Institut für Parallele und Verteilte
Höchstleistungsrechner (IPVR)
Universität Stuttgart

Gehe Informatik Services
GmbH & Co. KG (GEHIS)

Prüfer:	*Prof. Dr. B. Mitschang*
Betreuer:	*Dipl.-Inform. Holger Schwarz (IPVR)*
	Dipl.-Inform. (FH) Volker Wenzel (GEHIS)
Beginn am:	*03.12.1999*
Beendet am:	*16.06.2000*
CR-Nummer:	*H.2.8*

Diplomarbeit-Nr. 1821

Data Mining im Pharma-Großhandel

Joachim Kimmerle

Danksagungen

Hiermit möchte ich bei allen Personen bedanken, die zum Gelingen dieser Diplomarbeit beigetragen haben, vor allem bei den Mitarbeitern der GEHIS und der Gehe Pharma Handel GmbH. Mein besonderer Dank gilt meinen Betreuern, Herrn Volker Wenzel von der GEHIS und Herrn Holger Schwarz von der Universität Stuttgart, die mir stets mit Anregungen zur Seite standen. Aus den Unternehmensbereichen Marketing und Vertrieb gilt mein Dank Herrn Frisch und Herr Steinke, die mir zahlreiche Hinweise zu dem betriebswirtschaftlichen Teil meiner Diplomarbeit gegeben haben.
Zudem möchte ich mich bei Herrn Frank Theisen von der IBM Deutschland Informationssysteme GmbH und Herrn Lars Priebe von der Oracle Deutschland GmbH bedanken, die mich bei der praktischen Arbeit mit den Produkten Ihrer Unternehmen tatkräftig unterstützten.

Inhaltsverzeichnis

Kapitel 1. Einleitung 1
1.1. Motivation 1
1.2. Aufbau der Arbeit 2

Kapitel 2. Begriffsbestimmungen – Knowledge Discovery und Data Mining 4
2.1. Von OLTP zum Konzept des Data Warehouse 4
2.1.1. Online Transactional Processing (OLTP) 4
2.1.2. Online Analytical Processing (OLAP) 5
2.1.3. Gründe für Data Warehousing 6
2.2. Data Warehouse 7
2.2.1. Anforderungen an ein Data Warehouse 7
2.2.2. Architektur einer Data Warehouse Umgebung 8
2.3. Knowledge Discovery in Databases 11
2.3.1. Abgrenzung zu OLAP 11
2.3.2. Definition von Knowledge Discovery in Databases und Data Mining 12
2.3.3. Der KDD-Prozess 12
2.4. Data Mining-Methoden 15
2.4.1. Aufgaben von Data Mining 15
2.4.2. Überblick über die Data Mining-Methoden 15
2.5. Data Mining-Techniken 17
2.5.1. Entscheidungsbäume 18
2.5.2. Neuronale Netze 19
2.5.3. Assoziationsregeln 21
2.5.4. Verfahren zum Clustering 22
2.5.5. Statistische Verfahren 23

Kapitel 3. Evaluation der Data Mining-Tools 26
3.1. Vorauswahl der Data Mining-Tools 26
3.1.1. Marktübersicht 26
3.1.2 Kriterien für die Auswahl der Tools 27
3.1.3. Testumgebung der ausgewählten Tools 28
3.2. Beschreibung der Data Mining Tools 29
3.2.1. Intelligent Miner for Data v6.1 29
3.2.2. Enterprise Miner v3.0 32
3.2.3. Darwin v3.6.0 33
3.2.4. KnowledgeStudio v2.0 34
3.2.5. XpertRule Miner v1.46 36
3.3. Vorgehensweise bei der Evaluation 37
3.3.1. Vorstellung der Kriterienpunkte 37
3.3.2. Bewertung der Tools anhand der Einzelkriterien 41
3.3.3. Gewichtung der Einzelkriterien 41
3.4. Ergebnisse der Evaluation 42
3.4.1. Ergebnisse der Kriterienpunkte 42
3.4.2. Übersicht über die Gesamtbewertung der Tools 49

Kapitel 4.Anwendungsgebiete für Data Mining bei der Gehe Pharmahandel GmbH.....52
4.1. Unternehmensumfeld der Gehe Pharmahandel GmbH 52
4.1.1. Funktionen des Pharmazeutischen Großhandels 52
4.1.2. Markteigenschaften.....53
4.1.3. Beschreibung der Sortimente.....53
4.1.4. Abwicklung des Tagesgeschäftes 53
4.1.5. Niederlassungsstruktur der Gehe Pharmahandel GmbH 54
4.2. Struktur des Data Warehouse.....54
4.2.1. Transport der Auftragsdaten 54
4.2.2. Tabellenstruktur.....55
4.2.3. Verdichtung der Daten.....56
4.2.4. Daten der IMS 57
4.3. Einsatzgebiete für Data Mining bei der Gehe.....57
4.4. Beschreibung der Business Cases 59
4.4.1. Potentialermittlung von Kunden aus demografischen Merkmalen.....60
4.4.2. Entdeckung erfolgreicher Kunden aus kundenspezifischen Merkmalen.....61
4.4.3. Sortimentszusammensetzung.....61
4.4.4. Warenkorbanalyse 62
4.4.5. Zielgruppenidentifizierung für Online-Dienst POINT 63
4.4.6. Erfolgsfaktoren für Online-Dienst POINT 64
4.4.7. Einflussfaktoren für ausstehende Zahlungen.....64
4.4.8. Zuordnung der Business Cases zu Data Mining-Methoden 65
4.5. Auswahl der Business Cases zur Implementierung 66

Kapitel 5. Implementierung der Business Cases.....67
5.1. Reinigung der Kundendaten 67
5.2. Implementierung der Sortimentszusammensetzung.....68
5.2.1. Analyse anhand der demografischen Daten.....68
5.2.2. Analyse anhand einer Niederlassung.....70
5.2.3. Analyse auf der Ebene der ATC3-Warengruppen.....71
5.2.4. Interpretation der Ergebnisse.....73
5.3. Implementierung der Warenkorbanalyse.....74
5.3.1. Selektion der Analysedaten 74
5.3.2. Ergebnisse der Warenkorbanalyse auf Warengruppenebene.....76
5.3.3. Ergebnisse der Warenkorbanalyse auf Artikelebene 77
5.4. Implementierung der Potentialermittlung.....78
5.4.1. Berechnung der Zielgröße 78
5.4.2. Auswahl der Kundengruppe 79
5.4.3. Erstellung des Entscheidungsbaummodells.....80

Kapitel 6: Schlussfolgerungen aus der Arbeit und Ausblick.....83

Anhang A: Literaturverzeichnis 86
Anhang B: Übersicht über die Einzelkriterien 90
Anhang C: Darstellung der Gewichtung und Bewertung der Einzelkriterien 92
Einordnung des Herstellers 92
Systemarchitektur 93
Datenexploration 94
Datenvorverarbeitung 94
Funktionalität Klassifikation 95
Funktionalität Segmentierung 96
Funktionalität Abhängigkeitsanalyse 97
Funktionalität Änderungs- und Abweichungserkennung 97
Bedienung 98
Automatisierbarkeit 98
Anhang D: Darstellung der Gesamtbewertung 99
Intelligent Miner for Data v6.1 99
Enterprise Miner v3.0 100
Darwin v3.6.0 101
KnowledgeStudio v2.0 102
XpertRule Miner v1.46 103
Anhang E: Grundbegriffe der Statistik 104
Verteilungen 104
Absolute und relative Häufigkeiten einer Verteilung 104
Gliederung einer Verteilung 104
Kategorisierung von Daten 105
Skalenniveau von Merkmalen 105
Kennzeichnung des Zentrums 105
Kennzeichnung der Variabilität 106
Schiefe und Wölbung 106
Analysemethoden 107

Abbildungsverzeichnis

Abbildung 2.1: Übergang zur multidimensionalen Darstellung 6
Abbildung 2.2: Grobaufbau eines Data Warehouse-Systems 9
Abbildung 2.3: Analyseansätze im Überblick 11
Abbildung 2.4: Der KDD-Prozess 13
Abbildung 2.5: Verbindung zwischen Data Mining-Zielen und Data Mining-Methoden 17
Abbildung 2.6: Zuordnung von Data Mining-Techniken zu Data Mining-Methoden 17
Abbildung 2.7: Beispiel eines Entscheidungsbaumes zur Vorhersage des ROI 18
Abbildung 2.8: Schematische Darstellung eines Neurons 20
Abbildung 2.9: Streudiagramm-Varianten mit Regressionsgeraden und Dichte-Ellipse 25
Abbildung 3.1: Einteilung der Data Mining-Tools 27
Abbildung 3.2: Testumgebung der Data Mining-Tools 29
Abbildung 3.3: Ansicht der Mining Base beim Intelligent Miner von IBM 30
Abbildung 3.4: Ergebnisobjekt des demografischen Clusterings 31
Abbildung 3.5 : Benutzungsoberfläche beim Enterprise Miner 32
Abbildung 3.6 : Ergebnisse einer Modellsuche bei Oracle Darwin 34
Abbildung 3.7 : Oberfläche des KnowledgeStudio 35
Abbildung 3.8: Oberfläche des XpertRule Miner 36
Abbildung 3.9: Testergebnis einer Klassifikation beim XpertRule Miner 37
Abbildung 3.10: Architekturen für Data Mining-Systeme 39
Abbildung 3.11: Bewertung der Herstellereinordnung 43
Abbildung 3.12: Bewertung der Systemarchitektur 43
Abbildung 3.13: Bewertung der Datenexploration 44
Abbildung 3.14: Bewertung der Datenvorverarbeitung 45
Abbildung 3.15: Bewertung der Klassifikation 45
Abbildung 3.16: Bewertung der Segmentierung 46
Abbildung 3.17: Bewertung der Abhängigkeitsanalyse 47
Abbildung 3.18: Bewertung Änderungs- und Abweichungserkennung 47
Abbildung 3.19: Bewertung der Bedienung 48
Abbildung 3.20: Bewertung der Automatisierbarkeit 49
Abbildung 3.21: Vergleich der Gesamtbewertung Enterprise Miner – Intelligent Miner 50
Abbildung 3.22: Vergleich der Gesamtbewertung Darwin – KnowledgeStudio – XpertRule Miner 51
Abbildung 4.1: Übersicht über die Tabellenstruktur 55
Abbildung 4.2: Verdichtungsebenen des Data Warehouse 56
Abbildung 4.3: Zuordnung Anwendungsgebiete – Business Cases 60
Abbildung 4.4 : Zuordnung Business Cases zu Data Mining-Methoden 65
Abbildung 5.1: Transponierung der Umsatzzahlen 69
Abbildung 5.2: Zusammenstellung der Analysetabelle zur Sortimentszusammensetzung 69
Abbildung 5.3: Kundensegmente der Niederlassung Stuttgart 70
Abbildung 5.4: Sortimentsschwerpunkte in den Wintermonaten 72
Abbildung 5.5: Sortimentsschwerpunkte in den Sommermonaten 73
Abbildung 5.6: Einschränkungen bei der Datenmenge 75
Abbildung 5.7: Darstellung der Analysetabelle 75
Abbildung 5.8: Visualisierung der Assoziationsregeln 76
Abbildung 5.9: Ergebnisse der Warenkorbanalyse auf Artikelebene 77
Abbildung 5.10: Zusammenstellung der Analysetabelle zur Potentialermittlung 80
Abbildung 5.11: Ausschnitt aus dem Entscheidungsbaummodell 81
Abbildung 5.12: Darstellung der Genauigkeit des Entscheidungsbaummodells 81

Tabellenverzeichnis:

Tabelle 2.1: Charakteristika von operativen und informativen Datenbanken.....7
Tabelle 3.1: Bewertungsmöglichkeiten der Einzelkriterien.....41
Tabelle 3.2: Umrechnung der Bewertung in Punktzahlen 42
Tabelle 5.1: Kundenklassifikation 79
Tabelle 5.2: Übersicht über die Umsatzkategorien.....81

Kapitel 1. Einleitung

In diesem einführenden Kapitel werden zunächst die Gründe skizziert, die zu der Entstehung dieser Diplomarbeit führten. Anschließend werden die Ziele der Arbeit vorgestellt. Im nächsten Abschnitt wird eine Übersicht über den Aufbau der Diplomarbeit gegeben.

1.1. Motivation

Die Extraktion von verwertbarem Wissen aus Daten ist ein Thema, das angesichts der Menge des zur Zeit verfügbaren Datenmaterials mehr und mehr an Aktualität gewinnt. In vielen Unternehmen und wissenschaftlichen Institutionen existieren sehr große Datenbestände, deren genauere Analyse nutzbare Erkenntnisse verspricht. Mit dem Konzept des Data Warehouse und den Methoden des Online Analytical Processing (OLAP) stehen zwar erfolgversprechende Ansätze zur Speicherung und Selektion entscheidungsrelevanter Informationen zur Verfügung, doch decken diese beiden Instrumente häufig nicht den notwendigen Informationsbedarf. Daher steigt die Attraktivität von Informationssystemen, die große Datenbestände automatisch analysieren und komplexe Muster aufdecken können. Diese Verfahren und Werkzeuge zur ungerichteten Informationsselektion können unter dem Begriff Data Mining zusammengefasst werden.

Auch bei der Gehe Pharmahandel GmbH, einem Unternehmen, das im Bereich des pharmazeutischen Großhandels tätig ist, wurde vor einigen Jahren mit dem Aufbau eines Data Warehouse begonnen. Aufgrund der Gesundheitsreform und gesetzlichen Verordnungen nimmt der Konkurrenzdruck auf die in diesem Markt befindlichen Unternehmen immer mehr zu. Daher wächst der Bedarf an entscheidungsrelevanten Informationen, mit denen den zukünftigen Herausforderungen begegnet werden kann. Dazu gehört auch die Untersuchung, ob sich in den unternehmensinternen Datenbeständen bislang unentdeckte Erkenntnisse verbergen. In Zusammenarbeit mit der Universität Stuttgart konnten mit einer Beteiligung an dem europäischen Forschungsprojektes CRITIKAL schon erste Erkenntnisse im Bereich des Data Mining gewonnen werden. Um diese Erkenntnisse zu vertiefen wurde eine Diplomarbeit angeboten, in deren Rahmen die Anwendungsmöglichkeiten von Data Mining im Bereich des pharmazeutischen Großhandels näher untersucht werden sollten. Mit den Ergebnissen der Diplomarbeit sollte auch eine Grundlage für den produktiven Einsatz von Data Mining bei der Gehe geschaffen werden.

Die Gehe Pharmahandel GmbH gehört zu der Gehe AG, bei der es sich um eine Holding für verschiedene europäische Gesellschaften mit den strategischen Geschäftseinheiten Pharma-Großhandel und Pharma-Einzelhandel handelt. Durchgeführt wurde die Diplomarbeit bei der Gehe Informatik Services GmbH & Co. KG (GEHIS), bei der es sich um ein hundertprozentiges Tochterunternehmen der Gehe Pharmahandel GmbH mit rund 90 Mitarbeitern in Deutschland handelt. Die GEHIS wurde 1994 gegründet, um alle EDV-Aktivitäten der Gehe Pharmahandel und Gehe AG zu bündeln.

Zu den Aufgaben dieser Arbeit gehörte zunächst, eine Übersicht über das Themengebiet des Data Mining zu erstellen. Dabei stand im Vordergrund, welche Methoden und Techniken in diesem Bereich existieren und welche Fragestellungen sich damit bearbeiten lassen. Bei Überprüfung, ob sich im Unternehmen Anwendungsgebiete für Data Mining identifizieren lassen, sollten auch die Besonderheiten des Marktes für pharmazeutische Produkte berücksichtigt werden, was eine Einarbeitung in das Unternehmensumfeld der Gehe erforderte. Dabei war es besonders wichtig, dass zur Formulierung der konkreten Fragestellungen die Fachbereiche hinzugezogen wurden. Auch die Eignung der Daten, die im Data Warehouse vorliegen, sollte geprüft werden. Ein weiteres Ziel der Arbeit bestand darin, neben der Software aus dem CRITIKAL-Projekt mehrere kommerzielle Data Mining-Tools im Hinblick auf ihre Einsatzmöglichkeiten im Bereich des Pharma-Großhandels zu vergleichen. Zur Bewertung dieser Tools wurde ein Kriterienkatalog benötigt, der neben allgemeinen Bewertungskriterien auch Gehe-spezifische Kriterien enthält, die in Zusammenarbeit mit den Benutzern zu definieren waren. In einer Implementierungsphase sollte die Realisierung einiger ausgewählter Business Cases mit diesen Tools erfolgen. Aus den Ergebnissen dieser Implementierung sollten Erfahrungen und Hinweise gewonnen werden, die den erfolgreichen Einsatz von Data Mining im Unternehmen vorbereiten.

1.2. Aufbau der Arbeit

Im zweiten Kapitel wird zunächst eine allgemeine Einführung in die wichtigsten Begriffe gegeben, die es in dem Bereich Knowledge Discovery und Data Mining gibt. Dabei werden sowohl die Phasen des Data Mining-Prozesses beschrieben als auch eine Übersicht über einige Data Mining-Methoden gegeben. Anschließend erfolgt ein Einblick in die Arbeitsweise einiger weit verbreiteter Algorithmen, die zur Bearbeitung dieser Methoden verwendet werden. Im Mittelpunkt des folgenden Kapitel steht dann die Evaluation der Data Mining-Tools, die im Rahmen dieser Diplomarbeit stattfand. Aufbauend auf einem Überblick über den Markt für Data Mining-Produkte erfolgt eine Vorstellung der Tools, die für diese Evaluation betrachtet wurden. Die Bewertung der Tools wurde anhand eines Kriterienkatalogs vorgenommen, der neben allgemeinen Anforderungen an die Software auch Wünsche auf Seiten der Benutzer bei Gehe berücksichtigte. Im Anschluss daran werden die Ergebnisse der Evaluation näher betrachtet, wobei die Stärken und Schwächen der einzelnen Tools gegenübergestellt werden. Im vierten Kapitel wird auf die Besonderheiten des Data Mining im Bereich des pharmazeutischen Großhandels eingegangen. Dazu werden die einzelnen Fragestellungen, die es in den verschiedenen Unternehmensbereichen gibt, zu sogenannten Business Cases zusammengefasst und näher erläutert. Für diese Business Cases wird eine Zuordnung zu den einzelnen Data Mining-Methoden vorgenommen. Am Ende des Abschnittes erfolgt die Auswahl dreier Business Cases, die in einer Implementierungsphase näher betrachtet werden. Im fünften Kapitel wird diese Phase einer genaueren Betrachtung unterzogen. Zu jedem Business Case wird beschrieben, welche Vorgehensweise bei seiner Implementierung gewählt wurde und welche Ergebnisse hieraus resultieren. Im letzten Kapitel stehen die Schlussfolgerungen, die sich aus dieser Diplomarbeit für den Einsatz von

Data Mining bei der Gehe ergeben. Im Anhang finden sich noch zusätzliche Informationen zur Bewertung und Gewichtung der Einzelkriterien und eine grafische Darstellung der Gesamtbewertung der einzelnen Data Mining-Tools. Auch eine Übersicht über die statistischen Begriffe, die bei dieser Diplomarbeit verwendet wurden ist beigefügt.

Kapitel 2. Begriffsbestimmungen – Knowledge Discovery und Data Mining

Now that we have gathered so much data, what do we do with it? Diese Frage stellen sich immer mehr Unternehmen. Unsere Fähigkeiten, große Datenmengen zu verstehen und zu analysieren sind weit weniger ausgeprägt als die Möglichkeiten, diese Daten zu erzeugen und zu archivieren. Viele Unternehmen haben schon vor Jahr(zehnt)en damit begonnen, alle anfallenden Informationen wie relevante Marktkennzahlen, Mitbewerberdaten oder Kundendaten zu sammeln und zentral zu archivieren. Durch die Einführung von Barcodes und Scannerkassen sowie durch die zunehmende Computerisierung vieler Unternehmensbereiche, beispielsweise durch die Bezahlung mit Kredit- oder Kundenkarten, haben sich diese Datenmengen sprunghaft erhöht. Gleichzeitig sind die Hardwarepreise, vor allem für Prozessorleistung und Speicher kontinuierlich gesunken, was dazu führt, dass die Datenbanken in bislang unbekannte Dimensionen wachsen. Die US-amerikanische Warenhauskette Wal-Mart, mit mehr als 2000 Geschäften eines der größten Einzelhandelsunternehmen weltweit, speichert mehr als 20 Millionen Transaktionen *täglich*. Dieses Beispiel mag extrem erscheinen, aber auch mittelständische Unternehmen können schon Datenbanken mit beachtlichen Volumen vorweisen. Auch im wissenschaftlichen Bereich sind Datenbanken entstanden, die die Größe von einem Terabyte weit übersteigen[1]. Es ist klar, dass diese Datenmengen mit den traditionellen Methoden der Datenanalyse wie Berichten und ad-hoc Abfragen nicht mehr bewältigt werden können. In den folgenden Abschnitten soll ein kurzer Einblick gegeben werden, welche Methoden und Techniken entwickelt wurden, um dieses Problem zu lösen.

2.1. Von OLTP zum Konzept des Data Warehouse

In diesem Abschnitt werden zunächst die Begriffe Online Transactional Processing (OLTP) und Online Analytical Processing (OLAP) voneinander abgegrenzt. Anschließend werden anhand der Anforderungen moderner Analysewerkzeuge die Gründe verdeutlicht, die zur Entwicklung eines Data Warehouse-Konzeptes führen.

2.1.1. Online Transactional Processing (OLTP)

Im operativen Tagesgeschäft werden heute in nahezu allen Wirtschaftszweigen, zur Verwaltung der umfangreichen und ständig wachsenden Datenbeständen Datenbanksysteme eingesetzt. Sie sorgen für eine dauerhafte und fehlerrobuste Speicherung sowie einen effizienten Zugriff. Dieser Art des Einsatzes lässt sich mit dem Begriff des Online Transactional Processing (OLTP) umschreiben [Kem97]. Typische Beispiele hierfür sind

[1] Gemeint sind hier u.a. das Human Genome Database Project (Fasman, Cuticchia, & Kingsbury 1994) in dem Informationen über den genetischen Code des Menschen archiviert werden, sowie das NASA Earth Observing System (Way & Smith 1991), das Bilddaten von Satelliten verwaltet. Weitere Beispiele in [Fay96d].

Anwendungen wie „Buchung eines Fluges“ in einem Flugreservierungssystem oder „Verarbeitung einer Bestellung“ in einem Handelsunternehmen. OLTP-Anwendungen zeichnen sich dadurch aus, dass sie nur relativ kleine Datenmengen pro Transaktion zu verarbeiten haben und dass sie nur auf dem jüngsten, aktuell gültigen Zustand des Datenbestandes operieren. Die typische Datenstruktur besteht meist aus einem relationalen Datenbanksystem mit flachen, nicht hierarchischen Tabellen [Kim96].

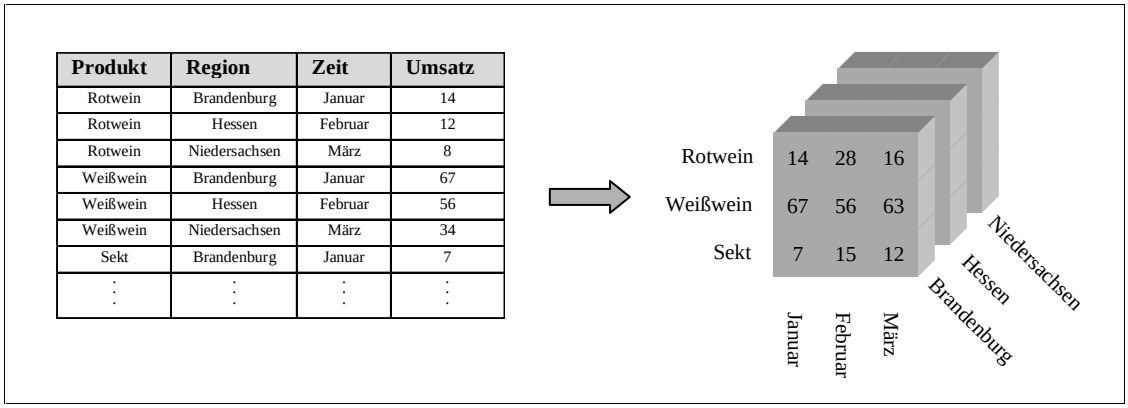
2.1.2. Online Analytical Processing (OLAP)

Während sich OLTP-Anwendungen auf die Verarbeitung von Transaktionen konzentrieren, benötigen viele Managementaufgaben Informationen aus unternehmensinternen Datenquellen, jedoch in einer aufbereiteten und verdichteten Form. Die dabei verwendeten Anwendungen greifen auf „historische“ Daten zurück, um daraus z.B. Rückschlüsse auf die Entwicklung des Unternehmens zu ziehen. Typische Anfragen bei einer Fluggesellschaft oder einem Handelsunternehmen sind:

- Wie hat sich die Auslastung der Transatlantikflüge über die letzten zwei Jahre entwickelt?
- Wie haben sich besonders offensive Marketingstrategien für bestimmte Produktlinien auf die Verkaufszahlen ausgewirkt?

Hier interessieren nicht mehr einzelne Geschäftsvorfälle, wie sie in den operationalen Systemen des Tagesgeschäftes abgebildet sind, sondern managementrelevante und aufbereitete Kenngrößen [Hol99]. Für diese Anforderungen sind die OLTP-Systeme mit ihren Strukturen nicht konzipiert worden. Darüber hinaus besteht in relationalen Systemen bei umfangreichen Datenbeständen oft eine unüberschaubare Anzahl von Tabellen, Views und Joins, deren Struktur nur von einem versierten Anwender beherrscht werden kann.
Aus diesem Grund stellte E.F. Codd im Jahr 1993 anhand von 12 Regeln einen neuen Ansatz vor, das Online Analytical Processing [Cod93]. Im Mittelpunkt dieses Ansatzes steht hierbei vor allem die multidimensionale Sicht auf die Daten. Damit können die gespeicherten Datenbestände auf vielfältige Art und Weise und von verschiedenen Seiten betrachtet werden. Aufgrund spezialisierter Abfragesprachen, einfacher Navigation im multidimensionalen Datenbestand, intuitiver Benutzungsoberflächen und geeigneter Ergebnispräsentation bleibt die tatsächliche Komplexität einer Abfrage dem Anwender verborgen [Cla98]. Zur Veranschaulichung der Multidimensionalität der Datenbasis wird diese oft als dreidimensionaler Datenwürfel abgebildet, da in einer Graphik nicht mehr Dimensionen darstellbar sind. Tatsächlich unterliegt sie aber keiner derartigen Beschränkung.
Abbildung 2.1 verdeutlicht die Zusammenhänge am Beispiel eines Weinhandels. Durch eine intuitive Datenbearbeitung, z.B. die Rotation des Würfels (*Slicing*) oder das Auswählen einer Teilmenge (*Data Dicing*), besteht für den Anwender eine erhöhte Flexibilität, so dass neue Ideen und Hypothesen schnell überprüft werden können. Sollen detailliertere Fragestellungen beantwortet werden, so kann der Verdichtungsgrad der einzelnen Dimension verändert werden [Hol99]. Will man beispielsweise den Jahresumsatz auf die einzelnen Monate abbilden so spricht man von einem *Drill down*, da dadurch eine weniger starke Verdichtung

der Daten stattfindet. Bei einem Wechsel zu einer höheren Verdichtungsstufe spricht man von einem *Roll up*.



Produkt	Region	Zeit	Umsatz
Rotwein	Brandenburg	Januar	14
Rotwein	Hessen	Februar	12
Rotwein	Niedersachsen	März	8
Weißwein	Brandenburg	Januar	67
Weißwein	Hessen	Februar	56
Weißwein	Niedersachsen	März	34
Sekt	Brandenburg	Januar	7
⋮	⋮	⋮	⋮

Abbildung 2.1: Übergang zur multidimensionalen Darstellung

Man kann drei Arten von OLAP-Anwendungen unterscheiden [Cla98]:

- **MOLAP – Multidimensionales OLAP:** Das Multidimensionale OLAP ist die klassische OLAP-Architektur. Die im RDBMS oder anderen Quellen vorliegenden Daten werden in die OLAP-Datenbank in Form einer optimierten, mehrdimensionalen Struktur geladen, ganz oder teilweise kalkuliert und stehen dann zur Abfrage mit einem Frontend bereit.

- **ROLAP – Relationales OLAP:** Die zur Analyse notwendigen Daten verbleiben im RDBMS und brauchen nicht zusätzlich geladen werden. Die zur Abfrage notwendigen SQL-Anweisungen werden zur Laufzeit erzeugt und auf die Transaktions- oder Aggregationstabellen abgesetzt. Relationales OLAP kann dadurch eine größere Anzahl von Dimensionen und Ausprägungen verarbeiten, kann dem Benutzer aber keine verlässliche Aussage über die zu erwartenden Antwortzeiten liefern.

- **HOLAP – Hybrides OLAP:** Das Hybride OLAP nutzt sowohl die Möglichkeit der multidimensionalen als auch der relationalen Speicherform. Hierdurch können die Vorteile beider Formen miteinander kombiniert werden, allerdings wird auch die Komplexität des Gesamtsystems erhöht.

2.1.3. Gründe für Data Warehousing

Betrachtet man die Eigenschaften von OLTP-Anwendungen und Analysewerkzeugen wie OLAP-Software, so wird schnell klar, dass beide nicht auf derselben physischen Datenbasis ausgeführt werden sollten. Operationale Datenbanken sind im Hinblick auf schnelle Transaktionsverarbeitung mit begrenzten Datenmengen optimiert. Die Datenbestände eines Unternehmens sind meist auf unterschiedlichen Datenbanken (oft von unterschiedlichen Herstellern) verteilt und sind bei größeren Unternehmen meist an die funktionellen Einheiten angegliedert [Gar98].

Demgegenüber benötigen Analyseanwendungen meist rein lesenden Zugriff auf große Datenmengen [Poe98], wobei die Informationen aus allen Unternehmensbereichen in einem einheitlichen Format vorliegen sollten. Bei vielen Auswertungen mit einem zeitlichen Bezug

werden historische Daten benötigt. Außerdem sind Analyseanfragen sehr komplex, so dass ihre Ausführung die Leistungsfähigkeit von OLTP-Anwendungen beeinträchtigen könnte.
Die Eigenschaften von operativen und informativen Datenbeständen sind in der folgenden Tabelle aufgeführt:

Charakteristika	Operative Datenbanken	Informative Datenbanken
Transaktionsvolumen	hohes Volumen	niedriges bis hohes Volumen
Antwortzeit	sehr schnell	normal
Update	hohe Frequenz, permanent	niedrige Frequenz
Betrachtungsperiode	aktuelle Periode	Vergangenheit bis Zukunft
Umfang	anwendungsintern	anwendungsübergreifend
Aktivitäten	operativ, detailliert	analytisch, taktisch
Abfragen	vorhersehbar, periodisch	unvorhersehbar, ad-hoc
Niveau der Daten	detailliert	aggregiert, aufbereitet
Verarbeitungseinheit	Datensatz (record), eindimensional	Matrizen (array), multi-dimensional, sachbezogen

Tabelle 2.1: Charakteristika von operativen und informativen Datenbanken [Kra98]

Aus den unterschiedlichen Eigenschaften von operativen und informativen Datenbanken können die Gründe abgeleitet werden, die zur Entwicklung des Data Warehouse-Konzeptes führten. Mit diesem Begriff werden Architekturansätze, Vorgehen bei der Projektarbeit und organisatorische Aspekte bezeichnet, mit denen eine unternehmensweite Datenbasis zur Beantwortung von betrieblichen Fragestellungen erreicht werden kann, was im nächsten Abschnitt einer genaueren Betrachtung unterzogen wird.

2.2. Data Warehouse

In einem ersten Schritt sollen zunächst die Anforderungen an ein Data Warehouse vorgestellt werden. Anschließend wird die typische Architektur einer Data Warehouse-Umgebung kurz erläutert.

2.2.1. Anforderungen an ein Data Warehouse

Im vorigen Kapitel konnte als Fazit festgehalten werden, dass die Daten in den operationalen Systemen in einer Form vorliegen, die den Anforderungen von Informationssystemen für schnelle und unkomplizierte Datenanalysen nicht entspricht. H.W. Inmon charakterisierte diese Anforderungen an die Datensammlung wie folgt [Inm96]:

- *Themenorientierung*: Während der klassische Datenbankentwurf eher die Funktionsorientierung der Transaktionen in den Vordergrund stellt, liegt die Betonung bei einer themenorientierten Darstellung auf eine Gliederung der Datensammlung nach den Kernbereichen und strategischen Zielen eines Unternehmens, z.B. Kunden, Produkte oder Umsätze. Operative Daten, die lediglich für die Prozessdurchführung wichtig sind und

nicht der Entscheidungsunterstützung dienen, finden in ein Data Warehouse keinen Eingang.

- *Vereinheitlichung*: Aufgrund des langjährigen Wachstums der operationalen Systeme und der Heterogenität von Rechnerarchitekturen in vielen Unternehmen sind Datenredundanzen und Inkonsistenzen kaum vermeidbar. Bei der Zusammenführung der in einem Unternehmen vorhandenen, vielfältigen Datenquellen müssen diese beseitigt werden. Diese Vereinheitlichung kann verschiedene Formen annehmen und bezieht sich häufig auf Namensgebung, Bemaßung und Kodierung. Das Ziel ist ein konsistenter Datenbestand, der sich stimmig und akzeptabel präsentiert.

- *Zeitraumbezug*: Die zeitpunktgenaue Betrachtung von Daten, wie sie in operationalen Systemen vorgenommen wird, ist für die Managementunterstützung von untergeordnetem Interesse. Vielmehr werden Daten und Informationen benötigt, welche die Entwicklung eines Unternehmens über längere und mittlere Zeiträume (Wochen, Monats- oder Jahresbetrachtungen) repräsentieren und zur Erkennung und Untersuchung von Trends herangezogen werden können.

- *Beständigkeit*: Veränderungen einmal eingebrachter Daten werden nur im geringem Umfang vorgenommen, Aktualisierungen ergänzen stets die Datenbasis. Für Auswertungen und Abfragen erfolgen also ausschließlich lesende Zugriffe. Die beständige Bevorratung von Daten über lange Zeiträume hinweg erfordert durchdachte, anwendungsgerechte Aggregationsverfahren und optimierte Speichertechniken, um den Umfang des zu speichernden Datenmaterials in erträgliche Grenzen zu halten.

Die inhaltliche Ausrichtung des Data Warehouse lässt sich damit unter Aufgabengesichtspunkten wie folgt zusammenfassen: Ein Data Warehouse hat die Aufgabe, themenorientierte und integrierte (im Sinne von vereinheitlichte) Informationen über lange Zeiträume und mit Zeitbezug aus unterschiedlichen Quellen periodisch zu sammeln. Diese Informationen werden zur Unterstützung von Entscheidern nutzungsbezogen aufbereitet und bedarfsgerecht zur Verfügung gestellt (siehe Cha98]). Von Stephen R. Gardener stammt eine Definition, die eine eher prozessorientierte Sichtweise in den Vordergrund stellt [Gar98]: *„Data warehousing is a process, not a product, for assembling and managing data from various sources for the purpose of gaining a single detailed view of part or all of a business“*. Zu den organisatorischen Herausforderungen zählt die Zusammenführung der im Laufe der Zeit gesammelten Datenbestände. Zu den technischen Herausforderungen gehört deren Nutzung zu zeitnahen Analysen und die Verwaltung während des Betriebes.

2.2.2. Architektur einer Data Warehouse Umgebung

Die Architektur eines Data Warehouse lässt sich, wie in Abbildung 2.2 dargestellt, in Form einer Schichtenarchitektur übersichtlich und verständlich darstellen.

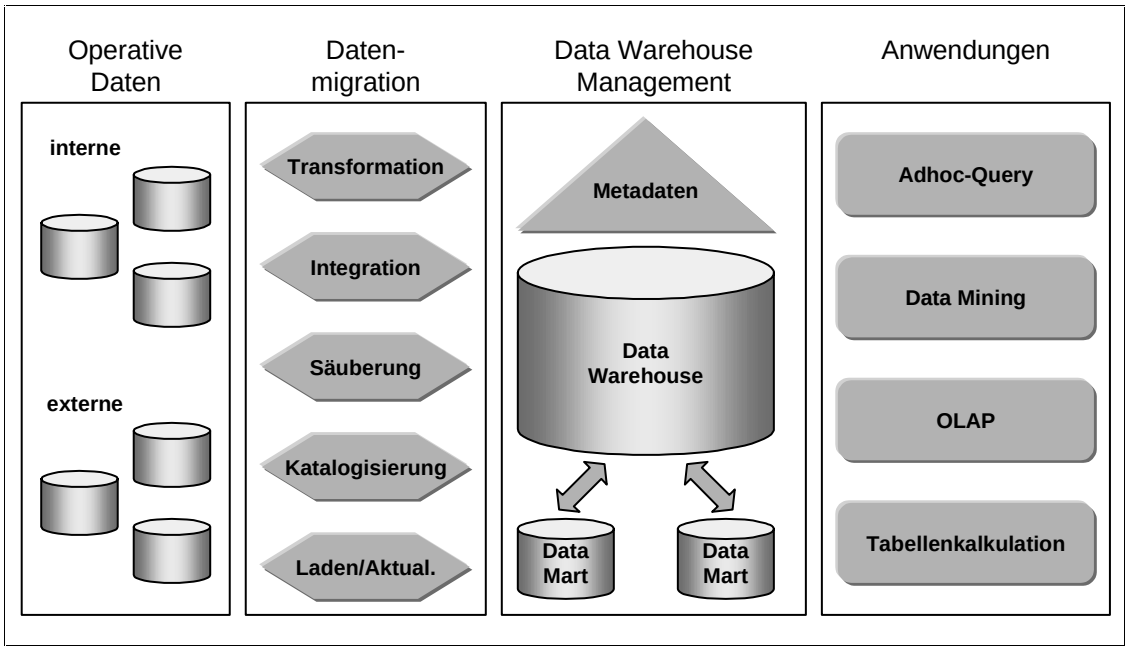


Abbildung 2.2: Grobaufbau eines Data Warehouse-Systems (nach [Bre98] und [Poe98])

Schicht 1: Operative Daten

Die operativen Datenhaltungssysteme bilden die Quellen, aus denen das Data Warehouse gespeist wird. Häufig liegen die Daten in den unterschiedlichsten Quellsystemen, angefangen von einfachen Dateisystemen über hierarchische und relationale Datenbanksysteme bis hin zu objekt-orientierten Datenbanksystemen. Für viele Analysen und Auswertungen werden auch Daten aus externen Quellen benötigt, beispielsweise demografische Daten. Diese sind mittlerweile in einer großen Anzahl aus überaus heterogenen Datenquellen verfügbar, z.B. politische Informationsdienste oder Marktforschungsinstitute..

Schicht 2: Datenmigration

Schicht 2 ist verantwortlich für die Überführung der Daten aus den operativen Systemen in den integrierten Datenbestand des Data Warehouse. Die dabei zu lösenden Probleme lassen sich grob fünf Kategorien zuordnen [Bre98]:

- *Transformation*: In einem ersten Schritt müssen die Unterschiede in den Datenstrukturen der Ausgangsdaten überbrückt werden. Dies bedarf zunächst der Überführung der unterschiedlichen Schemata (z.B. Dateien, hierarchische und relationale Schemata) in ein einheitliches Datenmodell.

- *Integration*: Für eine konsistente Datenhaltung im Data Warehouse muss eine Struktur- und Formatvereinheitlichung vorgenommen werden. So werden in einzelnen Datenbeständen die Datenfelder identischer Daten unterschiedlich benannt (Synonyme) oder unterschiedliche Datenfelder haben die gleiche Bezeichnung (Homonyme). Diese Unterschiede müssen für eine konsistente Datenhaltung bereinigt werden. Auch die unterschiedlichen Datenformate wie z.B. Kodierungen oder Maßeinheiten müssen angepasst werden, zumindest muss eine Umrechnungstabelle hinterlegt werden.

- *Säuberung*: Fehler in den operativen Daten sind unvermeidlich. Die Säuberung enthält daher Maßnahmen, um beispielsweise fehlende Werte durch Abgleich mit anderen Datenquellen oder durch entsprechende Default-Werte zu ergänzen oder um mittels übergreifende Integritätsregeln Datenfehler zu erkennen und zu beheben.

- *Katalogisierung*: Beim Zusammenführen müssen zu den operativen Daten Beschreibungen erzeugt werden. Dies führt zum Aufbau eines Meta-Datenbanksystems, in dem Informationen über alle Data Warehouse-Komponenten gehalten werden. Man spricht hier auch von Daten über Daten [Poe98]. Metadaten können als Landkarte gesehen werden, in der festgehalten ist, wo man welche Objekte finden kann und welche Bedeutung diese haben [Gar98].

- *Laden und Aktualisieren der Daten*: Zusätzlich sind noch Strategien für das Zusammenführen und Aktualisieren der Daten vorzusehen. Aufgrund der Größe der Datenbasen stoßen die herkömmlichen sequentiellen Methoden schnell an ihre Grenzen. Die meisten Datenbankhersteller bieten daher parallele und inkrementelle Lademöglichkeiten an. Beim Laden werden unter Umständen noch weitere Tätigkeiten zur Verarbeitung durchgeführt, beispielsweise Indexe angelegt oder Aggregierungen vorgenommen. Für das Aktualisieren ist festzulegen, in welchen Abständen und in welchem Umfang dies geschehen soll, da in der Regel nur ein relativ kleines Zeitfenster, beispielsweise in der Nacht, zur Verfügung steht.

Schicht 3: Data Warehouse Management

Das Data Warehouse als Ergebnis der Zusammenführung umfasst die gesamten für den Geschäfts- bzw. Entscheidungsprozess notwendigen Informationen über alle relevanten Bereiche, beispielsweise über Kunden, Produkte, Verkäufe und Personal. Diese werden in der Regel in einem relationalen Datenbanksystem verwaltet, ergänzt um die meist durch das gleiche Datenbanksystem verwalteten Metadaten. Um zu einer besseren Übersicht über die Daten zu gelangen, wird die Datenbasis eines Data Warehouse meist in kleinere Einheiten, den sogenannten Data Marts, aufgeteilt. Ein Data Mart kann beispielsweise die Kopie aller relevanten Daten einer Region, einer bestimmten Produktgruppe oder eines Zeitausschnitts beinhalten, zugeschnitten auf eine spezielle Gruppe von Endanwendern [Hol99].

Schicht 4: Aufsetzende Analyseanwendungen

Die vierte Schicht bilden die beim Benutzer angesiedelten Analyseanwendungen und Informationssysteme. Diese greifen über verschieden Protokolle und Netze auf die Daten des Data Warehouse zu. Die Palette reicht dabei von einfachen Visualisierungsverfahren und Tabellenkalkulation bis hin zu OLAP-Anwendungen und Data Mining-Systemen.

2.3. Knowledge Discovery in Databases

In diesem Abschnitt soll zunächst eine Abgrenzung des Begriffes Knowledge Discovery in Databases (abgekürzt KDD) zu dem OLAP-Begriff vorgenommen werden. Aufbauend auf einer Begriffsdefinition wird der Prozess des Knowledge Discovery in Databases dargestellt und erläutert.

2.3.1. Abgrenzung zu OLAP

Bei den eher traditionellen Methoden der Datenanalyse wie Reportgeneratoren aber auch bei OLAP-Systemen muss die zu prüfende Hypothese vom Anwender vorgegeben werden. Ein Experte wählt die geeignete Methode aus, mit der diese Hypothese geprüft wird, und bereitet sie dann so auf, dass der Anwender wiederum die Ergebnisse interpretieren kann. Dabei besteht jedoch die Gefahr, dass man sich bei der Suche nach neuen Zusammenhängen auf eher subjektive Einschätzungen verlässt. Ein großes Problem stellt auch die Anfrageformulierung dar [Fay98b]: Will man beispielsweise in dem Datenbestand einer Bank diejenigen Datensätze filtern, die auf einen Kreditkartenbetrug hinweisen, so lässt sich eine solche Anfrage kaum in SQL formulieren.
Im Gegensatz hierzu verfolgt die Forschungsrichtung Knowledge Discovery in Databases dieselben Absichten selbständig, d.h. ohne zusätzliche Interaktionen mit dem Benutzer. Dabei werden Hypothesen über mögliche Zusammenhänge, Muster oder Trends automatisch generiert und diese Hypothesen anhand von Daten überprüft. Aus der Vielzahl möglicher Hypothesen werden die gültigen als Ergebnis zurückgegeben. In Abbildung 2.3 werden diese Unterschiede noch einmal verdeutlicht.

