkeit zu einem bestimmten Cluster ist ein Kriterium und die Genauigkeit dieser Zugehörigkeit ein weiteres Merkmal.

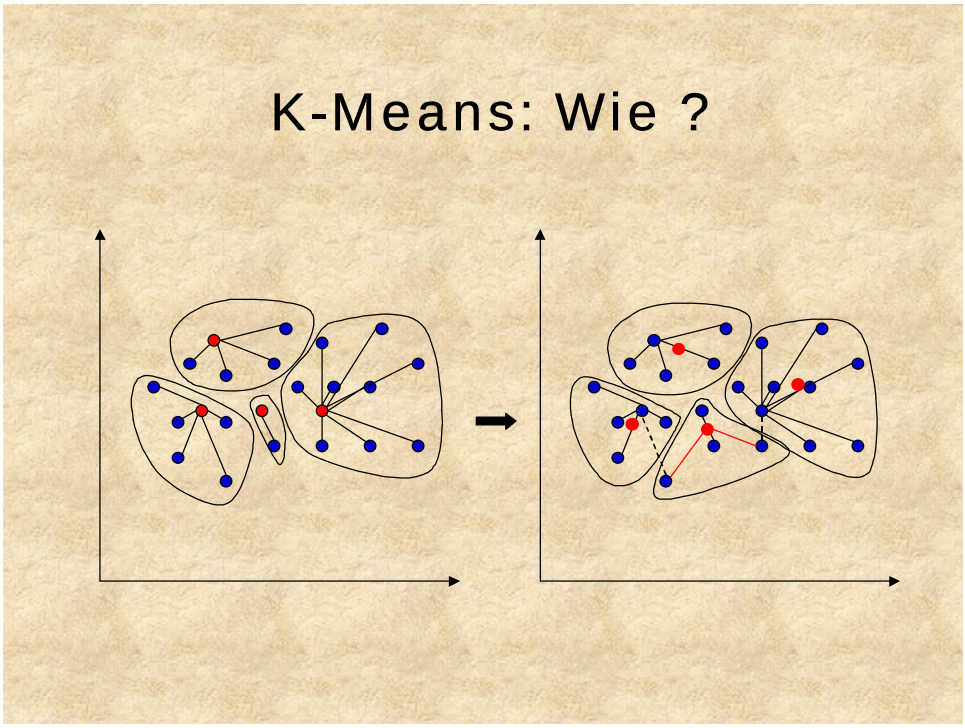
Einfache Cluster-Verfahren entstehen durch eine offensichtliche Gruppierung. Wenn Sie alle Werte mit X- und Y-Achse in eine Grafik als Punkt eintragen, werden diverse dicht besiedelte Bereiche entstehen. Die betrachten wir als Cluster.



Die Natur liefert ein schönes Beispiel einer 3-D Grafik – eine Weintraube kommt nicht alleine, sondern hängt an einer Art Cluster.

Wenn wir nun beginnen, dieses Thema wieder mathematischer zu betrachten, beschreiben einfache Cluster-Verfahren alle Punkte nach zwei Dimensionen wie eine einfache 2-D Liniengrafik aller Punkte (x_1, x_2).

Eins der bekanntesten und meist eingesetzten Algorithmen der Cluster-Gattung ist der **K-Means** von J.B. MacQueen aus dem Jahr 1967.



Dieses mathematische Verfahren teilt eine Anzahl von X Datensätzen in einer festzulegenden Anzahl Gruppen (K) ein. Die Genauigkeit jeder Zugehörigkeit zu einer Gruppe wird anschließend errechnet und eventuell geändert.

Die Anzahl geforderte Cluster ist das **K** in K-Means. Nach Festlegung dieser Anzahl wird eine vorläufige Schätzung der Mittelpunkte dieser Gruppen (Seeds) vorgenommen. Jeder Datensatz einer bestimmten Datenmenge wird dann seinem nächstgelegenen Schätzwert zugeordnet. Diese neuen Cluster haben eigene Mittelpunkte (centroid), die nun durchgerechnet werden. Diese Mittelpunkte oder Querschnitte (englisch: **Means**) gilt es festzumachen, bis jeder Datensatz passend eingeteilt ist. Dieser Vorgang wiederholt sich, jedoch bewegen sich die Grenzen der Cluster und nicht die Mittelpunkte.

Soviel zum Vorgang selber. Wozu soll er gut sein? Stellen Sie sich eine große Datenmenge vor, die aus lauter Kunden besteht. Obwohl es Kundensegmentierungen aus der Historie gibt, möchte man sehen, wie sich die Kunden in neue, unbedarfte Segmente einteilen lassen. Ob es Unterschiede zur üblichen Konvention des Hauses oder der Branche gibt.

Es sollen aus den Fakten der Datenmenge eine möglichst unvoreingenommene und nur durch die eigene Datenmenge getriebene Gruppierung erfolgen. Hierfür sind Cluster-Verfahren geeignet.

Solange man K-Means nicht mit zu vielen Variablen (oder gar zu wenig) belastet, entstehen hier schnelle und wunderbare Ergebnisse, die jederzeit an andere Mittel wie Entscheidungsbäume oder neuronale Netzwerke weiter gereicht werden können. K-Means kann problematisch wirken. Denn die notwendigen Prüfverfahren vor dem Beginn einer n-dimensionalen Berechnung (die ist ja komplex) führen oft dazu, dass das Verfahren nicht beginnt. Dies liegt meist an der inhaltlichen Überlappung einiger Felder oder der Tatsache, dass zu viele Variablen zur Untersuchung genannt werden. Hier können die vom Algorithmus erwartete Detailarbeit zur Frustration führen.

Expectation-Maximization ist sich weniger umständlich und zeigt sich gar robust in der Praxis.

Expectation-Maximization (EM), wie K-means, ist ein Algorithmus zur Clusterung einer Datenmenge in eine vordefinierte Anzahl Cluster (oder Gruppen) auf der Basis der Ähnlichkeit der Datensätze.

Anders als K-means basiert die Ähnlichkeit im EM Algorithmus auf der Theorie der Wahrscheinlichkeit. Ein Datensatz wird einem bestimmten Cluster dann zugewiesen, wenn er am wahrscheinlichsten durch die Wahrscheinlichkeitsverteilung des entsprechenden Clusters generiert würde – hierbei sind die Verteilungen pro Cluster unterschiedlich. Zum Beispiel, wenn Datensätze Kunden entsprechen, könnten unterschiedliche Verteilungen unterschiedliche Verhaltensarten bei den Kunden bedeuten.

Der Algorithmus beginnt durch die Wahl K unterschiedlicher Verteilungen. Jede Verteilung wird dann durch eine Anzahl Parameter beschrieben. Zusätzlich wird jeder Verteilung eine Gewichtung zugewiesen. Danach wird eine vorläufige Schätzung für alle Parameter vorgenommen, einschließlich Gewichtungen. Auf der Basis der Parameterwerte kann die Wahrscheinlichkeit der Betrachtung der Training Datenmenge errechnet werden. Der EM Algorithmus findet Parameterwerte, die diese Wahrscheinlichkeit maximiert. Bei jeder Iteration werden neue Parameterwerte unter Verwendung der alten Werte zur Maximierung der Zugehörigkeitswahrscheinlichkeit errechnet. Dieser Vorgang wird solange wiederholt, bis sich die Parameterwerte einen Spitzenwert erreichen, wo die Funktion der Wahrscheinlichkeit maximiert wird.

Schließlich wird beim Scoring eines neuen Datensatzes die Wahrscheinlichkeit, dass jener Datensatz von jeder K Verteilung generiert wird, unter Verwendung der endgültigen Parameterwerte und Bayes Formel der konditionalen Wahrscheinlichkeiten errechnet. Der Datensatz wird dann dem Cluster mit dieser maximalen Wahrscheinlichkeit zugewiesen.

Entscheidungsbaum

Sofern man über genügend Einblicke in die Materie verfügt, empfiehlt sich der Anfang einer Analyse meist mit einem Entscheidungsbaum. Wohl gemerkt ein „Zugang“ in Form einer möglichen abhängigen Variable muß bestehen.

Wenn Sie sich mit einer Datenmenge konfrontiert sehen, in der mehrere Felder sich als Zugang (abhängige Variable) eignen, sollte der erste Schritt ein erneutes Profiling oder gar ein Clusterverfahren werden. Wenn Sie allerdings eine spezielle Information untersuchen wollen, die in einem Feld vorhanden ist – Produktion erfolgreich ja/nein, Kundentyp, Markenwechsler ja/nein, diverse Perioden, Produktart, Marktsegment usw. -, können Sie direkt dieses Feld zu Ihrer abhängigen Variablen erklären und sofort schauen was Ihnen der Entscheidungsbaum bietet.

Es liegt in der Natur dieser Modelle, dass Sie zuerst einige Dinge entdecken, die Ihnen bereits bekannt sind. Zum Beispiel weiß man, dass ältere Menschen eher unter höherem Blutdruck leiden. Das Alter, als Faktor zur Erklärung der Umstände von hohem Blutdruck, ist wichtig, aber weder überraschend noch reicht das Alter allein als Erklärung aus. Warum der Blutdruck bei vielen älteren Menschen hoch ist, folgt als Verkettung von weiteren Umständen, die unterhalb der Rubrik Alter = Hoch erschienen. Die ersten interessanten Erkenntnisse tauchen meist in der Baumebene 3 bis 5 auf.

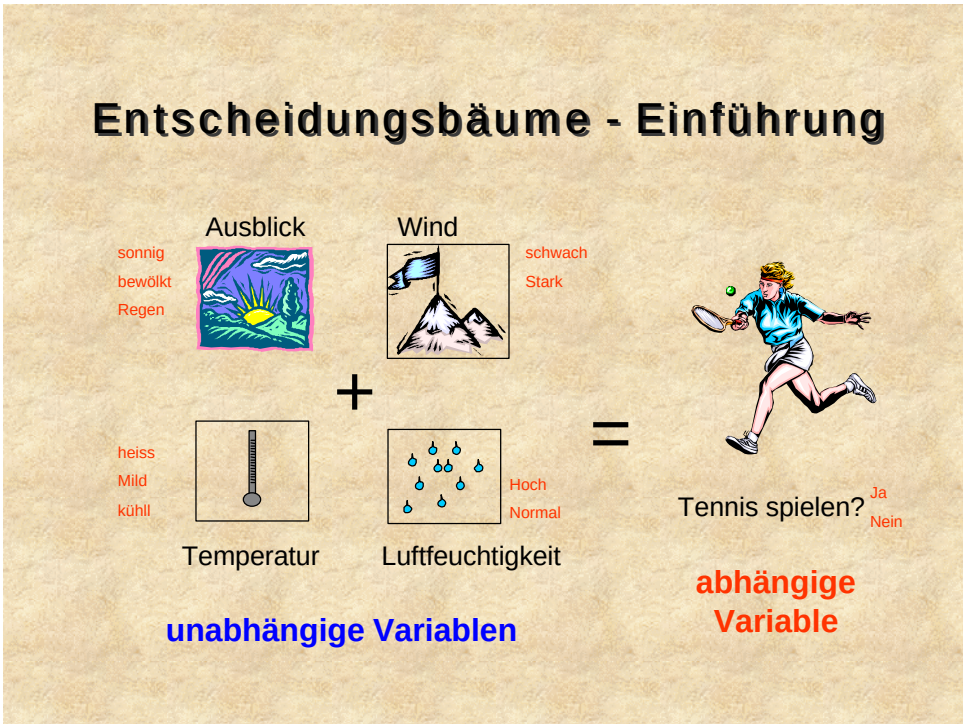
Welche Algorithmen gibt es bei Entscheidungsbäumen und welche darf ich am besten verwenden? Und vor allem was ist überhaupt ein Entscheidungsbaum? Diese Frage möchten wir vorerst erläutern, bevor wir uns den anderen Fragen zuwenden.

„Ein Entscheidungsbaum ist ein mächtiges Modell, das durch eine Klasse von Techniken erzeugt wird, zu der Klassifizierungs- und Regressionsbäume (CART) sowie Chi-Quadrat automatische Induktion (CHAID) gehören“ (Berry und Linoff – Data Mining Techniques 1997). Eine ausreichende Anzahl von Algorithmen sowie intelligente Defaultwerte mit Änderungsmöglichkeiten für den Anwender stehen zur Verfügung.

Wir wollen mit dem „Auto“ Entscheidungsbaum nun fahren und uns mit einer Neukonstruktion nicht aufhalten. Wichtig zu wissen, ist Ihre Vorgehensweise bzw. Möglichkeiten (Siehe auch Kapitel 13. Data Mining und herkömmliche Mittel im Vergleich).

Bei Entscheidungsbäumen brauchen Sie keine Fragen formulieren oder artikulieren – wie bei der Abfrage von Datenbanken -, um Antworten aus Ihren Datenbeständen zu erhalten. Starre Fragen werden durch flexible und eher dynamische Algorithmen ersetzt, die von der Data Mining Software eingeschaltet werden. Es werden danach Antworten produziert, die im Zusammenhang zueinander stehen. Es werden quasi Ketten von Ergebnissen gebildet, die als Gesamtheit eine Erläuterung darstellen. Wir als Anwender bleiben frei und unbelastet, uns auf die Interpretation und Beurteilung dieser Antworten zu konzentrieren. Somit entsteht ein völlig anderes Arbeitsgefühl sowie eine schnelle, gezielte sowie Ergebnis orientierte Arbeitsweise.

Im folgenden einfachen Beispiel möchten wir die Entscheidung Tennis zu spielen, anhand der diversen Wetterindikatoren herbeiführen. Die abhängige Variable ist also „Tennis spielen“ mit den kategorischen Werten „Ja“ oder „Nein.“



Die vier Faktoren Wetterausblick, Windstärke, Temperatur und die Luftfeuchtigkeit sollen in den diversen Ausprägungen die Entscheidung treffen helfen, ob Tennis gespielt werden soll oder nicht.

Sehen Sie hierzu das obige Bild, wo die Entscheidungsfindung bereits die strukturierte Form eines potentiellen Regelwerkes anzunehmen beginnt.

Jetzt muß man noch die diversen Möglichkeiten durchspielen, um zu einem Ergebnis bzw. zu einer vorbereiteten Entscheidung zu kommen. Alle Möglichkeiten werden

durchgerechnet. Mit relativ wenigen Kriterien und einer geringen Anzahl Datensätze (in diesem Fall sind es die letzten Tage, wo man Tennis gespielt oder eben nicht gespielt hat) kann man die möglichen Variationen fast im Kopf, auf einem Zettel oder schlimmstenfalls mit Hilfe eines kleinen Abfrage Programms durchspielen.

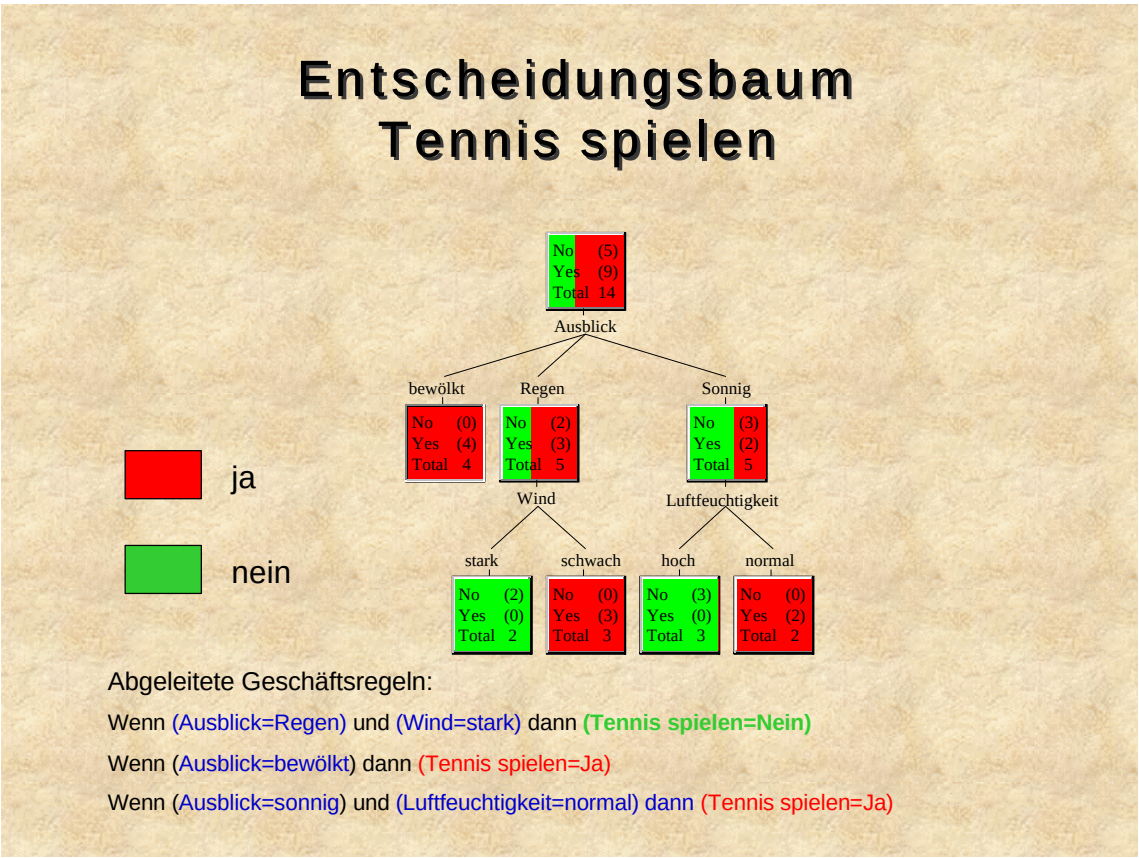
Bei großen Datenmengen und bei einer kaum zu überblickenden Anzahl von unabhängigen Variablen jeweils mit einer entsprechenden Anzahl Kategorien wird die Aufgabe deutlich komplexer und erreicht ein Status der Unmöglichkeit. Zumindest wird ein Stadium erreicht, wo die Aufgabe nicht mehr praktikabel scheint und unüberschaubar wird. Hier hilft der Entscheidungsbaum auch bei unserem Tennis Beispiel wie im folgenden Bild.

Als wichtigstes Kriterium wird der Wetterausblick in seinen drei vorhandenen Ausprägungen herangezogen. Bereits hier fällt eine deutliche Entscheidung fürs Tennis spielen, wenn der Ausblick bewölkt ist. Im Falle von Sonne oder Regen ist die Entscheidung noch nicht eindeutig. Sehen Sie hierzu das folgende Bild.

Bei Regen wird die Windstärke als weiteres Kriterium herangezogen. Ist der Wind stark fällt die Entscheidung eindeutig gegen das Tennisspiel aus. Ist der Wind jedoch schwach fällt die Entscheidung genauso eindeutig für das Tennisspiel aus.

Beim sonnigen Wetterausblick wird die Luftfeuchtigkeit als Kriterium zusätzlich herangezogen. Im Falle einer hohen Luftfeuchtigkeit fällt die Entscheidung gegen und bei einer normalen Luftfeuchtigkeit für das Tennisspiel aus.

Dieses klare zusammenhängende Bild der Variationen gibt uns insgesamt drei deutliche Fälle, wann das Tennisspiel als positive Entscheidung getroffen wird und zwei Fälle die dagegen sprechen.



Abgeleitete Regeln:

Wenn (Ausblick=Regen) und (Wind=stark) dann (Tennis spielen=Nein)
Wenn (Ausblick=Sonnig) und (Luftfeuchtigkeit=hoch) dann (Tennis spielen=Nein)

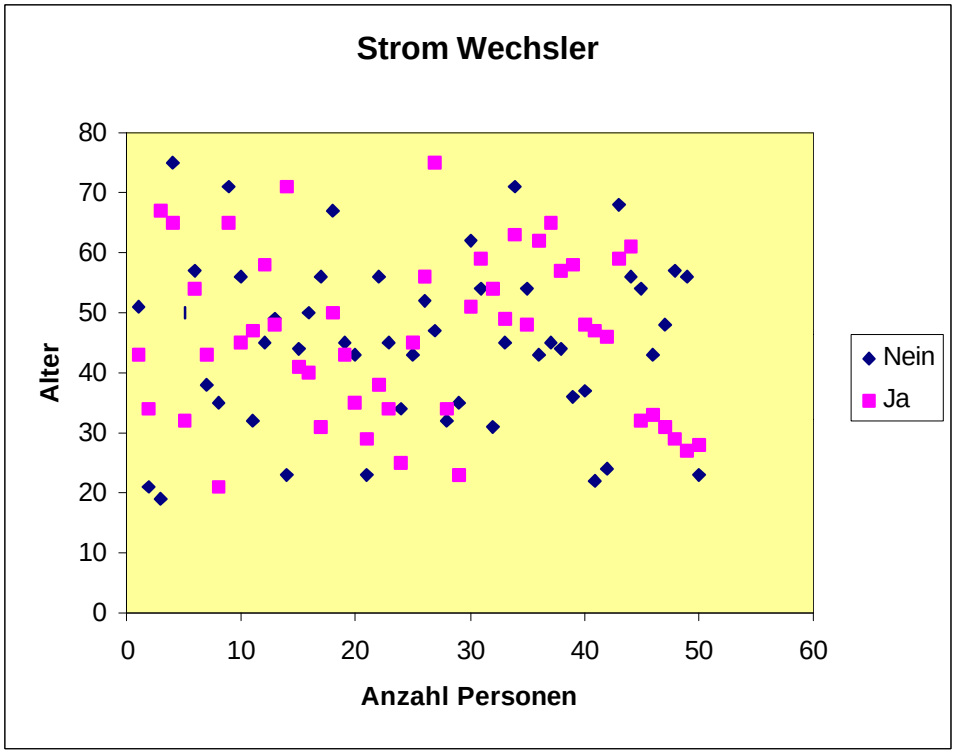
Wenn (Ausblick=bewölkt) dann (Tennis spielen=Ja)
Wenn (Ausblick=sonnig) und (Luftfeuchtigkeit=normal) dann (Tennis spielen=Ja)
Wenn (Ausblick=Regen) und (Wind=schwach) dann (Tennis spielen=Ja)

Auch in einem echten Geschäftsfall lassen sich Geschäftsregeln aus einem Wust von Daten über den Weg des Entscheidungsbaumes ableiten. Diese Regeln können dann ohne weiteres in bestehende Datenbank oder OLAP Applikationen eingegliedert werden. Ohne einen Entscheidungsbaum würde der Weg zu dieser Abfrageformulierung oder Regelwerk unüberblickbar sein und daher wahrscheinlich nie in Auftrag gegeben werden.

Neuronale Netzwerke

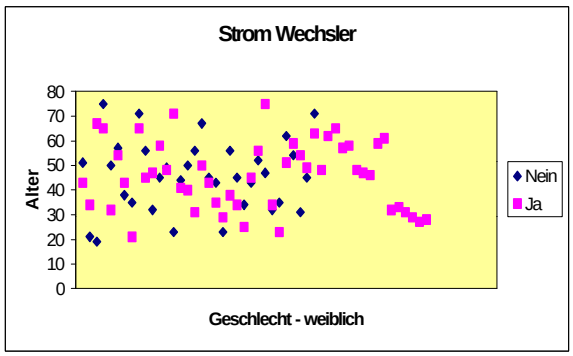
Diese Werkzeuge bilden den vielleicht größte Erklärung für die Mystik, die sich bei außenstehenden Beobachtern oft um Data Mining rankt. Diese „Black Box“ Werkzeuge basieren jedoch auf Vorgehensweisen und Algorithmen aus der Mathematik und Statistik, die wir an dieser Stelle aufbieten möchten.

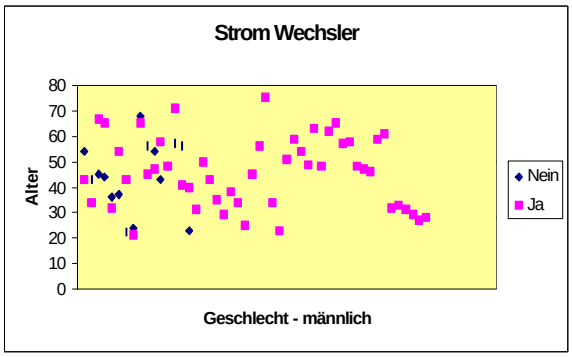
Als einfache Erklärung möchten wir die folgende einfache Grafik eines Scatter Charts betrachten. Es sind Kunden befragt worden, ob Sie den ehemaligen Monopolisten der Stromwirtschaft die Treue halten oder auf die verlockenden Angebote des „farbigen“ und billigeren Stroms einzugehen gedenken.



Die Ja-Antworten und die Nein-Antworten sind anhand der Achsen Alter und Anzahl der Befragten umrahmt. Die Aufgabe des neuronalen Netzwerkes ist es eine „Schichtung“ oder eine Gruppierung innerhalb dieses Bildes zu erzeugen, die möglichst jeden vorkommenden Fall „abdecken“ oder in der Praxis vorhersagen kann. Das heisst, es soll anhand der Angaben Alter, Anzahl der Befragten und Geschlecht (lauert im Hintergrund) vorhergesagt werden, ob die Antwort auf die Frage der Wechselwilligkeit eher „Ja“ oder „Nein“ ausfällt.

Die folgenden Bilder zeigen die gleiche Grafik nach Geschlecht getrennt – links weiblich und rechts männlich.



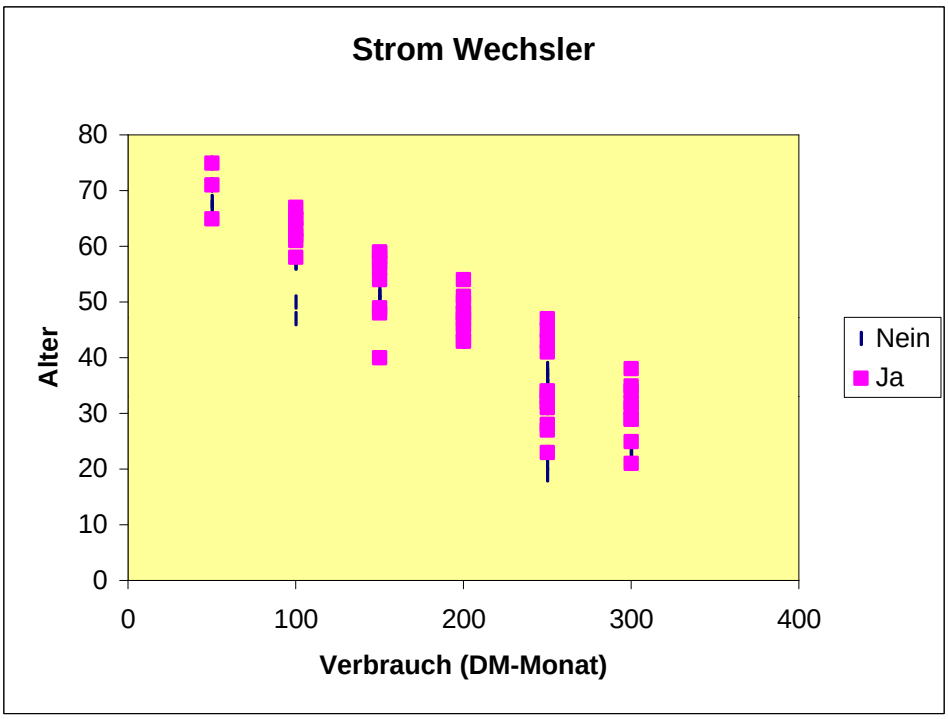


Auffällig ist die geringe Anzahl der dunklen Punkte beim männlichen Geschlecht. Dies würde aussagen, dass wenige Männer „Nein“ zum Wechsel sagen. Im Umkehrschluß ist die Neigung den Stromlieferanten zu wechseln bei den Männern hoch. Bei den Damen ist die Wechselbereitschaft deutlich geringer.

Weitere Interpretationen sind eigentlich überflüssig, denn dies sind Spieldaten, jedoch ist das Bild der Antworten stark verteilt, ob männlich oder weiblich, ob Ja oder Nein.

Um nun eine Gruppierung in Richtung Typierung (Vorhersage) machen zu können, werden weitere Indikatoren benötigt. Hierfür werden wir die Komponente Verbrauch hinzufügen.

Die Komponente Verbrauch haben wir der Einfachheit halber in D-Mark pro Monat ausgedrückt. Hier spiegeln sich einige Verhaltensweisen der unterschiedlichen Altersgruppen wieder. Zum Beispiel werden jüngere Personen eher viel Strom benötigen, weil sie Kinder haben, also mehrere Personen im Haushalt haben. Ältere Personen hingegen wenige Personen gar alleine. Ferner werden jüngere Personen eher viel Kochen, Sonnenlampen, HiFi Stereo Systeme, Computeranlagen, Heizlüfter, Sprudelbäder und sonstige stromfressende Spielzeuge der modernen Welt einsetzen als ältere Personen.



Durch die Eingabe des weiteren Indikationsfaktors Verbrauch (DM-Monat) bietet die Grafik bereits ein relativ deutliches Bild. Wir sehen ganz deutlich wie das durchschnittliche Alter je höher der Verbrauch sinkt. Gleichzeitig sehen wir innerhalb der 6 unterschiedlichen Verbrauchskategorien die loyalen Kunden eher jünger sind. Soll dies bedeuten, dass sich die älteren Kunden – unabhängig von Ihrer Verbauchskategorie – eher durch die bisherigen Monopolisten verprellt fühlen oder sind sie einfach preis sensibler? Zumindest können wir behaupten, je höher das Alter je geringer der
Verbrauch und pro Kategorie werden eher die Älteren wechseln wollen. Wechsel bedeutet in diesem Beispiel schließlich Ausgaben senken.

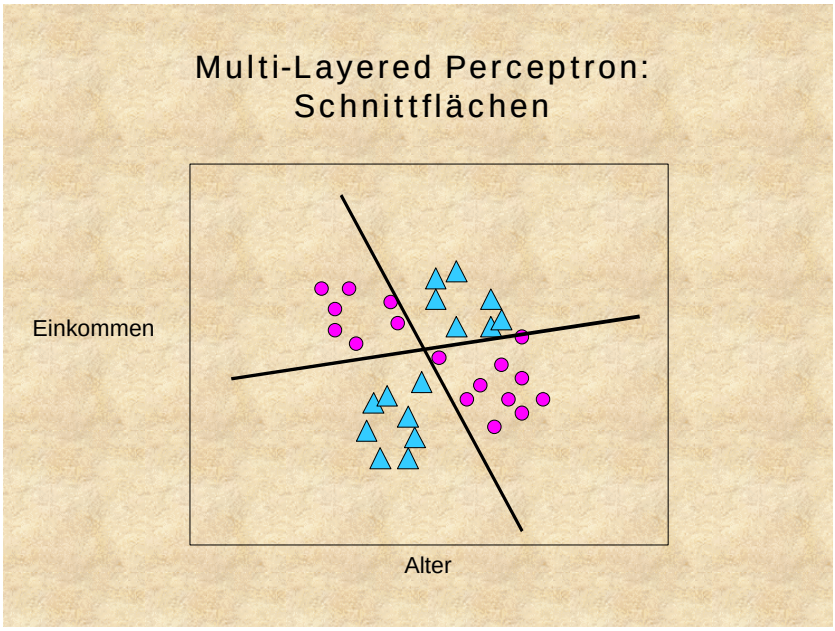
Neuronale Netzwerke können solche Gebilde, die wir zu Beispielszwecken mit wenigen Indikatoren, mit weit mehr Faktoren über eine größere Anzahl von Datensätzen aufbauen. Diese Vorhersagen können mit den tatsächlichen Werten verglichen und damit die Qualität des Modells bewertet werden. Dies nennt man **Validierung**.

Das fertige, validierte, neuronale Datenmodell kann dann anschließend dazu verwendet werden, weitere Datenmengen mit einer Vorhersage belegt zu werden. Diesen Vorgang nennt man **Scoring**.

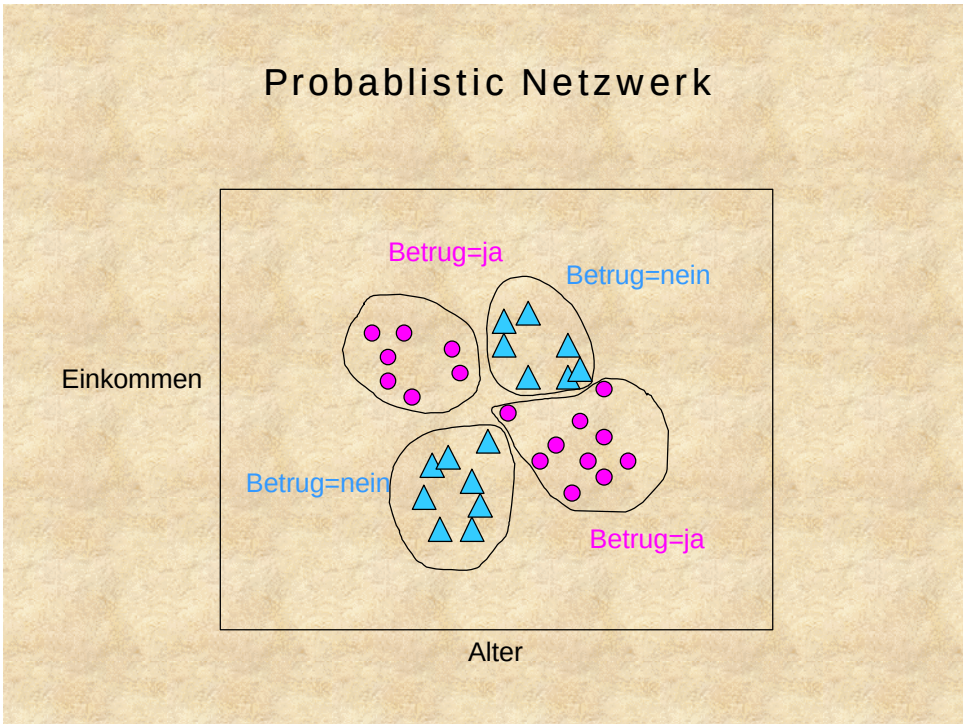
Einfluß auf die genaue Vorgehensweise eines bestimmten neuronalen Netzwerkes hat höchsten am Rande der Spezialist. In der Praxis soll man eine einfache Matrix zu Hilfe ziehen, die auf der Basis der Aufgabe oder Beschaffenheit der Datenmenge anzeigt, welcher Typus von neuronalem Netzwerk sich jeweils am besten für welche Aufgabe eignet.

	Multi-Layered Perceptron	**Probablistic**	**Radial Basis Function**
Geschwindigkeit beim Training	**OK**	**Super schnell**	**Sehr langsam bis langsam**
Geschwindigkeit beim Scoring	**OK**	**Langsam**	**Super schnell**
Genauigkeit	**OK**	**Hoch bei kategorischer AV Niedrig bei fortlaufender AV**	**Weniger bei kategorischer AV Hoch bei fortlaufender AV**
Größe der Datenmenge beim Training	**Durchschnitt**	**Große Mengen**	**Nur kleine Datenmengen**
Größe der Datenmenge beim Scoring	**Durchschnitt**	**Langsam**	**Schnell auch bei großen Mengen**

Hieraus geht hervor, dass der Multi-Layered Perceptron als zuverlässiges Arbeitstier ohne besondere Leistung zu betrachten ist. Er versucht seine Datenpositionen mit „geraden Schnitten“ zu tranchieren. Daher wird er immer zu einem erträglichen Ergebnis kommen.

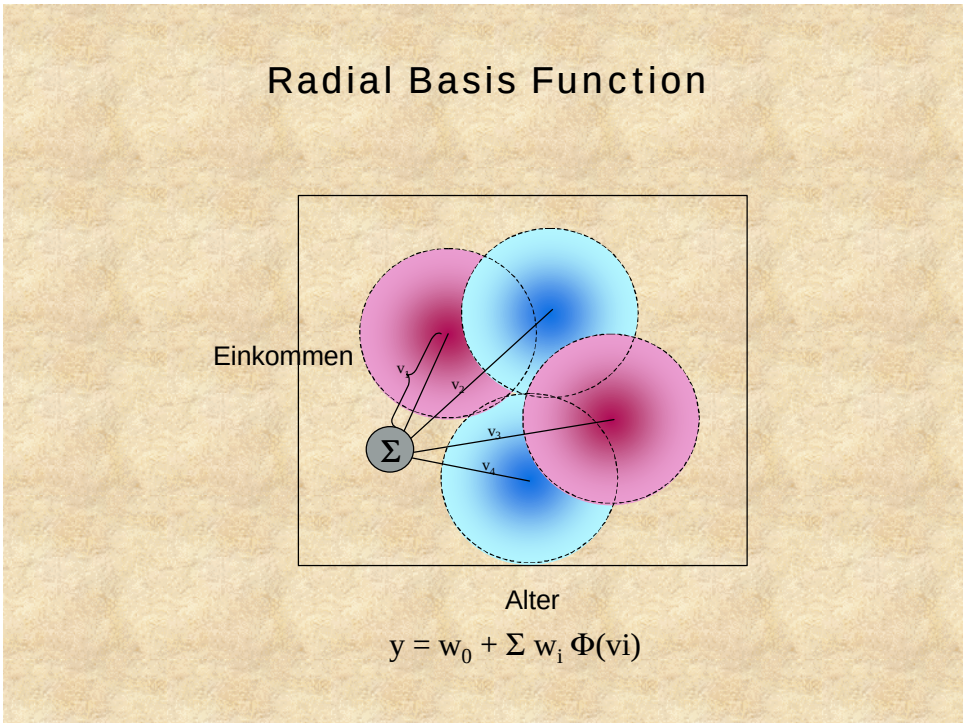


Ist die abhängige Variable kategorisch, wird das Probablistic Netzwerk versuchen „Gebiete“ abzustecken. Leider bei fortlaufender, abhängiger Variablen wird dieser Typus immer versagen, jedoch bei kategorischen, abhängigen Variablen werden auch bei stark gemischten und fehlerhaften Datenmengen gute Ergebnisse erzielt. Daher eignet sich dieser Typus gut zur Auswertung von Umfragedaten, da die Einflüsse in solchen Mengen oft weit verteilt und deutlich emotioneller und qualitativer Natur sind. Leistet schnelle Arbeit im Training jedoch langsam beim Scoring.



Als Kompromiß der beiden oberen Typen gibt es das Radial Basis Function Netzwerk. Dieser Typus versucht eine Summierung der Abstände der jeweiligen Datensätze (oder Punkte wie in unseren obigen Grafiken) von der Mitte der Gruppen wie im nächsten Bild und für die exakten Mitmenschen samt Formel zu berechnen.

Bildet also eine deutlich nicht-lineare Regression. Bietet daher Engpässe bei der Größe der Datenmenge im Training der Modelle. Beim Scoring legt er aber los. Gerade bei fortlaufenden, abhängigen Variablen liefert er eine große Genauigkeit.



Diese drei Formen von neuronalem Netzwerk möchten wir in den bisherigen grafischen Darstellungen mit einem tabellarischen Überblick über die jeweiligen Vor- und Nachteile so belassen. Die üblichen Darstellungen der Neuronen im Hirn eines Menschen oder als mehrstufiger Flußplan zur Verfolgung der verschiedenen Arbeitsschritte bei der Berechnung eines neuronalen Netzwerkes möchten wir wiederum allen Spezialisten überlassen.

Wir werden uns hier weiterhin auf die Data Mining Praxis konzentrieren und begnügen uns daher mit dieser Anleitung zur praktischen Anwendung dieser Werkzeuge mit einem kurzen grafischen Überblick über die Verfahrensweise des einzelnen Tools.

Regelbasierte Systeme

Könnte man einen Roboter mit einer riesigen Batterie voller Regeln für jede erdenkliche Situation ausrüsten, wäre er bestimmt für viele Situationen gut gerüstet. Beobachtet man die Entwicklung der Zeit vor FUZZY ist dies gerade bei solchen Maschinen die Art der Steuerung.

Leider müssen wir uns fragen, was macht der Roboter, wenn eine unbekannte Situation auftaucht. FUZZY erweitert seine Möglichkeiten durch eine gewisse Flexibilität.

Trotz allem fehlt eine größere Dynamik. Hat er die oben erläuterten Möglichkeiten zur Verfügung, dann kann er vielleicht eine Ähnlichkeit finden, ableiten oder mit einer gewissen Genauigkeit prognostizieren. Verfügt er jedoch lediglich über die besagte Batterie von Regeln, befindet er sich in einer statischen Lage. Er kann dann höchstens wie diverse, primitive Roboter aus der Filmindustrie mit „Unbekannter Vorgang!“ oder „Does not compute!“ oder gar mit einer Fehlermeldung „reagieren.“

Das ist genau der Krux der Sache von regelbasierten Systemen. Sie sind vielleicht schnell, jedoch eine Dynamik wie bei algorithmisch gesteuerten Systemen fehlt. Man muß abwägen, welches System für die genannte Aufgabe am besten paßt.

Trotzdem muß man die Frage bei regelbasierten Systemen stellen, wo die Regeln herkommen. Eine Prüfung sowie ein ständiger Ausbau des Regelwerkes kann klar mit den Mitteln des Data Minings präzise und schnell erfolgen.

Die Herkunft vieler Regeln wird oft mit „Erfahrung“ erklärt. In wie weit diese Erfahrung immer noch präzise zutrifft, kann zum Beispiel ein Entscheidungsbaum schnell testen. Man kann die Verzweigungen manuell nach der Vorgabe des vermeintlichen „Regelwerkes“ Schritt für Schritt nachstellen und die Erfolgswerte beobachten und/oder das System suchen lassen und die gefunden Variablen sowie Erfolgswerte gegeneinander vergleichen gar mit einem sogenannten Lift Chart vergleichen lassen.

Bei sich ständig ändernden Werten – wie Marktverhalten, Betrugsverhalten oder andere Situationen, wo das menschliche Verhalten die entscheidende Rolle spielt – muß eine Dauerbeobachtung der entscheidenden Einflüsse erfolgen. Nur somit wird sichergestellt, dass die Regeln im regelbasierten System aktuell und zutreffend bleiben.

Diese Dauerbeobachtung bietet Data Mining. Eine Kombination der beiden Disziplinen sorgt für eine sichere Sache.

12. Hauptprojekt

Der Übergang zum Hauptprojekt birgt in sich einige besondere Fallen.

Zuerst bringt der Begriff Hauptprojekt einige konzernpolitische Probleme mit sich. Daher muß das Hauptprojekt vielleicht vorerst Pilotprojekt oder Testphase genannt werden. Ansonsten wird eine Lawine von Folgeerscheinungen im Konzern ausgelöst, die dem Projekt im Wege stehen (Kontrollorgane, Hauptausschüsse müssen gehört werden) bzw. dem Start verhindern können (andere Entscheidungswege). Alleine der Begriff „Hauptprojekt“ erzeugt diese Schwierigkeiten, die rein organisationspolitischer Natur sind.

Desweiteren muß spätestens an dieser Stelle darüber nachgedacht werden, wie Data Mining in den praktischen Einsatz des Konzerns gebracht werden soll. Werden alle Stellen mit einem Produkt ausgestattet (das wäre ein Extrempol) oder werden die in Frage kommenden Einsatzbereiche im Endeffekt mit einer auf Data Mining basierende Anwender Applikation beglückt , die von Experten entwickelt wurde und vordergründig als Data Mining Werkzeug gar nicht erkennbar ist.

Irgendwo zwischen diesen beiden Extrempolen liegt die Wahrheit. Wie die individuelle Lösung dieser Frage aussieht, hängt von einer Menge Parameter ab, die jedem Konzern oder Institution besonders eigen sein wird. Hierfür gibt es keinen Optimum, außer man wägt die Realisierungsnotwendigkeit mit den Begebenheiten des Konzerns ab.

Wie sind die Entscheidungswege und wie stark sind die Neigungen zu Wildwuchs; denn ein Vollzeit Data Miner wird mehr und höherwertigere Erkenntnisse auf Dauer gewinnen als ein Gelegenheitsminer.

Wie die Entscheidung ausfällt, wird bei jedem Anwendungsgebiet eine andere Variante bilden, jedoch gibt es gar keine Entscheidung, kann man sich die Mühe eigentlich komplett sparen, Data Mining zu betreiben. Nur der Einsatz mit den Ergebnissen wird einen Vorteil bringen (Siehe Data Mining Projekt Rad).

Nichtsdestotrotz gibt es Produkte, die dem Business User weiterhelfen können, ohne dass die Produktivität oder gar die Effizienz für das Unternehmen leiden muß. ANGOSS KnowledgeSEEKER ist beispielsweise leicht zu erlernen und bringt dem Anwender leicht und schnell zu verwertbaren Ergebnissen. Es sind wiederholt innerhalb eines Tages „durchschnittlich begabte Polizeibeamte“(Zitat: LKA NRW 1996) erfolgreich eingewiesen worden.

Ferner können Data Mining getriebene Anwender-Applikationen hier schnelle Abhilfe schaffen. Spezialisten erstellen und pflegen Datenmodelle, die im Hintergrund einer Applikation – für den Anwender unsichtbar – verankert werden. Der Anwender

erzeugt (beschafft, importiert, erhält usw.) Werte, die er per Tastendruck bearbeiten läßt. Daraufhin erhält er eine einfach zu verstehende Antwort. Von Data Mining weiß dieser Anwender womöglich gar nichts.

Teilweise spielt die Branche bzw. der Unternehmensbereich sowie die Aufgabe hier eine Rolle. Nicht jede Aufgabe ist automatisierbar. Manche jedoch sehr. Ein Alarmsystem in der Produktion kann gut automatisiert werden, die Auswertung der Daten einer Webseite sind weniger leicht automatisierbar. Die Diagnose eines Patienten, der über Bauchweh klagt ist wiederum teils automatisierbar, jedoch die Erstellung eines individuellen Kündigungsbriefes kaum.

Ein Serviceunternehmen ist häufiger am Kunden als ein Hersteller von Investitionsgütern und muß daher schneller agieren. Automatisierbare Analyseformen kommen dem Serviceleister eher zugute als lang angelegte Untersuchungen. Im letzteren Fall werden Stabsfunktionen besser die Vordenke geleistet haben.

Welche Vision, welches Szenario?

Die Schwierigkeit ist hierbei meist die Vision oder griffiger gesagt das Szenario. Aus einem Verständnis für gewisse anwendbare Szenarien ergibt sich die Plausibilität für das Unternehmen sowie die Zielvorgabe.

Um den Prozess hierzu etwas zu erleichtern, möchten wir einige bekannte Szenarien aufführen. Zum einfachen Überblick möchten wir an dieser Stelle mit der Darstellung einiger vielleicht bizarren Szenarien beginnen, die ohne Data Mining überhaupt nicht vorstellbar wären.

1. Spesenabrechnung

Zum Anfang nehmen wir ein Szenario aus dem Controlling. Das Thema der Spesenabrechnung ist für viele Betroffene ein deutlicher Punkt des Ärgernisses.

Auf der einen Seite ärgert es dem Mitarbeiter, der Spesen verursacht. Er muß eine teilweise langwierige und veraltete Bürokratie durchlaufen, um sein Geld zurückerhalten zu können, die er zum Wohle der Firma bereits ausgegeben hat.

Auf der anderen Seite ärgert es dem Controller, weil die Spesenabrechnung für ihn personal intensiv und damit teuer ist.

Dazwischen hängt vielleicht noch ein direkter Vorgesetzter, der wirklich anderes zu tun haben sollte, als sich mit den Spesenkonten seiner Mitarbeiter herumzuschlagen.

Dies muß nicht so sein.

Es wurde ein Data Mining Projekt gestartet, mit dem Ziel, ein automatisches Scoring-System für Spesenkonten aufzubauen.

Die Einteilung sollte auf Mustern aufbauen, die eine unterschiedliche Kategorie von Risiko unterlagen. Beispiel: Score 10 = Risikolos und bedenkenlos weiterzuverarbeiten. Score 1 = Höchste Risiko, Chef vorlegen, Mitarbeiter sofort in Ketten zum Gespräch vorführen lassen.

Mit Hilfe von KnowledgeSTUDIO ist die Historie solcher Daten untersucht worden. Hierbei baute man ein solches Scoring System und stellte fest, dass über 3 Viertel aller Spesenabrechnungen im Bereich des Scores 10 lagen. Diese Abrechnungen wiesen weder Formfehler (Arithmetische oder Vorschriftsfehler) noch zeugten von einer Nicht-Berechtigung (also keine deutlichen Betrugsversuche).

Obwohl diese 100% korrekten Spesenabrechnungen den höchsten Grad der Genauigkeit boten, wurden sie trotzdem der gleichen teueren, zeitaufwendigen Prozedur unterzogen, wie alle anderen Spesenabrechnungen. Folglich wurden sie auch mit der gleichen „langsamen“ Prozedur abgewickelt. Der betroffene Mitarbeiter bekam sein Geld keine Stunde schneller zurück.

Bei den übrigen 25% der Spesenabrechnungen fielen fast alle anderen in die Score Kategorien 7 bis 9. Das bedeutete, wenige eher formelle Punkte waren in Fragen zu stellen. Ein Datum fehlte, eine Querrechnung stimmte nicht, oder in der Gesamtaddition hatte man sich vertan. Für das Unternehmen war das Risiko hierbei minimal. Trotzdem wurde der alt gebackene Weg strikt eingehalten und die Abwicklung blieb langsam und schwerfällig; damit teuer.

Es blieben etwa 1% EIN PROZENT aller Abrechnungen über, die ein bedenkliches Ergebnis aufwiesen, die eine nähere Bearbeitung bedürften.

Es folgten drei Vorgehensweisen.

1. Die Frage wie könne man schnell, bequem und korrekt die Abwicklung umstellen und gleichzeitig alle Formfehler auffangen.
2. Wie sollte man infolge der Identifikation mit den wenigen Risiko reichen Fällen umgehen.
3. Es blieb die Identifikation der „schwierigen“ Fällen

An dieser Stelle möchten wir anhalten.