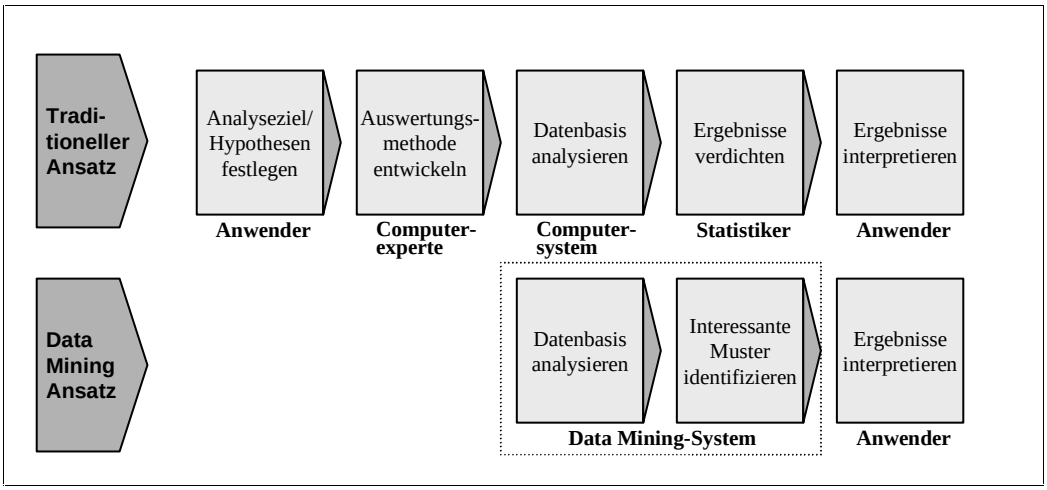


Abbildung 2.3: Analyseansätze im Überblick [Küp98]

Knowledge Discovery in Databases ist eine interdisziplinäre Forschungsrichtung, zu deren Einstehung und Entwicklung insbesondere die Forschungsgebiete Maschinelles Lernen, Mustererkennung, Datenbanksysteme und Statistische Datenanalyse beigetragen haben [Düs98]. Seit 1989 wurden jährlich von der American Association for Artificial Intelligence (AAAI) Workshops zu diesem Thema organisiert. Seit 1995 findet statt des Workshops eine eigene Konferenz statt, was verdeutlicht, dass dieses Thema immer mehr Interessenten findet (Informationen hierzu finden sich unter http://research.microsoft.com/datamine/kdd99).

2.3.2. Definition von Knowledge Discovery in Databases und Data Mining

Aus der ursprünglichen Begriffsvielfalt, die dieses Forschungsgebiet prägten, haben sich im Laufe der Zeit die beiden Begriffe *Data Mining* und *Knowledge Discovery in Databases (KDD)* durchgesetzt. Der Begriff Data Mining wurde vorwiegend von Statistikern, Datenanalysten und der MIS-Gemeinde[2], KDD von den Vertretern für Künstliche Intelligenz und dem Maschinelles Lernen verwendet [Fay96a]. Fayyad schlug 1996 eine heute weithin anerkannte Definition vor, in der die beiden Begriffe getrennt wurden (siehe [Fay96a] Seite 6 und 9):

> *Knowledge Discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.*
>
> *Data Mining is a step in the KDD process...*

Dieser in der Literatur auch als „Knowledge Extraction“ oder „Data Analysis“ bezeichnete Prozess ist darauf ausgerichtet, in umfangreichen Datenbeständen implizit vorhandenes Wissen zu entdecken und explizit zu machen. Der Prozessbegriff impliziert, dass mehrere iterative Schritte wie Datenvorverarbeitung, Mustererkennung und Evaluation notwendig sind um verwertbare Ergebnisse zu erhalten, was im nächsten Abschnitt noch einmal verdeutlicht wird. Data Mining wird als ein Schritt in diesem Prozess gesehen der die Methoden und Algorithmen umfasst, die zum entdecken der Muster Verwendung finden. Die gefundenen Muster müssen für einen möglichst großen Teil der Daten Geltung haben und bislang unbekannte, potentiell nützliche und leicht verständliche Zusammenhänge in den Daten zum Ausdruck bringen. Aus den ermittelten Beziehungsmustern wird schließlich durch Interpretation und Evaluation explizites Wissen abgeleitet.

2.3.3. Der KDD-Prozess

Abbildung 2.4 zeigt eine Darstellung der Schritte, die bei einem KDD-Prozess iterativ durchlaufen werden.

[2] MIS steht für Management-Informationssysteme

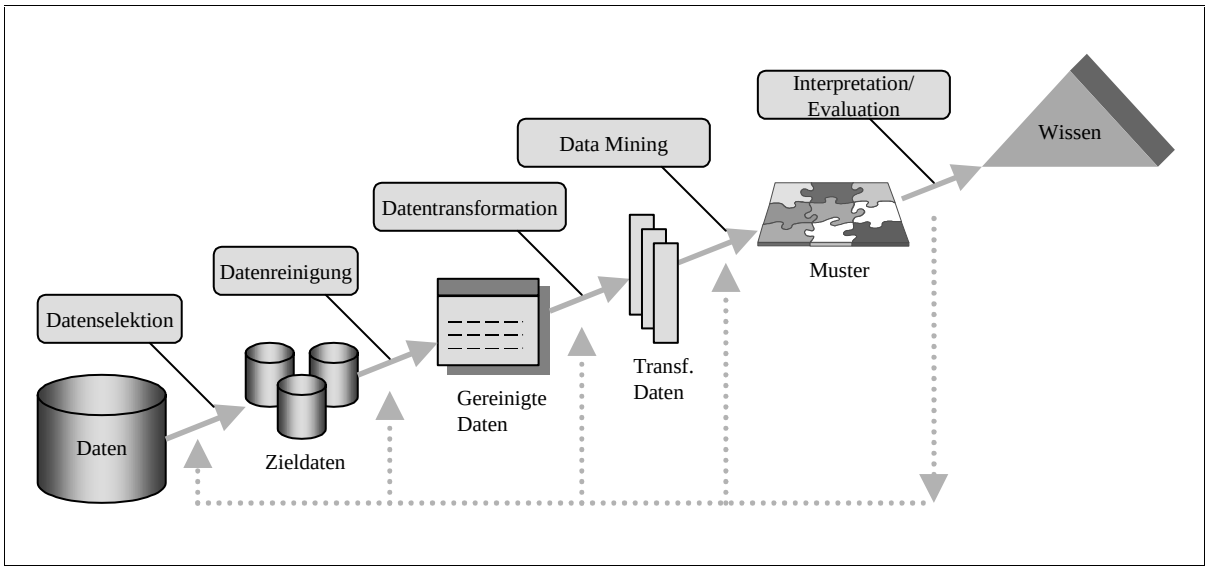


Abbildung 2.4: Der KDD-Prozess [Fay98a]

Der KDD-Prozess beginnt mit einer Vorstufe, in welcher relevantes und bereits vorhandenes Wissen über den gewünschten Anwendungsbereich gesammelt sowie die Zielsetzung der Anwendung festgelegt wird. Daran schließen sich folgende Schritte an (siehe [Fay96a], [Fay96b], [Bar98], [Ben99]):

- **Datenselektion:** Eine relevante Teilmenge der Daten wird ausgewählt, die als Basis für das weitere Vorgehen dient. Data Mining-Systeme können Sinnvollerweise mit vorhandenen Data Warehouse-Systemen gekoppelt werden, sind aber nicht auf den ausschließlichen Zugriff auf ein Data Warehouse beschränkt. So können die Daten auch direkt aus den operativen Datenbanken oder aus externen Datenquellen beschafft werden. In dieser Phase wird geprüft, welche Daten notwendig und verfügbar sind, um das gesetzte Ziel zu erreichen Es wird ebenfalls festgelegt, welcher Zeitrahmen betrachtet wird. Dadurch ergibt sich auch das Datenvolumen, das zur weiteren Verarbeitung ansteht.

- **Datenreinigung und Vorverarbeitung:** In dieser Phase wird die Datenqualität des selektierten Datenpools untersucht. Aufgrund technischer oder menschlicher Fehler können die Daten operativer Systeme fehlerhafte Elemente enthalten. In der Praxis wird damit gerechnet, dass 1-5% der Felder des Datenbestandes falsche Angaben aufweisen [Red98]. Eine häufige, leicht zu identifizierende Fehlerart besteht in fehlenden Werten. Zur Behandlung von fehlenden Werten stehen unterschiedliche Techniken zur Verfügung [Ber97]. So kann der Datensatz, der fehlende Werte aufweist, aus dem Datenbestand gelöscht werden. Dies kann jedoch zu einer Verfälschung der Analyseergebnisse führen. Eine weitere Möglichkeit besteht in der Schätzung oder nachträglichen Erhebung des fehlenden Wertes. Während die Schätzung bei quantitativen Daten über die Bildung des Mittelwertes technisch einfach zu realisieren ist, kann die nachträgliche manuelle Erhebung zu einem hohen Aufwand führen. Eine weitere potentielle Fehlerart wird durch Ausreißer hervorgerufen. Dabei handelt es sich um Wertausprägungen, die deutlich vom Niveau der übrigen Werte abweichen. Bei diesen Ausprägungen kann es sich um korrekt erfasste Daten handeln, die damit Eingang in die Analyse finden, oder aber um falsche Angaben, die nicht berücksichtigt werden dürfen und daher aus dem Datenbestand zu

löschen sind. Die Erkenntnisse, die der Nutzer in dieser Phase der Vorverarbeitung über den Datenbestand gewinnt, kann Hinweise auf die Verbesserung der Datenqualität der operativen Systeme geben.

- **Datentransformation:** In Abhängigkeit von der Art der gestellten Aufgabe wird versucht, eine geeignete und möglichst einfache Darstellung der Daten zu erreichen. Da viele Data Mining-Techniken nur mit einer einzigen Tabelle arbeiten können, werden in dieser Phase die Inhalte von mehreren Tabellen zu einer einzigen zusammengefasst. Dabei findet eventuell auch eine Voraggregierung der Daten statt, indem beispielsweise mehrere Bestellungen eines Kunden innerhalb eines Monats zu einem Datensatz zusammengefasst werden, der die Anzahl der Bestellungen und deren durchschnittlichen Wert enthält. Dieser Schritt kann schon deshalb notwendig sein, um die Datenmenge einzuschränken, die zur Verarbeitung ansteht. Eine Aufteilung in Trainings- oder Testdaten, welche von einigen Data Mining-Techniken benötigt werden, wird ebenfalls in dieser Phase vorgenommen.

- **Data Mining:** In dieser Phase erfolgt die Methodenauswahl und deren Einsatz zur Identifikation von Mustern auf der Basis des vorbereiteten Datenbestandes. In einem ersten Schritt wird zunächst entschieden, welche grundlegende Data Mining-Operation (Klassifizierung, Segmentierung o.ä.) zum Einsatz kommen soll. Daran schließt sich die Auswahl einer geeigneten Data Mining-Technik (z.B. Entscheidungsbäume, Neuronale Netze, ...) an. Mit dieser Technik wird nach der Angabe geeigneter Parameter in den Daten nach interessanten Mustern gesucht. Zwar dienen die der Mustererkennung vorgelagerten Phasen „nur“ der Datenvorbereitung, doch verbrauchen diese Aktivitäten einen erhebliche Teil der Gesamtressourcen des Data Mining-Prozesses. In der Praxis kann nach Expertenschätzungen die Datenreinigung und Vorverarbeitung bis zu 80% des Zeitaufwandes und Kosten des Data Mining betragen [Alp98].

- **Interpretation und Evaluation**: Hier erfolgt die Interpretation und Visualisierung der entdeckten Muster. Überflüssige oder irrelevante Muster werden aussortiert und die nützlichen in einer für den Benutzer verständlichen Weise dargestellt. Eine Evaluation der Resultate erfolgt meist in Zusammenarbeit mit den einzelnen Fachbereichen, wobei oft ein Vergleich mit den Testdaten vorgenommen wird. Typischerweise erfolgt in dieser Phase ein Rücksprung in eine der vorherigen Phasen. So ist meist eine Anpassung der Parameter notwendig oder die Auswahl einer anderen Data Mining-Technik kann wünschenswert sein. Es kann auch erforderlich sein, zum Schritt der Datenselektion zurückzukehren, falls festgestellt wird, dass sich die gewünschten Ergebnisse nicht mit den gewählten Daten erreichen lassen.

Am Ende des Prozesses soll das entdeckte Wissen soweit wie möglich in den Arbeitsablauf integriert werden. Dazu gehört die Dokumentation der Ergebnisse und die Kommunikation mit den betroffenen oder interessierten Benutzern. Auch ein Abgleich im Hinblick auf Widersprüche mit bisherigem oder erwartetem Wissen sollte vorgenommen werden.

2.4. Data Mining-Methoden

Nachdem im letzen Abschnitt die Phasen des Data Mining-Prozesses einer genaueren Betrachtung unterzogen wurden, sollen in diesem Abschnitt zunächst die Teilziele vorgestellt werden, die mit Data Mining verfolgt werden können. Daran anschließend soll anhand von vier grundlegenden Operationen ein Einblick in die Arbeitsweise eine Data Mining-Systems gegeben werden.

2.4.1. Aufgaben von Data Mining

Wie schon in der Beschreibung des KDD-Prozesses zu sehen, liegen die Aufgaben eines Data Mining-Systems in dem weitgehend automatischen Auffinden von bisher unbekannten Zusammenhängen in großen Datenmengen. Darauf aufbauend könne zwei Teilziele definiert werden: Vorhersage und Beschreibung [Fay96a].

Bei der Vorhersage werden ausgehend von den Rohdaten neue Muster oder Regeln extrahiert. Beispielsweise wird in Kundendaten einer Bank die Kreditwürdigkeit eines Kunden als zu vorhersagende Variable definiert. Diese Kreditwürdigkeit könnte in Form einer logischen Regel aus den anderen Kundenattributen erklärt werden: „Kreditwürdigkeit ist hoch, wenn Haushaltseinkommen größer als x DM pro Monat und keine Kontosperre im letzten Quartal und ...“. Aufgrund dieser Regeln kann die Kreditwürdigkeit von neuen Kunden anhand ihrer persönlichen Merkmale eingeschätzt werden.

Bei der Beschreibung werden eher verständliche Muster oder Abhängigkeiten in den Daten gesucht. Ein typisches Beispiel hierfür ist eine Warenkorbanalyse, wobei Gruppen von Produkten gefunden werden, die häufig gemeinsam gekauft werden. Diese Beziehungen zwischen den einzelnen Produkten können durch Wenn/Dann-Regeln beschrieben werden.

Die Unterscheidung zwischen beschreibenden und vorhersagenden Modellen ist allerdings nicht sehr ausgeprägt, da vorhersagende Modelle einerseits einen beschreibenden Charakter haben, andererseits beschreibende Modelle ebenfalls zu Vorhersage genutzt werden können.

2.4.2. Überblick über die Data Mining-Methoden

Im folgenden sollen vier grundlegende Operationen vorgestellt werden, die mit einem Data Mining-System erreicht werden können (siehe [Bar98], [Fay96a], [Fay96b], [Fay98b], [Kra98], [Nak98], [Man97]). Damit soll ein Einblick in die Arbeitsweise eines solchen Systems gegeben werden. Auch eine ungefähre Abschätzung der Art der produzierten Resultate ist damit möglich.

Segmentierung:

Die Segmentierung zielt auf die Aufspaltung der Daten in interessante und sinnvolle Teilmengen oder Klassen, sogenannte Segmente. Auf der Basis von Distanzmaßen soll dabei innerhalb eines Segments eine höchstmögliche Homogenität, zwischen den Segmenten eine größtmögliche Heterogenität erreicht werden. Der Grad der Homogenität und die Anzahl der Segmente kann vom Benutzer über Parameter bestimmt werden.

Die Segmentierung wird häufig zur Einteilung von Kunden in Zielgruppen verwendet, um eine möglichst zielgruppenorientierte Marketing-Aktivität zu realisieren. Dazu werden die Kundendaten und Transaktionen über einen gewissen Zeitraum analysiert, wobei ähnliche Verhaltensmuster identifiziert werden.

Klassifikation:
Bei dieser sehr häufig verwendeten Methode wird ein Datensatz in eine vordefinierte Klasse oder Gruppe eingeordnet. Dabei werden historische Daten verwendet, um ein Klassifikationsmodell zu entwickeln, das hilft, die bestehenden Daten und deren Verhalten in der Zukunft zu verstehen. Eine typische Anwendung ist beispielsweise die Einordnung von Versicherungsnehmern in Risikoklassen bei einer Autoversicherung.
Der Unterschied zwischen der Klassifikation und der Segmentierung liegt darin, dass bei der Klassifizierung vordefinierte Klassen verwendet werden, während bei der Segmentierung diese erst generiert werden. Aus diesem Grund können mit dieser Methode auch keine Klassen entdeckt werden, die zuvor noch unbekannt oder nicht definiert wurden.

Modellierung / Analyse von Abhängigkeiten:
Mit der Abhängigkeitsanalyse werden Modelle generiert, die Beziehungen und Abhängigkeiten zwischen Variablen beschreiben. Dabei können zwei Ebenen unterschieden werden: Auf der strukturellen Ebene werden die Zusammenhänge der Abhängigkeiten bestimmt, auf der quantitativen Ebene die (numerische) Stärke der Abhängigkeit. Bei einer Warenkorbanalyse kann ein Data Mining-System beispielsweise auf der strukturellen Ebene feststellen, dass zwischen dem Kauf von Brot und dem Kauf von Butter eine Abhängigkeit besteht. Auf der quantitativen Ebene kann die Wichtigkeit der Aussage festgestellt werden, indem die Abhängigkeit beispielsweise mit einer Wahrscheinlichkeit von 80% eintritt.
Ebenfalls in den Bereich der Assoziationen fallen sequentielle Assoziationen bzw. sequentielle Muster. Sequentielle Assoziationen setzen Datenhistorien voraus und zielen darauf ab, dass innerhalb eines Zeitraumes Beziehungen zwischen dem Auftreten verschiedener Ereignisse eintreten. Ein Einsatzgebiet ist das Aufdecken von Betrugsfällen im Kreditkartenbereich. Es wurde festgestellt, dass Kreditkartenbetrug oft nach regelmäßigen Schemata abläuft, was sich in der zeitlichen Abfolge bestimmter Transaktionen ausdrückt.

Änderungs- und Abweichungserkennung:
Die Änderungs- und Abweichungserkennung beschäftigt sich mit Objekten, die sich keinem Muster eindeutig zuordnen lassen. Bei diesen „Ausreißern“ kann es sich um fehlerfreie, interessante Merkmalsausprägungen handeln oder aber um fehlerhafte Daten, die keine realen Sachverhalte beschreiben. Die Zielsetzung der Abweichungsanalyse besteht darin, die Ursachen für die untypischen Merkmalsausprägungen des Ausreißers aufzudecken. Wird ein Ausreißer im Datenbestand identifiziert, so werden alle assoziierten Datenbestände durchsucht, um die Einflussfaktoren zu erklären, die zu einer abweichenden Merkmalsausprägung geführt haben. Auch signifikante Änderungen in Bezug auf vorher definierte oder gemessene Werte sollen erkannt werden. Dies wird vor allem im Bereich des Controlling verwendet, beispielsweise für Soll-Ist-Vergleiche.

Bei der Zeitreihenanalyse, einer Variante der Änderungs- und Abweichungserkennung, werden Strukturen untersucht, die bezogen auf einen bestimmten Zeitraum ein ähnliches Verhalten aufweisen. Dies ist vor allem für die Bestimmung von Trends nützlich.

In der folgenden Abbildung werden die Beziehungen zwischen den Zielen von Data Mining und den Methoden verdeutlicht.

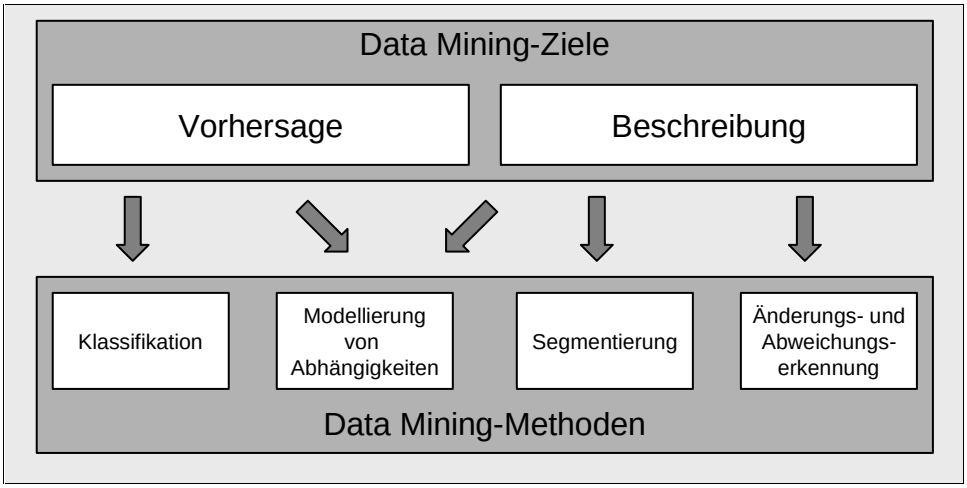


Abbildung 2.5: Verbindung zwischen Data Mining-Zielen und Data Mining-Methoden

2.5. Data Mining-Techniken

In diesem Abschnitt soll ein Überblick über die gängigsten Data Mining-Techniken gegeben werden. In Abbildung 2.6 ist eine Zuordnung dieser Techniken zu den weiter oben vorgestellten Methoden zu sehen.

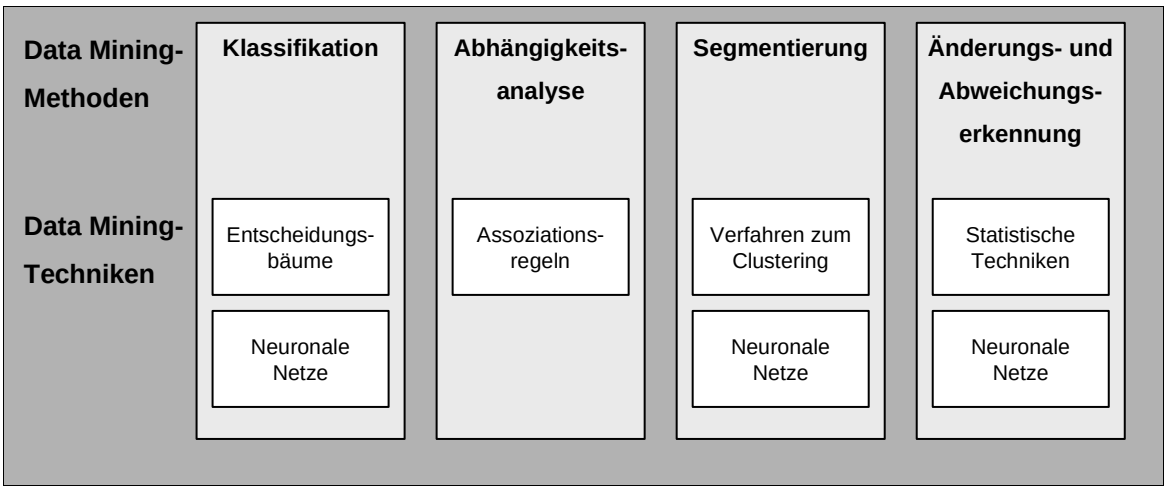


Abbildung 2.6: Zuordnung von Data Mining-Techniken zu Data Mining-Methoden

Einige Techniken können für mehr als eine der grundlegenden Methoden verwendet werden, wobei zur Lösung eines Problems oft einen Kombination von zwei oder mehreren Techniken verwendet wird. Darüber hinaus muss berücksichtigt werden, dass sich dieses

Forschungsgebiet ständig weiterentwickelt, so dass diese Darstellung keinen Anspruch auf Vollständigkeit erhebt.
Viele der auf dem Markt befindlichen Systeme unterstützen zumindest die grundlegenden Techniken, die im Folgenden kurz vorgestellt werden sollen (siehe [Bar98], [Fay98b], [Kra98], [Küp98], [Mar98], [Bac96], [Gri98], [Bor98], [Deg98])

2.5.1. Entscheidungsbäume

Entscheidungsbäume gehören zu den bei Data Mining-Systemen am häufigsten verwendeten Ansätzen, und werden meist zur Klassifikation verwendet. Die Menge der Datensätze wird dabei in einer baumähnlichen Struktur in Kategorien eingeordnet. Abbildung 2.7 zeigt einen solchen Entscheidungsbaum zur Vorhersage des ROI (Return of Investment). Der markierte Pfad zeigt eine Folge von Entscheidungen, die zu einem besseren Ergebnis führen. Der Algorithmus zum Aufbau eines Entscheidungsbaumes beginnt mit der Suche nach einer sinnvollen Unterteilung des Wurzelknotens. Dazu wird eine Menge von Trainingsdatensätzen herangezogen, in denen das abhängige Attribut (hier der ROI) bereits bekannt ist. Der Algorithmus sucht nun in den unabhängigen Attributen nach demjenigen, welches die Daten am besten in die gewünschten Kategorien unterteilt. Dieser Prozess wird für jeden entstehenden Baumknoten neu durchgeführt, bis kein passendes Attribut mehr zum Teilen gefunden werden kann.

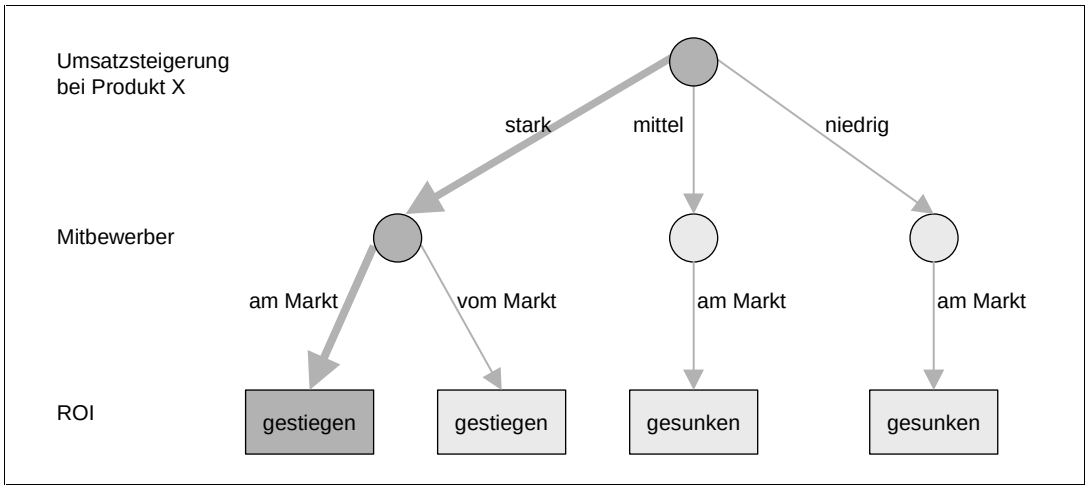


Abbildung 2.7: Beispiel eines Entscheidungsbaumes zur Vorhersage des ROI

Das Auswahlverfahren zur Bestimmung der Attribute ist für die Generierung des Baumes entscheidend. Häufig implementierte Algorithmen sind sogenannte CARTs (*classification and regression trees*) und CHAIDs *(chi-squared automatic interaction detection*):

- Bei einem CART-Algorithmus wird die Attributauswahl durch Maximierung des Informationsgehalts gesteuert. Dazu wird zu jedem Attribut ein Schwellwert gesucht, der eine optimale Trennung der Daten in Bezug auf die Klassifikation zulässt. Grundsätzlich lässt sich formulieren: Je höher der Informationsgehalt eines Attributs in bezug auf die Zielgröße, desto weiter oben im Entscheidungsbaum findet sich dieses. Eine Eigenart der CART-Bäume ist die Tatsache, dass durch die Teilung der Attribute durch einen festen

Schwellwert nur Binärbäume erzeugt werden können. Binärbäume sind dadurch gekennzeichnet, dass an jeder Verzweigung genau zwei Äste vorhanden sind.

- Bei CHAID-Algorithmen wird zur Attributauswahl der Chi-Quadrat-Unabhängigkeitstest verwendet. Dieser wird benutzt, um eine Aussage über die Abhängigkeit zweier Variablen zu treffen. Es wird dazu eine Kennzahl, der sogenannte „Chi-Quadrat-Abstand“, berechnet. Bei der Attributauswahl wird dann die Variable mit dem größten Abstand zur Zielgröße berücksichtigt. Im Gegensatz zum CART-Algorithmus können hier an jedem inneren Knoten des Baumes auch mehr als zwei Verzweigungen vorgenommen werden, wenn dadurch die Trennqualität erhöht wird. Die dadurch entstehenden Mehrfachbäume haben eine kompaktere Form der Darstellung und führen zumindest theoretisch zu besseren Ergebnissen. Allerdings ergeben sich bei der Verwendung von numerischen Daten bei großen Datenmengen automatisch viele unterschiedliche Ausprägungen, was zu einem schlechten Laufzeitverhalten führt. Deshalb müssen numerische Variablen in Kategorien eingeteilt werden, was einen erhöhten Aufwand bedeutet.

Mittels einer geeigneten grafischen Darstellung lassen sich Entscheidungsbäume leicht interpretieren. Die gefundenen Erkenntnisse können auch als logische Ausdrücke modelliert werden und lassen sich auf neue Datensätze anwenden. Bei vielen Problemen mit einer überschaubaren Komplexität lassen sich Entscheidungsbäume schnell und einfach generieren. Allerdings nehmen bei einem Anstieg der Dimensionalität die Bäume beträchtliche Ausmaße an, was die Interpretierbarkeit erschwert. Insbesondere in den tieferen Verzweigungen wird der Einfluss von zufälligen Elementen (fehlende Werte, Ausreißer in den Daten, ...) größer, was zu einer Übermodellierung des Entscheidungsbaumes führt. Zur Umgehung des Problems bietet sich an, nur eine bestimmte maximale Tiefe der Bäume zuzulassen oder eine Mindestanzahl der Objekte pro Knoten zu fordern.

2.5.2. Neuronale Netze

Neuronale Netze sind aus dem Wunsch heraus entstanden, das menschliche Gehirn mitsamt seiner Lernfähigkeit nachzubilden. In Analogie zu einem Neuron im Gehirn ist das Grundelement eines Neuronalen Netzes eine Verarbeitungselement, das mehrere gewichtete Eingänge, eine Aktivierungsfunktion und einen Ausgang besitzt. Die schematische Darstellung eines Neurons ist in Abbildung 2.8 zu sehen.
Die Lernfähigkeit besteht in der Anpassung der einzelnen Kantengewichte zwischen den einzelnen Neuronen. Dabei wird die Informationsverarbeitung in zwei Schritten durchgeführt:

1. Die Inputgrößen ($x_1, x_2, ..., x_n$) werden mit ($w_1, w_2, ..., w_n$) gewichtet und aufsummiert.

2. Der im 1. Schritt ermittelte Input führt mittels einer Aktivierungsfunktion zu einer Entscheidung, die an den Output angelegt wird.

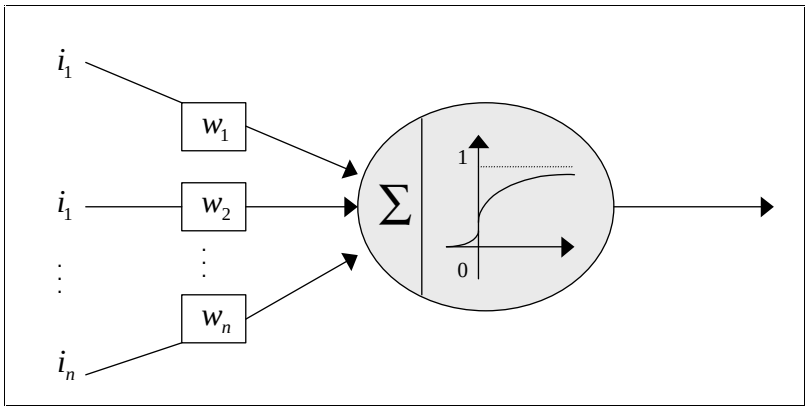


Abbildung 2.8: Schematische Darstellung eines Neurons

Bei den Aktivierungsfunktionen können im wesentlichen zwei Typen unterschieden werden:

- Bei der Klasse der Sigmoid-Funktionen wird das Neuron dann zu einer Reaktion veranlasst, wenn der Input einen gewissen Schwellwert überschreitet. Man sagt, das Neuron „feuert“. Wird der Schwellwert nicht überschritten, so feuert das Neuron nicht.

- Bei der Verwendung von Radialen Basis-Funktionen feuert das Neuron nur, wenn sich der Input in der Nähe eines bestimmten Schwerpunktes befindet.

Zur Bildung eines Neuronalen Netzes werden mehrere Verarbeitungseinheiten miteinander verbunden, üblicherweise in Ebenen, sogenannten *Layers*. Zwei Ebenen haben dabei eine Verbindung nach außen: Die Eingangsebene (*Input Layer*) und die Ausgangsebene (*Output Layer*). Die anderen Ebenen werden als versteckte Ebenen (*Hidden Layers*) bezeichnet. Zur Zeit werden neuronale Netze in drei Topologien eingeteilt:

1. Vorwärtsvermittelnde Netzwerke (*feed-forward*)

2. Partiell rekursive Netzwerke (*limited recurrent*)

3. Vollständig rekursive Netzwerke (*fully recurrent*)

Bei einem vorwärtsvermittelnden Netzwerk entspricht die Topologie einem azyklischen Graphen, das Netzwerk enthält also keine Zyklen. Demgegenüber sind rekursive Netze mit ihren Zyklen in der Lage, auch Zustände zu speichern.
Beim Betrieb eines neuronalen Netzes kann grundsätzlich zwischen überwachten (*supervised learning*) Lernen und unüberwachten Lernen (*unsupervised learning*) unterschieden werden:

- Überwachtes Lernen wird meist dazu verwendet, um Anwendungen zur Klassifizierung und Zeitreihenanalyse zu realisieren. Dazu werden in einer Lernphase mit Hilfe von Trainingsdaten die Verknüpfungen zwischen den Verarbeitungseinheiten angepasst, so dass bei einem vorgegebenem Eingangssignal ein bekanntes Ergebnis eintritt. Das Lernen des neuronalen Netzes wird durch Auswertung des Fehlers im Output und durch Anpassung der Netzparameter (Input-Gewicht und Schwellwerte bzw. Schwerpunkte der Aktivierungsfunktionen) derart bewerkstelligt, dass der Fehler beim nächsten Schritt kleiner wird. Diese Lernregel durch Fehlerrückrechnung wird *Backpropagation*-Lernregel

genannt. In einer zweiten Phase kann das trainierte neuronale Netz auf neue Daten angewandt werden.

- Unüberwachtes Lernen wird dann eingesetzt, wenn man eine Datenmenge untersuchen soll und die Frage, jedoch noch nicht die Antwort formulieren kann. Dies ist vor allem bei der Datensegmentierung der Fall, wo beispielsweise Fragestellungen der Form „Welche Kunden zeigen ein ähnliches Kaufverhalten?“ auftreten. Ein neuronales Netz versucht nun, Gemeinsamkeiten der präsentierten Eingangsmuster durch einen Ähnlichkeitsvergleich zu entdecken und seine Gewichtsstruktur danach auszurichten. Anhand dieser gefundenen Muster werden die Kunden in Zielgruppen segmentiert. Ein Beispiel für das unüberwachte Lernen sind die *Self Organizing Feature Maps* von Kohonen (siehe Abschnitt 2.5.4.).

Neuronale Netze bieten einen hohen Freiheitsgrad an Lernkapazität, da es sich einer sehr allgemeinen mathematischen Darstellung bedient und nicht an einschränkende Randbedingungen einer Wissensdarstellung in Form von scharfen Regeln gebunden ist. Neuronale Netze können sowohl nominal- als auch intervallskalierte Daten[3] verarbeiten und erreichen auch bei großen Datenmengen akzeptable Laufzeiten. Es kann auch mit unvollständigen Daten gearbeitet werden, was gerade für die Mustererkennung wichtig ist. Die hohe Flexibilität bezahlt man allerdings mit sehr aufwendigen und komplizierten Trainingsvorgängen. Ein weiterer Nachteil von neuronalen Netzen liegt in der fehlenden Erklärungskomponente. Es werden zwar Ergebnisse erzielt, jedoch kann der Weg, welcher zu dieser Lösung geführt hat, nicht erklärt werden. Zur Zeit bestehen auch noch keine allgemeinen Richtlinien für den Entwurf eines neuronalen Netzes, so dass sich der Modellierungsprozess (Anzahl der Schichten, Vergabe der Anfangsgewichte, ...) auf Erfahrungswerte des Entwicklers stützt.

2.5.3. Assoziationsregeln

Die algorithmische Umsetzung der Analyseverfahren zur Aufdeckung von Assoziationen basiert auf der Häufigkeitsbetrachtung von Attributkombinationen. Dazu werden zwei grundlegende Maße verwendet:

- Der Träger einer Attributmenge gibt an, wie prozentual häufig die Attribute gemeinsam innerhalb des gesamten Datenbestandes vorkommen.

$$\text{Träger (ProduktA, ProduktB)} = \frac{\text{Anzahl der Transaktionen, die A und B enthalten}}{\text{Anzahl aller Transaktionen}}$$

- Die Konfidenz einer Assoziationsregel A->B gibt prozentual an, wie oft bei Zutreffen von A auch tatsächlich B zutrifft.

$$\text{Konfidenz (ProduktA -> Produkt B)} = \frac{\text{Anzahl der Transaktionen, die A und B enthalten}}{\text{Anzahl aller Transaktionen, die A enthalten}}$$

[3] Eine Erläuterung des Skalenniveaus findet sich in Anhang E

Der Prozess zur Aufdeckung von Assoziationen lässt sich dabei in zwei Phasen gliedern [Agr93]:

- In einer ersten Phase wird die Datenbank nach Teilmengen, den sogenannten *Frequent Itemsets* durchsucht, deren Häufigkeit eine minimale untere Schranke überschreitet.

- In einer zweiten Phase werden aus diesen Teilmengen die eigentlichen Assoziationsregeln generiert. Auch dabei müssen für die einzelnen Regeln bestimmte Bedingungen erfüllt sein, beispielsweise dass die Konfidenz einer Assoziationsregel größer oder gleich einer bestimmten unteren Grenze ist.

Mit diesen Vorgaben von minimalen Trägern und Konfidenzen kann gesteuert werden, ab wann eine Assoziation als interessant angesehen wird. Denn sonst treten in einer umfangreichen Datenbasis fast beliebig viele Assoziationen auf. Dies hat auch einen positiven Einfluss auf die Rechenzeit des Verfahrens.
Im Gegensatz zu einem Entscheidungsbaum werden bei der Aufdeckung von Assoziationen hierarchische Struktur abgeleitet. Die Vorgabe eines Zielkriteriums ist nicht notwendig, vielmehr wird die Suche auf statistisch auffällige Muster beschränkt. Eine typische Anwendung dieser Methode ist die Warenkorbanalyse, bei der nach Produkten gesucht wird, die häufig gemeinsam gekauft werden. Daraus lassen sich Empfehlungen für die Regalanordnung oder Werbestrategien ableiten. Für Assoziationsregeln lassen sich aber auch leicht Anwendungen in anderen Bereichen finden, beispielsweise im Bankgewerbe. Dabei könnten die Finanzdienstleistungen identifiziert werden, die von einem Kunden gemeinsam in Anspruch genommen wurden.
Der Vorteil von Assoziationsregeln liegt darin, dass sie leicht implementiert werden können. Sie präsentieren häufig unerwartete Resultate und tragen so dazu bei, dass das Anwendungsgebiet genauer untersucht wird, und neue Fragestellungen auftauchen. Der Nachteil liegt in einem hohen Speicherplatzbedarf bei der Erzeugung der Trägermengen und der damit verbundenen langen Laufzeiten. Außerdem können die Ergebnisse sehr umfangreich und damit unüberschaubar werden.

2.5.4. Verfahren zum Clustering

Eines der am längsten eingesetzten Verfahren zur Segmentierung ist das *K-means Clustering* und wird im wesentlichen in zwei Schritten durchgeführt. Zunächst wird aus dem Datenbestand eine bestimmte Anzahl k von Objekten ausgewählt, die in der Folge als Clusterrepräsentanten anzusehen sind. Hierbei ist es wichtig, dass diese Auswahl möglichst repräsentativ ausfällt.
In einem zweiten Schritt wird nun jedes weitere Objekt demjenigen Cluster zugeordnet, zu dessen Repräsentanten die größte Ähnlichkeit besteht. In einem iterativen Prozess werden dann die Repräsentanten angepasst und bereits zugeordnete Objekte wiederum analysiert. Insgesamt entsteht ein wechselnder Prozess der Klassifizierung und der Klassenneudefinition.
Mittlerweile existieren viele Abwandlungen und Variationen des Verfahrens, wodurch eine hohe Anpassungsfähigkeit an das zu lösende Problem gegeben ist. Problematisch ist, dass die Anzahl der zu bestimmenden Cluster fest vorgegeben ist. Um eine optimale Anzahl zu

bestimmen müssen mehrere Durchläufe mit verschiedenen Werten für k durchgeführt und die Ergebnisse miteinander verglichen werden. Eine Erweiterung des Verfahrens ergibt sich durch die Anwendung von Fuzzy-Logik. Hierbei werden keine scharfen, sondern prozentuale Segmentszugehörigkeiten betrachtet. Das Verfahren heißt *Fuzzy-k-means*.

Ein weiteres, seit langem bekanntes Verfahren ist das hierarchische Clustering. Hierbei wird eine ganze Hierarchie möglicher Segmentierungen ermittelt. Es gibt zwei mögliche Vorgehensweisen:

- Jedes Objekt repräsentiert zunächst eine eigene Klasse. Es werden nun sukzessive Klassen miteinander verschmolzen, bis nur noch ein Cluster übrig bleibt. Dieses Vorgehen wird „agglomeratives hierarchisches Clustering“ genannt.

- Jedes Objekt gehört zunächst einem Cluster an. Es werden danach neue Teilcluster erzeugt, bis nur noch Cluster bestehend aus Einzelobjekten übrigbleiben. Bei diesem Vorgehen, welches „divises hierarchisches Clustering“ heißt, werden also bestehende Cluster in neue Cluster aufgeteilt.

Im Gegensatz zum k-means Clustering kann eine optimale Lösung aus dem Gesamtbild heraus gefunden werden, da durch die verschiedenen Hierarchiestufen mehr als eine Lösung präsentiert wird. Nachteilig ist, dass wegen der fehlenden Iteration anfangs getätigte Falschklassifizierungen nicht mehr rückgängig gemacht werden können.

Wie bereits in dem Abschnitt der Neuronalen Netze erwähnt, werden sogenannte *Kohonen-Netze* (Self-Organizing Feature Maps) zur Clusteranalyse verwendet [Koh95]. Kohonen-Netze bestehen in der Regel nur aus zwei Schichten, nämlich einer Inputschicht, die vollständig mit einer Outputschicht verknüpft ist. Im Unterschied zu den bisher betrachteten neuronalen Netzen besitzt hier ein Neuron weder Aktivierungsfunktion noch Schwellwert, es wird also nur durch die zugehörigen Gewichte (Code Book-Vektor) repräsentiert.
Beim Trainingsprozess wird zu einem Input das Output-Neuron mit dem höchsten Ähnlichkeitswert ermittelt. Dieses erhält dann das Recht, seine Gewichte an den Input anzupassen. Zusätzlich werden aber auch die Neuronen in der Nachbarschaft des Gewinner-Neurons adaptiert, wodurch zusammenhängende Bereiche ähnlich reagierender Neuronen in der Outputschicht entstehen.
Stärken und Schwächen der Kohonen-Netze sind ähnlich gelagert wie bei den herkömmlichen neuronalen Netzen. Positiv ist allerdings die einfachere Struktur zu bewerten, die die Anschaulichkeit des Verfahrens erhöht.

2.5.5. Statistische Verfahren

Wie schon bei den Entscheidungsbäumen und den Verfahren zum Clustering gesehen, können die statistischen Verfahren nicht vollständig vom Data Mining getrennt werden. Vielmehr sind sie ein elementarer Bestandteil, welcher zum einen klassische Analyseverfahren für Data Mining bereitstellt und zum anderen die Voraussetzungen für die Anwendbarkeit der modernen Verfahren schafft. Die Regressionsanalyse bildet dabei eines der flexibelsten und

am häufigsten eingesetzten statistischen Analyseverfahren. Sie dient zur Analyse von Beziehungen zwischen einer abhängigen Variablen (Regressand) und einer oder mehreren unabhängigen Variablen (Regressoren). Insbesondere wird sie eingesetzt, um Zusammenhänge zu erkennen und zu erklären und die Werte der abhängigen Variablen zu schätzen bzw. zu prognostizieren. Dabei wird eine lineare Beziehung zwischen den Variablen unterstellt, d.h. dass aus den empirischen Werten für Regressand und Regressor(en) eine lineare Beziehung errechnet wird, die durch den folgenden allgemeinen Ausdruck dargestellt werden kann:

$$y = ß_0 + ß_1 x_1 + ß_2 x_2 + \ldots + ß_n x_n$$

y = Regressand
$ß_0$ = konstantes Glied
$ß_n$ =Regressionskoeffizient des n-ten Regressors
x_n =n-ter Regressor

Generell liegt der Vorteil der Regressionsansätze darin, dass sie auf einer fundierten Theorie basieren, die dem Analytiker einen differenzierten Einblick in die Mechanismen des Modells ermöglicht. Hinzu kommt, dass die für die Modellierung verfügbaren numerischen Algorithmen sehr weit entwickelt sind und deshalb sehr schnell arbeiten.
Das Problem der linearen Verfahren ist, das sie komplexe Zusammenhänge oft zu einfach interpretieren, d.h. bestimmte Grundannahmen entsprechen nicht unbedingt der Realität. Zur Modellverbesserung können nichtlineare Verfahren genommen werden, jedoch erschweren die höhere Anzahl der Parameter und komplexere Funktionsformeln die Handhabung. Es sind mehr manuelle Eingriffe notwendig als bei den linearen Methoden, wodurch die Laufzeiten erhöht werden.
Auch Visualisierungstechniken spielen bei der Mustererkennung und der Entdeckung von Abweichungen eine große Rolle. Im wissenschaftlich-analytischen Bereich ist das Streudiagramm der am häufigsten verwendete grafische Darstellungstyp. In diesem werden die Merkmalsträger durch Punkte gekennzeichnet, deren Lage in Bezug zu einem rechtwinkligen Koordinatensystem beschrieben wird. Dieses Koordinatensystem wird aus der Ordinatenachse (Y-Achse) und der Abszissenachse (X-Achse) gebildet. Besitzt ein Merkmalsträger für das Zeilenmerkmal X die Merkmalsausprägung x und für das Spalten-Merkmal Y die Merkmalsausprägung y, so wird der ihm zugeordnete Punkt durch das Koordinatenpaar (x,y) gekennzeichnet. Allgemein verwendbar und in einer Vielzahl von Varianten verfügbar, kann mit einem Streudiagramm Art, Richtung und Intensität des Zusammenhangs zwischen zwei Merkmalen dargestellt werden. Darüber hinaus ist eine Identifikation ungewöhnlicher Beobachtungswerte sowie das Erkennen von Gruppenbildungen unter den Beobachtungswerten leicht möglich.
Streudiagramme können um zusätzliche Informationen ergänzt werden, die das Aufdecken von Mustern erleichtern. Vor allem bei einer großen Anzahl von Beobachtungswerten oder variierender Streuung ist es empfehlenswert, das Streudiagramm mit weiteren Informationen anzureichern, die Hinweise auf die zugrundeliegende Struktur geben können.

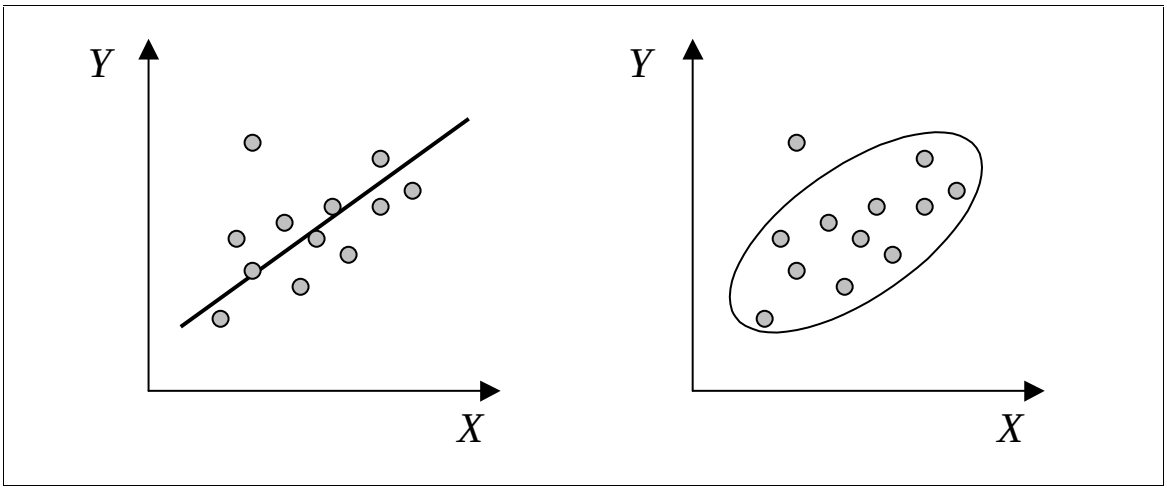


Abbildung 2.9: Streudiagramm-Varianten mit Regressionsgeraden und Dichte-Ellipse (siehe [Deg98])

Das einfachste Beispiel für derartige Verfahren ist das Einzeichnen einer Regressionsgeraden in die Punktwolke (siehe Abbildung 2.9), welches genau die Gerade ist, die von der Gesamtheit der Punkte den geringsten Abstand hat. Eine weitere nützliche Anwendung ist das Eintragen von sogenannten Dichte-Ellipsen (für 95% bzw. 99% des Datenmaterials) in das Streudiagramm.

Kapitel 3. Evaluation der Data Mining-Tools

Bei der Einführung einer neuen Technologie in einem Unternehmen, wie beispielsweise bei von Data Mining bei der Gehe Pharmahandel GmbH, stellt sich die Frage, welche Produkte mit welcher Funktionalität auf dem Markt erhältlich sind, die diese Technologie unterstützen und wie sich diese Produkte in das Unternehmen integrieren lassen. In dieser Diplomarbeit wurde im Rahmen einer Evaluation zunächst eine Sichtung des Marktes für Data Mining-Produkte durchgeführt. Daraufhin konnten insgesamt fünf Tools selektiert wurden, die von den Herstellern in Form von Testinstallationen zur Evaluation bereitgestellt wurden. Die Bewertung der Tools erfolgte anhand eines Kriterienkataloges, der neben allgemeinen Anforderungen an die Funktionalität der Tools auch die Erfahrungen berücksichtigt, die bei der praktischen Arbeit mit diesen Tools gemacht wurden. Aus der Gesamtbewertung konnte schließlich ein Vergleich der Tools erfolgen, wobei die Stärken und Schwächen der einzelnen Produkte verdeutlicht wurden.

3.1. Vorauswahl der Data Mining-Tools

Zunächst soll ein Überblick über die auf dem Markt befindlichen Produkte und Systeme gegeben werden. Darauf aufbauend werden die Kriterien betrachtet, die zur Vorauswahl der Data Mining-Tools benutzt wurden und es werden die Tools vorgestellt, die im Rahmen dieser Evaluation von den Herstellern zur Verfügung gestellt wurden. Anschließend wird die Testumgebung skizziert, in der die Tools zum Einsatz kamen.

3.1.1. Marktübersicht

In den letzten Jahren ist die Anzahl der existierenden Systeme für die Wissensentdeckung in Datenbanken ebenso wie die Anzahl der Veranstaltungen und Publikationen in diesem Bereich explosionsartig in die Höhe gestiegen. Die auf dem Markt verfügbaren Produkte können grob in drei Kategorien eingeteilt werden, was in der folgenden Abbildung zu sehen ist.

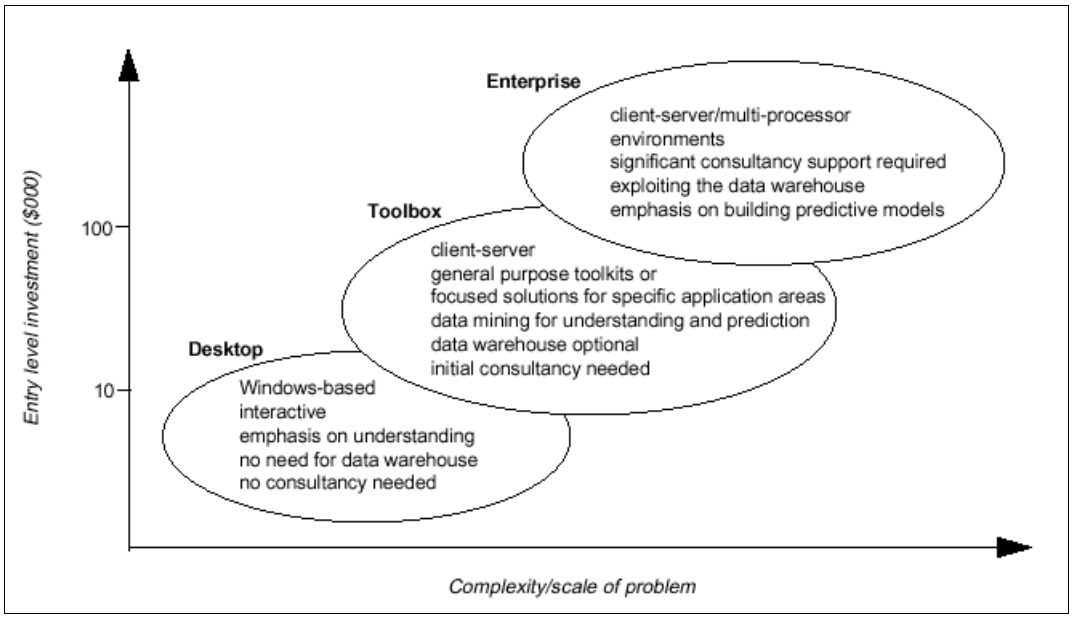


Abbildung 3.1: Einteilung der Data Mining-Tools [OV97]

Die Desktop- und Toolbox-Systeme zeichnen sich vor allem durch eine Spezialisierung auf bestimmte Anwendungsbereiche oder Data Mining-Techniken und durch eine begrenzte Anbindung an ein Data Warehouse aus. Die Enterprise-Systeme bieten eine breite Abdeckung von Methoden und Algorithmen und sind auf die Verarbeitung großer Datenmengen spezialisiert. Ein umfassender Überblick über die auf dem Markt befindlichen Produkte ist an dieser Stelle kaum möglich. Eine umfassende Zusammenstellung findet sich unter http://www.kdnuggets.com/sofware/index.html. Dort ist eine Vielzahl verschiedener Werkzeuge katalogisiert, kurz beschrieben und mit einer Kontaktadresse versehen.
Während in den letzten Jahren der Markt für Data Mining-Produkte ständig gewachsen ist, prophezeien führende Analysten wie die Gartner Group eine Stagnation und ab dem Jahr 2002 sogar eine Schrumpfung des Marktes [Dre98a]. Auch die Anzahl von Neuanbietern ist in den letzten zwei Jahren deutlich gesunken. Es wird befürchtet, dass sich in der nächsten Zukunft die Zahl der verfügbaren Produkte verkleinern wird, da vor allem kleinere Firmen mit ihren Nischenprodukten Schwierigkeiten bekommen werden, sich am Markt zu behaupten [Dre98b]. Die Gründe hierfür sind unter anderem in den zu hohen Erwartungen der Anwender und in den übertriebenen Versprechungen der Hersteller zu suchen. Während diese eine einfache Integration der Data Mining-Technologie in die Unternehmensstruktur und schnelle und umfassende Einblicke in Unternehmensdaten versprachen, hat sich in der Praxis herausgestellt, dass es sich um einen teuren und langwierigen Prozess handeln kann.

3.1.2 Kriterien für die Auswahl der Tools

Bei der Einführung von Data Mining in einem Unternehmen des pharmazeutischen Großhandels handelt meist um ein Projekt mit einer längerfristigen Ausrichtung. Zur Integration dieser Technologie in das Unternehmen wird von vielen Analysten die Gründung einer eigenen Abteilung empfohlen, die die Projekte in Zusammenarbeit mit den Fachbereichen durchführt und die Ergebnisse den einzelnen Unternehmensbereichen zur Verfügung stellt. Dafür kommen nur Produkte in Frage, die über einen bestimmten

Marktanteil und über einen gewisse Reputation des Herstellers verfügen. Damit soll gewährleistet werden, dass ein Unternehmen auch in den kommenden Jahren auf dem neuesten Stand der Data Mining-Technologie ist und eventuell auftretende Probleme in Zusammenarbeit mit dem Hersteller beseitigt werden können.
Ein weiteres Kriterium bei der Vorauswahl war die Flexibilität der Produkte und welche Phasen des Data Mining-Prozesses von ihnen unterstützt werden. Im pharmazeutischen Großhandel werden typischerweise große Datenmengen verarbeitet und es existieren viele unterschiedliche Problemstellungen, zu deren Lösung verschiedene Techniken benötigt werden. Da die Ergebnisse meist vielen Benutzern zugänglich gemacht werden sollen, müssen Mechanismen vorhanden sein, die eine komfortable Visualisierung der Ergebnisse und deren Weiterverarbeitung ermöglichen. Diesen Anforderungen werden derzeit nur wenige Produkte gerecht. Vor allem kleinere Anbieter haben meist Produkte im Angebot, die nur ein oder zwei Algorithmen unterstützen oder nur für bestimmte Problemlösungen geeignet sind [Bre97].
In verschiedenen Berichten der Gartner Group wurden 1999 die Hersteller beziehungsweise deren Produkte benannt, die die oben erwähnten Voraussetzungen bieten (siehe [Lin99], [Str99]). Diese Produkte wurden als Grundlage für eine weitere Bewertung genommen. Hinzugekommen ist noch der XpertRule Miner von Attar Software, der schon im Rahmen des CRITIKAL-Projektes bei der Gehe Pharmahandel zum Einsatz gekommen ist. Dies ergibt eine Vorauswahl von insgesamt fünf Produkten, die im folgenden näher betrachtet werden:

- Darwin von Oracle (http://www.oracle.com/datawarehouse/products/datamining/)
- Enterprise Miner von SAS Institute (http://www.sas.com/software/data_mining/)
- Intelligent Miner for Data von IBM (http://www.software.ibm.com/data/iminer/fordata/)
- KnowledgeStudio von Angoss (http://www.angoss.com)
- XpertRule Miner von Attar Software (http://www.attar.com)

3.1.3. Testumgebung der ausgewählten Tools

Für die Evaluation wurden von den Firmen Oracle und IBM sowohl die Client- als auch die Serverversionen der Data Mining-Tools zur Verfügung gestellt. Von SAS und Angoss waren lediglich die Clientversionen erhältlich, das System von Attar besitzt keine Serverkomponente. In der folgenden Abbildung ist eine schematische Darstellung der Testumgebung der einzelnen Tools zu sehen.

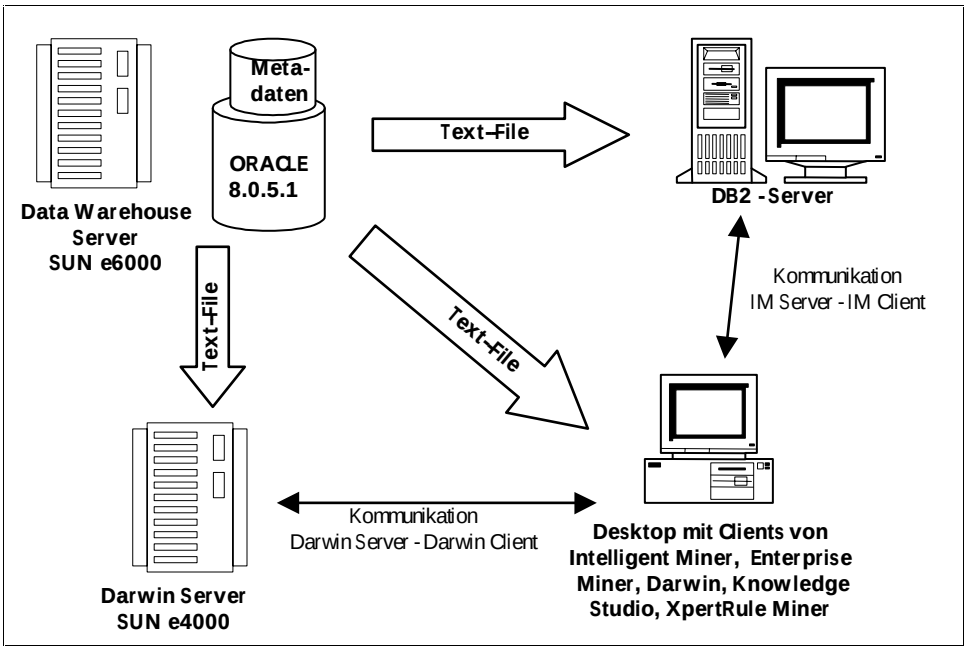


Abbildung 3.2: Testumgebung der Data Mining-Tools

Die Client-Installationen wurden alle auf einem Pentium 233MMX-Rechner mit dem Betriebssystem Windows NT 4.0 und einem Hauptspeicher von 96MB RAM vorgenommen. Mit keinem der betrachteten Tools war eine direkte Anbindung an das Data Warehouse der Gehe Pharmahandel möglich, da hierzu noch separate Produkte oder spezielle Treiber benötigt wurden, die im Rahmen dieser Evaluation nicht erhältlich waren. Aus diesem Grunde wurden die zu analysierenden Daten mittels Textimport die einzelnen Data Mining-Systeme eingelesen. Bei den Tools von Attar, Angoss und SAS erfolgte die Datenhaltung auf dem Clientrechner, während bei Darwin der Datenbestand auf dem Server verblieb. Für den Intelligent Miner von IBM war noch die zusätzliche Installation des Datenbanksystems DB2 nötig, da die Server-Komponente des Intelligent Miner ohne das Zusatzprodukt Data Joiner nur mit diesem Datenbanksystem zusammenarbeitet. Auch hier wurden die zu analysierenden Daten über Textdateien in die Datenbank importiert.

3.2. Beschreibung der Data Mining Tools

Im folgenden soll ein kurzer Überblick über die Werkzeuge gegeben werden, die im Rahmen dieser Evaluation betrachtet werden. Damit kann ein Einblick in die Arbeitsweise und die Funktionalität der einzelnen Systeme gegeben werden.

3.2.1. Intelligent Miner for Data v6.1

Einer der bekanntesten Vertreter auf dem Markt für Data Mining-Produkte ist das System Intelligent Miner for Data von IBM, das in der Version 6.1. vorliegt. Die Architektur des Systems orientiert sich am Client/Server-Prinzip und besteht aus den Komponenten des Intelligent Miner Server und des Intelligent Miner Client. Serverseitig werden alle Data Mining-Schritte von der Datenextraktion bis zur Methodenanwendung ausgeführt. Zur

Speicherung der Daten wird das hauseigene Datenbanksystem DB2 eingesetzt, die Anbindung an andere Datenbanksysteme erfolgt mittels eines separaten Produktes, dem IBM Data Joiner. Der Intelligent Miner Client ist eine Java-Anwendung, die Funktionen zur Administration, Visualisierung und zum Export von Ergebnissen bereitstellt. Die Methodenbank des Intelligent Miner erlaubt die Bearbeitung der folgenden Aufgabenstellungen:

- Assoziationsanalyse (Associations)
- Klassifikation (Classification)
- Clusteranalyse (Clustering)
- Sequenzanalyse (Sequential Patterns)
- Prognoseverfahren (Prediction)

Für die Bearbeitung dieser Aufgabenstellungen verfügt der Intelligent Miner über unterschiedliche Algorithmen, die in Abhängigkeit vom Anwendungskontext auszuwählen sind. Außerdem besitzt das System eine Reihe statistischer Funktionen (z.B. Faktorenanalyse, Regressionsverfahren, Hauptkomponentenanalyse), die auch in klassischen statistischen Analysesystemen zu finden sind. Die Verwaltung der Arbeitsumgebung erfolgt in Form einer sogenannten Mining Base (siehe Abbildung 3.3).

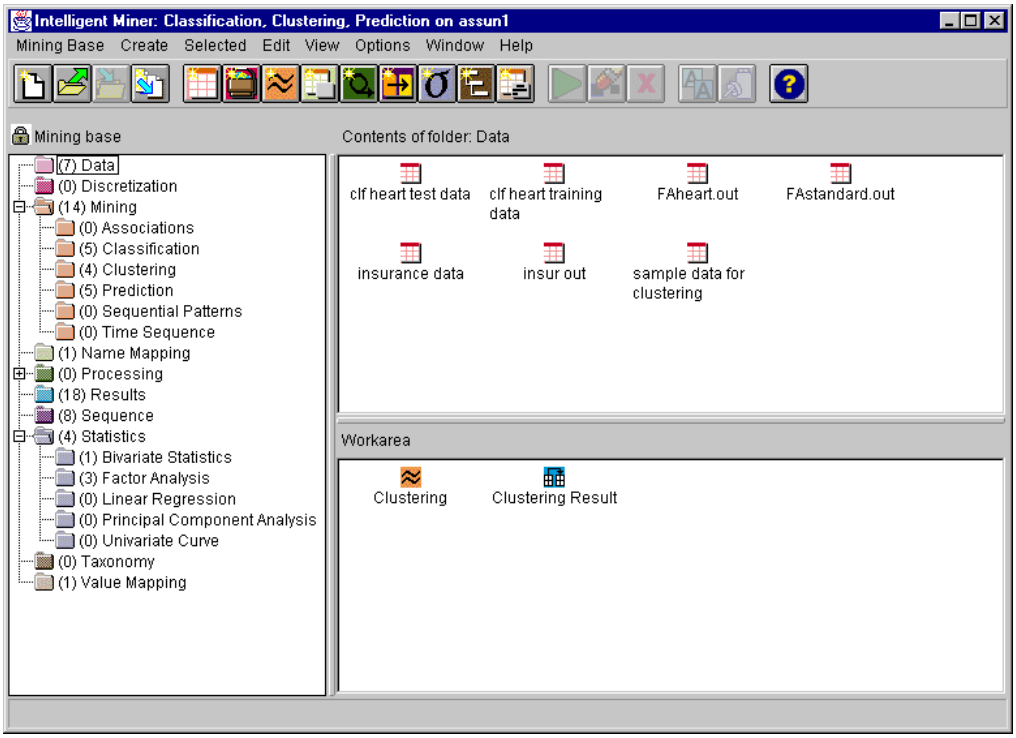


Abbildung 3.3: Ansicht der Mining Base beim Intelligent Miner von IBM

Eine Mining Base ist ein Objektcontainer, in dem der Anwender Objekte anlegt. Zu den zentralen Objekttypen sind Datenobjekte (Data), Analyseobjekte (Mining) und Ergebnisobjekte (Results) zu zählen. Bei der Definition legt der Anwender den Datenbestand für die Analyse fest. Die anzuwendende Methode wird in Form von Analyseobjekten (Associations, Classification, Clustering usw.) bestimmt und assistentengesteuert

parametrisiert. Außerdem wird festgelegt, welche Ergebnisobjekte generiert werden. Die Automatisierung von komplexen Arbeitsabläufen erfolgt durch das Anlegen von Sequenzobjekten, denen z.B. mehrere Analyseobjekte zugewiesen werden können.
Nach der Parametrisierung des Analyseobjekts erfolgt die Methodenanwendung. Das hier dargestellte Verfahren des Demographischen Clusterns berechnet die Ähnlichkeit der Objekte des Datenbestands anhand der Attributwerte und ermittelt die Anzahl der Cluster. Nach der Methodenanwendung erfolgt die Visualisierung des Ergebnisobjektes, das die Gesamtheit der identifizierten Cluster beschreibt. Bei der Anwendung in Abbildung 3.4 wurden insgesamt neun Cluster identifiziert, die im Ergebnisobjekt horizontal dargestellt werden.

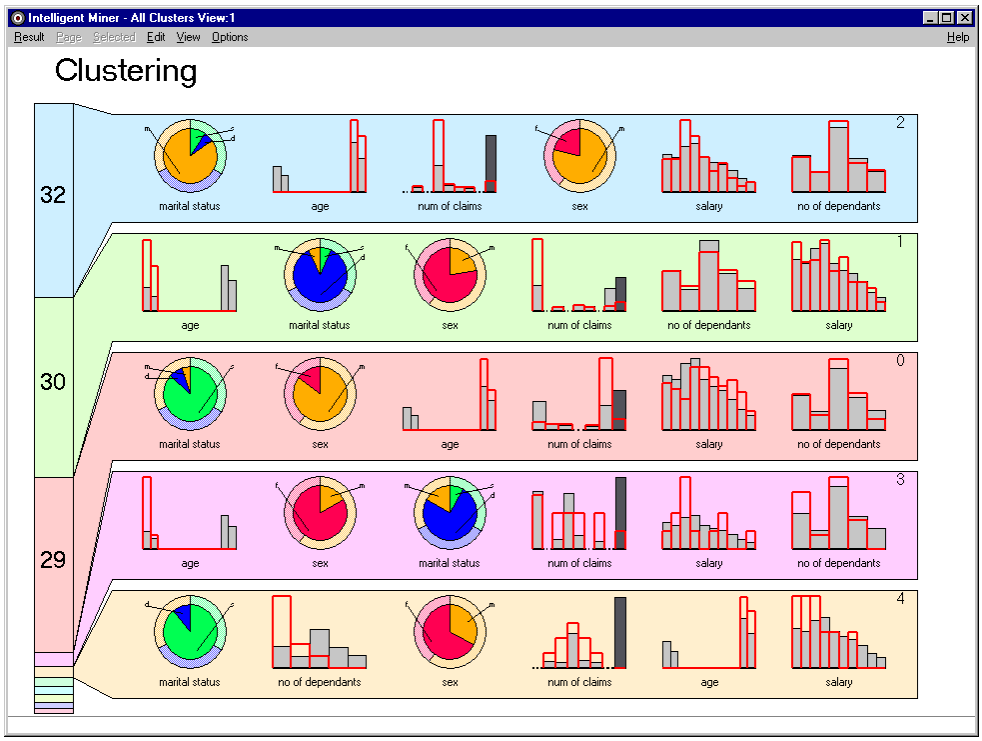


Abbildung 3.4: Ergebnisobjekt des demografischen Clusterings

Die Sortierung der Cluster erfolgt dabei nach ihrer relativen Größe. Für jeden Cluster werden außerdem die Attributausprägungen grafisch visualisiert. Die Darstellung kategorischer Attribute erfolgt als Kreisdiagramm, während numerische Attribute in Form zweidimensionaler Balkendiagramme präsentiert werden. Bei der Attributdarstellung werden in jedem Diagramm die Werte für den jeweiligen Cluster sowie für die Gesamtheit der Werte dargestellt. Die Sortierung der Attribute erfolgt in Abhängigkeit von dem Einfluss auf die Clusterbildung (mit abnehmenden Einfluss von links nach rechts). Zur Untersuchung einzelner Cluster verfügt der Intelligent Miner über eine Drill-Down-Mechanismus, der den Zugriff bis auf die elementaren Analyseergebnisse ermöglicht.
Um ein „automatisches“ Data Mining zu ermöglichen, steht eine API (**A**pplication **P**rogramming **I**nterface) zur Verfügung. Über diese Schnittstelle kann über di Programmiersprachen C oder C++ der Zugriff auf alle Funktionen des Programms erfolgen.

3.2.2. Enterprise Miner v3.0

Die Datenanalyse verkörpert für das SAS-Institute schon immer das Hauptgeschäft. Aus der langjährigen Erfahrung in diesem Bereich resultiert das SAS-eigene Vorgehensmodell zur Datenanalyse: SEMMA:

- **S**tichproben
- **E**xploration
- **M**odifaktion
- **M**odellierung
- **A**uswertung

Dieses Vorgehensmodell hat auch Auswirkungen auf den Aufbau der Benutzungsoberfläche (siehe Abbildung 3.5). Im linken Teil des Bildschirms sind zu den einzelnen Punkten des SEMMA-Modells die entsprechenden Arbeitsschritte aufgeführt, die angeboten werden.

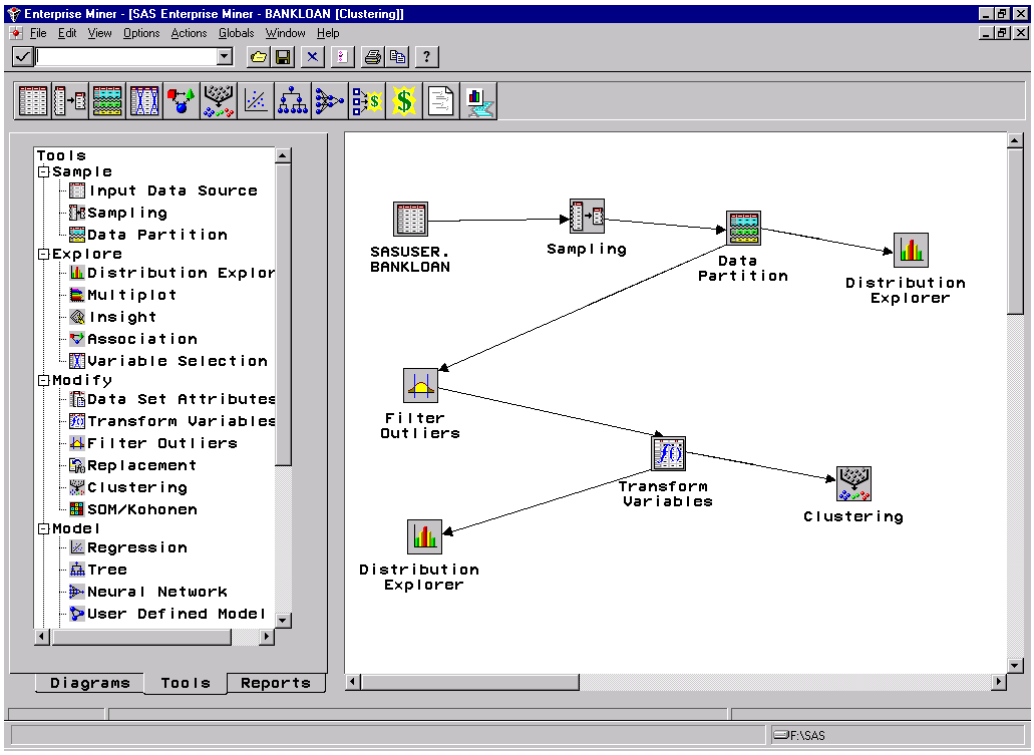


Abbildung 3.5 : Benutzungsoberfläche beim Enterprise Miner

Die graphische Oberfläche des Enterprise Miner erlaubt sehr einfach den Aufbau beliebiger Prozessfluss-Diagramme mit diesen Arbeitsschritten, was im rechten Teil zu sehen ist. Mit diesen Prozessfluss-Diagrammen wird der komplette Data Mining-Prozess gesteuert.

Die Datenbereitstellung (*Input Node*) ist der Ausgangspunkt aller Data Mining Projekte. Eine erste Visualisierung durch Häufigkeitsdiagramme ist möglich. Nach der Auswahl der Eingabedatei besteht die Option, Einzelheiten zu den Analysevariablen festzulegen. Dabei kann jeder Variablen eine sogenannte Rolle, bspw. als abhängige Variable bei

Entscheidungsbäumen (Target) oder als Transaktionsnummer bei der Assoziationsanalyse (Id) zugewiesen werden, die im weiteren Mining-Prozess beibehalten wird.
Die Datenteilung (*Data Partition Node*) sichert die Einteilung in Trainings-, Prüf- und Test-Daten, die für die Analysen eine wichtige Grundlage darstellen. Die Trainingsdatei dient der ersten Modellanpassung. Auf der Prüfdatei wird der Modellbildungsprozess verfolgt und die Modelle gegebenenfalls verbessert. Auf den Testdaten erfolgt der letztendliche Modellvergleich. Diese Daten können als geschichtete Stichproben erzeugt werden, wobei eine sogenannte Data Mining Database automatisch erzeugt wird. Sie dient der Beschleunigung aller Auswertungen. Diese Data Mining Databases bestehen aus zwei Teilen, eine Speicherplatz sparenden SAS-Datei, wobei Textvariable als Zahlen kodiert sind und einem Katalog von Metainformationen, die Ziel- und Input-Variablen beschreiben. Für die Modellierung stehen folgende Methoden zur Verfügung:

- Clusteranalyse
- Assoziationsanalyse
- Regressionsanalyse als „das Rückgrat des Data Mining“
- Entscheidungsbäume
- Neuronale Netze

Als Besonderheit bietet der SAS ein Verfahren zur Modellbewertung (*Assessment*) an. Dem Anwender sind die Ergebnisse der Modellbildung in einer interaktiven Auswertungsgraphik zugänglich. Diese enthält Güte-, Gewinn- und Aufwand-Nutzen-Graphiken. Dabei werden automatisch die betriebswirtschaftlich vorgegebenen Gewinn-Verlust-Koeffizienten mit einbezogen. Dies führt zu den Profit- bzw. ROI-Graphiken (ROI: Return on Investment).

3.2.3. Darwin v3.6.0

Im Vergleich zu den anderen Tools, die im Rahmen dieser Diplomarbeit betrachtet wurden, hat Darwin einen etwas ungewöhnlichen Werdegang hinter sich. Ursprünglich war das Tool eine interne Entwicklung des amerikanischen Großrechnerherstellers Thinking Machines Inc. um damit das Leistungsverhalten massiv paralleler Rechner zu testen. Daraus wurde mit beträchtlichem finanziellen Aufwand eine Data Mining-Anwendung entwickelt, die 1997 fertiggestellt wurde. Nachdem das Unternehmen in finanzielle Schwierigkeiten geriet, wurde das Produkt 1999 von dem Softwarehersteller Oracle zur Ergänzung seiner Produktpalette erworben. Für die Evaluation lag die Version 3.6 in einer Beta-Version vor.

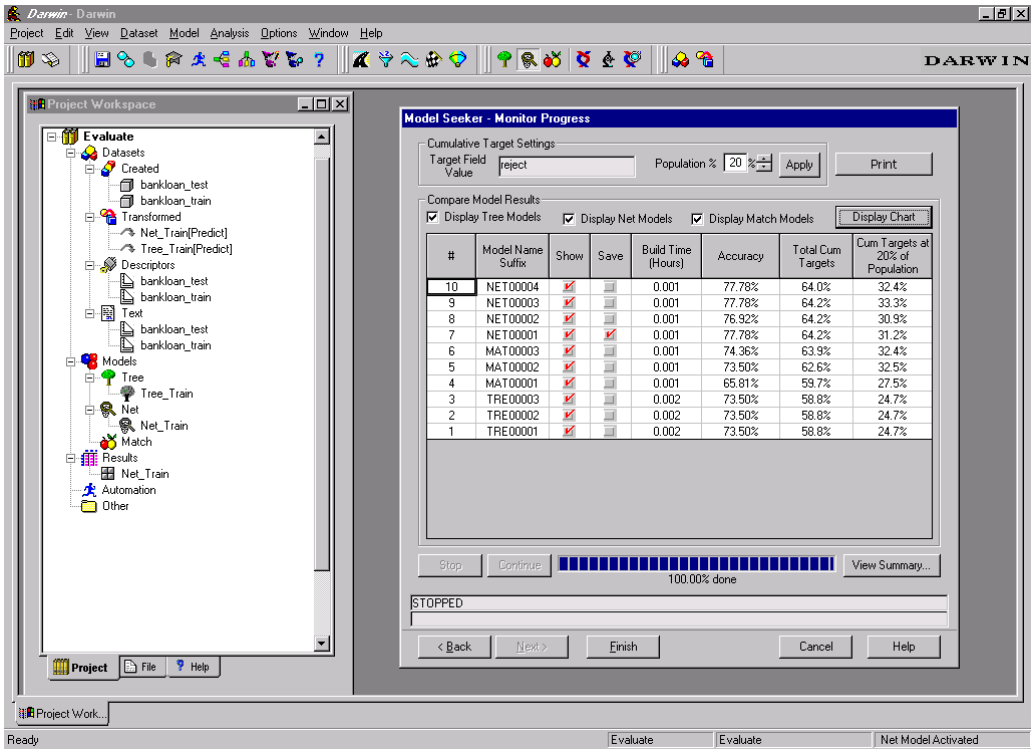


Abbildung 3.6 : Ergebnisse einer Modellsuche bei Oracle Darwin

Ähnlich wie bei Angoss erfolgt die Visualisierung der vorhandenen Datenobjekte, Modelle und Resultate in einer Explorer-ähnlichen Darstellung (siehe Abbildung 3.6). Bei der Funktionalität konzentriert sich Darwin auf die Generierung von Entscheidungsbäumen und die Segmentierung. Für die Klassifikation ist neben einem CART-Algorithmus und neuronalen Netzen auch ein Match-Modell auf der Basis der k-nächste-Nachbarn-Technik vorhanden. Zur Segmentierung wird ein Algorithmus auf Basis der Methode des k-means Clustering angeboten. Eine Besonderheit stellt bei Darwin der sogenannte *Model Seeker Wizard* dar. Damit können die einzelnen Modelle mit verschiedenen Parametern auf einen Trainingsdatenbestand angewandt werden. Die Anzahl der Modelle kann vom Benutzer vorgegeben werden, während die Parameter von dem Tool erzeugt werden. In der Abbildung 3.6 ist das Ergebnis eines solchen Laufes auf der rechten Seite zu sehen. Insgesamt wurden zehn Modelle verwendet, nämlich drei Entscheidungsbäume, drei Match-Modelle und vier neuronalen Netze. In der Darstellung kann sehr gut abgelesen werden, dass die neuronalen Netze bei dem vorliegenden Datenbestand am besten abschneiden. Das Modell mit dem höchsten Genauigkeitswert kann gespeichert werden und zur Klassifikation von neuen Datensätzen benutzt werden.

3.2.4. KnowledgeStudio v2.0

Das KnowledgeStudio von Angoss greift beim Desktop-Aufbau auf bewährte Windows-Konzepte zurück. So werden im linken Teil des Fensters das Projekt und alle zugehörigen Komponenten in Baumform dargestellt. Im Bildabschnitt oben rechts der Abbildung 3.7 sind die ausgewerteten Daten in Tabellenform zu sehen. An erster Stelle steht der Feldname, anschließend der Datentyp, die Anzahl fehlender Werte usw. Im unteren Teil der Grafik ist

schließlich ein Entscheidungsbaum zu sehen, der die Daten nach dem stärksten Unterscheidungskriterium klassifiziert.

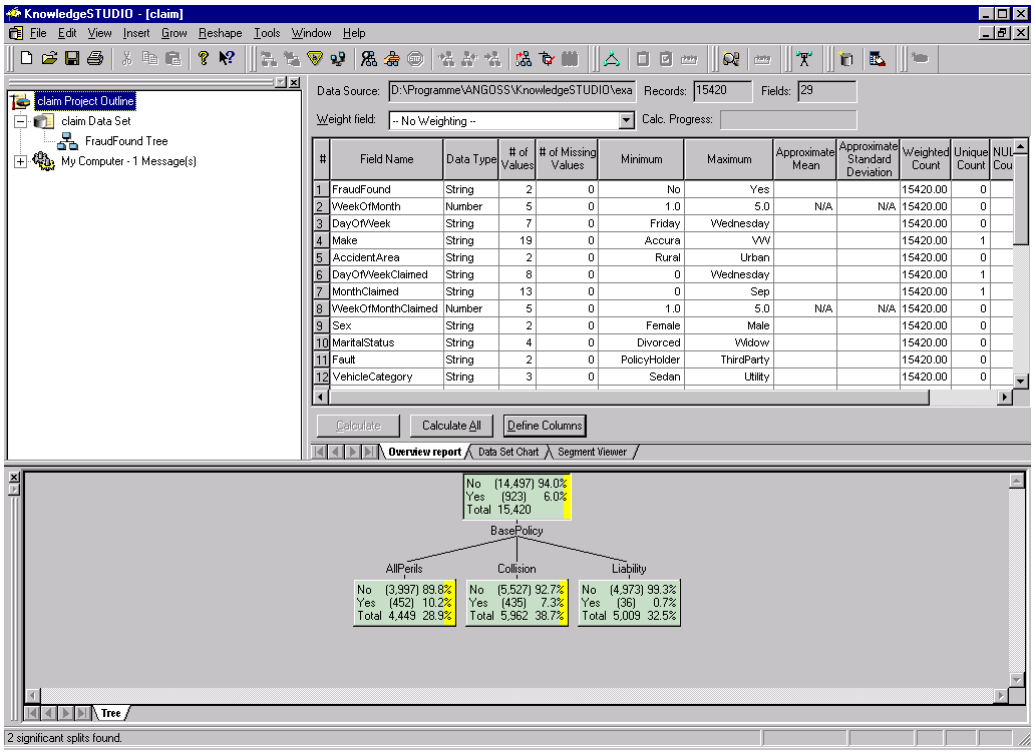


Abbildung 3.7 : Oberfläche des KnowledgeStudio

Um die Bedienung des Programms zu erlernen, steht in der Hilfe ein sehr übersichtliches Tutorial zur Verfügung, welches anhand zahlreicher Screenshots und ausführlicher Erklärungen eine sehr gute Einführung in das Programm bietet. Für den Einsteiger werden außerdem die theoretischen Grundlagen der Analysemethoden erklärt. Als weitere Erleichterung wird der Benutzer bei beinahe allen Aktionen von Assistenten begleitet, die Anfangswerte für die Algorithmen vorgeben und so den Prozess des Data Mining deutlich vereinfachen.

Nach dem Anlegen eines Projektes, wahlweise auf dem Server oder dem lokalen PC, ist zuerst eine Transformation der Rohdaten in das spezifische KnowledgeStudio-Format nötig. Neben dem ODBC-Import stehen eine Menge zusätzlicher Formate zur Verfügung, so zum Beispiel die Dateien mehrerer Tabellenkalkulationen (Lotus 1-2-3, Excel), Daten im Textformat oder SPSS-Dateien. Die importierten Daten werden in der bereits oben beschriebenen Tabellenform dargestellt und können anschließend analysiert werden. KnowledgeStudio stellt dabei den Entscheidungsbaum stark in den Mittelpunkt und bietet entsprechend viele visuelle Präsentationsmöglichkeiten (zwei- oder dreidimensionale Darstellung, versch. Diagrammarten, Farbauswahl).

Ein weiterer methodischer Schwerpunkt ist die Clusteranalyse, die mit den Methoden des K-Means-Clustering, des probabilistischen Clustering oder mit künstlichen neuronalen Netzen (Kohonen-Netze) durchgeführt werden kann. Die generierten Cluster werden anschließend in der gleichen Form wie die Entscheidungsbäume dargestellt.

Zur Integration der Data Mining-Funktionen in eigene Applikationen stellt Angoss eine gut dokumentierte ActiveX-Schnittstelle zur Verfügung.

3.2.5. XpertRule Miner v1.46

Der XpertRule Miner von Attar konzentriert sich auf die Generierung von Entscheidungsbäumen und die Assoziationsanalyse von fallbasierten und transaktionsorientierten Daten. Für beide Datenstrukturen gibt es ein eigenes Analysewerkzeug, den *Transaction Association Miner* und den *Case Association Miner*. Im Hauptfenster des Programms kann das aktuelle Projekt mit allen Ein- und Ausgangsdaten und den damit verbundenen Analysen betrachtet werden (siehe Abbildung 3.8).

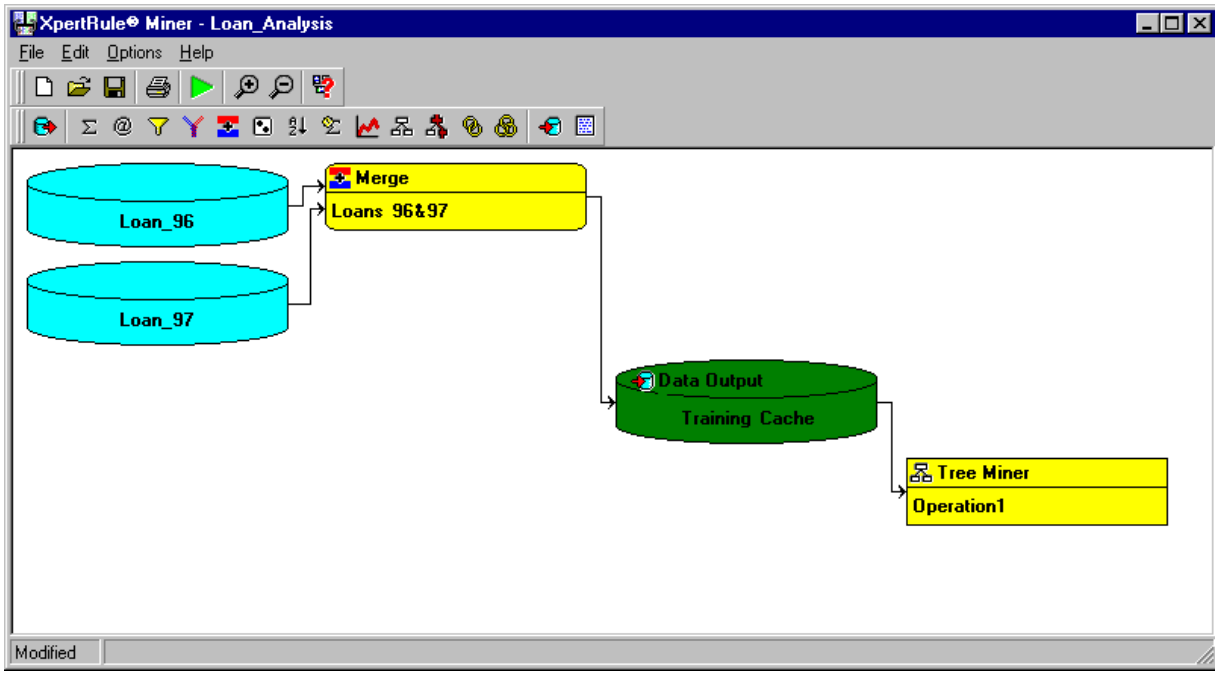


Abbildung 3.8: Oberfläche des XpertRule Miner

Die Darstellung der Beziehungen erfolgt mit Hilfe von Pfeilen. Beim Betätigen des Start-Buttons (grüner Pfeil) wird das Projekt von links nach rechts abgearbeitet.

Zur Datenanbindung stellt der XpertRule Miner eine ODBC-Schnittstelle und ein eigenes DMT-Format (**D**ata **M**ining **T**ables) zur Verfügung. Erfolgt der Datenzugriff über die ODBC-Schnittstelle, so werden die benötigten Daten während der Laufzeit angefordert. Bei größeren Datenmengen empfiehlt es sich, diese lokal in einem DMT-Cachefile abzuspeichern. Vorteilhaft für den Benutzer ist, dass er selbst entscheiden kann, ob die Datenanbindung immer auf dem aktuellem Stand sein muss, oder ein permanentes DMT-Cachefile angelegt werden soll.

Zur Generierung eines Entscheidungsbaumes kann eine abhängige Variable ausgewählt werden, anhand derer der XpertRule Miner einen Baum erzeugt. Dieses Modell wird zur Klassifikation von neuen Datensätze herangezogen. Die Qualität des Entscheidungsbaumes kann anhand von Validierungsdaten getestet werden. Die folgende Abbildung stellt das Ergebnis eines derartigen Test dar.