Unter Verwendung von KnowledgeSTUDIO ist die Grundlage einer „Kategorisierung“ ermittelt worden. Dies nennt man neudeutsch: Scoring oder Ranking. Auf einer Skala von 1 bis 10 sind unterschiedliche Grade von „Risiko“ definiert worden. Übrigens wird ein solches System von Kreditkarten / Plastikkarten Geldautomaten weltweit eingesetzt, um „entscheiden“ zu können, ob Sie Ihre geforderte Barsumme vom Geldapparat erhalten.

Es werden „WENN-Spiele“ durchgeführt, um die Auswirkungen von Zusammenlegungen ermittelter Gruppen prüfen zu können - sowie Aufsplittungen von anderen Gruppen ermitteln zu können. Die erneute Überprüfung solcher Gruppen ist ein Dauerzustand, der normalerweise einmal jährlich durchgeführt wird.

Da man Zeitdauer und letztendlich Kosten der Spesenabrechnung deutlich reduzieren wollte, fand man ein Vehikel in der modernen Kommunikation. Auf dem unternehmenseigenen Intra-/Internetsystem richtete man die Abwicklung aller Spesenabrechnungen Software gesteuert ein. Alle Belege wurden nachgereicht und per Stichprobe per Vergleich kontrolliert.

Die Online-Abrechnung beinhaltet einige „intelligente“ Abfragen, die direkt die meisten kritischen Punkte behandeln konnte. Diese Lösung wurde mit KnowledgeSTUDIO SDK erstellt und konnte somit auf verschiedene Algorithmen und Methoden des „Mutterpaketes“ zugreifen.

Ferner konnten fehlende Tagesdaten oder Fehler der Addition sofort verhindert werden. Diverse Muß-Felder sorgten für die passenden Eingaben und arithmetische Formel sorgten aus dem Hintergrund für eine korrekte Berechnung sowie Plazierung der Ergebnisse an der richtigen Stelle.

Es kamen folgende Ergebnisse heraus.

1. Die Kosten für dieses System haben sich prompt amortisiert und hatten fortan nur geringe personelle Ressourcen gebunden. Die Controller waren glücklich: Eine seltene Leistung.
2. Der Zeitverbrauch dieses Systems war gering. Somit waren Mitarbeiter und Controller glücklich. Außerdem gewann das Reporting aus diesem Bereich immens an Aktualität.
3. Die Einkäufer wurden glücklich, denn hier entstand eine aktuelle und prompte Aussage über die Belegung diverser Reise-, Hotel- und gastronomische Angebote. Die Verhandlungen mit solchen Dienstleistungsunternehmen wie Fluggesellschaften und Hotels verliefen auf einer anderen „intelligenteren“ Ebene.
4. Die „schlimmen“ Finger der Spesenabrechnung wurden schnell identifiziert und konnten meist zur Besserung bewegt, da ihre Versuche prompt transparent und damit auffällig wurden.

Bei diesem Beispiel wurde deutlich wie das Risiko des Mißbrauchs auf einer objektiven und in der Tat geringen Anteil des Ganzen reduziert werden konnte. Die Umsetzung einer neuen Abrechnungssystematik entstand als Folge der ersten Untersuchungen. Die Einteilung aller Fälle in Scores mit unterschiedlichen Risiken reduzierte die Menge der eigentlichen Arbeit auf einen Bruchteil dessen, was man eigentlich erwartet hatte und gewohnt war. Dies erleichterte die Fokussierung auf die wenigen Fälle, die Anomalien aufweisen.

Die „Angst“ alles kontrollieren zu müssen, beginnt vor diesem Hintergrund der objektiven Tatsachen durchleuchtet zu werden. Der Nebel wird transparent und die Angst wird kalkulierbar und griffig.

Warum einen großen Aufwand (Tausende von Mark) betreiben um DM 2,50 zu sparen. Der kaufmännische Sinn fehlt hier. Da hat eine Gesellschaft diese Tausender ein einziges Mal in ein System investiert und hat für immer ihre Ruhe.

Die Auswirkung auf die Personalbeziehungen ist vorhanden aber schwer zu berechnen. Jedoch dürfte die positive Wirkung auf die hausinternen Beziehungen (Arbeitsklima) spürbar geworden sein. Die PR Leute haben diese Verbesserung hoffentlich im Sinne der Gemeinsamkeit Unternehmen/Mitarbeiter verstanden herauszustellen.

2. Kreditprüfung

Viele kennen die Situation der Kreditprüfung aus persönlichen Beziehungen zu Banken aus der eigenen Vergangenheit. Wenn man die heutige Kreditprüfung durch Banken genau betrachtet, scheint alles eine Bauchentscheidung des Kreditberaters – oft im Gewand der bösen Zentrale. Untersuchungen, an der wir selber beteiligt waren, scheinen diese Annahme zu bestätigen.

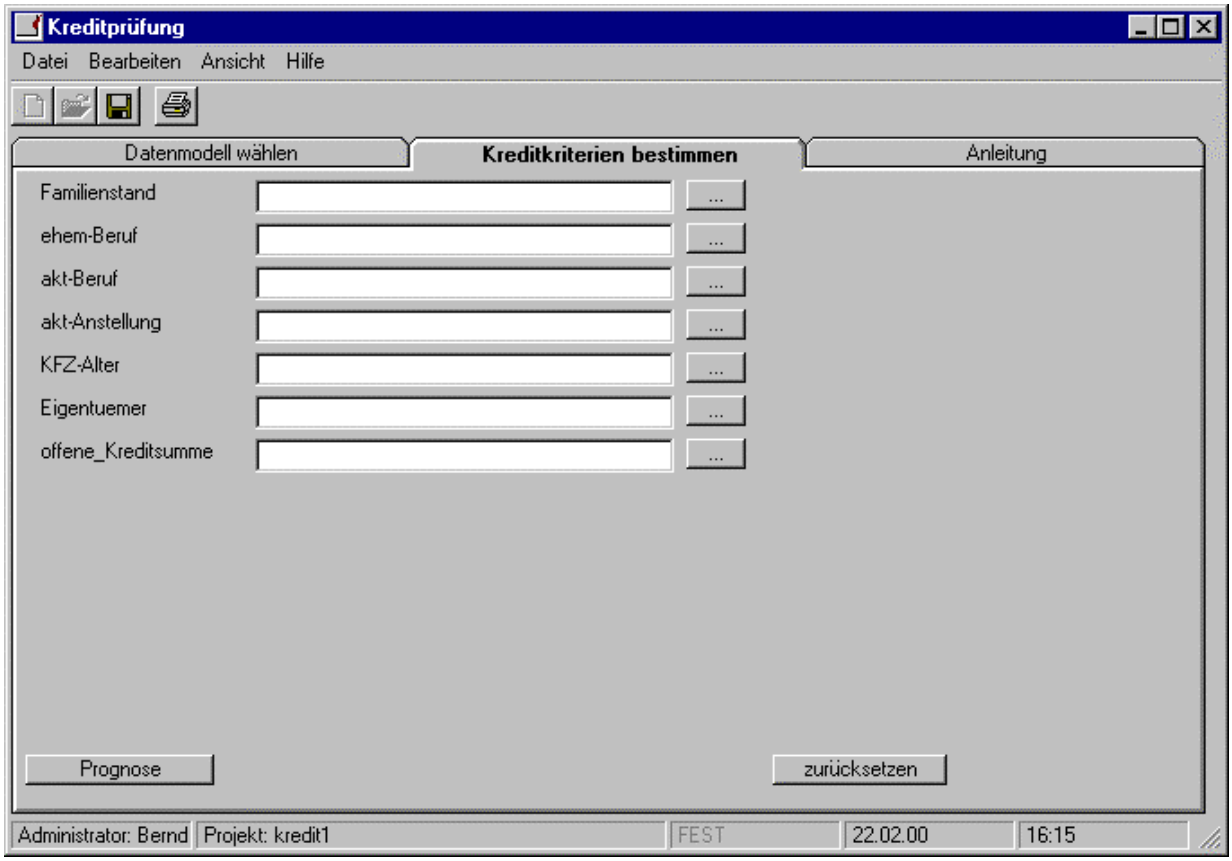
Die Ermittlung einer Kreditwürdigkeit ist die Antwort auf eine komplexe Frage, die Bestandteile beinhaltet, die kaum elegant in SQL zu fassen wären.

Data Mining baut eine Wahrscheinlichkeitsrechnung auf, die vergleicht - unter Verwendung bestimmter Algorithmen - die Daten eines bestimmten Falles mit den im aufgebauten Datenmodell abgebildeten Fällen. Das Datenmodell widerspiegelt in sich bekannte Fälle mit bekanntem Ausgang sowie trainierte neuronal gerechnete Ergebnisse. Welche Informationen von welcher Bedeutung sind, muß im Vorfeld analysiert werden.

Aus dem Wust von Informationen, die jedem Banker zur Verfügung stehen, ist es manchmal interessant zu sehen, welche Informationsteile als signifikant übrigbleiben. Zum Beispiel jemand der immer pünktlich seine Raten zahlt, wird unter bestimmten Voraussetzungen eher einen notleidenden Kredit verursachen als jemand der stets spät zahlt.

Dies widerspricht die üblichen Annahmen eines Bankers. Die Analyse der Tatsachen mit Hilfe der Möglichkeiten von Data Mining können diese eher traditionellen Annahmen eines Bankers als schlicht falsch bewiesen werden.

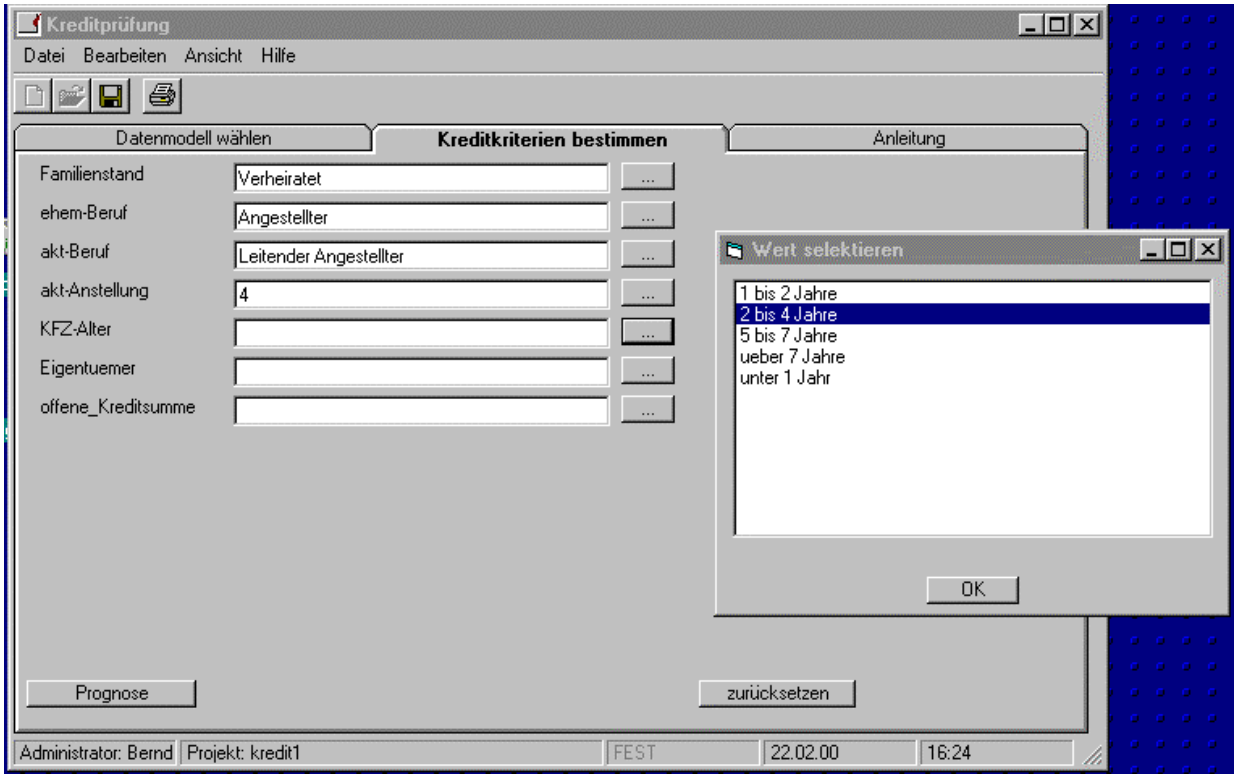
Eine geordnete Analyse der beim Banker vorliegenden Daten kann zur Feststellung einer Zahl von Kriterien führen, die immer eine wichtige Rolle bei der Kreditprüfung spielen. Hier im unteren Beispiel sind es 7 Kriterien, die links aufgeführt sind.



Das heißt, es müssen nur noch die jeweiligen, individuellen Ausprägungen eines Antragstellers eingegliedert werden. Im Hintergrund wartet das von Hausexperten gebaute Datenmodell. Wenn die individuellen Daten eines Kreditantrages zur Prüfung eingegeben worden sind, werden diese Daten (ein einziger Datensatz) mit dem Modell verglichen und das Resultat in Form einer leserlichen Ausgabe auf dem Bildschirm gebracht.

Selbst die Dateneingabe ist dermaßen erleichtert worden – wie uns das folgende Bild zeigt. Für einen geübten Kreditberater dürfte die Dateneingabe sowie die fundierte Aussage über einen Kreditantrag innerhalb von wenigen Minuten erfolgen können.

Die Berücksichtigung zentraler Informationen wie Kreditbestand, Risikofaktor, sowie der Grund für den Kreditantrag sowie spezielle Informationen über die Person – Schufa-Auskunft usw. – können ohne weiteres ins Modell bzw. im Programm verankert werden.



Jeder Mensch, der einigermaßen mit WINDOWS Applikationen umgehen kann, ist in der Lage, ein solches Programm in wenigen Minuten zu bedienen. Die Selektion des jeweils passenden Wertes entnimmt er seinen vorliegenden Kreditantrag und klickt den entsprechenden Wert aus der Selektion an.

Hat er dieses Formular komplett ausgefüllt – was das Programm von ihm verlangt -, klickt er Prognose. In wenigen Sekunden bekommt er eine deutliche Aussage, die er in Volltext erhält.

Vom IT-technischen Aufbau könnte dieses Prograämmle auf einem Laptop vor Ort beim Kunden oder auf einer Workstation in der Bank installiert sein. Diese Eingabegeräte könnten mit einem entsprechenden zentralen Server per Internet oder per LAN/WAN im Dialog stehen. Auf dem Server könnte das Modell geschützt vorliegen. Auf dem Server (oder auf einem anderen Server) würde die vollständige Berechnung durchgeführt werden und lediglich die resultierenden Informationen in Grafik und/oder Text an die auslösende Stelle zurückgegeben werden. Ein entsprechendes Protokoll würde im Hintergrund alle Vorgänge zu Kontrollzwecken festhalten. Somit entsteht auch weiteres, interessantes Datenmaterial liefern. Ferner könnten weiterführende Nachrichten an vorgesetzte oder nachgeschaltete Stellen kommuniziert werden.

Nach einem solchen System würde ein Kredit wirklich geprüft werden und die Schwachstellen für den Antragsteller - auch Kunde genannt – werden transparent.

In der Hauptsache hat die Bank diese wichtige Aufgabe rationalisiert und professionalisiert.

Dieses System entstand nachdem wir einige entsprechende Daten aus der Branche analysiert hatten. Jene Daten spiegelten Kreditwege vom Antrag bis zum Ende wieder. Informationen über Annahme oder Ablehnung des Antrages, Kreditsumme, Zahlungshistorie einschl. Mahnungen bis hin zur vollständigen Rückzahlung oder zum notleidenden Verfahren.

Bei dieser Analyse stellten wir fest, dass 7% aller 100%ig korrekt gelaufenen Kreditfälle nicht – ja nicht – genehmigt waren. Die Zahl schien zu deutlich, um mit einer Verwirrung in der Datenmenge erklärt zu werden.

Mit allen abgelehnten Kreditanträge führten wir eine Prognose durch. Dort fanden wir, dass 39% aller abgelehnten Kreditanträge als Kredit nach unserem Modell absolut korrekt durchgelaufen wären. Wiederum ist die Zahl so groß, dass wir nur von Fehlentscheidungen sprechen konnten. Diese Bank hatte 39% aller Anträge fälschlich abgelehnt und damit gute Kundschaft weggeschickt.

Diese erschreckenden Erkenntnisse haben uns dazu bewogen, entsprechende Modelle für die Kreditprüfung zu bauen.

3. Point-of-Sale Berechnung von Versicherungsprämien

Dieses englisch anmutende Applikationskonzept ist in Deutschland entstanden und betrifft die vor Ort Prämienberechnung sowie den Angebotsumfang einer Versicherung während oder am Ende des Beratungsgesprächs mit einem kaufwilligen Interessenten.

Das ehemalige Bild des Versicherungsagenten mit seinen dicken Mappen voller Tabellen hat sich nicht verändert. Die Tabellen haben sich in den meisten Fällen lediglich auf ein Laptop verlegt. Die vorgerechneten Tabellen existieren noch heute, jedoch nur in digitaler Form. Die Vorberechnung der Tabellen sowie die damit verbundene Zeitverzögerung (Dauer der Berechnung plus Dauer der Verteilung an die Vertriebsmitarbeiter) ist zwar verkürzt worden, jedoch nur um die Dauer der Verteilung. Die vorgelagerte Berechnungszeit ist nach wie vor gegeben.

Die Berechnung eines aktuellen, flexiblen Tarifs unter Berücksichtigung der Interessenteninformation, Marktsituation sowie Bestandssituation der Gesellschaft findet nicht statt.

Die Folgen hiervon sind jedoch gewaltig. Wenn man vor Ort eine Prämie kalkulieren kann, ist die Ermittlung vieler Kundeninformationen sinnvoll und für jeden verständlich. Daher leicht zu bekommen und zu behalten.

Die anschließende Berechnung eines individuellen Tarifs unter Berücksichtigung aller oben genannten Kriterien und mehr wird zum Kinderspiel degradiert. Die bisher übliche Vorkalkulation der Prämientabellen verliert seinen Sinn sowie seine Kostenberechtigung und Notwendigkeit.

Die Reaktionsfähigkeit auf plötzliche Veränderungen im Markt geht schneller als bisher von statten, da lediglich ein „Algorithmus“ oder ein Modell zentral angepaßt werden muß. Verbindungen vom Point-of-Sale zur Zentrale per Internet greifen somit sofort auf die neue Lage zu. Da der Berater vor Ort beim Kunden sowieso diese Modelle nicht sieht oder Algorithmen nicht kennt, kommt er in keinen negativen Vergleichsstrudel (der alte Tarif war günstiger, wie erkläre ich dies dem Kunden).

Die aktive Wirkung der Gesellschaft am Beratungsvorgang steigt. Die Möglichkeit „Cross-Selling“ Faktoren bzw. Kombifaktoren ins Angebot einzubauen, liegt auf der Hand bzw. im Rahmen der Kreativität der Visionäre im eigenen Konzern.

Zum Abschluß möchten wir einige provokative Fragen in Richtung Versicherungswirtschaft richten. Warum wird einer Familie nicht eine ganze Palette an Versicherungen zu einem „Kombipreis“ angeboten? Dies ist in der Wirtschaft bei Geschäftskunden üblich.

Liegt es nicht meistens daran, dass ein entsprechendes Werkzeug oder besser ein passendes Konzept fehlt und der Berater am „Point-of-Sale“ deswegen solche Offerten gar nicht in seinem Portfolio hat, weil keiner sie entwickelt bzw.pflegen kann?

Hierbei gehen positive Einflüsse auf weiteren kostenempfindlichen Themen wie Kundenloyalität, Stornoprophylaxe, Aktualität, Flexibilität usw. verloren. Anders gesprochen: Der Einsatz eines solchen Systems wie hier umrissen, behandelt gleichzeitig eine breite Palette von anderen mitentscheidenden Faktoren im Kampf um profitablen, sicheren Kunden.

4. Karriere Management

Fast jeder Konzern pflegt Daten über ihr eigenes Personal. Diese Daten können je nach Leistungsgruppe, Gehaltsklasse, aktiv/inaktiv usw. gestaffelt sein. Es werden Informationen über Eintritt in die Firma, Alter, Gehalt, Geschlecht, Adresse, Arbeitsstelle usw. gepflegt. Ebenso wird der Austritt und evtl. der Grund des Austritts vermerkt – sogar der Grund „Grund unbekannt“ wird von den gewissenhaften Personalverwaltern minuziös vermerkt. Es kommt eben „ alles in die Akte.“

Kaum jemand ist auf die Idee gekommen, dass diese „Personal“ einen Ressource d.h ein Aktiva der Firma darstellt. Manche Konzerne haben inzwischen den Wert ihrer Mitarbeiter längst erkannt und versuchen diese Fähigkeit als Leistungsmerkmal der Gesellschaft – obwohl sie nicht in der Bilanz erfaßt wird - zu pflegen.

Bisher scheint niemand auf die Idee gekommen zu sein, dass diese „Verwaltungsdaten“ die Grundlage eines völlig neuen Managementsystems bilden können.

Versuchen wir das Thema anhand einiger Fragen zu verdeutlichen.

Warum finden wir die Führungspersonen unserer Mitbewerber unter unseren ehemaligen Mitarbeitern wieder? Können wir unsere eigenen Führungspersönlichkeiten aufbauen oder kaufen wir sie ein? Warum verlassen uns unsere Middle Manager eher aus unbekannten Gründen? Was kostet uns der Weggang eines Managers über die Gehaltszahlungen und Abfindungen hinaus? Wie hoch ist der Bestandteil weiblicher Kräfte im Hause? Wie verhält sich diese Zahl zum weiblichen Anteil an den verschiedenen Managementstufen? Welche Komponente entscheidet über den Erfolg eines Menschen im Konzern – die akademische Bildung oder die erwiesene Leistungsfähigkeit? Wird Leistungsfähigkeit mit akademischer Bildung gleichgesetzt? Usw.

Dieser Fragenkomplex soll die eigentliche Thematik des Personal Managements nach Erledigung der Personalverwaltung verdeutlichen. Wer hier erfolgreich arbeitet, kann das menschliche Treiben in seinem Konzern im Sinne der ausgegebenen und praktizierten Unternehmenskultur bündeln. Dies wirkt an der Formulierung der Ziele, der Strategien, der Produkt- und Konzeptentwicklung usw. und fließt zurück in die Unternehmenskultur. **Hierzu muß er die richtigen Leute, zur richtigen Zeit, an den richtigen Stellen, richtig motiviert haben.**

Im komplexen, vielschichtigen Treiben einer Gesellschaft von Personen ist diese Aufgabe immens schwierig und schwer durchschaubar. Wo Menschen sind, ist sie auch gar nicht vorhersehbar!

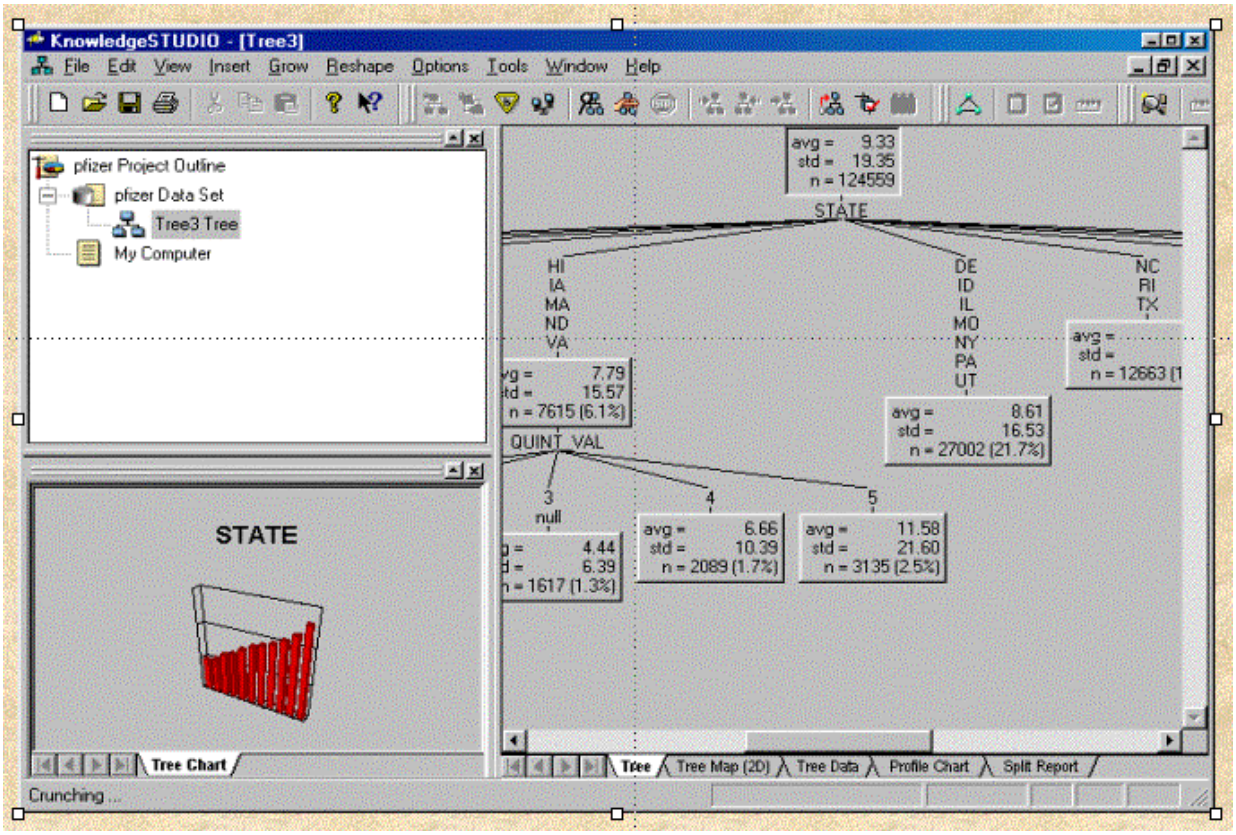
Stimmt nicht! Wenn ein junger Manager nach 10 Jahren Firmenzugehörigkeit kündigt, tut er dies meist nicht aus einer Laune heraus. Der Auslöser liegt oft weit zurück.

5. Fazit

Als Fazit dieser Szenarien möchten wir ein Wort ziehen – Fokus. In jedem Fall gibt es einen Wust von Daten und einen Nebel vom Verständnis. Data Mining erzeugt einen Fokus auf die entscheidenden Faktoren und schafft somit ein schnelles, aktuelles, transparentes, vollständiges und klares Bild der jeweiligen Situation.

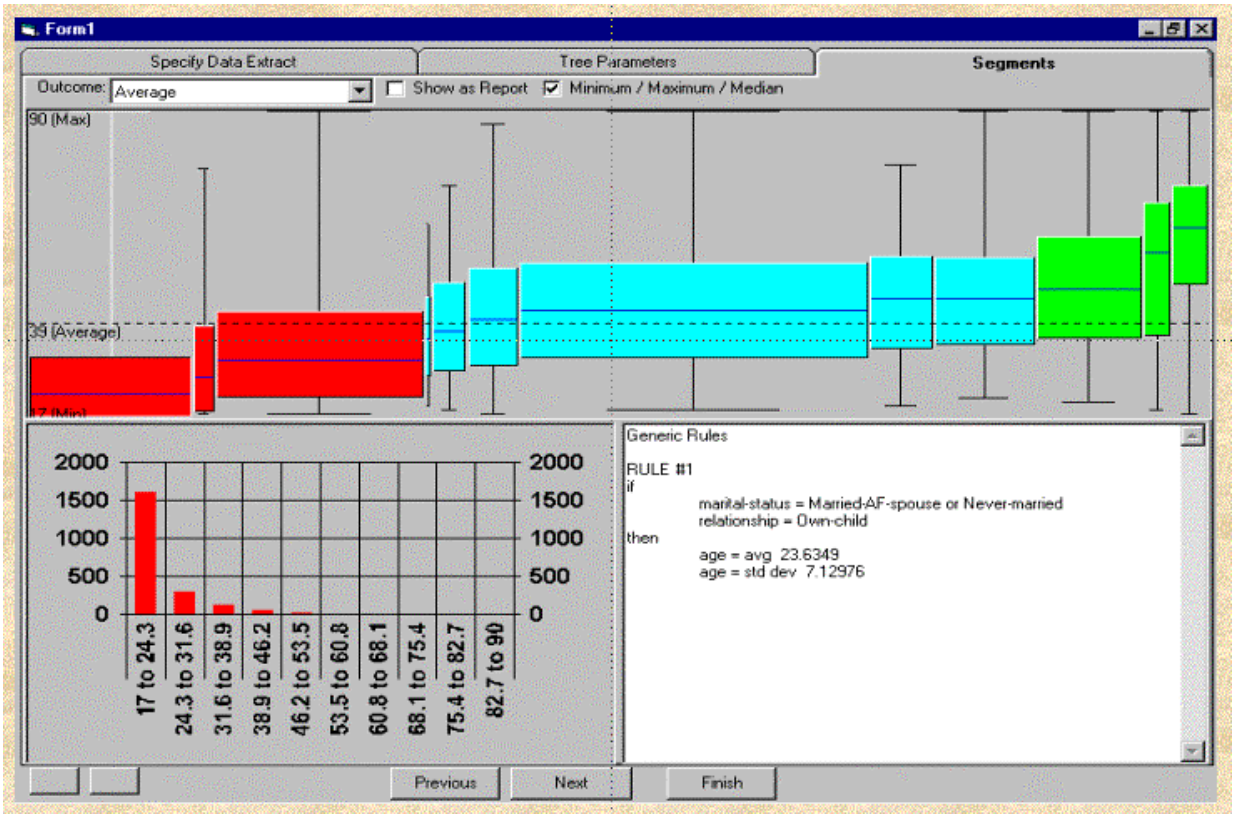
Im Falle der Spesenabrechnung sagt Data Mining welche Fälle bedenkenlos bearbeiten werden können und welche weniger. Bei der Kreditprüfung wird vorausgesagt, welche Kredite gesund laufen werden. „Last but not least“ erfahren wir die Grundlage eines neuen vom Kunden getriebenen Vertriebssystems in der Assekuranz.

Diese Szenarien sollen beispielhaft für viele sein. Die Diagnose in der Medizin, Cross-Selling Vorschläge im Call Center, Alarmsysteme in der Fertigung oder in der Betrugsaufdeckung bei der Schadensabteilung der Assekuranz oder Telekommunikation oder gar die Simulation von Marketing Kampagnen sind alle denkbar sowie bereits exerzierte Szenarien. Dies wird natürlich nicht vom Betreiber „an die große Glocke gehängt.“ Machbar sind sie alle und liefern den entsprechenden Gewinn beim Betreiber ab.



Dieses Bild zeigt die Erstellung eines Entscheidungsbaumes zur Ermittlung der entsprechenden Erkenntnisse zum Ablauf einer Marketingaktion. Gesellschaftsdaten sowie die Einteilung in Profitwerten sowie geographische Gebiete wird in einer Form vorgenommen, die manchen etwas schwierig erscheinen mag.

Das folgende Bild zeigt die gleiche Situation jedoch in einer Form, die von Database Marketing Leute leicht umgesetzt werden könnte.



Die zwei dimensionale Balkengrafik, der Regelbericht sowie die leicht zu überblickende Segmentierungen sind alle für einen Marketier vertraute oder leicht verständliche Hilfsmittel. Sieht aber nicht nach Data Mining aus, oder?

Wer entscheidet und wer verantwortet ?

Diese Rubrik bildet bereits vor Beginn des Hauptprojektes bereits eine existentielle Frage. Ohne Entscheidung gibt es keine Genehmigung und daher kein Geld. Also kein Startschuß.

Manchmal hängt es davon ab, dass niemand die Verantwortung tragen will und manchmal wird anschließend um die Verantwortung gedrückt. Egal wie, solche Ereignisse bestimmen den Werdegang mancher Data Mining Projekte.

Wir haben bereits gesehen, dass alleine die Terminologie des Projektes bereits zu Schwierigkeiten führen kann. Die Besetzung der Zuständigkeiten ist wiederum noch wichtiger.

In einem bekannten deutschen Unternehmen wurde der Beginn eines Hauptprojektes um über ein Jahr verzögert, weil der Verkaufsdirektor verstanden hat, das Data Mining ein Werkzeug sei, womit man Transparenz in jede Lage bringen könnte. Diese Transparenz sah er so gedreht, dass man ihm Fehlentscheidungen würde nachweisen können und damit entscheidend an seinem Stuhl werde sägen können. Die brisante Marktchance war nicht bestimmend, sondern rein der Selbsterhaltungstrieb eines scheinbar alternden Konzernlöwen.

Markant war auch die Fortführung dieser Geschichte. Auch der Hauptgegner dieses Herren nahm kurzfristig seinen Hut. Wiederum ein Jahr später wurde die positive Entscheidung gefällt. Der „neue“ Mann, der die Entscheidung zum Beginn des Projektes traf, war bis zum tatsächlichen Beginn des Projektes nicht mehr an seinem Platz. Alles Zufall?

Sowohl Entscheidung als auch Verantwortung muß vor Beginn des Projektes geklärt sein.

Die Entscheidung und die Verantwortung sollte man möglichst nicht trennen, jedoch gibt es einen sinnvollen und gangbaren Ausweg. Jemand im Projekt Team sollte die Entscheidungen fällen dürfen und möglichst über den notwendigen Rang der Firmen Hierarchie verfügen. Die Verantwortung sollte möglichst hoch angesiedelt sein. Am besten wird das Projekt von „ganz oben“ protegiert.

Da Data Mining als Prozeß seinen eigenen Arbeitstempi hat und aus der Sicht anderer eher bekannter Disziplinen teilweise seltsame Wege geht gar gehen muß, um entsprechend wertvolle Ergebnisse liefern zu können, ist die schützende, helfende Hand eines „Chefs“ für den täglichen Erfolg eines Projektes im Dickicht der Kompetenzen oft von unschätzbarem Wert.

13. Data Mining und herkömmliche Mittel im Vergleich

Als wichtige Säule eines Pilotprojektes kann ein Vergleich zwischen diesem neuen Verfahren Data Mining und der bisherigen Vorgehensweise gezogen werden.

„Bisher wurden Aufträge mit Fragestellungen erst in präziser Form entgegengenommen.“ Aus dieser Aussage entnimmt man, dass die bisherige Methode eine Frage/Antwort Systematik zur Basis hat. Mit anderen Worten wurden Antworten mit einer Reihe von Schritten „Frage-Antwort-Verifizieren“ unter großem Einsatz des „Bio-Computers Mensch“ beantwortet.

Jeder einzelne Schritt mußte sorgfältig überlegt werden. Jede Frage mußte in der Hoffnung artikuliert werden, dass die Antwort weitere Aufdeckung möglicher Ergebnisse beinhalte.

Der erste auffällige Unterschied ist die Möglichkeit mit Data Mining „unscharf“ formulierte Frage bearbeiten zu können.

Versetzen Sie sich in die Lage eines Analytikers, der von einem „Auftraggeber“ gebrieft wird, ein bestimmtes Thema zu untersuchen. In der Hauptsache geht es darum, das Thema in präzise Fragestellungen zu bringen, die der Analytiker beantworten soll. Diese Formulierung von Fragen ist schwierig. Nehmen wir uns ein Beispiel vor: Sie möchten Ihren potentiellen Kundenkreis nach möglichen Käufern, nach Umsatzumfang und Profitabilität sowie nach geographischen und demoskopischen Kriterien geordnet wissen. Diese diverse Merkmale müssen Sie vordefinieren, damit der Analytiker seine Aufgabe überhaupt durchführen kann.

Keine einfache Aufgabe, vor allem bildet sie ein Hindernis, denn wer möchte Antworten auf Fragen wiederum durch Fragen blockiert bekommen. Wenn Sie diese Folgefragen des Analytikers nicht beantworten, erhalten Sie Ihrerseits keine Antwort auf Ihre Fragen. Wäre es nicht schön, Ihre Frage zu stellen und ein Vorgeschmack auf mögliche Antworten zu erhalten, aus dem Sie weitere exaktere Anforderungen stellen können?

Zweitens verlangt das Data Mining eine saubere Datenmenge, denn die Rolle des Bio-Computers Mensch als ausgleichender Faktor im Analyseprozess, der Schwächen im Dateninhalt kompensieren kann, fällt bei Data Mining aus.

Die am Anfang des Data Mining Prozesses liegende Datenaufbereitung wird teilweise schnell zurückgezahlt, denn Ihr „Hauptprozeß“ wird durch einige meist interessante Nebeneffekte begleitet, die umsonst erscheinen.

Diese Nebeneffekte können als Einblicke in unwichtige Datenfelder, unsaubere bzw. falsche Dateninhalte liegen als auch in der Aufdeckung weiterer potentiellen Untersuchungsgebiete (die einem sonst nicht aufgefallen wären), die durch den Fokus der Data Mining Werkzeuge auf die wesentlichen Dateninhalte entstehen.

Drittens die Geschwindigkeit in der die passenden Data Mining Verfahren nachvollziehbare Ergebnisse erzeugt, ist im Vergleich schlicht berauschend.

Herkömmliche Untersuchungen werden in Wochen und Monaten angelegt. Oft sind die daraus folgenden Ergebnisse so gut wie nutzlos. Man erinnere sich in Deutschland Ende der Achtziger an die letzte Volkszählung, die ca. 18 Monate später erst Erkenntnisse hätte liefern sollen. Die Frage nach der Nützlichkeit und des Wirkungsgrades als Entscheidungsgrundlage solcher „alt gewordenen“ Erkenntnisse liegt auf der Hand.

Ähnlich sieht es im geschäftlichen Umfeld aus. Strategische Untersuchungen über Markt- oder Produktchancen können kaum 1 Jahr nach der Erhebung noch einen aktuellen Bezug besitzen. Hier kommt es eher auf die Erhaltung der Brisanz eines aktuellen Bezuges an. Erkenntnis die in Tagen oder Stunden gewonnen werden, können den entscheidenden Unterschied im Wettbewerb bedeuten. Data Mining liefert Ergebnisse in Minuten, Stunden oder Tagen.

Die Wirkung des früheren Einsatzes dieser gewonnen Erkenntnisse wird gepaart mit einer einfacheren Planung der notwendigen Analyseprozesse im Falle des Data Mining Einsatzes. Der kürzere Zeitraum ist leicht zu überschauen und daher aus Managementsicht leichter planbar und eigentlich leichter zu genehmigen. Aktuelle Informationen in Form von beherrschbarem Wissen stehen nun zur Verfügung, die ansonsten vielleicht gar nicht in Auftrag gegeben würden. Ein entsprechendes Return-on-Investement ist ebenfalls früher oder gar überhaupt zu erwarten.

Viertens die beinahe spielerische Art wie mit manchen Data Mining Verfahren umzugehen ist, erzeugt in sich eine Spannung, die sich als weitere Motivation äußert.

Data Mining macht Spaß. Sie ist spannend. Solche Aussagen hört man immer von Menschen, die Data Mining praktiziert haben. Die Ergebnisse stehen selten im Vorfeld fest. Es ist sogar fast unmöglich zu ahnen, welche Ergebnisse zu Tage gefördert werden. Man wirkt gezwungen, sich mit „anderen“ Erkenntnissen auseinander zu setzen und erhält dadurch einen erfrischenden Blick für Vorfälle, die sich vielleicht über Jahre etwas eingefahren und langweilig gewirkt haben.

Fünftens, vorausgesetzt die Hardware Plattform läßt es zu, versteht die Arbeit mit manchen Data Mining Verfahren keine störenden Grenzen im Umfang der Datenmengen.

Große Datenmengen en bloc zu behandeln, kann spannend sein und ist manchmal notwendig.

Die Suche nach Anomalien – z.B. eine Betrugsaufdeckung – in einer Datenmenge lassen kaum durch Sampling bewerkstelligen. Der Ausschnitt aus einer Datenmenge erhöht die Wahrscheinlich, das ein Ausreißer gar nicht gefunden wird. Wenn er überhaupt in der Datenmenge existiert, wird er nur mit Sicherheit gefunden, wenn die vollständige Datenmenge bearbeitet wird. Sampling ist hier tödlich.

Ferner ist der Gesamtblick manchmal notwendig, um die Gewißheit zu haben, dass die gefundenen Erkenntnisse eines Datenausschnittes auch in der großen Menge ebenfalls zustimmen.

Fazit

In der Natur einer Frage/Antwort Systematik liegt die Gefahr, dass Fragen schwer aufgedeckt und formuliert werden können. Eine gewinnbringende Garantie in der Antwort ist nicht gegeben. Daher müssen Fragen oft umformuliert werden und die jeweils folgende Abfrage an eine Datenmenge zeitraubend wiederholt werden, usw.

Dieses schrittweise Herantasten an eine mögliche Auflösung ist deutlich zeitraubend. Für den Menschen bedeutet diese Vorgehensweise eine Verbrennung seiner Kraft bei der Lösungssuche. Mit Data Mining konzentriert er sich auf die Interpretation vorhandener Ergebnisse.

14. Knowledge Management und Data Mining

Knowledge Management und Data Mining sind auf jeden Fall alliierte Disziplinen. Data Mining behandelt die Datenressourcen eines Unternehmens (implizites Wissen) und Knowledge Management versucht das Wissen aus den Köpfen der Mitarbeiter zu extrahieren und verfügbar zu machen (stillschweigendes Wissen).

Knowledge Management

Das stillschweigende Wissen als Basis des Knowledge Management muß ans Tageslicht gebracht werden und für andere Mitarbeiter verfügbar gemacht werden.

Nach langer Zeit der Wissensverschwendung durch Personalabbau (Controller gesteuerte Entlassungen um Ersparnisse beim Personal im G+V zu erzeugen) sind die Gesellschaften aufgewacht und entdecken den Wert des Wissens in den Köpfen ihrer Mitarbeiter.

Knowledge Management oder KM ist in aller Munde im Zusammenhang mit Begriffen wie Wissensdatenbank (Knowledge Base) als neues Heilmittel auf der Suche Wettbewerbsvorteilen. Nachdem Firmen sich jahrelang mit dem Rotstift gesundet haben, entdeckten einige wie sie ihre eigene Konkurrenz durch Stellenabbau geschafft hatten. Dabei sind „Wissensträger“ dem Rotstift zum Opfer gefallen.

Diese Wissensträger bauten eigene Firmen auf oder gingen notgedrungen zum Wettbewerb über. Deren Wissen stellte sich als wichtiger und mal entscheidender strategischer Faktor im Branchenwettbewerb heraus.

Das auf und ab von hi-tech Firmen am Beispiel der Formel Eins Teams (heute sind sie längst alle eigenständige Gesellschaften) wird nicht nur vom Fahrerwechsel, sondern vom Transferspiel der Designer, Ingenieure und Entwickler geprägt.

Der Wechsel von Ross Brawn von Williams zu Ferrari hat bestimmt deutlich zum Aufholjagd in der Qualität der Ferrari-Fahrzeuge im Kampf mit McLaren-Mercedes wie zum Abflauen der Erfolgswelle von den momentan Williams Erfolgen geführt.

Die Wiedergeburt von Chrysler während der 80er Jahre unter dem Präsidenten Lee Iacocca kennen inzwischen viele Betriebswirte und Manager. Wieviele wissen wie Iacocca als Erfolgsverkäufer bei Ford durch den noch aktiven Henry Ford II geradezu gefeuert wurde. Das Wissen von Iacocca um die Branche als auch um viele gute

persönliche Kontakte wie auch den Willen es Ford zu „zeigen," trugen unter Garantie zum Erfolg von Iacocca beim scheintoten Chrysler bei.

Hätte man das Wissen von Iacocca in wiederverwendbarer Form bei Ford gespeichert, wäre der Schaden vielleicht geringer ausgefallen. Aber gerade an diesem Punkt entdeckt man die Kunst des Knowledge Management, nämlich das nützliche Wissen vom persönlichen Wert des Wissensträgers zu trennen. Der Nutzen von Wissen ist oft mit der Umsetzungsfähigkeit einer Person verbunden oder mit der Kraft einer Persönlichkeit vereint.

Am Beispiel des Chrysler Erfolges muß man fragen, ob das bei Ford gebliebene und dort umgesetzte Wissen den Iacocca überhaupt aufgehalten hätte? Denn das Wissen kann man von Personen – in diesem Fall Iacocca – kaum zurückverlangen und erst recht bei ihnen nicht löschen. Inwieweit Ford als Gesellschaft den Erfolg von Chrysler z.B. mit eigenem Marktanteil bezahlen mußte, ist bestimmt schwer zu ermitteln.

Als Ford jedoch selber damit begann – nach Abtritt des Henry Ford -, ihre eigene Erfolgsstory unter der Leitung der Geschäftsphilosophie von Tom Peters aufzubauen, fragt man sich wiederum, was hat als das Wissen des Mannes mit dem berühmten Namen Ford überhaupt genutzt.

Schließlich war es das Wissen eines Fremden, eines Externen, die dafür sorgte, dass die Ford Motor Company nach vorne getrieben wurde. „Everything we do is driven by you." Doppelte Bedeutung: Alles was wir machen, wird durch sie angetrieben / durch sie gefahren. Diese Genialität ging in der deutschen Übersetzung unter.

Diese Beziehung zwischen dem Wissen der Person und die Kraft der Persönlichkeit sollte eine der schwierigsten Punkte wenn nicht die Crux des Knowledge Management in der Praxis darstellen. Beim Data Mining spielt die Person keine geringe Rolle, jedoch überwiegen die Prozesse sowie die elektronische Prozeßunterstützung.

Wenn jedoch die Wissensessenzen mit Data Mining entdeckt und geprüft sind und damit zu Knowledge geworden sind, greifen die selben Schwierigkeiten, die aus dem Knowledge Management direkt stammen. Denn jetzt sind diese Essenzen Knowledge geworden und in den Köpfen irgendwelcher Mitarbeiter. Da gehört ein Sicherheitssystem her.

Data Mining im Knowledge-Umfeld

Das implizierte Wissen muß zuerst aus den Datenquellen hinauf gefördert werden, um danach daraus einen Mehrwert erzielen zu können. Diesen Prozeß nennen wir Data Mining.

Data Mining ist also mit der Entdeckung von Wissen aus bestehenden Datenquellen beschäftigt sowie mit der Umsetzung dieses Wissen in unternehmerischer Vorteil. Nachdem jedoch das Wissen entdeckt und verfügbar gemacht wurde, fällt es zum großem Teil unter den Mantel von Knowledge Management. Eine wichtige Auswirkung dieser Situation wollen wir im nächsten Kapitel behandeln, nämlich „Knowledge und die Sicherheit."

Der Einsatz bzw. wiederholter Einsatz von Knowledge aus den Data Mining Projekten ist jedoch eine Thematik, wo beide Disziplinen des Erfolges Willen Hand in Hand miteinander arbeiten müssen.

Fazit

Diese Disziplinen sind unterschiedliche Techniken, die aber das gleiche große Ziel verfolgen – Gewinnsteigerung, laufender Verfahrensverbesserung, strategische Vorteile .....). Teilweise greifen sie ineinander und teilweise arbeiten sie liiert aber getrennt voneinander.

Der Data Miner betrachtet Knowledge Management als Folge seine Arbeit. Kein Data Mining, kein Knowledge folglich kein Knowledge Management.

Der Knowledge Manager wiederum sieht sein Personal als wichtigste Komponente des Knowledge Reservoirs, die er pflegen, anzapfen und verwalten muß. Data Mining ist für ihn eine zusätzliche Wissensquelle, die am Rand seiner Tätigkeit operiert. Dort werden gelegentlich neue Aspekte gewonnen, die in sein Knowledge Reservoir hinein fließen können.

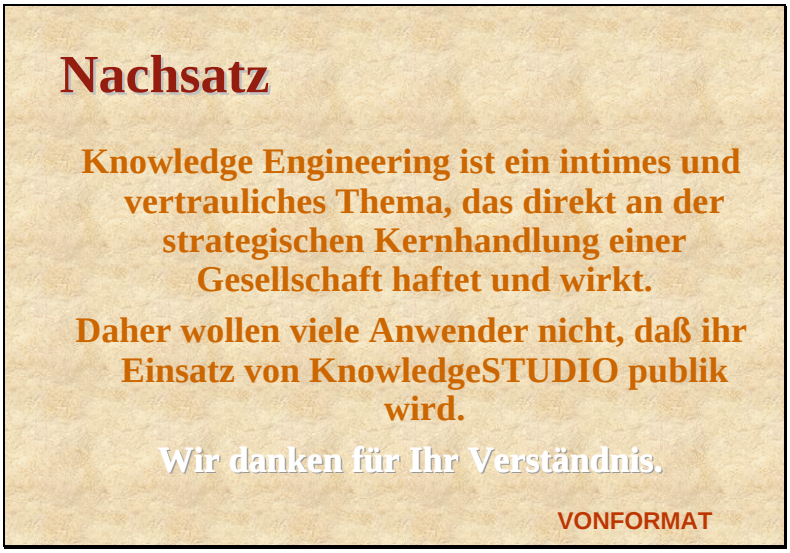
Aus der Sicht des Unternehmens sollten diese Disziplinen ineinander greifen können, denn Knowledge Management und Data Mining sind im Erfolgsfall auf jeden Fall eines feste Allianz.

15. Knowledge und die Sicherheit

„Wissen ist Macht“ glauben die meisten Mitmenschen, die auch die Sicherheitsbemühungen der meisten Institutionen beherrschen. Diese Macht wird vorsichtig verteilt und meist wenigen Mitarbeitern zugänglich gemacht.

Software Werkzeuge, die essentiell Wissen aufdecken bzw. Wissen verständlich machen, sind in sich kein Problem. Die Ergebnisse, die diese Werkzeuge produzieren, sind dagegen äußerst problematisch.