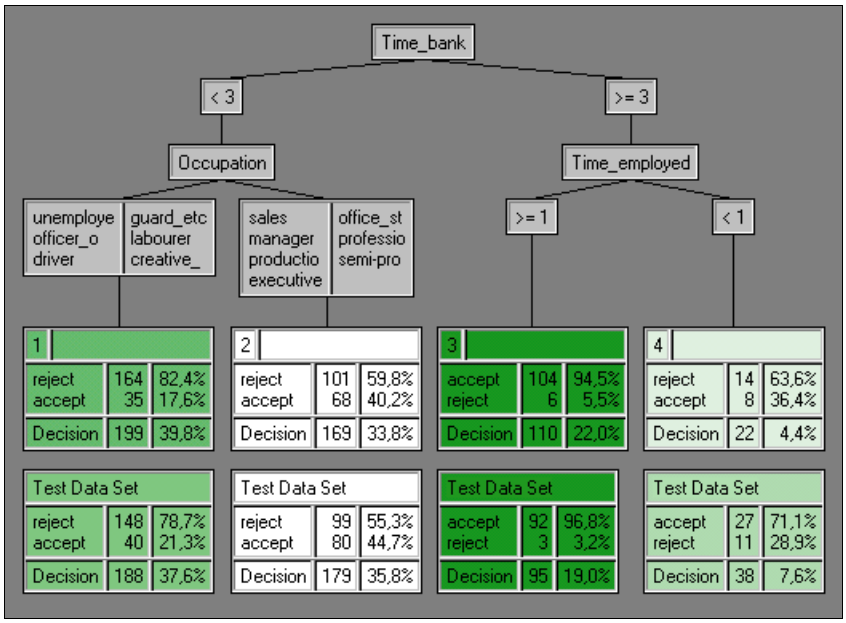


Abbildung 3.9: Testergebnis einer Klassifikation beim XpertRule Miner

Im unteren Teil der Abbildung 3.9 ist das Ergebnis für die Testdatensätze zu sehen. Eine Zeile darüber wird das Klassifikationsergebnis für die Trainingsdatensätze dargestellt.. Abgesehen von dem in der Abbildung ganz rechts stehenden Teil liefert der Entscheidungsbaum auch für die Testdaten ein gutes Klassifikationsergebnis.

Zur Integration in eigene Anwendungen bietet der XpertRule Miner die Möglichkeit, eine sogenannte XTM-Datei zu generieren, die mit Hilfe von ActiveX in beliebige Programme eingebunden werden kann. Beispiele für die Programmierung unter Visual Basic, Delphi und HTML sind in der Hilfedatei ebenso beschrieben wie die einzelnen Prozeduren.

3.3. Vorgehensweise bei der Evaluation

Hier wird vorgestellt, anhand welcher Kriterien und mit welcher Vorgehensweise die Evaluation durchgeführt wurde. Anhand von insgesamt sieben sogenannten Kriterienpunkten werden zunächst die Anforderungen an die Data Mining-Tools beschrieben. Die Bewertung der Tools erfolgte anhand von Einzelkriterien, die zu jedem dieser Kriterienpunkte erstellt wurden, was im nächsten Abschnitt beschrieben ist. Daran anschließend wird die Bewertung und Gewichtung der Einzelkriterien dargestellt, aus denen sich die Gesamtbewertung des Kriterienpunktes errechnet.

3.3.1. Vorstellung der Kriterienpunkte

Zur Sammlung der hier vorgestellten Anforderungen wurden unterschiedliche Quellen benutzt. Einen ersten Anhaltspunkt stellten die Veröffentlichungen dar, die in diesem Bereich erhältlich waren (siehe [Abb], [Dre99], [Eld], [Fer99], [OV97]). Ergänzt wurde dies durch die Erfahrungen bei der praktischen Arbeit mit den Tools, durch Gespräche mit den Vertretern der Hersteller und durch die Anforderungen, die durch die Verantwortlichen des IT-Bereiches der Gehis definiert wurden.

Einordnung des Herstellers

Da es sich bei der Einführung der Data Mining-Technologie bei der Gehe Pharmahandel um ein Projekt mit einem längerfristigen Bezug handelt, muss gewährleistet sein, dass der Hersteller auch in den nächsten Jahren am Markt vorhanden ist und das ausgewählte Produkt unterstützt. Für diese Anforderungen wurde versucht, Kriterien wie Umsatz, Unternehmensgröße oder Existenzdauer des Unternehmens zu finden, die diese Anforderungen widerspiegeln. Da es sich beim Data Mining um eine relativ neue Forschungsrichtung handelt, werden vor allem im Bereich der Algorithmen Fortschritte erzielt. Diese Neuerungen sollten in dem Produkt implementiert werden, so dass es stets auf dem neuesten Stand der Technik ist. Hierfür können die Ausgaben für Forschung und Entwicklung und das Vorhandensein einer eigenen Entwicklungsabteilung einen Indikator bilden. Für die Schulung von Mitarbeitern ist es vorteilhaft, wenn eine geografische Nähe zu dem Hersteller vorhanden ist, zumindest sollte eine Niederlassung in Deutschland existieren. Bei der Projekteinführung oder bei spezifischen Problemen ist wichtig, dass eine Analysten- oder Consultingabteilung zur Verfügung steht, die bei der Bewältigung dieser Aufgaben herangezogen werden kann. Das Vorhandensein einer solchen Abteilung stellt ebenfalls ein Kriterium dar.

Systemarchitektur

In Abbildung 3.10 sind die möglichen Konfigurationen von Data Mining-Systemen dargestellt. Bei der 1-Schicht-Architektur befinden sich der zu analysierende Datenbestand und das Data Mining-Tool auf dem gleichen Rechner. Hier muss der gesamte Datenbestand über Flat Files oder Binärdateien importiert werden. Dies stellt die einfachste und kostengünstigste Alternative dar, eignet sich aber nur für die Analyse von kleineren Datenbeständen. Bei der Verarbeitung von großen Datenmengen, wie sie bei der Gehe vorhanden sind, wird eine Verteilung der Komponenten auf zwei oder drei Schichten angestrebt. Hier wird der notwendige Import erst dann vorgenommen, wenn die Daten vom Analysealgorithmus benötigt werden. Dies wird durch einen direkten Zugriff auf die Datenbank realisiert.

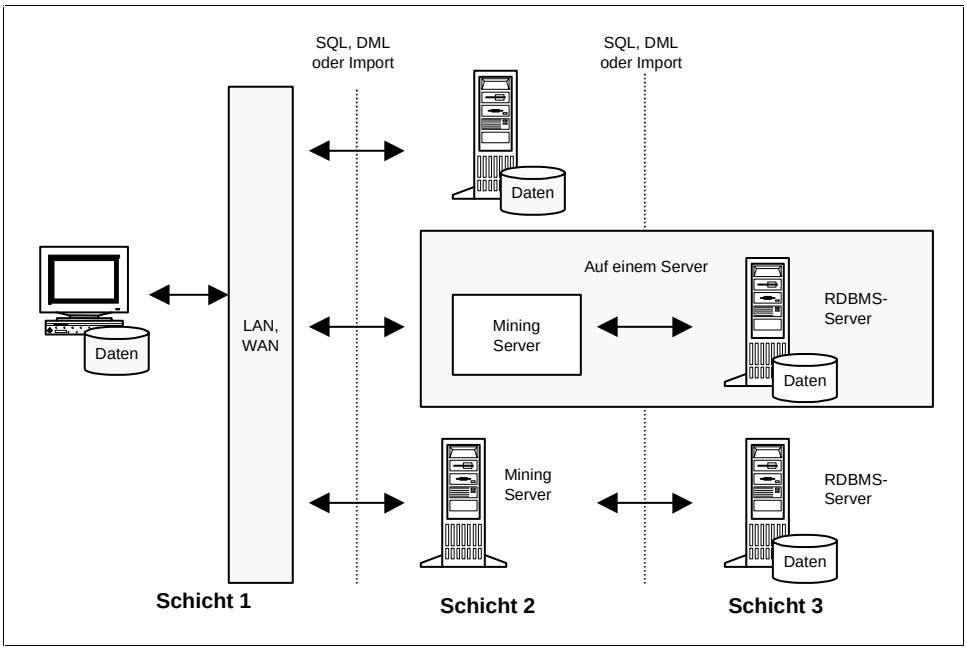


Abbildung 3.10: Architekturen für Data Mining-Systeme

Neben der Architektur stellt bei der Analyse von großen Datenbeständen eine Unterstützung von mehreren Prozessoren ein entscheidendes Kriterium dar, da dies die Verarbeitungsgeschwindigkeit deutlich erhöhen kann. Sowohl die Client- als auch die Server-Komponente sollten auf den gängigsten Betriebssystemen verfügbar sein, damit bei einem eventuellen Systemwechsel eine gewisse Flexibilität erhalten bleibt. Auch die Anforderungen an die Hardware, die das Programm stellt, sollen hier betrachtet werden.

Datenexploration

Bei der Analyse von großen Datenbeständen ist es für den Anwender zunächst wichtig, ein Gefühl für die Struktur und das Aussehen der Daten zu gewinnen. Hierzu kann ein Data Mining Tool grundlegende statistische Werte wie Schiefe, Wölbung oder Standardabweichung u.ä. anbieten, um dem Benutzer einen ersten Eindruck von der Verteilung der Daten zu vermitteln. Für das Erkennen von Ausreißern können Werte wie Minimum oder Maximum eine Anhaltspunkt bilden. Auch zur Entdeckung von fehlenden Werten sollten entsprechende Funktionen vorhanden sein. Für einen schnellen und intuitiven Einblick in den Datenbestand ist eine auch geeignete grafische Darstellung wünschenswert, die an eigene Bedürfnisse angepasst werden kann.

Datenvorverarbeitung

Zunächst wird untersucht, welche Möglichkeiten das Programm bei der Datenselektion anbietet, beispielsweise das Filtern von Datensätzen und das Verknüpfen (Join) oder Verschmelzen (Merge) von Tabellen. Zwei während des Data Mining-Prozesses häufig benötigte Funktionalitäten sind die Ziehung von Stichproben und die Aufteilung der Analysedaten in Trainings- Test- oder Validierungsdaten. Die Möglichkeiten hierzu sollen ebenfalls betrachtet werden. Bei der Datentransformation wird unter anderem untersucht, inwieweit das Berechnen neuer Feldinhalte oder das Vornehmen von Aggregierungen erleichtert wird. Von Interesse ist dabei vor allem, welche Möglichkeiten zur Kategorisierung

von Variablen zur Verfügung stehen, also inwieweit Attributausprägungen zu Gruppen zusammengefasst werden können. Einen weiteren Kriterienpunkt stellen die Behandlungsoptionen für fehlende Werte und Ausreißer dar. Zu diesem Zweck können beispielsweise statistische Funktionen vorhanden sein, mit denen die Ersetzung dieser Werte ermöglicht wird.

Verfügbarkeit der Methoden

Für jede Data Mining-Methode wie Klassifizierung, Segmentierung, Erkennung von Assoziationen und die Änderungs- und Abweichungserkennung wird hier eine separate Bewertung vorgenommen. Dabei wird zunächst geprüft, ob das Tool für diese Methode überhaupt eine algorithmische Unterstützung anbietet. Getestet wird, welche Möglichkeiten das Tool zur Parametrisierung der einzelnen Algorithmen anbietet und ob das Tool anhand der Datenbasis Hinweise oder Vorschläge anbietet, die den Benutzer bei dieser Parametrisierung unterstützen (beispielsweise bei der Architekturbestimmung von neuronalen Netzen). Es soll auch bewertet werden, wie verständlich und intuitiv die Ergebnisse der Mining-Läufe visualisiert werden und welche Optionen zu deren Export angeboten werden. Dies kann beispielsweise in einer ausführbaren Form mittels SQL-Statements oder Java-Skripten oder in bildlicher Form mittels Excel-Dateien oder HTML geschehen.

Bedienung

Damit soll der Grad der Benutzerfreundlichkeit und die Verständlichkeit der Oberfläche bewertet werden. Unter anderem fließt in die Bewertung mit ein, ob der Benutzer in den einzelnen Phasen des Data Mining-Prozesses durch ein geeignetes Hilfesystem unterstützt wird und welche Qualität dieses Hilfesystem besitzt. Dabei soll auch betrachtet werden, ob für die einzelnen Operationen Assistenten zur Verfügung stehen und ob sich für unerfahrene Anwender bestimmte Funktionalitäten des Programms verbergen lassen wodurch der Einsatz des Tools auch außerhalb einer reinen Analystenabteilung ermöglicht wird. Für die Bedienung ist unter anderem auch maßgeblich, ob während eines Data Mining-Laufes Fortschrittsanzeigen eingeblendet werden und ob sich diese Läufe abbrechen lassen. Schließlich wird noch festgehalten, ob während der praktischen Arbeit mit dem System Programmfehler oder Abstürze aufgetreten sind.

Automatisierbarkeit

Wie schon in Abschnitt 2.3.3. gesehen, weist das Data Mining einen hohen Grad von Prozessorientierung auf, so dass ein Data Mining-Lauf im Normalfall aus mehreren Schritten wie Datenselektion, Datentransformation und Methodenanwendung besteht. Mit diesem Kriterienpunkt soll nun geklärt werden, inwieweit die sequentielle Ausführung von Operationen in den Programmablauf integriert bzw. welche Möglichkeiten das Tool hierzu anbietet. Damit verbunden ist auch die Fragestellung, ob eine eigene Programmier- oder Makrosprache in das Tool integriert ist, womit die Flexibilität um beispielsweise periodisch auftretende Abläufe oder Problemstellungen zu automatisieren, deutlich erhöht wird. Des weiteren soll betrachtet werden, ob die erstellten Modelle oder Resultate in anderen Applikationen weiterverwendet werden können. So könnte es beispielsweise bei einem

Klassifikationsproblem möglich sein, das Entscheidungsbaummodell als SQL-Statement oder C-Code abzuspeichern und in eine eigene Anwendung zu integrieren.

3.3.2. Bewertung der Tools anhand der Einzelkriterien

In einem nächsten Schritt wurden die Fragestellungen, die sich hinter den im vorigen Abschnitt vorstellten Kriterienpunkte verbergen, in Einzelkriterien abgebildet. Zu jedem Kriterienpunkt wurden dabei zwischen drei und sieben Einzelkriterien formuliert, die in Anhang B aufgelistet sind. Für jedes Tool wurde nun anhand dieser Einzelkriterien eine Bewertung vorgenommen. Dabei ergaben sich folgende Bewertungsmöglichkeiten:

Bewertung	Bedeutung
- -	Das Tool unterstützt das Kriterium / die Funktionalität gar nicht oder nur äußerst unzureichend.
-	Das Kriterium / die Funktionalität wird schlecht oder leicht mangelhaft unterstützt.
o	Das Kriterium / die Funktionalität wird eher durchschnittlich unterstützt.
+	Das Kriterium / die Funktionalität wird gut unterstützt.
++	Das Kriterium / die Funktionalität wird hervorragend unterstützt.

Tabelle 3.1: Bewertungsmöglichkeiten der Einzelkriterien

Die Bewertung der Einzelkriterien erfolgte anhand der Kenntnisse, welche bei der praktischen Arbeit mit diesen Tools erworben wurden. Bei der Evaluation der Bedienbarkeit der Data Mining-Tools wurden auch die Erfahrungen berücksichtigt, die von Mitarbeitern der Gehe im Rahmen einer kurzen Einarbeitung in die Funktionalität der einzelnen Tools gemacht wurden. Ergänzt wurde die Bewertung durch Informationen der Hersteller in Form von Handbüchern und Veröffentlichungen im Internet und durch persönliche Gespräche mit Consultants der einzelnen Anbieter.
Kriterien, bei denen die Verfügbarkeit auf den verschiedenen Betriebsystemen, die Hardwareausstattung und die Sprachunterstützung hinterfragt wurde, flossen in die Wertung nicht mit ein, da diese einen rein informativen Charakter besitzen. Auch für die Algorithmen, mit denen die Tools die verschiedenen Data Mining-Methoden unterstützen, wurde keine Bewertung vorgenommen. Diese Kriterien sind in der Bewertung der Einzelkriterien in Anhand C mit einem „*“ gekennzeichnet.

3.3.3. Gewichtung der Einzelkriterien

Je nach Bedeutung der einzelnen Kriterien für das Data Mining bei der Gehe Pharmahandel wurden diese mit einer Gewichtung von 1 (eher unwichtig) bis 3 (sehr wichtig) versehen. Für die Festlegung der Gewichtungen wurden einerseits die Anforderungen der Fachabteilungen berücksichtigt, so dass beispielsweise die Möglichkeiten zur Datenselektion geringer gewichtet wurde, da sich diese Aufgabe mit den im Unternehmen vorhandenen OLAP-Werkzeugen schon erledigen lässt. Auf der anderen Seite wurden die Erfahrungen, die sich aus der praktischen Arbeit mit den Tools ergaben stark berücksichtigt. So wurde bspw. die

Unterstützung der Kategorisierung der Variablen sehr stark gewichtet, da diese Funktion bei der praktischen Arbeit immer wieder benötigt wurde. Mit den Gewichtungen war es auch möglich, bei der differenzierten Betrachtung der Data Mining-Methoden Schwerpunkte zu setzen. So wurde bei der Segmentierung die Visualisierung der Resultate wesentlich höher bewertet, während bei der Klassifikation die Überprüfung der Modellgenauigkeit im Vordergrund steht.
Aus der Bewertung der einzelnen Kriterien und deren Gewichtung errechnet sich die Bewertung des gesamten Kriterienpunktes. Dazu werden die oben vorgestellten Bewertungen in ein Punktesystem umgerechnet, was in der folgenden Tabelle zu sehen ist.

Bewertung	Punktzahl
- -	1
-	2
o	3
+	4
++	5

Tabelle 3.2: Umrechnung der Bewertung in Punktzahlen

Diese Punktzahlen werden mit den Gewichtungen multipliziert. Dieses Ergebnis wird durch die Anzahl der Gewichtungen geteilt, woraus sich nach dem Runden eine Bewertung des Kriterienpaketes zwischen 1 und 5 ergibt. Bei der Bewertung der Data Mining-Methoden kann sich auch eine 0 ergeben, für den Fall, dass das Data Mining-Tools keine Unterstützung für diese Methode oder Funktionalität anbietet. Eine Übersicht über die Gewichtung der Einzelkriterien und deren Bewertungen ist in Anhang C zu sehen.

3.4. Ergebnisse der Evaluation

In diesem Abschnitt werden die Ergebnisse der Evaluation näher beschrieben. Zunächst werden die Bewertungen, die von den Tools in den oben vorgestellten Kriterienpunkten erreicht wurden, in Form von Balkendiagrammen dargestellt. In einer textlichen Beschreibung wird kurz skizziert, welches die maßgeblichen Gründe waren, die zu dieser Einordnung führten. Anschließend wird eine Übersicht über die Gesamtbewertung der Tools gegeben, wobei auf die Stärken und Schwächen der einzelnen Produkte eingegangen wird.

3.4.1. Ergebnisse der Kriterienpunkte

Einordnung des Herstellers:

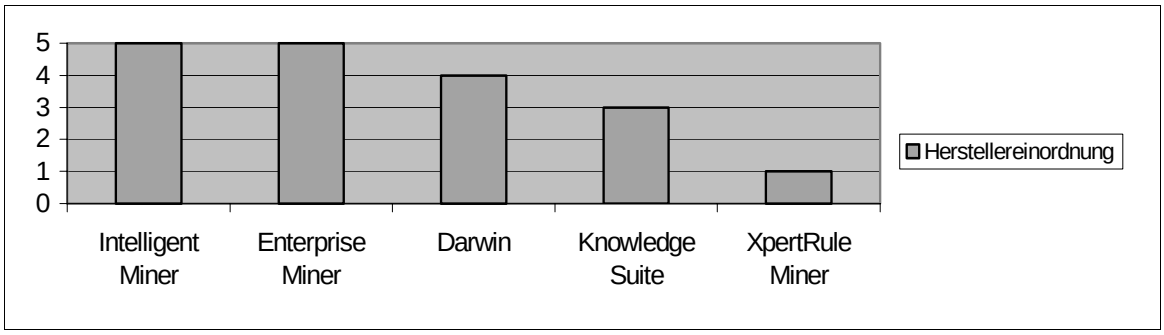


Abbildung 3.11: Bewertung der Herstellereinordnung

Sowohl IBM, SAS als auch Oracle sind aufgrund ihrer Unternehmensgröße und ihrer Marktpräsenz in der Lage, ihr Produkt auf dem neuesten Stand der Technik zu halten und den Anforderungen des Markt auch in den nächsten Jahren zu begegnen. Für IBM spricht die Tatsache, dass im Unternehmen eine eigene Abteilung existiert, die Grundlagenforschung im Bereich der Algorithmen betreibt. Genau wie bei SAS wird ein umfangreiches Schulungs- und Trainingsprogramm für das Produkt angeboten. Bei SAS wirken sich die jahrelangen Erfahrungen positiv aus, die das Unternehmen im Bereich der statistischen Analyse besitzt. Außerdem investiert das Unternehmen im Vergleich zu anderen Herstellern einen überproportional hohen Anteil in die Bereiche Forschung und Entwicklung. Bei Angoss handelt es sich zwar um eine relativ kleines Unternehmen, es hat sich aber im Gegensatz zu den anderen Herstellern mittlerweile auf die Entwicklung von Data Mining-Anwendungen konzentriert.

Systemarchitektur:

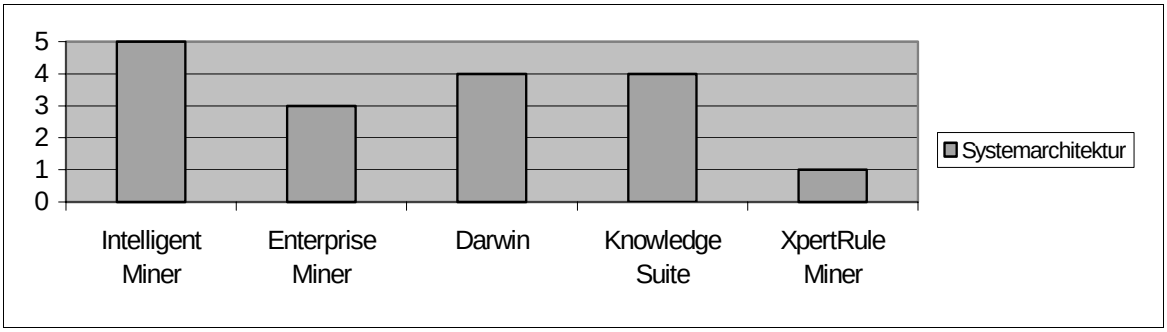


Abbildung 3.12: Bewertung der Systemarchitektur

Der XpertRule Miner von Attar ist im Rahmen dieser Evaluation das einzige Tools, das keine Server-Komponente besitzt. Zur Anbindung an andere Datenbanken wird lediglich eine ODBC-Schnittstelle angeboten, was bei großen Datenmengen aufgrund der geringen Geschwindigkeit unzureichend ist. Der Intelligent Miner von IBM bietet neben einem breiten Zugriff auf verschiedne Datenbanksysteme auch eine umfassende Unterstützung von Multiprozessoren an. Damit kann eine gleichzeitige Verwendung von mehreren Verarbeitungsknoten erreicht werden. Diese Verarbeitungsknoten können auch auf anderen Rechner eingerichtet werden. Zusätzlich lässt sich über eine Zeit- oder Festplattenoptimierung im Gegensatz zu anderen Tools der Umgang mit temporären Dateien sehr gut kontrollieren. Bei Enterprise Miner von SAS führt die mangelnde Unterstützung von Multiprozessoren und der hohe Bedarf an temporärem Speicherplatz zu einer etwas schlechteren Bewertung. Von

Darwin werden zwar in der vorliegenden Beta-Version nur Oracle-Datenbanken direkt unterstützt, dafür sind jedoch alle Algorithmen vollständig parallelisiert, was zu einer sehr schnellen Verarbeitung beiträgt.

Datenexploration:

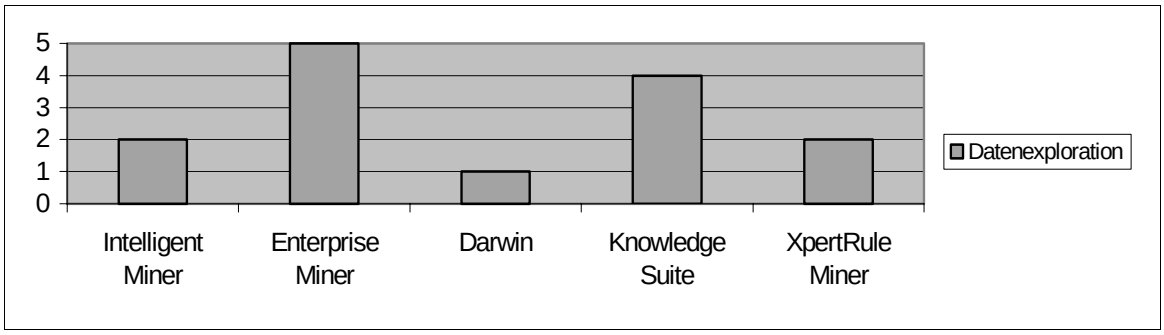


Abbildung 3.13: Bewertung der Datenexploration

Da das SAS-System als eines der am häufigsten eingesetzten Werkzeuge zur statistischen Datenanalyse die Basis für den Enterprise Miner bildet, bietet das Tool von SAS die besten Möglichkeiten, um einen schnellen und umfassenden Einblick in die Struktur des Datenbestandes zu gewinnen. Dazu werden bei der Definition einer Datenquelle anhand einer Stichprobe von maximal 2000 Datensätzen grundlegende statistische Werte wie beispielsweise Schiefe und Wölbung[4] errechnet. Zur Datenvisualisierung werden Balken- und Streudiagramme sowie 3-dimensionale Darstellungen angeboten, wobei noch das Statistik-Paket INSIGHT eingebunden werden kann, mit dem zahlreiche univariate und multivariate Analysemethoden zur Verfügung gestellt werden. Auch beim Knowledge Studio von Angoss stehen zahlreiche Diagrammtypen zur Verfügung wobei diese durch eine umfassenden Parametrisierung und zahlreiche Optionen fast beliebig an individuelle Wünsche angepasst werden können. Etwas schwach präsentiert sich der Intelligent Miner von IBM bei der Datenexploration. Hier fallen vor allem die langen Wartezeiten bei der Berechnung der Statistik auf. Zur Visualisierung der Häufigkeitsverteilung steht lediglich die Darstellung zur Verfügung, die für die Präsentation der Clusteringresultate eingesetzt wird (siehe Abschnitt 3.2.1.). Darwin bietet von sich aus keine Möglichkeiten zur Datenexploration an. Zur Datenvisualisierung wird MS Excel eingesetzt, was sich bei großen Datenmengen aufgrund der Speicheranforderungen und den damit verbundenen langen Laufzeiten als völlig ungeeignet erweist.

Datenvorverarbeitung:

[4] Eine Erläuterung dieser Begriffe findet sich in Anhang E

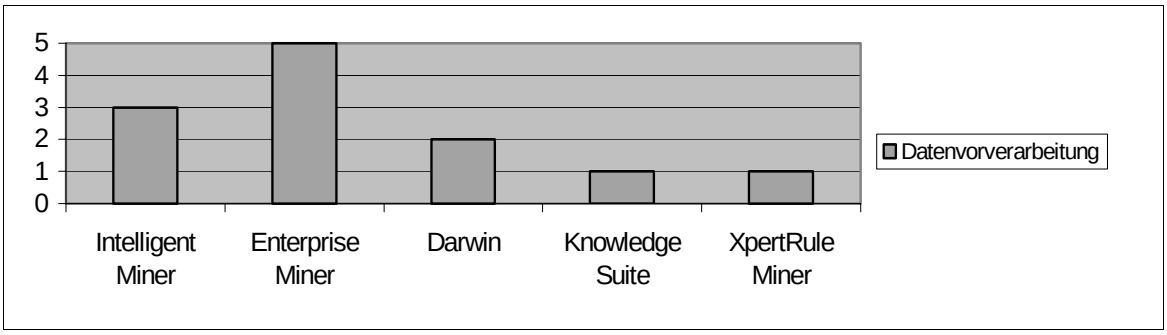


Abbildung 3.14: Bewertung der Datenvorverarbeitung

Auch bei der Datenvorverarbeitung erreicht der Enterprise Miner von SAS von allen Tools die höchste Wertung. Neben zahlreichen arithmetische Operationen zum Berechnen von neuen Feldinhalten steht zur Aufteilung in Test-, Trainings- oder Validierungsdaten eine eigene Operation zur Verfügung. Auch bei der Stichprobenziehung werden zahlreiche Optionen angeboten, mit denen die Auswahl der Datensätze beeinflusst werden kann. So besteht beispielsweise die Möglichkeit der Ziehung einer geschichteten Stichprobe. Dazu können Variablen ausgewählt werden, anhand derer die Datensätze gruppiert werden. Je höher die Häufigkeit einer Gruppe ist, desto mehr Datensätze werden aus ihr entnommen. Innerhalb einer Gruppe findet wieder eine normale Zufallsauswahl statt. Damit kann die Aufteilung besser an die realen Charakteristika der Daten angepasst werden. Für das Ersetzen fehlender Werte stehen unter anderem Verteilungsfunktionen zur Verfügung, mit denen fehlende Werte aufgrund der Verteilungen der übrigen Variablen berechnet werden können. Auch der Intelligent Miner bietet eine komfortable Stichprobenziehung und gute Ersetzungsmöglichkeiten für fehlende Werte an, allerdings ist die Kategorisierung von Variablen etwas umständlich. Am schlechtesten bewertet wurden die Tools von Attar und Angoss. Insbesondere die fehlenden Möglichkeiten zur Kategorisierung von Variablen und zur Stichprobenziehung bzw. zur Aufteilung der Datensätze fielen negativ auf.

Funktionalität Klassifikation:

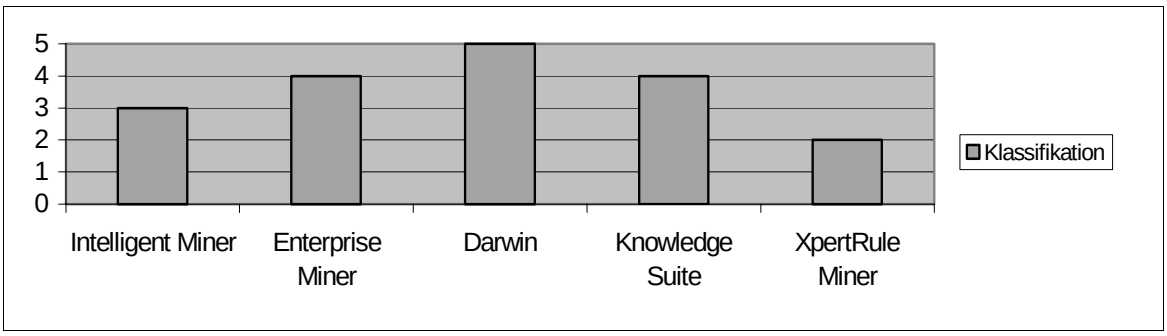


Abbildung 3.15: Bewertung der Klassifikation

Die Klassifikation ist die einzige Data Mining-Methode, die von allen Tools, typischerweise mit Entscheidungsbäumen und neuronalen Netzen unterstützt wird. Am besten bewertet wurde Darwin, da sich neben einer breiten algorithmischen Unterstützung auch die Genauigkeit der einzelnen Modelle hervorragend verfolgen lässt. Besonders positiv fiel bei der praktischen Arbeit der Assistent zur Modellsuche auf, mit dem nach dem Modell gesucht werden kann, das anhand von Trainings- und Testdaten die höchste Genauigkeit aufweist.

Sowohl der Enterprise Miner als auch das KnowledgeStudio zeichnen sich durch einen hohen Grad an Parametrisierbarkeit der einzelnen Algorithmen und durch eine umfassende und verständliche Visualisierung der Entscheidungsbäume aus. Beim Intelligent Miner fällt auf, dass nur sehr wenige Parameter eingestellt werden können. So ist beispielsweise eine Verwendung von unterschiedlichen Trennmaßen nicht möglich. Auch die Visualisierung der Ergebnisse wirkt im Vergleich mit den anderen Tools etwas rudimentär. Beim XpertRule Miner von Attar besteht das Problem, dass keine Übersicht über die Genauigkeit des Modells vorhanden ist. Diese muss vielmehr vom Benutzer aus der Missklassifikationsrate in den einzelnen Knoten errechnet werden.

Funktionalität Segmentierung:

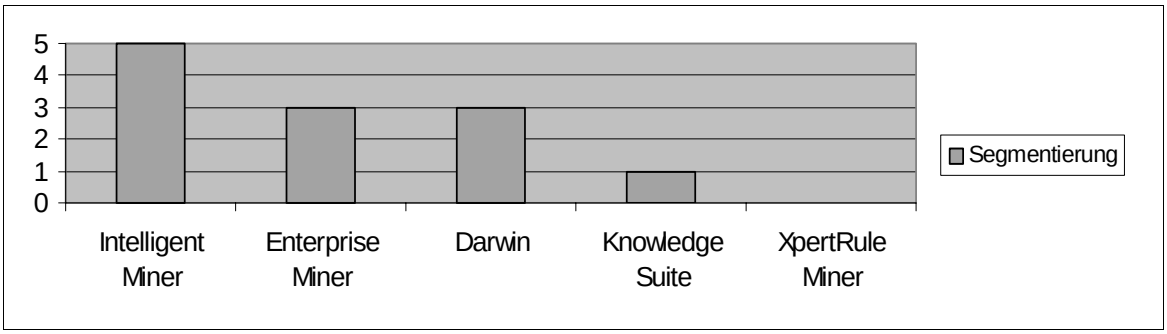


Abbildung 3.16: Bewertung der Segmentierung

Wird beim Data Mining das Ziel der Segmentierung verfolgt, dann stellt das Produkt von IBM die erste Wahl dar. Vor allem bei der Visualisierung der Resultate bietet der Intelligent Miner die übersichtlichste und verständlichste Darstellung aller betrachteten Tools. Im Vergleich zum Enterprise Miner lassen sich zwar weniger Parameter einstellen, diese sind aber sehr verständlich und lassen sich für den Benutzer sehr gut nachvollziehen. Für SAS stellt die Segmentierung keine eigenen Data Mining-Methode dar, sondern wird lediglich als ein Schritt in der Datenvorverarbeitung zur Aufteilung der Daten in einzelne handhabbare Gruppen betrachtet. Daraus resultiert eine Darstellung der Segmentierung, in der die Charakteristika der einzelnen Cluster nur äußerst schwer nachvollzogen werden kann. Bei Darwin wird zwar eine etwas bessere Darstellung der einzelnen Cluster erreicht, allerdings wirkt sich hier die sehr eingeschränkte Parametrisierung negativ aus. Auch die Unterschiede zwischen den einzelnen Clustern lassen sich nur schwer erkennen. Beim KnowledgeStudio wird diese Methode nur ansatzweise unterstützt. Mit Hilfe der Segmentierung kann lediglich eine Einteilung der Datensätze in die gefundenen Cluster vorgenommen werden. Eine Visualisierung der Eigenschaften der einzelnen Cluster findet nicht statt. Von Attar wird keine Algorithmus zur Segmentierung angeboten.

Funktionalität Abhängigkeitsanalyse:

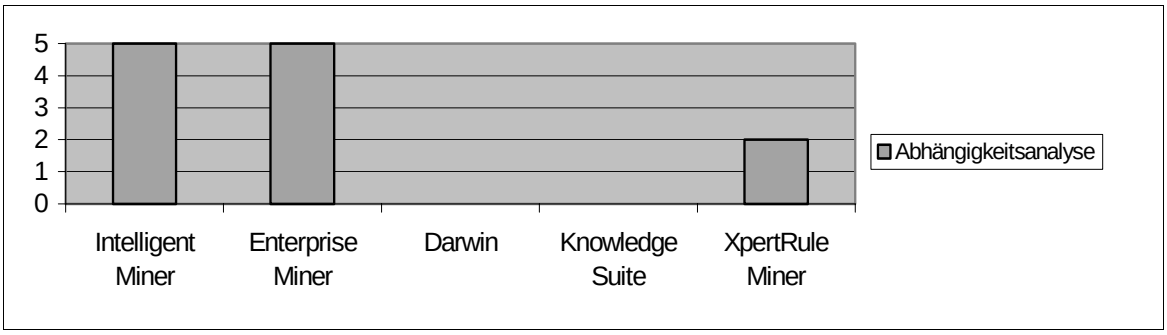


Abbildung 3.17: Bewertung der Abhängigkeitsanalyse

Die Abhängigkeitsanalyse wird sowohl vom Intelligent Miner als auch vom Enterprise Miner sehr gut unterstützt. Beide bieten die Erkennung von Assoziationen für transaktionsorientierte Daten und die Erkennung von zeitlichen sequentiellen Mustern an. Bei der Parametrisierung der Algorithmen ergeben sich keine großen Unterscheide zwischen den betrachteten Tools. Über die Angaben zu dem minimalen Träger, der minimalen Konfidenz und einer maximalen Regellänge kann die Menge der gefundene Muster gut gesteuert werden. Zur Visualisierung benutzt der Intelligent Miner ein Konzept der verbundenen Sichten. Dabei kann in einer grafischen Darstellung eine bestimmte Anzahl von Regeln selektiert werden, worauf diese in einer zweiten Ansicht mit Hilfe von Punktobjekten und Pfeilen grafisch dargestellt werden. Diese Darstellung kann vom Benutzer fast beliebig an eigene Wünsche angepasst werden. Auch der Enterprise Miner von SAS bietet eine sehr gute Visualisierung der Resultate. Hier kann die Anzahl der Regeln, die angezeigt werden sollen, auf grafischem Wege äußerst komfortabel eingestellt werden. Betrachtetet man die Genauigkeit der Resultate, so ergeben sich zwischen den Tools, die die Abhängigkeitsanalyse anbieten, keinerlei Unterschiede. Alle drei produzieren exakt die gleichen Resultate mit denselben Werten für Support und Konfidenz. Beim XpertRule Miner von Attar wird allerdings keine grafische Visualisierung der Ergebnisse vorgenommen. Auch die Darstellung der Ergebnisse kann nicht eingeschränkt werden, was sich bei vielen gefundenen Assoziationsregeln als äußerst nachteilig erweist. Von den Produkten von Oracle und Angoss wird diese Methode nicht unterstützt.

Funktionalität Änderungs- und Abweichungserkennung:

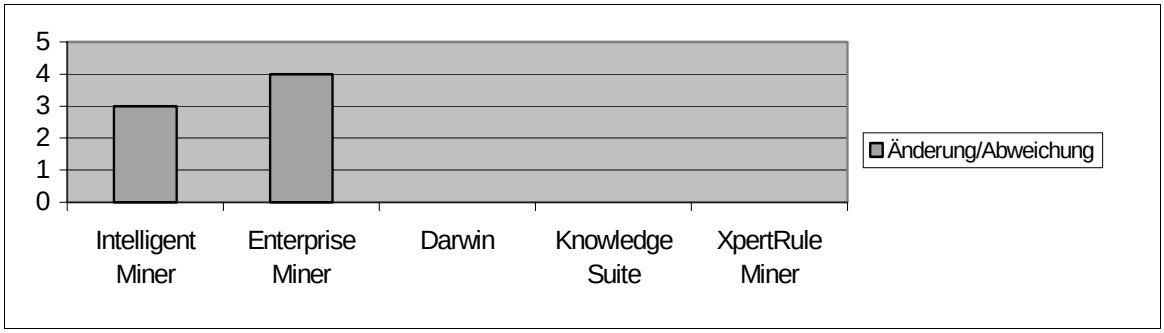


Abbildung 3.18: Bewertung Änderungs- und Abweichungserkennung

Für die Änderungs- und Abweichungserkennung biete sowohl der Intelligent Miner als auch der Enterprise Miner die klassische Regressionsanalyse an. Beim Enterprise Miner bieten sich hier mehr Optionen zur Einstellung der Parameter und Möglichkeiten bei der Visualisierung der Ergebnisse. Anhand einer Klassifizierungsmatrix und zahlreichen statistischen Angaben

kann die Genauigkeit des Modells sehr gut überprüft werden. Zur Abweichungserkennung bietet der Enterprise Miner im Gegensatz zum Intelligent Miner auch Streudiagramme an, die sich durch zahlreiche Optionen und Erweiterungsmöglichkeiten auszeichnen. Das Tool von IBM bietet als einziges die Möglichkeit einer Zeitreihenanalyse, womit sich in einem zeitlichen Kurvenverlauf Ähnlichkeiten in Form von identischen Untersequenzen erkennen lassen. Integriert ist auch eine Faktoren- und Hauptkomponentenanalyse. Diese lassen sich zum Zusammenfassen von Daten und zum Identifizieren linearer Beziehungen zwischen Variablen verwenden. Das Ziel besteht darin, eine größere Menge von beobachtbaren Variablen, auf möglichst wenige hypothetische Variablen, die Faktoren, zurückzuführen.

Bedienung:

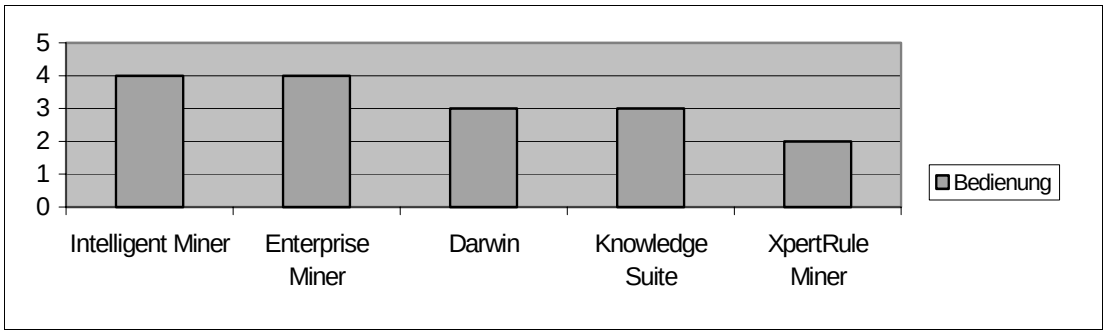


Abbildung 3.19: Bewertung der Bedienung

Der Enterprise Miner von SAS bietet von allen Tools die übersichtlichste Benutzerführung und Bedienung. Die einzelnen Operationen sind auf der Oberfläche anhand der Reihenfolge im Data Mining-Prozess angeordnet und können mittels Drag & Drop sehr komfortabel zur Verwendung selektiert werden. Negativ wirkt sich hier nur die mangelnde Unterstützung für unerfahrene Benutzer aus, da bei den meisten Operationen eine sehr komplexe Parametrisierung vorgenommen werden muss. Darüber hinaus können die Data Mining-Läufe nicht abgebrochen werden. Auch die Oberfläche des Intelligent Miner macht einen sehr ausgereiften Eindruck. Pluspunkte sind hier vor allem die sehr konsistente Benutzerführung, was über die durchgängige Verwendung von Assistenten erreicht wird. Darüber hinaus ist ein kontextsensitive Hilfesystem mit vielen praktischen Beispielen vorhanden und es besteht die Möglichkeit zum jederzeitigen Abbruch der Data Mining-Operationen. Im Gegensatz hierzu macht die Bedienung bei Darwin und beim KnowledgeStudio einen etwas unausgereiften Eindruck. Beide bieten keine durchgängige Benutzerführung und erwecken eher den Eindruck, als seien sie aus Einzelkomponenten zusammengesetzt worden. Allerdings ist das KnowledgeStudio durch eine starke Anlehnung ans das MS Office Look & Feel für viele Benutzer leicht zu erlernen, während Darwin durch eine sehr ausführliche Dokumentation und ein gutes Hilfesystem hervorsticht. Bei Attar findet sich wie beim Enterprise Miner eine übersichtlich aufgebaute Oberfläche mit einer einfachen Bedienung durch Drag & Drop-Operationen. Allerdings bietet der Desktop weniger Möglichkeiten diese anzuordnen und wird bei größeren Projekten sehr schnell unübersichtlich. Am negativsten fiel jedoch beim XpertRule Miner das etwas instabile Programmverhalten auf. So traten während der praktischen Arbeit mehrfach Fehler und Programmabstürze auf.