Das folgende Bild faßt die Thematik zusammen:-



Die Situation ist allzu nachvollziehbar. Die Frage wie geht man organisatorisch damit um, stellt sich von alleine.

In der Tat handelt es sich jeweils um das genaue Szenario, das hier behandelt wird. Im äußersten Fall müßte man einen strengen Unterschied zwischen „roten“ und „schwarzen“ Informationswege entsprechend der Systeme der heutigen Militärs schaffen. Diese Herrschaften betreiben solcher Systeme seit Jahren und haben zum Beispiel im NATO präzise Normen zur Einhaltung der Konventionen der Informationszugriffe entwickelt und aufgestellt. Der MOSSAD – berüchtigter und angesehener Geheimdienst in Israel – hat sogar spezieller Hardware entwickeln lassen, die eine Trennung dieser üblichen Informationskreise flexibel und intelligent unterstützt. Diese Hardware ist mittlerweile kommerziell im Markt zu haben.

Im Business Umfeld sind die Anforderung meist nicht so streng, jedoch man kann sich bei den Militärs um die Systeme bedienen. Meist kann man sich gut an Management Konventionen der Hausorganisation bzw. der Zulassungen für den Zugang zu finanzieller Informationen der Bilanzart anlehnen und übernehmen.

Lehnt man sich an Management Konventionen des Hauses an, lassen sich Data Mining Erkenntnisse und Systeme in diese bestehende Ordnung eingliedern.

Es entsteht sogar ein Plädoyer bzw. eine feste Grundlage für den Einsatz von Data Mining Werkzeuge innerhalb dieser Kreise. Der Business Manager, der diesem „inneren“ Kreis meist angehört, löst meist in seiner Person automatisch auch alle Sicherheitsüberlegungen, die aus der Arbeit mit Data Mining Werkzeugen enstehen könnten. Durch eine solche Allianz mit eingewiesenen Business Managern hat der Data Miner gleichzeitig die notwendige politische Rückendeckung zur Bewältigung etlicher Hürden im Hause, insbesondere bei der Datenbeschaffung oder später beim Einsatz oder Testeinsatz von Data Mining basierten Applikationen. Ein Manager, der bereits den Wert dieser Arbeit kennt, wird sie meist nicht aufhalten wollen.

In besonderen Fällen wie Revision, Finanzbuchhaltung usw. muß das Data Mining auf die dortigen Fachleute übertragen werden. Der Einsatz eines „Fremden“ wird an diese neuralgischen Stelle in der Regel nicht geduldet und erst recht niemand aus dem eigenen Hause, der dieser Abteilung nicht angehört. Es besteht somit die Notwendigkeit ein für Anwender freundliches Paket auszusuchen. Andernfalls muß die Abteilung im Gespräch – auf der Basis von Daten mit exakter Struktur aber mit fiktiven Inhalt eine spezielle Data Mining Applikation entwickeln lassen. Beide Varianten in der Anforderung können bereits nach dem heutigen Stand im Markt komplett erfüllt werden.

Nachdem Regelsätze entwickelt und Applikationen inkl. Modelle erstellt sind, muß die Pflege dieser Wissenskerne sowie den Zugriff geregelt werden. Der Zugriff wird meist über entsprechende programmierte bzw. Systemrechte geregelt. Ein „innerer Kreis“ hat das alleinige Schreibrecht auf solcher Projektdateien, denn sie müssen für die Aktualisierung sorgen.

Der Zugriff von Unberechtigten an dieser Stelle muß absolut sicher gestellt werden, denn unerlaubte Änderungen an dieser Stelle könnte fatale Folgen haben. Stellen Sie sich vor, ein Bankhaus prüft alle Transaktionen an ihren Bankautomaten mit einem Data Mining gestützten Programm, das jemand wie folgt manipuliert hat. Alle am Automat angegeben und für den Benutzer sichtbaren Zahlen, die zur Auszahlung führen – also DM 300 – DM 400 usw. -, werden durch das Eurozeichen ersetzt und genehmigt. Plötzlich werden alle Auszahlung gegenüber der geltenden Konvention automatisch verdoppelt.

Das Sicherheitsprüfsystem, was meist eine Scoring Tabelle ist, wird ersetzt und verändert. Plötzlich stimmen die Prüfkriterien des Vorhersagemodells nicht mehr. Das Chaos wäre kurzfristig Herr der Lage.

Weitere Horrorszenarien sind auszudenken. Insbesondere können Intelligente Systeme mit „intelligenten“ Veränderungen versehen werden. Das heißt, die Veränderung wie in unserem Szenario am Bankautomat dürfte relativ schnell auffallen. Spätestens dann wenn der Automat in kürzerer Zeit als sonst leer ist. Die Veränderung des Modells um einen kleinen Faktor wie 5% statt die 200% wie oben dürfte eine ganze Zeit länger unentdeckt bleiben und für den Kriminellen deutlich lohnender ausfallen.

In vielerlei Hinsicht decken sich die Sicherheitsanforderungen mit den üblichen Erwartungen aller brisanter Daten, jedoch bei Data Mining ist nicht nur die Brisanz der Daten wichtig, sondern auch die Werkzeuge sowie die Leute, die diese Werkzeuge führen, mit denen das brisante Wissen erzeugt wird. Hier geht es also nicht nur darum die Diamanten zu bewachen, sondern gleichzeitig die Diamantenmine.

16. Rund um das Data Mining

Vor dem Beginn der Data Mining Arbeit als auch danach gibt es eine Reihe von artverwandten sowie anschließenden Aufgaben, die zu bedenken sind.

Das Data Mining findet nicht im luftleeren Raum statt, sondern benötigt Daten in einem bestimmten Format und Zustand, um überhaupt funktionieren zu können und in einer in die Data Mining Ansicht transformierten Form, um besser arbeiten zu können. Obendrein muß das ganze Data Mining Unterfangen einen meßbaren Vorteil bringen, ansonsten kann man sich die Aufgabe schenken.

Dieser Vorteil sollte zur Steigerung des „Gewinns“ nicht nur in den Einsatz gebracht werden, sondern muß auch innerhalb einer Institution in die Kommunikation einfließen.

Die Verbreitung entsprechenden Wissens, dass aus diversen Aufgaben des Data Minings erzeugt und verständlich gemacht wird, kann manchmal ausschließlich durch die Kenntnis von Menschen in Gewinn umgewandelt werden, ohne Aufbau irgendwelcher Vorhersagesysteme.

Als Beispiel haben wir in einem Handelsunternehmen entdeckt, dass alle Kaufinteressenten mit einem bestimmten Merkmal, gar keine Kaufabsichten hegten und alle Kaufinteressenten mit fünf bestimmten Merkmalen besonders starke Kaufabsichten aufzeigten.

Solche Informationen müssen unter der Verkaufsmannschaft schnell zu Allgemeinwissen werden; ansonsten werden diese brisanten Wissensperlen nicht in die Tat umgesetzt und die Mannschaft verpulvert ihre Energie in der Jagd nach Interessenten ohne Kaufabsicht, statt sich gewinnbringend auf die zweite, wichtige Interessentengruppe zu konzentrieren.

Zuerst beschäftigen wir uns mit der Vorbereitung für das Data Mining.

Am Anfang war eine Flat File

Die Hauptanbieter von universellen Data Mining Tools haben zwei völlig unterschiedliche Philosophien, die prägend auf die Anfangsarbeit wirken. Alle verlangen eine Flatfile, aber danach trennen sich einige Wege.

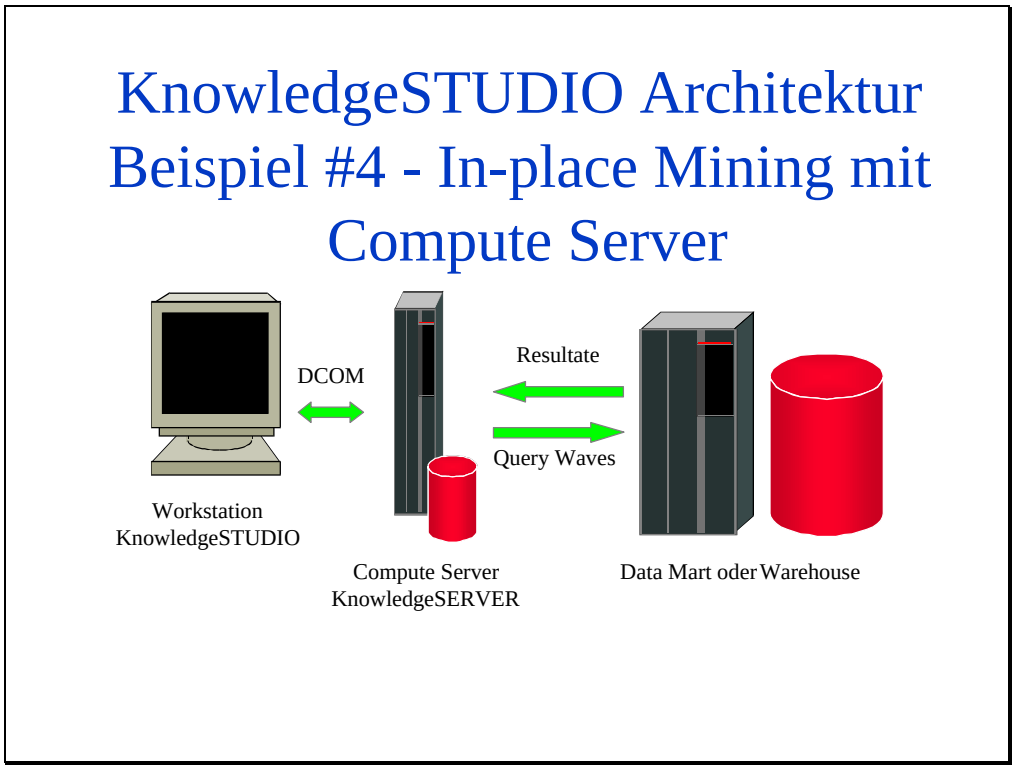
IBM, SAS, SPSS und andere bevorzugen hauptsächlich ihre eigene ziemlich geschlossene Datenwelt. Fremde Datenformate finden hier keine Anwendung. Sie müssen die Daten in Form X liefern, ansonsten können Sie einpacken oder müssen

auf einige starke Eigenschaften des Pakets verzichten oder schließlich dann doch sich der vorgeschriebenen Datenwelt anpassen.

ANGOSS – als einziger führender ausgesprochener Spezialist dieser Sparte – bietet auch die Möglichkeit einer Reihe von bekannten Import-, Export- und Regelformate sowie das In-Place Mining™.

In-Place Mining

In-Place Mining™ ermöglicht Data Mining Prozesse ohne die zugrunde liegenden Daten überhaupt importieren zu müssen. Die Daten werden in Fremdformate wie ORACLE, INFORMIX, SYBASE, MS-SQL 7.0, InfoCharger, WhiteCross usw. direkt im Sinne des Data Minings verarbeitet.



Rechts außen sehen wir ein drittes Tier abgebildet, das die Daten im Format eines In-Place Mining Datensystems abbildet. Dort findet auch die Verarbeitung statt. Von der Data Mining Software werden lediglich Query Waves (Abfragen) an die Daten geschickt, die dort in „kilometerlangen“ SQL-Statements umgewandelt werden. Ausschließlich die Ergebnisse werden zurückgeschickt. Die Belastung für den Netzverkehr ist gering.

Hierbei kommen die verschiedenen Sonderfähigkeiten – vor allem die Spezial DB-Systeme für Analyse wie WhiteCross oder InfoCharger – besonders durch die starke

Leistung und im Falle InfoCharger durch die ETL-Fähigkeiten in einer heterogenen Datenwelt zur Geltung.

Datenbeschaffung

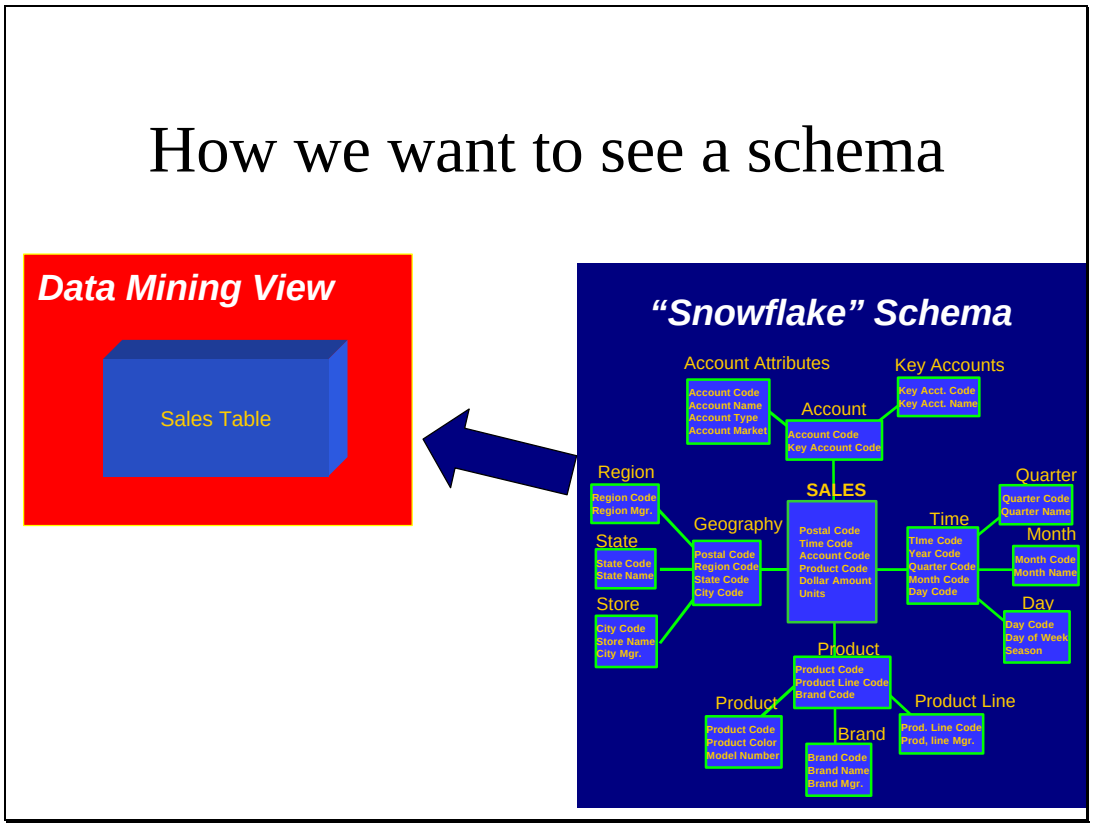
Hier stoßen wir auf weitere praktische Probleme des Data Miners, die eigentlich nichts mit dem Data Mining Softwarepaket zu tun haben. Wo und in welchem Zustand liegen die Daten? Eine reelle Situation ist die absolute Verteilung der Daten in verschiedenen Formaten, auf unterschiedlichen Rechnersystemen sowie unter unterschiedlichen Betriebssystemen. Ferner im Rahmen eines Data Warehouse können die vom Data Miner geforderten Tabellenspalten in einem bestimmten Schema verteilt. Wie lösen?

Der langwierige Weg ist die Bildung einer zentralen Data Mining Datenhaltung in Form eines Data Marts, wo mit großem Aufwand und durch Unterstützung der IT-Technik viele Transformationen, SQL-Joinvorgänge, Daten-umwandlungen und unter Aufbau großer Redundanzen, sowie entsprechend erschreckende Steigerung des Netzwerkverkehrs und Vergeudung von CPU-Zeit.

Ein solches Szenario ist technisch machbar. In der Praxis wird es meist an der politischen Zustimmung unter den vorgeschobenen hohen Kosten scheitern. IT-Leute haben „wichtigeres“ zu tun. Außerdem ist das ganze nicht notwendig.

Eine spezielle Datenbank wie InfoCharger wurde zur Unterstützung von Analyseaufgaben wie Data Mining extra entwickelt. InfoCharger wird eine universelle virtuelle Tabelle (Flat File) aus den oben genannten völlig heterogenen Datenumwelt bilden und sogar jedes Snowflake Schema plätten.

Das folgende Schaubild bietet eine einfache Darstellung des Bildes vom Data Warehouse Umfeld zur Data Mining Ansicht.



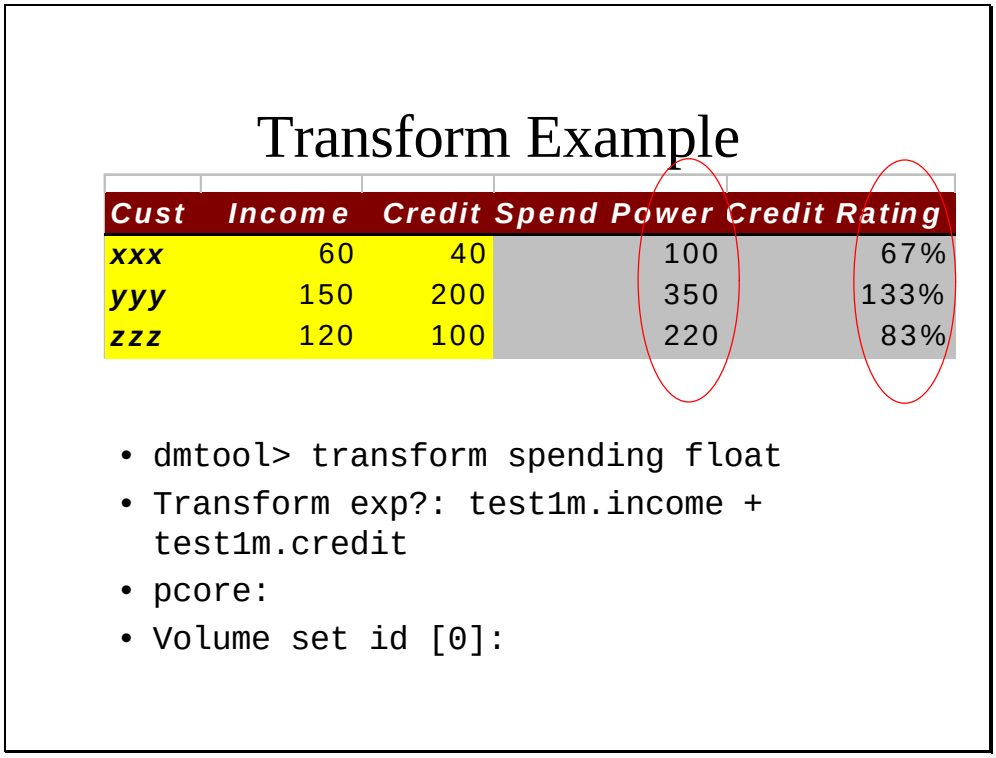
Da der Data Miner oft seine Datengrundlage bei Projektbeginn erst bilden muß sowie einen Arbeitstempo und einen Arbeitsrhythmus pflegt, die dem üblichen Fortgang der IT-Welt sprengt, bietet es sich nach der getanen Pflicht der Datensammlung, eine unabhängige Grundlage zu bilden. InfoCharger bietet diese Grundlage als Datenbank und ermöglicht sämtliche Datensäuberungen, Erweiterungen und Transformationen durchzuführen, die zur den ersten Schritten des eigentlichen Data Minings gehören. Eine SQL-ähnliche Struktur sowie grafische Anwenderoberfläche erleichtert diese Arbeit.

Diese Veränderungen an der Datenquelle erleichtert die Aufgabe des Data Miners, insbesondere dann wenn er nicht erneut importieren muß. In-Place Mining™ von ANGOSS KnowledgeSTUDIO ermöglicht dies. Daher ist lediglich ein erneutes Laden der Daten erforderlich, da sie nicht importiert werden müssen.

Datentransformation und -erweiterung

Transformationen zur Steigerung des Datenwertes können leicht vorgenommen werden. Das untere Beispiel bildet in Zeile 1 eine neue Spalte Spend Power (dt. Kaufkraft) aus der simplen Addition der Spalte Income (dt. Einkommen) und Credit (dt. Kredit). Ferner wird die Spalte Credit (dt. Kredit) durch die Spalte Income (dt. Einkommen) dividiert und als Prozentwert in der neuen Spalte Credit Rating (dt.

Kreditanteil) gebildet. Aus der Sicht des Data Miners steigern diese neuen Spalten die Aussagekraft solcher Datenmenge vielleicht entscheidend.



Zum Schluß hat der Data Miner oft das Problem der möglichen Größe seiner Datenmengen. Nicht immer kann er mit Sampling arbeiten. Die Untersuchung einer Datenmenge nach Anomalien (wenigen Ausnahmen) läßt sich mit einem Ausschnitt der Daten kaum bewerkstelligen. Es könnte schließlich sein, dass die gesuchten Datensätze just außerhalb des Datenmusters liegen. Also muß man sicherheitshalber alle Daten der Menge verarbeiten können. Dies könnte eine Menge von X 100 Millionen gar Milliarden Datensätze bedeuten.

InfoCharger kann diese Mengen durch Komprimierung und parallele Verarbeitung auf relativ bezahlbarer- wenn nicht bereits vorhandener - Standardhardware (z.B. NT oder SUN Maschinen) erstaunlich schnell verarbeiten. ANGOSS KnowledgeSTUDIO kann wiederum per In-Place Mining solche Mengen ebenfalls unterbringen. Erstaunlich aber sehr wahr.

Der Data Miner löst damit viele Probleme mit einer Entscheidung und kann ohne Umschweife klein anfangen und immens denken.

Die Data Mining Ansicht

An verschiedenen Stellen haben wir von der Data Mining Ansicht gesprochen. Vielleicht sollten wir diese Vorstellung etwas verdeutlichen.

Wenn Daten in Verwaltungssystemen abgelegt sind, haben Sie eine rein informatorische Aufgabe. Ein Geburtsdatum in der Personal Verwaltung wird als 19.10.1952 abgelegt und erfüllt komplett und präzise seine dortige Aufgabe. Ein Web Logfile enthält einen Datum- und Zeitstempel eines Zugangs zur Website in der Form „2000:06:28:11:50:23.“ Kompletter und präziser geht´s nimmer. Die Kundendatenbank enthält präzise Informationen über Kundennummer, Adresse mit Straße, Hausnummer, PLZ/Ort usw. Alles wunderbar, besonders wenn diese Daten vollständig, korrekt und einheitlich sind. Da kann wunderbar Briefe mit erstellen.

In dieser jeweiligen Form sind diese Daten für den Data Miner nutzlos. Was soll er mit einem Geburtsdatum, Zeitstempel oder Kundennummer errechnen? Alle Ergebnisse wären Unsinn.

Aus dem Geburtsdatum kann man jedoch Altersgruppe errechnen und diese neuen Ergebnisse in eine neue Spalte schreiben. Mit Altersgruppen kann ein Data Miner wunderbar agieren. Aus einem Zeitstempel kann man den Monat, die Woche, den Wochentag, eine Periode während des Tages sowie - mit weiterer Kalenderhilfe – eine Einteilung in Wochentag und Sonn- und Feiertag vornehmen. So schafft er aus regulären Verwaltungsdaten eine Data Mining Sicht. Er bildet eine andere Sicht der Datengrundlage, mit der sich rechnen läßt. Selbst mit der PLZ lassen sich regionale bzw. Kombinationen und Zuordnungen von Wohnort zu geographisch unterteilten Kaufkraftgebiete oder Nähe zu bestimmten Handelszentren usw. ermitteln.

Hier zählt das jeweilige Szenario sowie die eigene Phantasie des Data Miners wie er solche Wertsteigerungen in seine Datenmenge einbaut. Auf jeden Fall wird er mit solchen Werten zu brauchbaren Ergebnissen kommen, wenn er sich die Mühe macht, die Daten in die Data Mining Sicht um- und auszubauen.

Wir müssen hier betonen, dass keine Veränderung der Daten in der Operativen vorgenommen wird. Diese Veränderungen werden nur auf dem unabhängigen Data Mining Plattform durchgeführt. Selbstverständlich sollte man bestrebt sein, diese zwei unterschiedlichen Sichtweisen der Daten näher aneinander zu bringen. Dies wird nicht immer möglich oder wünschenswert sein.

Ferner muß betont werden, dass die Datenmengen nicht verändert werden, im Sinne der Verfälschung. Es wird lediglich die Sichtweise verändert. Wenn der 28.6.2000 in der Tat ein Dienstag war brauchen wir uns nicht fürchten, jenes Datum als Dienstag auszudrücken und anzusehen. Dieser Dienstag so wie jeder Dienstag kann in der Regel ohne Probleme zur Rubrik Arbeitstag zugeordnet werden – es sein denn es fällt ein Feiertag auf einem Dienstag. Davon gibt es in Deutschland außer dem ersten und zweiten Weihnachtstag nur die Möglichkeiten Sylvester und der 3. Oktober.

Hier liegt auch eine gewisse Kunst des Data Miners vor Beginn seines Projektes als auch während der Arbeit immer ein visionäres, wachsames Auge auf die möglichen Verbesserung seiner Datengrundlage aus der Data Mining Sicht zu halten. Just dieses Visionäre fällt fielen schwer, die sich sowieso mit dem virtuellen Aspekt der Data Mining Arbeit bereits schwer tun.

17. Data Mining und der Web

Die Möglichkeiten des E-Commerce für das Data Mining sind seit wenigen Jahren erkannt und in der Zwischenzeit in Produktkonzepten zur Bewältigung der resultierenden Informationsproblematik eingeflossen.

24x7- der Laden beherrscht mich

Zum Verständnis müssen wir uns die Situation des E-Commerce vor Augen als Hintergrundbild führen. Der „Laden“ ist 24x7 geöffnet und im Vergleich zum herkömmlichen Laden sieht man die Kundschaft nicht oder hat kaum eine geographische Verbindung zum Kunden. Die persönliche Beziehung zum Kunden im herkömmlichen Laden fehlt im Web fast gänzlich.

Das Internet ist global und unpersönlich. Die Kunden zeigt das Gesicht nicht, wir wissen kaum wo sie herkommen, jedoch trotzdem müssen wir uns um ihnen „kümmern,“ wenn wir ihnen als Kunden gewinnen und behalten möchten.

Der „Markt“ als Ort der Annäherung von Angebot und Nachfrage hat sich vom physischen Raum eines Marktplatzes oder eines Ladens in einen virtuellen Raum namens Internet mit seinen Websites verlagert. Dieses neue Medium erfordert neue Mittel und Verhaltensformen, um mit der neuartigen Informationsflut fertig zu werden, um erfolgreich zu werden oder zu bleiben.

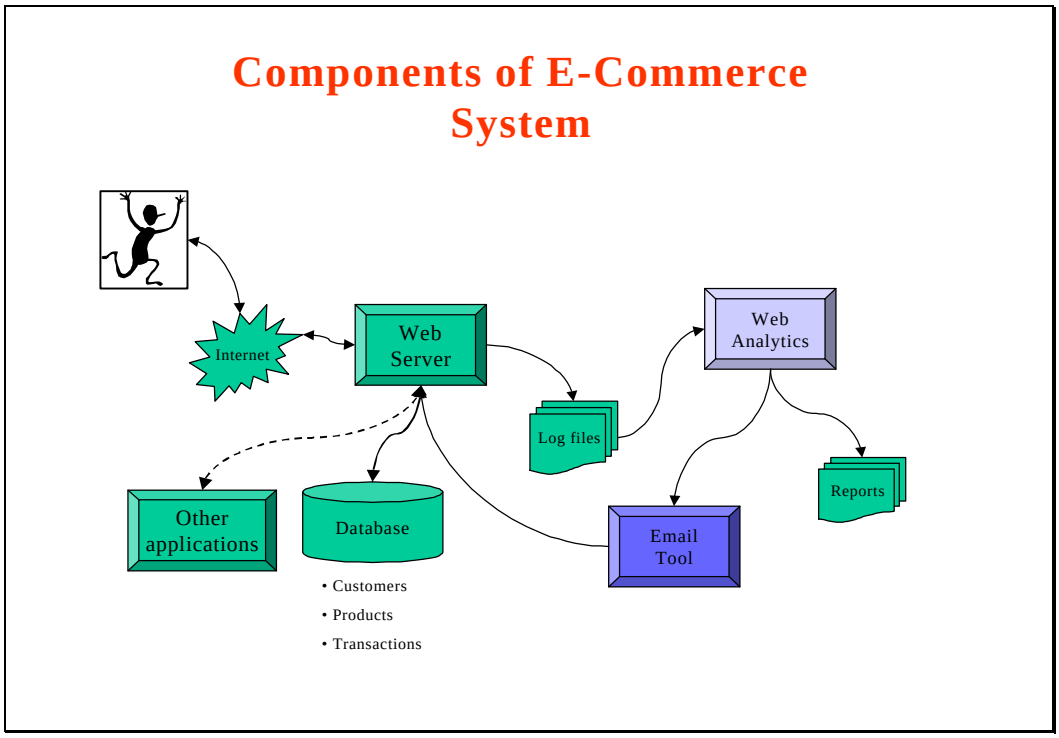
Der Fokus der Internet Technik wandelt sich schnell von der bisherigen Aufzählung von Hits bis zu einem Clickstream – wo Logfiles angelegt werden – bis hin zur nächsten noch unbekannten Möglichkeit.

Abgesehen von den momentanen technischen Möglichkeiten ist man bestrebt, ein „Browser“ (dt. jemand der sich umsieht) in ein „Transaktor“ (dt. ein Geschäftstätigen) umzuwandeln. In diesem Zusammenhang ist ein Surfer jemand, der nur die Homepage betrachtet. Verläßt er die Homepage und steigt er in die Seite ein, wird er bereits zum Browser. Der Transaktor schließlich ist der Kunde, der bereitwillig ist, Geschäfte zu tätigen.

Welche Möglichkeiten hat man überhaupt zur Verfügung, einen Browser zum Transaktor zu bewegen?

Sofern man seine Website entsprechend programmiert hat, verfügt man über ein Logfile, der solche Informationen wie Zeit- und Datumsstempel, Internet-Host des Browsers usw. liefert. Wiederum entsprechend der Programmierung des Websites können die Bewegungen des Browsers von Seite zu Seite mit Verweildauer (anhand des Zeitstempels) ermittelt werden.

Das folgende Schaubild zeigt eine herkömmliches Bild des E-Commerce und zeichnet die „üblichen“ Werkzeuge der Informationsaufbereitung, ihre Quelle sowie weitere vorhandene Quellen im Zusammenspiel, die genutzt werden könnten.



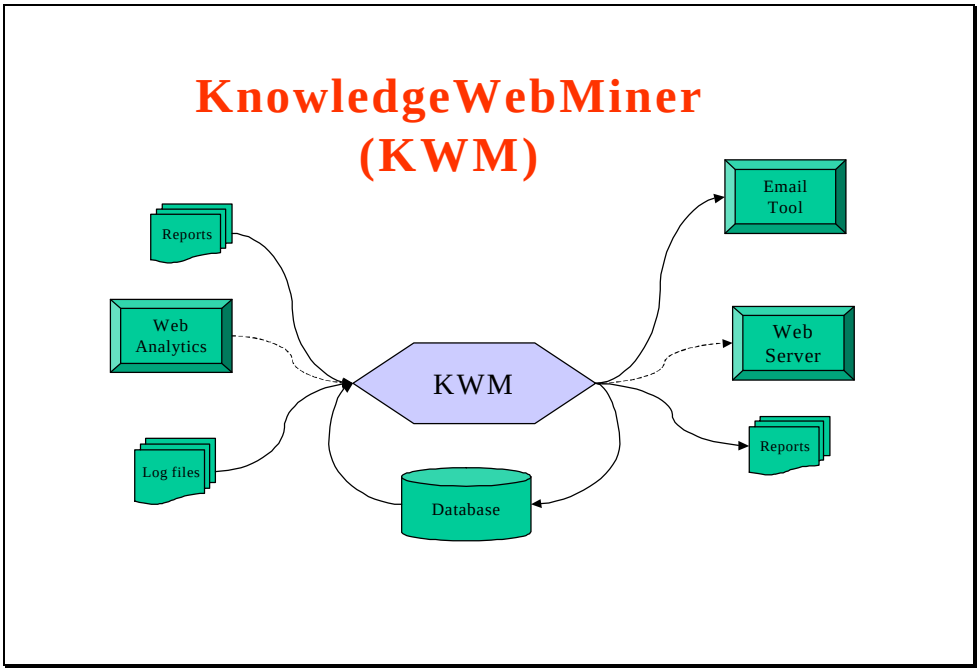
Die hier aufgeführten Web Analytics meinen die üblichen „zählbaren“ Information, wie Anzahl Besucher, Verkehrszahlen in diversen Seiten usw. Es dreht sich hier um Reports, die das Zählbare aufführen und keine Anstalten machen, die versteckten Trends im Verhalten der Besucher aufzudecken. Erst recht wird kein Versuch unternommen, jene Trends in Geschäftsregeln oder andere geschäftliche Maßnahmen umzusetzen. Hieraus an eine Steigerung in der Qualität der Website aus der Sicht möglicher Geschäftsabschlüsse oder einer verbesserten Kundenbedienung ist nicht zu denken.

Der Ansatz von Web Mining (Data Mining im Umfeld des Web) geht also weit über das Zählen von Ereignissen hinaus und untersucht die Zusammenhänge, die wiederum in nutzbare Formen umgesetzt werden. Hier einige Beispiele zur Verdeutlichung.

Web Analytics liefert Antworten auf die Fragen welche Besucher am längsten im Website verweilen und wo Sie herkommen (Land, Domain....).Web Mining dagegen erlaubt – hier das Szenario zum maximalen Ende gedacht - die online Vorhersage, ob ein Besucher ein oder mehrere der angebotenen Produkte kaufen wird und auch welche Vorschläge ihm unterbreitet werden sollten.

Integration verschiedener Datenbestände

Dies möchten wir anhand des Konzeptproduktes von ANGOSS Software verdeutlichen, KnowledgeWebMiner. Er nimmt eine zentrale Position im oberen Gebilde ein und sorgt damit als Schalt- sowie Umschaltstelle für die passende Um- und Übersetzung sowie Zusammenführung der Informationen aus den diversen betroffenen Quellen. (Siehe folgendes Bild)



KnowledgeWebMiner bietet die komplette Technologie und Methodologie zur Herstellung dieser Schaltstelle. Hierbei werden alle E-Business Themen der aktuellen Generation behandelt. Das flexible Design rechnet bereits mit den Neuentwicklungen der zukünftigen Generationen bei den allerbesten Optionen in der Skalierung – von kleinen Webumgebungen bis zu den größten Portalsites. Natürlich gehört ein Knowledge Transfer in der Methodologie zum Lieferumfang.

Das Business Modell des KnowledgeWebMiner gibt es in drei Variationen:-

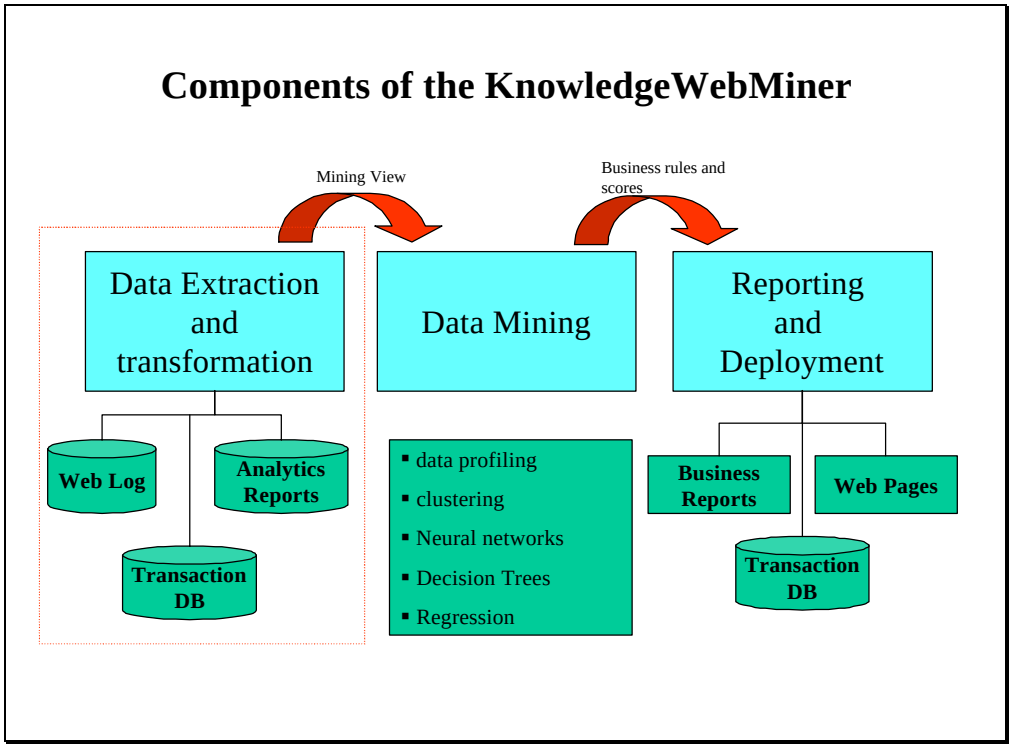
- eine umfangreiche schlüsselfertige Lösung,
- auf Pilotprojekt/Proof of Concept Basis
- als Service durch ANGOSS Mitarbeiter

Die Leichtigkeit der Anwendung, die Leistungsfähigkeit und die bezahlbare Variationen machen KnowledgeWebMiner zu einer äußerst interessanten Neuigkeit, die in einigen Sparten bereits für Furore sorgt.

KnowledgeWebMiner verwendet die KnowledgeSTUDIO Architektur, die seit Jahren eine erprobt Technologie darstellt. Diese Lösung baut auf 3 Kernkomponenten, die wir ebenfalls seit Jahren beherrschen:-

+ Datenextraktion und –transformation
+ Profiling, Mining und Visualisierung
+ Reporting und Datenmodelleinsatz

Im Zusammenhang werden die Haupt- und Nebenanforderungen von Geschäfts- sowie Gelegenheitsanwendern gleichermaßen bedient.



Das obige Bild zeigt die Zusammenführung der Komponente aus der Sicht der 3 Kernkomponente des KnowledgeWebMiner. Langsam entpuppt sich KnowledgeWebMiner als der besondere Einsatz eines „normalen“ Data Mining Projektes in einer Webumgebung. Dies ist auch so, jedoch gibt es einige prägnante Charakteristika des Web Mining, die eine besondere Bezeichnung rechtfertigen.

Die zwei wichtigsten Komponenten des Web Mining im Vergleich zum regulären Data Mining sind die zwingende Notwendigkeit der Datentransformation und Datenzusammenführung. Diese Aktion ist hier unabdingbar, um überhaupt arbeiten zu können, da das „Rohmaterial“ an Daten nicht nur schlecht, sondern völlig ohne Aussagekraft angeliefert wird.

Die Anreicherung der Daten durch vorhandene Daten wir Kundendaten stellt den möglichen Bezug dieser Logdaten zu „reellen“ Personen her. Hierdurch beginnt diese unpersönliche Browser das Gesicht eines persönlichen Menschen anzunehmen, der faßbare Bedürfnisse und Vorlieben hat.
Ohne diese ersten Schritte ist keine Analyse geschweige denn irgendeine Arbeit mit vorhersagenden Zielen möglich.

Die zweite besondere Komponente des Web Mining ist der beherrschende Zeitfaktor. Wenn eine Vorhersage Online erfolgen soll, müssen die Modelle wirklich pronto Ergebnisse liefern, um den Online Anspruch zu erfüllen.

Ferner spielt dieses Szenario jeden Tag rund um die Uhr und 7 Tage pro Woche. Je nach Popularität des Websites kann es sich hierbei um eine gewaltige Menge von Informationen handeln, die bearbeitet werden sollen.

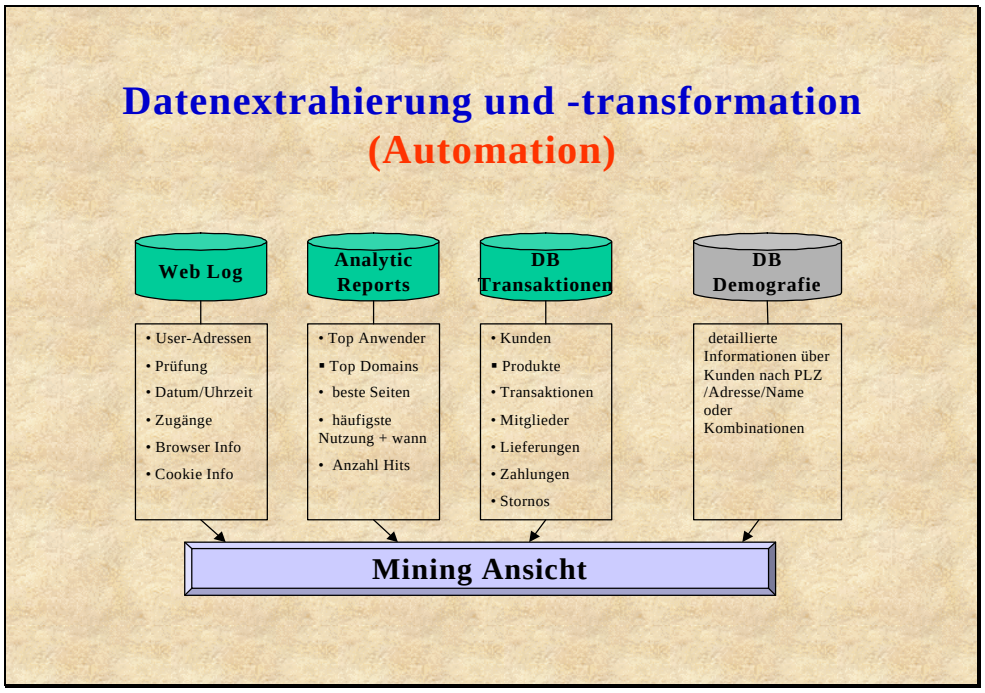
Dieses Zusammentreffen von einer großen unvorhersehbaren Menge von Daten, die in einer kurzen Zeit verarbeitet werden müssen, erzeugt einen Druck, der bei den meisten „üblichen“ Data Mining Projekten in dieser Schärfe nicht auftritt.

Der Druck der Datenbewältigung in Zusammenhang mit der – aus der Sicht des Data Minings – absolut nutzlosen Aussagekraft der anfallenden Daten bildet eine besondere Lage in der Welt des Data Minings und rechtfertigt durchaus die eigenständige Bezeichnung Web Mining.

Nun zurück zum E-Business. Im Prinzip wird ein E-Business besonders erfolgreich, wenn das gesamte Handeln in eine Mining Ansicht mündet. Dies bedeutet:-

- ♥ Die Integration und Automation der Datenextraktion und Transformation für das Data Mining
- ♥ Die Integration mit Web Log Analysen wie Net*Gen, MarketWave usw.
- ♥ Die Integration mit relevanten Kundendatenquellen, ob Web oder Operativ, die dem Geschäftszweck dienlich sind
- ♥ Integrierte Zusatzinformationen aus externen Angeboten wie Dun&Bradstreet, AZ, PanAdress, Microm, GfK, Claritas usw.

Um diese Anforderungen zu erfüllen, kommen alle wichtigen Data Mining Techniken zum Einsatz.

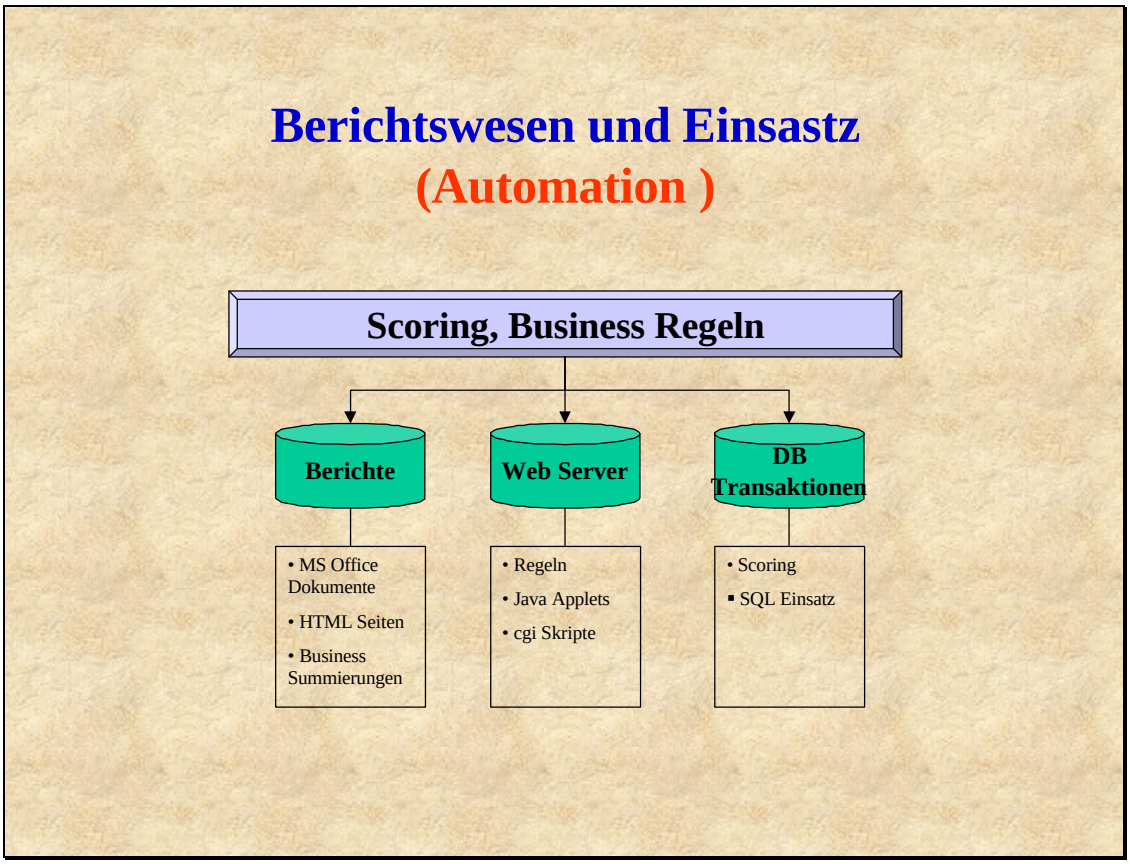


Das Ziel der Data Mining Techniken ist die Erzeugung der Data Mining Sicht, aus der die Vorteile für das Geschäft abgeleitet werden können. Im oberen Bild haben wir 4 Säulen, die als Datenlieferanten im Maximalfall dienen können gar dienen müssen, um die konstruktive überhaupt zu ermöglichen.

Online Angebote mit Vorhersagemodellen

Nach der Zusammenführung und Aufbereitung dieser Informationen in die Data Mining Ansicht wird unter Verwendung der diversen Techniken eine Mischung aus Berichtswesen zum Thema wie auch das Rüstzeug eines „Deployment“ (dt. in den Einsatz bringen) erzeugt. Hiermit werden Managementanforderungen erfüllt sowie entsprechende „Verbesserungen“ des Website und direkte Wirkungen auf den „Transaktor ermöglicht.

Diese Scorings dienen zur Klassifizierung der Kunden – vielleicht im Sinne der Kaufwahrscheinlichkeit, des Kaufumfanges und des Kaufzeitpunktes - als Fokus auf die „wichtigsten“ Besucher. Die Geschäftsregeln steigern den Fokus auf die gewinnbringenden und Umsatz wahrscheinlichen Besucher sowie alle möglichen Maßnahmen im Webumfeld, die zum Erreichen dieser entscheidenden Ziele dienen.



Die Integration mit Office 2000 sowie mit der Regelgenerierung in vielen vorhandenen Formaten wie SQL, Java u.a. sowie in natürlicher Sprache wie englisch oder deutsch erlauben eine sofortige Eingliederung in die IT-Umwelt des Kunden.

Das Scoring von bestehenden Datenmengen erlaubt die Kategorisierung weiterer Datenmengen mit den Mitteln des Data Mining.

Wie setzt man so was um? Auch hier werden die Schritte aus dem Data Mining Projekt Rad befolgt.

Vertraulichkeitserklärung
Bestimmungen festlegen
Prototyp/Demo mit Projektplan entwickeln
Hauptziele und Technologie Plan festlegen
Kundenorientierte KnowledgeWebMiner liefern
sowie Szenarien bestimmen

Es folgt nach jedem Schritt die Entscheidung über den Lieferumfang. Zuerst wird ein Pilot als Proof of Concept in Auftrag gegeben und nach dem Abschluß des Prototyps das Hauptprojekt entschieden. Es folgt danach die Entscheidung über eine schlüsselfertige Lösung oder der Kunde möchte die Entwicklung selber betreiben und

bestellt eine Service Lösung zur Unterstützung beim Aufbau seiner Einsatzapplikation.

Was bietet die Zukunft?

Die Zukunft des Web Mining hängt im wesentlichen von zwei Komponenten ab. Die erste ist die eigene Entwicklung und den eigenen Einsatz von Web Mining Applikationen und Konzepten auf der Basis der gewonnenen Erfahrungen und der wachsenden Anforderungen durch anwendende Institutionen. Die zweite Komponente hängt von der technischen Entwicklung im Bereich Internet Technologie im Hardware- und Softwarebereich ab.

Die eigene Entwicklung und Erfahrung beschränkt sich teilweise auf die Beherrschung eines bisher unbewältigten Terrains. Man ist froh alles in den Griff bekommen zu haben. Jedoch ist heute bereits abzusehen, dass die Internet Technologie teilweise der Maßstab für alle Netzwerke – intern oder extern, offen oder geschlossen – werden wird. Die kostengünstige, universelle Technik, die auf allen diesen Gebieten einsetzbar ist, wird das Argument liefern.