Automatisierbarkeit:

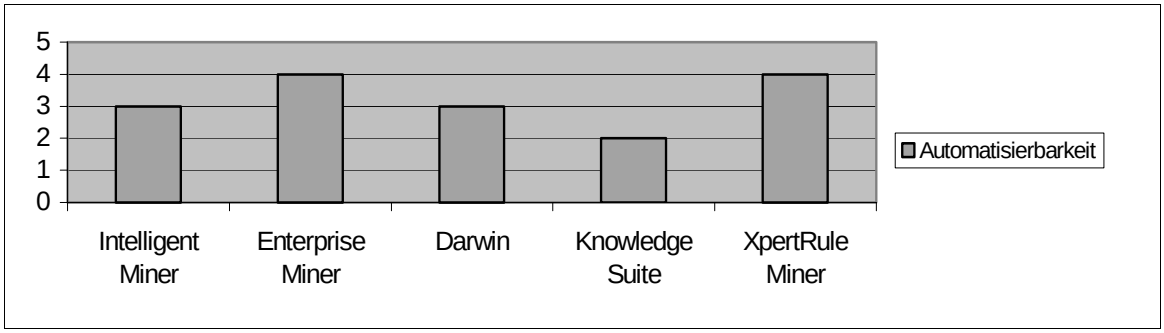


Abbildung 3.20: Bewertung der Automatisierbarkeit

Betrachtet man die Fähigkeiten der Tools, über eine sequentielle Ausführung der Operationen immer wiederkehrende Abläufe zu automatisieren, so bietet der Enterprise Miner hierfür die besten Vorraussetzungen. Aufgrund des implementierten SEMMA-Ansatzes wird eine sequentielle Anordnung der Einzeloperationen in Form von Ablaufdiagrammen förmlich erzwungen. Bei einer Änderung der Ausgangsdaten wird einfach das zugrundeliegende Datenobjekt ausgetauscht, um den Data Mining-Lauf mit den veränderten Daten durchzuführen. Zudem bietet das Tool durch eine eigene Programmiersprache, den SAS-Code, mit dem vor allem im Bereich der Datenvorverarbeitung die Flexibilität deutlich erhöht werden kann. Eine Weiterverwendung der erstellten Modelle in andere Applikationen wird allerdings nicht angeboten. Neben einer ähnlich guten Unterstützung der sequentiellen Ausführung bietet der XpertRule Miner von Attar über ActiveX und HTML die Möglichkeit an, die Entscheidungsbaummodelle in andere Applikationen (beispielsweise Visual Basic) zu integrieren. Eine eigene Programmiersprache ist ebenfalls vorhanden, kann aber nur zum Berechnen neuer Feldinhalte benutzt werden. Bei den Tools von IBM und Oracle muss eine Hintereinanderausführung von Operationen etwas umständlich vom Benutzer manuell vorgenommen werden. Eine eigene Programmiersprache wird hier nicht angeboten. Allerdings biete der Intelligent Miner die Möglichkeit, über eine eigene API auf die Funktionen des Intelligent Miners zuzugreifen, während bei Darwin ein Export der Entscheidungsbaummodelle in C++ oder Java-Code möglich ist. Vom KnowledgeStudio wird weder eine sequentielle Ausführung noch eine Programmiersprache angeboten, dafür lassen sich die Modelle in Form von ActiveX-Komponenten in andere Anwendungen einbinden.

3.4.2. Übersicht über die Gesamtbewertung der Tools

Betrachtet man die in dem vorherigen Abschnitt vorgestellte Bewertung der Kriterienpunkte, so fällt auf, dass keines der im Rahmen der Evaluation betrachteten Tools eine durchgängig hohe Bewertung bei allen Kriterien erreicht. Vielmehr hat jedes Tools seine eigenen Schwerpunkte und weist dagegen Schwächen in anderen Bereichen auf. Von allen Tools machen die Produkte von IBM und SAS den ausgereiftesten Eindruck. Beide bieten eine breite Abdeckung von Data Mining-Methoden und unterstützen mehrere Phasen des Data Mining-Prozesses. Der Enterprise Miner setzt dabei auf eine sehr umfangreiche Parametrisierbarkeit der einzelnen Algorithmen, wodurch sich für den Benutzer viele Möglichkeiten zur Optimierung der Resultate bieten. Allerdings verlangt dies vom Benutzer

ein hohes Maß an Verständnis für die Arbeitsweise der Algorithmen und stellt hohe Anforderungen an die statistischen Kenntnisse des Anwenders. Der Intelligent Miner bietet dagegen eine vergleichsweise einfache Parametrisierung an, die für den Benutzer sehr verständlich ist und anhand des direkten Einflusses auf die Ergebnisse sehr gut nachvollzogen werden kann. Zur Darstellung der Schwerpunkte der Tools wurden die Ergebnisse aus der Bewertung der Kriterienpunkte in ein 2-dimensionales Diagramm eingetragen und durch Linien verbunden. Auf der senkrechten Achse sind die einzelnen Kriterienpunkte aufgetragen, auf der waagrechten Achse die Bewertung, die dabei erreicht wurde. Aus Gründen der Übersichtlichkeit wird der Darstellung der Bewertung der Tools auf zwei Diagramme verteilt. Eine Übersicht für alle Tools findet sich im Anhang D. In der folgenden Abbildung sind die Produkte von IBM und SAS im direkten Vergleich zu sehen.

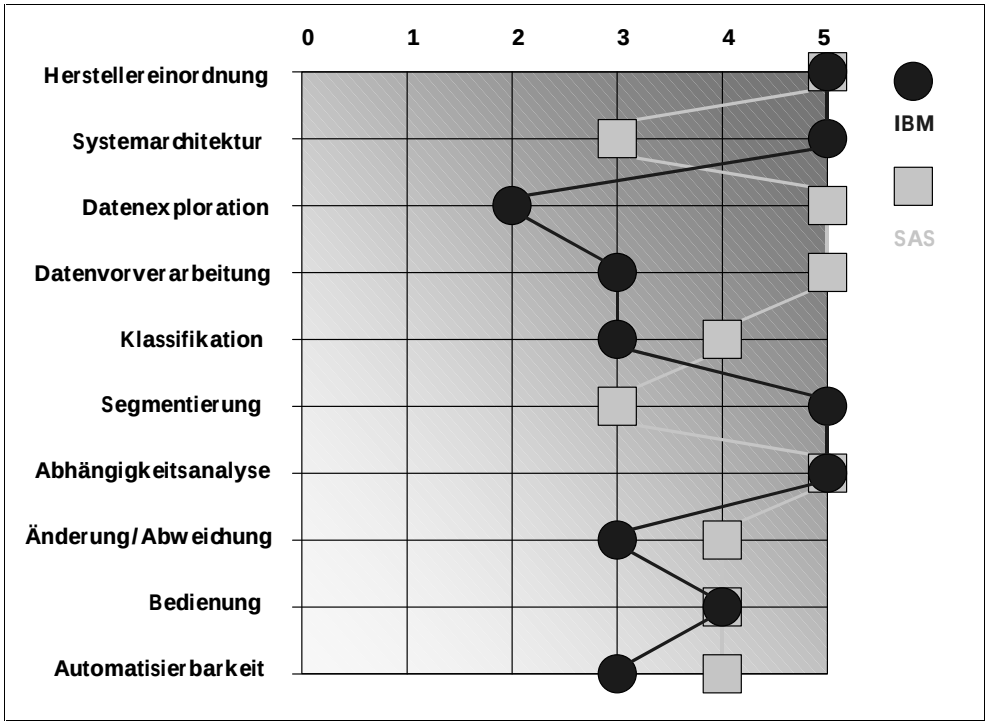


Abbildung 3.21: Vergleich der Gesamtbewertung Enterprise Miner – Intelligent Miner

In der Abbildung ist zu sehen, dass eine der Stärke des Intelligent Miner die Segmentierung darstellt, was durch eine hervorragende Visualisierung der Resultate erreicht wird. Schwachpunkte sind allerdings bei der Klassifikation und der Datenexploration erkennbar. Die Stärken des Enterprise Miners liegen, aufgrund des statistischen Schwerpunktes, bei der Datenexploration und der Datenvorverarbeitung, während die Segmentierung nur durchschnittlich unterstützt wird. In der folgenden Abbildung sind die Produkte der anderen Hersteller zu sehen.

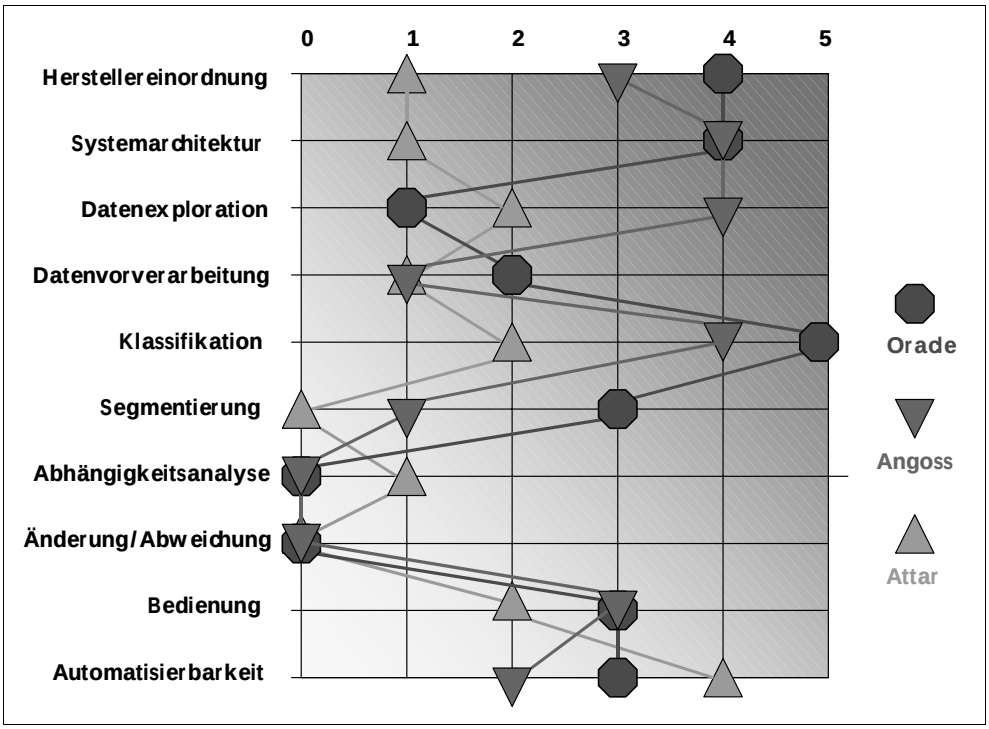


Abbildung 3.22: Vergleich der Gesamtbewertung Darwin – KnowledgeStudio – XpertRule Miner

Dank eines Assistenten zur Modellsuche und durch sehr gute Möglichkeiten zur Überprüfung der Modellgenauigkeit nimmt das Tool Darwin von Oracle einen Spitzenplatz ein. Auch die Segmentierung wird zumindest noch befriedigend unterstützt. Durch die Parallelisierung sämtlicher Algorithmen sind auch die Voraussetzungen zur Bearbeitung großer Datenmengen vorhanden. Allerdings wird die Datenexploration nur unzureichend unterstützt. Das KnowledgeStudio von Angoss hat sich ebenfalls auf die Klassifikation spezialisiert, bietet jedoch weniger Möglichkeiten um die Genauigkeit der Modelle zu überprüfen und unterstützt keine Operationen zur Datenvorverarbeitung. Durch eine 1-Schicht-Architektur bietet der XpertRule Miner von Attar die schlechtesten Voraussetzungen zur Bearbeitung großer Datenmengen. Im Gegensatz zu den anderen Tools werden hier auch weniger Möglichkeiten angeboten, um beispielsweise die Ergebnismenge bei den Assoziationsregeln einzuschränken oder die Genauigkeit des Klassifikationsmodells zu verbessern. Zudem traten während der praktischen Arbeit häufiger Abstürze auf, was sich negativ auf die Bewertung der Bedienung auswirkte.

Kapitel 4. Anwendungsgebiete für Data Mining bei der Gehe Pharmahandel GmbH

Um Anwendungsgebiete für Data Mining in einem Unternehmen festzulegen, werden zunächst Informationen darüber benötigt, wie das Tagesgeschäft in den einzelnen Geschäftsbereichen abläuft und welche Fragestellungen oder Probleme möglicherweise schon vorhanden sind. Auf der anderen Seite wird auch Wissen darüber benötigt, welche Datenbestände überhaupt im Unternehmen vorhanden sind, auf denen eine Analyse aufsetzen kann. Im Rahmen dieser Diplomarbeit wurde deshalb das Unternehmensumfeld der Gehe Pharmahandel GmbH einer näheren Betrachtung unterzogen, wobei die Besonderheiten des Marktes für pharmazeutische Produkte und die Abwicklung des Tagesgeschäftes skizziert werden. Danach erfolgte eine Einarbeitung in die Tabellen und Strukturen, die im Data Warehouse vorliegen. Aufgrund dieser Erfahrungen konnten in Zusammenarbeit mit den Fachbereichen vier beispielhafte Anwendungsgebiete für Data Mining bei der Gehe festgelegt werden, die sich für eine Bearbeitung im Rahmen einer Diplomarbeit anboten. Die Fragestellungen, die sich hinter diesen Anwendungsgebieten verbergen wurden in einem nächsten Schritt in mehreren Business Cases formuliert, von denen insgesamt drei zur Implementierung ausgewählt wurden.

4.1. Unternehmensumfeld der Gehe Pharmahandel GmbH

In dem folgenden Abschnitt soll zunächst skizziert werden, welche Aufgaben der pharmazeutische Großhandel übernimmt und durch welche grundlegenden Eigenschaften sich der Markt für pharmazeutische Produkte in Deutschland auszeichnet. Anschließend wird beschrieben, in welche Kategorien sich diese Produkte aufteilen lassen und wie sich die Abwicklung des Tagesgeschäftes mit den Kunden gestaltet.

4.1.1. Funktionen des pharmazeutischen Großhandels

Der Markt für pharmazeutische Produkte in der Bundesrepublik Deutschland ist von zwei Merkmalen gekennzeichnet: Auf der einen Seite werden von der Industrie in diesem Bereich über 250.000 Artikeln angeboten, was weltweit wohl einmalig sein dürfte. Keine noch so große Apotheke ist aufgrund der begrenzten Lagerkapazität in der Lage, auch nur einen Bruchteil dieser Angebotspalette vorrätig zu halten. Auf der anderen Seite sorgen ca. 22.000 Apotheken und Drogerien für den Absatz dieser Produkte. Eine so große Anzahl von Abnehmern kann von der Industrie nur schwer betreut werden. Diesen Tatsachen verdankt der pharmazeutische Großhandel seine Existenzberechtigung. Er übernimmt sozusagen eine Mittlerfunktion zwischen den Herstellern und den Apotheken.

4.1.2. Markteigenschaften

Der Markt zeichnet sich vor allem durch eine Angebotsoligopol aus. Er wird von fünf Großhändlern dominiert, die zusammen ungefähr 80% Marktanteil besitzen. Die restlichen 20% entfallen auf kleine, meist regionale Anbieter. Zu den Branchengrößen zählt die Phönix AG mit ca. 30% Marktanteil, gefolgt von der Gehe Pharmahandel GmbH mit ca. 20%. Die Anteile der einzelnen Anbieter sind dabei relativ stabil und variieren im Normalfall nur gering. Größere Verschiebungen sind nur durch die Übernahme eines Konkurrenten möglich, was allerdings durch kartellrechtliche Beschränkungen erschwert wird.
Eine Besonderheit beim Handel mit pharmazeutischen Produkten stellt die sogenannte Preisspannenverordnung dar. Mit ihr werden vom Gesetzgeber die Preise auf dem Weg von der Industrie zum Endverbraucher geregelt. Die Verordnung besagt, dass der pharmazeutische Großhandel auf die Preise der Industrie eine maximale Gewinnspanne aufschlagen darf. Diese ist abhängig von dem Preis des Artikels und beträgt durchschnittlich ca. 15.58%. Tatsächlich wird ein Großteil dieser Gewinnspanne in Form von Rabatten an den Apotheker weitergereicht. Daraus resultiert eine sehr kleine Umsatzrendite (bei der Gehe Pharmahandel GmbH ca. 1,8%) die sich aus dem Umsatz, geteilt durch die anfallenden Kosten ergibt. Auf diese Weise erwirtschaftet die Gehe Pharmahandel GmbH bei einen Umsatz von 5,3 Mrd. DM einen Gewinn von etwas über 100 Mio. DM.

4.1.3. Beschreibung der Sortimente

Man kann drei verschiedene Kategorien von Produkten unterscheiden. Die größte Gruppe bilden dabei die verschreibungspflichtigen oder ethischen Artikel. Sie können vom Endverbraucher nur unter Vorlage eines ärztlichen Rezeptes bei einer Apotheke bezogen werden. Ihr Anteil am Gesamtumsatz beträgt ca. 60-70%. Die zweite Gruppe setzt sich aus den apothekenpflichtigen Artikeln zusammen. Obwohl diese Artikel (beispielsweise Kopfschmerzmittel) auch ohne Rezept bezogen werden können, ist vom Gesetzgeber eine Beratungspflicht vorgeschrieben, so dass der Verkauf nur von einem ausgebildeten Apotheker vorgenommen werden darf. Die Platzierung dieser Produkte erfolgt meist in Regalen im Sichtbereich des Kunden, weshalb diese Artikel auch unter dem Begriff Sichtwahlsortiment zusammengefasst werden. Durch den immer größer werdenden Trend zur Selbstmedikation ist der Anteil dieser Produkte von derzeit 20-30% im Wachstum begriffen. Die nicht verschreibungspflichtigen Artikel bilden die dritte Gruppe. Sie werden auch als Griffwahlsortiment bezeichnet, da sie in der Regel direkt im frei zugänglichen Verkaufsbereich, also im Griffbereich des Kunden befinden. Der Verkauf dieser Artikel ist nicht ausschließlich an Apotheken gebunden, sie dürfen auch in Drogerien oder Supermärkten abgesetzt werden. Ihr Anteil liegt bei ca. 5%.

4.1.4. Abwicklung des Tagesgeschäftes

Jede Apotheke hat in der Regel einen Hauptlieferanten, mit dem sie ungefähr 80% - 90% ihres Umsatzes tätigt. Nur bei Artikeln, die beim Hauptlieferanten nicht, oder nur mit einer bestimmten Lieferfrist vorrätig sind, wendet sie sich an einen Zweit- oder Drittlieferanten. Zum Kundenstamm der Gehe Pharmahandel GmbH gehören momentan ca. 12.000

Apotheken, wovon etwa 2.000 als sogenannte Hauptkunden anzusehen sind, da sie einen überwiegenden Anteil ihres Sortimentes bei der Gehe bestellen.
Fast alle Kunden verwalten ihr Warenlager mit Hilfe eines Warenwirtschaftssystems, mit dem sie ihre Sortimentszusammensetzung und ihre Lagerhaltung kontrollieren. Dieses System unterstützt den Apotheker auch bei seinem Bestellwesen, indem es Aufträge zusammenstellt, die elektronisch mit Hilfe der Daten-Fernübertragung mit dem pharmazeutischen Großhandel austauscht werden. Beim Großhandel werden die regelmäßig und nach genauen Zeitplänen eingehenden Aufträge automatisch in den Geschäftsprozess eingespeist. Im Lager werden die Aufträge automatisiert oder manuell kommissioniert und verlassen nach weniger als 60 Minuten das Großhandelslager. Mit der Hilfe kleinerer Lieferfahrzeuge werden die Aufträge dann direkt der jeweiligen Apotheke zugestellt. In extremen Fällen erfolgen bis zu sechs Lieferungen pro Tag und Apotheke.

4.1.5. Niederlassungsstruktur der Gehe Pharmahandel GmbH

Um eine hohe Kundennähe zu erreichen, wurde das Bundesgebiet von der Gehe in sechs Regionen unterteilt, zu denen jeweils eine oder mehrere Niederlassungen mit den entsprechenden Lagereinrichtungen gehören. Die Bestellung eines Kunden wird zunächst an diejenige Niederlassung übermittelt, die die größte geografische Nähe zu dem betreffenden Kunden aufweist. Zur Abwicklung der Bestellung besitzt der Kunde eine eindeutige Nummer zur Identifikation, den sogenannte Kundenhaupteintrag. Sind einer oder mehrere bestellte Artikel nicht in der aktuellen Niederlassung vorhanden, so wird die Bestellung an eine sogenannte Zuliefererniederlassung übermittelt, welche die Lieferung der fehlenden Artikel übernimmt. Dazu erhält der Kunden eine weitere Nummer, den sogenannten Kundennebeneintrag. Hat auch die Zuliefererniederlassung die bestellten Waren nicht im Angebot, so wendet sich der Apotheker mit seiner Bestellung an einen anderen Lieferanten.

4.2. Struktur des Data Warehouse

Zunächst soll beschrieben werden, wie die Auftragsdaten in das Data Warehouse gelangen, und aus welchen Komponenten sich dieses zusammensetzt. Danach soll ein kurzer Überblick über die im Data Warehouse vorhandenen Tabellen und deren Aggregierungen gegeben werden, die für die verschieden Analyseanwendungen im Unternehmen benötigt werden. Dabei wird im besonderen Maße auf die Daten und Aggregierungen eingegangen, denen im Laufe dieser Diplomarbeit eine große Bedeutung zukam, wie beispielsweise den ATC-Warengruppen und den Daten der IMS.

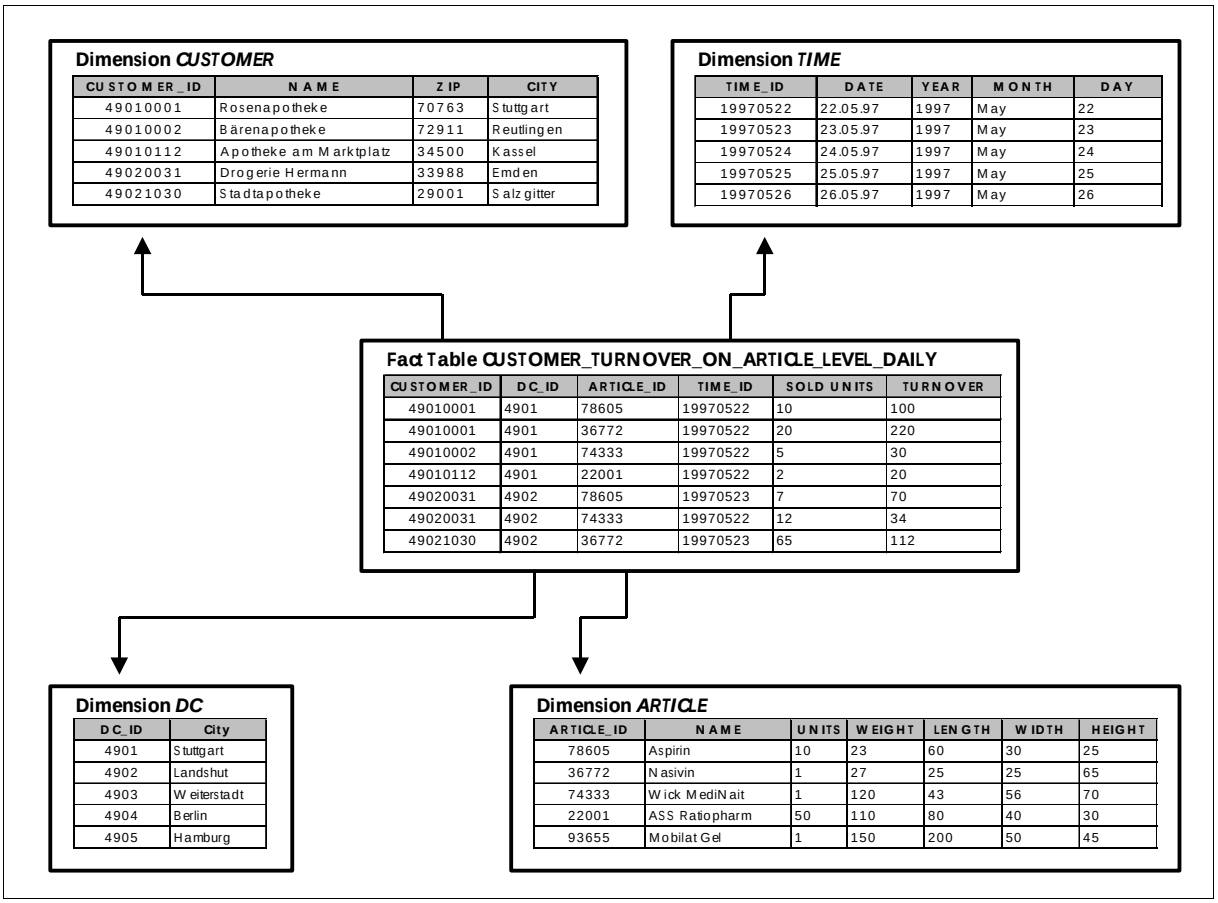
4.2.1. Transport der Auftragsdaten

In jeder Niederlassung sorgt eine Software für den täglichen Transport der Auftragsdaten in das Data Warehouse der Gehe, das sich in der Hauptverwaltung in Stuttgart befindet. Pro Woche werden dabei mehrere Millionen Datensätze übermittelt, was einem täglichen Volumen von 20-65MB entspricht. Diese Daten werden zunächst in Ladetabellen überführt,

aus denen durch PL/SQL-Transformationsprogramme und Unix-Skripten die eigentliche Faktentabelle generiert wird. Auf dieser Tabelle setzen die in Abschnitt 4.2.3. beschriebenen Aggregierungen auf. Das Data Warehouse befindet sich seit Anfang 1998 im produktiven Einsatz. Durch die Vorhaltezeit der Daten von zwei Jahren hat die Tabelle mittlerweile ein Volumen von mehreren hundert Millionen Datensätzen erreicht. Für die Speicherung kommt ein massiv paralleler Rechner vom Typ SUN Ultra Enterprise 6000 mit insgesamt 6 Prozessoren, 2GB RAM und einer Festplattenkapazität von einem Terabyte zum Einsatz. Auf der Datenbankebene wird das relationale Datenbanksystem Oracle 8.0.5.1. verwendet. Auf das Data Warehouse greifen verschiedene Informations- und Analyseanwendungen zu, wie beispielsweise die Tools von Micro Strategy.

4.2.2. Tabellenstruktur

Das Data Warehouse der Gehe Pharmahandel GmbH besteht aus einer Faktentabelle mit den getätigten Verkäufen, um die herum sich die Dimensionstabellen wie Artikel, Kunden und Lieferanten gruppieren. In der folgenden Abbildung ist eine schematische Darstellung dieser Beziehungen zu sehen, wobei natürlich nicht alle Tabellen und Verknüpfungen berücksichtigt werden konnten.



Dimension *CUSTOMER*

CUSTOMER_ID	NAME	ZIP	CITY
49010001	Rosenapotheke	70763	Stuttgart
49010002	Bärenapotheke	72911	Reutlingen
49010112	Apotheke am Marktplatz	34500	Kassel
49020031	Drogerie Hermann	33988	Emden
49021030	Stadtapotheke	29001	Salzgitter

Dimension *TIME*

TIME_ID	DATE	YEAR	MONTH	DAY
19970522	22.05.97	1997	May	22
19970523	23.05.97	1997	May	23
19970524	24.05.97	1997	May	24
19970525	25.05.97	1997	May	25
19970526	26.05.97	1997	May	26

Fact Table CUSTOMER_TURNOVER_ON_ARTICLE_LEVEL_DAILY

CUSTOMER_ID	DC_ID	ARTICLE_ID	TIME_ID	SOLD UNITS	TURNOVER
49010001	4901	78605	19970522	10	100
49010001	4901	36772	19970522	20	220
49010002	4901	74333	19970522	5	30
49010112	4901	22001	19970522	2	20
49020031	4902	78605	19970523	7	70
49020031	4902	74333	19970522	12	34
49021030	4902	36772	19970523	65	112

Dimension *DC*

DC_ID	City
4901	Stuttgart
4902	Landshut
4903	Weiterstadt
4904	Berlin
4905	Hamburg

Dimension *ARTICLE*

ARTICLE_ID	NAME	UNITS	WEIGHT	LENGTH	WIDTH	HEIGHT
78605	Aspirin	10	23	60	30	25
36772	Nasivin	1	27	25	25	65
74333	Wick MediNait	1	120	43	56	70
22001	ASS Ratiopharm	50	110	80	40	30
93655	Mobilat Gel	1	150	200	50	45

Abbildung 4.1: Übersicht über die Tabellenstruktur

Die Tabelle CUSTOMER_TURNOVER_ON_ARTICLE_LEVEL_DAILY bildet in diesem Beispiel die Faktentabelle, in der Angaben zu den verkauften Produkten wie bestellte,

gelieferte und defekte bzw. nicht verfügbare Mengen enthält. Ebenfalls aufgeführt sind der Apothekeneinkaufspreis, Apothekenverkaufspreis und der Herstellerabgabepreis.
Diese Tabelle hat Verbindungen mit der Dimensionstabelle ARTICLE, in der Angaben zu den mehr als 250.000 Artikel gespeichert sind. Neben den Preisangaben sind auch eher technische Daten wie Größe, Gewicht und Darreichungsform gespeichert. Eine Zuordnung zu den einzelnen Sortimenten wird hier ebenfalls vorgenommen. In der Kundentabelle CUSTOMER sind neben den Adresse des Kunden noch Angaben wie Alter und Geschlecht des Apothekers, Gründungsjahr der Apotheke, Anzahl und Qualifikation der Angestellten, Größe der Verkaufsflächen u.ä. gespeichert. Zusätzlich sind noch zahlreiche Angaben zu den komplizierten Rabatten vorhanden, die dem betreffenden Kunden eingeräumt werden. In der Tabelle DC wird festgehalten, zu welcher Niederlassung, welcher Region und welchem Land dieser Kunde gehört. Über die Dimensionstabelle TIME kann die Zuordnung der Verkaufszahlen zu dem jeweiligen Jahr, Quartal oder Woche entnommen werden.

4.2.3. Verdichtung der Daten

Da die Faktentabelle mittlerweile ein Volumen von über zweihundert Gigabyte aufweist, wird diese nach unterschiedlichen Gesichtspunkten verdichtet, wodurch eine schnellere Verarbeitung der im Unternehmen bestehenden Auswertungen und Analysen erreicht wird. In der Abbildung 4.2 sind die verschiedenen Verdichtungsebenen zu sehen, die auf die Faktentabelle angewendet werden.

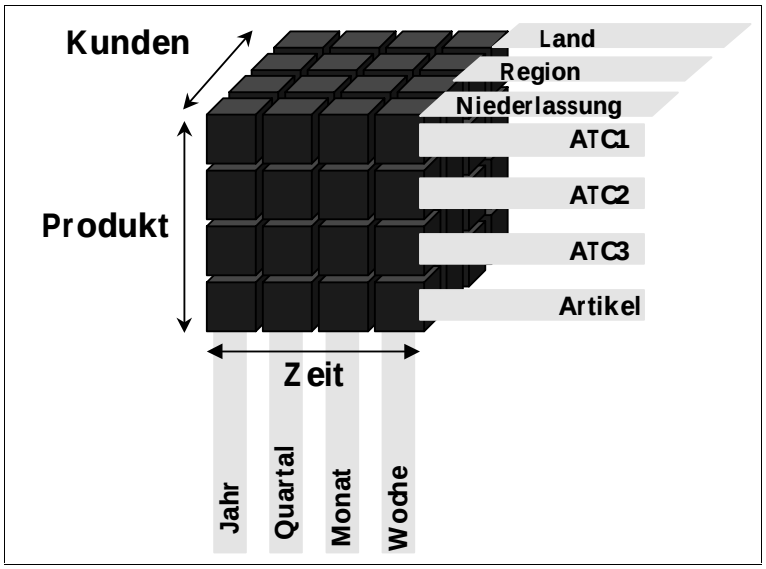


Abbildung 4.2: Verdichtungsebenen des Data Warehouse

Auf der Zeitachse werden die Daten jeweils auf Jahres-, Quartals-, Monats-, und Wochenebene verdichtet. Bei den Produkten bildet die Artikeldimension die unterste Ebene. Darauf aufbauend erfolgt eine Verdichtung anhand von Warengruppen, die sich an einem bestimmten Schlüssel, dem ATC-Schlüssel orientieren.
Dieser ATC-Schlüssel wurde von der Weltgesundheitsorganisation WHO aufgestellt und nimmt eine Aufteilung der Artikel nach eher anatomischen bzw. therapeutischen Gesichtspunkten vor. Die Warengruppen liegen dabei in 3 Detaillierungsstufen vor: Auf der höchsten Ebene, auch ATC1 genannt, existieren 30 verschiedene Warengruppen. Auf der

nächsten Ebene, der ATC2, werden diese in 230 verschiedene Gruppen unterteilt, während es auf der niedrigsten Verdichtungsstufe, der ATC3, 959 verschieden Warengruppen gibt. In der Kundendimension schließlich wird auf den Ebenen der Niederlassungen, Regionen und Länder verdichtet.

4.2.4. Daten der IMS

Bei der IMS (**I**nstitut für **M**edizinische **S**tatistik) handelt es sich um einen externen Dienstleister, der Abverkaufdaten des gesamten pharmazeutischen Marktes in Deutschland sammelt, aufbereitet und kommerziell vertreibt. Zu diesem Zweck wird bei den Apotheken eine geografische Zuordnung in sogenannte RPM-Kreise vorgenommen (RPM steht für **R**egionaler **P**harmazeutischer **M**arkt). Das Bundesgebiet ist hierzu in ungefähr 1.900 dieser Kreise aufgeteilt. Ein RPM-Kreis setzt sich aus mehreren benachbarten Postleitzahlengebieten zusammen, wobei zu einem Kreis mindestens 4-5 Apotheken gehören müssen. Vom pharmazeutischen Großhandel werden der IMS nun die Abverkaufszahlen in den einzelnen RPM-Kreisen, zusätzlich aufgeschlüsselt nach den ATC-Warengruppen, gemeldet.
Von der IMS werden aber auch Daten zur Verfügung gestellt. Beispielsweise werden demografische Daten zu den einzelnen RPM-Kreisen übermittelt. Diese Daten liegen jeweils für ein Kalenderjahr vor. Darin enthalten sind u.a. folgende Angaben:

- Einwohner im RPM-Kreis
- Anzahl der weiblichen Einwohner
- Anzahl der männlichen Einwohner
- Anzahl der Apotheken im RPM-Kreis
- Anzahl der niedergelassenen Ärzte
- Anzahl der Krankenhäuser
- Anzahl der Planbetten in den Krankenhäusern

Bei der Anzahl der Ärzte in den einzelnen RPM-Kreisen findet zusätzlich eine Aufschlüsselung der Ärzte nach Fachbereichen wie Internisten, Neurologen, Orthopäden usw. statt.

4.3. Einsatzgebiete für Data Mining bei der Gehe

Um konkrete Anwendungsgebiete für das Data Mining bei der Gehe zu erhalten, wurden zunächst die im Unternehmen vorhandenen Probleme und offenen Fragestellungen gesammelt und aufbereitet. Dabei wurde auch das Unternehmensumfeld der Gehe Pharmahandel GmbH und die Besonderheiten des pharmazeutischen Marktes berücksichtigt. In Zusammenarbeit mit den Mitarbeitern aus den Bereichen Marketing und Vertriebsunterstützung wurden

daraufhin vier beispielhafte Einsatzbereiche für Data Mining identifiziert, die im Rahmen dieser Diplomarbeit betrachtet wurden.

Erhöhung des Marktanteils:
Der Markt für pharmazeutische Produkte in Deutschland ist vor allem dadurch gekennzeichnet, dass die Marktanteile der einzelnen Wettbewerber relativ stabil sind. Selbst Verschiebungen um Prozentpunkte können für ein Unternehmen entscheidend sein, da eine Erhöhung des Marktanteils durch Firmenzukäufe aufgrund von kartellrechtlichen Beschränkungen fast unmöglich ist. In diesem Zusammenhang kommt der Frage nach umsatzstarken Kunden eine besondere Bedeutung zu. Eine Identifizierung von zukünftig erfolgreichen bzw. umsatzstarken Apotheken könnte hierzu einen entscheidenden Beitrag leisten. Wenn das Kundenportfolio der Gehe überproportionales Wachstum aufweist, dann wächst auch langfristig der Marktanteil der Gehe. Sind diese Apotheken einmal identifiziert, kann mit einem zielgerichteten Marketing versucht werden, diese Apotheken verstärkt anzusprechen und in den Kundenstamm der Gehe einzugliedern. Bisher ist die Gehe Pharmahandel GmbH bei der Einschätzung des Potentials einer Apotheke auf die eher subjektiven Eindrücke der Vertriebsmitarbeiter vor Ort angewiesen. Diese Einschätzungen könnten durch den Einsatz von Data Mining ergänzt bzw. erweitert werden.

Umsatzmaximierung / Kundenspezifische Ansprache:
Im Rahmen der Marketingaktivitäten der Gehe wird vor allem eine zielgerichtete Kundenansprache verfolgt. Im Zusammenhang mit einer Einteilung der Apotheken stehen bisher allerdings fast ausschließlich monetäre Größen im Vordergrund. Dazu wird eine Kundenklassifizierung mit einer Einteilung in Groß- Mittel- und Kleinkunden, sowie eine Schätzung des Umsatzpotential verwendet. Gerade in Bezug auf Bestrebungen, vertikal verstärkt mit der Industrie in Kooperationen oder Systemen zusammenarbeiten zu wollen, stellt sich aber die Frage nach sortiments- oder indikationsspezifischen Einteilungen der Apotheken. Damit verbunden ist auch die Fragestellung, ob sich aus der Analyse der Bestelldaten der Apotheken Erkenntnisse gewinnen lassen, mit denen eine kundenspezifische Ansprache unterstützt wird.

Wirksamkeit von Dienstleistungsprogrammen:
Betrachtet man die Leistungen der einzelnen Großhändler, so ergeben sich keine wesentlichen Unterschiede. Jeder Großhändler besitzt ein ähnliches Warenangebot, liefert pünktlich, zuverlässig und aufgrund der gesetzlich festgelegten Preisspannenverordnung nahezu zum gleichen Preis. Die Gehe versucht deshalb, über Marketingdienstleistungen dem Kunden einen Zusatznutzen zu bieten und diesen somit langfristig zu binden. Dazu gehört beispielsweise der Onlinedienst POINT, der in Form eines Informations-, Kommunikations- und Dialogsystems den Apotheker bei der Bewältigung seiner täglichen Aufgaben unterstützt. Mit diesem Dienst werden dem Apotheker über das Internet unter anderem aktuelle Informationen über den Markt für pharmazeutische Produkte und die Möglichkeit zur Produktbestellung angeboten. Darüber hinaus wird mit dem Programm Galleria eine Platzierungs- und Präsentationsunterstützung für Produkte des Sichtwahlsortiments in

Apotheken angeboten. Bisher gibt es keine Möglichkeiten, den Erfolg dieser Dienstleistungsprogramme transparent zu machen, beziehungsweise deren Akzeptanz beim Kunden zu überprüfen. Mit Hilfe von Data Mining soll versucht werden, den Erfolg dieser Programme überprüfbarer zu machen und deren Inhalte zielgerichtet auf den Kunden abzustimmen.

Verringerung von Zahlungsausfällen:
Die Gehe Pharmahandel GmbH hat monatlich Zahlungsausfälle in nicht unerheblichem Ausmaß zu verkraften. Diese Ausfälle äußern sich in geplatzten Bankeinzügen bzw. Wechseln. Des weiteren verschaffen sich Apotheker zusätzliche Liquidität durch die Bitte nach Zahlungszielverlängerungen oder Stundungen ihrer Sammelrechnungen. Die Folgen für Gehe sind zusätzliche Zinskosten und ein erhöhtes Zahlungsausfallrisiko. Notleidende Apotheken können erst dann identifiziert werden, wenn deren prekäre wirtschaftliche Situation schon eingetreten ist. Durch die üblichen Zahlungs- und Mahnungsprozedere im pharmazeutischen Großhandel können pro Apotheke leicht 300.000 – 500.000 DM Forderungen auflaufen, bevor die Warenlieferungen eingestellt werden. Auch die Übereignung des Warenlagers deckt diesen Betrag nur noch zu einem Teil.
Es bestehen zur Zeit nur eingeschränkte Möglichkeiten, potentiell gefährdete Apotheken zu identifizieren, um eine mögliche Notsituation mit auftretenden Zahlungsschwierigkeiten im Vorfeld zu erkennen oder zu erahnen. Die Gehe reagiert nur auf die eintretenden Wünsche oder Fakten von Apothekenseiten, sie kann nicht agieren oder vorbeugen. Mit Data Mining soll versucht werden, aus den historischen Daten der Kunden ein Modell zu generieren, das als eine Art Frühwarnsystem zur Erkennung von Zahlungsausfällen dienen kann.

4.4. Beschreibung der Business Cases

Die Fragestellungen, die sich hinter den oben vorgestellten Anwendungsgebieten verbergen, wurden in einem nächsten Schritt in konkrete Fragestellungen umgesetzt, den sogenannten Business Cases. In Zusammenarbeit mit Mitarbeitern aus dem IT-Bereich und den Abteilungen Marketing und Vertriebsunterstützung wurden für jedes Anwendungsgebiet ein oder zwei Business Cases definiert, die sich für eine Implementierung im Rahmen einer Diplomarbeit eigneten. Diese Zuordnung kann aus der folgenden Abbildung entnommen werden.

Abbildung 4.3: Zuordnung Anwendungsgebiete – Business Cases

Im folgenden wird jeder Business Case mit Angaben zu seiner Zielsetzung, dem potentiellen Nutzen für die Gehe und einer groben Skizzierung der Daten, die hierzu benötigt werden, vorgestellt.

4.4.1. Potentialermittlung von Kunden aus demografischen Merkmalen

Ziel:

Es soll geprüft werden, ob sich aus den Umsatzzahlen der Apotheken, die zum Kundenstamm der Gehe Pharmahandel GmbH gehören und den demografischen Daten, die von der IMS bereitgestellt werden, Indikatoren erkennen lassen, mit denen der Umsatz einer Apotheke erklärt werden kann. Beispielsweise könnte die Einwohnerzahl und die Anzahl der Ärzte oder Krankenhäuser in der Umgebung einer Apotheke den Umsatz nachhaltig beeinflussen. Das Ziel besteht darin, anhand dieser demografischen Daten ein Modell zu generieren, mit dem Apotheken hinsichtlich ihres Umsatzpotentials eingeschätzt werden können. Dieses Modell kann auf Apotheken angewandt werden, die bisher noch nicht zum Kundenstamm der Gehe gehören oder mit denen die Gehe nur geringe Umsätzen tätigt. Die jährliche Aktualisierung der demografischen Daten kann auch zur Verifizierung des gefundenen Modells benutzt werden. So kann die Gültigkeit des Modells das aufgrund der Daten des Jahres 1998 generiert wurde, anhand der Daten des Jahres 1999 überprüft werden.

Nutzen:

Durch die gefundenen Regeln sollen die eher subjektive Potentialschätzungen durch den Vertrieb abgelöst oder ergänzt werden. Es besteht dann die Möglichkeit, besonders umsatzstarke Apotheken zu identifizieren, bei denen die Gehe bisher nur einen kleinen Anteil an den Warenlieferungen besitzt, oder die bisher ausschließlich von Konkurrenten beliefert werden. Mit einem zielgerichteten Marketing kann versucht werden, diese Apotheken verstärkt anzusprechen und in den Kundenstamm der Gehe einzugliedern.

Benötigte Daten:

Den Ausgangspunkt für diese Analyse bilden die Umsatzzahlen der Apotheken, die zum Kundenstamm der Gehe Pharmahandel GmbH gehören. Diese liegen im Data Warehouse auf einer monatlichen Basis in der Tabelle CUST_TURNOVER_MT vor. Daraus muss der Gesamtumsatz eines Kunden berechnet werden, der die Zielgröße dieser Analyse darstellt. In den demografischen Daten, die von der IMS bereitgestellt werden, liegen Angaben zu der Einwohnerstruktur und Angaben zu den Krankenhäusern innerhalb eines RPM-Kreises jeweils für ein Kalenderjahr vor. Jede Apotheke kann eindeutig einem RPM-Kreis und damit den entsprechenden demografischen Daten zugeordnet werden.

4.4.2. Entdeckung erfolgreicher Kunden aus kundenspezifischen Merkmalen

Ziel:

Im Data Warehouse der Gehe Pharmahandel GmbH wird zu jedem Kunden eine große Anzahl von Merkmalen wie Alter des Apothekers, Größe der Verkaufsflächen, Anzahl und Qualifikation der Angestellten, Lage der Apotheke u.ä. festgehalten. Aus diesen kundenbezogenen Merkmalen einer Apotheke soll herausgefunden werden, ob sich hierin Indikatoren für einen hohen Gesamtumsatz oder eine wie auch immer geartete Umsatzentwicklung finden lassen. Neben dem Gesamtumsatz ist hier auch der Einfluss der kundenspezifischen Merkmale auf die Umsätze in den einzelnen Warengruppen von besonderem Interesse. Aus den möglichen Einflussfaktoren sollen diejenigen identifiziert werden, die den stärksten Einfluss auf den Umsatz haben.

Nutzen:

Die gefundenen Merkmale können dazu dienen, im Rahmen bestehender Dienstleistungsangebote Empfehlungen oder Hinweise für den Apotheker auszusprechen, was die Gestaltung seiner Verkaufsflächen oder seinen Personalpolitik betrifft. Auf der anderen Seite können Regeln, die Zusammenhänge zwischen der Lage einer Apotheke und dem Umsatz herstellen, zum Überprüfen des Kundenportfolios und zur Potentialermittlung von neuen Kunden benutzt werden.

Benötigte Daten:

Den Ausgangspunkt bildet die Tabelle mit den Kundenstammdaten, in der alle Apotheken in Deutschland gespeichert sind. Zu jeder Apotheke, die zu den Kunden der Gehe zählt, sind kundenspezifische Merkmale abgelegt, wobei der Füllgrad und der Pflegezustand der Daten je nach Niederlassung und Region variiert. Für einen ersten Ansatzpunkt wird wie im obigen Fall der Gesamtumsatz als Zielvariable benötigt, der für die Überprüfung des Einflusses auf die Umsätze in den einzelnen Warengruppen auch auf dieser Ebene dem Data Warehouse entnommen werden kann.

4.4.3. Sortimentszusammensetzung

Ziel:

Die Fragestellung, ob eine Segmentierung von Apotheken einen Beitrag zur zielgerichteten Kundenansprache leisten kann, ist das Ziel dieses Business Case. Es soll überprüft werden, ob sich Gruppen von Apotheken identifizieren lassen, die bestimmte Sortimentsschwerpunkte besitzen. Zur Ermittlung dieser Schwerpunkte können einerseits geografische Faktoren, wie beispielsweise die Zugehörigkeit eines Kunden zu einer bestimmten Niederlassung oder Region herangezogen werden. Andererseits soll bei dieser Analyse auch der Einfluss der demografischen Daten, wie beispielsweise Informationen über die Einwohner- oder Ärztestruktur ermittelt werden. Von besonderen Interesse ist ebenfalls, ob sich saisonale Schwerpunkten erkennen lassen.

Nutzen:
Mit den Ergebnissen der Sortimentszusammensetzung können Apotheken identifiziert werden, die sich bestimmten Themengebieten verschrieben haben. Bei gemeinsamen Aktionen mit der Pharma-Industrie können diese Apotheken-Zielgruppen artikel- oder indikationsspezifisch angesprochen werden. Denkbar wären Aktionspakete, die speziell auf die einzelnen Apothekengruppen zugeschnitten sind, so dass mit verbesserten Konditionen oder kundenspezifischen Marketing-Aktionen eine höhere Kundenbindung erreicht wird.

Benötigte Daten:
Im Data Warehouse liegen die Abverkaufsdaten pro Apotheke für die letzten 24 Monate vor. Diese können nach dem oben vorgestellten ATC-Warengruppenschlüssel selektiert werden, der im Data Warehouse in den Detaillierungsstufen ATC1 und ATC3 vorliegt. Die Tabellen SALES_CUS_ATC1_MT und SALES_CUS_ATC3_MT enthalten die Umsätze der Apotheken in den entsprechenden Warengruppen. Diese Umsätze liegen dabei monatsweise vor, so dass hierauf eine Betrachtung nach saisonalen Schwerpunkten erfolgen kann. Die Zuordnung zu den demografischen Daten erfolgt über den RPM-Kreis, in dem die Apotheke liegt, während die Zugehörigkeit zu einer Niederlassung bzw. Region aus der Kundentabelle entnommen werden kann.

4.4.4. Warenkorbanalyse

Ziel:
Bei der Warenkorbanalyse steht die Frage im Vordergrund, ob sich aus dem Bestellverhalten der Apotheker Muster oder Regeln erkennen lassen. Dies kann auf der Ebene der Artikel geschehen, d.h. welche Artikel werden häufig gemeinsam bestellt. Die gleiche Analyse kann aber auch auf der Ebene der Warengruppen erfolgen. Betrachtet werden hierbei ausschließlich Artikel im Griffwahl- oder Sichtwahlbereich, da aufgrund der Preisspannenverordnung bei den verschreibungspflichtigen Artikeln keinerlei Spielraum bei der Preisgestaltung von Seiten der Gehe besteht.

Nutzen:
Lassen sich aus dieser Analyse Erkenntnisse über gemeinsam bestellte Artikel oder Warengruppen gewinnen, so kann der Nutzen für die Gehe in der Bildung von Aktionspaketen für das Marketing und den Zentraleinkauf bestehen. Für Sonderangebote und

gezielten Preisreduzierungen bzw. Preiserhöhungen können neue Impulse gewonnen werden. Diese neuen Erkenntnisse können auch zur Optimierung der Lager in den Niederlassungen eingesetzt werden, indem häufig gemeinsam bestellte Artikel in unmittelbarere Nähe platziert werden, um so die Lagerdurchgangsrate zu senken.

Benötigte Daten:
Für diesen Business Case ist der Zugriff auf die Auftragsdaten notwendig, in der die Bestellungen aller Kunden der Gehe der letzen zwei Jahre vorliegen. Für die Assoziationsanalyse auf der Artikelebene wird die Bestellnummer und die Artikelnummer aus dieser Tabelle benötigt. Eine Zuordnung zu den Artikeln im Griff- oder Sichtwahlbereich ist über die Angaben in der Artikeltabelle möglich. Bei einer Analyse auf der Ebene der Warengruppen wird statt der Artikelnummer die Bezeichnung der entsprechenden ATC-Warengruppe verwendet, zu der der Artikel gehört. Dazu wird die Warengruppe mit der höchsten Detaillierungsstufe, die ATC3-Warengruppe verwendet.

4.4.5. Zielgruppenidentifizierung für den Online-Dienst POINT

Ziel:
Mit dieser Analyse soll der Versuch unternommen werden, innerhalb der Kunden des POINT-Systems Benutzergruppen zu identifizieren, die aufgrund ihres Navigationsverhaltens Ähnlichkeiten aufweisen. Von Interesse ist dabei, ob es Benutzergruppen gibt, welche das POINT-System eher als Informationsmedium oder als Bestellsystem benutzen und ob sich aus der Anwahl der Seiten Themenschwerpunkte erkennen lassen. Neben der Art der Nutzung soll auch festgehalten werden, ob sich aus der Gesamtgruppe der POINT-Nutzer Merkmale erkennen lassen, mit denen sich POINT-Nutzer von Nicht-POINT-Nutzern unterscheiden lassen.

Nutzen:
Die gefundenen Erkenntnisse können dazu benutzt werden, Inhalte und Navigation besser auf die Kundenbedürfnisse einzustellen, um so einen Zusatznutzen für den Apotheker zu schaffen. Für den Fall, dass sich Merkmale zur Erkennung von POINT-Nutzern identifizieren lassen, können diese dazu benutzt werden, um im übrigen Kundenportfolio die potentiellen POINT-Nutzer besser identifizieren zu können.

Benötigte Daten:
Für diesen Business Case werden Daten über das Navigationsverhalten der Kunden innerhalb von POINT benötigt. Für die Aufzeichnung dieser Daten müssen allerdings noch die Voraussetzungen auf der Seite des POINT-Systems geschaffen werden, um beispielsweise die Häufigkeit der Anmeldungen pro Tag oder den Zeitraum zwischen der Anmeldung und Abmeldung im System zu ermitteln. Mögliche Einflussfaktoren wären auch die Anzahl der angeforderten Informations- oder Bestellseiten, die Verweildauer auf diesen Seiten oder kundenspezifische Merkmale wie Alter und Geschlecht des Apothekers und seiner Angestellten oder das Gründungsdatum der Apotheke.

4.4.6. Erfolgsfaktoren für den Online-Dienst POINT

Ziel:

Mit Hilfe dieses Business Case soll überprüft werden, ob sich Faktoren finden lassen, die den Wert einer Bestellung über das POINT-System beeinflussen und wie stark diese Auswirkungen sind. Von besonderem Interesse ist dabei, ob sich aus den thematischen Schwerpunkten der Seiten, die ein Benutzer besucht, Präferenzen für bestimmte Sortimente oder Warengruppen ableiten lassen. Mögliche Einflussfaktoren hierfür wären beispielsweise, welche Seiten der Benutzer wie oft besucht hat und wie lange er sich insgesamt im System bewegt.

Nutzen:

Die gewonnen Erkenntnisse können dazu dienen, den Aufbau des POINT-Systems zu optimieren. Sind die Faktoren einmal identifiziert, die zu einer erhöhten Umsatz auf bestimmten Bestellseiten bewirken, wird das Benutzerverhalten wesentlich transparenter gemacht. Dadurch lässt sich der Aufbau und der Inhalt der Seiten besser an die Wünsche und Bedürfnisse der Benutzer anpassen, so dass die Akzeptanz des Online-Dienstes gesteigert wird.

Benötigte Daten:

Grundlage dieser Analyse sind die Umsatzzahlen, die mit dem POINT-System erreicht werden. Als Zielvariable kann entweder der Gesamtumsatz eines Kunden oder der Umsatz in den einzelnen Warengruppen oder Sortimenten genommen werden. Genau wie bei dem vorherigen Business Case werden allerdings Daten über das Benutzungsverhalten des POINT-Systems benötigt. Hierzu müssen die Voraussetzungen geschaffen werden, um die Anzahl der Mausclicks eines Benutzers auf die einzelnen Seiten und die Verweildauer im Gesamtsystem aufzuzeichnen. Zusätzlich wird auch eine Einteilung der Webseiten anhand von bestimmten Themenschwerpunkten benötigt.

4.4.7. Einflussfaktoren für ausstehende Zahlungen

Ziel:

Aus der Analyse der Daten von Apotheken, die in der Vergangenheit in eine Zahlungsunfähigkeitssituation geraten sind, sollen typische, vergleichbare Muster herausgefiltert werden. Dieses kann helfen, für die Zukunft eine Art Frühwarnindikator zu entwickeln. Dieser kann permanent über den Gesamtdatenbestand wachen und im Falle des Eintretens dieses Indikators eine Warnmeldung an entsprechender Stelle tätigen.

Nutzen:

Die Ergebnisse dieser Analyse können für die Gehe eine Reduktion der Zahlungsausfälle und damit eine direkte Kosteneinsparung bedeuten. Für den Vertrieb würden sich in diesen Fällen unterschiedlichste Möglichkeiten bieten wie beispielsweise erhöhte Aufmerksamkeit, vorbeugende Kundenbesuche, Anpassung von Zahlungsmodalitäten oder die Festlegung einer Auftragshöchstgrenze.

Benötigte Daten:
Die für diesen Business Case wesentlichen Daten zu Zahlungsströmen kommen aus dem Buchhaltungssystem welches zumindest zur Zeit noch nicht in das Data Warehouse integriert ist. Dennoch können diese Daten bereitgestellt werden. In diesem Buchhaltungssystem ist bei jeden Kunden ein sogenannter Zahlungsverhaltensindex gespeichert, der (als relativ abstrakter numerischer Wert) die Zahlungsmoral des jeweiligen Kunden angibt. Er ist abhängig von der Höhe der offenen Posten und der Zeitdauer, die bis zu deren Begleichung verstreicht.

4.4.8. Zuordnung der Business Cases zu Data Mining-Methoden

Aus den Zielen der einzelnen Business Cases lässt sich eine erste Zuordnung zu den in Abschnitt 2.4.2 vorgestellten Data Mining Methoden erkennen, was in der folgenden Abbildung dargestellt ist.

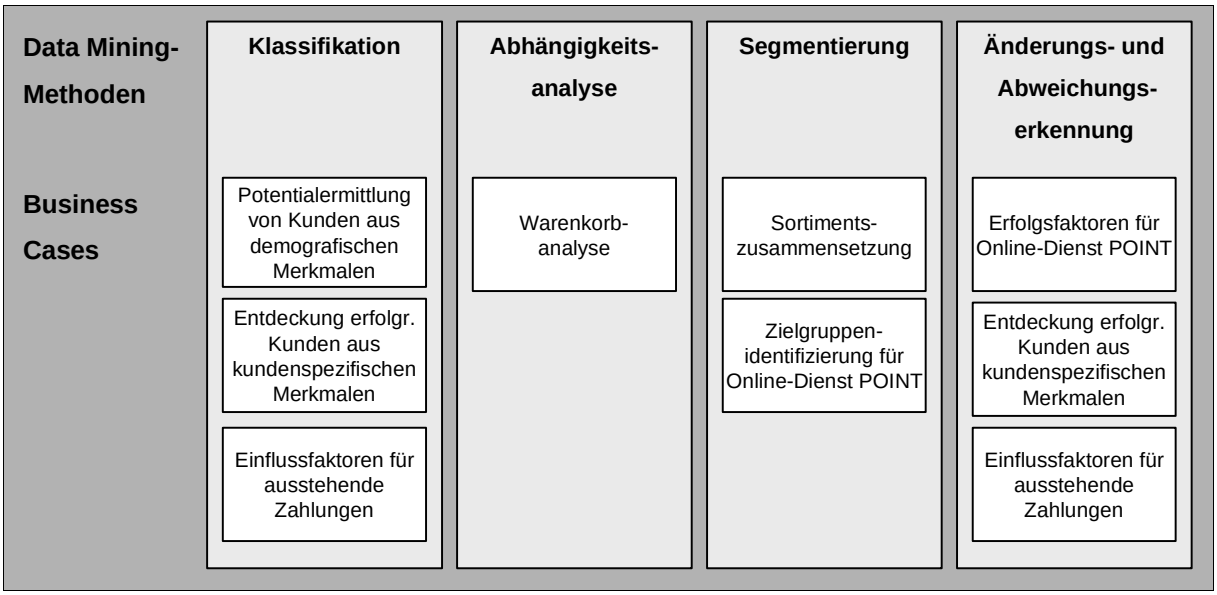


Abbildung 4.4 : Zuordnung Business Cases zu Data Mining-Methoden

Bei einigen Business Cases ist diese Zuordnung eindeutig, wie bei der Warenkorbanalyse, die ganz klar mit einer Abhängigkeitsanalyse durchgeführt wird. Bei der Potentialenmittlung von Kunden aus den demografischen Merkmalen soll ein Modell erstellt werden, dass den Gesamtumsatz eines Kunden vorhersagt, was eindeutig als eine Klassifikationsaufgabe interpretiert werden kann. Sowohl bei der Sortimentszusammensetzung als auch bei der Zielgruppenidentifizierung wird eine Gruppenbildung angestrebt, was mit der Methode der Segmentierung erreicht wird. Bei den Erfolgsfaktoren für den Online-Dienst POINT wird nach der Stärke der einzelnen Faktoren auf den Wert einer Bestellung gesucht, was mit den klassischen Verfahren der Statistik, beispielsweise der Regressionsanalyse erreicht wird.
Bei den Einflussfaktoren für ausstehende Zahlungen und der Entdeckung erfolgreicher Kunden aus den kundenspezifischen Merkmalen ist diese Zuordnung etwas schwieriger. Aufgrund ihres vorhersagenden Charakters können beide als Klassifikationsaufgabe gedeutet werden. Allerdings kann die Stärke eines Zusammenhangs zwischen den kundenspezifischen Merkmalen und dem Gesamtumsatz auch mit den statistischen Verfahren bei einer

Änderungs- und Abweichungserkennung ermittelt werden. Bei den Einflussfaktoren für ausstehende Zahlungen kann auch eine Zeitreihenanalyse durchgeführt werden, mit der sich im Zahlungsverhalten von Kunden ähnliche Sequenzen entdecken lassen, die zu einem Zahlungsausfall führen. Auf jeden Fall wird die oben dargestellte Zuordnung dazu benötigt, um einen ersten Anhaltspunkt zur Auswahl der richtigen Data Mining-Methode zu liefern

4.5. Auswahl der Business Cases zur Implementierung

Aus den im vorigen Abschnitt vorgestellten Business Cases wurden insgesamt drei ausgewählt, die in einer Implementierungsphase mit den zur Verfügung stehenden Data Mining-Tools bearbeitet wurden. Diese Vorgehensweise wird im nächsten Kapitel ausführlich beschrieben. Kriterien für die Auswahl waren einerseits die Verfügbarkeit der Daten, d.h. wie leicht die Daten, die für die Analyse benötigt werden, aus dem Data Warehouse extrahiert werden können. Auf der anderen Seite basierte die in Kapitel 3 beschriebene Evaluierung auch auf den Erfahrungen, die bei dieser Implementierung gemacht wurden, weshalb eine möglichst breite Methodenabdeckung wünschenswert war. Bei den folgenden drei Business Cases waren diese Kriterien erfüllt:

- Warenkorbanalyse
- Sortimentszusammensetzung
- Potentialermittlung von Kunden aus demografischen Merkmalen

Damit wurde eine Abdeckung der Data Mining-Methoden Abhängigkeitsanalyse, Segmentierung und Klassifikation erreicht. Zudem liefern diese Business Cases durch ihre eher allgemeinen Fragestellungen grundlegende Hinweise, was den Einsatz von Data Mining bei der Gehe betrifft.

Gegen die Auswahl der anderen Business Cases sprachen verschiedene Gründe: Bei der Entdeckung erfolgreicher Kunden aus kundenspezifischen Merkmalen bestand die Problematik in dem Füllgrad und dem Pflegezustand der betreffenden Merkmale, die von jeder Niederlassung übermittelt werden. Je nach Niederlassung waren diese Felder teilweise nur zu 20% gefüllt oder die Daten waren veraltet. Bei den Business Cases, die auf dem POINT-Systems aufsetzen, müssen zunächst die Voraussetzungen geschaffen werden, um Informationen über das Navigationsverhalten der POINT-Nutzer aufzuzeichnen. Zusätzlich wurde eine Einteilung der Web-Seiten in bestimmte Kategorien und eine Berechnung der Verweildauer der Benutzer auf diesen Seiten benötigt, was wegen des hohen zeitlichen Aufwandes im Rahmen der Diplomarbeit nicht zu realisieren war. Bei der Entdeckung von Zahlungsausfällen bestand die Problematik darin, dass die Daten des Buchhaltungssystem momentan noch nicht in das Data Warehouse integriert sind, so dass für die Bereitstellung der Daten ein unverhältnismäßig hoher Zeitaufwand erforderlich gewesen wäre.

Kapitel 5. Implementierung der Business Cases

Als Ergebnis des vorigen Abschnitts wurden drei Business Cases ausgewählt, die sich für eine Implementierung mit den vorhandenen Data Mining-Tools eigneten. In dem folgenden Kapitel soll nun beschrieben werden, welche Vorgehensweise bei der Implementierung dieser Business Cases gewählt wurde und welche Schlussfolgerungen aus den Ergebnissen zu ziehen sind. Zunächst wird beschrieben, welche grundlegenden Maßnahmen zur Reinigung des Kundendatenbestandes notwendig waren, die den Ausgangspunkt für die folgenden Analysen bildete. Anschließend werden für jeden Business Case die jeweiligen Implementierungsschritte näher erläutert und welche Probleme oder Besonderheiten bei deren Bearbeitung auftraten. Zu jedem Business Case werden auch die Ergebnisse skizziert, die sich aus dieser Implementierung ableiten ließen.

5.1. Reinigung der Kundendaten

Die Tabelle mit den Kundendaten bildet den Ausgangspunkt aller nun folgenden Analysen. In einem ersten Schritt wurde diese einer näheren Betrachtung unterzogen. Wie schon in Abschnitt 4.2.3 beschrieben, sind in dieser Tabelle zu einer Kundennummer nähere Angaben zu der Apotheke und deren Besitzer gespeichert. Da diese Daten aus den operativen Systemen der Niederlassungen stammen, befinden sich hier Datenbestände, die um einer Ergebnisverfälschung vorzubeugen, von der Analyse ausgeschlossen werden müssen:

- *Einrichterkunden*: Wird eine Apotheke oder Drogerie neu eröffnet, so bekommt sie von ihrem Lieferanten einen Grundbestand an Waren übereignet. Die dabei erzielten Umsätze werden auf eine separate Kundennummer, die sogenannte Einrichterkundennummer gebucht. Da diesen hohen, punktuellen Umsätze nicht den Ablauf des üblichen Tagesgeschäftes wiederspiegeln, wurden diese Kundennummern aus der Analyse entfernt.

- *Testkunden*: Dabei handelt es sich um Kunden, die zu Testzwecken in den operativen Systemen der Niederlassungen angelegt wurden. Um zu verhindern, dass eventuell vorhandene Umsätze dieser Kunden in die Ergebnisse einfließen wurden diese Testkunden entfernt. Dies konnte über den Ausschluss bestimmter Niederlassungsbereiche und die manuelle Suche nach leeren Feldinhalten beim Apothekename und Adressangaben erreicht werden.

- *Spezialkunden*: Dabei handelt es sich einerseits um Kunden im Ausland oder um Kunden, die nur selten Bestellungen tätigen, wie beispielsweise die Wilhema, dem Zoologisch-Botanischen Garten in Stuttgart. Da das Bestellverhalten dieser Kunden sicherlich von der Norm abweicht, wurden sie ebenfalls ausgeschlossen. Zur Entdeckung dieser Datenbestände konnten ungültige Einträge bei der Nummer des RPM-Kreises benutzt werden.

- *Ausschluss von Kundengruppen*: Es wurden nur Verkäufe an Apotheken betrachtet. Warenlieferungen an Drogerien wurden von der Analyse ausgeschlossen, da bei diesen eine Einschränkung auf nicht apothekenpflichtige Artikel besteht. Warenlieferungen zwischen den Niederlassungen und Verkäufe an Mitarbeiter und Privatpersonen wurden ebenfalls ausgeschlossen, da diese keinen Beitrag zu einem repräsentativen Gesamtbild leisten.

Die auf diese Weise gereinigte Kundentabelle wurde als Ausgangspunkt für die Implementierung der Business Cases benützt.

5.2. Implementierung der Sortimentszusammensetzung

Bei der Umsetzung dieses Business Cases steht im Vordergrund, dass mehrere Iterationen notwendig waren, um zu verwertbaren Ergebnissen zu gelangen. Diese verschiedenen Iterationen werden im folgenden beschrieben. Den ersten Ausgangspunkt für diese Analyse bildete die Vermutung, dass zwischen der Anzahl der Ärzte in der Umgebung einer Apotheke und den Umsatzzahlen dieser Apotheke in bestimmten Warengruppen ein Zusammenhang besteht. Da sich diese Vermutung nicht bestätigte, wurde ein Zurückgehen in die Phase der Datenselektion vollzogen, indem die Analyse auf eine bestimmte Niederlassung und eine Warengruppe mit einer geringeren Detaillierungsstufe eingeschränkt wurde. Aufbauend auf diesen Erkenntnissen konnten im einem nächsten Schritt saisonale Schwerpunkte bei den Sortimenten identifiziert werden.
Zunächst werden diese drei Iterationen einer näheren Betrachtung unterzogen. Anschließend erfolgt eine Zusammenfassung und Interpretation der Ergebnisse.

5.2.1. Analyse anhand der demografischen Daten

Das Ziel bei diesem Vorgehen bestand in der Überprüfung, ob sich aus den Umsätzen der Kunden in den einzelnen Warengruppen und der Anzahl der Fachärzte des RPM-Kreises, in dem diese Kunden liegen, einen Zusammenhang herstellen lässt. So könnte beispielsweise eine hohe Anzahl von Augenärzten in einem RPM-Kreis einen erhöhten Umsatz in bestimmten Warengruppen zur Folge haben. Diese Erkenntnisse können eine Basis für eine kundenspezifische Ansprache liefern.
Benötigt wurden hierzu der durchschnittliche jährliche Umsatz der Kunden in den ATC3-Warengruppen im Kalenderjahr 1998, da die vorhandenen demografischen Daten sich auf diesen Zeitraum bezogen. Die Durchschnittsbildung wurde vorgenommen, um die Bandbreite der numerischen Werte bei der Ergebnisdarstellung einzuschränken. In der Tabelle SALES_CUS_ATC3_MT im Data Warehouse liegen die monatlichen Umsätze der Kunden auf dieser Warengruppenebene vor, woraus der durchschnittliche Jahresumsatz pro Kunde errechnet werden konnte. In Abbildung 5.1 ist im linken Teil ein Ausschnitt aus dieser Tabelle dargestellt. Damit sich bei der grafischen Ergebnisdarstellung einer Segmentierung

Schwerpunkte in den Umsätzen der Warengruppen besser erkennen lassen, wurde allerdings für jede Warengruppe eine eigene Spalte benötigt, was im rechten Teil der Abbildung zu sehen ist.

CUSTOMER_ID	IMS_ATC3_DESC	SALES
49810001	BETA-Blocker	1.276,78
49810001	Antibiotika	23.412,56
49810112	Antitussiva	2.896,45
49810112	Antibiotika	12.856,34
49810112	Grippemittel	5.897,55

Transponierung →

CUSTOMER_ID	Beta-Blocker	Antibiotika	Antitussiva
49810001	1.276,78	23.412,56	0
49810112	0	12.856,34	2.896,45
49810156	546,89	1.290,67	0
49820034	567,45	3.786,56	123,34
49820045	0	490,55	0

Abbildung 5.1: Transponierung der Umsatzzahlen

Diese sogenannte Transponierung, also das Umlegen von Zeilen- in Spaltenwerte wurde mit MS Access durchgeführt, da keines der Data Mining-Tools diese Funktionalität anbot. Für jeden Kunden lagen also nun die Umsätze in den Warengruppen in einer Zeile vor. Eine Zuordnung dieser Umsatzzahlen zu den demografischen Daten konnte anhand der Tabelle CUSTOMER über das Feld IMS_DISTRICT_ID vorgenommen werden, in dem die Zugehörigkeit eines Kunden zu dem jeweiligen RPM-Kreis festgehalten wird. Eine schematische Darstellung dieser Zuordnungen ist in Abbildung 5.2 zu sehen.

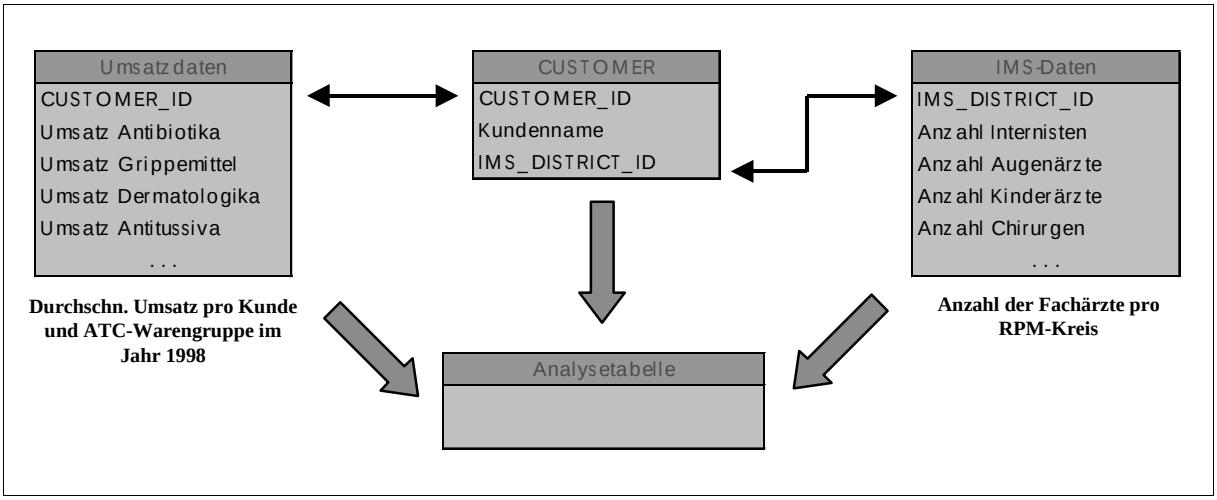


Abbildung 5.2: Zusammenstellung der Analysetabelle zur Sortimentszusammensetzung

Auf der somit erzeugten Analysetabelle wurde die Segmentierung durchgeführt. Es konnte allerdings keine Clusterbildung erreicht werden, aus der sich ein sinnvoller Zusammenhang zwischen der Anzahl der Fachärzte und den Umsätzen in den Warengruppen ergeben hätte. Die gefundenen Cluster setzten sich vielmehr aus Kunden zusammen, die über konstant niedrige bzw. hohe Umsätze in einigen Warengruppen verfügen, ohne dass bei der Anzahl der Fachärzte Schwerpunkte erkennbar waren. Durch die Verwendung von Feldgewichtungen, mit denen den demografischen Daten eine größere Bedeutung bei der Clusterbildung zugemessen wurde, konnten zwar Cluster mit einer hohen bzw. niedrigen Ärztezahl gebildet werden, allerdings ließ sich bei den Umsatzzahlen in den Clustern keinerlei Tendenz zu erhöhten oder geringeren Umsätzen feststellen.

5.2.2. Analyse anhand einer Niederlassung

Aufgrund der Erfahrungen bei der bisherigen Vorgehensweise wurde in einem nächsten Schritt der Betrachtungshorizont sowohl bei den Kunden als auch bei den Warengruppen eingeschränkt, um zunächst zu grundlegenden Aussagen über eine Sortiments-zusammensetzung zu gelangen. In einer erneuten Phase der Datenselektion wurden alle Kunden der Niederlassung Stuttgart mit ihren durchschnittlichen jährliche Umsätzen in den ATC1-Warengruppen aus dem Data Warehouse extrahiert. Die Niederlassung Stuttgart wurde gewählt, da sich durch die Nähe zu den Mitarbeitern die Ergebnisse der Analysen am einfachsten überprüfen ließen. Wie im letzten Abschnitt beschrieben, wurde auch hier eine Transponierung der Daten durchgeführt.
Ein Ausschnitt aus dem Ergebnis der Segmentierung, die mit dem Intelligent Miner von IBM mit dem Verfahren des demografischen Clusterings vorgenommen wurde, ist in der folgenden Abbildung 5.3 zu sehen.

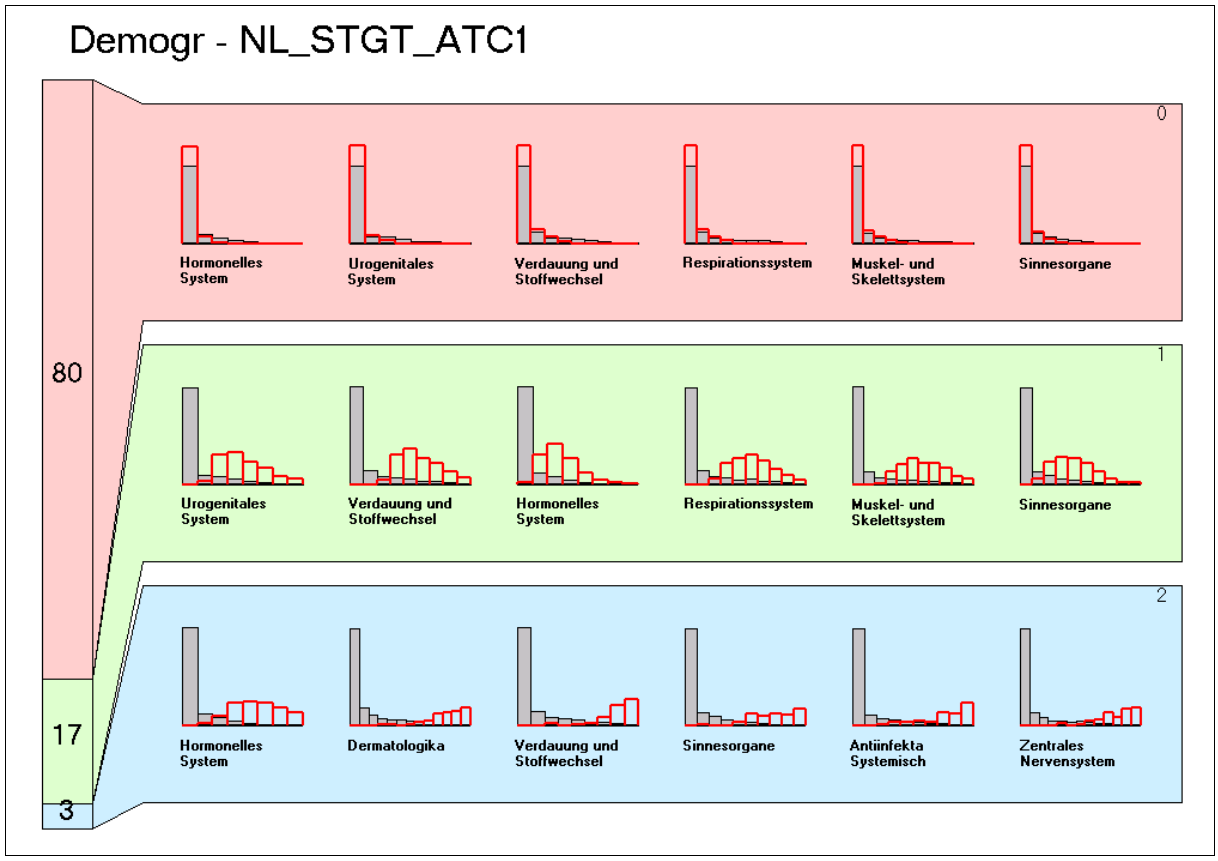


Abbildung 5.3: Kundensegmente der Niederlassung Stuttgart

Es wurden insgesamt drei Cluster identifiziert, in denen jeweils 80% bzw. 17% und 3% der Gesamtzahl der Datensätze eingeordnet wurden. In den Diagrammen sind die Umsätze der Kunden in den einzelnen Warengruppen aufgetragen, wobei die Werteverteilung im Gesamtdatenbestand grau hinterlegt ist. Auf der X-Achse sind die Umsätze von links nach rechts in den Bereichen niedrig bis hoch aufgetragen. Für jede Warengruppe ergibt sich daraus im Gesamtdatenbestand der typische Verlauf mit einer großen Anzahl von Kunden, die Umsätze im niedrigen Bereich aufweisen. Hingegen gibt es nur wenige Kunden mit mittleren und hohen Umsätzen.

Mit der roten Linie wird in der Abbildung die Werteverteilung im jeweiligen Cluster beschrieben. Daraus lässt sich erkennen, dass sich im Cluster 0 vor allem Kunden befinden, deren Umsätze durchgängig durch alle Warengruppen im niedrigen Bereich liegen. Es handelt sich also hierbei um Nebenkunden, die nur sporadisch oder im geringem Umfang bei der Gehe bestellen. Im Cluster 1 finden sich dagegen kaum Kunden mit niedrigen Umsätzen. Hier liegt hier der Schwerpunkt deutlich im mittleren und hohen Bereich. Noch extremer zeigt sich dies im kleinsten Cluster, in dem sich ausschließlich Kunden mit Umsätzen im hohen Bereich befinden. Fasst man die letzten beiden Cluster zusammen, so ergibt sich eine Kundengruppe mit einem Anteil von 20% am Kundenbestand der Niederlassung Stuttgart. Mit diesen sogenannten Hauptkunden tätigt die Gehe einen Großteil des Umsatzes.
Auffällig bei dieser Analyse ist, dass sich keine Schwerpunkte in einzelnen Warengruppen erkennen lassen. Vielmehr bewegen sich die Umsätze durchgängig durch alle Warengruppen im niedrigen bzw. hohen Bereich. Da auf der Ebene der ATC1-Warengruppen lediglich eine Einteilung der Artikel in 30 verschiedene Kategorien vorliegt, wurde in einem nächsten Schritt die Betrachtung auf eine Warengruppe mit einer höheren Detaillierungsstufe durchgeführt.

5.2.3. Analyse auf der Ebene der ATC3-Warengruppen

Den Ausgangspunkt hierfür bildeten die Hauptkunden der Niederlassung Stuttgart, die im Rahmen der vorherigen Analyse identifiziert werden konnten. Da diese Kunden regelmäßig bestellen und ihren gesamten Warenbestand fast ausschließlich von der Gehe beziehen, bilden sie ein repräsentatives Gesamtbild zur Erkennung einer Sortimentszusammensetzung. Für diese Kundengruppe wurden die Umsätze auf der Ebene der ATC3-Warengruppen mit 959 verschiedenen Ausprägungen selektiert. Die Betrachtung erfolgte nicht wie bisher auf dem durchschnittlichen Jahresumsatz, vielmehr wurden die Umsätze auf einer monatlichen Basis bereitgestellt. Damit sollte die Identifikation von eventuell vorhandenen saisonalen Schwerpunkten ermöglicht werden.
Nach der Anwendung des Segmentierungsalgorithmus ließen sich mehrere Cluster identifizieren, von denen im folgenden die zwei interessantesten näher betrachtet werden. In Abbildung 5.4 ist das erste dieser Segmente dargestellt.

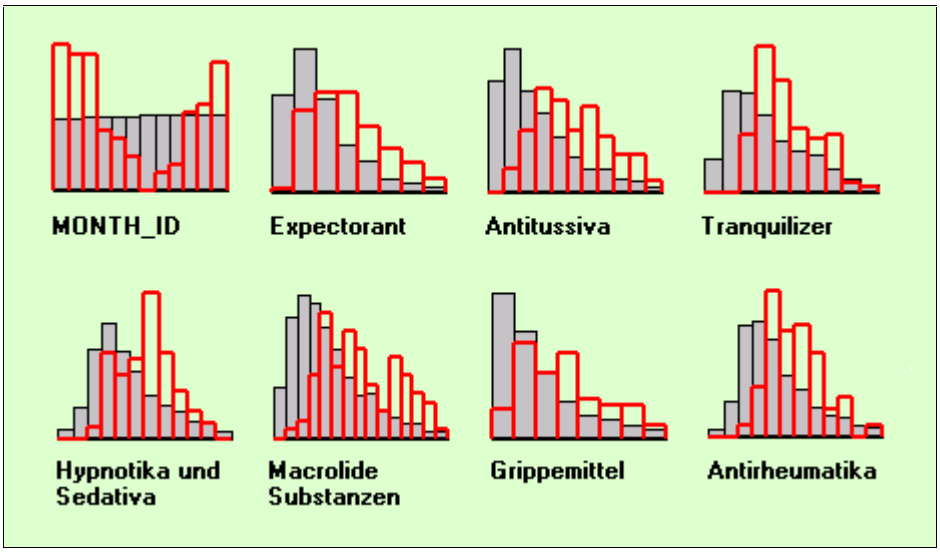


Abbildung 5.4: Sortimentsschwerpunkte in den Wintermonaten

In dem Feld MONTH_ID ist von links nach rechts die Verteilung der Datensätze in den Monaten Januar bis Dezember dargestellt, wobei sich im Gesamtdatenbestand eine Gleichverteilung ergibt. Anhand der roten Linie kann festegestellt werden, das sich in diesem Cluster überproportional viele Umsatzdaten aus den Herbst- und Wintermonaten befinden. Bei den Umsätzen dieser Apotheken in den einzelnen Warengruppen lassen sich durchweg Schwerpunkte im mittleren und hohen Bereich feststellen. Bei den Warengruppen handelt es sich teilweise um Medikamente, die zur Bekämpfung von Erkältungskrankheiten eingesetzt werden wie Grippemittel und Antitussiva, also Hustenmittel. Bei Xpectorant und Macrolide Substanzen handelt es sich um Mittel gegen Infektionskrankheiten. Auch die Warengruppe, hinter der sich Artikel zur Linderung von Rheumabeschwerden verbergen, weist einen erhöhten Umsatz auf. Hypnotika, Sedativa und Tranquilizer werden normalerweise gegen Depressionskrankheiten eingesetzt. Die überproportional hohen Umsätze bei den Erkältungs- und Antirheumamitteln lassen sich durch das gehäufte Auftreten dieser Beschwerden in den Wintermonaten erklären. Bei den Antidepressiva liegt die Vermutung nahe, dass sich dahinter die typische Herbst- oder Winterdepression verbirgt.

Bei dem zweiten Cluster handelt es sich um Umsatzdaten, die schwerpunktmäßig aus den Sommermonaten stammen, was sich aus dem Werteverlauf des Feldes MONTH_ID in der folgenden Abbildung entnehmen lässt.

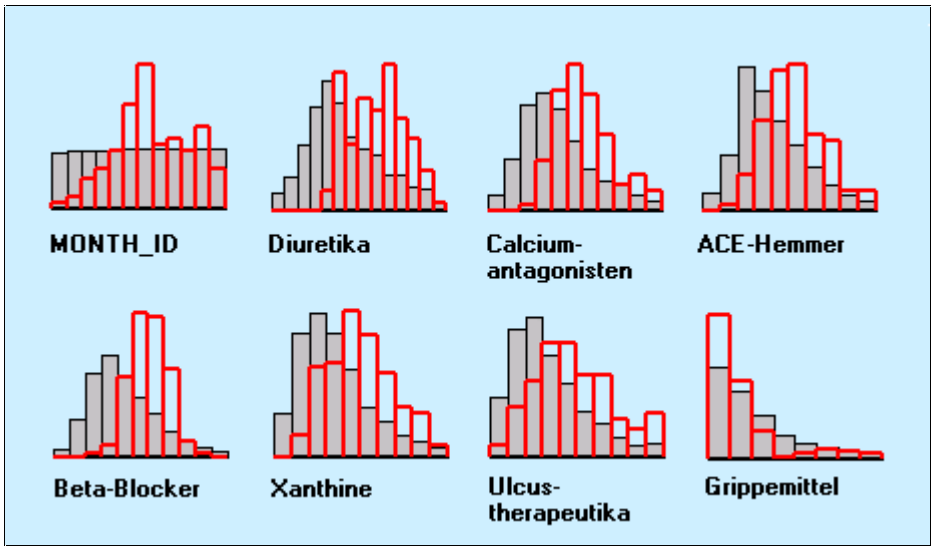


Abbildung 5.5: Sortimentsschwerpunkte in den Sommermonaten

Hier zeigt sich, das der Umsatz bei den Grippemitteln in diesem Zeitraum eher im niedrigen Bereich liegt, was in den Sommermonaten sicherlich kein Überraschung darstellt. Dagegen weisen Mittel gegen Herz- und Kreislauferkrankungen wie Beta-Blocker, Calciumantagonisten und ACE-Hemmer einen deutlich erhöhten Umsatz auf. Bei Diuretika handelt es sich um Mittel gegen Bluthochdruck, während Xanthine zur Behandlung von Asthmaerkrankungen eingesetzt werden. Ulcustherapeutika hingegen werden vor allem gegen Magenerkrankungen und Durchfall eingesetzt. Die erhöhten Umsätze in diesen Warengruppen lassen sich einerseits auf die erhöhten Temperaturen in den Sommermonaten zurückführen, die vor allem bei älteren Menschen Herz- und Kreislaufprobleme auslösen. Auch bei Asthmatikern entsteht in diesem Zeitraum durch die erhöhte Pollen- und Ozonbelastung ein erhöhter Behandlungsbedarf. Hintergrund für die hohen Umsätze bei den Ulcustherapeutika könnte eine erhöhte Reiseaktivität in den Sommermonaten in ferne Länder mit den entsprechenden Folgen für die Urlauber sein.

5.2.4. Interpretation der Ergebnisse

Aufgrund der Ergebnisse, die sich aus der Segmentierung der Umsatzdaten ergaben, kann eine Aufteilung in verschiedene Kundengruppen und das Erkennen von saisonalen Schwerpunkten in den Warensortimenten der Kunden erreicht werden.

Mit einer Analyse auf der Ebene der Niederlassungen und einer kleinen Anzahl von Warengruppen auf der ATC1-Ebene kann eine Aufteilung der Apotheken in Haupt- und Nebenkunden erreicht werden. Dies kann separat für jede Niederlassung durchgeführt werden, so dass die Kundenzusammensetzung von verschiedenen Niederlassungen miteinander verglichen werden kann. Durch Betrachtungen in regelmäßigen Zeitabständen können Veränderungen im Kundenportfolio erkannt und entsprechende Maßnahmen eingeleitet werden. Diese Kundensegmentierung kann auch zur Überprüfung der im Data Warehouse vorhandenen Kundenklassifizierung eingesetzt werden.

Bei den saisonalen Schwerpunkten wurden zwar keine grundlegend neue Erkenntnisse gewonnen, vielmehr waren die Zusammenhänge zwischen den Umsätzen bei den Artikeln und der Jahreszeit aufgrund der Absatzzahlen durchaus bekannt. Es konnten aber Gruppen

von Apotheken gefunden werden, die diesem Trend besonders deutlich folgen. Auf diese Kunden kann mit optimierten saisonalen Angeboten spezifisch eingegangen werden, wodurch eine größere Kundenbindung erreicht werden kann.
Den Hintergrund für die fehlenden Zusammenhänge zwischen den demografische Daten und den Umsätzen in den Warengruppen bildet wahrscheinlich die Tatsache, dass die RPM-Kreise zu großräumig sind. Durch die Aufteilung des Bundesgebietes in die ca. 1.900 Kreise und die Mindestzugehörigkeit von 4-5 Apotheken pro RPM-Kreis entstehen Flächen, die teilweise aus mehreren Postleitzahlengebieten zusammengesetzt sind. Für dieses Gebiet werden dann die Ärztezahlen erhoben. Für die Sortimentsschwerpunkte von Apotheken ist es allerdings eher weniger entscheidend, welche Ärzte sich in welcher Anzahl in einem Umkreis von mehreren Kilometern befinden, sondern eher deren Vorhandensein in der unmittelbaren Umgebung der Apotheke, beispielsweise im gleichen Gebäude. Damit erscheinen die demografischen Daten in der vorliegenden Form wenig geeignet, um eine Sortimentszusammensetzung zu ermitteln.

5.3. Implementierung der Warenkorbanalyse

Die Warenkorbanalyse wurde auf zwei Ebenen durchgeführt. Zum einen auf der Ebene der ATC3-Warengruppen, um mit diesen Zusammenhängen neue Erkenntnisse für Marketing-Aktionen und Angebotszusammenstellungen zu gewinnen. Zum anderen auf Artikelebene, um Auswirkungen auf die Lageranordnung und Logistik zu überprüfen. Im folgenden wird zunächst beschrieben, welcher Datenbestand für die Analyse selektiert wurde und wie die Analysetabellen zusammengestellt wurden. Anschließend werden die Ergebnisse der beiden Analysen dargestellt.

5.3.1. Selektion der Analysedaten

Für eine Warenkorbanalyse werden zwei grundlegende Attribute benötigt. Erstens eine Transaktionsnummer, wie beispielsweise eine Bestell- oder Rechnungsnummer, unter der alle Artikel gebucht werden, die ein Kunde bestellt oder gekauft hat. Zweitens wird eine Bezeichnung für den Artikel benötigt, beispielsweise eine Produktnummer oder ein Produktname. Anhand dieser zwei Attribute werden häufig gemeinsam bestellte oder gekaufte Artikel vom Algorithmus berechnet.
In dem vorliegenden Fall wird hierzu der Zugriff auf die Auftragsdaten benötigt. Hier sind alle Bestellungen der Apotheker der letzten zwei Jahre unter Angabe einer Bestellnummer und eines Artikelnamens, bzw. der Bezeichnung der entsprechenden Warengruppe gespeichert. Da sich in dieser Tabelle mittlerweile mehrere Millionen Datensätze befinden, wurde die zu analysierende Datenmenge eingeschränkt, da mit den vorliegenden Testinstallationen der Tools eine solche Datenmenge nicht bearbeitet werden konnte. In Abbildung 5.6 ist eine schematische Darstellung dieser Einschränkungen zu sehen.