Auf dieser Basis werden immense Mengen von Information jedem zur Verfügung stehen, solange er einen Computer, eine Anbindung und Zugriffsrechte hat. Die Verteilung von Informationen – insbesondere auf der Basis intelligenter Systeme – wird wachsen. Börseninformationen mit Hinweisen über Kauf- und Verkaufsabsichten wie ein elektronischer Makler, Cross-Selling Service – Konzertkarte zusammen mit Anfahrtsinformation, Bahnkarten, Hotelreservierungen usw. oder individuelle Tarifgestaltung bei Versicherungsangeboten wie ein elektronischer Versicherungsmakler usw.

Unsere Servicegesellschaft bekommt also einen weiteren Schub. Die anfallende Menge von Daten sowie das Verlangen nach höherer Intelligenz im System schreit nach Web Mining und/oder Data Mining. In der Tat solcher Systeme befinden sich heute bereits in der Entwicklung und stehen kurz davor, in den Einsatz zu gehen,

Fazit

Web Mining – am Beispiel einer der wenigen Produkte in diesem Sektor auf dem Markt von ANGOSS – ermöglicht einen flexiblen und individuellen Ansatz zur Bewältigung der versteckten und aus dem Virtuellen kommenden Erzeugung von Grunddaten und baut maximal eine Modell gestützte Einsatzapplikation zur Online Identifikation und zur Vorhersage der Anforderungen eines neu eingeloggten Transaktors in der Website.

Auf der Basis von Web Mining kann jede Website effizienter gestaltet und jeder Manager stets und pronto über die Gewinnträchtigkeit sowie Popularität seiner Produkte informiert sein. Diese Sondersparte des Data Minings wird sich in den

kommenden Jahren mit dem Internet sowie mit unseren eigenen Erfahrungen unter Garantie gewaltig weiterentwickeln.

18. Data Mining Modelle im Einsatz

Es bestehen unterschiedliche Möglichkeiten Data Mining in den Einsatz zu bringen. Sie unterscheiden sich nach Szenario, nach Anforderung und nach Zielsetzung. Wir haben hier den Überblick in drei Kategorien eingeteilt, die im Regelfall eine Steigerung im Umfang der Mehrwerterzeugung bedeuten.

Regelsätze erzeugen

Die folgenden Aufstellungen gehören zu einer Datenmenge über die Einflüsse, die auf den Zustand im Bereich Blutdruck (mit den Werten Niedrig, Normal und Hoch) wirken. Hier ging es darum, heraus zu finden, warum 21.3889 Prozent der Datenmenge (also Personen) unter hohem Blutdruck leiden.

Es wurde einen Entscheidungsbaum erstellt und folgende Regelsätze in englischer Sprache, als Generische Regeln (leicht für IBM IS einzusetzen) sowie SQL erzeugt. Die Ausgabe der Regeln erscheint anders, obwohl die Grundlage in jedem der drei Fälle identisch ist.

Als erstes haben wir den Regelsatz in englischer Sprache ausgegeben. Die Regeln 4 und 5 geben das kompletteste Bild der Auswirkungen auf das höchste Bild des Blutdruckes. Diese Ausgabe könnte jeder Manager oder Arzt oder sonstige Person verstehen, ohne überhaupt was von Informationstechnologie oder Data Mining zu verstehen. Ob sie es glauben und akzeptieren, liegt auf einem anderen Blatt.

English Language Rule # 1:
There is a 18.3333 percent chance that Blutdruck will be Niedrig, a 60.2778 percent chance that Blutdruck will be Normal, and a 21.3889 percent chance that Blutdruck will be Hoch.

English Language Rule # 2:
If Alter is equal to 3 then there is a 5.43478 percent chance that Blutdruck will be Niedrig, a 45.6522 percent chance that Blutdruck will be Normal, and a 48.913 percent chance that Blutdruck will be Hoch.

English Language Rule # 3:
If Alter is equal to 3 and Größe(cm) is between 611 and 675 then there is a 0 percent chance that Blutdruck will be Niedrig, a 45.8333 percent chance that Blutdruck will be Normal, and a 54.1667 percent chance that Blutdruck will be Hoch.

English Language Rule # 4:
If Alter is equal to 3 and Größe(cm) is between 611 and 675 and Salzverbrauch is equal to 1 or 2 then there is a 0 percent chance that Blutdruck will be Niedrig, a 61.2903 percent chance that Blutdruck will be Normal, and a 38.7097 percent chance that Blutdruck will be Hoch.

English Language Rule # 5:

If Alter is equal to 3 and Größe(cm) is between 611 and 675 and Salzverbrauch is equal to 1 or 2 and SportAktivität is equal to 5 then there is a 0 percent chance that Blutdruck will be Niedrig, a 33.3333 percent chance that Blutdruck will be Normal, and a 66.6667 percent chance that Blutdruck will be Hoch.

(Die Werte der Variablen Größe sollten nicht zu ernst genommen werden. Nach 7 Jahren hat der Autor noch keinen Sinn daraus machen können. Bitte um Nachsicht.)

Generische Regeln bilden eine Art Brücke zwischen der IT-Welt und Managern. Manche IT-Leute verwenden diese Regelsatzform, um ihre eigenen Ausgabe z.B. für die IBM Hostwelt zu erzeugen. Manche Manager finden diese Form leichter – denn mit weniger Sprache – zu verstehen, als die spezielle Ausgabe in einer natürlichen Sprache.

Generic Rules

RULE #1
(Whole Tree)
Blutdruck = Niedrig 0.183333333333
Blutdruck = Normal 0.602777777778
Blutdruck = Hoch 0.213888888889

RULE #2
if
Alter = 3
then
Blutdruck = Niedrig 0.054347826087
Blutdruck = Normal 0.45652173913
Blutdruck = Hoch 0.489130434783

RULE #3
if
Alter = 3
Größe(cm) = (611,675]
then
Blutdruck = Niedrig 0
Blutdruck = Normal 0.458333333333
Blutdruck = Hoch 0.541666666667

RULE #4
if
Alter = 3
Größe(cm) = (611,675]
Salzverbrauch = 1 or 2
then
Blutdruck = Niedrig 0
Blutdruck = Normal 0.612903225806
Blutdruck = Hoch 0.387096774194

RULE #5
if

```
        Alter = 3
        Größe(cm) = (611,675]
        Salzverbrauch = 1 or 2
        SportAktivität = 5
then
        Blutdruck = Niedrig  0
        Blutdruck = Normal  0.333333333333
        Blutdruck = Hoch  0.666666666667
```

Zum Schluß bieten wird das gleiche Bild in der Form SQL-Abfrage.

```
-- SQL Rule #: 1 (Root Node)
SELECT * FROM %table%;

-- SQL Rule #: 2
SELECT * FROM %table%
WHERE       Alter = 3;

-- SQL Rule #: 3
SELECT * FROM %table%
WHERE       Alter = 3 and
        (Größe(cm) >= 611 and Größe(cm) < 675);

-- SQL Rule #: 4
SELECT * FROM %table%
WHERE       Alter = 3 and
        (Größe(cm) >= 611 and Größe(cm) < 675) and
        (Salzverbrauch = 1 or Salzverbrauch = 2);

-- SQL Rule #: 5
SELECT * FROM %table%
WHERE       Alter = 3 and
        (Größe(cm) >= 611 and Größe(cm) < 675) and
        (Salzverbrauch = 1 or Salzverbrauch = 2) and
        SportAktivität = 5;
```

Der letzte Regelsatz kann (sofern die Variablennamen und die Feldnamen der SQL-Tabellen nicht verändert wurden und entsprechend editiert werden müssen) sofort der ursprünglichen Datenbank mit der gleichen Struktur zugeführt werden.

Die Abfrage muß man nicht mehr formulieren. Im Entscheidungsbaum sind die Antworten interpretiert und gefunden worden, Die Fragen sind durch die Algorithmen gestellt. Der Mensch hat die Angeboten Antworten selektiert und die passenden Antworten ausgesucht. Die Erstellung dieser Regelsätze ist dann anschließend eine Leichtigkeit von wenigen Tastendrucken und Minuten.

Wie lange hätte man mit herkömmlichen Mitteln hierfür benötigt. Oft reichen diese Regelsätze oder einfache Erkenntnisse von Zusammenhängen mehr als aus, um den notwendigen Fokus zu schaffen oder den wichtigen Hinweis auf der Suche nach einer Lösung zu bieten, wenn man auf der Suche nach Klärung in einem scheinbar nebulösen undeutlichen Fall.

Scoring von (auch) großen Datenmengen

Die meisten besseren Data Mining Werkzeuge sind in der Lage nach dem Aufbau eines Modells, diese enthaltenen Werte in der Form eines Scoring eine weitere – ungesehene – Datenmenge zu beschreiben. Dieses Scoring – oder manchmal trifft man auf den Begriff Ranking (engl. In Stufen einteilen) – kann zwei unterschiedliche Aufgabe haben.

1. *Die Datenmenge besteht aus einem Datensatz* – Man möchte die Information aus einer Dateneingabe – am Geldautomaten, bei Zugang zu einer Website usw. – klassifizieren. Der Antrieb im ersten Fall ist die Prüfung, ob diese Karte für die erforderliche Summe Bargeld gut ist und im zweiten möchten man den Kunden möglicherweise identifizieren, klassifizieren und damit ein passendes , individuelles Angebot online unterbreiten.

   Obwohl die Motivation unterschiedlich ist, wird aus der Data Mining Ecke beinahe identisch vorgegangen.

2. *Die Datenmenge besteht aus mehreren Datensätzen bis in die Milliarden* – Sie verfügen über eine immense Kundendatenbank (Amazon muß etwa 17 Millionen lt. Presse haben) oder eine Transaktionsdatenbank (die Deutsche Telekom verfügt im Festnetz über etwa. 250 Millionen Transaktionen täglich) und möchten eine gewisse Ordnung erzeugen.

   Amazon möchte seine Kunden vielleicht nach bibliothekarischen Gruppen oder Kaufwahrscheinlichkeit, -häufigkeit, Zahlungsform, Geographisch oder, oder ordnen. Das Projekt Kaufwahrscheinlichkeit könnte auf einem Data Mining Modell basieren. Nach dem Erstellen des Modell muß die ungesehenen Datenmenge (Kundendatenbank) mit dem Data Mining Modell verarbeitet werden.

   Genauso möchte die Telekom vielleicht wissen, welche Art Anrufe potentielle Betrüger darstellen, um Telefonbuden aufspüren zu können, bevor die Betreiber in alle Winde zerstreut sind und die entstandenen Verluste in die Millionen gehen.

   Auch hier geht es darum, die neu gesammelten Daten mit einem Data Mining Modell zu versehen.

In beiden Fällen gibt es mehrere Variationen.

Die einfachste Variation ist die Ordnung nach einem Regelsatz. Diese kann erfolgen (siehe oben - Regelsätze erzeugen) einem Beschreibungsmodell in verschiedenen Formaten. Als Vorhersagemodell kann z.B. ein XML Programm (siehe unten – XML/PMML) diese Arbeit leisten.

Die besseren Softwarepakete liefern ein Scoring Feature, das sogar auf externe Dateien anwendbar ist. Hinter diesem Feature wird aus dem bei der Erstellung des Modells angewendeten Algorithmus, die neue, ungesehene Datei mit den Scoring Informationen laut hinterlegtem Vorhersagemodell erweitert. Die Ordnung der Tabelle nach den Werten der neuen Scoring Spalte ist dann ein leichtes.

Die Methodologie sowie Beispiel Programme zum Scoring von externen Fremdformaten liegen vor und können leicht nach Bedarf in andere Sprachen und Format umgesetzt werden. Welche Größe Ihrer Datenmenge hat, ist egal. Ihre Datenverwaltungssoftware sowie Ihre Hardware muß lediglich in der Lage sein, diese große Menge zu halten.

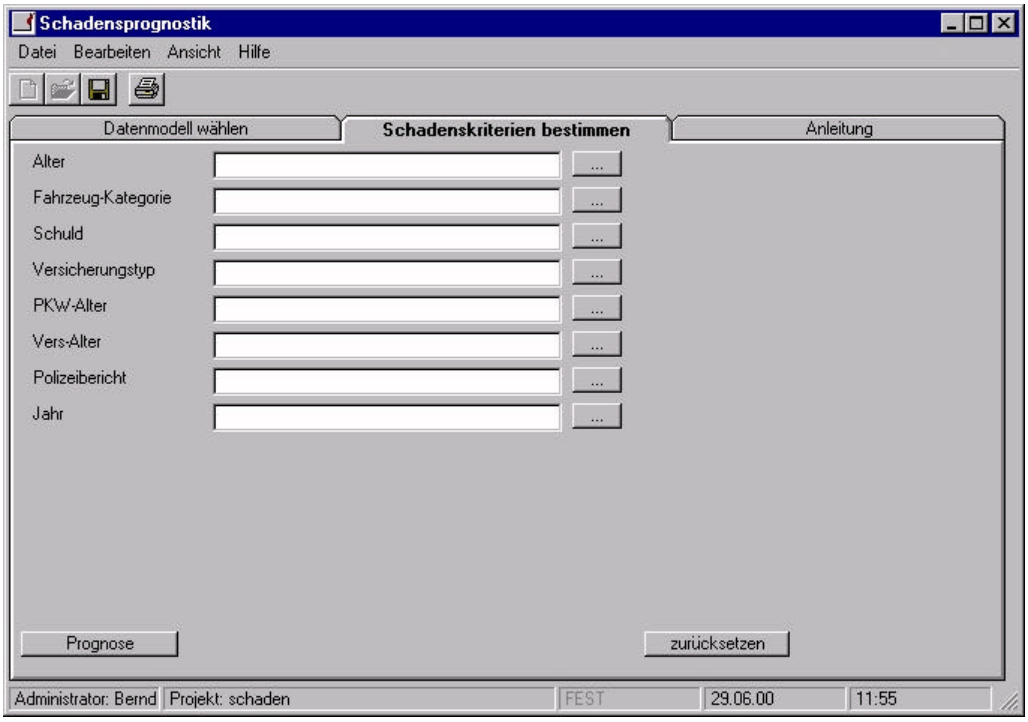
Stellen Sie sich vor, das ganze wird in einer Routine eingebunden und es läuft automatisch ab. Amazon braucht sich nur noch auf wenige Kunden bewußt konzentrieren, die am meisten, am häufigsten und am profitabelsten Ausgeben und eine hohe Kaufwahrscheinlichkeit aufweisen. Oder die Telekom erwischt die bösen Buben der „Telefonbuden“ prompt bzw. während das Verbrechen anläuft. Diese beiden Konzerne würden eine Menge Kosten einsparen und größere Umsätze erzeugen bzw. Verluste in Millionenhöhe umschiffen. Und das ständig.
Jedes Mal wenn Sie Bargeld von einem Geldautomaten holen, überlegen Sie was die Maschine eigentlich tut, bevor sie anfängt zu rattern und Ihr Geld auszuzählen und bereitzustellen. Es dürfte ein solches Scoring sein, das anhand der Informationen auf dem Magnetstreifen Ihrer Karte sowie Ihrer Eingaben plus Erweiterungen oder Transformationen zu einem Scoring geführt haben!

Data Mining gesteuerte Applikationen

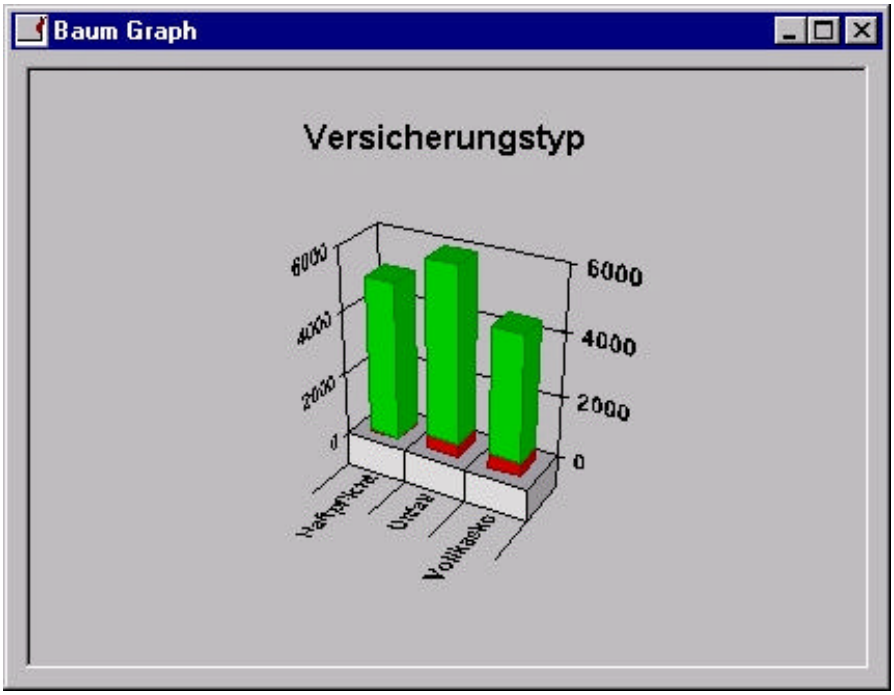
Die folgende Bildausschnitte stammen aus einer Applikation, die in Visual Basic auf der Basis von ANGOSS KnowledgeSTUDIO SDK und einem Beschreibungsmodell in Form eines Entscheidungsbaumes als Beispiel geschrieben wurde, in der Schadensabteilung einer Kfz-Versicherung eingesetzt werden soll.

Hier soll in wenigen Sekunden – direkt vom Schadenssachbearbeiter – jede Schadensmeldung vorab auf Betrugswahrscheinlichkeit überprüft werden (Mehr zum Hintergrund des Szenarios lesen im Anhang unter Data Mining Challenge).

Beim Start des Programms werden einige Informationen vom Sacharbeiter verlangt, die er von der jeweiligen Schadensmeldung ablesen kann.

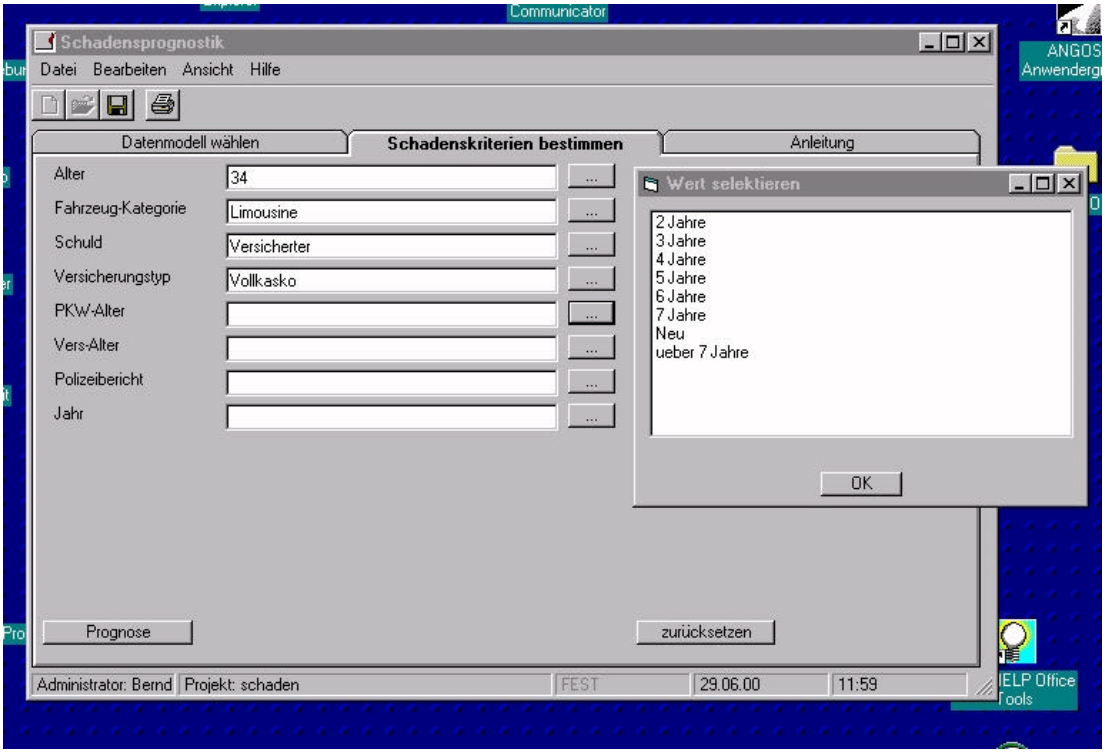


Die wenigen Informationen sind von Data Miner im Entscheidungsbaum als entscheidenden Indikatoren ermittelt worden. Im Hintergrund lagert der Baum und KnowledgeSTUDIO treibt den Vergleich dieser Eingaben (als einziger Datensatz der Datenmenge) mit dem hinterlegten Entscheidungsbaum.



Ein Administrator – keine Sachbearbeiter der Schadensabteilung, sondern ein Data Miner sieht in der oberen Grafik, das die Betrugsfälle (rot) im Bereich Unfall und Vollkasko (mittlere und rechte Säule) am deutlichsten vorkommen.

Der Sachbearbeiter füllt lediglich seine Felder mit einem Mausklick auf dem jeweiligen Feldbutton und einer Auswahl aus dem jeweils aufgeblendeten Angebot und zum Schluß betätigt er den Button „Prognose“ unten links.



Die Antwort erfolgt in Form eines Dialogs „Die Betrugswahrscheinlichkeit beträgt X% oder die Information einer Warnung erscheint. Sonst kann der Sachbearbeiter normal weiterarbeiten, Erhält er eine vordefinierte hohe Wahrscheinlichkeit oder die Warnung erscheint, muß er den Fall weitergeben. Protokolle werden zur Kontrolle im Hintergrund geschrieben und später ebenfalls ausgewertet.

Eine solche Applikation kostet wenige Sekunden und kann in wenigen Minuten installiert und die Mitarbeiter damit vertraut machen. Die möglichen Ersparnisse gehen in die Millionen – jährlich.

Andere Applikationen haben wir im Kapitel „10. Hauptprojekt – Welche Vision, welches Szenario“ aufgeführt. In jedem Fall werden hier einfache Datensätze mit einem Modell an der Front von Sachbearbeitern verarbeitet. Was Data Mining überhaupt ist, braucht dieser Sachbearbeiter gar nicht zu wissen. Er betreibt ein kleines Windows Programm.

Solche Applikationen können unterschiedlich gestaltet sein. Der Grad der Automation hängt vom Szenario ab. Die Überwachung einer Produktion in der Herstellung, in der Telekom-Technik oder in jedem Leitstand Szenario erlauben eher einen hohen Grad der Automation. Die Bearbeitung individueller Fälle wie oben, in der Kreditprüfung, in der Spesenabrechnung, bei der optimalen Auslastung eines Flugzeuges, Hotels, Routenplanung einer Spedition wie FedEx usw. bedürfen eher die Einwirkung eines Menschen. Trotz allem kann ein solches Programm so einfach zu bedienen sein, dass die Macht des Data Mining sicher im Hintergrund bleiben kann und die Erzeugung des Mehrwertes ohne zusätzliches Probleme zu erzeugen, an der Front ungehindert und im maximalen Maße erfolgen kann.

XML

„Extensible Markup Language, abgekürzt XML, beschreibt eine Klasse Datenobjekte, die XML-Dokumente heißen, und beschreibt zum Teil die Arbeitsweise von Computerprogrammen, die sie verarbeiten.“

Eine Markup Sprache ist lediglich ein Mechanismus zur Identifizierung von Strukturen in einem Dokument. Die XML Spezifikation definiert einen Standard, wonach die Struktur (oder Markup) Dokumenten zugeführt werden kann.

XML is ein Applikationsprofil oder eingeschränkte Form des SGML, the Standard Generalized Markup Language ISO 8879. Im Aufbau Format entsprechen XML Documente auch SGML Dokumente.

XML Dokumente bestehen aus Speichereinheiten, die entweder zergliederte oder nicht zergliederte Daten enthalten. Zergliederte Daten bestehen aus Charakteren, die teilweise Charakterdaten bilden und andere, die zur Struktur (oder Markup) gehören. Der Markup codiert eine Beschreibung des Speicherformats und der logischen Struktur des Dokumentes. XML bietet ein Mechanismus zur Beschränkung dieser beiden Punkte.

Strukturierte Information enthält Inhalte (ie Wörter, Bilder usw.) und einen Ansatz darüber, welche Rolle welcher Inhalt zu spielen hat. Zum Beispiel der Inhalt einer Überschrift hat eine andere Bedeutung als der Inhalt einer Fußnote oder einer Bildbeschreibung oder Inhalt einer DB-Tabelle. Fast alle Dokumente haben irgendeine Struktur.

Ursprung und Ziele

XML wurde durch eine XML Arbeitsgruppe (ursprünglich als SGML Editorial Review Board bekannt) entwickelt, die 1996 durch den World Wide Web Consortium (W3C) gegründet wurde.

Die Ziele für XML sind:

1. XML soll einfach im Internet einsetzbar sein.
2. XML soll eine große Auswahl Applikationen unterstützen.
3. XML soll SGML kompatibel sein.
4. Die Erstellung von Programmen zur Verarbeitung von XML Dokumenten soll einfach sein.
5. Die Anzahl optionaler Features in XML soll auf einem Minimum gehalten werden, im Idealfall bei Null.
6. XML Dokumente sollen vom Menschen lesbar und deutlich sein.
7. Das XML Design soll schnell vorbereitet werden.
8. Das XML Design soll formell und kurz sein.
9. XML Dokumente soll einfach in der Erstellung sein.
10. Die Kürze in der XML Struktur ist von minimaler Wichtigkeit.

Der exakte Grund für die Entwicklung von XML liegt darin, Dokumente mit einer umfangreichen Struktur im Web verwenden zu können. Die einzigen Alternativen waren HTML und SGML, die zu diesem Zweck nicht geeignet sind.

Der W3C in Zusammenarbeit mit Browser Hersteller und der WWW Gemeinde arbeitet ständig daran, die Definition von HTML mit neuen Möglichkeiten zu erweitern, um mit der aufkommenden Technologie Schritt zu halten und weitere Darstellungsvarianten zu bieten. Die Änderungen werden durch die Beschränkungen der Browsertechnologie behindert, denn eine Rückwärtskompatibilität ist vordergründig. Diejenigen, die jedoch Informationen weit verbreiten möchten, kommen mit den neuesten Versionen von Netscape oder Internet Explorer nicht weiter.

XML ist in Wirklichkeit eine Metasprache, zur Beschreibung von Strukturen – also Markup Sprachen. XML bietet eine Grundlage Tags und ihre strukturierten Verbindungen zu definieren. Da es keine vordefinierten Tags gibt, gibt es ebenfalls keine vorgefaßte Semantik. Jegliche Semantik eines XML Dokumentes wird durch die Applikationen definiert, die ihn verarbeitet oder durch angehängte Stilblätter.

XML und Data Mining

Nachdem Sie die letzten Seiten zum Thema XML verdaut haben, fragen Sie sich langsam, was XML mit Data Mining zu tun hat. XML definiert eine Strukturform innerhalb der Internet Technologie. Das heißt, der Data Miner hat eine Grundlage, um seine Vorhersagemodelle im Internet einfach (lt. Definition der XML Ziele) unterbringen zu können.

Im Sinne der Data Mining Anforderungen wurde u.a. XML erzeugt. Das Ziel dieser Arbeit war es, eine Anbindung an der W3C (World Wide Web Consortium) zu schaffen und damit die Anzahl Mitglieder dahingehend auszubauen, dass alle wichtige Anbieter von Data Mining Werkzeugen und Applikationen dazugehören.

Ein Ziel von XML ist es Applikationen, online analytische Verarbeitungstools bis hin zu Modellen, die aus mehreren Quellen stammen, zu erlauben, ohne die einzelnen Unterschiede zwischen diesen Quellen berücksichtigen zu müssen.

Ein weiteres Ziel ist der kombinierte Einsatz von einer großen Anzahl einzelner Modelle sowie Ansammlungen von Modellen, die auf geschäftliche Anforderungen sowie mathematische Prinzipien fußen.

Die wichtigsten Wörter hier sind *online*, *analytisch* sowie *kombinierter Einsatz*. Wir erinnern uns die Kerneigenschaften des Web Mining. Online als Geschwindigkeitsanforderung war ein Thema. Ferner ging es um Menge der Daten. Eine ungeahnte Steigerung der Intelligenz von Web Mining ist mit der relativ einfach zu bewerkstelligenden Kombination von Vorhersagemodellen greifbar.

Diese Fähigkeiten sind geradezu grundsätzlich beim effektiven Einsatz von Data Mining Modellen in einem kommerziellen Umfeld im Internet. XML erfüllt diese dringenden und dramatisch gestiegenen Bedürfnisse der Geschäftswelt.

Ein XML Dokument bietet eine Definition von parametrisierten Modellen sowie genügend Steuerungsinformation, damit sie in einer Applikation eingesetzt werden können.

Unter Verwendung eines XML Parsers kann die Applikation alle Input- und Outputdaten der Modelle, die detaillierte Formen der Modelle feststellen, und wie - nach den Standards der Data Mining Terminologie – deren Ergebnisse zu verstehen sind.

Ein Beispiel eines XML Programms finden Sie in der Anlage III.

Moderne Data Mining Tools werden in der Lage sein, solche Codes wie „Regression Model“ oder „Tree Model“ oben auf der Basis eines Vorhersagemodels erstellen zu können. Die Einbindung in ein solches Programm ist dann ein Leichtes. Letztlich wird alles mit über einem XML Parser durch die HTML/SGML Fähigkeiten des Internet ausgeführt.

In Einzelheiten werden sich die Modellformen nach Typ unterscheiden, aber sie sind letztlich alle Textdefinitionen. In geparster Form werden sie genügend Information liefern, um damit ein Programm zu generieren oder eine interpretive Ausführung des Modells auf der Basis eines Parse-Baums durchführen lassen.

Die Version 1.0 des Standards bietet einen kleinen Befehlssatz von DTDs, die Einheiten und Attribute zur Dokumentation von Entscheidungsbäumen und logistische Regressionsmodellen spezifizieren können.

Dies ist in keinem Fall vollständig, jedoch soll hiermit einen Anfang gemacht werden, worauf später gebaut werden kann. Die XML-Gruppe arbeitet ständig weiter am Ausbau der Funktionalität der Sprache. Die Grundsätze von XML wären jedoch damit demonstriert. Es läßt sich leicht und realistisch vorstellen, welche umfangreiche und reichhaltige Modellfähigkeiten aufkommen kann.

Bewertung der Modelle

Nachdem man sich die Mühe gemacht hat, Modelle im Data Mining zu erstellen, möchte man natürlich wissen, was die Modelle wert sind. Eine erste Möglichkeit finden wir in der Validierung.

Mit der Validierung können wir alle in der Erstellung des Vorhersagemodells verwendeten Variablen mit dem Ergebnis dieser Vorhersage, den Vergleich zum bekannten Wert (Inhalt wahr oder falsch) und dem mathematischen Sicherheitswert in eine neue Datenmenge geben. Diese Datenmenge kann man über die abhängige Variable „Wahr/Falsch“ zum Beispiel in einem Entscheidungsbaum auswerten. Man sieht deutlich, welche Wirkung diverse unabhängige Variablen auf die Werte des abhängigen Variablen ausübt. Gleichzeitig kann man den schlimmsten Fall kontrollieren, dass das System die gleiche Aussage bei jeder Vorhersage vorgenommen hat. Dies kann normalerweise kaum in der Tat sein.

Nach der Validierung möchten man vielleicht Modelle nicht nur bewerten, sondern untereinander vergleichen. Um dies machen zu können, möchten wir das Konzept Lift einführen.

LIFT ist ein Verhältnis, das am häufigsten zum Vergleich von Klassifizierungsmodellen verwendet wird. In der Tat wird die Veränderung in der Konzentration einer bestimmten Klasse gemessen, wenn das Modell zur Selektion eines gestellten Musters aus der Gesamtheit der Datenmenge verwendet wird.

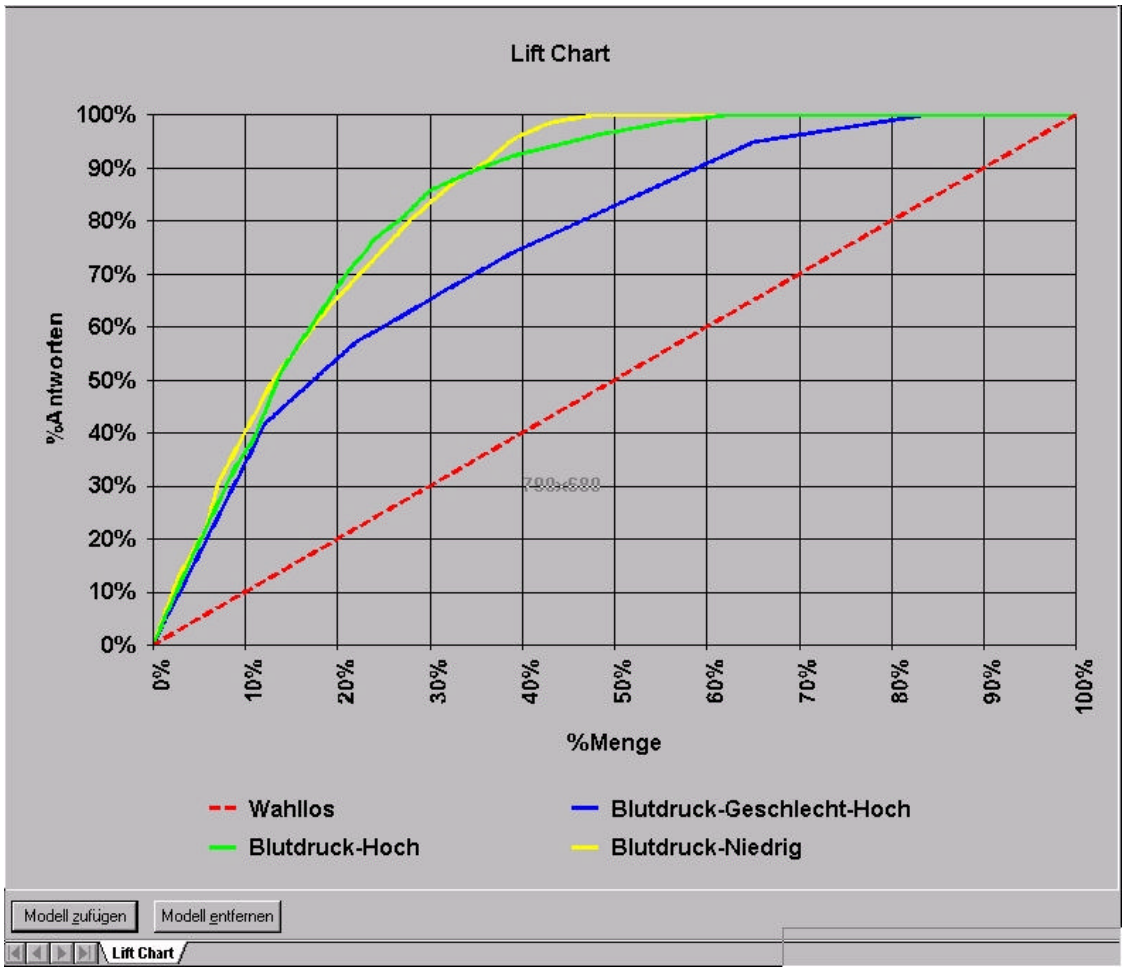
Lift = P(Klasse, | Muster) / P(Klasse, | Gesamtheit)

Wir bedienen uns eines Beispiels aus dem sogenannten Direct Marketing, woher diese Bezeichnung ursprünglich stammt. Wir haben ein Modell gebaut, wo die Wahrscheinlichkeit einer Antwort auf ein Mailing gemessen werden soll. Aus Erfahrungsdaten klassifizieren wir die Datensätze nach Antwort „Ja“ oder „Nein.“ Hieraus bauen wir ein Modell. Nun können wir den Liftwert errechnen.

Die Gruppe mit der Markierung „Ja“ sollte eine höhere Proportion von tatsächlichen Antworten erhalten als die Gesamtmenge. Die Gruppe mit der Markierung „Ja“ wird unser gestelltes Muster. Wenn die Evaluierungsmenge 10% tatsächliche Antworten und das Muster 50% tatsächliche Antworten enthält, bietet unser Model einen Liftwert von 50 / 10 = 5.

Sollte allerdings die Klassifizierung 5% tatsächliche Antworten finden und dies in 80% der Fälle erfolgreich zutreffen, haben wir eine Lift von 80 / 5 = 16. Ob wir dadurch ein besseres Modell haben, muß noch untersucht werden. Denn wenn man diese Beispiel fortführt und eine Mailingliste in Betracht zieht, wird der Liftwert sinken, je länger die Mailingliste wird. Mit anderen Worten je mehr Personen angeschrieben werden, je mehr sinkt der Liftwert.

Dies hängt damit zusammen, dass die Berechnung des Lift eine Funktion der Mustergröße ist. Die folgende Grafik – Lift Chart genannt – stellt in der Tat die kumulative Antwort dar.



Bei unserem früheren Beispiel der Blutdruckwerte sehen die Wirkung unseres Modells im Falle hohen Blutdrucks (grün) und niedrigen Blutdrucks (gelb).
Die rote gestrichelte Linie stellt die Auswahl nach dem Zufallsprinzip her. Jede Modellkurve, die unter dieser Linie liegt, benötigen wir nicht, denn eine zufällige Selektion würde bessere Ergebnisse liefern.

In dieser Grafik sehen wir, dass wir nach etwa die Hälfte der Datenmenge eine vollständige Trefferquote erreiche, wobei die gelbe Linie früher die Vollständigkeit erreicht. Der zweite Baum der den hohen Blutdruck auch klassifiziert (blaue Linie), erreicht die Vollständigkeit erst bei über 80% der Gesamtmenge. Es scheint deutlich, dass die grüne Linie erfolgreicher die Indikatoren des hohen Blutdruckes als das Modell mit der blaue Linie gefunden hat.

Wackelige Kurven zeugen meist von unfertigen Modelle. Kurven, die plötzlich abfallen und wieder steigen, enthalten Widersprüche, die solche Modelle nutzlos machen.

Mit solchen Lift Charts kann man die Leistung verschiedener Modelle im Vergleich zueinander hervorragend messen. Im Falle die abhängige Variable besitzt fortlaufende Werte, wird diese Arbeit in einem Genauigkeitsmodell errechnet. Aber ob im Sinne des Geschäfts die Erstellung ein solches Modell sich lohnt, d.h. einen wirtschaftlichen Sinn ergibt, ist hiermit nicht beantwortet.

Die Antwort auf diese Frage gibt es nur wenn Sie das geschäftliche Vorhaben damit vergleichen. Beim Beispiel des Mailings kann es sein, dass der mögliche Gewinn durch den Aufbau eines Klassifizierungsmodells mehr kostet in Zeit und Geld, als ohne Modell direkt mit einem Massenmailing loszulegen. Im Falle der Betrugsaufdeckung im anderen Extremfall ergibt die Erstellung des Modells wahrscheinlich einen deutlich höheren Gewinn als wenn man gar nichts macht.
In jedem Fall sollen die Kosten eines Modellaufbaus zum möglichen Gewinn des Projektes gesetzt werden. Fällt der Gewinn nicht höher aus, als die Kosten und/oder höher als gar nichts zu tun (abhängig vom Szenario), kann man sich die ganze Mühe des Data Minings sparen.

19. Mystik des Data Mining

Während der nunmehr 8 Jahre seit dem der Autor sich mit Data Mining beschäftigt – aus einer Zeit bevor der Begriff Data Mining in Mode gekommen war – sind einige Vorstellungen über Data Mining aufgekommen, die sich wirklich nur noch mit den Märchen zur Zeit der Eisenbahnpioniere vergleichen lassen.

Damals hieß es, der Mensch können keine 50Km/h aushalten. Er würde den Druck nicht standhalten und ersticken. Oder Kühe würden mehrköpfige Kälber zur Welt bringen. Dieses schnaufende Ungeheuer (die Dampflok) wäre schließlich aus der Hölle emporgekommen und alles böses Teufelswerk.

Nun Data Mining als Teufelswerk bezeichnet, wäre neu und schwer haltbar, aber einige Äußerungen gehen gewaltig an die Substanz. Zum Beispiel nach einer kurzen Erläuterung über Data Mining – „das kann ich doch alles mit meiner Tabellenkalkulation.“ Oder Data Mining wird als „Analyseprogramm“ abgewertet und es folgt: „Das programmiere ich mir lieber selber.“ Man merke, was nun programmiert werden soll, wird nicht mitgeteilt.

Datenbank Mensch zweifeln am Verstand und am Nutzwert des Data Miners, der behauptet, er können Antworten von Datenmengen erhalten, ohne Fragen formulieren zu müssen. Der Datenbankler weiß nach dem EVA (Eingabe Verarbeitung Ausgabe) Prinzip, dass seine Abfragen die notwendige Eingabe des Grundprinzips der Datenverarbeitung erfüllen und jetzt bekommt er es regelrecht mit dem Mitleid zu tun, wenn der Data Miner scheinbar diese Prinzip entgegen spricht.

Auf den Hinweis, man muß doch Fragen stellen, behauptet der Data Miner, diese Fragen algorithmisch stellen zu lassen. Da diese Algorithmen für den Datenbankler verschlossen sind, beginnt er den Data Miner bestenfalls mit argwöhnischen Augen zu betrachten. In der Regel sehen die Datenbankler Data Miner als Spinner oder Spieler an, die eigentlich keine Ahnung von der echten Welt der reellen Informationstechnologie haben. Schließlich müssen Daten erfaßt, gepflegt und intelligent zur Verfügung gestellt werden. Das sind schließlich richtige IT-Leute und die heißen Datenbankler,

Ob man diese Meinung teilt oder nicht, sie baut deutlich auf Unverständnis und werden oft nach dem meist späten Einsehen mit einer gewaltigen Portion Angst gelenkt.

Gründe der Mystik

Unwissen und Angst paaren sich zum Aufbau einer mystischen Umhüllung von allen Dingen, die wir schwer oder nicht begreifen oder gar nicht erklären können. Data Mining gehört für viele Menschen in diese Rubrik.

Obendrein arbeitet Data Mining mit vielen „Black Boxes." Manche Anbieter von Data Mining pflegen und steigern dieses Image und wollen Ihr spezielles Angebot durch die Beibehaltung des Blackbox schützen.

Man kann jedes Blackbox durch eingehende Erklärung öffnen. Alleine den Vorschlag reicht den meisten Managern. Sie verstehen die Blackbox als Motor in einem Vehikel, das sie fahren möchten, ohne direkt wissen zu müssen, wie dieser spezieller Motor im Detail funktioniert.

Leider glauben – was nicht wissen bedeutet – zu viele Menschen, dass es so was wie Data Mining gar nicht gibt. Es ist in der Tat meist eine „backroom" Disziplin, das im Hintergrund von einigen Eierköpfen praktiziert wird und daher kaum in der Öffentlichkeit frei zu erleben. Ferner werden diese Mechanismen oft in kritischen Bereichen eingesetzt, was ebenfalls nicht gerade dazu führt, das der Einsatz publik wird. Die Ergebnisse der Arbeit kommen zum Einsatz und der Weg dorthin bleibt im Verborgenen.

Hinzu kommt die virtuelle Natur von Data Mining, das für dieses Erscheinungsbild einen nebulösen und noch schwerer verständlichen Rahmen bildet. Damit ist die Mystik längst entstanden. Für manche Menschen wirkt dies geradezu wie einen Killervirus der Art Krebs oder AIDS oder gar aus Hollywood, den sie um sich wähnen, aber schutzlos gegenüber zu stehen glauben.

Solche Reaktionen äußern sich meist in Skepsis. Erklärungen der höchsten Institutionen der Welt nützen hier nichts und helfen keineswegs ein wirkliches Verständnis aufzubauen. Kurzum man kann reden was man will, die Lage wird dadurch nicht gebessert.

Lediglich der praktische Umgang mit Data Mining kann den Skeptiker zum rechten Glauben bekehren bzw. den Verständnis vermitteln. Man kann während einiger Stunden beobachten, wie der Ungläubiger seine Bekehrung vollzieht. Auf Skepsis folgt Staunen. Auf Staunen folgen Fragen. Da auf die Fragen Antworten folgen, beginnt der Ungläubiger ein Wissensgerüst im Rahmen seiner Akzeptanz zu errichten. Es kommen weitere Phasen des Staunens und des Fragens, bis aus dem Wissensgerüst langsam ein Gebäude errichtet wird. Dann folgt endlich die Motivation durch die Erkenntnis der bisher ungeahnten Möglichkeiten des Data Minings.

Es darf schließlich das nicht sein, was nicht ist. Heißt hier, es darf das nicht sein, was ich nicht verstehe (Voltaire möge dies verzeihen). Wird die Situation durch die erlebte Praxis umgekehrt, dann heißt es, was ich verstehe, setze ich ein.

Skurriles aus dem Fachbereich

In jedem Fachbereich gibt es nette Geschichte und Anekdoten. Wir möchten hier einige eher skurrile Erlebnisse wiedergeben.

Im Anfang hat man versucht, möglichst viele Kontakte zur Anwendung dieses wunderbare „Analyse-Tools“ zu finden. Ein gewisser Systemadminstrator eines großes Umweltunternehmens wollte de Autor – der gerade zum vereinbarten Termin erschienen war – doch nicht empfangen. Ein Ausnahmebeispiel an Unhöflichkeit. Auf Verlangen erschien der man, jedoch und teilte kurzerhand mit, dass fremde Personen nicht an seine Daten herankommen.

Er stellte sich heraus, dass niemand in dem Unternehmen recht wußte, irgendwas mit dem Thema anzufangen. Da es mit Computern zu tun hatte, landete diese heiße Kartoffel bei den Systemleuten. Besagter Mann schien zu verstehen, das wir ein Wunderwerkzeug besaßen, das seine Daten nach Redundanzen untersuchen sollte, um sie sofort zu vernichten. Dies wollte er mit aller Macht verhindern und die Ehre seiner Leute und die Unversehrtheit seiner Daten bis zum Ende verteidigen. Bis heute ist es ungewiß wie er zu dieser Annahme gekommen ist.

Bei der Polizei kam man nach einigen Gesprächen mit teilweise hochrangigen Menschen zum Ergebnis, die Untersuchung der kriminalstatistischen Daten mit Data Mining Werkzeugen, könnte ungeahnte Vorteile bringen. Vor allem erwartete man einen immensen Zeitgewinn bei der Erkenntnis zusammenhängender Informationen zur Bekämpfung der geplanten bis organisierten Kriminalität.

Ferner sollten Stellen der öffentlichen Vertretung wie Stadträte, Landräte, Landestage usw. mit fundierten und schnellen Informationen zu den wiederkehrenden politischen Problempunkte wir Jugendkriminalität, Kriminalität gegen Frauen und Kinder sowie Drogen- und Ausländerkriminalität.

Bei einem bestimmten Polizeipräsidium wurde ein gestandener Kripobeamter um die 50 dazu auserkoren, diese Thematik zu betreiben und sich in die Geheimnisse des Data Minings einarbeiten zu lassen. Dieser Mann hatte scheinbar keine größeren Karrieresprünge vor sich und richtete sich auf einer möglichst ruhigen Fortführung seiner Dienstzeit bis zur Pension ein. Nun plötzlich Data Mining. Er hatte deutlich Angst.

Nach einem Tag in der Einweisung war dieser beruflich abgeschaltete, gestandene Mann zu einem voll motivierten Data Miner mutiert, der mit Wonne seine neue Arbeit

verrichtete und immer wieder Kollegen mit seinen erstaunlichen Fähigkeiten verblüffen konnte.

Einweiteres bemerkenswertes Ereignis passierte während der Gespräche zum Beginn eines Pilotprojektes. Unsere Hilfe wurde angefordert, da das Top-Management eines großen Konzernes wissen wollte, welche Fragen sie nun auf der Basis stellen könne. Die befragte Stelle meinte: „Wir wissen nicht, wie wir diese Fragen beantworten sollen, können Sie uns helfen?“

Selten ist eine solche präzise Anforderung zu Beginn einer Zusammenarbeit formuliert worden. Meist verstecken sich die Leute hinter Ihrem Unwissen. Dabei ist es nur selbstverständlich, dass man einen solche Frage ohne Erfahrung gar nicht beantworten kann.

Vielleicht die skurrilsten Geschichten stammen aus dem deutschen Bankwesen. Die Herren dort bilden eine eigene Gemeinde scheinbar mit eigenen Riten und Terminologie. Die Chance auch in Deutschland mögliche Betrügereien z.B. mit Plastikgeld aufzuspüren, schien reell zu sein, da im Ausland solche Erfolge bereits verbucht werden konnten und die Szenarien bekannt waren. In Deutschland reichte die Reaktion von Null bis zur Aussage, das machen wir selbst. Es stellte sich nach einige Jahren heraus, dass die Terminologie verkehrt war. „Betrugsaufdeckung“ kann man bei deutschen Banken nicht betreiben, denn dort gibt es keinen Betrug. Zumindest auf diesem Begriff wird so allergisch reagiert, dass man das Wort nicht hört. Das Thema heißt „Risikomanagement“ und mit Risiko gehen die Herrschaften seit Jahren erfolgreich um. Klinkt nach einem ziemlichen „closed shop.“

Peu á peu setzt sich Data Mining durch. Was manche Anbieter Firmen damit machen, sehen wir im nächsten Abschnitt.

Ein großes Märchen

„Software that can think:“ („Software, die denken kann“). Haben Sie jemals so was fragwürdiges gehört? Es stellt sich heraus, das der Hintergrund dieses Werbeslogans eine Produktpalette eines angesehenen, weltweit agierenden Software und Dienstleistungsunternehmens im Data Mining zu finden ist.

Egal wie super die Regelsätze, oder astronomisch die Algorithmen gar weltbewegend die neuronalen Fähigkeiten dieses Angebots, seit wann kann Software, die auf herkömmlichen Computern läuft, denken. Wir hier der Begriff des Denkens nicht etwas umgeschrieben?