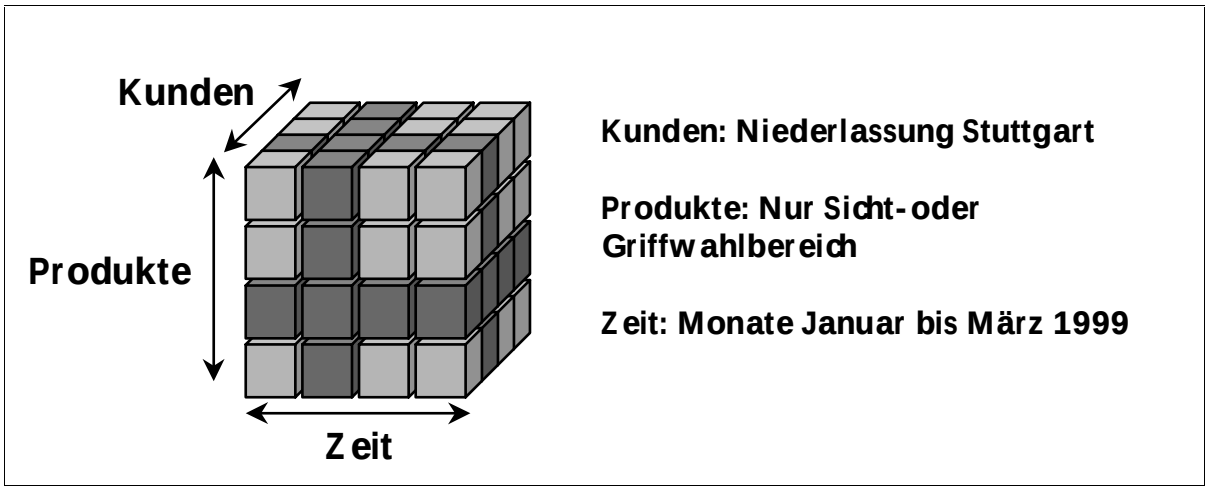


Abbildung 5.6: Einschränkungen bei der Datenmenge

In der Kundendimension wurden nur die Hauptkunden der Niederlassung Stuttgart betrachtet, die aus der Kundensegmentierung des vorherigen Business Cases gewonnen werden konnten. Bei dieser Kundengruppe besteht Grund zur der Annahme, dass deren Bestellverhalten repräsentativ ist, zudem lassen sich die gewonnen Erkenntnisse durch die Nähe zu den Mitarbeitern der Niederlassung Stuttgart leichter überprüfen.
In der Produktdimension wurden nur Artikel betrachtet, die sich im Griff- oder Sichtwahlbereich des Kunden befinden. Hintergrund hierfür bildet die Tatsache, dass im Gegensatz zu den apothekenpflichtigen Artikeln für diese Produktgruppen keine gesetzliche Preisspannenverordnung vorgeschrieben ist. Daher bietet sich für die Gehe nur bei diesen Produktgruppe ein Spielraum bei der Preisgestaltung, was eine Voraussetzung zur Bildung von Aktionspaketen oder Sonderangeboten bildet.
Um die Datenmenge einzuschränken, wurde die Datenselektion auf den Zeitraum Januar bis März 1999 eingegrenzt. Daraus ergab sich ein Analysevolumen von mehr als einer Million Datensätzen, was ungefähr 70.000 einzelnen Bestellungen entspricht.
Da die Analyse einmal auf der Ebene der ATC3-Warengruppen und einmal auf Artikelebene stattfand, wurden zwei Tabellen erstellt.

ORDER_ID	IMS_ATC3_SHORT_DESC
bid2784	HAARWASCHMITTEL
bid2784	ANTIRHEUMATIKA, TOPISCH
bid2784	MULTIVITAMINE O.MINERAL
bid2785	BONBONS/SONST.ZUCKERWAR
bid2785	SONSTIGE ANTIBIOTIKA

Abbildung 5.7: Darstellung der Analysetabelle

In Abbildung 5.7 ist ein Ausschnitt aus der Tabelle für die Assoziationsanalyse der Warengruppen dargestellt. Zu sehen sind die Bestellnummer und eine Kurzbeschreibung der Warengruppe. Das Aussehen der Artikeltabelle ist entsprechend. Auf diesen Tabellen wurde die Assoziationsanalyse mit den Tools von IBM, SAS und Attar durchgeführt, deren Ergebnisse in den folgenden Abschnitten präsentiert werden.

5.3.2. Ergebnisse der Warenkorbanalyse auf Warengruppenebene

Wie in Abschnitt 2.5.3 bereits beschrieben, wird über die Vorgaben eines minimalen Trägers und einer minimalen Konfidenz die Anzahl der Assoziationsregeln gesteuert, die vom Algorithmus gefunden werden. Bei der vorliegenden Implementierung wurde mit den Standardwerten begonnen, welche die Tools für diese Parameter anboten. In diesem Fall betrug der minimale Träger 5 und die minimale Konfidenz 10. Allerdings wurde damit vom Algorithmus eine sehr große Anzahl von Regeln identifiziert, so dass diese Werte schrittweise erhöht wurden, bis sich mit einem minimalen Träger von 10 und einer minimalen Konfidenz von 35 eine überschaubare Menge von Regeln ergab. In der folgenden Abbildung ist eine grafische Visualisierung der Regeln dargestellt, wie sie vom Enterprise Miner von SAS vorgenommen wird.

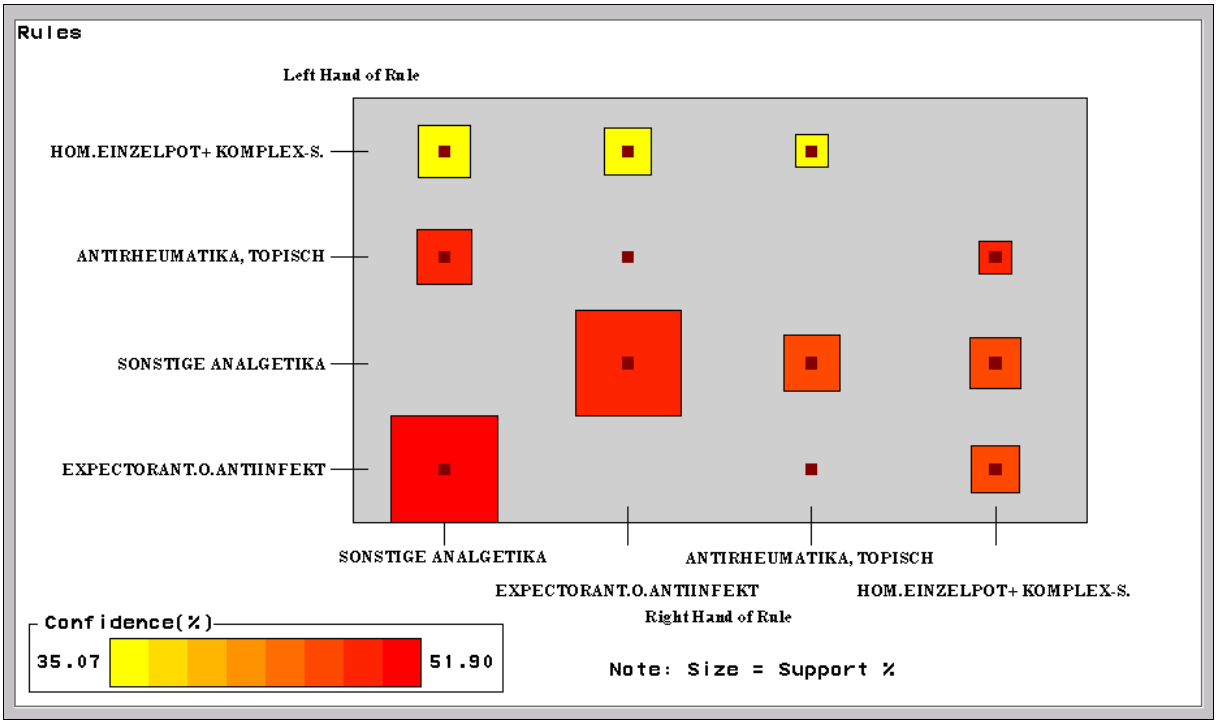


Abbildung 5.8: Visualisierung der Assoziationsregeln

Auf der Y-Achse sind die Warengruppen dargestellt, die die linke Seite der Assoziationsregel bilden. Von oben nach unten sind dies Homöopathische Produkte, Mittel gegen Rheumaerkrankungen, Schmerzmittel und Medikamente gegen Infektionskrankheiten. Die X-Achse bildet entsprechend die rechte Seite der Regel. Mit einem Quadrat wird jeweils dargestellt, dass zwischen zwei Warengruppen eine Abhängigkeit besteht. Die Größe gibt dabei die Häufigkeit an, mit der diese Regel in den Bestellungen auftritt, also den Träger, der in diesem Fall zwischen 10% und 15% liegt. Farblich dargestellt ist die Konfidenz, also die Sicherheit der Assoziationsregel.

Aus der Abbildung kann entnommen werden, dass Mittel gegen Infektionskrankheiten und Schmerzmittel häufig gemeinsam bestellt werden. Hintergrund hierfür ist sicherlich der Beobachtungszeitraum, der in den Wintermonaten liegt. Auffallend ist auch Zusammenhang zwischen den Homöopathischen Produkten und den Schmerzmitteln beziehungsweise Antiinfektika. Die Ursache hierfür ist, dass die Niederlassung Stuttgart im Bereich der

Homöopathischen Artikel von allen Niederlassungen den höchsten Umsatz tätigt. Diese Tatsache konnte auch in Zusammenarbeit mit den Fachbereichen verifiziert werden. Aus der Darstellung kann auch entnommen werden, dass zwischen der Bestellung von Antirheumatika und Homöopathischen Produkten ein etwas schwächerer Zusammenhang besteht und dass zwischen den Warengruppen Antirheumatika und Antiinfektika keine Abhängigkeiten bestehen.

Diese Erkenntnisse lassen sich in verschiedener Weise umsetzten. Einerseits kann durch Kooperationen mit der Pharma-Industrie die Voraussetzungen für besonders attraktive Sonderangebote für den Apotheker geschaffen werden, die an diesen Zusammenhängen orientiert sind. Diese Analyse kann separat für jede Niederlassung und für ausgewählte Zeiträume durchgeführt werden, so dass diese Angebote anhand der Marktsituation und den saisonalen Gegebenheiten optimiert sind. Auch zur Gewinnung von Neukunden kann diese Strategie eingesetzt werden. Für die Regalanordnung im Verkaufsbereich der Apotheken können sich ebenfalls neue Impulse ergeben, indem beispielsweise ausgewählte Artikel aus diesen Warengruppen näher beieinander platziert werden, um so für den Verbraucher einen höheren Kaufanreiz zu schaffen.

5.3.3. Ergebnisse der Warenkorbanalyse auf Artikelebene

Bei dieser Analyse konnte mit den vorgegebenen Werten für den minimalen Träger und der minimalen Konfidenz zunächst keine verwertbaren Resultate erzielt werden. Vielmehr musste der Träger schrittweise verkleinert werden, bis bei einem Wert von 0.5 schließlich Assoziationsregeln in dem Datenbestand identifiziert werden konnten. Ein Ausschnitt aus diesen Regeln ist in Abbildung 5.9 zu sehen.

Support(%)	Confidence(%)	Rule
0,46	49,04	VICHY SELBSTBR GESICHT DUNKEL == SINPRO N SCHMERZL GRAN
0,46	39,63	SINPRO N SCHMERZL GRAN == VICHY SELBSTBR GESICHT DUNKEL
0,34	100	TEEBAUMOEL AARTEMARYS == SANO HAUSTEE FILTER BTL
0,34	100	RESPIFLO HC O2-ADAP WARM == TEEBAUMOEL AARTEMARYS
0,34	98,56	SANO HAUSTEE FILTER BTL == TEEBAUMOEL AARTEMARYS
0,34	100	TEEBAUMOEL AARTEMARYS == RESPIFLO HC O2-ADAP WARM
0,34	100	RESPIFLO HC O2-ADAP WARM == SANO HAUSTEE FILTER BTL
0,34	98,56	SANO HAUSTEE FILTER BTL == RESPIFLO HC O2-ADAP WARM
0,33	97,09	TEEBAUMOEL AARTEMARYS == MEPHA ROYAL UNGEZ BIO 65CM
0,33	100	MEPHA ROYAL UNGEZ BIO 65CM == TEEBAUMOEL AARTEMARYS
0,33	100	MEPHA ROYAL UNGEZ BIO 65CM == SANO HAUSTEE FILTER BTL
0,33	97,09	RESPIFLO HC O2-ADAP WARM == MEPHA ROYAL UNGEZ BIO 65CM
0,33	95,69	SANO HAUSTEE FILTER BTL == MEPHA ROYAL UNGEZ BIO 65CM
0,33	95,69	SANO HAUSTEE FILTER BTL == SINPRO N SCHMERZL GRAN
0,33	28,21	SINPRO N SCHMERZL GRAN == SANO HAUSTEE FILTER BTL
0,33	100	MEPHA ROYAL UNGEZ BIO 65CM == RESPIFLO HC O2-ADAP WARM

Abbildung 5.9: Ergebnisse der Warenkorbanalyse auf Artikelebene

In den beiden Spalten auf der linken Seite ist der Support und die Konfidenz der auf der rechten Seite dargestellten Assoziationsregel zu sehen. Mit diesen Angaben lässt sich beispielsweise die erste Zeile in folgende Aussage umsetzen:

„Bestellt ein Apotheker den Artikel Vichy Selbstbräunungscreme, dann bestellt er mit einer Sicherheit von 49% auch den Artikel Sinpro Schmerzmittel. Diese Kombination beider Produkte tritt in 0.46% aller Bestellungen auf.“

Auffallend bei diesen Regeln ist die Tatsache, dass sie zwar sehr selten im Datenbestand auftreten, dafür aber teilweise über eine hohe Sicherheit verfügen. So bedeutet eine Sicherheit von 100%, dass jedes Mal, wenn der Apotheker den Artikel auf der linken Seite der Regel bestellt, auch den Artikel auf der rechten Seite bestellt. Zudem handelt es sich bei allen Artikeln in diese Assoziationen um sogenannte Langsamdreher, also Artikel, die eher seltener von den Apothekern bestellt werden.
Daher stellte sich die Frage, welche Kunden für diese Bestellungen verantwortlich waren. Aus der Analyse des Ausgangsdatenbestandes konnte herausgefunden werden, dass genau eine Apotheke in gewissen Abständen die Bestellungen tätigte, in der diese Artikel vorhanden waren.
Als Ergebnis dieser Analyse kann festhalten werden, dass zumindest im dem betrachteten Zeitrahmen keinerlei Zusammenhänge auf der Artikeleben existieren, die sich für eine Umsetzung im Bereich der Lageranordnung oder Logistik eignen. Hierfür ist das Auftreten der Regeln in dem Datenbestand und der Warenumschlag bei diesen Artikeln zu gering. Allerdings lassen sich hiermit Kunden identifizieren, die ein etwas ungewöhnliches Bestellverhalten zeigen. Bei einer Erweiterung der Analyse auf eine größere Kundengruppe besteht die Möglichkeit, dass sich hier Kundengruppen identifizieren lassen, die ein zielgerichtetes Marketing rechtfertigen.

5.4. Implementierung der Potentialermittlung

Bei diesem Business Case stehen vor allem die Maßnahmen zur Datentransformation im Vordergrund, die zur Berechnung des Gesamtumsatzes eines Kunden und zur Auswahl der Kundengruppe unternommen wurden. Anschließend wird betrachtet, auf welche Weise das Entscheidungsbaummodell generiert wurde und welche Schlussfolgerungen sich aus der Genauigkeit dieses Modells ableiten lassen.

5.4.1. Berechnung der Zielgröße

Um anhand der demografischen Daten zu einer Einschätzung des Umsatzpotentials zu gelangen, wird zunächst der Gesamtumsatz eines Kunden benötigt. Wie schon in Abschnitt 4.1.5. angedeutet, werden die Umatzzahlen bei der Gehe in Zusammenhang mit einer Niederlassung erfasst. Tätigt eine Apotheke Bestellungen bei mehreren Niederlassungen, so erhält sie bei jeder Niederlassung eine eigene Kundennummer, auf der die Umsätze gebucht werden. Um den Gesamtumsatz einer Apotheke zu berechnen, müssen die Beträge auf den einzelnen Kundennummern also zusammengefasst werden.
Hierzu wird für jede Apotheke ein eindeutiger Identifikator benötigt. Zu diesem Zweck ist im Data Warehouse die Kundenhauptnummer vorhanden, über die es möglich sein soll, einen Kunden eindeutig zu identifizieren. Es wurde jedoch festgestellt, dass die Zuordnung über

diese Nummer fehlschlägt, da eine große Anzahl Kunden vorhanden ist, die keinen Hauptkundeneintrag haben. Als nächstes wurde die BGA-Nummer des Apothekers als Identifikationsmerkmal herangezogen. Diese Nummer bekommt jeder Apotheker bei seiner Registrierung beim Bundesministerium für Gesundheit. Da in Deutschland jeder Apotheker nur eine Apotheke betreiben darf, sollte diese Zuordnung eindeutig sein. Allerdings wurde festgestellt, dass im Data Warehouse mehrere Fälle vorliegen, bei denen zwei Apotheken die gleiche BGA-Nummer besitzen. Hintergrund hierfür ist wahrscheinlich, dass bei einem Umzug des Apothekers oder einem Besitzerwechsel die Aktualisierung der BGA-Nummer nicht vorgenommen wurde. Diese Fehler wurden manuell bereinigt, so dass der Umsatz pro BGA-Nummer für das Kalenderjahr 1998 ermittelt werden konnte.

5.4.2. Auswahl der Kundengruppe

Wie schon in Abschnitt 4.1.4. beschrieben, ist ein Apotheker bei seinen Bestellungen nicht an einen einzigen Großhändler gebunden. Vielmehr besitzt jede Apotheke einen Stammlieferanten, bei dem sie ca. 80-90% ihrer Waren einkauft, während sich der Rest der meist kleinen Bestellungen auf zwei oder drei Großhändler verteilt. Um zu einer Einschätzung des Umsatzpotentiales eines Kunden zu gelangen, müssen aus dem im vorigen Abschnitt beschriebenen Datenbestand nun genau diejenigen Apotheken identifiziert werden, die ihre Waren fast ausschließlich bei der Gehe bestellen, die sogenannten Hauptkunden. Bei diesen Apotheken kann davon ausgegangen werden, dass deren Umsatz mit der Gehe zumindest beinahe dem Gesamtumsatz entspricht, den die betreffende Apotheke mit allen Großhändlern tätigt.

Den ersten Ansatzpunkt zur Identifizierung dieser Kunden bildete die im Data Warehouse vorliegende Kundenklassifikation. Jeder Kunde der Gehe wird hierbei aufgrund seines Umsatzes in eine Kategorie eingeordnet. Eine Übersicht über diese Kategorien ist in der folgenden Tabelle zu sehen:

Kundenkategorie	Bedeutung
A	Großkunde
B	Normalkunde
C	Kleinkunde
D	Kleinstkunde

Tabelle 5.1: Kundenklassifikation

Aufgrund dieser Klassifikation wurden zunächst nur die Kunden selektiert, die der Kategorie A oder B angehörten. Die Schlussfolgerung, dass diese die Hauptkunden der Gehe darstellen, ist allerdings nicht zulässig. So kann es bei einer großen Apotheke durchaus der Fall sein, dass sie der Kategorie A angehört, obwohl sie nur einen kleinen Teil ihres Umsatzes mit der Gehe tätigt. Zur Identifikation der Hauptkunden wurden deshalb Informationen über die Umsatzanteil der Konkurrenten einbezogen.

In der Tabelle CUST_COMPETITOR_TURNOVER_MT sind für jede Apotheke die geschätzten Anteile der Gehe und deren Konkurrenten am Gesamtumsatz des Kunden gespeichert. Diese Schätzungen werden von den Vertriebsmitarbeitern vor Ort aufgrund von

Erfahrungswerten vorgenommen. Aufgrund der Angaben in dieser Tabelle konnten diejenigen Kunden bestimmt werden, bei denen der Anteil der Gehe mindestens 80% von deren Gesamtumsatz beträgt. Die somit erstellte Tabelle der Hauptkunden enthielt insgesamt ca. 2.000 Einträge, was mit den Erfahrungen der Vertriebsmitarbeiter übereinstimmte. In der folgenden Abbildung ist die Zusammenstellung der Analysetabelle zu sehen.

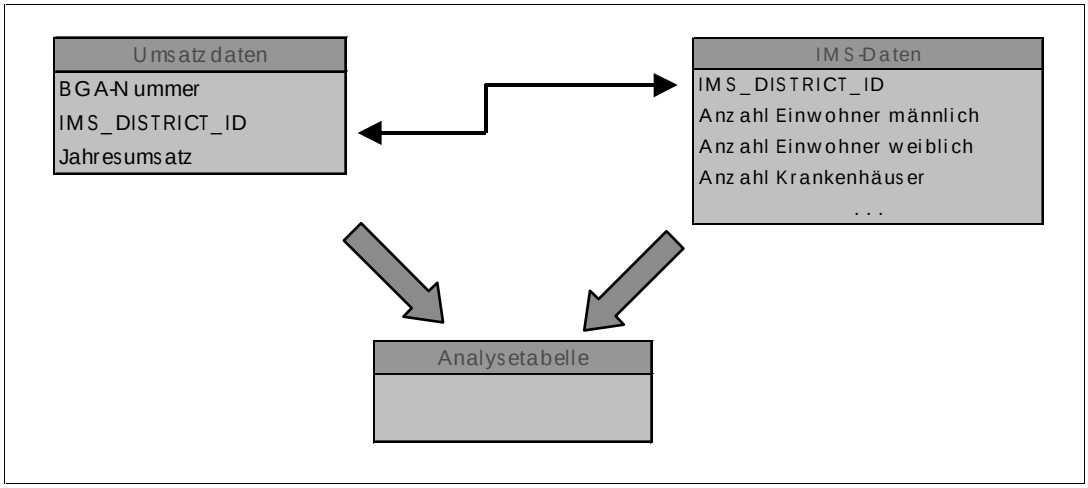


Abbildung 5.10: Zusammenstellung der Analysetabelle zur Potentialermittlung

Auf der linken Seite ist unter der BGA-Nummer der Jahresumsatz der Hauptkunden der Gehe zu sehen. Auf der rechten Seite sind die demografischen Daten wie die Einwohnerzahlen, die Anzahl der Krankenhäuser und Fachärzte pro RPM-Kreis abgebildet, die zur Modellerstellung benutzt werden. Über die IMS_DISTRICT_ID wird die Zuordnung der Apotheken zu den demografischen Daten vorgenommen.

5.4.3. Erstellung des Entscheidungsbaummodells

Zunächst wurde durch eine Zufallsauswahl eine Aufteilung der Analysetabelle in Trainings- und Testdaten vorgenommen. Anhand der Trainingsdaten erfolgte die Modellerstellung, woraufhin die Genauigkeit des Modells anhand der Testdaten überprüft werden konnte. Diese Aufgabe wurde mit verschiedenen Data Mining-Tools ausgeführt. Bei der Modellgenauigkeit konnten allerdings keine größeren Unterschiede zwischen den Tools festgestellt werden.

In einem ersten Schritt wurde der Gesamtumsatz als numerischer Wert zur Zielvariable erklärt. Das hierfür erstellt Entscheidungsbaummodell wies zwar bei den Trainingsdaten eine hohe Genauigkeit auf, bei der Überprüfung anhand der Testdaten zeigte sich jedoch, dass die Fehlerrate bei der Vorhersage des Umsatzes doch relativ groß ist. Hintergrund ist eine Übermodellierung, die bei diesen numerischen Werten einsetzt. In jeder Klasse, die vom Entscheidungsbaum generiert wird, liegen typischerweise ein oder zwei Datensätze, so dass ein sehr großer Baum entsteht, der für die Trainingsdaten optimiert ist. Allerdings stellt solch eine Baum kein allgemein gültiges Modell dar.

Um diese Übermodellierung zu vermeiden wurde in einem nächsten Schritt die Zielvariable „Gesamtumsatz“ in vier Kategorien eingeteilt, nämlich in klein, mittel, hoch und sehr hoch. Eine Übersicht über diese Umsatzbereiche ist in Tabelle 5.2 zu sehen. Für die Bestimmung der Intervallgrenzen wurde eine Einteilung in Quartilen (siehe Anhang E) vorgenommen, so dass sich in jeder der Kategorien genau 25% aller Datensätze befinden.

Kategorie	**Umsatz von ... bis ...**
1 (kleiner Umsatz)	0 - 879.469DM
2 (mittlerer Umsatz)	879.469 – 1,16 Mio. DM
3 (hoher Umsatz)	1,16 – 1,56 Mio. DM
4 (sehr hoher Umsatz)	1,56 Mio. DM - Maximum

Tabelle 5.2: Übersicht über die Umsatzkategorien

Mit diesen Kategorien als Zielvariable wurde die Entscheidungsbaumgenerierung erneut durchgeführt. Abbildung 5.11 zeigt einen Ausschnitt aus diesem Baum, wobei die Blätter in der Darstellung farblich markiert sind.

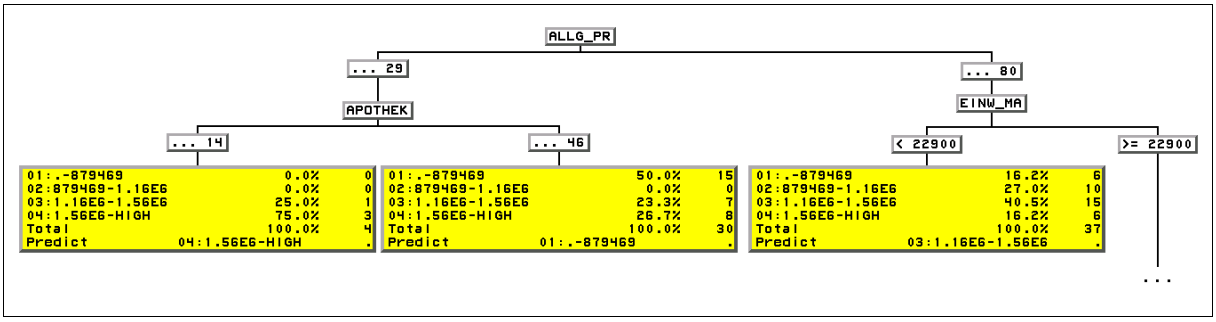
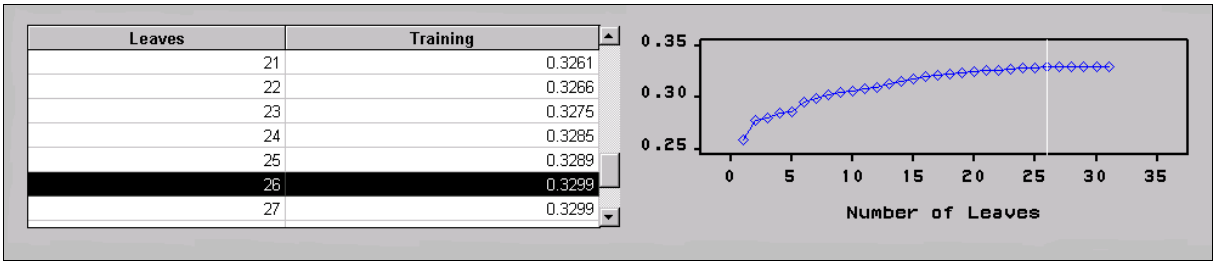


Abbildung 5.11: Ausschnitt aus dem Entscheidungsbaummodell

Die Datensätze, die in jedem Blatt liegen, können genau einer Umsatzkategorie zugeordnet werden, was aus der letzten Zeile in jedem Blatt abgelesen werden kann. Darüber wird die Gesamtzahl der Datensätze in diesem Blatt und deren Verteilung auf die einzelnen Kategorien angezeigt. Folgt man dem Wurzelknoten zu den einzelnen Blättern, so können auf dem dazugehörigem Pfad die Attribute abgelesen werden, die zur Einteilung eines Datensatzes in das entsprechende Blatt führen. Betrachtet man beispielsweise den Weg zu dem linken Blatt, so werden Apotheken, in deren Umgebung sich zwischen 0 und 29 allgemeine Ärzte und zwischen 0 und 14 konkurrierende Apotheken befinden, in die Umsatzkategorie „sehr hoch“ eingeteilt.

Aus den Angaben in den Blättern kann auch eine erste Abschätzung der Genauigkeit des Modells erfolgen. In der vorherigen Abbildung sind im ganz linken Blatt drei Datensätze in die richtige Kategorie eingeteilt worden, während bei einem eine Fehlklassifikation vorliegt. Bei dem Blatt auf der rechten Seite ist dieses Verhältnis noch ungünstiger. Eine Übersicht über die Genauigkeit des gesamten Modells kann folgender Abbildung entnommen werden:



Leaves	Training
21	0.3261
22	0.3266
23	0.3275
24	0.3285
25	0.3289
26	0.3299
27	0.3299

Abbildung 5.12: Darstellung der Genauigkeit des Entscheidungsbaummodells

In dieser Abbildung ist der Genauigkeitsverlauf des Modells in Abhängigkeit von der Anzahl der Blätter abgebildet, links in einer tabellarischen und rechts in einer grafischen Darstellung. Hierbei zeigt sich, dass lediglich eine maximale Genauigkeit von rund 33% erreicht werden

kann. Auch erneute Analysen mit einer höheren Anzahl von Umsatzkategorien brachte keine Verbesserung der Ergebnisse.

Wie schon bei der Sortimentszusammensetzung liegt auch hier die wahrscheinlichste Ursache in der Größe der RPM-Kreise, so dass sich der Umsatz einer Apotheke zumindest nicht ausschließlich aus den vorliegenden demografischen Daten erklären lässt. Vielmehr werden auch hier zusätzliche Daten benötigt, mit denen die Genauigkeit des Modells verbessert werden kann. Dazu gehören sicherlich auch Informationen über die Lage einer Apotheke, d.h. ob sie sich beispielsweise in einer Einkaufspassage, einem Bahnhof oder in eine Nebenstraße befindet.

Aus den Ergebnissen und Erfahrungen, die bei der Implementierung dieser Business Cases gewonnen wurden, lassen sich nun Empfehlungen und Hinweise für den Einsatz von Data Mining bei der Gehe aussprechen, was im letzten Kapitel näher erläutert wird.

Kapitel 6: Schlussfolgerungen aus der Arbeit und Ausblick

Aus den Ergebnissen, die sich aus der Implementierung der Business Cases ergaben, lässt sich eine prinzipielle Eignung der Datenbestände im Data Warehouse der Gehe Pharma-Handel GmbH für das Data Mining erkennen. Bei der Sortimentszusammensetzung konnte auf der Ebene der ATC1-Warengruppen eine Trennung in Haupt- und Nebenkunden vorgenommen werden, was zur einer Überwachung des Kundenportfolios und zur Überprüfung der bestehenden Kundenklassifikation eingesetzt werden kann. Auf der Ebene der ATC3-Warengruppen konnten Kundensegmente identifiziert werden, die saisonale Umsatzschwerpunkte in bestimmten Warengruppen aufweisen. Diese Kundensegmente können als Ausgangspunkt für kundenspezifische Marketing-Aktionen und saisonale Angebote dienen. Mit Hilfe der Warenkorbanalyse lassen sich Warengruppen im Sicht- und Griffwahlbereich entdecken, die von einem Apotheker häufig gemeinsam bestellt werden. Aus diesen Ergebnissen lassen sich neue Impulse für kundenspezifische Sonderangebote gewinnen.

Bei der Sortimentszusammensetzung und der Potentialermittlung von Kunden wurde allerdings auch festgestellt, dass sich die demografischen Daten zumindest in der vorliegenden Form nicht für eine Analyse eignen. Bei beiden Business Cases liegt die Schlussfolgerung nahe, dass die Größe der RPM-Kreise aus denen die Daten erhoben wurden einen Zusammenhang mit den Umsatzdaten verhindert. Um anhand von demografischen Daten neue Erkenntnisse über das Kundenverhalten zu erhalten, müssen diese in einem möglichst kleinen Umkreis um die einzelnen Kunden erhoben werden. Können die Daten auf dieser Ebene gewonnen werden, dann lassen sich auch zusätzliche Datenbestände in das Data Warehouse aufnehmen, wie beispielsweise Informationen über die Alters- oder Beschäftigungsstruktur.

Eng verbunden damit ist die Tatsache, dass der Datenbestand im Data Warehouse der Gehe Pharma Handel GmbH sehr stark auf die Umsatzzahlen fokussiert ist. So hat sich gezeigt, dass zum jetzigen Zeitpunkt relativ wenig kundenspezifische Informationen im Data Warehouse vorhanden sind. Bei der Sortimentszusammensetzung wären beispielsweise Daten über die Anzahl der Fachärzte im gleichen Gebäude sicherlich von Nutzen gewesen. Hinzu kommt, dass der Pflegezustand und Füllgrad der existierenden Daten für eine Analyse nur bedingt geeignet war. Gerade diese Datenbestände sind für Data Mining aber von besonderem Interesse, da sich aus ihnen oft Zusammenhänge erkennen lassen, die mit den herkömmlichen Analysewerkzeugen im Unternehmen nicht hergestellt werden können.

Neben den demografischen und kundenspezifischen Daten bieten sich auch noch andere Datenbestände an, mit denen der Einsatz von Data Mining bei der Gehe Pharma Handel GmbH erweitert werden kann. So ist es denkbar, zusätzlich zu den ATC-Warengruppen einen anderen Warengruppenschlüssel in das Data Warehouse zu integrieren, mit dem die Analysen unter einem anderen Betrachtungswinkel durchgeführt werden können.

Im Verlauf der Implementierung hat sich auch gezeigt, dass teilweise umfangreiche Maßnahmen zur Datenreinigung und Datentransformation notwendig sind, um zu verwertbaren Ergebnissen zu gelangen. Bereinigt werden musste vor allem die Kundentabelle, aus der unter anderem die Test- und Einrichterkunden entfernt wurden, die ansonsten die Ergebnisse verfälscht hätten. Im Rahmen der Datentransformation wurde bei der Sortimentszusammensetzung eine Transponierung der Umsatzzahlen und bei der Potentialermittlung eine Berechnung des Gesamtumsatzes pro Apotheke vorgenommen. Damit wird die in der Literatur häufig genannte Tatsache bestätigt, dass ca. 80% des Zeitaufwandes in diese Phasen investiert wird. Um diesen Zeitaufwand bei zukünftigen Analysen zu reduzieren, sollte beispielsweise die Identifizierung von Test- und Einrichterkunden erleichtert werden. Auch die Berechnung der transponierten Umsatztabelle kann schon im Data Warehouse stattfinden, um einen schnellen und flexiblen Zugriff auf die Ausgangsdaten zu gewährleisten.

Mit den Transformationsschritten, die bei der Potentialermittlung von Kunden vorgenommen werden mussten, wurde auch eine Verbesserung der Datenqualität im Data Warehouse erreicht. Hintergrund hierfür bildet die Tatsache, dass die Daten bei einer Data Mining-Anwendung typischerweise aus einem anderen Blickwinkel betrachtet werden. Bei diesem Business Case stand nicht der Umsatz eines Kunden mit einer Niederlassung im Vordergrund, wie er für die im Unternehmen vorhandenen Analysen und Auswertungen benötigt wird. Stattdessen wurde der Gesamtumsatz eines Kunden berechnet, woraufhin doppelt vorhandene BGA-Nummern und Apotheken mit einer Zuordnung zu mehreren verschiedenen RPM-Kreisen festgestellt werden konnten.

Aus der Evaluation der Data Mining-Tools konnten ebenfalls grundlegende Erfahrungen für den Einsatz dieser Tools im pharmazeutischen Großhandel gewonnen werden. Mit der Ausnahme des XpertRule Miner von Attar bieten alle im Rahmen dieser Evaluation betrachteten Tools die Vorraussetzungen, um die Datenmengen, die im Data Warehouse der Gehe Pharmahandel vorliegen, zu verarbeiten. Alle vier Produkte bieten eine Client- / Server-Architektur, die Möglichkeit des direkten Zugriffes auf das Data Warehouse und sind auf den Rechnerplattformen verfügbar, die im Unternehmen existieren. Betrachtet man die Fragestellung, welche Data Mining-Methoden das Tool unterstützen soll, so wurde deutlich, dass der pharmazeutische Großhandel eine ganze Reihe unterschiedlichster Fragestellungen bietet. Um die Einsatzmöglichkeiten von Data Mining bei der Gehe nicht von vorneherein einzuschränken wird daher eine breite Methodenabdeckung benötigt. Als Basis kann hierzu beispielsweise eines der Produkte von IBM, SAS oder Oracle genommen werden. Um die Einsatzmöglichkeiten für das Data Mining zu erweitern, kann dieses Basisprodukt durch ein weiteres Tools ergänzt werden. Als Bewertungsgrundlage und zur Ermittlung der Funktionalität kann der in dieser Diplomarbeit vorgestellte Kriterienkatalog hinzugezogen werden.

Betrachtet man die Möglichkeiten zur organisatorischen Einbettung von Data Mining in ein Unternehmen, so lassen sich die Tools unter diesem Gesichtspunkt in zwei Kategorien einteilen. Die Produkte von SAS, Oracle und Angoss erfordern vom Anwender aufgrund ihrer hohen Parametrisierbarkeit und einer etwas komplexeren Bedienung ein hohes Maß an statistische Grundkenntnissen und detailliertem Wissen über die Arbeitsweise von Data

Mining-Algorithmen. Für sie kommt am ehesten ein Einsatz in einer eigenen Analystenabteilung in Frage, in der sich speziell geschulte Mitarbeiter mit dieser Thematik beschäftigen. Die Tools von IBM und Attar besitzen dagegen eine wesentlich höheren Abstrahierungsgrad und lassen sich einfacher bedienen. Für sie wäre ein Einsatz denkbar, bei dem Data Mining bis hinein in die Fachbereiche dezentralisiert wird.

Abschließend lässt sich bemerken, dass der pharmazeutische Großhandels ein breites Spektrum von Anwendungsgebieten für den Einsatz von Data Mining bietet. Bei der Implementierung der Business Cases wurden in dieser Diplomarbeit schwerpunktmäßig Fragestellungen aus den Bereichen Marketing und Vertrieb betrachtet. Damit konnten neue Impulse für kundenspezifische Marketing-Aktionen und saisonale Angebote gewonnen werden. Als wesentliche Erkenntnis aus dieser Arbeit ergab sich aber auch, dass sich der pharmazeutische Großhandel im Gegensatz zu den Unternehmen im Versicherungs- oder Bankwesen bislang zu wenig mit der Sammlung von kundenspezifischen Daten beschäftigt hat. Für ein erfolgreiches Data Mining werden mehr Informationen über den Kunden, beispielsweise persönliche Daten oder Informationen über sein Zahlungs- und Bestellverhalten benötigt. Dazu gehört auch eine sorgfältige Pflege der gesammelten Datenbestände. Sind diese Vorraussetzungen erst einmal erfüllt, erschließen sich auch in den Unternehmensbereichen Einkauf, Finanzen und Logistik neue Anwendungsgebiete für das Data Mining. Für die schnelle Integration dieser neuen Technologie im Unternehmen können die im Rahmen dieser Diplomarbeit gewonnenen Erkenntnisse und die daraus resultierenden Empfehlungen einen wertvollen Beitrag leisten.

Anhang A: Literaturverzeichnis

[Abb] Abbott, Dean W.; Matkovsky, Philip I.; Elder, John F.: An Evaluation of High-end Data Mining Tools for Fraud Detection; White Paper

[Agr93] Agrawal, R.; Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases; in *Proceedings of the ACM SIGMOD Conference on Management of Data*; Washington D.C.; Mai 1993; Seite 207-216

[All99] Allen, Paul: D2.4 Data Warehouse Design; Critikal Consortium; 1998

[Alp98] Alpar, P., Grob, H.L., Winter, R.: Unternehmensorientierte Wirtschaftsinformatik, Braunschweig; 1998

[Bac96] Backhaus, Klaus; Erichson, Bernd; Plinke, Wulff; Weiber, Rolf: Multivariate Analysemethoden: Eine anwendungsorientierte Einführung; 8. Auflage; Springer-Verlag; Berlin; 1996

[Bar98] Barth, Tilmann: D2.3 Guidelines for the Data Mining Process; Critikal Consortium; 1998

[Ben99] Bensberg, Frank; Grob, Heinz Lothar: Das Data Mining-Konzept; Arbeitsbericht Nr. 8 der Reihe Computergestütztes Controlling; Institut für Wirtschaftsinformatik der Westfälischen Wilhelms-Universität Münster; Juni 1999

[Ber97] Berry, M.J.A., Linoff, G.: Data Mining Techniques: For Marketing, Sales and Customer Support; New York; 1997

[Bon98] Bontempo, Charles; Zagelow, George: The IBM Data Warehouse Architecture; in *Communications of the ACM*; Seite 38-48; Vol. 41 No. 9; September 1998

[Bor98] Borgelt, Christian; Kruse, Rudolf: Attributauswahl für die Induktion von Entscheidungsbäumen: Ein Überblick; in *Data Mining –Theoretische Aspekte und Anwendungen*; Physica-Verlag; Heidelberg; 1998; Seite 77-98

[Bra96] Brachman, Ronald J.; Khabaza, Tom; Kloesgen, Willi; Piatetsky-Shapiro, Gregory; Simoudis, Evangelos: Mining Business Databases; in *Communications of the ACM*; Vol. 39 No. 11; November 1996;Seite 42-48

[Bre98] Breitner, C.A.; Lockemann, P.C.; Schlösser, J.A.: Die Rolle des Informationsverwaltung im KDD-Prozess; in *Data Mining –Theoretische Aspekte und Anwendungen*; Physica-Verlag; Heidelberg; 1998; Seite 35-60

[Bre97] Brethenoux, E.: Data Mining Market: In Search of Order; White Paper; Gartner Group; 1998

[Cha98] Chamoni, Peter, Gluchowski, Peter: Analytische Informationssysteme – Einordnung und Überblick; in *Analytische Informationssysteme*;; Springer-Verlag; Berlin; 1998; Seite 4-25

[Cla98] Clausen, Nils: OLAP – Multidimensionale Datenbanken: Produkte, Markt, Funktionsweise und Implementierung; Addison Wesley Longman Verlag GmbH; Bonn; 1998

[Cod93] Codd, E.F.; Codd, S.B.; Salley, C.T.: Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate; White Paper; Codd & Date Inc.; 1993

[Deg98] Degen, Horst: Statistische Methoden zur visuellen Exploration mehrdimensionaler Daten; in *Analytische Informationssysteme*; Springer-Verlag; Berlin; 1998; Seite 388-400

[Dre98a] Dresner, H.: The Demise of Data Mining?; White Paper; Gartner Group; 1998

[Dre98b] Dresner, H.: Rumors of Data Mining's Demise are greatly Exaggerated; White Paper; Gartner Group; 1998

[Dre99] Dresner, H; Linden, A.: Criteria for Evaluating Generic Data Mining Tools; White Paper; Gartner Group; 1999

[Düs98] Düsing, Roland: Knowledge Discovery in Databases und Data Mining; in *Analytische Informationssysteme*; Springer-Verlag; Berlin; 1998; Seite 291-299

[Eld] Elder, John F.; Gomolka, Brian; King, Michel A.; Schmidt, Eric; Summers, Marguerite; Toop, Kevin: Evaluation of Fourteen Desktop Data Mining Tools; White Paper, Department of System Engineering; University of Virginia; Charlottesville

[Fay96a] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic; Uthurusamy, Ramasamy: Advances in Knowledge Discovery and Data Mining; AAAI Press, California; 1996

[Fay96b] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic: The KDD Process for Extracting Useful Knowledge from Volumes of Data; in *Communications of the ACM*; Vol. 39 No. 11; November 1996; Seite 27-34

[Fay96c] Fayyad, Usama; Uthurusamy, Ramasamy: Data Mining and Knowledge Discovery in Databases; in *Communications of the ACM*; Vol. 39 No. 11; November 1996; Seite 24-26

[Fay96d] Fayyad, Usama; Haussler, David; Stolorz, Paul: Mining Scientific Data; in *Communications of the ACM*; Vol. 39 No. 11; November 1996; Seite 51-57

[**Fay98a]** Fayyad, Usama: Taming the Giants and the Monsters: Mining Large Databases for Nuggets of Knowledge; in *Database Programming and Design magazine*; März 1998

[Fay98b] Fayyad, Usama: Mining Databases: Towards Algorithmus for Knowledge Discovery; in *Bulletin of the Technical Committee on Data Engineering*; Vol. 21 No.1; März 1998; Seite 39-48

[Fer99] Ferguson, Mike: Evaluating and Selecting Data Mining Tools; in *InfoDB*; Volume 11, Number 2

[Gar98] Gardner, Stephen R.: Building the Data Warehouse; in *Communications of the ACM*; Vol. 41 No. 9; September 1998; Seite 52-60

[**Gly98]** Glymour, Clark; Madigan, David; Pregibon, Daryl; Smyth, Padhraic: Statistical Inference and Data Mining; in *Communications of the ACM*; Vol. 41 No. 9; September 1998; Seite 35-41

[Gri98] Grimmer, Udo; Much, Hans-Joachim: Datensegmentierung mittels Clusteranalyse; in *Data Mining –Theoretische Aspekte und Anwendungen*; Physica-Verlag; Heidelberg; 1998; Seite 109-141

[Hol99] Holthuis, Jan: Der Aufbau von Data Warehouse-Systemen: Konzeption – Datenmodellierung – Vorgehen; Deutscher Universitäts-Verlag, Wiesbaden und Gabler Verlag, Wiesbaden; 2. Aufl. 1999

[Imi96] Imielinski, Tomasz; Mannila, Heikki: A Database Perspektive on Knowledge Discovery; in *Communications of the ACM*; Vol. 39 No. 11; November 1996; Seite 58-64

[Inm96] Inmon, W.H.: The Data Warehouse and Data Mining; in *Communications of the ACM*; Vol. 39 No. 11; November 1996; Seite 49-50

[Kem97] Kemper, Alfons: Datenbanksysteme: Eine Einführung; R. Oldenbourg Verlag, München; 2. Aufl. 1997

[Kim96] Kimball, Ralph: The Data Warehouse Toolkit; John Wiley & Sons Inc., New York, Chicester, Brisbane, Toronto, Singapore; 1996

[Koh95] Kohonen, T.: Self-Organizing Maps; Springer Verlag Berlin/Heidelberg; 1995

[Kra98] Krahl, Daniela; Windheuser, Ulrich; Zick, Friedrich-Karl: Data Mining: Einsatz in der Praxis; Addison Wesley Longman Verlag, Bonn; 1998

[Küp98] Küppers, Bertram: Data Mining in der Praxis: Ein Ansatz zur Nutzung der Potentiale von Data Mining im betrieblichen Umfeld; Europäische Hochschulschriften: Reihe 5; Volks- und Betriebswirtschaft Bd. 2373; Lang Verlag, Frankfurt am Main; 1998

[Lin99] Linden, A.: The 1999 Magic Quadrant on Data Mining Workbenches; White Paper; Gartner Group; 1999

[Man97] Mannila, Heikki: Methods and Problems in Data Mining; in Proceedings of International Conference on Database Theory, Delphi, Greece; Springer-Verlag; Januar 1997

[Mar98] Martin, Wolfgang: Data Warehousing; ITP GmbH, Bonn; 1. Aufl. 1998

[Mit99] Mitschang, B.: Datenbankbasierte Anwendungen – Modellierungs-, Entwicklungs- und Verarbeitungskonzepte; Foliensammlung zur Vorlesung im Sommersemester 1999; Universität Stuttgart, IPVR; 1999

[Nak98] Nakhaeizadeh, Gholamreza; Reinartz, Thomas; Wirth, Rüdiger: Wissenentdeckung in Datenbanken und Data Mining: Ein Überblick; in *Data Mining –Theoretische Aspekte und Anwendungen*; Physica-Verlag; Heidelberg; 1998: Seite 3-29

[OV97] O. V.: Ovum Evaluates: Data Mining; White Paper; 1997

[Poe98] Poe, Vidette; Klauer, Patricia; Brobst, Stephen: Building a Data Warehouse for Decision Support; Prentice Hall, New Jersey; 2. Aufl. 1998

[Red98] Redman, T.C.: The Impact of Poor Data Quality on the Typical Enterprise; in *Communications of the ACM*; Vol. 41 No. 2; Februar 1998; Seite 79-82

[Str99] Strange, K.: Data Mining Tools: A Classification; White Paper; Gartner Group; 1999

Anhang B: Übersicht über die Einzelkriterien

Einordnung des Herstellers:
Wie lange ist das Produkt schon am Markt verfügbar?
Wie lange existiert der Hersteller?
Welche Unternehmensgröße liegt vor (Mitarbeiter / Umsatz)?
Entwickelt der Hersteller eigene Algorithmen?
Ausgaben für Forschung und Entwicklung?
Ist eine Analysten- oder Consultingabteilung verfügbar?
Hat der Hersteller Niederlassungen in Deutschland?