Da dieser Spruch scheinbar im englisch sprechenden Ausland erzeugt wurde, bedienen wir uns im ehrwürdigen englischen Wörterbuch des Oxford Dictionary zum Thema „think“ (dt. denken). Da heißt es:

1. Überlegen, Meinung haben
2. Vorhaben, Erwarten
3. Vorstellung von etwas bilden
4. Dasein von etwas erkennen
5. Auf spezifizierte Bedingung reduzieren
6. Gehirn anders benutzen, als nur passiv die Ideen anderer anzunehmen
7. Halbes Vorhaben besitzen
8. Die Praktikabilität von etwas ausloten

Soweit der Concise Oxford Dictionary zum sprachlichen Thema „think“ (dt. denken). Welche heutige Software soll nun in der Lage sein, überhaupt eine dieser Teilbedeutung eigenständig ausführen zu können. Ohne den Input des Menschen – ob automatisch oder manuell ist an dieser Stelle irrelevant – sind Data Mining Werkzeuge hilflos. Die Interpretation der Ergebnisse ist ohne einen menschlichen Fachmann unmöglich. Auf der Basis der Interpretation eines Menschen – bei Aufbau eines beschreibenden Modells – kann es zum Vorhersagemodell weitergehen.

Bei der Betrachtung des Ablaufes vom Einsatz Data Mining Werkzeuge ist die Unterstützung durch den Menschen mit seinem Denkvermögen überall und vor allem an entscheidenden und wesentlichen Punkten des Vorganges unerläßlich.

Es sind zwar intelligente Werkzeuge, aber von dieser Art Intelligenz zum Denken lt. obiger Beschreibung ist es noch weit.

Wohlgemerkt, die Kritik hier ist keineswegs moralisch oder sonstwie geistig gemeint, sondern eine handfeste Vorstellung dessen was denken überhaupt ist und das Data Mining diese Bedingungen nicht erfüllt.

Teilt man diese Meinung, dass Software nicht denken kann, muß man zum Ergebnis kommen, mit welcher Macht die Menschen hier für dumm verkauft werden sollen.

Eine TV-Werbekampagne – N.B. im Massenmedium Fernsehen – will uns diese Botschaft vermitteln und scheinbar reagiert niemand. Entweder hört keiner zu oder die Botschaft kommt aus irgendeinem anderen Grund nicht an. Denn eigentlich müßten die Mensch euphorisch gar panisch reagieren, wenn es auf einem Mal denkende Software gibt. Es müßte eine heftige Reaktion wie die berühmte Orson Wells Radioübertragung, wo es um Außerirdische Besucher ging und es eine Massenflucht aus der Stadt gab. Aber nichts!

Vielleicht haben die Menschen eine Antenne für diesen Unsinn und ignorieren solche Sprüche. Jedenfalls möchte ich nicht in der Lage der Berater stehen, die diese Aussage vertreten müssen. Sie werden in die Not der Rechtfertigung kommen oder alles als Werbespruch abwälzen müssen. Im ersten Fall werden Sie es schwer haben

und im zweiten Fall einen Imageverlust erleiden; denn dies ist ein billiger (obwohl aus der Sicht der Werbekampagne einen recht teuerer) Trick. Dies ist ein Märchen.

20. Zusammenfassung

Während der Entstehung dieses Werkes sind bereits einige wichtigen Komponenten für ein Nachfolgewerk aufgeworfen worden. Aus Gründen der Übersicht bei diesem Werk und aus den praktischen Anforderungen der zeitigen Erscheinung dieses Opus haben wir uns entschieden, weitere Punkte für eine weitere Ausgabe aufzuheben. In einigen Fällen werden wir in einem Nachfolgewerk einige Modernisierungen vornehmen müssen, da die Entwicklung droht, uns hier zu überholen.

Insbesondere haben wir hier versucht, die wichtigsten Fallen sowie einige Freuden der Projektarbeit aus dem Feld des Data Minings herauszustellen.

Die größte Falle

Die größte Falle – haben wir festgestellt – ist die Annahme, dass ein Student - egal wie fähig, ob Diplomarbeit oder Promotion – in der Regel nicht in der Lage sein kann, die praktische Einsatzfähigkeit von Data Mining Tools für irgendein Unternehmen erfolgreich zu untersuchen. Die Praxis eines Unternehmens mag der Student verstehen und berücksichtigen können, jedoch die praktischen Möglichkeiten des Data Minings sind noch nicht Bestandteil seines Wissensspektrums.

Als analoges Beispiel würde man kaum einen Studenten der Automobiltechnik dafür einsetzen, den Wirkungsgrad eines Formel Eins Automobils nach den unterschiedlichen Fähigkeiten der verfügbaren Fahrer und Mannschaften zu untersuchen! Die Werbe- oder Marketingwirksamkeit dieses Unterfangens für ein jeweiliges Unternehmen wird man wohl auch keinem Studenten überlassen.

Beide Beispiele sollen nicht als Plädoyer dafür dienen, Studenten keine wichtigen Aufgaben zu übertragen. Im Gegenteil können Studenten meist für kleines Geld wichtige Indizien für viele Unternehmen ausarbeiten und gleichzeitig eine essenzielle Integration mit der Wirtschaft leben. Jedoch können solche strategischen Entscheidungen wie der Einsatz von Data Mining im Unternehmen höchstens durch Studenten sinnvoll unterstützt werden.

Die Vorbereitung der Entscheidung, ob Data Mining oder nicht, müssen durch professionelle Praxis vorgelebt und ausgewertet werden. Denn Data Mining sowie seine Werkzeuge sind nur so wertvoll wie ein Formel Eins Auto. Die Erfolge müssen durch Mannschaften und äußerst fähige Fahrer eingefahren werden. Dies trifft auch für Data Mining zu.

Unser Ziel war es diese Praxis von Data Mining herauszustellen und betrachten das obige Plädoyer als letzte Falle der Praxisarbeit. Die Auswertung eines Tools nach den

akademischen Vorstellungen eines Studenten, kann wohl kaum die Stolperdrähte der Praxis und die Anlehnung an die Werkzeuge überblicken. Zum Beispiel, wie wichtig ist das Tool der Association Rules, wenn man auf anderem Wege vielleicht leichter und schneller zum Ziel kommt?

Die Freude

Data Mining ist eine Disziplin, die auf vielen Werkzeugen zurückgreifen kann. Aus der Sicht der praktischen Arbeit ist es nicht entscheidend, welche Anzahl Werkzeuge oder Algorithmen zur Verfügung stehen, sondern eher die Fähigkeit umfassende Beschreibungs- und Vorhersagemodelle erstellen zu können sowie die notwendige Flexibilität zu bieten, die eine Systemwelt und Datenwelt verlangen könnte. Ferner sollten alle möglichem Wege offenstehen, gefundene Modelle in den erfolgreichen, Mehrwert erzeugenden Einsatz bringen zu können – wieder unter Berücksichtigung aller Erfordernisse der Betriebssysteme, des Datenumfeldes (Größe der Datenmenge und Komplexität) sowie in der Hauptsache des größtmöglichen Erfolges.

Der Erfolg wird in vielerlei Art gemessen, aber in der Tat wird der Gewinn in einer einfachen Formel gemessen. Data Mining Projekt sind erst zu genehmigen solange:-

Mehrwert – Kosten > Ist-Zustand

Wenn der Mehrwert eines Data Mining Projektes weniger geplanter Kosten keine meßbare Verbesserung gegenüber den aktuellen Zustand bildet, braucht man das Projekt gar nicht zu starten.

Hiernach ist der eigentliche Erfolg Sache der Meßlatte. Wie hoch soll der Mehrwert sein? Betrugsaufdeckung Szenarien bilden 15-20% Gewinn. Direct Mailing Mehrwert wird entweder eine höhere Anzahl Reaktionen als bisher, eine geringere Größe von Sendungen sowie einen höheren Prozentsatz an Reaktionen als bisher. Der Fokus auf die gewinnbringenden bzw. wahrscheinlichen Reaktionen erzeugt schließlich diese geringeren Kosten (der Streuverlust muß nicht aufgebracht werden) und/oder die erhöhte Trefferquote.

Egal welches Szenario sollte der Gewinn schließlich an erste Stelle stehen. Sollte der Gewinn in der Zeitfrage liegen, wird er schwer zu messen sein. Jedoch liegt er zwischen „nice to have“ und eine deutliche Konzentration der Kräfte durch früher erkannten Indikatorenketten und daraus entstehenden Warnungen. Als Beispiel ist die Warnung einer bevorstehenden Fehlproduktion in einem Herstellungsprozeß bestimmt im voraus nützlicher als nach dem der Ausschuß bereits begonnen hat. Die frühe Erkennung irgendeines wichtigen Zusammenhangs in z.B. einer medizinischen, pharmazeutischen, juristischen, kriminalpolizeilichen oder labortechnischen Situation kann den Unterschied zwischen Erfolg und Mißerfolg bedeuten.

Den Erfolg von solchen Projekten in Geld zu messen, ist meist möglich aber oft müßig, denn der wahre Gewinn befindet sich jenseits den Wert des Geldes.

Das richtige Vehikel (Horses for Courses)

Jedes Szenario sowie jedes Aufgabenumfeld erfordert die passenden Werkzeuge. Genauso wie jedes Formel Eins Auto je nach Rennstrecke umgebaut werden muß, so benötigt man jeweils den passenden Data Mining Vehikel für bestimmte Aufgaben.

Die Festhaltung an bestimmter Werkzeuge oder Algorithmen widerspricht die grundsätzliche Einstellung des Data Minings, das keine Person und keine Erfahrung bedeutender ist als die Macht der gesamten Disziplin.

Die Eier legende Wollmichsau

Dieses Tier gibt es nicht. Im Data Mining gibt es ebenfalls kein Allheilmittel, außer Sie suchen sich die für Ihre Aufgaben und für Ihr Umfeld passende Werkzeuge und Mannschaft. Hierbei sollten Sie den üblichen Schutz gegen Betriebsblindheit sowie den Wert des Expertenwissens in Ihrem Kalkül einbeziehen.

Die Vision

Wo wollen wir mit unserem Projekt ankommen und wer kann dieses Ziel entdecken, artikulieren und anschließend für uns erreichen? Diese zusammenhängenden Fragen sollten den Antrieb für die Entscheidung von Data Mining Projekten bilden. Was wollen/können wir erreichen (Gibt es vielleicht mehr?), setzt alles andere voraus. Dies ist der Startpunkt aller Erfolgsberechnungen. Wie sollen wir das Ziel beschreiben? Was gehört dazu und was nicht? Schließlich wird ein Projekt durch Mannschaftsmitglieder und einen Kapitän/Trainer zum Erfolg geführt. Wer soll diese Rolle spielen.

Alle diese Komponente gehören zum Erfolgsrezept. Der erfahrenen Durchblick der führenden Person bildet die Schnittstelle zum Erfolg. Die Kosten für noch so gute Software und noch kräftigere Hardware werden bei der Auswirkung der falschen Projektbelegung erblassen. Auch in der Formel Eins trifft zu, ein noch so gutes Auto wird nicht gewinnen, wenn die Mannschaft sowie der Fahrer von der Fähigkeit und der Zusammenarbeit nicht stimmen.

Zum Abschluß

Wir haben uns bemüht, Ihnen das Rüstzeug zur Bewältigung eines erfolgreichen Data Mining Projektes unter Umschiffung der größten Fallen zu bieten. Es gibt hier keine

einfache Antwort, trotz allem waren wir bestrebt, die Antwort systematisch und deutlich aufzubauen.

Gleichzeitig hoffen wir Ihnen eine gewisse Unterhaltung geboten zu haben, denn das praktische Data Mining ist nur von der Wirkung faszinierend, sondern vor allem in der Sache spannend. Von dieser Spannung möchten wir etwas herüber gebracht haben sowie eine Faszination für Data Mining erzeugt zu haben. Vielen Dank.

Wir danken für die freundliche Freigabe dieser Dokumente

Anlage I

Vertraulichkeitserklärung

Hiermit erklären wir, die Daten aus Ihrem Hause, die wir demnächst per Diskette in XXXX-Format erhalten, ausschließlich zu Zwecken der Test und Demonstration in Zusammenarbeit mit Ihnen persönlich oder Ihrem Hause zu gebrauchen.

Die zu übermittelden Daten werden streng vertraulich behandelt und Dritten nicht zugänglich gemacht. Dies gilt auch für Präsentationen auf Messen oder bei anderen Firmen. Die Verwendung dieser Daten dient allein dem Pilotprojekt mit ANGOSS KnowledgeSTUDIO für die FIRMA.

Den Beginn des Pilotprojektes haben wir für den DATUM um UHRZEIT in unserem Hause vereinbart. Nach Abschluß des Pilotprojektes werden die Daten in unserem Hause auf unseren Computern gänzlich gelöscht. Es werden keine Kopien oder sonstige Duplikate egal in welcher Form – komplett oder als Ausschnitt – erstellt.

In der Erwartung einer aufregenden Analyse, verbleiben wir

Diese Vertraulichkeitserklärung kann auch ganz anders klingen. Hier ein Ausschnitt:-

Wir wurden insbesondere darüber belehrt, dass es uns untersagt ist, die übergebenen geschützten personenbezogenen Daten zu einem anderen als dem vereinbarten Projekt zu verarbeiten, bekanntzugeben, zu übermitteln, zugänglich zu machen oder sonst zu nutzen oder an Dritte weiterzugeben. Außerdem ist es uns untersagt, die überreichten anonymisierten Daten zu entschlüsseln. Der Versuch allein kann strafrechtliche Folgen mit sich bringen.

Wir ermächtigen Sie hiermit, soweit notwendig, im Rahmen des Umgangs mit den Datenträgern zur Datenverarbeitung **eine Kopie** der Ihnen überreichten Datenträger anzufertigen.

Nach Abschluß des Projektes sind die Originaldatenträger zurückzugeben, sowie alle angefertigten Kopien unwiederbringlich zu löschen bzw. zu vernichten.

Uns ist bekannt, dass Verstöße gegen das Datengeheimnis nach den einschlägigen Rechtsvorschriften mit Freiheits- oder Geldstrafen geahndet werden können; arbeitsrechtliche Folgen werden dadurch nicht

ausgeschlossen. Eine Verletzung des Datengeheimnisses wird in den meisten Fällen gleichzeitig ein Verstoß gegen die arbeitsvertragliche Schweigepflicht darstellen, auch kann in ihr eine Verletzung spezieller Geheimhaltungspflichten liegen.

Diese Version stammt eher vom Juristen und zitiert gleichzeitig diverse Gesetze, vor allem Bundesdatenschutzgesetz und Strafgesetzbuch.

Die folgende Version gleicht eine fertige Vereinbarung, die auch den Austausch von Daten während einer Vorbereitungsphase von Präsentationen sowie einer späteren Zusammenarbeit in Projekten abdeckt.

Vertraulichkeitsvereinbarung

zwischen

Name und Adresse des Dateninhabers (nachfolgend „INHABER“ genannt)

und

Name und Adresse des Data Miners (nachfolgend „MINER“ genannt)

Die Parteien wollen in Projekten zusammenarbeiten, bei denen vertrauliche Informationen gegenseitig ausgetauscht werden. Für alle künftigen Projekte zwischen den Parteien wird zum Schutze vertraulicher Informationen die nachfolgende Vereinbarung getroffen.

(1) Die Parteien behandeln alle Informationen vertraulich, die ihnen von der anderen Partei mündlich, schriftlich oder in elektronischer Form übermittelt werden. Vertrauliche Informationen im Sinne dieser Vereinbarung sind diejenigen Informationen, die dem Geschäft einer Partei zuzurechnen sind und

(i) von einer Partei als vertraulich oder ihr gehörend bezeichnet oder markiert werden, oder

(ii) aufgrund ihres Charakters oder ihrer Natur von einer verständigen Person als vertraulich behandelt werden würden.

Zu den vertraulichen Informationen von INHABER gehören insbesondere alle .....-daten von INHABER und die daraus erzeugten Analysen.

(2) Die Parteien treffen alle angemessenen Maßnahmen, um sicherzustellen, dass ihre Mitarbeiter diese Informationen ebenfalls vertraulich behandeln. Außerdem werden die Parteien die erforderlichen technischen und organisatorischen Maßnahmen zum Schutz personenbezogener Daten nach §9 Bundesdatenschutzgesetz treffen. Die Mitarbeiter sind entsprechend dieser Vereinbarung auf die Einhaltung der Vertraulichkeit und der geltenden Datenschutzvorschriften zu verpflichten.

(3) Außerdem verpflichtet sich der MINER:

(i) die zur Verfügung gestellten vertraulichen Informationen ausschließlich für die mit INHABER vereinbarten Zwecke zu nutzen und nicht kommerziell oder auf andere Weise zu verwerten oder sie zu kopieren oder auf andere Weise ohne vorherige Zustimmung von INHABER weiterzugeben;

(ii) die vertraulichen Informationen, die im Rahmen dieser Verspätung zur Verfügung gestllt wurden, auf Anforderung unverzüglich, spätestens aber innerhalb von einer Woche zurückzugeben, wenn und soweit sich aus den einzelnen projektbezogenen Verträgen nicht etwas anderes ergibt.

(4) Die Vertraulichkeitsverpflichtung gilt nicht für Informationen, die

(i) zum Zeitpunkt des Abschlusses dieser Vereinbarung bereits öffentlich bekannt waren und weder durch eine Handlung noch durch eine Unterlassung der anderen Partei öffentlich bekannt wurden;

(ii) von einer Partei unabhängig erzeugt wurden;

(iii) sich bereits vor Beginn der Geschäftsbeziehung im Besitz der Partei befanden;

(iv) aufgrund einer rechtsgültigen Anordnung eines zuständigen Gerichts oder Behörde offengelegt werden müssen. In diesem Fall muß die offenlegende Partei der anderen Partei unverzüglich – und wenn möglich vor der Offenlegung – davon unterrichten und vom Empfänger eine vertrauliche Behandlung der Informationen verlangen.

(5) Diese Vereinbarung kann drei Jahre nach dem Ende des letzten Vertragsverhältnisses zwischen INHABER und MINER gekündigt werden.

(6) Ist eine Bestimmung dieser Vereinbarung ganz oder teilweise unwirksam, so bleibt die Wirksamkeit der übrigen Bestimmungen hiervon unberührt. Die Parteien verpflichten sich in

diesem Fall, die unwirksame Bestimmung durch diejenige wirksame Bestimmung zu ersetzen, die dem wirtschaftlichen Zweck der unwirksamen Bestimmung am nächsten kommt.

(7) Die Vereinbarung unterliegt dem Recht der Bundesrepublik Deutschland . Ausschließlicher Gerichtsstand für alle Streitigkeiten ist .....

Ort, den _____ Ort, den_____

_____ _____

INHABER MINER

Anlage II

Algorithmen

i) Der Brenda Algorithmus

Dies ist eine Tiling Technik, die eine fortlaufende unabhängige Variable (UV) verzweigt, eine binäre, diskrete abhängige Variable (AV) vorausgesetzt:

Liste0 = Liste aller UV Werte für den "0" Wert der AV
Liste1 = Liste aller UV Werte für den "1" Wert der AV

definiere Median(Liste) als Medianwert der UV Werteliste, und
Leer(Liste) als wahr, wenn es keine Elemente in der Liste gibt, und
definiere Mutatoren größer(), die Liste abkürzt gegenüber Median Wert höher, und kleiner(), die Liste abkürzt gegenüber Median Wert niedriger, dann:

```
/*
 * erste Mediane
 */
Median0 = Median(Liste0)
Median1 = Median(Liste1)

/*
 • tausche Listen so dass Liste0 Median ist < Liste1 Median
 */
wenn(Median0 < Median1)
        tausche(Liste0, Liste1)

/*
 * wenn weitere Listen zu betrachten
 */
solange(!Leer(Liste0) && !Leer(Liste1)) {
        /*
         * Mediane der übrigen Listen aufgreifen
         */
        x0 = Median(Liste0)
        x1 = Median(Liste1)

        /*
         * kreuzen sich die Mediane, sind wir fertig
         */
        wenn(x0 >= x1)
                abbrechen;
```

```
            /*
             * sonst Listen abkürzen und fortfahren
             */
            größer(Liste0)
            kleiner(Liste1)
    }

    /*
     * die Antwort
     */
    Median0 = x0
    Median1 = x1;
```

ist der Brenda Algorithmus. Die Ausgabe besteht aus zwei Werten N und H, die als die Grenzen der UV Verzweigung dienen:

(-INF, N]
[N, H)
[H, +INF)

Wenn N und H nicht gleich sind. Ansonsten:

(-INF, N]
(N, +INF)

Wenn N und H gleich sind.

Die Aufgabe des Brenda Algorithmus ist den Plan für die UV zu berechnen. Es ist eine andere Form des Tiling.

Es geht darum, die N und H Werte zu finden, wo der

(-INF, N]

Zweig hauptsächlich ein AV Wert (0 oder 1 -- binär diskret) enthält, und

[H, +INF)

hauptsächlich den anderen AV Wert enthält. Der mittlere Bereich:

[N, H]

besitzt eine "Mittel" AV.

Wenn man sich zwei normale Verteilungen vorstellt, - ein pro Wert der AV - Die auf den entsprechenden UV Werten basieren. Die Mediane jeder Verteilung sind die Spitzen des Normalen. Der Brenda Algorithmus schneidet die übrigen Hälften der Verteilungen erfolgreich ab, bis sich die verbleibenden Mediane treffen: Das heißt, wo die normalen Kurven sich kreuzen, wenn im gleichen Maßstab geplottet.

Hatten die Verteilungen die gleiche Varianz, würde der Algorithmus fast sofort zum Abschluß kommen und der Intervall zwischen den Durchschnittswerten der Verteilung würde der [N, H) Bereich wie oben sein.

Bemerken Sie, dass die Verwendung einer normalen Verteilung nur den Zweck des Beispiels dient. Die Technik ist nicht parametrisch.

Entscheidungsbaum Algorithmen – KnowledgeSEEKER / HeatSEEKER

(mit Dank an ANGOSS Software International)

Die Notation

Y – abhängige Variable (AV);
N – Anzahl Datensätze des Knoten;

Für eine bestimmte unabhängige Variable (UV):

Kategorie – eine Gruppe Werte, die eine Variable annehmen kann (reicht von einem einzigen Wert bis zu allen möglichen Werten);
r – Anzahl nicht zusammenhängender Kategorien;
v_i - *i*-te Kategorie;
n_i - Anzahl Datensätze im Knoten mit UV in v_i ;

Numerische AV

$y_{i1},...,y_{in_i}$ - Werte der AV für Datensätze mit UV in v_i ;

$$y_i = \sum_{j=1}^{n_i} y_{ij} \ ;$$

$$\bar{y}_i = \frac{y_i}{n_i} \ ;$$

$$y = \sum_{i=1}^{r} y_i = \sum_{i=1}^{r} \sum_{j=1}^{n_i} y_{ij} \ ;$$

$$\bar{y} = \frac{y}{N} ;$$

$$STD = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \ ;$$

$$SSTR = \sum_{i=1}^{r} n_i (\bar{y}_i - \bar{y})^2 \ ;$$

$$SSE = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 ;$$

$$MSTR = \frac{SSTR}{r-1} ;$$

$$MSE = \frac{SSE}{N-r} ;$$

$$F^* = \frac{MSTR}{MSE} ;$$

$$PV = F(F^*; r-1, N-r) ;$$

Wenn *i* und *k* zwei Kategorien der UV sind, dann:

$$MSTR_{\{i,k\}} = n_i\left(\bar{y}_i - \frac{y_i + y_k}{n_i + n_k}\right)^2 + n_k\left(\bar{y}_k - \frac{y_i + y_k}{n_i + n_k}\right)^2 ;$$

$$MSE_{\{i,k\}} = \frac{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{j=1}^{n_k}(y_{kj} - \bar{y}_k)^2}{n_i + n_k - 2} ;$$

$$F^*_{\{i,k\}} = \frac{MSTR_{\{i,k\}}}{MSE_{\{i,k\}}} ;$$

Kategorische AV

d – Anzahl eindeutiger Werte, die eine AV aufnehmen kann;

w_j - *j*te Wert der AV;

q_{ij} - Anzahl Datensätze mit AV= w_j und UV in v_i ;

m_j - Anzahl Datensätze mit AV= w_j ;

$\bar{}_{ij}$ $\overline{}^{i}$;

Wir haben:

$$m_j = \sum_{i=1}^{r} q_{ij} , \quad n_i = \sum_{j=1}^{d} q_{ij} , \quad N = \sum_{j=1}^{d} m_j = \sum_{i=1}^{r} n_i = \sum_{i=1}^{r}\sum_{j=1}^{d} q_{ij} ;$$

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{d}\frac{(q_{ij} - \bar{q}_{ij})^2}{\bar{q}_{ij}} ;$$

$$X^2_{\{i,k\}} = \sum_{j=1}^{d}\frac{(q_{ij} - \bar{q}_{ij})^2}{\bar{q}_{ij}} + \sum_{j=1}^{d}\frac{(q_{kj} - \bar{q}_{kj})^2}{\bar{q}_{kj}} ;$$

$$PV = c^2(X^2; df) \text{ wo } df = (d-1)(r-1) ;$$

$$E_i = -\frac{1}{\ln d}\sum_{j=1}^{d}\frac{q_{ij}}{n_i}\ln\frac{q_{ij}}{n_i};$$

$$E = -\frac{1}{\ln d}\sum_{j=1}^{d}\frac{m_j}{N}\ln\frac{m_j}{N};$$

$$G = 1 - \frac{\sum_{j=1}^{d} m_j^2}{N^2};$$

$$G_i = 1 - \frac{\sum_{j=1}^{d} q_{ij}^2}{n_i^2};$$

Algorithmen

Behandlung einer numerischen UV

Um eine numerische UV in einem Entscheidungsbaum verwenden zu können, muß sie kategorisch gemacht werden. KnowledgeSTUDIO bricht den Wertebereich der UV in eine Anwender kontrollierte Anzahl Intervalle (per Default 10) auf, damit die Anzahl Datensätze für jedes Intervall gleich ist. Jedes Intervall wird dann als Kategorie angesehen. the range of values of the IV into user-controlled number of intervals (set by default to 10) so that the number of records for each interval is the same. Each interval is then considered a category.

KnowledgeSEEKER

Verzweigung einer bestimmten UV finden

Die Verzweigung wird durch die iterative Gruppierung von Kategorien der UV in neue Kategorien gefunden. Ein passendes Paar Kategorien wird in eine einzelne Kategorie zusammengefaßt, wenn sich die beiden “ähneln” im Sinne der AV. Bei einer ungeordneten UV sind alle Paare passend, während bei einer geordneten UV sind nur die benachbarten Paare passend. Es bestehen zwei Methoden der Gruppierung: Erschöpfend und Cluster.

1. Die erschöpfende Methode beginnt mit jedem Wert der UV als getrennte Kategorie. Bei jeder Iteration:

   - für jedes passende Kategorienpaar (i,k) wird die Ähnlichkeitsstatistik errechnet. Bei numerischer AV ist die Statistik $F^*_{\{i,k\}}$ und bei kategorischen AV ist sie $X^2_{\{i,k\}}$;
   - das Paar mit der höchsten Ähnlichkeitsstatistik wird in eine einzelne Kategorie zusammengeführt, daher wird die Gesamtanzahl Kategorien *r* durch 1 reduziert. Die neue Menge Kategorien ist die nächste mögliche Verzweigung und der P-Wert wird errechnet. Für numerische AV ist der P-Wert $F(F^*; r-1, N-r)$ und für kategorische AV ist er $c^2(X^2; df)$;

   Diese beiden Schritte werden wiederholt bis $r = 1$. Von allen möglichen Verzweigungen wird der mit dem höchsten **Info** (siehe unten) gewählt. Wenn **Info** oberhalb des vom Anwender gesetzten Grenzwertes liegt, wird die Verzweigung für jene UV angezeigt. Ansonsten können keine signifikante Verzweigungen für die UV gefunden werden.

2. Die Cluster Methode beginnt auch mit jedem Wert der UV als getrennte Kategorie. Dann bei jeder Iteration:

   - für jedes passende Kategorienpaar (i,k) wird die Ähnlichkeitsstaistik errechnet. Für numerische AV sit die Statistik $F^*_{\{i,k\}}$ und für kategorische AV ist sie $X^2_{\{i,k\}}$;
   - das Paar mit der höchsten Ähnlichkeitsstatistik wird in eine einzelne Kategorie zusammengeführt, daher wird die Gesamtanzahl Kategorien *r* durch 1 reduziert. Die neue Menge Kategorien ist die nächste mögliche Verzweigung und der P-Wert wird errechnet. Für numerische AV ist der P-Wert $F(F^*; r-1, N-r)$ und für kategorische AV ist er $c^2(X^2; df)$;

   Diese beiden Schritte werden wiederholt bis der P-Wert des Paares mit der höchsten Ähnlichkeitsstatistik, die während des ersten Schrittes gefunden wurde, den vom Anwender gesetzten Grenzwert übersteigt. (Der Defaultwert ist 0.05) oder bis $r=1$. Der obere P-Wert ist gleich $F(F^*_{\{i,k\}}; 1, N-2)$ für eine numerische AV und gleich $c^2(X^2_{\{i,k\}}; d-1)$ für eine kategorische AV. Von allen möglichen Verzweigungen wird die mit dem höchsten **Info** gewählt. Nachdem diese Verzweigung gefunden ist, versucht der Algorithmus die Signifikanz zu verbessern. Er nimmt einzelne AV Werte der jeweils zugeordneten Gruppen und gruppiert ihnen mit anderen Kategorien. Im Falle einer Verbesserung, wird die betroffene Verzweigung zum neuen Kandidaten. Dieser Prozess der erneuten Verzweigung ergibt möglicherweise eine unterschiedliche Verzweigung im Vergleich zur Zusammenführung. Wenn der **Info** über dem vom Anwender gesetzten Grenzwert liegt, wird die Verzweigung für jene UV angezeigt. Ansonsten können keine signifikante Verzweigungen für die UV gefunden werden.

Messung des Informationsgehalts der Verzweigungen von jeder UV

Im ersten Teil des Algorithmus ist höchstens eine Verzweigung pro UV gefunden worden. Der Informationsgehalt dieser Verzweigungen wird nach der aktiven Verzweigungsmessung errechnet. Verzweigungsmessungen sehen Sie unten in der Beschreibung. Eine Verzweigung ist informativ, wenn sie den vom Anwender gesetzten Grenzwert passiert. Der Anwender kann jede informative Verzweigung auswählen. Die informativste Verzweigung wird per Default ausgewählt.

Wachstumskontrollparameter

Das Wachstum des Baumes wird durch eine Raihe Anwender gesetzte Parameter gesteuert, die wie folgt definiert werden:

- **Maximale Anzahl UV Kategorien.** Wenn eine UV mehr Kategorien besitzt als diese Zahl, wird sie in der Verzweigung nicht benutzt;
- **Minimale Knotenerzeugungsgröße.** Wenn, am Ende eines Verzweigungsalgorithmus (Cluster oder Erschöpfend), ein Knoten in der gefundenen Verzweigung weniger Datensätze als diese Zahl enthält, wird jener Knoten (für geordnete UV) mit einem benachbarten Knoten zusammengeführt oder (für ungeordnete UV) mit dem statistisch ähnlichsten Knoten zusammengeführt;
- **Anzahl Verzweigung Cache pro Knoten.** Dieser Parameter wird zur Verwaltung des Arbeitsspeichers verwendet. Er bestimmt wieviele Verzweigungen (eine pro UV) für jeden Knoten gespeichert werden. Wenn der auf einen früheren Verzweigungsknoten zurückkommen möchte, um seine Verzweigung auf eine andere UV zu sehen, muß jene Verzweigung erneut errechnet werden, wenn sie nicht gespeichert wird;

Die folgenden zwei Parameters greifen, wenn der Baum automatisch aufgebaut wird.

- **Automatische Wachstumstopgröße (Minimum Datensätze).** Wenn einen Knoten weniger Datensätze enthäkt, als in diesem Parameter angegeben, wird jener Knoten nicht verzweigt.
- **Automatisches Wachstum – Maximale Baumtiefe.** Wenn der Pfad vom obersten Knoten zu einem bestimmten Knoten soviele Knoten enthält, wie in diesem Parameter festgelegt, wird jener bestimmte Knoten nicht verzweigt. Wenn dieser Parameter gleich 0 ist, (wie per Default) wird das Baumwachstum nicht begrenzt.

HeatSEEKER

Verzweigung einer bestimmten UV finden

Dieser Teil des Algorithmus verwendet die Cluster Methode wie oben bei KnowledgeSEEKER beschrieben.

Messung des Informationsgehalts der Verzweigungen von jeder UV

Dieser Abschnitt wird exakt wie bei KnowledgeSEEKER ausgeführt.

Wachstumskontrollparameter

- **Maximale Baumtiefe.** Wie oben Automatisches Wachstum – Maximale Baumtiefe
- **Maximale Baumbreite.** Steuert die Verzweigungsprozedur. Wenn eine mögliche Verzweigung mehr Knoten als diese Anzahl aufweist, werden Knoten nach der statistischen Ähnlichkeit zusammengeführt;
- **Minimale Knotenbevölkerung.** Wie oben minimale Knotenerzeugungsgröße;
- **Minimale passende Objektive.** Analog der minimalen Knotenbevölkerung***;***
- **Diskreter Informationsverlust**
- **Anzahl Verzweigungscache pro Knoten.** Wie in KnowledgeSEEKER;

Parameter, die das automatische Wachstum steuern, unterliegen die gleichen Definitionen wie in KnowledgeSEEKER.

Voting

Automatisch beim Wachstum ermöglicht den Aufbau mehrfacher Bäume. KnowledgeSTUDIO wählt *t* informativste Verzweigungsvariablen beim obersten Knoten. Für jede dieser Variablen wird ein separater Baum automatisch aufgebaut. Die *t* Bäume werden im Voting Vorgang dann zusammengebunden, wenn ein Vorhersagemodell erstellt wird. Wenn eine Vorhersage für einen Datensatz vorgenommen wird, wird sie zuerst von jedem erstellten Baum vorgenommen, wo er in einem der Blattknoten vorkommt. Dann haben wir für jenen Datensatz *t* mögliche Blattknoten, einer pro Baum. Einer wird ausgewählt. Wenn AV numerisch ist, wird der Knoten mit der niedrigsten Standardabweichung gewählt. Wenn AV kategorisch ist, wird es der Knoten mit dem höchsten Modus sein. Der Modus ist die höchste Proportion eines einzelnen Wertes der AV in der Knotendatemmenge. Dann bildet der Baum mit dem ausgewählten Knoten die Vorhersage für diesen Datensatz.

Messung des Informationsgehalts der Verzweigungen

Statistiken der Verzeweigungen

Bei der Evaluierung der Signifikanz einer Verzweigung errechnet KnowledgeSTUDIO eine Anzahl Statistiken (neben der Datensatzzählung und Porportionen). Welche Statistiken berichtet werden ist Anwender abhängig, außer **Info,** die bei jeder Messung und AV Typ angegeben wird. Unten werden eine Liste der Statistiken mit dem jeweiligen mathematischen Ausdruck pro Messung und AV Typ.

Statistische Messungen

Eine Verzweigung ist eine Partition des Knotendatenmenge in Untermengen nach dem Wert der unabhängigen Variable. Nach der statistischen Messung ist die beste Verzweigung diejenige, die Untermengen generiert, die mit der wenigsten Wahrscheinlichkeit aus der selben Verteilung kommen werden. Die Wahrscheinlichkeit wird nach standardmäßigen statistischen Tests gemessen wird. Eine Verzweigung wird dann angezeigt, wenn eine vordefinierte Signigfikanzebene passiert wird. Diese Ebene kann auf zwei Arten erreicht werden:

1. Mit der nicht angepaßter P-Wert Messung wird die gewünschte Signifikanzeben direkt bei der besten Verzweigung angewendet.
2. Angepaßte Messung berücksichtigt, dass eine Anzahl unterschiedlicher Verzweigungen für jede UV errechnet wird und lediglich eine wird als signifikant angesehen. Diese Verzweigung muß eine strengere Signifikanzprüfung bestehen, als wenn eine einfache Verzweigung versucht wurde. Die Auswahl dieser Prüfung wird innerhalb des Bonferroni Rahmens durch Anpassung des P-Wertes einer Verzweigung vorgenommen. Exaktere technishe Details dieser Anpassung können auf Anfage bei ANGOSS erhalten werden.

Nicht angepaßter P-Wert

1. Numerische AV

   - $Info = 100 \times C$;
   - $df1 = N - r$;
   - $df2 = r - 1$;
   - $F = F^* = \frac{MSTR}{MSE}$;
   - $P = PV = F(F^*; df2, df1)$;
   - $C = 1 - P$;

2. Kategorishe AV

   - $Info = 100 \times C$;
   - $df = (d-1)(r-1)$;
   - $Chi2 = X^2 = \sum_{i=1}^{r}\sum_{j=1}^{d} \frac{(q_{ij} - \overline{q}_{ij})^2}{\overline{q}_{ij}}$;
   - $P = PV = c^2(X^2; df)$;
   - $C = 1 - P$;

Angepaßter P-Wert

1. Numerische AV

   - $Info = 100 \times C$;

- $df1 = N - r$;
- $df2 = r - 1$;
- $F = F^* = \frac{MSTR}{MSE}$;
- $uP = PV = F(F^*; df2, df1)$;
- $uC = 1 - uP$;
- $P = B \times uP$;
- $C = 1 - P$;
- $B = B_a B_m$;
- B_m ist der zusätzliche Anpassungsfaktor, der vom User eingerichtet werden kann. Der Default ist 1;
- B_a ist der Bonferroni Anpassungsfaktor, der automatischa vom Algorithmus bestimmt wird;

2. Kategorische AV

- $Info = 100 \times C$;
- $df = (d-1)(r-1)$;
- $Chi2 = X^2 = \sum_{i=1}^{r} \sum_{j=1}^{d} \frac{(q_{ij} - \overline{q}_{ij})^2}{\overline{q}_{ij}}$;
- $uP = PV = c^2(X^2; df)$;
- $uC = 1 - uP$;
- $P = B \times uP$;
- $C = 1 - P$;
- $B = B_a B_m$;
- B_m ist der zusätzliche Anpassungsfaktor, der vom User eingerichtet werden kann. Der Default ist 1;
- B_a ist der Bonferroni Anpassungsfaktor, der automatischa vom Algorithmus bestimmt wird;

Nicht statistische Messungen

Entropie Varianz

Diese Quantität besagt wieviel Information über einen AV Wert in einem wahllosen Auszug aus einer Knotendatenmenge enthalten ist. Wenn AV Werte gleichmäßig innerhalb jener Menge veteilt waren, würde man durch den wahllosen Auszug eines Datensatzes keine Information erhalten, da jeder Wert der AV gleich wahrscheinlich sein würde. Solltenjedoch alle Datensätze den gleichen AV Wert besitzen, würde ein wahlloser Auszug perfekte Information liefern. Hohe Entropie bedeutet einen niedrigen Informationswert und umgekehrt. Entropie *E* erreicht einen Maximum von 1 bei einer gleichmäßigen Verteilung und einen Minimum von 0 bei einer degenerierten (zugespitzter) Verteilung.

1. Numerische AV

- $Info = 100 \times RatioVariance$;

- $InputVariance = STD$;
- $OutputVariance = SSE$;
- $RatioVariance = 1 - \frac{OutputVariance}{InputVariance}$;

2. Kategorische AV

- $Info = 100 \times RatioEntropy$;
- $InputEntropy = E$;
- $OutputEntropy = \sum_{i=1}^{r} \frac{n_i}{N} E_i$;
- $RatioEntropy = 1 - \frac{OutputEntropy}{InputEntropy}$;

Gini Varianz

Gini Varianz ähnelt Entropie Varianz insofern als sie ebenfalls die Verteilung der AV Werte (gleichmäßig oder nicht) in der Datenmenge mißt.

1. Numerische AV

- $Info = 100 \times RatioEntropy$;
- $InputVariance = STD$;
- $OutputVariance = SSE$;
- $RatioVariance = 1 - \frac{OutputVariance}{InputVariance}$;

2. Kategorische AV

- $Info = 100 \times RatioIndex$;
- $InputIndex = G$;
- $OutputIndex = \sum_{i=1}^{r} \frac{n_i}{N} G_i$;
- $RatioIndex = 1 - \frac{OutputIndex}{InputIndex}$;

Grenzwerte des Informationsgehaltes von Verzweigungen

In der obigen Beschreibung der Algorithmen wird gesagt, dass nur informative Verzweigungen (d.h. jene deren **Info** den Grenzwert passiert) angezeigt werden. Dieser Grenzwert wird je nach Messungstyp unterschiedlich gesetzt.

Statistische Messungen

Für eine statistische Messung wird der **Info** Grenzwert aus der Signifikanzebene gebildet, die der Anwender einstellen kann. Verzweigungen bei der *P* unterhalb jener Signifikanzebene liegt, (und somit **Info** ist über den abgeleiteten Grenzwert) werden angezeigt. Der Defaultwert der Signifikanzebene ist 0.05.

Nicht-statistische Messungen

Bei nicht-statistischer Messungen kann der Anwender Maximum und Minimum Grenzwerte für **Info** setzen.

Expectation-Maximization

Expectation-Maximization (EM), wie K-means, ist ein Algorithmus zur Clusterung einer Datenmenge in eine vordefinierte Anzahl Cluster (oder Gruppen) auf der Basis der Ähnlichkeit der Datensätze.

Anders als K-means basiert die Ähnlichkeit im EM Algorithmus auf der Theorie der Wahrscheinlichkeit. Ein Datensatz wird einem bestimmten Cluster dann zugewiesen, wenn er am wahrscheinlichsten durch die Wahrscheinlichkeitsverteilung des entsprechenden Clusters generiert würde – hierbei sind die Verteilungen pro Cluster unterschiedlich. Zum Beispiel, wenn Datensätze Kunden entsprechen, könnten unterschiedliche Verteilungen unterschiedliche Verhaltensarten bei den Kunden bedeuten.

Der Algorithmus beginnt durch die Wahl K unterschiedlicher Verteilungen. Jede Verteilung wird dann durch eine Anzahl Parameter beschrieben. Zusätzlich wird jeder Verteilung eine Gewichtung zugewiesen. Danach wird eine vorläufige Schätzung für alle Parameter vorgenommen, einschließlich Gewichtungen. Auf der Basis der Parameterwerte kann die Wahrscheinlichkeit der Betrachtung der Training Datenmenge errechnet werden. Der EM Algorithmus findet Parameterwerte, die diese Wahrscheinlichkeit maximiert. Bei jeder Iteration werden neue Parameterwerte unter Verwendung der alten Werte zur Maximierung der Zugehörigkeitswahrscheinlichkeit errechnet. Dieser Vorgang wird solange wiederholt, bis sich die Parameterwerte einen Spitzenwert erreichen, wo die Funktion der Wahrscheinlichkeit maximiert wird.

Schließlich wird beim Scoring eines neuen Datensatzes die Wahrscheinlichkeit, dass jener Datensatz von jeder K Verteilung generiert wird, unter Verwendung der endgültigen Parameterwerte und Bayes Formel der konditionalen Wahrscheinlichkeiten errechnet. Der Datensatz wird dann dem Cluster mit dieser maximalen Wahrscheinlichkeit zugewiesen

Regressionen

Lineare oder logistische Regression wird zur Vorhersage des Wertes einer abhängigen Variablen (AV) unter Verwendung der Werte einer Menge numerischer oder kategorischer unabhängiger Variablen (UV).

Die Einbeziehung einer kategorischen Variable als UV in einer Regression wird wie folgt organisiert. Betrachten Sie eine kategorische Variable mit *n* eindeutigen Werten. Diese Werte werden lexikographisch nach ihren Namen geordnet. Der erste Wert dieser Ordnung dient als Referenz. Für jeden anderen Wert wird eine Design Variable erzeugt. Somit werden *n*-1 Design Variable der Regression hinzugefügt. Der Defaultwert des Referenzwertes kann geändert werden.

Notation:

N – Anzahl Beobachtungen in der Regression;

p – Anzahl der Variablen in der Regression (ausschließlich Kreuzung);

$$X = \begin{bmatrix} x_{11} & \dots & x_{1p} & 1 \\ x_{21} & \dots & x_{2p} & 1 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{N1} & \dots & x_{Np} & 1 \end{bmatrix}$$ - Matrix der UV Werte (Design Matrix; in Modellen ohne Kreuzung wird die letzte Spalte weggelassen);

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_N \end{bmatrix}$$ - Spaltenvektor der AV Werte;

$$b = \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ . \\ b_p \end{bmatrix}$$ - Spaltenvektor der geschätzten Parameter (in Modellen ohne Kreuzung wird b_0 weggelassen);

$$J = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 1 & 1 & \dots & 1 \end{bmatrix}$$ - $N \times N$ Matrix von 1s;

Lineare Regression

Bei einer linearen Regression muß die AV numerisch sein. ANGOSS KnowledgeSTUDIO schätzt die lineare Regression durch ordinäre Mindestquadrate. Ergo gleicht der Vektor der Parameterschätzungen

$$b = \left(X'X\right)^{-1} X'Y$$

Die Berechnung erfolgt unter Verwendung der singulären Wertauflösung von *X*.

Regressionsausgabe hat bei ANGOSS KnowledgeSTUDIO den folgenden Format:

Sequenz Statistik			
Abhängige Variable:	*Name*		
Varianz erläutert:	$\frac{SSR}{SSTO}$	Angepaßte Varianz Erläutert:	$1-\frac{MSE}{MSO}$
F-Verhältnis:	$\frac{MSR}{MSE}$	F-Verhältnis Freiheitsgrad 1 / 2:	*RDF* /
P-Wert:	$F\left(\frac{MSR}{MSE}; RDF, EDF\right)$		

Analyse der Varianz

Quelle	Summe der Quadrate	DF	Mean-Square
REGRESSION			$\frac{SSR}{RDF}$
FEHLER			$\frac{SSE}{EDF}$
SUMME			$\frac{SSTO}{TDF}$

Unabhängige Variable Statistik

Variable Name / Wert	Modell Parameter	Parameter Standard Fehler	Wald Statistik	Signifikanz
Werte für kategorische Variablen	b_i	$SE(b_i) = \left(MSE \cdot \left(X'X\right)_{ii}^{-1}\right)^{1/2}$	$\left[\frac{b_i}{SE(b_i)}\right]^2$	$c^2\left(\left[\frac{b_i}{SE(b_i)}\right]^2, 1\right)$

Oben:

$TDF = N$ - Anzahl Beobachtungen in der Regression;

$RDF = p$ - Anzahl Variablen in der Regression ausschließlich der Kreuzung;

$EDF = TDF - RDF$;

$SSR = b'X'Y - \frac{1}{N}Y'JY;$

$SSTO = Y'Y - \frac{1}{N}Y'JY = y_1^2 + ... + y_N^2 - \frac{1}{N}(y_1 + ... + y_N)^2;$

$SSE = SSTO - SSR = (Y - Xb)'(Y - Xb);$

Nachdem die Menge der für die Regression zu betrachtenden Variablen ausgewählt wurde, können noch einzelne Variablen selektiert werden. Regressionen die während der Varaiblenselektion ausgeführt werden, werden Sequenzen genannt und ihre Ausgabe kann angezeigt werden (siehe oben). Bei ANGOSS KnowledgeSTUDIO stehen sechs Methoden zur Verfügung:

1. **Vorwärts Ordnung** beginnt mit dem Grundmodell (beinhaltet Kreuzung und alle vom Anwender ausgewählten Variablen, die ins Modell gezwingen werden sollen) und fügt Variablen nach der eingestellten Ordnung ein – mit Beginn bei der Ersten.
2. **Rückwärts Ordnung** beginnt mit dem vollständigen Modell (alle Variablen der Menge) und löscht Variablen nach der vom Anwender eingestellten – mit Beginn bei der Letzten.
3. **Schrittweise Selektion** beginnt mit dem Basismodell und – bei jedem Schritt – fügt die Variable hinzu, die den höchsten Beitrag zur Aussagekraft des vom vorherigen Schritt gewählten Modells nach dem partialen *F* Test bietet, wenn dieser Beitrag höher als die voreingestellte Ebene ausfällt, die simultan mit der Variablenselektionsmethode gewählt wurde.Es nennt sich **Signifikanz Grenzen – Eingang** und wird per Default bei 0.15 eingestellt. Nach der Addition jeder Variable wird vorher geprüft und die Variable mit dem kleinsten Beitrag wird fallen gelassen, wenn seine Wald Statistik unter der voreingestelletn Ebene ausfällt (**Signifikanz Grenzen – Ausgang** mit dem Defaultwert 0.15). Die Selektion endet auch, wenn sich eine Variable 3 Mal (d.h. einbezogen und ausgeschlossen) kreist.

4. **Vorwärts Variable Selektion** ist eine Vereinfachung der **Schrittweise Selektion,** wo keine Variablen nach ihrer Addition fallen gelassen werden.
5. **Rückwärts Variable Selektion** ähnelt der **Vorwärts Variable Selektion** aber der Vorgang beginnt mit dem vollständigen Modell und Variablen werden nach den Kriterien der **Schrittweisen Selektion** fallen gelassen.
6. **R-Quadrat Selektion** führt Regressionen auf allen Untermengen der ausgewählten Variablenmenge aus.

Logistische Regression

Für logistische Regression muß die AV kategorisch sein. ANGOSS KnowledgeSTUDIO schätzt binäre oder polytomöse logistische Regression. Parameter werden nach der maximalen Wahrscheinlichektismethode errechnet.