Systemarchitektur:
Über welche Komponenten verfügt das Tool und welche Architektur wird damit erreicht?
Wie wird der Zugriff auf die Daten realisiert?
Welche Datenbanksysteme werden unterstützt?
Werden Multiprozessoren unterstützt?
Auf welchen Betriebssystemen sind die Client- und die Server-Komponente verfügbar?
Über welche Hardwareausstattung (Prozessor, Speicher, Festplattenspeicherplatz) müssen die Client- und Serverrechner verfügen?
Für welche Sprachen wird das Programm angeboten?

Datenexploration:
Welche grundlegenden statistischen Werte können angezeigt werden?
Welche Möglichkeiten gibt es zur Visualisierung der Daten?
Wie komfortabel können fehlende Daten, Ausreißer oder ungültige Werte erkannt werden?
Auf welche Datentypen können die Eingabedaten abgebildet werden?

Datenvorverarbeitung:
Welche Möglichkeiten gibt es zur Selektion der Daten?
Welche Möglichkeiten gibt es zur Datentransformation?
Welche Möglichkeiten gibt es zur Behandlung von fehlenden Werten, Ausreißern und ungültigen Werten?
Inwieweit wird das Ziehen von Stichproben unterstützt?
Welche Möglichkeiten gibt es zum Aufteilen der Datensätze in Trainings- oder Testdaten?
Welche Möglichkeiten gibt es zur Kategorisierung von Variablen?

Funktionalität Klassifikation:
Welche Algorithmen werden für diese Operation angeboten?
Gibt es Einschränkungen hinsichtlich der Werte, die vorhergesagt werden sollen (nur binäre oder kategorische Werte)?
Werden Default-Werte vorgegeben oder aus den Daten berechnet?

Welche Möglichkeiten bietet das Programm zur Optimierung der Resultate?
Welche Möglichkeiten gibt es zur Vorhersage von Werten?
Wie werden bei den einzelnen Algorithmen die Resultate visualisiert?
Wie kann die Genauigkeit der Modelle überprüft werden?
Wie können die Resultate exportiert werden?

Funktionalität Segmentierung:
Welche Algorithmen werden für diese Operation angeboten?
Welche Parameter können eingegeben werden, um die Genauigkeit der Resultate zu verbessern (Anzahl der Cluster, Gewichtungen)?
Welche Stoppkriterien können für den Algorithmus eingestellt werden?
Werden Default-Werte vorgegeben oder aus den Daten berechnet?
Wie intuitiv werden die gefundenen Cluster visualisiert?
Wie können die Resultate exportiert werden?

Funktionalität Abhängigkeitsanalyse:
Für welche Datenstruktur wird die Assoziationsanalyse angeboten?
Welche Parameter können zur Einschränkung der Regelmenge angegeben werden (Konfidenz, Support)?
Werden Default-Werte vorgegeben oder aus den Daten berechnet?
Wie werden die gefundenen Regeln / Assoziationen visualisiert?
Wie können die Resultate exportiert werden?

Funktionalität Änderungs- und Abweichungserkennung:
Welche Algorithmen werden für diese Operation angeboten?
Inwieweit sind die einzelnen Algorithmen parametrisierbar?
Werden Default-Werte vorgegeben oder aus den Daten berechnet?
Wie werden bei den einzelnen Algorithmen die Resultate visualisiert?
Wie können die Resultate exportiert werden?

Bedienung:
Welchen Eindruck macht die Benutzeroberfläche (Intuition)?
Qualität des Hilfesystems?
Stehen Assistenten zur Verfügung?
Werden bei den einzelnen Operationen Fortschrittsanzeigen eingeblendet?
Sind Programmfehler oder Abstürze aufgetreten?

Automatisierbarkeit:
Inwieweit wird die sequentielle Ausführung der einzelnen Operationen unterstützt?
Steht eine Makrosprache oder eine eigene Programmiersprache zur Verfügung?
Welche Möglichkeiten bietet das Tool, um die erstellten Modelle oder Ergebnisse in eigene Applikationen einzubinden?

Anhang C: Darstellung der Gewichtung und Bewertung der Einzelkriterien

Einordnung des Herstellers

	Gewichtung	Intelligent Miner for Data v6.1	Enterprise Miner v3.0	Darwin v3.6.0	Knowledge Studio v2.0	XpertRule Miner v3.0.6
Wie lange ist das Produkt schon am Markt verfügbar?	3	+ +	+ +	o	+	-
Wie lange existiert der Hersteller?	1	+ +	+ +	+	o	o
Welche Unternehmensgröße liegt vor (Mitarbeiter / Umsatz)?	2	+ +	+ +	+ +	-	- -
Entwickelt der Hersteller eigene Algorithmen?	2	+ +	o	o	-	- -
Ausgaben für Forschung und Entwicklung?	3	+	+ +	+	o	*
Ist eine Analysten- oder Consultingabteilung verfügbar?	3	+ +	+ +	+	-	-
Hat der Hersteller Niederlassungen in Deutschland?	2	+ +	+	+ +	-	- -
Gesamtbewertung		5	5	4	3	1

Systemarchitektur

	Gewichtung	Intelligent Miner for Data v6.1	Enterprise Miner v3.0	Darwin v3.6.0	Knowledge Studio v2.0	XpertRule Miner v3.0.6
Über welche Komponenten verfügt das Tool und welche Architektur wird damit erreicht?	3	+ +	+ +	o	+ +	- -
Wie wird der Zugriff auf die Daten realisiert?	2	+ +	+	+	o	-
Welche Datenbanksysteme werden unterstützt?	2	+	+	+ +	+	- -
Werden Multiprozessoren unterstützt?	3	+ +	- -	o	+ +	- -
Auf welchen Betriebssystemen sind die Client- und die Server-Komponente verfügbar?	*	*	*	*	*	*
Über welche Hardwareausstattung müssen die Client- und Serverrechner verfügen?	*	*	*	*	*	*
Für welche Sprachen wird das Programm angeboten?	*	*	*	*	*	*
Gesamtbewertung		5	3	4	4	1

Datenexploration

	Gewichtung	Intelligent Miner for Data v6.1	Enterprise Miner v3.0	Darwin v3.6.0	Knowledge Studio v2.0	XpertRule Miner v3.0.6
Welche grundlegenden statistischen Werte können angezeigt werden?	2	-	+ +	-	o	o
Welche Möglichkeiten gibt es zur Visualisierung der Daten?	3	-	+ +	- -	+	o
Wie komfortabel können fehlende Daten, Ausreißer oder ungültige Werte erkannt werden?	3	o	+	-	+ +	-
Auf welche Datentypen können die Eingabedaten abgebildet werden?	1	+	+ +	o	o	- -
Gesamtbewertung		2	5	1	1	2

Datenvorverarbeitung

	Gewichtung	Intelligent Miner for Data v6.1	Enterprise Miner v3.0	Darwin v3.6.0	Knowledge Studio v2.0	XpertRule Miner v3.0.6
Welche Möglichkeiten gibt es zur Selektion der Daten?	1	o	-	+ +	- -	o
Welche Möglichkeiten gibt es zur Datentransformation?	3	+	+	o	- -	o
Welche Möglichkeiten gibt es zur Behandlung von fehlenden Werten, Ausreißern und ungültigen Werten?	3	+	+ +	o	- -	- -
Inwieweit wird das Ziehen von Stichproben unterstützt?	3	+ +	+ +	-	-	- -
Welche Möglichkeiten gibt es zum Aufteilen der Datensätze in Trainings- oder Testdaten?	2	-	+ +	o	o	- -
Welche Möglichkeiten gibt es zur Kategorisierung von Variablen?	3	-	+ +	- -	- -	-
Gesamtbewertung		3	5	2	1	1

Funktionalität Klassifikation

	Gewichtung	Intelligent Miner for Data v6.1	Enterprise Miner v3.0	Darwin v3.6.0	Knowledge Studio v2.0	XpertRule Miner v3.0.6
Welche Algorithmen werden für diese Operation angeboten?	*	*	*	*	*	*
Gibt es Einschränkungen hinsichtlich der Werte, die vorhergesagt werden sollen?	1	-	+ +	+ +	+ +	+
Werden Default-Werte vorgegeben oder aus den Daten berechnet?	2	+	-	+ +	+ +	o
Welche Möglichkeiten bietet das Programm zur Optimierung der Resultate?	3	+	+ +	+ +	+	-
Welche Möglichkeiten gibt es zur Vorhersage von Werten?	2	-	+	+	-	o
Wie werden bei den einzelnen Algorithmen die Resultate visualisiert?	1	-	+	- -	+	o
Wie kann die Genauigkeit der Modelle überprüft werden?	3	-	+ +	+ +	o	- -
Wie können die Resultate exportiert werden?	2	- -	o	+	+	o
Gesamtbewertung		3	4	5	4	2

Funktionalität Segmentierung

	Gewichtung	Intelligent Miner for Data v6.1	Enterprise Miner v3.0	Darwin v3.6.0	Knowledge Studio v2.0	XpertRule Miner v3.0.6
Welche Algorithmen werden für diese Operation angeboten?	*	*	*	*	*	*
Welche Parameter können eingegeben werden, um die Genauigkeit der Resultate zu verbessern?	3	+ +	+ +	o	o	*
Welche Stoppkriterien können für den Algorithmus eingestellt werden?	2	+	+	-	-	*
Werden Default-Werte vorgegeben oder aus den Daten berechnet?	2	o	-	o	- -	*
Wie intuitiv werden die gefundenen Cluster visualisiert?	3	+ +	-	o	- -	*
Wie können die Resultate exportiert werden?	2	+	o	o	- -	*
Gesamtbewertung		**5**	**3**	**3**	**1**	**0**

Funktionalität Abhängigkeitsanalyse

	Gewichtung	Intelligent Miner for Data v6.1	Enterprise Miner v3.0	Darwin v3.6.0	Knowledge Studio v2.0	XpertRule Miner v3.0.6
Für welche Datenstruktur wird die Assoziationsanalyse angeboten?	*	*	*	*	*	*
Welche Parameter können zur Einschränkung der Regelmenge angegeben werden?	2	+ +	+ +	*	*	o
Werden Default-Werte vorgegeben oder aus den Daten berechnet?	3	+	o	*	*	-
Wie werden die gefundenen Regeln / Assoziationen visualisiert?	2	+ +	+ +	*	*	-
Wie können die Resultate exportiert werden?	2	+	+ +	*	*	-
Gesamtbewertung		5	5	0	0	2

Funktionalität Änderungs- und Abweichungserkennung

	Gewichtung	Intelligent Miner for Data v6.1	Enterprise Miner v3.0	Darwin v3.6.0	Knowledge Studio v2.0	XpertRule Miner v3.0.6
Welche Algorithmen werden für diese Operation angeboten?	*	*	*	*	*	*
Inwieweit sind die einzelnen Algorithmen parametrisierbar?	2	o	+ +	*	*	*
Werden Default-Werte vorgegeben oder aus den Daten berechnet?	2	-	-	*	*	*
Wie werden bei den einzelnen Algorithmen die Resultate visualisiert?	3	+	+	*	*	*
Wie können die Resultate exportiert werden?	2	-	o	*	*	*
Gesamtbewertung		3	4	0	0	0

Bedienung

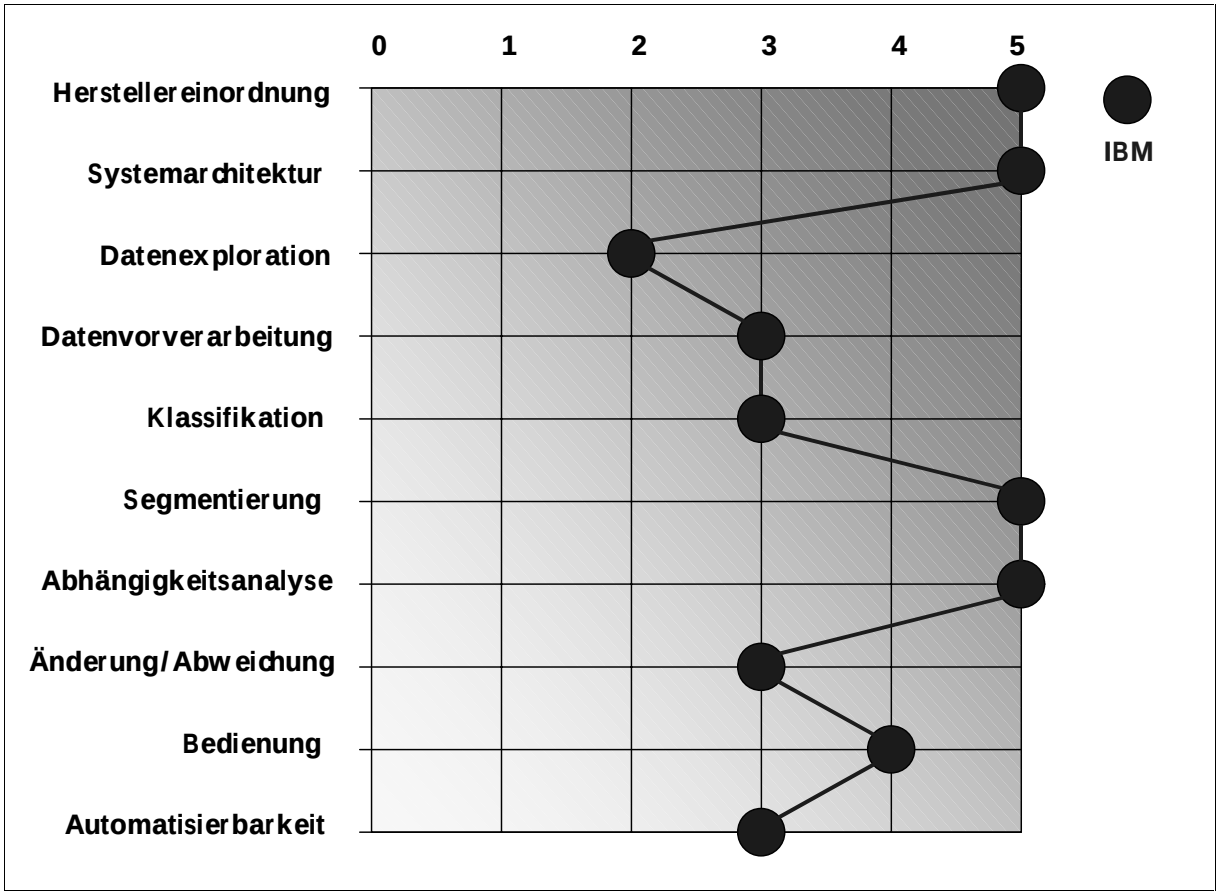
	Gewichtung	Intelligent Miner for Data v6.1	Enterprise Miner v3.0	Darwin v3.6.0	Knowledge Studio v2.0	XpertRule Miner v3.0.6
Welchen Eindruck macht die Benutzeroberfläche (Intuition)?	3	+	+ +	-	-	o
Qualität des Hilfesystems?	2	+	o	+	-	-
Stehen Assistenten zur Verfügung?	1	+ +	- -	+	o	o
Werden bei den einzelnen Operationen Fortschrittsanzeigen eingeblendet?	2	+	o	o	+ +	+
Sind Programmfehler oder Abstürze aufgetreten?	3	+	+	-	-	- -
Gesamtbewertung		4	4	3	3	2

Automatisierbarkeit

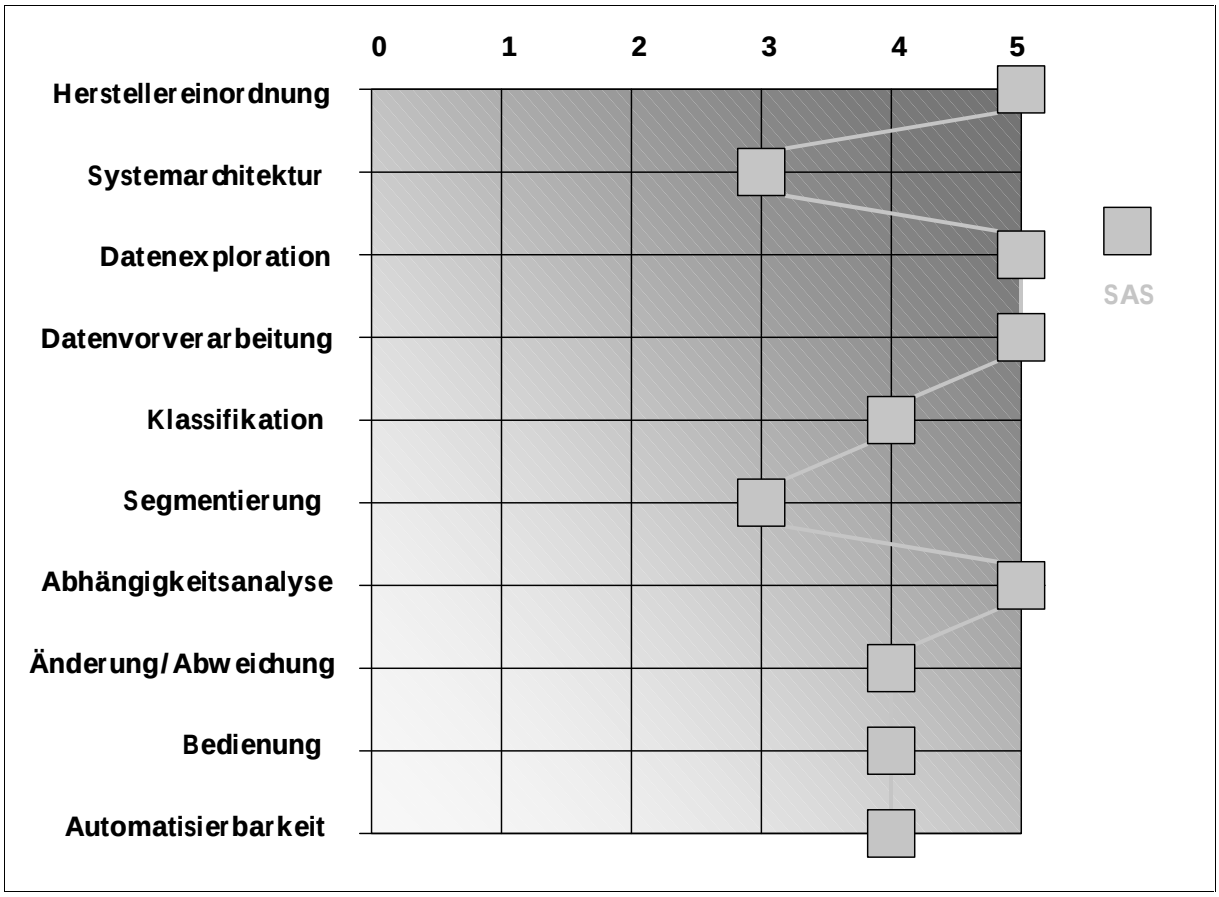
	Gewichtung	Intelligent Miner for Data v6.1	Enterprise Miner v3.0	Darwin v3.6.0	Knowledge Studio v2.0	XpertRule Miner v3.0.6
Inwieweit wird die sequentielle Ausführung der einzelnen Operationen unterstützt?	3	o	+ +	o	- -	+
Steht eine Makrosprache oder eine eigene Programmiersprache zur Verfügung?	2	-	+	-	- -	+
Welche Möglichkeiten bietet das Tool, um die erst. Modelle oder Ergebnisse in eigene Applikationen einzubinden?	1	o	-	o	+	+
Gesamtbewertung		3	4	3	2	4

Anhang D: Darstellung der Gesamtbewertung

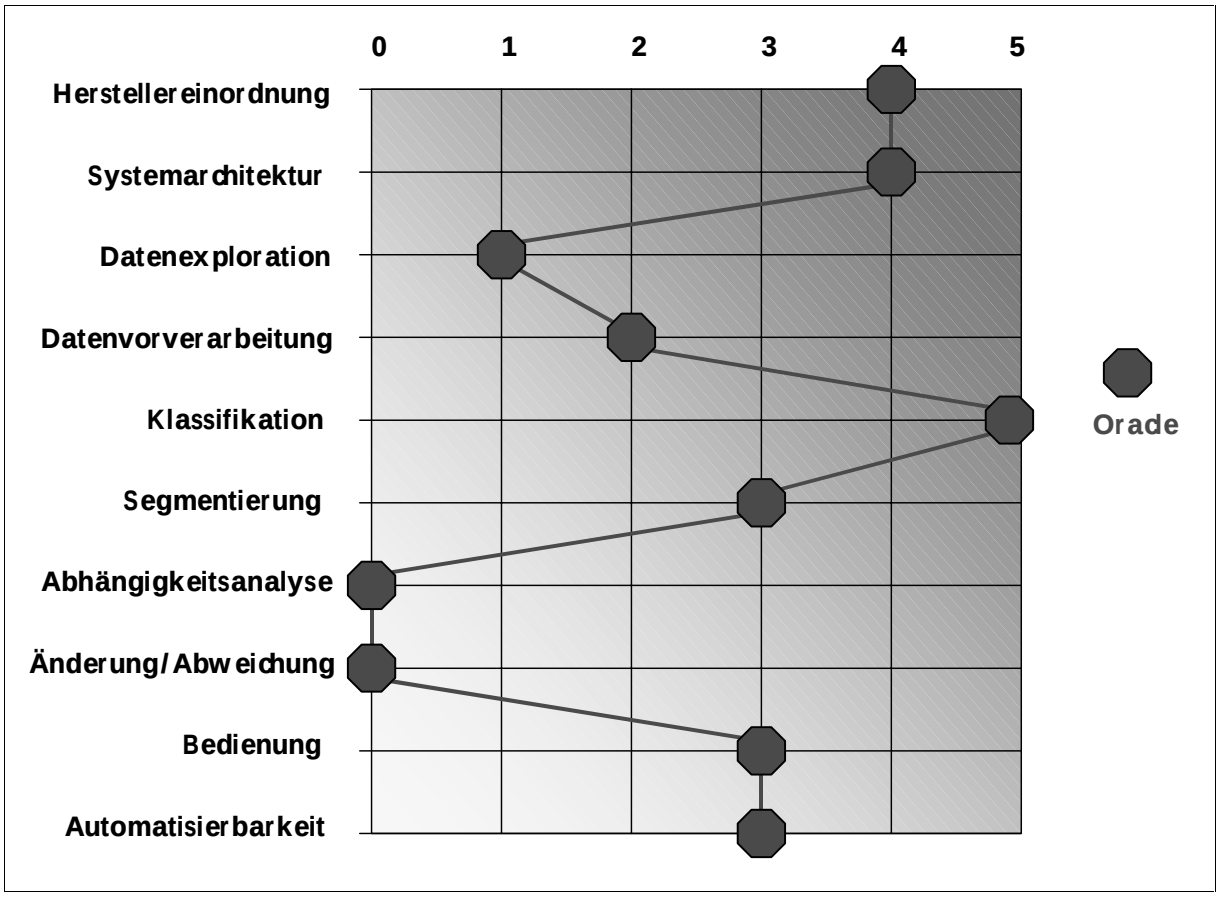
Intelligent Miner for Data v6.1



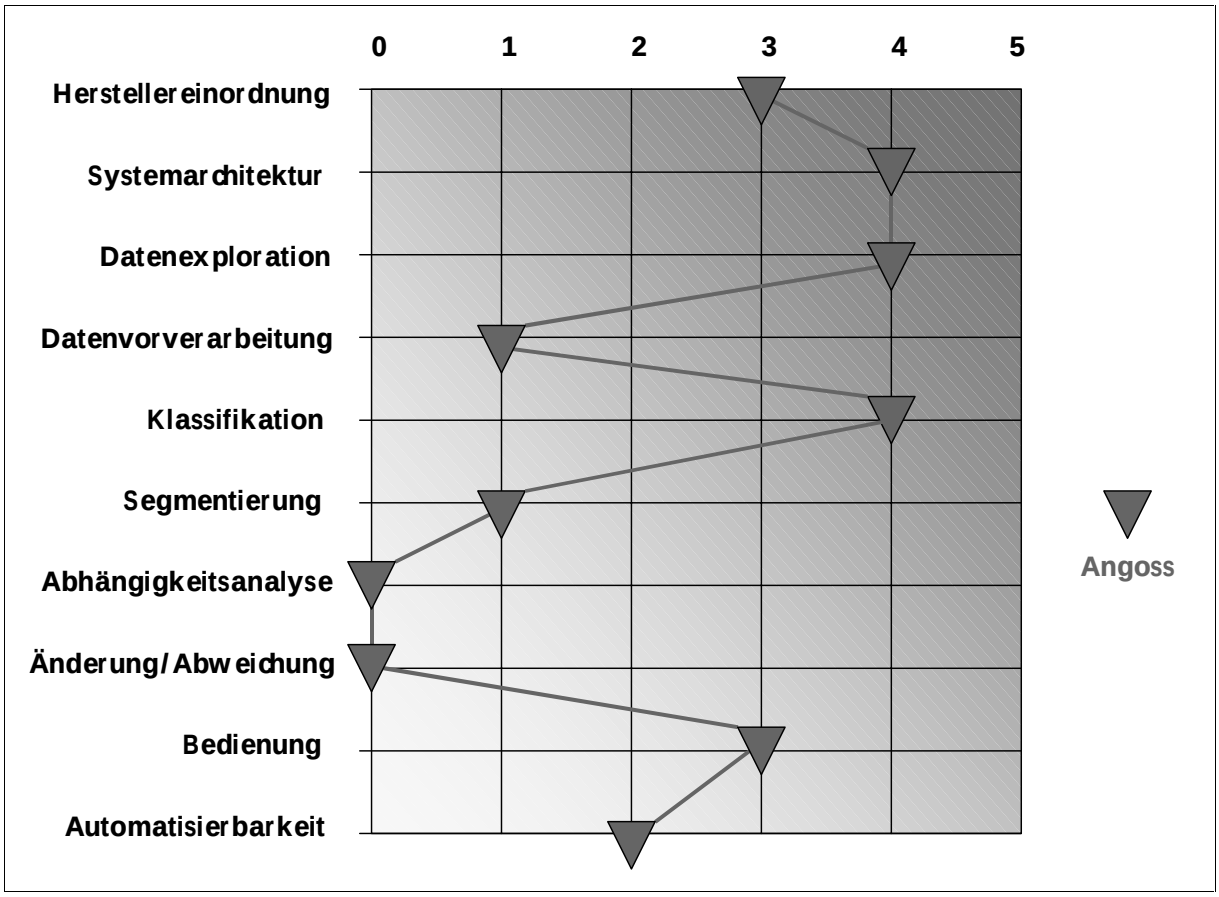
Enterprise Miner v3.0



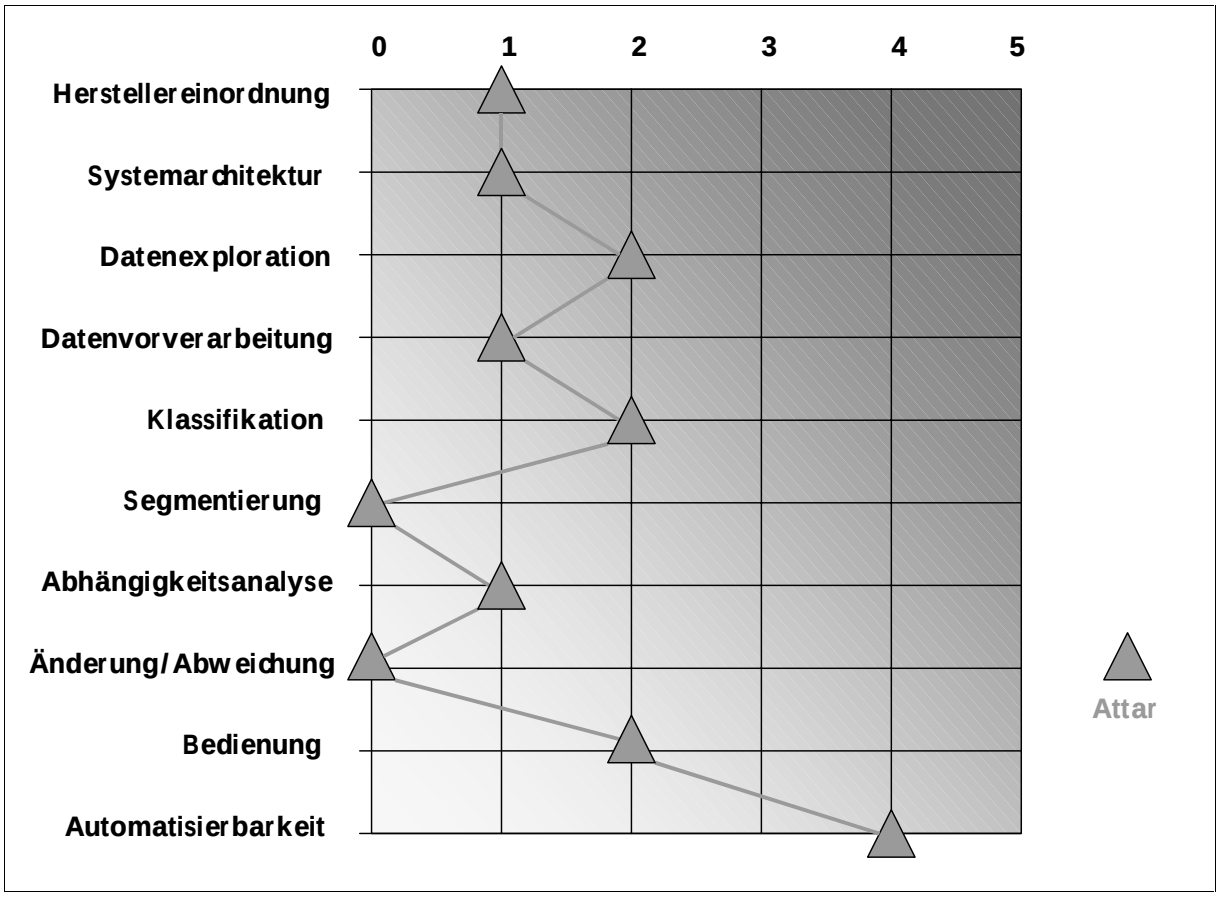
Darwin v3.6.0


0
1
2
3
4
5
Herstellereinordnung
Systemarchitektur
Datenexploration
Datenvorverarbeitung
Klassifikation
Segmentierung
Abhängigkeitsanalyse
Änderung/Abweichung
Bedienung
Automatisierbarkeit
Orade

KnowledgeStudio v2.0



XpertRule Miner v1.46



Anhang E: Grundbegriffe der Statistik

In diesem Anhang soll eine kurze Einführung in die Grundbegriffe der Statistik gegeben werden, von denen einige in der Diplomarbeit verwendet wurden. Damit kann auch die Einarbeitung in die Data Mining-Tools erleichtert werden, da hierzu zumindest statistische Grundkenntnisse verlangt werden. Aus diesem Grund ist bei einigen Begriffen auch die englische Übersetzung beigefügt.

Verteilungen

Statistische Verfahren werden als Werkzeug innerhalb empirischer Wissenschaften eingesetzt, d.h. in Bereichen, in denen das Wissen aus Erfahrungen, unter anderem durch Beobachtungen oder Antworten auf standardisierte Fragen gewonnen wird. Dabei wird versucht, auf der Basis der ermittelten Daten Aussagen über die Eigenschaften der jeweils vorliegenden Untersuchungseinheiten zu erlangen. Die zu untersuchenden Eigenschaften werden als **Merkmale** (Variablen) und die jeweils möglichen Ergebnisse als **Merkmalsausprägungen** (Werte) bezeichnet. Sollen beispielsweise Informationen über die zeitliche Belastung eines Schülers durch den Unterricht gewonnen werden, so ist das Merkmal „Unterrichtsstunden“ zu untersuchen. Die zugehörigen Merkmalsausprägungen sind ganzzahlige Werte und können beispielsweise zwischen 29 und 40 variieren. Die Häufigkeit der einzelnen Merkmalsausprägungen können entweder in einer Tabelle oder in einem Balkendiagramm dargestellt werden. Dabei wird von einer **Häufigkeitsverteilung** (engl. *frequency distribution*), abgekürzt zu Verteilung, gesprochen.

Absolute und relative Häufigkeiten einer Verteilung

Damit man die Bedeutung eines Wertes im Hinblick auf das Auftreten der anderen Werte ohne grafische Hilfsmittel erkennen kann, sind geeignete Maßzahlen festzulegen, mit denen die ermittelten Häufigkeiten der einzelnen Werte zu einer Bezugsgröße in Beziehung gesetzt werden können. Die **absolute Häufigkeit** (engl. *frequencies*) ergibt sich dabei aus der Anzahl der Ausprägungen eines Wertes. Die **relative Häufigkeit** (engl. *percentage*) ist definiert als die absolute Häufigkeit eines Wertes geteilt durch die Gesamtzahl der Werte.

Gliederung einer Verteilung

Um eine detaillierte Aussage über die Lage eines Wertes im Hinblick auf sämtliche anderen Werte einer Variablen zu erhalten, kann eine grobe Gliederung einer Verteilung in vier gleich große Verteilungsflächen (**Quartilen** Q, ebenso **Perzentilen** P sowie **Dezilen** D) vorgenommen werden. Sie schneiden eine Verteilung in vier bzw. einhundert oder zehn Teile. Der 1. Quartilwert hat also die Eigenschaft, dass mindestens 25% der Werte kleiner oder gleich diesem Wert und mindestens 75% der Werte größer oder gleich diesem Wert sind. Für den 2. (3.) Quartilwert gilt, dass mindestens 50% (75%) der Werte kleiner oder gleich diesem Wert und mindestens 50% (25%) der Werte größer oder gleich diesem Wert sind.

Kategorisierung von Daten

Bei Merkmalen mit sehr vielen unterschiedlichen Merkmalsausprägungen besteht die Gefahr, dass die Häufigkeitsverteilung keinen Gesamteindruck vermittelt, da viele Werte nur mit einer sehr kleinen Häufigkeit auftreten. Es gelingt nur dann eine Einsicht in die Art der Verteilung, wenn eine **Kategorisierung** (engl. *grouped data*), auch Gruppierung oder Klassierung genannt, vorgenommen wird. Sollen beispielsweise beim Merkmal „Schulleistung“ die Punktwerte von 1 bis 9 kategorisiert werden, so liegt eine Zusammenfassung in Form der Klassen „gering“ (1-3), „mittel“ (4-6) und „hoch“ (7-9) nahe. Im Hinblick auf eine Kategorisierung ist es von Bedeutung, ob es sich um ein diskretes oder um ein kontinuierliches Merkmal handelt.

- Als diskretes oder diskontinuierliches Merkmal (engl. *discrete variable*) bezeichnet man ein Merkmal, das nur eine endlich viele Anzahl von Zahlencodes oder Ausprägungen annimmt, das also innerhalb eines bestimmten Bereiches Lücken aufweist. Beispiel: Kinderzahl.

- Als kontinuierliches oder stetiges Merkmal (engl. *continous variable*) bezeichnet man ein Merkmal, das unendlich viele Werte annehmen kann. Beispiel: Gewicht, Zeit.

Skalenniveau von Merkmalen

Von großer Bedeutung ist die Frage, ob man die Werte eines Merkmals ohne weiteres summieren darf, und ob es zulässig ist, mit diesen Gesamtwerten bestimmte numerische und statistische Operationen auszuführen. Dazu kann ein Merkmal einem von drei verschiedenen Skalenniveaus zugeordnet werden.

- Bei einer **Nominalskala** können die Merkmalsausprägungen keiner Rangordnung unterzogen werden (z.B. Religionszugehörigkeit mit den Ausprägungen evangelisch, römisch-katholisch...).

- Bei einer **Ordinalskala** kann eine Rangordnung festgelegt werden, man kann jedoch keine Abstände messen. Ein Beispiel sind Schulnoten. Man kann zwar sagen, dass eine eins besser wie eine zwei ist, aber man kann nicht behaupten, dass eine zwei doppelt so gut ist wie eine vier.

- Bei einer **Intervallskala** können auch Abstände gemessen werden (z.B. Jahreseinkommen).

Kennzeichnung des Zentrums

- Für ein intervallskaliertes Merkmal wird das Zentrum durch das **arithmetisches Mittel** „$\bar{x}$“ (engl. *mean*) gekennzeichnet. Es ist definiert als die Summe aller Werte einer Häufigkeitsverteilung, dividiert durch die Anzahl der Werte.

- Bei einer Ordinalskala wird der **Median** (engl. *median*) zur Kennzeichnung des Zentrums verwendet. Er ist gleich dem 2. Quartilwert.
- Bei einer Nominalskala wird der **Modus** (engl. *mode*), auch Modalwert genannt, verwendet. Er ist als der Wert festgelegt, der die größte Häufigkeit besitzt.

Kennzeichnung der Variabilität

Nachdem Statistiken zur Beschreibung des Zentrums einer Verteilung vorgestellt wurden, stehen jetzt Statistiken im Vordergrund, mit denen sich angeben lässt, wie konzentriert bzw. wie weitläufig die Werte um das Zentrum gelagert sind. Derartige Maßzahlen werden als Statistiken der Variabilität (engl. *dispersion*) bezeichnet. Für ein intervallskaliertes Merkmal lässt sich die Variabilität (in einem ersten Ansatz) durch die **Variation** kennzeichnen. Diese Statistik beschreibt die quadratische Abweichung vom Mittelwert (engl. *sum of squares*) und wird wie folgt errechnet:

$$Variation = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Die Quadrierung erfolgt deswegen, um den sich gegenseitig aufhebenden Effekt negativer und positiver Vorzeichen zu umgehen. Die Gruppengröße wird bei der Berechnung der Variation allerdings nicht berücksichtigt. Um eine Statistik zu erhalten, mit der unterschiedlich große Gruppen von Merkmalsträgern vergleichbar gemacht werden können, teilt man die Variation durch eine Größe, die von der Anzahl der Werte abhängt, die in die Summation einbezogen werden. Eine derartige Normierungsgröße stellt die Anzahl der **Freiheitsgrade** (engl. *degrees of freedom*) der Variation dar. Eine Datenbasis mit n Werten besitzt n-1 Freiheitsgrade. Daraus errechnet sich die Varianz (engl. *variance*):

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Ein solcher Wert ist inhaltlich allerdings kaum interpretierbar. Die **Standardabweichung** s (engl. *standard deviation*), die definiert ist als die Quadratwurzel der Varianz, weist diesen Nachteil nicht auf. Sie drückt die Streuung in denselben Maßeinheiten wie die Beobachtungswerte aus.

Für nicht intervallskalierte Merkmale ist es nicht sinnvoll, die Varianz oder die Standardabweichung zur Kennzeichnung der Variabilität zu verwenden. Als geeignete Statistiken lassen sich für mindestens ordinalskalierte Variablen beispielsweise der **minimale Wert** (engl. *minimum*), der **maximale Wert** (engl. *maximum*) und die **Spannweite** (engl. *range*) als Differenz dieser beiden Werte festlegen.

Schiefe und Wölbung

Um eine Einschätzung darüber zu erhalten, ob eine Verteilung symmetrisch ist, kann die Maßzahl der **Schiefe** (engl. *skewness*) in der folgenden Form ermittelt werden:

$$Schiefe = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right)^3$$

Man teilt also die Differenzen zwischen den einzelnen Werten und dem Mittelwert durch die Standardabweichung, erhebt diese Quotienten in die 3. Potenz, summiert diese Größen und teilt die resultierende Summe durch die Anzahl der Merkmalsträger. Ergibt sich der Wert der Schiefe nahezu bei 0, so liegt Symmetrie vor. Ist das Ergebnis ein (größerer) positiver Wert, so ist die Verteilung linksschief, d.h. die Verteilungskurve fällt nach rechts langsamer als nach links ab. Ergibt sich dagegen ein (größerer) negativer Wert, so handelt es sich um eine rechtsschiefe Verteilung.
Ein Vergleich einer Verteilung mit einer entsprechenden Normalverteilung lässt sich dadurch vornehmen, dass die Wölbung (engl. *kurtosis*) in der folgenden Form berechnet wird:

$$Wölbung = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^4 - 3$$

Ist eine Verteilung genauso gewölbt wie die mit ihr korrespondierende Normalverteilung, so ergibt sich der Wert der Wölbung annähernd zu 0. Bei einem (größeren) positiven Wert ist die Verteilung mehr um das Zentrum konzentriert, als bei einer vergleichbaren Normalverteilung. Bei einem (größeren) negativen Wert verläuft die Verteilungskurve vergleichsweise flacher.

Analysemethoden

- Die **univariate Analyse** wird zur Gewinnung von Informationen über Variablen eingesetzt, und zwar einer jeden für sich, unabhängig von den anderen. Hier kommen hauptsächlich grafische Darstellungen (Histogramme, Boxplots, Sektorendiagramme, Stabdiagramme, Kartogramme, ...) zur Anwendung.

- Die **bivariate Analyse** untersucht die wechselseitige Abhängigkeit zweier Variablen; diese Untersuchung erlaubt, Kausalbeziehungen in Betracht zu ziehen. Die Analyseart hängt vom Variablentyp (qualitativ oder quantitativ) ab. Verwendung finden hauptsächlich grafische Darstellungen (Streudiagramme, schraffierte Tableaus, Stereogramme, Regressionskurven, ...).

- Die **multivariate Analyse** gibt eine globale Darstellung des Zusammenhangs von mehreren Variablen. Voraussetzung für die Anwendung der Verfahren ist, dass der Anwender vorab eine Vorstellung hat, welche Variablen auf andere einwirken. Er muss also die von ihm betrachteten Variablen in abhängige und unabhängige Variablen einteilen können.

Erklärung

Ich versichere hiermit, dass ich diese Arbeit selbstständig verfasst und nur die angegebenen Hilfsmittel verwendet habe.

.....
Unterschrift