Y nimmt Werte in $\{v_0,...,v_d\}$ mit v_0 als Referenzwert:

$p_j(x_i) = P(Y = v_j \mid x_i)$, j=0,…,p;

$$g_j(x_i) = \ln\frac{p_j(x_i)}{p_0(x_i)} = b_{j0} + b_{j1}x_{i1} + ... + b_{jp}x_{ip},\ j=1,\ldots,p;$$

wo $x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \cdot \\ \cdot \\ \cdot \\ x_{ip} \end{bmatrix}$;

Ergo,

$$p_0(x_i) = \frac{1}{1+\sum_{k=1}^{p} e^{g_k(x_i)}} \quad \text{und} \quad p_j(x_i) = \frac{e^{g_j(x_i)}}{1+\sum_{k=1}^{p} e^{g_k(x_i)}} \quad \text{für } j=1,\ldots,p;$$

m_i ist der Wertindex von y_i, d.h. $m_i = m$ if $y_i = v_m$. Dann ist die Log-Wahrscheinlichkeitsfunktion

$$L(b) = \sum_{i=1}^{N} \ln p_{m_i}(x_i) = \sum_{i=1}^{N} \{g_{m_i}(x_i) - \ln[1+\sum_{k=1}^{p} e^{g_k(x_i)}]\}$$

Diese Funktion wird nach der zugeordneten Gefälle Methode maximiert.

Wenn die AV binär ist, hat die Log-Wahrscheinlichkeit eine einfache Form:

$$L(b) = \sum_{i=1}^{N} \{y_i \ln p(x_i) + (1-y_i)\ln[1-p(x_i)]\}$$

Die Ausgabe einer logistischen Regression hat bei ANGOSS KnowledgeSTUDIO den folgenden Format:

Sequenz Statistik			
Abhängige Variable:	*Name*		
Entropie Erläutert:	$1-\frac{L(b_0)}{L(b)}$		
Chi-Quadrat:	$G=-2[L(b_0)-L(b)]$	Chi-Square Freiheitsgrad:	*p*
P-Wert:	$c^2(G;p)$		

Zusammenfassung der Modellpassung

Quelle	Negative 2 Log-Wahrscheinlichkeit	DF
Nur Kreuzung	$-2L(b_0)$	-
Volles Modell	$-2L(b)$	p

Unabhängige Variable Statistik

Abhängiger Variablenwert				
Variable Name / Wert	**Modell Parameter**	**Parameter Standard Fehler**	**Wald Statistik**	**Signifikanz**
Werte für kategorische Variablen	b_i	$SE(b_i)=\Sigma(b)_{ii}$	$\left[\frac{b_i}{SE(b_i)}\right]^2$	$c^2\left(\left[\frac{b_i}{SE(b_i)}\right]^2,1\right)$

Oben:

$L(b_0)$ ist die Log-Wahrscheinlichkeit für die Regression lediglich mit Kreuzung;

$\Sigma(b) = -\left[\frac{\partial^2 L}{\partial b^2}\right]^{-1}$ wo $\frac{\partial^2 L}{\partial b^2}$ ist der geschätzte Hessian der Log-Wahrscheinlichkeit;

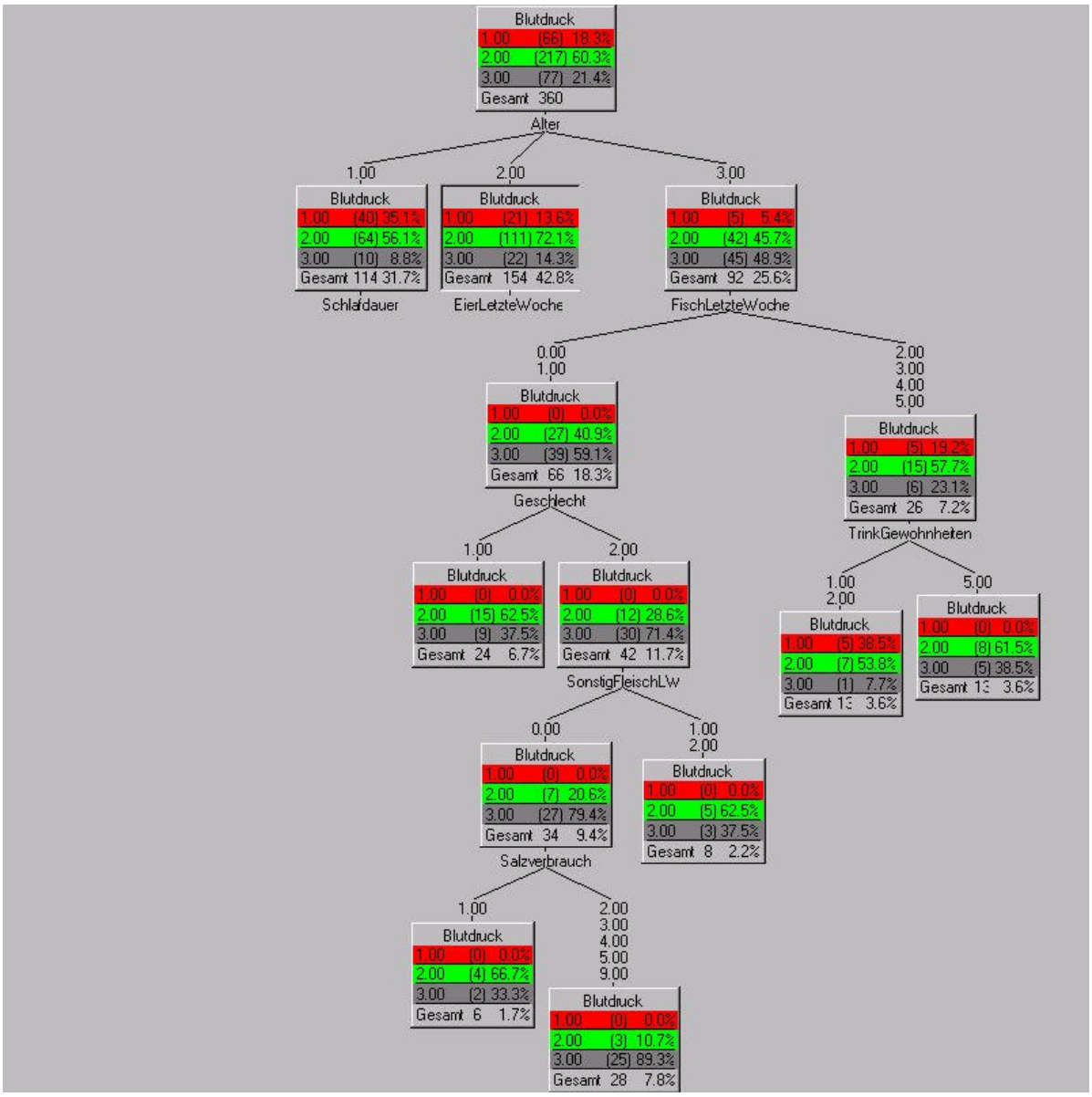
Wenn AV mehr als zwei Werte annimmt wird die Tabelle der **Unabhängigen Variable Statistik** für jeden Wert ausgegeben, außer der Referenz, die per Default die erste nach der alphabetischen Reihenfolge einnimmt.

Variable Selektionsmethoden sind mit den methoden der linearen Regression identisch, außer dass der Test Wahrscheinlichkeitsverhältnis zur Einbeziehung und zum Ausschluß der Varaiblen verwendet wird.

Anlage III

XML Beispiel

Der unten aufgeführte Entscheidungsbaum behandelt eine Datenmenge zur Untersuchung der Einflüsse auf den Blutdruck. Die Werte der abhängigen Variablen 1,2,3



verstehen sich als niedrig, mittel und hoch. Der Baum ist in der Darstellung auf die Verzweigung beschränkt, die die signifikantesten Einflüsse auf den „hohen“ Blutdruck bilden. Auf der Basis eines Vorhersagemodells (Entscheidungsbaum) wurde der folgende XML-Code generiert.

```
<?xml version="1.0"?>
<!DOCTYPE predictive-model [
 <!ELEMENT predictive-model (parameters,variables+,attribute-info,(neural-
model|RBF-model|PNN-model|cluster-analysis|kmeans-analysis|tree-model))>
 <!ATTLIST predictive-model
  is-cluster CDATA #REQUIRED
  is-valid CDATA #REQUIRED
 >
 <!ELEMENT parameters (parameter*)>
 <!ELEMENT parameter (value,limits)>
```

```
 <!ATTLIST parameter
  name CDATA #REQUIRED
  read-only CDATA #IMPLIED
 >
 <!ELEMENT value (#PCDATA|open|closed)*>
 <!ATTLIST value
  name CDATA #IMPLIED
  type (missing|null|other|bool|integer|real|string|time|time-of-
day|angle|color|float) #REQUIRED
 >
 <!ELEMENT open (#PCDATA)>
 <!ELEMENT closed (#PCDATA)>
 <!ELEMENT limits (value,value,value)>
 <!ELEMENT variables (variable*)>
 <!ATTLIST variables
  class (independent|dependent|predicted) #REQUIRED
 >
 <!ELEMENT variable (use,base,index,map,normalize,weight-type,denormal-
type)>
 <!ATTLIST variable
  name CDATA #REQUIRED
 >
 <!ELEMENT use (#PCDATA)>
 <!ELEMENT base (#PCDATA)>
 <!ELEMENT index (#PCDATA)>
 <!ELEMENT map (map-entry*)>
 <!ATTLIST map
  name CDATA #IMPLIED
  type (data|color|cost|unknown) #REQUIRED
 >
 <!ELEMENT map-entry (value,value)>
 <!ELEMENT normalize (type,width,missing-values,offset*,scale*,power-
exponent*,log-base*,clamp-lo*,clamp-hi*,data-summary*,value*,set-value*,res-
value*)>
 <!ELEMENT type (#PCDATA)>
 <!ELEMENT width (#PCDATA)>
 <!ELEMENT missing-values (#PCDATA)>
 <!ELEMENT offset (#PCDATA)>
 <!ELEMENT scale (#PCDATA)>
 <!ELEMENT power-exponent (#PCDATA)>
 <!ELEMENT log-base (#PCDATA)>
 <!ELEMENT clamp-lo (#PCDATA)>
 <!ELEMENT clamp-hi (#PCDATA)>
 <!ELEMENT data-summary (histogram*,data-summary-statistics*)>
 <!ATTLIST data-summary
  name CDATA #IMPLIED
 >
 <!ELEMENT data-summary-statistics (data-summary-statistic*)>
 <!ELEMENT data-summary-statistic (valid,value)>
 <!ATTLIST data-summary-statistic
  name CDATA #REQUIRED
 >
 <!ELEMENT valid (#PCDATA)>
 <!ELEMENT histogram (hist-entry*)>
 <!ELEMENT hist-entry (value,weighted,unweighted)>
 <!ELEMENT weighted (#PCDATA)>
 <!ELEMENT unweighted (#PCDATA)>
 <!ELEMENT set-value (#PCDATA)>
 <!ELEMENT res-value (#PCDATA)>
 <!ELEMENT weight-type (#PCDATA)>
```

```
 <!ELEMENT denormal-type (#PCDATA)>
 <!ELEMENT attribute-info (cluster-metric,metric-power,dv-complexity,map-
collection,attribute*)>
 <!ELEMENT cluster-metric (#PCDATA)>
 <!ELEMENT metric-power (#PCDATA)>
 <!ELEMENT dv-complexity (#PCDATA)>
 <!ELEMENT map-collection (map*)>
 <!ELEMENT attribute (state,original-name?,maps?,weight?,attribute-
settings*)>
 <!ATTLIST attribute
  name CDATA #IMPLIED
  index CDATA #IMPLIED
 >
 <!ELEMENT state (#PCDATA)>
 <!ELEMENT original-name (#PCDATA)>
 <!ELEMENT maps (mapref*)>
 <!ELEMENT weight (#PCDATA)>
 <!ELEMENT mapref (#PCDATA)>
 <!ATTLIST mapref
  type CDATA #REQUIRED
 >
 <!ELEMENT attribute-settings
(include,grouping,intervals,signifigance,display,missing-values,break-
apart,maps*)>
 <!ATTLIST attribute-settings
  state (independent|dependent|weight) #REQUIRED
 >
 <!ELEMENT include (#PCDATA)>
 <!ELEMENT grouping (#PCDATA)>
 <!ELEMENT intervals (#PCDATA)>
 <!ATTLIST intervals
  type (unspecified|static|dynamic|user-defined|binary|invalid) #REQUIRED
 >
 <!ELEMENT signifigance (#PCDATA)>
 <!ELEMENT display (#PCDATA)>
 <!ELEMENT break-apart (#PCDATA)>
 <!ELEMENT tree-model (node-count,total-recs,((average,stddev,root-
variance,total-variance)|vector)?,accuracy,error,value-map,node*)>
 <!ELEMENT node-count (#PCDATA)>
 <!ELEMENT total-recs (#PCDATA)>
 <!ELEMENT vector (#PCDATA)>
 <!ATTLIST vector
  type CDATA #REQUIRED
  dim CDATA #REQUIRED
  name CDATA #IMPLIED
 >
 <!ELEMENT accuracy (#PCDATA)>
 <!ELEMENT error (#PCDATA)>
 <!ELEMENT value-map (value*)>
 <!ELEMENT node (node-dist,vector,branch-values,split*)>
 <!ATTLIST node
  iv CDATA #REQUIRED
 >
 <!ELEMENT node-dist (total,continuous-dist,discrete-dist)>
 <!ELEMENT total (#PCDATA)>
 <!ELEMENT continuous-dist (unweighted,weighted,mean,stddev)>
 <!ELEMENT mean (#PCDATA)>
 <!ELEMENT stddev (#PCDATA)>
 <!ELEMENT discrete-dist (value*)>
 <!ELEMENT branch-values (value*)>
```

```
 <!ELEMENT split (split-map,node*)>
 <!ATTLIST split
  IV CDATA #REQUIRED
 >
 <!ELEMENT split-map (value*)>
 <!ATTLIST split-map
  branches CDATA #REQUIRED
 >
]>
<predictive-model is-cluster="false" is-valid="true">
 <parameters>
  <parameter name="RecordCount" read-only="false">
   <value name="val" type="integer">0</value>
   <limits>
    <value name="min" type="integer">20</value>
    <value name="def" type="integer">0</value>
    <value name="max" type="integer">1000000</value>
   </limits>
  </parameter>
  <parameter name="MemorySoftLimit" read-only="false">
   <value name="val" type="integer">8388608</value>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="integer">8388608</value>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="MemoryHardLimit" read-only="false">
   <value name="val" type="integer">67108864</value>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="integer">67108864</value>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="RecordSelectFirstN" read-only="false">
   <value name="val" type="integer">0</value>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="integer">0</value>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="DataDumpFile" read-only="false">
   <value name="val" type="string"></value>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="string"></value>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="ModelDumpFile" read-only="false">
   <value name="val" type="string"></value>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="string"></value>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="ValidationHoldBack" read-only="false">
```

```
 <value name="val" type="real">0</value>
 <limits>
  <value name="min" type="real">0</value>
  <value name="def" type="real">0</value>
  <value name="max" type="real">0.99</value>
 </limits>
</parameter>
<parameter name="ValidationRate" read-only="false">
 <value name="val" type="integer">0</value>
 <limits>
  <value name="min" type="integer">0</value>
  <value name="def" type="integer">0</value>
  <value name="max" type="missing"/>
 </limits>
</parameter>
<parameter name="ValidationStopCount" read-only="false">
 <value name="val" type="integer">25</value>
 <limits>
  <value name="min" type="integer">1</value>
  <value name="def" type="integer">25</value>
  <value name="max" type="missing"/>
 </limits>
</parameter>
<parameter name="ValidationScoreSmoothing" read-only="false">
 <value name="val" type="real">0</value>
 <limits>
  <value name="min" type="real">0</value>
  <value name="def" type="real">0</value>
  <value name="max" type="real">0.9999</value>
 </limits>
</parameter>
<parameter name="MeasureCovariance" read-only="false">
 <value name="val" type="bool">false</value>
 <limits>
  <value name="min" type="missing"/>
  <value name="def" type="bool">false</value>
  <value name="max" type="missing"/>
 </limits>
</parameter>
<parameter name="MeasureIVCollinearity" read-only="false">
 <value name="val" type="bool">false</value>
 <limits>
  <value name="min" type="missing"/>
  <value name="def" type="bool">false</value>
  <value name="max" type="missing"/>
 </limits>
</parameter>
<parameter name="MeasureDVCollinearity" read-only="false">
 <value name="val" type="bool">false</value>
 <limits>
  <value name="min" type="missing"/>
  <value name="def" type="bool">false</value>
  <value name="max" type="missing"/>
 </limits>
</parameter>
<parameter name="RandomSeed" read-only="false">
 <value name="val" type="missing"/>
 <limits>
  <value name="min" type="missing"/>
  <value name="def" type="missing"/>
```

```
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="DecisionTree" read-only="false">
   <value name="val" type="unknown"/>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="missing"/>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="__DecisionTreePointer__" read-only="false">
   <value name="val" type="missing"/>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="missing"/>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="Algorithm" read-only="false">
   <value name="val"
type="string">KnowledgeTREE.Algorithms.KnowledgeSEEKER.1</value>
   <limits>
    <value name="min" type="missing"/>
    <value name="def"
type="string">KnowledgeTREE.Algorithms.KnowledgeSEEKER.1</value>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="Measure" read-only="false">
   <value name="val" type="string">KnowledgeTREE.Measure.Adjusted.1</value>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="string">KnowledgeTREE.Measure.Adjusted.1</value>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="Accuracy" read-only="true">
   <value name="val" type="real">0</value>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="real">0</value>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="ErrorRate" read-only="true">
   <value name="val" type="real">0</value>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="real">0</value>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="MultiTree" read-only="false">
   <value name="val" type="bool">false</value>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="bool">true</value>
    <value name="max" type="missing"/>
   </limits>
```

```
  </parameter>
  <parameter name="VoteMethod" read-only="false">
   <value name="val" type="string">best</value>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="string">best</value>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="ModelGUID" read-only="false">
   <value name="val" type="string">&#123;EE191993-A234-11D1-8B75-
00C0F01035FC&#125;</value>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="missing"/>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
  <parameter name="LastTrainTime" read-only="false">
   <value name="val" type="time">3.6493e+007</value>
   <limits>
    <value name="min" type="missing"/>
    <value name="def" type="time">3.6493e+007</value>
    <value name="max" type="missing"/>
   </limits>
  </parameter>
 </parameters>
 <variables class="independent">
  <variable name="MilchTyp">
   <use>true</use>
   <base>-1</base>
   <index>0</index>
   <map name="" type="data"/>
   <normalize>
    <type>rank</type>
    <width>1</width>
    <missing-values>use</missing-values>
    <data-summary>
     <histogram>
      <hist-entry>
       <value type="integer">1</value>
       <weighted>85</weighted>
       <unweighted>85</unweighted>
      </hist-entry>
      <hist-entry>
       <value type="integer">2</value>
       <weighted>231</weighted>
       <unweighted>231</unweighted>
      </hist-entry>
      <hist-entry>
       <value type="integer">3</value>
       <weighted>19</weighted>
       <unweighted>19</unweighted>
      </hist-entry>
      <hist-entry>
       <value type="integer">4</value>
       <weighted>2</weighted>
       <unweighted>2</unweighted>
      </hist-entry>
      <hist-entry>
```

```
      <value type="integer">5</value>
      <weighted>23</weighted>
      <unweighted>23</unweighted>
     </hist-entry>
    </histogram>
    <data-summary-statistics>
     <data-summary-statistic name="cardinality">
      <valid>true</valid>
      <value type="integer">5</value>
     </data-summary-statistic>
     <data-summary-statistic name="missing">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
     <data-summary-statistic name="min">
      <valid>true</valid>
      <value type="integer">1</value>
     </data-summary-statistic>
     <data-summary-statistic name="max">
      <valid>true</valid>
      <value type="integer">5</value>
     </data-summary-statistic>
     <data-summary-statistic name="mean">
      <valid>true</valid>
      <value type="real">2.01944444444</value>
     </data-summary-statistic>
     <data-summary-statistic name="stddev">
      <valid>true</valid>
      <value type="real">0.942443763628</value>
     </data-summary-statistic>
     <data-summary-statistic name="unique">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
     <data-summary-statistic name="maxReal">
      <valid>true</valid>
      <value type="real">5</value>
     </data-summary-statistic>
     <data-summary-statistic name="minReal">
      <valid>true</valid>
      <value type="real">1</value>
     </data-summary-statistic>
    </data-summary-statistics>
   </data-summary>
   <value name="ref-value" type="missing"/>
   <set-value>1</set-value>
   <res-value>0</res-value>
  </normalize>
  <weight-type>none</weight-type>
  <denormal-type>integer</denormal-type>
 </variable>
 <variable name="LindLetzteWoche">
  <use>true</use>
  <base>-1</base>
  <index>2</index>
  <map name="" type="data"/>
  <normalize>
   <type>rank</type>
   <width>1</width>
   <missing-values>use</missing-values>
```

```
<data-summary>
 <histogram>
  <hist-entry>
   <value type="integer">0</value>
   <weighted>28</weighted>
   <unweighted>28</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">1</value>
   <weighted>84</weighted>
   <unweighted>84</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">2</value>
   <weighted>106</weighted>
   <unweighted>106</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">3</value>
   <weighted>66</weighted>
   <unweighted>66</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">4</value>
   <weighted>45</weighted>
   <unweighted>45</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">5</value>
   <weighted>16</weighted>
   <unweighted>16</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">6</value>
   <weighted>6</weighted>
   <unweighted>6</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">7</value>
   <weighted>9</weighted>
   <unweighted>9</unweighted>
  </hist-entry>
 </histogram>
 <data-summary-statistics>
  <data-summary-statistic name="cardinality">
   <valid>true</valid>
   <value type="integer">8</value>
  </data-summary-statistic>
  <data-summary-statistic name="missing">
   <valid>true</valid>
   <value type="real">0</value>
  </data-summary-statistic>
  <data-summary-statistic name="min">
   <valid>true</valid>
   <value type="integer">0</value>
  </data-summary-statistic>
  <data-summary-statistic name="max">
   <valid>true</valid>
   <value type="integer">7</value>
  </data-summary-statistic>
```

```
      <data-summary-statistic name="mean">
       <valid>true</valid>
       <value type="real">2.36944444444</value>
      </data-summary-statistic>
      <data-summary-statistic name="stddev">
       <valid>true</valid>
       <value type="real">1.54583084626</value>
      </data-summary-statistic>
      <data-summary-statistic name="unique">
       <valid>true</valid>
       <value type="real">0</value>
      </data-summary-statistic>
      <data-summary-statistic name="maxReal">
       <valid>true</valid>
       <value type="real">7</value>
      </data-summary-statistic>
      <data-summary-statistic name="minReal">
       <valid>true</valid>
       <value type="real">0</value>
      </data-summary-statistic>
     </data-summary-statistics>
    </data-summary>
    <value name="ref-value" type="missing"/>
    <set-value>1</set-value>
    <res-value>0</res-value>
   </normalize>
   <weight-type>none</weight-type>
   <denormal-type>integer</denormal-type>
  </variable>
  <variable name="FleischLetzteW">
   <use>true</use>
   <base>-1</base>
   <index>3</index>
   <map name="" type="data"/>
   <normalize>
    <type>rank</type>
    <width>1</width>
    <missing-values>use</missing-values>
    <data-summary>
     <histogram>
      <hist-entry>
       <value type="integer">0</value>
       <weighted>133</weighted>
       <unweighted>133</unweighted>
      </hist-entry>
      <hist-entry>
       <value type="integer">1</value>
       <weighted>130</weighted>
       <unweighted>130</unweighted>
      </hist-entry>
      <hist-entry>
       <value type="integer">2</value>
       <weighted>55</weighted>
       <unweighted>55</unweighted>
      </hist-entry>
      <hist-entry>
       <value type="integer">3</value>
       <weighted>22</weighted>
       <unweighted>22</unweighted>
      </hist-entry>
```

```
  <hist-entry>
   <value type="integer">4</value>
   <weighted>9</weighted>
   <unweighted>9</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">5</value>
   <weighted>7</weighted>
   <unweighted>7</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">6</value>
   <weighted>2</weighted>
   <unweighted>2</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">7</value>
   <weighted>2</weighted>
   <unweighted>2</unweighted>
  </hist-entry>
 </histogram>
 <data-summary-statistics>
  <data-summary-statistic name="cardinality">
   <valid>true</valid>
   <value type="integer">8</value>
  </data-summary-statistic>
  <data-summary-statistic name="missing">
   <valid>true</valid>
   <value type="real">0</value>
  </data-summary-statistic>
  <data-summary-statistic name="min">
   <valid>true</valid>
   <value type="integer">0</value>
  </data-summary-statistic>
  <data-summary-statistic name="max">
   <valid>true</valid>
   <value type="integer">7</value>
  </data-summary-statistic>
  <data-summary-statistic name="mean">
   <valid>true</valid>
   <value type="real">1.11944444444</value>
  </data-summary-statistic>
  <data-summary-statistic name="stddev">
   <valid>true</valid>
   <value type="real">1.27747126705</value>
  </data-summary-statistic>
  <data-summary-statistic name="unique">
   <valid>true</valid>
   <value type="real">0</value>
  </data-summary-statistic>
  <data-summary-statistic name="maxReal">
   <valid>true</valid>
   <value type="real">7</value>
  </data-summary-statistic>
  <data-summary-statistic name="minReal">
   <valid>true</valid>
   <value type="real">0</value>
  </data-summary-statistic>
 </data-summary-statistics>
</data-summary>
```

```
  <value name="ref-value" type="missing"/>
  <set-value>1</set-value>
  <res-value>0</res-value>
 </normalize>
 <weight-type>none</weight-type>
 <denormal-type>integer</denormal-type>
</variable>
<variable name="Gefl&#252;gelLetzteW">
 <use>true</use>
 <base>-1</base>
 <index>4</index>
 <map name="" type="data"/>
 <normalize>
  <type>rank</type>
  <width>1</width>
  <missing-values>use</missing-values>
  <data-summary>
   <histogram>
    <hist-entry>
     <value type="integer">0</value>
     <weighted>57</weighted>
     <unweighted>57</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="integer">1</value>
     <weighted>137</weighted>
     <unweighted>137</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="integer">2</value>
     <weighted>93</weighted>
     <unweighted>93</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="integer">3</value>
     <weighted>51</weighted>
     <unweighted>51</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="integer">4</value>
     <weighted>15</weighted>
     <unweighted>15</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="integer">5</value>
     <weighted>2</weighted>
     <unweighted>2</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="integer">6</value>
     <weighted>3</weighted>
     <unweighted>3</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="integer">7</value>
     <weighted>2</weighted>
     <unweighted>2</unweighted>
    </hist-entry>
   </histogram>
   <data-summary-statistics>
```

```
      <data-summary-statistic name="cardinality">
       <valid>true</valid>
       <value type="integer">8</value>
      </data-summary-statistic>
      <data-summary-statistic name="missing">
       <valid>true</valid>
       <value type="real">0</value>
      </data-summary-statistic>
      <data-summary-statistic name="min">
       <valid>true</valid>
       <value type="integer">0</value>
      </data-summary-statistic>
      <data-summary-statistic name="max">
       <valid>true</valid>
       <value type="integer">7</value>
      </data-summary-statistic>
      <data-summary-statistic name="mean">
       <valid>true</valid>
       <value type="real">1.60555555556</value>
      </data-summary-statistic>
      <data-summary-statistic name="stddev">
       <valid>true</valid>
       <value type="real">1.22188594457</value>
      </data-summary-statistic>
      <data-summary-statistic name="unique">
       <valid>true</valid>
       <value type="real">0</value>
      </data-summary-statistic>
      <data-summary-statistic name="maxReal">
       <valid>true</valid>
       <value type="real">7</value>
      </data-summary-statistic>
      <data-summary-statistic name="minReal">
       <valid>true</valid>
       <value type="real">0</value>
      </data-summary-statistic>
     </data-summary-statistics>
    </data-summary>
    <value name="ref-value" type="missing"/>
    <set-value>1</set-value>
    <res-value>0</res-value>
   </normalize>
   <weight-type>none</weight-type>
   <denormal-type>integer</denormal-type>
  </variable>
  <variable name="FischLetzteWoche">
   <use>true</use>
   <base>-1</base>
   <index>5</index>
   <map name="" type="data"/>
   <normalize>
    <type>rank</type>
    <width>1</width>
    <missing-values>use</missing-values>
    <data-summary>
     <histogram>
      <hist-entry>
       <value type="integer">0</value>
       <weighted>150</weighted>
       <unweighted>150</unweighted>
```

```
  </hist-entry>
  <hist-entry>
   <value type="integer">1</value>
   <weighted>121</weighted>
   <unweighted>121</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">2</value>
   <weighted>50</weighted>
   <unweighted>50</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">3</value>
   <weighted>24</weighted>
   <unweighted>24</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">4</value>
   <weighted>7</weighted>
   <unweighted>7</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">5</value>
   <weighted>4</weighted>
   <unweighted>4</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">6</value>
   <weighted>3</weighted>
   <unweighted>3</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">7</value>
   <weighted>1</weighted>
   <unweighted>1</unweighted>
  </hist-entry>
 </histogram>
 <data-summary-statistics>
  <data-summary-statistic name="cardinality">
   <valid>true</valid>
   <value type="integer">8</value>
  </data-summary-statistic>
  <data-summary-statistic name="missing">
   <valid>true</valid>
   <value type="real">0</value>
  </data-summary-statistic>
  <data-summary-statistic name="min">
   <valid>true</valid>
   <value type="integer">0</value>
  </data-summary-statistic>
  <data-summary-statistic name="max">
   <valid>true</valid>
   <value type="integer">7</value>
  </data-summary-statistic>
  <data-summary-statistic name="mean">
   <valid>true</valid>
   <value type="real">1.01666666667</value>
  </data-summary-statistic>
  <data-summary-statistic name="stddev">
   <valid>true</valid>
```

```
      <value type="real">1.21950261563</value>
     </data-summary-statistic>
     <data-summary-statistic name="unique">
      <valid>true</valid>
      <value type="real">1</value>
     </data-summary-statistic>
     <data-summary-statistic name="maxReal">
      <valid>true</valid>
      <value type="real">7</value>
     </data-summary-statistic>
     <data-summary-statistic name="minReal">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
    </data-summary-statistics>
   </data-summary>
   <value name="ref-value" type="missing"/>
   <set-value>1</set-value>
   <res-value>0</res-value>
  </normalize>
  <weight-type>none</weight-type>
  <denormal-type>integer</denormal-type>
 </variable>
 <variable name="SonstigFleischLW">
  <use>true</use>
  <base>-1</base>
  <index>7</index>
  <map name="" type="data"/>
  <normalize>
   <type>rank</type>
   <width>1</width>
   <missing-values>use</missing-values>
   <data-summary>
    <histogram>
     <hist-entry>
      <value type="integer">0</value>
      <weighted>256</weighted>
      <unweighted>256</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">1</value>
      <weighted>72</weighted>
      <unweighted>72</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">2</value>
      <weighted>22</weighted>
      <unweighted>22</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">3</value>
      <weighted>9</weighted>
      <unweighted>9</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">7</value>
      <weighted>1</weighted>
      <unweighted>1</unweighted>
     </hist-entry>
    </histogram>
```

```
    <data-summary-statistics>
     <data-summary-statistic name="cardinality">
      <valid>true</valid>
      <value type="integer">5</value>
     </data-summary-statistic>
     <data-summary-statistic name="missing">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
     <data-summary-statistic name="min">
      <valid>true</valid>
      <value type="integer">0</value>
     </data-summary-statistic>
     <data-summary-statistic name="max">
      <valid>true</valid>
      <value type="integer">7</value>
     </data-summary-statistic>
     <data-summary-statistic name="mean">
      <valid>true</valid>
      <value type="real">0.416666666667</value>
     </data-summary-statistic>
     <data-summary-statistic name="stddev">
      <valid>true</valid>
      <value type="real">0.79605573635</value>
     </data-summary-statistic>
     <data-summary-statistic name="unique">
      <valid>true</valid>
      <value type="real">1</value>
     </data-summary-statistic>
     <data-summary-statistic name="maxReal">
      <valid>true</valid>
      <value type="real">7</value>
     </data-summary-statistic>
     <data-summary-statistic name="minReal">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
    </data-summary-statistics>
   </data-summary>
   <value name="ref-value" type="missing"/>
   <set-value>1</set-value>
   <res-value>0</res-value>
  </normalize>
  <weight-type>none</weight-type>
  <denormal-type>integer</denormal-type>
 </variable>
 <variable name="EierLetzteWoche">
  <use>true</use>
  <base>-1</base>
  <index>9</index>
  <map name="" type="data"/>
  <normalize>
   <type>rank</type>
   <width>1</width>
   <missing-values>use</missing-values>
   <data-summary>
    <histogram>
     <hist-entry>
      <value type="integer">0</value>
      <weighted>82</weighted>
```

```
  <unweighted>82</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">1</value>
  <weighted>101</weighted>
  <unweighted>101</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">2</value>
  <weighted>103</weighted>
  <unweighted>103</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">3</value>
  <weighted>38</weighted>
  <unweighted>38</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">4</value>
  <weighted>14</weighted>
  <unweighted>14</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">5</value>
  <weighted>3</weighted>
  <unweighted>3</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">6</value>
  <weighted>4</weighted>
  <unweighted>4</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">7</value>
  <weighted>13</weighted>
  <unweighted>13</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">9</value>
  <weighted>2</weighted>
  <unweighted>2</unweighted>
 </hist-entry>
</histogram>
<data-summary-statistics>
 <data-summary-statistic name="cardinality">
  <valid>true</valid>
  <value type="integer">9</value>
 </data-summary-statistic>
 <data-summary-statistic name="missing">
  <valid>true</valid>
  <value type="real">0</value>
 </data-summary-statistic>
 <data-summary-statistic name="min">
  <valid>true</valid>
  <value type="integer">0</value>
 </data-summary-statistic>
 <data-summary-statistic name="max">
  <valid>true</valid>
  <value type="integer">9</value>
 </data-summary-statistic>
```

```
      <data-summary-statistic name="mean">
       <valid>true</valid>
       <value type="real">1.73611111111</value>
      </data-summary-statistic>
      <data-summary-statistic name="stddev">
       <valid>true</valid>
       <value type="real">1.67891276296</value>
      </data-summary-statistic>
      <data-summary-statistic name="unique">
       <valid>true</valid>
       <value type="real">0</value>
      </data-summary-statistic>
      <data-summary-statistic name="maxReal">
       <valid>true</valid>
       <value type="real">9</value>
      </data-summary-statistic>
      <data-summary-statistic name="minReal">
       <valid>true</valid>
       <value type="real">0</value>
      </data-summary-statistic>
     </data-summary-statistics>
    </data-summary>
    <value name="ref-value" type="missing"/>
    <set-value>1</set-value>
    <res-value>0</res-value>
   </normalize>
   <weight-type>none</weight-type>
   <denormal-type>integer</denormal-type>
  </variable>
  <variable name="Fleisch2MalT&#228;glLW">
   <use>true</use>
   <base>-1</base>
   <index>10</index>
   <map name="" type="data"/>
   <normalize>
    <type>rank</type>
    <width>1</width>
    <missing-values>use</missing-values>
    <data-summary>
     <histogram>
      <hist-entry>
       <value type="integer">1</value>
       <weighted>22</weighted>
       <unweighted>22</unweighted>
      </hist-entry>
      <hist-entry>
       <value type="integer">2</value>
       <weighted>41</weighted>
       <unweighted>41</unweighted>
      </hist-entry>
      <hist-entry>
       <value type="integer">3</value>
       <weighted>29</weighted>
       <unweighted>29</unweighted>
      </hist-entry>
      <hist-entry>
       <value type="integer">4</value>
       <weighted>23</weighted>
       <unweighted>23</unweighted>
      </hist-entry>
```

```
  <hist-entry>
   <value type="integer">5</value>
   <weighted>26</weighted>
   <unweighted>26</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">6</value>
   <weighted>3</weighted>
   <unweighted>3</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">7</value>
   <weighted>19</weighted>
   <unweighted>19</unweighted>
  </hist-entry>
  <hist-entry>
   <value type="integer">9</value>
   <weighted>197</weighted>
   <unweighted>197</unweighted>
  </hist-entry>
 </histogram>
 <data-summary-statistics>
  <data-summary-statistic name="cardinality">
   <valid>true</valid>
   <value type="integer">8</value>
  </data-summary-statistic>
  <data-summary-statistic name="missing">
   <valid>true</valid>
   <value type="real">0</value>
  </data-summary-statistic>
  <data-summary-statistic name="min">
   <valid>true</valid>
   <value type="integer">1</value>
  </data-summary-statistic>
  <data-summary-statistic name="max">
   <valid>true</valid>
   <value type="integer">9</value>
  </data-summary-statistic>
  <data-summary-statistic name="mean">
   <valid>true</valid>
   <value type="real">6.49166666667</value>
  </data-summary-statistic>
  <data-summary-statistic name="stddev">
   <valid>true</valid>
   <value type="real">3.0272554841</value>
  </data-summary-statistic>
  <data-summary-statistic name="unique">
   <valid>true</valid>
   <value type="real">0</value>
  </data-summary-statistic>
  <data-summary-statistic name="maxReal">
   <valid>true</valid>
   <value type="real">9</value>
  </data-summary-statistic>
  <data-summary-statistic name="minReal">
   <valid>true</valid>
   <value type="real">1</value>
  </data-summary-statistic>
 </data-summary-statistics>
</data-summary>
```

```
   <value name="ref-value" type="missing"/>
   <set-value>1</set-value>
   <res-value>0</res-value>
  </normalize>
  <weight-type>none</weight-type>
  <denormal-type>integer</denormal-type>
 </variable>
 <variable name="Salzverbrauch">
  <use>true</use>
  <base>-1</base>
  <index>12</index>
  <map name="" type="data"/>
  <normalize>
   <type>rank</type>
   <width>1</width>
   <missing-values>use</missing-values>
   <data-summary>
    <histogram>
     <hist-entry>
      <value type="integer">1</value>
      <weighted>102</weighted>
      <unweighted>102</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">2</value>
      <weighted>123</weighted>
      <unweighted>123</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">3</value>
      <weighted>101</weighted>
      <unweighted>101</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">4</value>
      <weighted>21</weighted>
      <unweighted>21</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">5</value>
      <weighted>7</weighted>
      <unweighted>7</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">9</value>
      <weighted>6</weighted>
      <unweighted>6</unweighted>
     </hist-entry>
    </histogram>
    <data-summary-statistics>
     <data-summary-statistic name="cardinality">
      <valid>true</valid>
      <value type="integer">6</value>
     </data-summary-statistic>
     <data-summary-statistic name="missing">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
     <data-summary-statistic name="min">
      <valid>true</valid>
```

```
      <value type="integer">1</value>
     </data-summary-statistic>
     <data-summary-statistic name="max">
      <valid>true</valid>
      <value type="integer">9</value>
     </data-summary-statistic>
     <data-summary-statistic name="mean">
      <valid>true</valid>
      <value type="real">2.28888888889</value>
     </data-summary-statistic>
     <data-summary-statistic name="stddev">
      <valid>true</valid>
      <value type="real">1.30773950235</value>
     </data-summary-statistic>
     <data-summary-statistic name="unique">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
     <data-summary-statistic name="maxReal">
      <valid>true</valid>
      <value type="real">9</value>
     </data-summary-statistic>
     <data-summary-statistic name="minReal">
      <valid>true</valid>
      <value type="real">1</value>
     </data-summary-statistic>
    </data-summary-statistics>
   </data-summary>
   <value name="ref-value" type="missing"/>
   <set-value>1</set-value>
   <res-value>0</res-value>
  </normalize>
  <weight-type>none</weight-type>
  <denormal-type>integer</denormal-type>
 </variable>
 <variable name="ButterHaltiges">
  <use>true</use>
  <base>-1</base>
  <index>13</index>
  <map name="" type="data"/>
  <normalize>
   <type>rank</type>
   <width>1</width>
   <missing-values>use</missing-values>
   <data-summary>
    <histogram>
     <hist-entry>
      <value type="integer">1</value>
      <weighted>298</weighted>
      <unweighted>298</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">2</value>
      <weighted>39</weighted>
      <unweighted>39</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">3</value>
      <weighted>23</weighted>
      <unweighted>23</unweighted>
```

```
     </hist-entry>
    </histogram>
    <data-summary-statistics>
     <data-summary-statistic name="cardinality">
      <valid>true</valid>
      <value type="integer">3</value>
     </data-summary-statistic>
     <data-summary-statistic name="missing">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
     <data-summary-statistic name="min">
      <valid>true</valid>
      <value type="integer">1</value>
     </data-summary-statistic>
     <data-summary-statistic name="max">
      <valid>true</valid>
      <value type="integer">3</value>
     </data-summary-statistic>
     <data-summary-statistic name="mean">
      <valid>true</valid>
      <value type="real">1.23611111111</value>
     </data-summary-statistic>
     <data-summary-statistic name="stddev">
      <valid>true</valid>
      <value type="real">0.555876570826</value>
     </data-summary-statistic>
     <data-summary-statistic name="unique">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
     <data-summary-statistic name="maxReal">
      <valid>true</valid>
      <value type="real">3</value>
     </data-summary-statistic>
     <data-summary-statistic name="minReal">
      <valid>true</valid>
      <value type="real">1</value>
     </data-summary-statistic>
    </data-summary-statistics>
   </data-summary>
   <value name="ref-value" type="missing"/>
   <set-value>1</set-value>
   <res-value>0</res-value>
  </normalize>
  <weight-type>none</weight-type>
  <denormal-type>integer</denormal-type>
 </variable>
 <variable name="Schlafdauer">
  <use>true</use>
  <base>-1</base>
  <index>15</index>
  <map name="" type="data"/>
  <normalize>
   <type>rank</type>
   <width>1</width>
   <missing-values>use</missing-values>
   <data-summary>
    <histogram>
     <hist-entry>
```

```
 <value type="real">3</value>
 <weighted>3</weighted>
 <unweighted>3</unweighted>
</hist-entry>
<hist-entry>
 <value type="real">5</value>
 <weighted>1</weighted>
 <unweighted>1</unweighted>
</hist-entry>
<hist-entry>
 <value type="real">6</value>
 <weighted>1</weighted>
 <unweighted>1</unweighted>
</hist-entry>
<hist-entry>
 <value type="real">7</value>
 <weighted>7</weighted>
 <unweighted>7</unweighted>
</hist-entry>
<hist-entry>
 <value type="real">8</value>
 <weighted>22</weighted>
 <unweighted>22</unweighted>
</hist-entry>
<hist-entry>
 <value type="real">9</value>
 <weighted>6</weighted>
 <unweighted>6</unweighted>
</hist-entry>
<hist-entry>
 <value type="real">10</value>
 <weighted>42</weighted>
 <unweighted>42</unweighted>
</hist-entry>
<hist-entry>
 <value type="real">11</value>
 <weighted>8</weighted>
 <unweighted>8</unweighted>
</hist-entry>
<hist-entry>
 <value type="real">12</value>
 <weighted>99</weighted>
 <unweighted>99</unweighted>
</hist-entry>
<hist-entry>
 <value type="real">13</value>
 <weighted>125</weighted>
 <unweighted>125</unweighted>
</hist-entry>
<hist-entry>
 <value type="real">14</value>
 <weighted>18</weighted>
 <unweighted>18</unweighted>
</hist-entry>
<hist-entry>
 <value type="real">15</value>
 <weighted>17</weighted>
 <unweighted>17</unweighted>
</hist-entry>
<hist-entry>
```

```
     <value type="real">17</value>
     <weighted>9</weighted>
     <unweighted>9</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="real">20</value>
     <weighted>1</weighted>
     <unweighted>1</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="real">22</value>
     <weighted>1</weighted>
     <unweighted>1</unweighted>
    </hist-entry>
   </histogram>
   <data-summary-statistics>
    <data-summary-statistic name="cardinality">
     <valid>true</valid>
     <value type="integer">15</value>
    </data-summary-statistic>
    <data-summary-statistic name="missing">
     <valid>true</valid>
     <value type="real">0</value>
    </data-summary-statistic>
    <data-summary-statistic name="min">
     <valid>true</valid>
     <value type="real">3</value>
    </data-summary-statistic>
    <data-summary-statistic name="max">
     <valid>true</valid>
     <value type="real">22</value>
    </data-summary-statistic>
    <data-summary-statistic name="mean">
     <valid>true</valid>
     <value type="real">12.0055555556</value>
    </data-summary-statistic>
    <data-summary-statistic name="stddev">
     <valid>true</valid>
     <value type="real">2.25405168399</value>
    </data-summary-statistic>
    <data-summary-statistic name="unique">
     <valid>true</valid>
     <value type="real">4</value>
    </data-summary-statistic>
    <data-summary-statistic name="maxReal">
     <valid>true</valid>
     <value type="real">22</value>
    </data-summary-statistic>
    <data-summary-statistic name="minReal">
     <valid>true</valid>
     <value type="real">3</value>
    </data-summary-statistic>
   </data-summary-statistics>
  </data-summary>
  <value name="ref-value" type="missing"/>
  <set-value>1</set-value>
  <res-value>0</res-value>
 </normalize>
 <weight-type>none</weight-type>
 <denormal-type>real</denormal-type>
```

```
</variable>
<variable name="Rauchen">
 <use>true</use>
 <base>-1</base>
 <index>16</index>
 <map name="" type="data"/>
 <normalize>
  <type>rank</type>
  <width>1</width>
  <missing-values>use</missing-values>
  <data-summary>
   <histogram>
    <hist-entry>
     <value type="integer">1</value>
     <weighted>114</weighted>
     <unweighted>114</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="integer">2</value>
     <weighted>11</weighted>
     <unweighted>11</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="integer">3</value>
     <weighted>140</weighted>
     <unweighted>140</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="integer">4</value>
     <weighted>94</weighted>
     <unweighted>94</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="integer">8</value>
     <weighted>1</weighted>
     <unweighted>1</unweighted>
    </hist-entry>
   </histogram>
   <data-summary-statistics>
    <data-summary-statistic name="cardinality">
     <valid>true</valid>
     <value type="integer">5</value>
    </data-summary-statistic>
    <data-summary-statistic name="missing">
     <valid>true</valid>
     <value type="real">0</value>
    </data-summary-statistic>
    <data-summary-statistic name="min">
     <valid>true</valid>
     <value type="integer">1</value>
    </data-summary-statistic>
    <data-summary-statistic name="max">
     <valid>true</valid>
     <value type="integer">8</value>
    </data-summary-statistic>
    <data-summary-statistic name="mean">
     <valid>true</valid>
     <value type="real">2.61111111111</value>
    </data-summary-statistic>
    <data-summary-statistic name="stddev">
```

```
      <valid>true</valid>
      <value type="real">1.21682214954</value>
     </data-summary-statistic>
     <data-summary-statistic name="unique">
      <valid>true</valid>
      <value type="real">1</value>
     </data-summary-statistic>
     <data-summary-statistic name="maxReal">
      <valid>true</valid>
      <value type="real">8</value>
     </data-summary-statistic>
     <data-summary-statistic name="minReal">
      <valid>true</valid>
      <value type="real">1</value>
     </data-summary-statistic>
    </data-summary-statistics>
   </data-summary>
   <value name="ref-value" type="missing"/>
   <set-value>1</set-value>
   <res-value>0</res-value>
  </normalize>
  <weight-type>none</weight-type>
  <denormal-type>integer</denormal-type>
 </variable>
 <variable name="TrinkGewohnheiten">
  <use>true</use>
  <base>-1</base>
  <index>17</index>
  <map name="" type="data"/>
  <normalize>
   <type>rank</type>
   <width>1</width>
   <missing-values>use</missing-values>
   <data-summary>
    <histogram>
     <hist-entry>
      <value type="integer">1</value>
      <weighted>213</weighted>
      <unweighted>213</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">2</value>
      <weighted>43</weighted>
      <unweighted>43</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">4</value>
      <weighted>11</weighted>
      <unweighted>11</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">5</value>
      <weighted>93</weighted>
      <unweighted>93</unweighted>
     </hist-entry>
    </histogram>
    <data-summary-statistics>
     <data-summary-statistic name="cardinality">
      <valid>true</valid>
      <value type="integer">4</value>
```

```
      </data-summary-statistic>
      <data-summary-statistic name="missing">
       <valid>true</valid>
       <value type="real">0</value>
      </data-summary-statistic>
      <data-summary-statistic name="min">
       <valid>true</valid>
       <value type="integer">1</value>
      </data-summary-statistic>
      <data-summary-statistic name="max">
       <valid>true</valid>
       <value type="integer">5</value>
      </data-summary-statistic>
      <data-summary-statistic name="mean">
       <valid>true</valid>
       <value type="real">2.24444444444</value>
      </data-summary-statistic>
      <data-summary-statistic name="stddev">
       <valid>true</valid>
       <value type="real">1.72841957605</value>
      </data-summary-statistic>
      <data-summary-statistic name="unique">
       <valid>true</valid>
       <value type="real">0</value>
      </data-summary-statistic>
      <data-summary-statistic name="maxReal">
       <valid>true</valid>
       <value type="real">5</value>
      </data-summary-statistic>
      <data-summary-statistic name="minReal">
       <valid>true</valid>
       <value type="real">1</value>
      </data-summary-statistic>
     </data-summary-statistics>
    </data-summary>
    <value name="ref-value" type="missing"/>
    <set-value>1</set-value>
    <res-value>0</res-value>
   </normalize>
   <weight-type>none</weight-type>
   <denormal-type>integer</denormal-type>
  </variable>
  <variable name="T&#228;glichAlkohol">
   <use>true</use>
   <base>-1</base>
   <index>18</index>
   <map name="" type="data"/>
   <normalize>
    <type>rank</type>
    <width>1</width>
    <missing-values>use</missing-values>
    <data-summary>
     <histogram>
      <hist-entry>
       <value type="integer">1</value>
       <weighted>77</weighted>
       <unweighted>77</unweighted>
      </hist-entry>
      <hist-entry>
       <value type="integer">2</value>
```

```
      <weighted>136</weighted>
      <unweighted>136</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">9</value>
      <weighted>147</weighted>
      <unweighted>147</unweighted>
     </hist-entry>
    </histogram>
    <data-summary-statistics>
     <data-summary-statistic name="cardinality">
      <valid>true</valid>
      <value type="integer">3</value>
     </data-summary-statistic>
     <data-summary-statistic name="missing">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
     <data-summary-statistic name="min">
      <valid>true</valid>
      <value type="integer">1</value>
     </data-summary-statistic>
     <data-summary-statistic name="max">
      <valid>true</valid>
      <value type="integer">9</value>
     </data-summary-statistic>
     <data-summary-statistic name="mean">
      <valid>true</valid>
      <value type="real">4.64444444444</value>
     </data-summary-statistic>
     <data-summary-statistic name="stddev">
      <valid>true</valid>
      <value type="real">3.64225009615</value>
     </data-summary-statistic>
     <data-summary-statistic name="unique">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
     <data-summary-statistic name="maxReal">
      <valid>true</valid>
      <value type="real">9</value>
     </data-summary-statistic>
     <data-summary-statistic name="minReal">
      <valid>true</valid>
      <value type="real">1</value>
     </data-summary-statistic>
    </data-summary-statistics>
   </data-summary>
   <value name="ref-value" type="missing"/>
   <set-value>1</set-value>
   <res-value>0</res-value>
  </normalize>
  <weight-type>none</weight-type>
  <denormal-type>integer</denormal-type>
 </variable>
 <variable name="Alter">
  <use>true</use>
  <base>-1</base>
  <index>19</index>
  <map name="" type="data"/>
```

```
<normalize>
 <type>rank</type>
 <width>1</width>
 <missing-values>use</missing-values>
 <data-summary>
  <histogram>
   <hist-entry>
    <value type="integer">1</value>
    <weighted>114</weighted>
    <unweighted>114</unweighted>
   </hist-entry>
   <hist-entry>
    <value type="integer">2</value>
    <weighted>154</weighted>
    <unweighted>154</unweighted>
   </hist-entry>
   <hist-entry>
    <value type="integer">3</value>
    <weighted>92</weighted>
    <unweighted>92</unweighted>
   </hist-entry>
  </histogram>
  <data-summary-statistics>
   <data-summary-statistic name="cardinality">
    <valid>true</valid>
    <value type="integer">3</value>
   </data-summary-statistic>
   <data-summary-statistic name="missing">
    <valid>true</valid>
    <value type="real">0</value>
   </data-summary-statistic>
   <data-summary-statistic name="min">
    <valid>true</valid>
    <value type="integer">1</value>
   </data-summary-statistic>
   <data-summary-statistic name="max">
    <valid>true</valid>
    <value type="integer">3</value>
   </data-summary-statistic>
   <data-summary-statistic name="mean">
    <valid>true</valid>
    <value type="real">1.93888888889</value>
   </data-summary-statistic>
   <data-summary-statistic name="stddev">
    <valid>true</valid>
    <value type="real">0.755030585732</value>
   </data-summary-statistic>
   <data-summary-statistic name="unique">
    <valid>true</valid>
    <value type="real">0</value>
   </data-summary-statistic>
   <data-summary-statistic name="maxReal">
    <valid>true</valid>
    <value type="real">3</value>
   </data-summary-statistic>
   <data-summary-statistic name="minReal">
    <valid>true</valid>
    <value type="real">1</value>
   </data-summary-statistic>
  </data-summary-statistics>
```

```
  </data-summary>
  <value name="ref-value" type="missing"/>
  <set-value>1</set-value>
  <res-value>0</res-value>
 </normalize>
 <weight-type>none</weight-type>
 <denormal-type>integer</denormal-type>
</variable>
<variable name="Ausbildungsdauer">
 <use>true</use>
 <base>-1</base>
 <index>20</index>
 <map name="" type="data"/>
 <normalize>
  <type>rank</type>
  <width>1</width>
  <missing-values>use</missing-values>
  <data-summary>
   <histogram>
    <hist-entry>
     <value type="real">1</value>
     <weighted>16</weighted>
     <unweighted>16</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="real">2</value>
     <weighted>259</weighted>
     <unweighted>259</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="real">3</value>
     <weighted>11</weighted>
     <unweighted>11</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="real">4</value>
     <weighted>24</weighted>
     <unweighted>24</unweighted>
    </hist-entry>
    <hist-entry>
     <value type="real">5</value>
     <weighted>50</weighted>
     <unweighted>50</unweighted>
    </hist-entry>
   </histogram>
   <data-summary-statistics>
    <data-summary-statistic name="cardinality">
     <valid>true</valid>
     <value type="integer">5</value>
    </data-summary-statistic>
    <data-summary-statistic name="missing">
     <valid>true</valid>
     <value type="real">0</value>
    </data-summary-statistic>
    <data-summary-statistic name="min">
     <valid>true</valid>
     <value type="real">1</value>
    </data-summary-statistic>
    <data-summary-statistic name="max">
     <valid>true</valid>
```

```
      <value type="real">5</value>
     </data-summary-statistic>
     <data-summary-statistic name="mean">
      <valid>true</valid>
      <value type="real">2.53611111111</value>
     </data-summary-statistic>
     <data-summary-statistic name="stddev">
      <valid>true</valid>
      <value type="real">1.14362780478</value>
     </data-summary-statistic>
     <data-summary-statistic name="unique">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
     <data-summary-statistic name="maxReal">
      <valid>true</valid>
      <value type="real">5</value>
     </data-summary-statistic>
     <data-summary-statistic name="minReal">
      <valid>true</valid>
      <value type="real">1</value>
     </data-summary-statistic>
    </data-summary-statistics>
   </data-summary>
   <value name="ref-value" type="missing"/>
   <set-value>1</set-value>
   <res-value>0</res-value>
  </normalize>
  <weight-type>none</weight-type>
  <denormal-type>real</denormal-type>
 </variable>
 <variable name="Einkommen">
  <use>true</use>
  <base>-1</base>
  <index>21</index>
  <map name="" type="data"/>
  <normalize>
   <type>rank</type>
   <width>1</width>
   <missing-values>use</missing-values>
   <data-summary>
    <histogram>
     <hist-entry>
      <value type="integer">0</value>
      <weighted>1</weighted>
      <unweighted>1</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">3</value>
      <weighted>1</weighted>
      <unweighted>1</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">4</value>
      <weighted>10</weighted>
      <unweighted>10</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">5</value>
      <weighted>11</weighted>
```

```
 <unweighted>11</unweighted>
</hist-entry>
<hist-entry>
 <value type="integer">6</value>
 <weighted>18</weighted>
 <unweighted>18</unweighted>
</hist-entry>
<hist-entry>
 <value type="integer">7</value>
 <weighted>41</weighted>
 <unweighted>41</unweighted>
</hist-entry>
<hist-entry>
 <value type="integer">8</value>
 <weighted>19</weighted>
 <unweighted>19</unweighted>
</hist-entry>
<hist-entry>
 <value type="integer">9</value>
 <weighted>32</weighted>
 <unweighted>32</unweighted>
</hist-entry>
<hist-entry>
 <value type="integer">10</value>
 <weighted>28</weighted>
 <unweighted>28</unweighted>
</hist-entry>
<hist-entry>
 <value type="integer">11</value>
 <weighted>38</weighted>
 <unweighted>38</unweighted>
</hist-entry>
<hist-entry>
 <value type="integer">12</value>
 <weighted>50</weighted>
 <unweighted>50</unweighted>
</hist-entry>
<hist-entry>
 <value type="integer">13</value>
 <weighted>17</weighted>
 <unweighted>17</unweighted>
</hist-entry>
<hist-entry>
 <value type="integer">14</value>
 <weighted>10</weighted>
 <unweighted>10</unweighted>
</hist-entry>
<hist-entry>
 <value type="integer">15</value>
 <weighted>28</weighted>
 <unweighted>28</unweighted>
</hist-entry>
<hist-entry>
 <value type="integer">16</value>
 <weighted>22</weighted>
 <unweighted>22</unweighted>
</hist-entry>
<hist-entry>
 <value type="integer">17</value>
 <weighted>8</weighted>
```

```
  <unweighted>8</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">18</value>
  <weighted>7</weighted>
  <unweighted>7</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">19</value>
  <weighted>12</weighted>
  <unweighted>12</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">20</value>
  <weighted>4</weighted>
  <unweighted>4</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">21</value>
  <weighted>1</weighted>
  <unweighted>1</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">23</value>
  <weighted>1</weighted>
  <unweighted>1</unweighted>
 </hist-entry>
 <hist-entry>
  <value type="integer">98</value>
  <weighted>1</weighted>
  <unweighted>1</unweighted>
 </hist-entry>
</histogram>
<data-summary-statistics>
 <data-summary-statistic name="cardinality">
  <valid>true</valid>
  <value type="integer">22</value>
 </data-summary-statistic>
 <data-summary-statistic name="missing">
  <valid>true</valid>
  <value type="real">0</value>
 </data-summary-statistic>
 <data-summary-statistic name="min">
  <valid>true</valid>
  <value type="integer">0</value>
 </data-summary-statistic>
 <data-summary-statistic name="max">
  <valid>true</valid>
  <value type="integer">98</value>
 </data-summary-statistic>
 <data-summary-statistic name="mean">
  <valid>true</valid>
  <value type="real">11.3222222222</value>
 </data-summary-statistic>
 <data-summary-statistic name="stddev">
  <valid>true</valid>
  <value type="real">6.06570849617</value>
 </data-summary-statistic>
 <data-summary-statistic name="unique">
  <valid>true</valid>
```

```
      <value type="real">5</value>
     </data-summary-statistic>
     <data-summary-statistic name="maxReal">
      <valid>true</valid>
      <value type="real">98</value>
     </data-summary-statistic>
     <data-summary-statistic name="minReal">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
    </data-summary-statistics>
   </data-summary>
   <value name="ref-value" type="missing"/>
   <set-value>1</set-value>
   <res-value>0</res-value>
  </normalize>
  <weight-type>none</weight-type>
  <denormal-type>integer</denormal-type>
 </variable>
 <variable name="Geschlecht">
  <use>true</use>
  <base>-1</base>
  <index>22</index>
  <map name="" type="data"/>
  <normalize>
   <type>rank</type>
   <width>1</width>
   <missing-values>use</missing-values>
   <data-summary>
    <histogram>
     <hist-entry>
      <value type="integer">1</value>
      <weighted>198</weighted>
      <unweighted>198</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">2</value>
      <weighted>162</weighted>
      <unweighted>162</unweighted>
     </hist-entry>
    </histogram>
    <data-summary-statistics>
     <data-summary-statistic name="cardinality">
      <valid>true</valid>
      <value type="integer">2</value>
     </data-summary-statistic>
     <data-summary-statistic name="missing">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
     <data-summary-statistic name="min">
      <valid>true</valid>
      <value type="integer">1</value>
     </data-summary-statistic>
     <data-summary-statistic name="max">
      <valid>true</valid>
      <value type="integer">2</value>
     </data-summary-statistic>
     <data-summary-statistic name="mean">
      <valid>true</valid>
```

```
      <value type="real">1.45</value>
     </data-summary-statistic>
     <data-summary-statistic name="stddev">
      <valid>true</valid>
      <value type="real">0.498186124899</value>
     </data-summary-statistic>
     <data-summary-statistic name="unique">
      <valid>true</valid>
      <value type="real">0</value>
     </data-summary-statistic>
     <data-summary-statistic name="maxReal">
      <valid>true</valid>
      <value type="real">2</value>
     </data-summary-statistic>
     <data-summary-statistic name="minReal">
      <valid>true</valid>
      <value type="real">1</value>
     </data-summary-statistic>
    </data-summary-statistics>
   </data-summary>
   <value name="ref-value" type="missing"/>
   <set-value>1</set-value>
   <res-value>0</res-value>
  </normalize>
  <weight-type>none</weight-type>
  <denormal-type>integer</denormal-type>
 </variable>
</variables>
<variables class="dependent">
 <variable name="Blutdruck">
  <use>true</use>
  <base>-1</base>
  <index>26</index>
  <map name="" type="data"/>
  <normalize>
   <type>1-of-N</type>
   <width>3</width>
   <missing-values>use</missing-values>
   <data-summary>
    <histogram>
     <hist-entry>
      <value type="integer">1</value>
      <weighted>66</weighted>
      <unweighted>66</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">2</value>
      <weighted>217</weighted>
      <unweighted>217</unweighted>
     </hist-entry>
     <hist-entry>
      <value type="integer">3</value>
      <weighted>77</weighted>
      <unweighted>77</unweighted>
     </hist-entry>
    </histogram>
    <data-summary-statistics>
     <data-summary-statistic name="cardinality">
      <valid>true</valid>
      <value type="integer">3</value>
```

```
      </data-summary-statistic>
      <data-summary-statistic name="missing">
       <valid>true</valid>
       <value type="real">0</value>
      </data-summary-statistic>
      <data-summary-statistic name="min">
       <valid>true</valid>
       <value type="integer">1</value>
      </data-summary-statistic>
      <data-summary-statistic name="max">
       <valid>true</valid>
       <value type="integer">3</value>
      </data-summary-statistic>
      <data-summary-statistic name="mean">
       <valid>true</valid>
       <value type="real">2.03055555556</value>
      </data-summary-statistic>
      <data-summary-statistic name="stddev">
       <valid>true</valid>
       <value type="real">0.630390710717</value>
      </data-summary-statistic>
      <data-summary-statistic name="unique">
       <valid>true</valid>
       <value type="real">0</value>
      </data-summary-statistic>
      <data-summary-statistic name="maxReal">
       <valid>true</valid>
       <value type="real">3</value>
      </data-summary-statistic>
      <data-summary-statistic name="minReal">
       <valid>true</valid>
       <value type="real">1</value>
      </data-summary-statistic>
     </data-summary-statistics>
    </data-summary>
    <value name="ref-value" type="missing"/>
    <set-value>1</set-value>
    <res-value>0</res-value>
   </normalize>
   <weight-type>none</weight-type>
   <denormal-type>integer</denormal-type>
  </variable>
 </variables>
 <variables class="predicted">
  <variable name="Blutdruck">
   <use>true</use>
   <base>-1</base>
   <index>26</index>
   <map name="" type="data"/>
   <normalize>
    <type>1-of-N</type>
    <width>3</width>
    <missing-values>use</missing-values>
    <data-summary>
     <histogram>
      <hist-entry>
       <value type="integer">1</value>
       <weighted>66</weighted>
       <unweighted>66</unweighted>
      </hist-entry>
```

```
      <hist-entry>
       <value type="integer">2</value>
       <weighted>217</weighted>
       <unweighted>217</unweighted>
      </hist-entry>
      <hist-entry>
       <value type="integer">3</value>
       <weighted>77</weighted>
       <unweighted>77</unweighted>
      </hist-entry>
     </histogram>
     <data-summary-statistics>
      <data-summary-statistic name="cardinality">
       <valid>true</valid>
       <value type="integer">3</value>
      </data-summary-statistic>
      <data-summary-statistic name="missing">
       <valid>true</valid>
       <value type="real">0</value>
      </data-summary-statistic>
      <data-summary-statistic name="min">
       <valid>true</valid>
       <value type="integer">1</value>
      </data-summary-statistic>
      <data-summary-statistic name="max">
       <valid>true</valid>
       <value type="integer">3</value>
      </data-summary-statistic>
      <data-summary-statistic name="mean">
       <valid>true</valid>
       <value type="real">2.03055555556</value>
      </data-summary-statistic>
      <data-summary-statistic name="stddev">
       <valid>true</valid>
       <value type="real">0.630390710717</value>
      </data-summary-statistic>
      <data-summary-statistic name="unique">
       <valid>true</valid>
       <value type="real">0</value>
      </data-summary-statistic>
      <data-summary-statistic name="maxReal">
       <valid>true</valid>
       <value type="real">3</value>
      </data-summary-statistic>
      <data-summary-statistic name="minReal">
       <valid>true</valid>
       <value type="real">1</value>
      </data-summary-statistic>
     </data-summary-statistics>
    </data-summary>
    <value name="ref-value" type="missing"/>
    <set-value>1</set-value>
    <res-value>0</res-value>
   </normalize>
   <weight-type>none</weight-type>
   <denormal-type>integer</denormal-type>
  </variable>
 </variables>
 <attribute-info>
  <cluster-metric>euclidean</cluster-metric>
```

```
<metric-power>0.5</metric-power>
<dv-complexity>250</dv-complexity>
<map-collection>
 <map name="Gewicht(kg)" type="data">
  <map-entry>
   <value type="real"><closed>89</closed><open>102</open></value>
   <value type="real"><closed>89</closed><open>102</open></value>
  </map-entry>
  <map-entry>
   <value type="real"><closed>102</closed><open>110</open></value>
   <value type="real"><closed>102</closed><open>110</open></value>
  </map-entry>
  <map-entry>
   <value type="real"><closed>110</closed><open>114</open></value>
   <value type="real"><closed>110</closed><open>114</open></value>
  </map-entry>
  <map-entry>
   <value type="real"><closed>114</closed><open>119</open></value>
   <value type="real"><closed>114</closed><open>119</open></value>
  </map-entry>
  <map-entry>
   <value type="real"><closed>119</closed><open>122</open></value>
   <value type="real"><closed>119</closed><open>122</open></value>
  </map-entry>
  <map-entry>
   <value type="real"><closed>122</closed><open>125</open></value>
   <value type="real"><closed>122</closed><open>125</open></value>
  </map-entry>
  <map-entry>
   <value type="real"><closed>125</closed><open>130</open></value>
   <value type="real"><closed>125</closed><open>130</open></value>
  </map-entry>
  <map-entry>
   <value type="real"><closed>130</closed><open>133</open></value>
   <value type="real"><closed>130</closed><open>133</open></value>
  </map-entry>
  <map-entry>
   <value type="real"><closed>133</closed><open>136</open></value>
   <value type="real"><closed>133</closed><open>136</open></value>
  </map-entry>
  <map-entry>
   <value type="real"><closed>136</closed><open>140</open></value>
   <value type="real"><closed>136</closed><open>140</open></value>
  </map-entry>
  <map-entry>
   <value type="real"><closed>140</closed><open>141</open></value>
   <value type="real"><closed>140</closed><open>141</open></value>
  </map-entry>
  <map-entry>
   <value type="real"><closed>141</closed><open>142</open></value>
   <value type="real"><closed>141</closed><open>142</open></value>
  </map-entry>
  <map-entry>
   <value type="real"><closed>142</closed><open>144</open></value>
   <value type="real"><closed>142</closed><open>144</open></value>
  </map-entry>
  <map-entry>
   <value type="real"><closed>144</closed><open>147</open></value>
   <value type="real"><closed>144</closed><open>147</open></value>
  </map-entry>
```

```
<map-entry>
 <value type="real"><closed>147</closed><open>149</open></value>
 <value type="real"><closed>147</closed><open>149</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>149</closed><open>150</open></value>
 <value type="real"><closed>149</closed><open>150</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>150</closed><open>152</open></value>
 <value type="real"><closed>150</closed><open>152</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>152</closed><open>153</open></value>
 <value type="real"><closed>152</closed><open>153</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>153</closed><open>155</open></value>
 <value type="real"><closed>153</closed><open>155</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>155</closed><open>156</open></value>
 <value type="real"><closed>155</closed><open>156</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>156</closed><open>157</open></value>
 <value type="real"><closed>156</closed><open>157</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>157</closed><open>160</open></value>
 <value type="real"><closed>157</closed><open>160</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>160</closed><open>161</open></value>
 <value type="real"><closed>160</closed><open>161</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>161</closed><open>162</open></value>
 <value type="real"><closed>161</closed><open>162</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>162</closed><open>163</open></value>
 <value type="real"><closed>162</closed><open>163</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>163</closed><open>164</open></value>
 <value type="real"><closed>163</closed><open>164</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>164</closed><open>167</open></value>
 <value type="real"><closed>164</closed><open>167</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>167</closed><open>169</open></value>
 <value type="real"><closed>167</closed><open>169</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>169</closed><open>171</open></value>
 <value type="real"><closed>169</closed><open>171</open></value>
</map-entry>
```

```
<map-entry>
 <value type="real"><closed>171</closed><open>172</open></value>
 <value type="real"><closed>171</closed><open>172</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>172</closed><open>174</open></value>
 <value type="real"><closed>172</closed><open>174</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>174</closed><open>176</open></value>
 <value type="real"><closed>174</closed><open>176</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>176</closed><open>177</open></value>
 <value type="real"><closed>176</closed><open>177</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>177</closed><open>180</open></value>
 <value type="real"><closed>177</closed><open>180</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>180</closed><open>182</open></value>
 <value type="real"><closed>180</closed><open>182</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>182</closed><open>185</open></value>
 <value type="real"><closed>182</closed><open>185</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>185</closed><open>186</open></value>
 <value type="real"><closed>185</closed><open>186</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>186</closed><open>187</open></value>
 <value type="real"><closed>186</closed><open>187</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>187</closed><open>188</open></value>
 <value type="real"><closed>187</closed><open>188</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>188</closed><open>191</open></value>
 <value type="real"><closed>188</closed><open>191</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>191</closed><open>195</open></value>
 <value type="real"><closed>191</closed><open>195</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>195</closed><open>201</open></value>
 <value type="real"><closed>195</closed><open>201</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>201</closed><open>204</open></value>
 <value type="real"><closed>201</closed><open>204</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>204</closed><open>208</open></value>
 <value type="real"><closed>204</closed><open>208</open></value>
</map-entry>
```

```
 <map-entry>
  <value type="real"><closed>208</closed><open>211</open></value>
  <value type="real"><closed>208</closed><open>211</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>211</closed><open>215</open></value>
  <value type="real"><closed>211</closed><open>215</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>215</closed><open>220</open></value>
  <value type="real"><closed>215</closed><open>220</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>220</closed><open>232</open></value>
  <value type="real"><closed>220</closed><open>232</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>232</closed><open>245</open></value>
  <value type="real"><closed>232</closed><open>245</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>245</closed><closed>284</closed></value>
  <value type="real"><closed>245</closed><closed>284</closed></value>
 </map-entry>
</map>
<map name="Gr&#246;&#223;e(cm)" type="data">
 <map-entry>
  <value type="real"><closed>565</closed><open>587</open></value>
  <value type="real"><closed>565</closed><open>587</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>587</closed><open>597</open></value>
  <value type="real"><closed>587</closed><open>597</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>597</closed><open>601</open></value>
  <value type="real"><closed>597</closed><open>601</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>601</closed><open>602</open></value>
  <value type="real"><closed>601</closed><open>602</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>602</closed><open>604</open></value>
  <value type="real"><closed>602</closed><open>604</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>604</closed><open>605</open></value>
  <value type="real"><closed>604</closed><open>605</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>605</closed><open>607</open></value>
  <value type="real"><closed>605</closed><open>607</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>607</closed><open>608</open></value>
  <value type="real"><closed>607</closed><open>608</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>608</closed><open>610</open></value>
```

```
 <value type="real"><closed>608</closed><open>610</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>610</closed><open>611</open></value>
 <value type="real"><closed>610</closed><open>611</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>611</closed><open>613</open></value>
 <value type="real"><closed>611</closed><open>613</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>613</closed><open>617</open></value>
 <value type="real"><closed>613</closed><open>617</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>617</closed><open>620</open></value>
 <value type="real"><closed>617</closed><open>620</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>620</closed><open>622</open></value>
 <value type="real"><closed>620</closed><open>622</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>622</closed><open>624</open></value>
 <value type="real"><closed>622</closed><open>624</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>624</closed><open>626</open></value>
 <value type="real"><closed>624</closed><open>626</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>626</closed><open>629</open></value>
 <value type="real"><closed>626</closed><open>629</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>629</closed><open>631</open></value>
 <value type="real"><closed>629</closed><open>631</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>631</closed><open>633</open></value>
 <value type="real"><closed>631</closed><open>633</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>633</closed><open>635</open></value>
 <value type="real"><closed>633</closed><open>635</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>635</closed><open>637</open></value>
 <value type="real"><closed>635</closed><open>637</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>637</closed><open>640</open></value>
 <value type="real"><closed>637</closed><open>640</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>640</closed><open>643</open></value>
 <value type="real"><closed>640</closed><open>643</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>643</closed><open>646</open></value>
```

```
 <value type="real"><closed>643</closed><open>646</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>646</closed><open>649</open></value>
 <value type="real"><closed>646</closed><open>649</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>649</closed><open>650</open></value>
 <value type="real"><closed>649</closed><open>650</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>650</closed><open>655</open></value>
 <value type="real"><closed>650</closed><open>655</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>655</closed><open>659</open></value>
 <value type="real"><closed>655</closed><open>659</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>659</closed><open>663</open></value>
 <value type="real"><closed>659</closed><open>663</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>663</closed><open>665</open></value>
 <value type="real"><closed>663</closed><open>665</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>665</closed><open>667</open></value>
 <value type="real"><closed>665</closed><open>667</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>667</closed><open>669</open></value>
 <value type="real"><closed>667</closed><open>669</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>669</closed><open>671</open></value>
 <value type="real"><closed>669</closed><open>671</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>671</closed><open>674</open></value>
 <value type="real"><closed>671</closed><open>674</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>674</closed><open>675</open></value>
 <value type="real"><closed>674</closed><open>675</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>675</closed><open>677</open></value>
 <value type="real"><closed>675</closed><open>677</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>677</closed><open>678</open></value>
 <value type="real"><closed>677</closed><open>678</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>678</closed><open>680</open></value>
 <value type="real"><closed>678</closed><open>680</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>680</closed><open>681</open></value>
```

```
  <value type="real"><closed>680</closed><open>681</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>681</closed><open>683</open></value>
  <value type="real"><closed>681</closed><open>683</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>683</closed><open>687</open></value>
  <value type="real"><closed>683</closed><open>687</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>687</closed><open>689</open></value>
  <value type="real"><closed>687</closed><open>689</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>689</closed><open>692</open></value>
  <value type="real"><closed>689</closed><open>692</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>692</closed><open>696</open></value>
  <value type="real"><closed>692</closed><open>696</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>696</closed><open>698</open></value>
  <value type="real"><closed>696</closed><open>698</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>698</closed><open>703</open></value>
  <value type="real"><closed>698</closed><open>703</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>703</closed><open>706</open></value>
  <value type="real"><closed>703</closed><open>706</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>706</closed><open>709</open></value>
  <value type="real"><closed>706</closed><open>709</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>709</closed><open>723</open></value>
  <value type="real"><closed>709</closed><open>723</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>723</closed><closed>746</closed></value>
  <value type="real"><closed>723</closed><closed>746</closed></value>
 </map-entry>
</map>
<map name="zWerte" type="data">
 <map-entry>
  <value type="real"><closed>0</closed><open>1.84</open></value>
  <value type="real"><closed>0</closed><open>1.84</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>1.84</closed><open>2.32</open></value>
  <value type="real"><closed>1.84</closed><open>2.32</open></value>
 </map-entry>
 <map-entry>
  <value type="real"><closed>2.32</closed><open>2.52</open></value>
  <value type="real"><closed>2.32</closed><open>2.52</open></value>
 </map-entry>
```

```
<map-entry>
 <value type="real"><closed>2.52</closed><open>2.76</open></value>
 <value type="real"><closed>2.52</closed><open>2.76</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>2.76</closed><open>2.88</open></value>
 <value type="real"><closed>2.76</closed><open>2.88</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>2.88</closed><open>2.91</open></value>
 <value type="real"><closed>2.88</closed><open>2.91</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>2.91</closed><open>3.23</open></value>
 <value type="real"><closed>2.91</closed><open>3.23</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.23</closed><open>3.26</open></value>
 <value type="real"><closed>3.23</closed><open>3.26</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.26</closed><open>3.4</open></value>
 <value type="real"><closed>3.26</closed><open>3.4</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.4</closed><open>3.48</open></value>
 <value type="real"><closed>3.4</closed><open>3.48</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.48</closed><open>3.5</open></value>
 <value type="real"><closed>3.48</closed><open>3.5</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.5</closed><open>3.51</open></value>
 <value type="real"><closed>3.5</closed><open>3.51</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.51</closed><open>3.54</open></value>
 <value type="real"><closed>3.51</closed><open>3.54</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.54</closed><open>3.62</open></value>
 <value type="real"><closed>3.54</closed><open>3.62</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.62</closed><open>3.66</open></value>
 <value type="real"><closed>3.62</closed><open>3.66</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.66</closed><open>3.75</open></value>
 <value type="real"><closed>3.66</closed><open>3.75</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.75</closed><open>3.81</open></value>
 <value type="real"><closed>3.75</closed><open>3.81</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.81</closed><open>3.83</open></value>
 <value type="real"><closed>3.81</closed><open>3.83</open></value>
</map-entry>
```

```
<map-entry>
 <value type="real"><closed>3.83</closed><open>3.87</open></value>
 <value type="real"><closed>3.83</closed><open>3.87</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.87</closed><open>3.93</open></value>
 <value type="real"><closed>3.87</closed><open>3.93</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.93</closed><open>3.96</open></value>
 <value type="real"><closed>3.93</closed><open>3.96</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>3.96</closed><open>4.04</open></value>
 <value type="real"><closed>3.96</closed><open>4.04</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.04</closed><open>4.12</open></value>
 <value type="real"><closed>4.04</closed><open>4.12</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.12</closed><open>4.16</open></value>
 <value type="real"><closed>4.12</closed><open>4.16</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.16</closed><open>4.18</open></value>
 <value type="real"><closed>4.16</closed><open>4.18</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.18</closed><open>4.23</open></value>
 <value type="real"><closed>4.18</closed><open>4.23</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.23</closed><open>4.28</open></value>
 <value type="real"><closed>4.23</closed><open>4.28</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.28</closed><open>4.31</open></value>
 <value type="real"><closed>4.28</closed><open>4.31</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.31</closed><open>4.35</open></value>
 <value type="real"><closed>4.31</closed><open>4.35</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.35</closed><open>4.38</open></value>
 <value type="real"><closed>4.35</closed><open>4.38</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.38</closed><open>4.41</open></value>
 <value type="real"><closed>4.38</closed><open>4.41</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.41</closed><open>4.43</open></value>
 <value type="real"><closed>4.41</closed><open>4.43</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.43</closed><open>4.46</open></value>
 <value type="real"><closed>4.43</closed><open>4.46</open></value>
</map-entry>
```

```
<map-entry>
 <value type="real"><closed>4.46</closed><open>4.49</open></value>
 <value type="real"><closed>4.46</closed><open>4.49</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.49</closed><open>4.53</open></value>
 <value type="real"><closed>4.49</closed><open>4.53</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.53</closed><open>4.6</open></value>
 <value type="real"><closed>4.53</closed><open>4.6</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.6</closed><open>4.67</open></value>
 <value type="real"><closed>4.6</closed><open>4.67</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.67</closed><open>4.7</open></value>
 <value type="real"><closed>4.67</closed><open>4.7</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.7</closed><open>4.76</open></value>
 <value type="real"><closed>4.7</closed><open>4.76</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.76</closed><open>4.84</open></value>
 <value type="real"><closed>4.76</closed><open>4.84</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.84</closed><open>4.85</open></value>
 <value type="real"><closed>4.84</closed><open>4.85</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.85</closed><open>4.89</open></value>
 <value type="real"><closed>4.85</closed><open>4.89</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.89</closed><open>4.91</open></value>
 <value type="real"><closed>4.89</closed><open>4.91</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.91</closed><open>4.97</open></value>
 <value type="real"><closed>4.91</closed><open>4.97</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>4.97</closed><open>5.03</open></value>
 <value type="real"><closed>4.97</closed><open>5.03</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>5.03</closed><open>5.13</open></value>
 <value type="real"><closed>5.03</closed><open>5.13</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>5.13</closed><open>5.23</open></value>
 <value type="real"><closed>5.13</closed><open>5.23</open></value>
</map-entry>
<map-entry>
 <value type="real"><closed>5.23</closed><open>5.34</open></value>
 <value type="real"><closed>5.23</closed><open>5.34</open></value>
</map-entry>
```

```
   <map-entry>
    <value type="real"><closed>5.34</closed><open>5.5</open></value>
    <value type="real"><closed>5.34</closed><open>5.5</open></value>
   </map-entry>
   <map-entry>
    <value type="real"><closed>5.5</closed><closed>5.72</closed></value>
    <value type="real"><closed>5.5</closed><closed>5.72</closed></value>
   </map-entry>
  </map>
  <map name="27" type="color">
   <map-entry>
    <value type="integer">1</value>
    <value type="integer">255</value>
   </map-entry>
   <map-entry>
    <value type="integer">2</value>
    <value type="integer">65280</value>
   </map-entry>
   <map-entry>
    <value type="integer">3</value>
    <value type="integer">8421504</value>
   </map-entry>
  </map>
 </map-collection>
 <attribute name="MilchTyp" index="0">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="TiefgefrorenLW" index="1">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
```

```
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="LindLetzteWoche" index="2">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="FleischLetzteW" index="3">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
```

```
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="Gefl&#252;gelLetzteW" index="4">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="FischLetzteWoche" index="5">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
```

```
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="LammLetzteWoche" index="6">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="SonstigFleischLW" index="7">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
```

```
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="K&#228;seLetzteWoche" index="8">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="EierLetzteWoche" index="9">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
```

```
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="Fleisch2MalT&#228;glLW" index="10">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="SalzImEssen" index="11">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
```

```
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="Salzverbrauch" index="12">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="ButterHaltiges" index="13">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
```

```
   <signifigance>0.05</signifigance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="SportAktivit&#228;t" index="14">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="Schlafdauer" index="15">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
```

```
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="Rauchen" index="16">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="TrinkGewohnheiten" index="17">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
```

```
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="T&#228;glichAlkohol" index="18">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="Alter" index="19">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
```

```
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="Ausbildungsdauer" index="20">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="Einkommen" index="21">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
```

```
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="Geschlecht" index="22">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>all</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <significance>0.05</significance>
   <display>present</display>
   <missing-values>use</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="Gewicht(kg)" index="23">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>no</include>
   <grouping>continuous</grouping>
   <intervals type="static">10</intervals>
```

```
   <signifigance>0.05</signifigance>
   <display>range</display>
   <missing-values>ignore</missing-values>
   <break-apart>yes</break-apart>
   <maps>
    <mapref type="discrete">Gewicht(kg)</mapref>
    <mapref type="continuous">Gewicht(kg)</mapref>
   </maps>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>continuous</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>all</display>
   <missing-values>ignore</missing-values>
   <break-apart>yes</break-apart>
   <maps>
    <mapref type="discrete">Gewicht(kg)</mapref>
    <mapref type="continuous">Gewicht(kg)</mapref>
   </maps>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>continuous</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>range</display>
   <missing-values>ignore</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="Gr&#246;&#223;e(cm)" index="24">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>no</include>
   <grouping>continuous</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>range</display>
   <missing-values>ignore</missing-values>
   <break-apart>yes</break-apart>
   <maps>
    <mapref type="discrete">Gr&#246;&#223;e(cm)</mapref>
    <mapref type="continuous">Gr&#246;&#223;e(cm)</mapref>
   </maps>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>continuous</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>all</display>
   <missing-values>ignore</missing-values>
   <break-apart>yes</break-apart>
   <maps>
    <mapref type="discrete">Gr&#246;&#223;e(cm)</mapref>
    <mapref type="continuous">Gr&#246;&#223;e(cm)</mapref>
   </maps>
  </attribute-settings>
```

```
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>continuous</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>range</display>
   <missing-values>ignore</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="zWerte" index="25">
  <state>independent</state>
  <attribute-settings state="independent">
   <include>no</include>
   <grouping>continuous</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>range</display>
   <missing-values>ignore</missing-values>
   <break-apart>yes</break-apart>
   <maps>
    <mapref type="discrete">zWerte</mapref>
    <mapref type="continuous">zWerte</mapref>
   </maps>
  </attribute-settings>
  <attribute-settings state="dependent">
   <include>yes</include>
   <grouping>continuous</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>all</display>
   <missing-values>ignore</missing-values>
   <break-apart>yes</break-apart>
   <maps>
    <mapref type="discrete">zWerte</mapref>
    <mapref type="continuous">zWerte</mapref>
   </maps>
  </attribute-settings>
  <attribute-settings state="weight">
   <include>yes</include>
   <grouping>continuous</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>range</display>
   <missing-values>ignore</missing-values>
   <break-apart>yes</break-apart>
  </attribute-settings>
 </attribute>
 <attribute name="Blutdruck" index="26">
  <state>dependent</state>
  <maps>
   <mapref type="color">27</mapref>
  </maps>
  <attribute-settings state="independent">
   <include>yes</include>
   <grouping>ordered</grouping>
   <intervals type="static">10</intervals>
   <signifigance>0.05</signifigance>
   <display>present</display>
   <missing-values>use</missing-values>
```

```
    <break-apart>yes</break-apart>
   </attribute-settings>
   <attribute-settings state="dependent">
    <include>yes</include>
    <grouping>ordered</grouping>
    <intervals type="static">10</intervals>
    <signifigance>0.05</signifigance>
    <display>all</display>
    <missing-values>use</missing-values>
    <break-apart>yes</break-apart>
   </attribute-settings>
   <attribute-settings state="weight">
    <include>yes</include>
    <grouping>ordered</grouping>
    <intervals type="static">10</intervals>
    <signifigance>0.05</signifigance>
    <display>present</display>
    <missing-values>use</missing-values>
    <break-apart>yes</break-apart>
   </attribute-settings>
  </attribute>
 </attribute-info>
 <tree-model>
  <node-count>50</node-count>
  <total-recs>360</total-recs>
  <vector type="real" dim="3" name="training-dist">
   66 217 77
  </vector>
  <accuracy>0.752777777778</accuracy>
  <error>0.247222222222</error>
  <value-map>
   <value name="0" type="integer">1</value>
   <value name="1" type="integer">2</value>
   <value name="2" type="integer">3</value>
  </value-map>
  <node iv="-1">
   <node-dist>
    <total>360</total>
    <continuous-dist>
     <unweighted>0</unweighted>
     <weighted>0</weighted>
     <mean>0</mean>
     <stddev>0</stddev>
    </continuous-dist>
    <discrete-dist>
     <value name="0.183333333333" type="integer">1</value>
     <value name="0.602777777778" type="integer">2</value>
     <value name="0.213888888889" type="integer">3</value>
    </discrete-dist>
   </node-dist>
   <vector type="real" dim="3" name="training-dist">
    66 217 77
   </vector>
   <branch-values/>
   <split IV="19">
    <split-map branches="3">
     <value name="0" type="integer">1</value>
     <value name="1" type="integer">2</value>
     <value name="2" type="integer">3</value>
    </split-map>
```

```
<node iv="19">
 <node-dist>
  <total>114</total>
  <continuous-dist>
   <unweighted>0</unweighted>
   <weighted>0</weighted>
   <mean>0</mean>
   <stddev>0</stddev>
  </continuous-dist>
  <discrete-dist>
   <value name="0.350877192982" type="integer">1</value>
   <value name="0.561403508772" type="integer">2</value>
   <value name="0.0877192982456" type="integer">3</value>
  </discrete-dist>
 </node-dist>
 <vector type="real" dim="3" name="training-dist">
  40 64 10
 </vector>
 <branch-values>
  <value type="integer">1</value>
 </branch-values>
 <split IV="15">
  <split-map branches="3">
   <value name="0" type="real">3</value>
   <value name="0" type="real">7</value>
   <value name="0" type="real">8</value>
   <value name="0" type="real">9</value>
   <value name="0" type="real">10</value>
   <value name="1" type="real">12</value>
   <value name="2" type="real">13</value>
   <value name="2" type="real">14</value>
   <value name="2" type="real">15</value>
  </split-map>
  <node iv="15">
   <node-dist>
    <total>26</total>
    <continuous-dist>
     <unweighted>0</unweighted>
     <weighted>0</weighted>
     <mean>0</mean>
     <stddev>0</stddev>
    </continuous-dist>
    <discrete-dist>
     <value name="0.269230769231" type="integer">1</value>
     <value name="0.5" type="integer">2</value>
     <value name="0.230769230769" type="integer">3</value>
    </discrete-dist>
   </node-dist>
   <vector type="real" dim="3" name="training-dist">
    7 13 6
   </vector>
   <branch-values>
    <value type="real">3</value>
    <value type="real">7</value>
    <value type="real">8</value>
    <value type="real">9</value>
    <value type="real">10</value>
   </branch-values>
   <split IV="16">
    <split-map branches="2">
```

```
    <value name="0" type="integer">1</value>
    <value name="0" type="integer">3</value>
    <value name="1" type="integer">4</value>
   </split-map>
   <node iv="16">
    <node-dist>
     <total>21</total>
     <continuous-dist>
      <unweighted>0</unweighted>
      <weighted>0</weighted>
      <mean>0</mean>
      <stddev>0</stddev>
     </continuous-dist>
     <discrete-dist>
      <value name="0.333333333333" type="integer">1</value>
      <value name="0.571428571429" type="integer">2</value>
      <value name="0.0952380952381" type="integer">3</value>
     </discrete-dist>
    </node-dist>
    <vector type="real" dim="3" name="training-dist">
     7 12 2
    </vector>
    <branch-values>
     <value type="integer">1</value>
     <value type="integer">3</value>
    </branch-values>
   </node>
   <node iv="16">
    <node-dist>
     <total>5</total>
     <continuous-dist>
      <unweighted>0</unweighted>
      <weighted>0</weighted>
      <mean>0</mean>
      <stddev>0</stddev>
     </continuous-dist>
     <discrete-dist>
      <value name="0" type="integer">1</value>
      <value name="0.2" type="integer">2</value>
      <value name="0.8" type="integer">3</value>
     </discrete-dist>
    </node-dist>
    <vector type="real" dim="3" name="training-dist">
     0 1 4
    </vector>
    <branch-values>
     <value type="integer">4</value>
    </branch-values>
   </node>
  </split>
 </node>
 <node iv="15">
  <node-dist>
   <total>37</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
    <stddev>0</stddev>
   </continuous-dist>
```

```
 <discrete-dist>
  <value name="0.162162162162" type="integer">1</value>
  <value name="0.837837837838" type="integer">2</value>
  <value name="0" type="integer">3</value>
 </discrete-dist>
</node-dist>
<vector type="real" dim="3" name="training-dist">
 6 31 0
</vector>
<branch-values>
 <value type="real">12</value>
</branch-values>
<split IV="2">
 <split-map branches="2">
  <value name="0" type="integer">0</value>
  <value name="0" type="integer">1</value>
  <value name="1" type="integer">2</value>
  <value name="1" type="integer">3</value>
  <value name="1" type="integer">4</value>
  <value name="1" type="integer">5</value>
  <value name="1" type="integer">6</value>
  <value name="1" type="integer">7</value>
 </split-map>
 <node iv="2">
  <node-dist>
   <total>7</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
    <stddev>0</stddev>
   </continuous-dist>
   <discrete-dist>
    <value name="0.714285714286" type="integer">1</value>
    <value name="0.285714285714" type="integer">2</value>
    <value name="0" type="integer">3</value>
   </discrete-dist>
  </node-dist>
  <vector type="real" dim="3" name="training-dist">
   5 2 0
  </vector>
  <branch-values>
   <value type="integer">0</value>
   <value type="integer">1</value>
  </branch-values>
 </node>
 <node iv="2">
  <node-dist>
   <total>30</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
    <stddev>0</stddev>
   </continuous-dist>
   <discrete-dist>
    <value name="0.0333333333333" type="integer">1</value>
    <value name="0.966666666667" type="integer">2</value>
    <value name="0" type="integer">3</value>
   </discrete-dist>
```

```
</node-dist>
<vector type="real" dim="3" name="training-dist">
 1 29 0
</vector>
<branch-values>
 <value type="integer">2</value>
 <value type="integer">3</value>
 <value type="integer">4</value>
 <value type="integer">5</value>
 <value type="integer">6</value>
 <value type="integer">7</value>
</branch-values>
<split IV="13">
 <split-map branches="2">
  <value name="0" type="integer">1</value>
  <value name="1" type="integer">2</value>
 </split-map>
 <node iv="13">
  <node-dist>
   <total>24</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
    <stddev>0</stddev>
   </continuous-dist>
   <discrete-dist>
    <value name="0" type="integer">1</value>
    <value name="1" type="integer">2</value>
    <value name="0" type="integer">3</value>
   </discrete-dist>
  </node-dist>
  <vector type="real" dim="3" name="training-dist">
   0 24 0
  </vector>
  <branch-values>
   <value type="integer">1</value>
  </branch-values>
 </node>
 <node iv="13">
  <node-dist>
   <total>6</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
    <stddev>0</stddev>
   </continuous-dist>
   <discrete-dist>
    <value name="0.166666666667" type="integer">1</value>
    <value name="0.833333333333" type="integer">2</value>
    <value name="0" type="integer">3</value>
   </discrete-dist>
  </node-dist>
  <vector type="real" dim="3" name="training-dist">
   1 5 0
  </vector>
  <branch-values>
   <value type="integer">2</value>
  </branch-values>
```

```
    </node>
   </split>
  </node>
 </split>
</node>
<node iv="15">
 <node-dist>
  <total>51</total>
  <continuous-dist>
   <unweighted>0</unweighted>
   <weighted>0</weighted>
   <mean>0</mean>
   <stddev>0</stddev>
  </continuous-dist>
  <discrete-dist>
   <value name="0.529411764706" type="integer">1</value>
   <value name="0.392156862745" type="integer">2</value>
   <value name="0.078431372549" type="integer">3</value>
  </discrete-dist>
 </node-dist>
 <vector type="real" dim="3" name="training-dist">
  27 20 4
 </vector>
 <branch-values>
  <value type="real">13</value>
  <value type="real">14</value>
  <value type="real">15</value>
 </branch-values>
 <split IV="13">
  <split-map branches="2">
   <value name="0" type="integer">1</value>
   <value name="0" type="integer">2</value>
   <value name="1" type="integer">3</value>
  </split-map>
  <node iv="13">
   <node-dist>
    <total>44</total>
    <continuous-dist>
     <unweighted>0</unweighted>
     <weighted>0</weighted>
     <mean>0</mean>
     <stddev>0</stddev>
    </continuous-dist>
    <discrete-dist>
     <value name="0.568181818182" type="integer">1</value>
     <value name="0.409090909091" type="integer">2</value>
     <value name="0.0227272727273" type="integer">3</value>
    </discrete-dist>
   </node-dist>
   <vector type="real" dim="3" name="training-dist">
    25 18 1
   </vector>
   <branch-values>
    <value type="integer">1</value>
    <value type="integer">2</value>
   </branch-values>
   <split IV="18">
    <split-map branches="3">
     <value name="0" type="integer">1</value>
     <value name="1" type="integer">2</value>
```

```
 <value name="2" type="integer">9</value>
</split-map>
<node iv="18">
 <node-dist>
  <total>20</total>
  <continuous-dist>
   <unweighted>0</unweighted>
   <weighted>0</weighted>
   <mean>0</mean>
   <stddev>0</stddev>
  </continuous-dist>
  <discrete-dist>
   <value name="0.65" type="integer">1</value>
   <value name="0.35" type="integer">2</value>
   <value name="0" type="integer">3</value>
  </discrete-dist>
 </node-dist>
 <vector type="real" dim="3" name="training-dist">
  13 7 0
 </vector>
 <branch-values>
  <value type="integer">1</value>
 </branch-values>
</node>
<node iv="18">
 <node-dist>
  <total>14</total>
  <continuous-dist>
   <unweighted>0</unweighted>
   <weighted>0</weighted>
   <mean>0</mean>
   <stddev>0</stddev>
  </continuous-dist>
  <discrete-dist>
   <value name="0.214285714286" type="integer">1</value>
   <value name="0.714285714286" type="integer">2</value>
   <value name="0.0714285714286" type="integer">3</value>
  </discrete-dist>
 </node-dist>
 <vector type="real" dim="3" name="training-dist">
  3 10 1
 </vector>
 <branch-values>
  <value type="integer">2</value>
 </branch-values>
</node>
<node iv="18">
 <node-dist>
  <total>10</total>
  <continuous-dist>
   <unweighted>0</unweighted>
   <weighted>0</weighted>
   <mean>0</mean>
   <stddev>0</stddev>
  </continuous-dist>
  <discrete-dist>
   <value name="0.9" type="integer">1</value>
   <value name="0.1" type="integer">2</value>
   <value name="0" type="integer">3</value>
  </discrete-dist>
```

```
        </node-dist>
        <vector type="real" dim="3" name="training-dist">
         9 1 0
        </vector>
        <branch-values>
         <value type="integer">9</value>
        </branch-values>
       </node>
      </split>
     </node>
     <node iv="13">
      <node-dist>
       <total>7</total>
       <continuous-dist>
        <unweighted>0</unweighted>
        <weighted>0</weighted>
        <mean>0</mean>
        <stddev>0</stddev>
       </continuous-dist>
       <discrete-dist>
        <value name="0.285714285714" type="integer">1</value>
        <value name="0.285714285714" type="integer">2</value>
        <value name="0.428571428571" type="integer">3</value>
       </discrete-dist>
      </node-dist>
      <vector type="real" dim="3" name="training-dist">
       2 2 3
      </vector>
      <branch-values>
       <value type="integer">3</value>
      </branch-values>
     </node>
    </split>
   </node>
  </split>
 </node>
 <node iv="19">
  <node-dist>
   <total>154</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
    <stddev>0</stddev>
   </continuous-dist>
   <discrete-dist>
    <value name="0.136363636364" type="integer">1</value>
    <value name="0.720779220779" type="integer">2</value>
    <value name="0.142857142857" type="integer">3</value>
   </discrete-dist>
  </node-dist>
  <vector type="real" dim="3" name="training-dist">
   21 111 22
  </vector>
  <branch-values>
   <value type="integer">2</value>
  </branch-values>
  <split IV="9">
   <split-map branches="3">
    <value name="0" type="integer">0</value>
```

```
 <value name="0" type="integer">1</value>
 <value name="0" type="integer">2</value>
 <value name="1" type="integer">3</value>
 <value name="1" type="integer">4</value>
 <value name="2" type="integer">5</value>
 <value name="2" type="integer">6</value>
 <value name="2" type="integer">7</value>
 <value name="2" type="integer">9</value>
</split-map>
<node iv="9">
 <node-dist>
  <total>121</total>
  <continuous-dist>
   <unweighted>0</unweighted>
   <weighted>0</weighted>
   <mean>0</mean>
   <stddev>0</stddev>
  </continuous-dist>
  <discrete-dist>
   <value name="0.148760330579" type="integer">1</value>
   <value name="0.760330578512" type="integer">2</value>
   <value name="0.0909090909091" type="integer">3</value>
  </discrete-dist>
 </node-dist>
 <vector type="real" dim="3" name="training-dist">
  18 92 11
 </vector>
 <branch-values>
  <value type="integer">0</value>
  <value type="integer">1</value>
  <value type="integer">2</value>
 </branch-values>
 <split IV="5">
  <split-map branches="2">
   <value name="0" type="integer">0</value>
   <value name="0" type="integer">1</value>
   <value name="0" type="integer">2</value>
   <value name="1" type="integer">3</value>
   <value name="1" type="integer">5</value>
   <value name="1" type="integer">6</value>
  </split-map>
  <node iv="5">
   <node-dist>
    <total>112</total>
    <continuous-dist>
     <unweighted>0</unweighted>
     <weighted>0</weighted>
     <mean>0</mean>
     <stddev>0</stddev>
    </continuous-dist>
    <discrete-dist>
     <value name="0.142857142857" type="integer">1</value>
     <value name="0.794642857143" type="integer">2</value>
     <value name="0.0625" type="integer">3</value>
    </discrete-dist>
   </node-dist>
   <vector type="real" dim="3" name="training-dist">
    16 89 7
   </vector>
   <branch-values>
```

```
 <value type="integer">0</value>
 <value type="integer">1</value>
 <value type="integer">2</value>
</branch-values>
<split IV="0">
 <split-map branches="3">
  <value name="0" type="integer">1</value>
  <value name="0" type="integer">2</value>
  <value name="1" type="integer">3</value>
  <value name="2" type="integer">4</value>
  <value name="2" type="integer">5</value>
 </split-map>
 <node iv="0">
  <node-dist>
   <total>94</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
    <stddev>0</stddev>
   </continuous-dist>
   <discrete-dist>
    <value name="0.106382978723" type="integer">1</value>
    <value name="0.81914893617" type="integer">2</value>
    <value name="0.0744680851064" type="integer">3</value>
   </discrete-dist>
  </node-dist>
  <vector type="real" dim="3" name="training-dist">
   10 77 7
  </vector>
  <branch-values>
   <value type="integer">1</value>
   <value type="integer">2</value>
  </branch-values>
  <split IV="10">
   <split-map branches="2">
    <value name="0" type="integer">1</value>
    <value name="0" type="integer">2</value>
    <value name="1" type="integer">3</value>
    <value name="1" type="integer">4</value>
    <value name="1" type="integer">5</value>
    <value name="1" type="integer">6</value>
    <value name="1" type="integer">7</value>
    <value name="1" type="integer">9</value>
   </split-map>
   <node iv="10">
    <node-dist>
     <total>14</total>
     <continuous-dist>
      <unweighted>0</unweighted>
      <weighted>0</weighted>
      <mean>0</mean>
      <stddev>0</stddev>
     </continuous-dist>
     <discrete-dist>
      <value name="0.428571428571" type="integer">1</value>
      <value name="0.5" type="integer">2</value>
      <value name="0.0714285714286" type="integer">3</value>
     </discrete-dist>
    </node-dist>
```

```
 <vector type="real" dim="3" name="training-dist">
  6 7 1
 </vector>
 <branch-values>
  <value type="integer">1</value>
  <value type="integer">2</value>
 </branch-values>
</node>
<node iv="10">
 <node-dist>
  <total>80</total>
  <continuous-dist>
   <unweighted>0</unweighted>
   <weighted>0</weighted>
   <mean>0</mean>
   <stddev>0</stddev>
  </continuous-dist>
  <discrete-dist>
   <value name="0.05" type="integer">1</value>
   <value name="0.875" type="integer">2</value>
   <value name="0.075" type="integer">3</value>
  </discrete-dist>
 </node-dist>
 <vector type="real" dim="3" name="training-dist">
  4 70 6
 </vector>
 <branch-values>
  <value type="integer">3</value>
  <value type="integer">4</value>
  <value type="integer">5</value>
  <value type="integer">6</value>
  <value type="integer">7</value>
  <value type="integer">9</value>
 </branch-values>
 <split IV="3">
  <split-map branches="2">
   <value name="0" type="integer">0</value>
   <value name="0" type="integer">1</value>
   <value name="0" type="integer">2</value>
   <value name="1" type="integer">3</value>
   <value name="1" type="integer">4</value>
   <value name="1" type="integer">6</value>
  </split-map>
  <node iv="3">
   <node-dist>
    <total>72</total>
    <continuous-dist>
     <unweighted>0</unweighted>
     <weighted>0</weighted>
     <mean>0</mean>
     <stddev>0</stddev>
    </continuous-dist>
    <discrete-dist>
     <value name="0.0416666666667" type="integer">1</value>
     <value name="0.916666666667" type="integer">2</value>
     <value name="0.0416666666667" type="integer">3</value>
    </discrete-dist>
   </node-dist>
   <vector type="real" dim="3" name="training-dist">
    3 66 3
```

```
</vector>
<branch-values>
 <value type="integer">0</value>
 <value type="integer">1</value>
 <value type="integer">2</value>
</branch-values>
<split IV="2">
 <split-map branches="2">
  <value name="0" type="integer">0</value>
  <value name="0" type="integer">1</value>
  <value name="0" type="integer">2</value>
  <value name="0" type="integer">3</value>
  <value name="0" type="integer">4</value>
  <value name="1" type="integer">5</value>
  <value name="1" type="integer">6</value>
  <value name="1" type="integer">7</value>
 </split-map>
 <node iv="2">
  <node-dist>
   <total>65</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
    <stddev>0</stddev>
   </continuous-dist>
   <discrete-dist>
    <value name="0.0461538461538" type="integer">1</value>
    <value name="0.938461538462" type="integer">2</value>
    <value name="0.0153846153846" type="integer">3</value>
   </discrete-dist>
  </node-dist>
  <vector type="real" dim="3" name="training-dist">
   3 61 1
  </vector>
  <branch-values>
   <value type="integer">0</value>
   <value type="integer">1</value>
   <value type="integer">2</value>
   <value type="integer">3</value>
   <value type="integer">4</value>
  </branch-values>
  <split IV="21">
   <split-map branches="2">
    <value name="0" type="integer">3</value>
    <value name="0" type="integer">4</value>
    <value name="0" type="integer">6</value>
    <value name="1" type="integer">7</value>
    <value name="1" type="integer">8</value>
    <value name="1" type="integer">9</value>
    <value name="1" type="integer">10</value>
    <value name="1" type="integer">11</value>
    <value name="1" type="integer">12</value>
    <value name="1" type="integer">13</value>
    <value name="1" type="integer">14</value>
    <value name="1" type="integer">15</value>
    <value name="1" type="integer">16</value>
    <value name="1" type="integer">19</value>
    <value name="1" type="integer">20</value>
    <value name="1" type="integer">23</value>
```

```
</split-map>
<node iv="21">
 <node-dist>
  <total>6</total>
  <continuous-dist>
   <unweighted>0</unweighted>
   <weighted>0</weighted>
   <mean>0</mean>
   <stddev>0</stddev>
  </continuous-dist>
  <discrete-dist>
   <value name="0.333333333333" type="integer">1</value>
   <value name="0.666666666667" type="integer">2</value>
   <value name="0" type="integer">3</value>
  </discrete-dist>
 </node-dist>
 <vector type="real" dim="3" name="training-dist">
  2 4 0
 </vector>
 <branch-values>
  <value type="integer">3</value>
  <value type="integer">4</value>
  <value type="integer">6</value>
 </branch-values>
</node>
<node iv="21">
 <node-dist>
  <total>59</total>
  <continuous-dist>
   <unweighted>0</unweighted>
   <weighted>0</weighted>
   <mean>0</mean>
   <stddev>0</stddev>
  </continuous-dist>
  <discrete-dist>
   <value name="0.0169491525424" type="integer">1</value>
   <value name="0.966101694915" type="integer">2</value>
   <value name="0.0169491525424" type="integer">3</value>
  </discrete-dist>
 </node-dist>
 <vector type="real" dim="3" name="training-dist">
  1 57 1
 </vector>
 <branch-values>
  <value type="integer">7</value>
  <value type="integer">8</value>
  <value type="integer">9</value>
  <value type="integer">10</value>
  <value type="integer">11</value>
  <value type="integer">12</value>
  <value type="integer">13</value>
  <value type="integer">14</value>
  <value type="integer">15</value>
  <value type="integer">16</value>
  <value type="integer">19</value>
  <value type="integer">20</value>
  <value type="integer">23</value>
 </branch-values>
 <split IV="4">
  <split-map branches="2">
```

```
                        <value name="0" type="integer">0</value>
                        <value name="1" type="integer">1</value>
                        <value name="1" type="integer">2</value>
                        <value name="1" type="integer">3</value>
                        <value name="1" type="integer">4</value>
                       </split-map>
                       <node iv="4">
                        <node-dist>
                         <total>6</total>
                         <continuous-dist>
                          <unweighted>0</unweighted>
                          <weighted>0</weighted>
                          <mean>0</mean>
                          <stddev>0</stddev>
                         </continuous-dist>
                         <discrete-dist>
                          <value name="0" type="integer">1</value>
                          <value name="0.833333333333" type="integer">2</value>
                          <value name="0.166666666667" type="integer">3</value>
                         </discrete-dist>
                        </node-dist>
                        <vector type="real" dim="3" name="training-dist">
                         0 5 1
                        </vector>
                        <branch-values>
                         <value type="integer">0</value>
                        </branch-values>
                       </node>
                       <node iv="4">
                        <node-dist>
                         <total>53</total>
                         <continuous-dist>
                          <unweighted>0</unweighted>
                          <weighted>0</weighted>
                          <mean>0</mean>
                          <stddev>0</stddev>
                         </continuous-dist>
                         <discrete-dist>
                          <value name="0.0188679245283"
type="integer">1</value>
                          <value name="0.981132075472" type="integer">2</value>
                          <value name="0" type="integer">3</value>
                         </discrete-dist>
                        </node-dist>
                        <vector type="real" dim="3" name="training-dist">
                         1 52 0
                        </vector>
                        <branch-values>
                         <value type="integer">1</value>
                         <value type="integer">2</value>
                         <value type="integer">3</value>
                         <value type="integer">4</value>
                        </branch-values>
                        <split IV="20">
                         <split-map branches="2">
                          <value name="0" type="real">1</value>
                          <value name="0" type="real">2</value>
                          <value name="0" type="real">4</value>
                          <value name="1" type="real">5</value>
                         </split-map>
```

```
          <node iv="20">
           <node-dist>
            <total>48</total>
            <continuous-dist>
             <unweighted>0</unweighted>
             <weighted>0</weighted>
             <mean>0</mean>
             <stddev>0</stddev>
            </continuous-dist>
            <discrete-dist>
             <value name="0" type="integer">1</value>
             <value name="1" type="integer">2</value>
             <value name="0" type="integer">3</value>
            </discrete-dist>
           </node-dist>
           <vector type="real" dim="3" name="training-dist">
            0 48 0
           </vector>
           <branch-values>
            <value type="real">1</value>
            <value type="real">2</value>
            <value type="real">4</value>
           </branch-values>
          </node>
          <node iv="20">
           <node-dist>
            <total>5</total>
            <continuous-dist>
             <unweighted>0</unweighted>
             <weighted>0</weighted>
             <mean>0</mean>
             <stddev>0</stddev>
            </continuous-dist>
            <discrete-dist>
             <value name="0.2" type="integer">1</value>
             <value name="0.8" type="integer">2</value>
             <value name="0" type="integer">3</value>
            </discrete-dist>
           </node-dist>
           <vector type="real" dim="3" name="training-dist">
            1 4 0
           </vector>
           <branch-values>
            <value type="real">5</value>
           </branch-values>
          </node>
         </split>
        </node>
       </split>
      </node>
     </split>
    </node>
    <node iv="2">
     <node-dist>
      <total>7</total>
      <continuous-dist>
       <unweighted>0</unweighted>
       <weighted>0</weighted>
       <mean>0</mean>
       <stddev>0</stddev>
```

```
          </continuous-dist>
          <discrete-dist>
           <value name="0" type="integer">1</value>
           <value name="0.714285714286" type="integer">2</value>
           <value name="0.285714285714" type="integer">3</value>
          </discrete-dist>
         </node-dist>
         <vector type="real" dim="3" name="training-dist">
          0 5 2
         </vector>
         <branch-values>
          <value type="integer">5</value>
          <value type="integer">6</value>
          <value type="integer">7</value>
         </branch-values>
        </node>
       </split>
      </node>
      <node iv="3">
       <node-dist>
        <total>8</total>
        <continuous-dist>
         <unweighted>0</unweighted>
         <weighted>0</weighted>
         <mean>0</mean>
         <stddev>0</stddev>
        </continuous-dist>
        <discrete-dist>
         <value name="0.125" type="integer">1</value>
         <value name="0.5" type="integer">2</value>
         <value name="0.375" type="integer">3</value>
        </discrete-dist>
       </node-dist>
       <vector type="real" dim="3" name="training-dist">
        1 4 3
       </vector>
       <branch-values>
        <value type="integer">3</value>
        <value type="integer">4</value>
        <value type="integer">6</value>
       </branch-values>
      </node>
     </split>
    </node>
   </split>
  </node>
  <node iv="0">
   <node-dist>
    <total>8</total>
    <continuous-dist>
     <unweighted>0</unweighted>
     <weighted>0</weighted>
     <mean>0</mean>
     <stddev>0</stddev>
    </continuous-dist>
    <discrete-dist>
     <value name="0.75" type="integer">1</value>
     <value name="0.25" type="integer">2</value>
     <value name="0" type="integer">3</value>
    </discrete-dist>
```

```
      </node-dist>
      <vector type="real" dim="3" name="training-dist">
       6 2 0
      </vector>
      <branch-values>
       <value type="integer">3</value>
      </branch-values>
     </node>
     <node iv="0">
      <node-dist>
       <total>10</total>
       <continuous-dist>
        <unweighted>0</unweighted>
        <weighted>0</weighted>
        <mean>0</mean>
        <stddev>0</stddev>
       </continuous-dist>
       <discrete-dist>
        <value name="0" type="integer">1</value>
        <value name="1" type="integer">2</value>
        <value name="0" type="integer">3</value>
       </discrete-dist>
      </node-dist>
      <vector type="real" dim="3" name="training-dist">
       0 10 0
      </vector>
      <branch-values>
       <value type="integer">4</value>
       <value type="integer">5</value>
      </branch-values>
     </node>
    </split>
   </node>
   <node iv="5">
    <node-dist>
     <total>9</total>
     <continuous-dist>
      <unweighted>0</unweighted>
      <weighted>0</weighted>
      <mean>0</mean>
      <stddev>0</stddev>
     </continuous-dist>
     <discrete-dist>
      <value name="0.222222222222" type="integer">1</value>
      <value name="0.333333333333" type="integer">2</value>
      <value name="0.444444444444" type="integer">3</value>
     </discrete-dist>
    </node-dist>
    <vector type="real" dim="3" name="training-dist">
     2 3 4
    </vector>
    <branch-values>
     <value type="integer">3</value>
     <value type="integer">5</value>
     <value type="integer">6</value>
    </branch-values>
   </node>
  </split>
 </node>
 <node iv="9">
```

```
<node-dist>
 <total>23</total>
 <continuous-dist>
  <unweighted>0</unweighted>
  <weighted>0</weighted>
  <mean>0</mean>
  <stddev>0</stddev>
 </continuous-dist>
 <discrete-dist>
  <value name="0" type="integer">1</value>
  <value name="0.565217391304" type="integer">2</value>
  <value name="0.434782608696" type="integer">3</value>
 </discrete-dist>
</node-dist>
<vector type="real" dim="3" name="training-dist">
 0 13 10
</vector>
<branch-values>
 <value type="integer">3</value>
 <value type="integer">4</value>
</branch-values>
<split IV="2">
 <split-map branches="2">
  <value name="0" type="integer">0</value>
  <value name="0" type="integer">1</value>
  <value name="1" type="integer">2</value>
  <value name="1" type="integer">3</value>
  <value name="1" type="integer">4</value>
  <value name="1" type="integer">5</value>
 </split-map>
 <node iv="2">
  <node-dist>
   <total>9</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
    <stddev>0</stddev>
   </continuous-dist>
   <discrete-dist>
    <value name="0" type="integer">1</value>
    <value name="1" type="integer">2</value>
    <value name="0" type="integer">3</value>
   </discrete-dist>
  </node-dist>
  <vector type="real" dim="3" name="training-dist">
   0 9 0
  </vector>
  <branch-values>
   <value type="integer">0</value>
   <value type="integer">1</value>
  </branch-values>
 </node>
 <node iv="2">
  <node-dist>
   <total>14</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
```

```
        <stddev>0</stddev>
       </continuous-dist>
       <discrete-dist>
        <value name="0" type="integer">1</value>
        <value name="0.285714285714" type="integer">2</value>
        <value name="0.714285714286" type="integer">3</value>
       </discrete-dist>
      </node-dist>
      <vector type="real" dim="3" name="training-dist">
       0 4 10
      </vector>
      <branch-values>
       <value type="integer">2</value>
       <value type="integer">3</value>
       <value type="integer">4</value>
       <value type="integer">5</value>
      </branch-values>
     </node>
    </split>
   </node>
   <node iv="9">
    <node-dist>
     <total>10</total>
     <continuous-dist>
      <unweighted>0</unweighted>
      <weighted>0</weighted>
      <mean>0</mean>
      <stddev>0</stddev>
     </continuous-dist>
     <discrete-dist>
      <value name="0.3" type="integer">1</value>
      <value name="0.6" type="integer">2</value>
      <value name="0.1" type="integer">3</value>
     </discrete-dist>
    </node-dist>
    <vector type="real" dim="3" name="training-dist">
     3 6 1
    </vector>
    <branch-values>
     <value type="integer">5</value>
     <value type="integer">6</value>
     <value type="integer">7</value>
     <value type="integer">9</value>
    </branch-values>
   </node>
  </split>
 </node>
 <node iv="19">
  <node-dist>
   <total>92</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
    <stddev>0</stddev>
   </continuous-dist>
   <discrete-dist>
    <value name="0.054347826087" type="integer">1</value>
    <value name="0.45652173913" type="integer">2</value>
    <value name="0.489130434783" type="integer">3</value>
```

```
 </discrete-dist>
</node-dist>
<vector type="real" dim="3" name="training-dist">
 5 42 45
</vector>
<branch-values>
 <value type="integer">3</value>
</branch-values>
<split IV="5">
 <split-map branches="2">
  <value name="0" type="integer">0</value>
  <value name="0" type="integer">1</value>
  <value name="1" type="integer">2</value>
  <value name="1" type="integer">3</value>
  <value name="1" type="integer">4</value>
  <value name="1" type="integer">5</value>
 </split-map>
 <node iv="5">
  <node-dist>
   <total>66</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
    <stddev>0</stddev>
   </continuous-dist>
   <discrete-dist>
    <value name="0" type="integer">1</value>
    <value name="0.409090909091" type="integer">2</value>
    <value name="0.590909090909" type="integer">3</value>
   </discrete-dist>
  </node-dist>
  <vector type="real" dim="3" name="training-dist">
   0 27 39
  </vector>
  <branch-values>
   <value type="integer">0</value>
   <value type="integer">1</value>
  </branch-values>
  <split IV="22">
   <split-map branches="2">
    <value name="0" type="integer">1</value>
    <value name="1" type="integer">2</value>
   </split-map>
   <node iv="22">
    <node-dist>
     <total>24</total>
     <continuous-dist>
      <unweighted>0</unweighted>
      <weighted>0</weighted>
      <mean>0</mean>
      <stddev>0</stddev>
     </continuous-dist>
     <discrete-dist>
      <value name="0" type="integer">1</value>
      <value name="0.625" type="integer">2</value>
      <value name="0.375" type="integer">3</value>
     </discrete-dist>
    </node-dist>
    <vector type="real" dim="3" name="training-dist">
```

```
  0 15 9
 </vector>
 <branch-values>
  <value type="integer">1</value>
 </branch-values>
</node>
<node iv="22">
 <node-dist>
  <total>42</total>
  <continuous-dist>
   <unweighted>0</unweighted>
   <weighted>0</weighted>
   <mean>0</mean>
   <stddev>0</stddev>
  </continuous-dist>
  <discrete-dist>
   <value name="0" type="integer">1</value>
   <value name="0.285714285714" type="integer">2</value>
   <value name="0.714285714286" type="integer">3</value>
  </discrete-dist>
 </node-dist>
 <vector type="real" dim="3" name="training-dist">
  0 12 30
 </vector>
 <branch-values>
  <value type="integer">2</value>
 </branch-values>
 <split IV="7">
  <split-map branches="2">
   <value name="0" type="integer">0</value>
   <value name="1" type="integer">1</value>
   <value name="1" type="integer">2</value>
  </split-map>
  <node iv="7">
   <node-dist>
    <total>34</total>
    <continuous-dist>
     <unweighted>0</unweighted>
     <weighted>0</weighted>
     <mean>0</mean>
     <stddev>0</stddev>
    </continuous-dist>
    <discrete-dist>
     <value name="0" type="integer">1</value>
     <value name="0.205882352941" type="integer">2</value>
     <value name="0.794117647059" type="integer">3</value>
    </discrete-dist>
   </node-dist>
   <vector type="real" dim="3" name="training-dist">
    0 7 27
   </vector>
   <branch-values>
    <value type="integer">0</value>
   </branch-values>
   <split IV="12">
    <split-map branches="2">
     <value name="0" type="integer">1</value>
     <value name="1" type="integer">2</value>
     <value name="1" type="integer">3</value>
     <value name="1" type="integer">4</value>
```

```
    <value name="1" type="integer">5</value>
    <value name="1" type="integer">9</value>
   </split-map>
   <node iv="12">
    <node-dist>
     <total>6</total>
     <continuous-dist>
      <unweighted>0</unweighted>
      <weighted>0</weighted>
      <mean>0</mean>
      <stddev>0</stddev>
     </continuous-dist>
     <discrete-dist>
      <value name="0" type="integer">1</value>
      <value name="0.666666666667" type="integer">2</value>
      <value name="0.333333333333" type="integer">3</value>
     </discrete-dist>
    </node-dist>
    <vector type="real" dim="3" name="training-dist">
     0 4 2
    </vector>
    <branch-values>
     <value type="integer">1</value>
    </branch-values>
   </node>
   <node iv="12">
    <node-dist>
     <total>28</total>
     <continuous-dist>
      <unweighted>0</unweighted>
      <weighted>0</weighted>
      <mean>0</mean>
      <stddev>0</stddev>
     </continuous-dist>
     <discrete-dist>
      <value name="0" type="integer">1</value>
      <value name="0.107142857143" type="integer">2</value>
      <value name="0.892857142857" type="integer">3</value>
     </discrete-dist>
    </node-dist>
    <vector type="real" dim="3" name="training-dist">
     0 3 25
    </vector>
    <branch-values>
     <value type="integer">2</value>
     <value type="integer">3</value>
     <value type="integer">4</value>
     <value type="integer">5</value>
     <value type="integer">9</value>
    </branch-values>
   </node>
  </split>
 </node>
 <node iv="7">
  <node-dist>
   <total>8</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
```

```
        <stddev>0</stddev>
       </continuous-dist>
       <discrete-dist>
        <value name="0" type="integer">1</value>
        <value name="0.625" type="integer">2</value>
        <value name="0.375" type="integer">3</value>
       </discrete-dist>
      </node-dist>
      <vector type="real" dim="3" name="training-dist">
       0 5 3
      </vector>
      <branch-values>
       <value type="integer">1</value>
       <value type="integer">2</value>
      </branch-values>
     </node>
    </split>
   </node>
  </split>
 </node>
 <node iv="5">
  <node-dist>
   <total>26</total>
   <continuous-dist>
    <unweighted>0</unweighted>
    <weighted>0</weighted>
    <mean>0</mean>
    <stddev>0</stddev>
   </continuous-dist>
   <discrete-dist>
    <value name="0.192307692308" type="integer">1</value>
    <value name="0.576923076923" type="integer">2</value>
    <value name="0.230769230769" type="integer">3</value>
   </discrete-dist>
  </node-dist>
  <vector type="real" dim="3" name="training-dist">
   5 15 6
  </vector>
  <branch-values>
   <value type="integer">2</value>
   <value type="integer">3</value>
   <value type="integer">4</value>
   <value type="integer">5</value>
  </branch-values>
  <split IV="17">
   <split-map branches="2">
    <value name="0" type="integer">1</value>
    <value name="0" type="integer">2</value>
    <value name="1" type="integer">5</value>
   </split-map>
   <node iv="17">
    <node-dist>
     <total>13</total>
     <continuous-dist>
      <unweighted>0</unweighted>
      <weighted>0</weighted>
      <mean>0</mean>
      <stddev>0</stddev>
     </continuous-dist>
     <discrete-dist>
```

```
          <value name="0.384615384615" type="integer">1</value>
          <value name="0.538461538462" type="integer">2</value>
          <value name="0.0769230769231" type="integer">3</value>
         </discrete-dist>
        </node-dist>
        <vector type="real" dim="3" name="training-dist">
         5 7 1
        </vector>
        <branch-values>
         <value type="integer">1</value>
         <value type="integer">2</value>
        </branch-values>
       </node>
       <node iv="17">
        <node-dist>
         <total>13</total>
         <continuous-dist>
          <unweighted>0</unweighted>
          <weighted>0</weighted>
          <mean>0</mean>
          <stddev>0</stddev>
         </continuous-dist>
         <discrete-dist>
          <value name="0" type="integer">1</value>
          <value name="0.615384615385" type="integer">2</value>
          <value name="0.384615384615" type="integer">3</value>
         </discrete-dist>
        </node-dist>
        <vector type="real" dim="3" name="training-dist">
         0 8 5
        </vector>
        <branch-values>
         <value type="integer">5</value>
        </branch-values>
       </node>
      </split>
     </node>
    </split>
   </node>
  </split>
 </node>
</tree-model>
</predictive-model>
```

Dieser Code erwartet einen Parser, um die Nutzbarkeit des Programmes zu ermöglichen.

Anlage IV

Glossar

- A -

abhängige Variable

Das Feld, das Sie analysieren möchten.

Algorithmus

- Eine Prozedur zur Lösung eines wiederkehrenden mathematischen Problems.

analytisches Modell

Eine Struktur und ein Prozeß zur Analyse einer Datenmenge. Zum Beispiel, eine Entscheidungsbaum ist ein Modell zur Klassifizierung einer Datenmenge.

angepaßte Signifikanz

Eine Signifikanzprüfung, bei der die Testanzahl zur Bestimmung des Signifikanzgrades angepaßt wurde. Diese Anpassung verhindert Zufallsergebnisse im Rahmen der Signifikanz**.**

ANGOSS

A New Generation of Software Systems. ANGOSS ist eine spezialisierte Data Mining Software Herstellerfirma mit Sitz in Toronto, Kanada. Siehe unter im Web unter www.angoss.com. Hersteller von KnowledgeSTUDIO und KnowledgeSEEKER sowie KnoweldgeEXCELerator, KnowledgeACCESS, KnowledgeWebMiner.

anomale Daten

Daten, die aus Fehlern resultieren (zum Beispiel, Dateneingabe Fehler) oder die ungewöhnliche Ereignisse darstellen. Anomale Daten sollten vorsichtig untersucht werden, da sie vielleicht wichtige Informationen enthalten.

ANOVA

ANalysis Of VAriance. Eine Prozedur mit der statistisch signifikante Einflüsse einer unabhängigen Variable auf einer fortlaufenden abhängigen Variable entdeckt werden. Die ANOVA Prozedur verwendet einen F-Test, um die Unterschiede unter einer bestimmten Menge Durchschnitte zu messen, wo F = Durchschnittsquadrat für Auswirkung durch Durchschnittsquadrat für Fehler (F = MST/MSE).

Ansicht

In KnowledgeSTUDIO sind Ansichten eine Unterabteilung des Objektes. Zum Beispiel ein Baumobjekt hat eine Baumansicht und u.a. einer Grafikansicht. Ansichten bietet unterschiedliche Betrachtungs- und Handlungsweisen im Umgang mit Objekten. Ansicht Tabs am unteren Rand des Objektfensters werden zur Navigation im Objekt verwendet.

ASCII

Akronym für American Standard Code for Information Interchange. Das American National Standards Institute legte Standard Zeichencodes fest, zum Austausch zwischen Computern oder Computern und Peripheriegeräte. KnowledgeSTUDIO kann ASCII-Dateien importieren, wenn die Datei strukturiert ist – entweder mit festen Feldlängen oder mit Feldern, die per Sonderzeichen getrennt sind.

Außenwert

Eine Datenposition, deren Wert außerhalb der Grenzen liegt, die die meisten Werte im Muster umschließt. Kann Datenanomalien aufzeigen. Sollte sorgfältig untersucht werden, denn hier können wichtige Hinweise lagern

- B -

B Faktor

Bonferroni Anpassung für den P Wert.

bonferroni Anpassungen

Minimiert die Entdeckung von Zufallsverbindungen in einer Datenmenge. Bonferroni Anpassungen passen den Signifikanzgrad automatisch bei der Validierungsprüfung einer Verbindung an.

Browser

Neben der Bedeutung der Software Art, die als Zugang zu HTML Dateien sprich Internet ermöglicht, kommt dieser Terminus aus dem Web Mining. Ein Browser ist jemand, der sich in einem Website herumschaut.

- C -

CART

Classification and Regression Trees. Eine Entscheidungsbaumtechnik zur Klassifizierung einer Datenmenge. Bietet einen Regelsatz, den Sie einer neuen (nicht klassifizierten) Datenmenge zuführen können, um den Ausgang bestimmter Datensätze vorherzusagen. Segmentiert eine Datenmenge durch die Erstellung von Doppelverzweigungen. Benötigt weniger Datenvorbereitung als CHAID.

CHAID

Chi Square Automatic Interaction Detection. Eine Entscheidungsbaumtechnik zur Klassifizierung einer Datenmenge. Bietet einen Regelsatz, den Sie einer neuen (nicht klassifizierten) Datenmenge zuführen können, um den Ausgang bestimmter Datensätze vorherzusagen. Segmentiert eine Datenmenge durch Chi Quadrattests, die Mehrwegverzweigungen erzeugen. Benötigt eine größere Datenvorbereitung als CART.

chi2

Chi Square. Ein Test zur Messung der statistischen Verbindung zwischen zwei kategorischen Variablen

Clustertyp

Eine Möglichkeit Feldwerte zu behandeln oder zu manipulieren. Im allgemeinen wenn die Werte eines Feldes monotonisch (1, 2, 3, ... oder niedrig, mittel und hoch) dann sind sie geordnet. Wenn keine Ordnung vorliegt, z.B. rot, blau, grün, sind sie ungeordnet.

Clusterung

Der Prozess der Datenaufteilung in gegenseitig exklusive Gruppen, so dass die Mitglieder jeder Gruppe so „nah“ aneinander wie möglich sind und die unterschiedlichen Gruppen so „weit“ wie möglich auseinander sind, wo die Distanz in Relation zu allen verfügbaren Variablen gemessen wird.. KnowledgeSTUDIO verwendet die Clusterung im Modul Cluster Analyse.

Code

Feldwerte enthalten oft Code. Zum Beispiel, der Code für eine Region könnte sein Ost, West Süd, Nord und Mitte oder für ein ZIP Feld könnte ein Beispiel 60601, 90931 und 56041 sein. Während der Datenanalyse werden diese Codes oft in unterschiedliche Kombinationen zusammengefaßt, um Kernverbindungen zu identifizieren oder herauszustellen.

COM

Programmierungsumgebung. COM betont die Schnittstelle oder Verbindung zwischen den Komponenten.

- D -

Data Mining

Die Aufdeckung versteckter Vorhersageinformationen aus großen Datenbanken.

Data Warehouse

Ein System womit Riesenmengen von Daten gespeichert und zur Verfügung gestellt werden.

Datenbank

Zusammengetragene Informationen, die eng verbunden sind. Die meisten Datenbanken bestehen aus Felder oder Spalten, die Informationseinheiten enthalten und Datensätze oder Reihen, die Feldmengen enthalten.

Datenbereinigung

Der Vorgang, wobei alle Werte einer Datenmenge korrekt und folgerichtig aufgeführt sind.

Datennavigation

Der Betrachtungsvorgang der unterschiedlichen Dimensionen, Scheiben und Detailebenen einer multidimensionalen Datenbank. Siehe OLAP.

Datenmenge

In KnowledgeSTUDIO bilden Datenmengen die Quelldaten der Entscheidungsbäume und Modelle eines Projektes. Datenmengen enthalten direkt keine Daten, sondern zeigen auf die Datenquelle. Daher stimmen Datenmengen unabhängig vom Quellentyp stets überein.

Datensätze

Bezieht sich auf Informationen der Datenbank, die aus einer Eingabe für jedes Feld der Datenbank steht. Zum Beispiel ein Mitarbeiter Datenbank enthält einen Datensatz pro Mitarbeiter Manchmal Reihe genannt.

Datenvisualisierung

Die visualle Interpretation von komplexen Verbindungen bei multidimensionalen Daten.

DCOM

Programmierungsumgebung COM betont die Schnittstelle oder Verbindung zwischen den Komponenten. DCOM erweitert diese Leistung durch die transparente Umstellung von Komponenten in einem Netzwerk.

df

Freiheitsgrade (Degrees of freedom).

Dimension

In einer flachen oder relationalen Datenbank stellt jedes Feld eines Datensatzes eine Dimension dar. In einer multidimensionalen Datenbank ist eine Dimension eine Menge ähnliche Einheiten; zum Beispiel könnte eine multidimensionale Verkaufsdatenbank die Dimensionen Produkt, Zeit und Stadt enthalten.

diskrete Felder

Felder, die unterschiedliche Kategorien von Feldwerten aufweisen (z.B. sonnig, wolkig oder regnerisch). Auch als kategorische Felder bekannt.

Distanz Algorithmus

Mit diesem Algorithmus wird Ähnlichkeit in einem memory-based reasoning Modell gemessen.

- E -

Endknoten

Unterste oder Endknoten einer Analyse.

Entscheidungsbaum

Eine baumförmige graphische Darstellung von Verbindungen zwischen einer abhängigen Variable und einer Menge unabhängiger Variablen. Typischerweise wird die abhängige Variable oben oder links des Baumes positioniert (oberste Knoten), wobei die unabhängigen Variablen (Knoten) und ihre Verbindungen als Äste des Baumes angezeigt werden. Dies ist auch als Klassifizierungsbaum bekannt.

erschöpfende Partitionierung

Im Vergleich zur heuristischen Partitionierung wird die erschöpfende Partitionierung eher die Partitionierung mit der höchsten Signifikanz finden, denn es werden mehr Gruppierungen von Werten nach der abhängigen Variable gebildet. Erschöpfende Partitionierung benötigt mehr Zeit aber die gebildeten Partitionen sind empirisch stärker als heuristisch gebildete Partitionen. Bäume, die nach der erschöpfender Methode aufgebaut werden, neigen dazu, mehr Äste zu haben, als Bäume, die mit der heuristischen Methode aufgebaut wurden.

- F -

Feld

Eine Spalte in einer Datenbank, die für jeden Datensatz gleiche Informationsart enthält. Zum Beispiel, an Feld Alter enthält das Alter jeder Person der Datenbank. Manchmal Spalte genannt.

fehlende Werte

Werte, die in einem Feld nicht vorkommen, da kein Wert vorhanden ist. Können in der Analyse als drei Fragezeichen (???) erscheinen.

Fließen

Eine Option der Clusterung, die es ermöglicht, fehlende Werte eines geordneten Feldes mit anderen ähnlichen Werten zu gruppieren, d.h., sie haben einen ähnlichen Einfluß auf die abhängige Variable wie die Mitglieder ihrer Gruppe.

fortlaufende Felder

Felder, die einen numerischen oder geordneten Wertbereich enthalten, wie Temperaturwerte, z.B., 25, 26, 27 usw.

F-Verhältnis Statistisch

Ein errechneter Wert als Teil der ANOVA Prozedur. Je größer die Zahl, je größer der Abstand zwischen den Durchschnittswerten der Verzweigung. Siehe ANOVA.

- G -

genetische Algorithmen

Optimierungstechniken, die Prozesse wie genetische Kombination, Mutation und natürliche Selektion in einem auf den Konzepten der natürlichen Evolution basierten Design verwenden.

geordnet

Eine Clusteroption, die eine Wertmenge als geordnete Sequenz behandelt, und nur die Gruppierung von angrenzenden Werten erlaubt.

- H -

heuristische Partitionierung

Eine Methode der Felddatenpartitionierung, optimale Verzweigung auf der Basis der Heuristik oder der statistischen Schätzung. Die Zeitdauer ist geringer als erschöpfende Partitionierung und neigt dazu, weniger Äste zu erzeugen.

- I -

ID3

Ein Algorithmus aus dem maschinellen Lernen.

Induktion

Eine Methode mit der Aussagen über einer geordneten Datenmenge geprüft werden. Es bedeutet Schlußfolgerung aus den Besonderheiten auf das Allgemeine oder vom Individuellen auf das Universelle.

Interaktionseffekt

Ein Effekt der Verhältnisse zwischen zwei (oder mehrere) Variablen, wo die Richtung der Verhältnisse (d.h. positive oder negative) vom Wert einer anderen Variablen abhängt. Zum Beispiel, das Verhältnis zwischen Diät und Blutdruck ändert sich je nach Altersgruppe.

Intervall

Ein definierter Wertebereich.

Intervallgrenzen

Schnittpunkte in einem fortlaufenden Feld, die Intervalle einteilen.

- K -

kategorische Felder

Felder, die unterschiedliche Kategorien von Feldwerten aufweisen (z.B. sonnig, wolkig oder regnerisch). Auch als diskrete Felder bekannt.

Klassifizierung

Der Prozess der Datenaufteilung in gegenseitig exklusive Gruppen, so dass die Mitglieder jeder Gruppe so „nah“ aneinander wie möglich sind und die unterschiedlichen Gruppen so „weit“ wie möglich auseinander sind, wo die Distanz in Relation zu den zu vorhersagenden Variablen gemessen wird. Zum Beispiel, ein typisches Klassifizierungsproblem ist die Einteilung einer Firmendatenbank in möglichst homogene Gruppen nach einer Variable „Kreditwürdigkeit“ mit den Werten "Gut" und "Schlecht."

Klassifizierungsbaum

Eine graphische Darstellung von Verbindungen zwischen einer abhängigen Variable und eine Menge unabhängigen Variablen. Typischerweise wird die abhängige Variable oben oder links des Baumes positioniert (oberste Knoten), wobei die unabhängigen Variablen (Knoten) und ihre Verbindungen als Äste des Baumes angezeigt werden. Dies ist auch als Entscheidungsbaum bekannt.

K-means

Eine Technik, die jeden Datensatz einer Datenmenge nach einer Kombination der Klassen der K Sätze klassifiziert, mit der größten Ähnlichkeit des Datensatzes zu einer historischen Datenmenge (wo k größer gleich 1ist). Manchmal auch nächste Nachbar- oder k-nächste Nachbartechnik genannt.

Knowledge Discovery

Der Prozeß der Wissensgenerierung aus einer Datensammlung. Bei der Anwendung an elektronischen Daten, enthält der Prozeß – mit dem Akronym KDD bekannt – im Kern der Sache Data Mining. Im allgemeinen enthält Knowledge Discovery die Schritte Datenvorbereitung Mustersuche, Evaluierung, Aufwertung und Wiederholung. Letztendlich werden die entdeckten Muster gültig, nützlich, verständlich und bisher in der Datenmenge unbekannt sein.

Knoten

Eine Lokation, die durch Verzweigungsattribute eines Baumes definiert wird. Alle Verzweigungen stammen vom obersten Knoten. Die untersten Knoten sind die Abschlußknoten des Baumes.

Konfusionsmatrix

In KnowledgeSTUDIO bildet diese Grafik eine Kreuztabelle des echten Ausgangs gegenüber dem vorhergesagten Ausgang.

- L -

lineares Modell

Ein analytisches Modell, das lineare Beziehungen unter den Koeffizienten der untersuchten Variablen zugrunde legt.

lineare Regression

Eine statistische Technik, zur Findung der besten linearen Beziehung zwischen einer Zielvariablen (abhängigen Variablen und ihre Indikatoren (unabhängigen Variablen).

logistische Regression

Eine lineare Regression, die die Proportionen einer kategorischen Zielvariablen vorhersagt, sowie Kundentyp, in einer Menge.

- M -

Modell

Abkürzung für Vorhersagemodell – eine Struktur und ein Prozeß zur Vorhersage von Werten bestimmter Variablen einer Datenmenge.

multidimensionale Datenbank

Eine Datenbank für On-line Analytical Processing (OLAP). Strukturiert als multidimensionale Hypercube mit einer Achse pro Dimension.

multiprozessoriger Computer Fehler! Textmarke nicht definiert.

Ein Computer, der mehrere Prozessoren nebeneinander (auch im Netzwerk) enthält. Siehe Parallelverarbeitung. Siehe auch SMP.

- N -

nächster Nachbar

K-means Technik. Eine Technik, die jeden Datensatz einer Datenmenge nach einer Kombination der Klassen der K Sätze klassifiziert, mit der größten Ähnlichkeit des Datensatzes zu einer historischen Datenmenge (wo k größer gleich 1ist). Manchmal auch k-nächste Nachbartechnik genannt.

neuronales Netzwerk

Nicht-lineare Vorhersagemodelle, die durch Training lernen und in der Struktur, biologische Neuronen Netzwerke ähneln.

nicht-lineares Modell

Ein analytisches Modell, das nicht von linearen Beziehungen unter den Koeffizienten der untersuchten Variablen ausgeht.

Null Kategorien

Kategorien, die keinen entsprechenden Inhalt in einem Feld eines Unterknotens eines Baumes haben. Zum Beispiel, bei einer Serie „manche, kleine und sehr kleine“ erscheint kleine als Ast im Baum, auch wenn es keine tatsächlichen Inhalte für diese Kategorie im Knoten gibt. In Bäumen ist dies optional.

- O -

oberster Knoten

Der Knoten an der obersten Stelle in einer hierarchischen Baumanzeige. Dieser Knoten stellt auch die Werte der abhängigen Variable dar.

Objekt

In KnowledgeSTUDIO Sprache ist ein Objekt ein primärer Bestandteil eines Projektes. Zum Beispiel, Datenmengen, Entscheidungsbäume und Vorhersagemodelle sind Objekte. Ein Objekte kann mehr als eine Ansicht haben.

OLAP

On-line analytical processing. Hier sind array-orientierte Datenbankapplikationen gemeint, die es Anwendern ermöglichen, multidimensionale Datenbanken zu sehen, zu navigieren, zu manipulieren und zu analysieren. Besser als SQL bei der Erstellung von mulitdimensionaler Zusammenfassungen.

- P -

P Wert

Bonferroni-anpaßter P Wert.

parallele Verarbeitung

Die koordinierte Verwendung von mehreren Prozessoren, um Verarbeitungen zu leisten. Parallele Verarbeitung kann in einem Multiprozessorcomputer oder in einem Netzwerk von PCs oder Arbeitsstationen erfolgen. Siehe auch SMP.

Partitionierung

1. In KnowledgeSTUDIO wird der Partition Befehl verwendet, um Daten in zwei oder mehrere Datenmengen aufzuteilen. Hierbei wird meist eine Trainings- und eine Validierungs-(Trainings-)menge aus einer Datenmenge erstellt, wo die Werte der abhängigen Variable bekannt sind.
2. Die Aufteilung einiger Felder in diskreter Gruppen, auf der Basis der Ähnlichkeit entsprechend der abhängigen Variable nach einem Test der statistischen Signifikanz.

Projekt Arbeitsfläche

In KnowledgeSTUDIO ist das grundlegende Objekt das Projekt. – die Fläche selber heißt Projekt Arbeitsfläche. Es bindet die Objekte einer Analyse zusammen.

- R -

Reihe

In einer Datenbank stellt eine Reihe einen einzigen Datensatz dar. In einer Tabellenkalkulation ist sie eine waagerechte Serie von Zellen von einer Zellhöhe über die ganze Breite des Arbeitsblattes.

Regelinduktion

Die Extrahierung von nützlichen if-then Regeln aus den Daten nach der statistischen Signifikanz.

Regression

Siehe lineare oder logistische Regression unter L im Glossar sowie Regression unter Algorithmen in diesem Anhang.

rückblickende Datenanalyse

Datenanalyse, die Einblicke in bereits geschehene Trends, Verhaltensweisen oder Ereignisse liefert.

- S -

Score

Kategorisiert Datensätze aufgrund eines Modells. Vorhersagen werden hiermit gemacht.

Signifikanz

Ein Maß der Stärke einer Beziehung unter Musterelementen nach der statistischen Wahrscheinlichkeit.

SMP

Symmetric MultiProcessor. Eine Art multiprozessorigen Computers, wo der Speicherplatz unter den Prozessoren geteilt wird.

Spalte

In einer Datenbank stellt eine Spalte ein einziges Feld dar. In einer Tabellenkalkulation bedeutet sie einen senkrechten Zellenausschnitt mit einer einzelligen Breite mit der Länge des ganzen Arbeitsblattes.

Standardabweichung

Der Quadrat der Varianz. Das Maß der Unterschiedlichkeit in einer Datenmenge. Je höher der Wert je größer die Unterschiedlichkeit.

- T -

Transaktor

Begriff aus dem Web Mining. Besucher im Website, der auch Geschäfte abwickelt, d.h. er führt seine Kaufabsichten aus

Typenbibliothek

Programmierungsumgebung. Typenbibliotheken, bieten Programmierer den Zugang zu Objekten, wie die in KnowledgeSTUDIO enthalten sind.

- U -

unabhängige Variable

Ein Feld, das es Ihnen ermöglicht, die Variationen zu beschreiben oder vorherzusagen, die in den Werten einer abhängigen Variablen oder vorkommen.

ungeordnet

Eine Clusteroption, die es ermöglicht, Werte eines Feldes mit jedem anderen Wert im Feld zu verbinden, d.h., sie können mit jedem anderen Feld ungeachtet der Ordnung gruppiert werden.

untersuchende Datenanalyse

Die Verwendung von graphischen und beschreibenden statistischen Techniken, um über die Struktur einer Datenmenge lernen zu können.

uP

Unadjusted P-value für Chi2 und F Tests.

- V -

Varianz

Der zweite Moment um den Durchschnitt. Der erwartete Wert des Quadrats der Abweichungen einer wahllosen Variable von ihrem Durchschnittswert.

Verzweigung

Eine Partition bei einer Menge Feldwerte.

Vorhersagemodell

Eine Struktur und ein Vorgang zur Vorhersage der Werte von bestimmten Variablen der Datenmenge.

vorhersagende Datenanalyse

Datenanalyse, die zukünftige Trends, Verhaltensweisen oder Ereignisse auf der Basis historischer Daten vorhersagt.

- W -

Web Mining

Web Mining ist eine besondere Form des Data Mining. Web Mining beschäftigt sich mit der Aufbereitung und Analyse von Daten aus dem E-Business und bereitet sie auf. Es können Vorhersagemodelle z.B. zur Unterbreitung von individuellen Online Angeboten entwickelt und eingesetzt werden. Siehe Kapitel „Data Mining und der Web."

- X -

XML

XML ist eine Sprache (Extensible Markup Language) für Dokumente, die strukturierte Informationen enthalten. Strukturierte Informationen enthalten sowohl Inhalt (Wörter, Bilder usw.) als auch einen Hinweis darauf, welche Rolle dieser Inhalt spielt (z.B. Inhalt in einer Kopfzeile hat eine andere Bedeutung als in einer Fußzeile, was wiederum anders ist als eine Schaubildbezeichnung order Inhalt einer DB-Tabelle usw.).

Eine Markup Language ist ein Mechanismus zur Identifikation von Strukturen in einem Dokument. Die XML Spezifikation definiert einen Standard, Dokumente mit diesem Mechanismus Markup zu versehen.

- Z -

Zeichencode

Ein numerisches System zur Darstellung von Zeichen, die am Bildschirm oder in einer Druckdatei erscheinen.

Zeitserie Analyse

Die Analyse einer Sequenz von Messungen, die zu bestimmten Zeiten erfolgt sind. Hierbei ist die Zeit meist die dominierende Dimension der Datenmenge.

Zelle

Grundeinheit der Tabellenkalkulation, in der Sie Daten und Formeln speichern Eine Zelle wird bei der Kreuzung der Spalte und der Reihe gebildet.

Anlage V

Literatur

ANGOSS Software Corporation – diverse Publikationen

Berry / Linoff – Data Mining Techniques

Oxford at the Clarendon Press – The Concise Oxford Dictionary

Tantau Software Corporation – diverse Publikationen

Westphal / Blaxton – Data Mining Solutions