Universität-Gesamthochschule Paderborn

Fachbereich 17

Umsatzprognose im Lebensmitteleinzelhandel mit Hilfe von Data Mining Methoden

Diplomarbeit im
Fachbereich Informatik

vorgelegt von: Mischa Kuchinke

1. Gutachter: Prof. Dr. Hans Kleine Büning

2. Gutachter: Prof. Dr. Wilfried Hauenschild

Inhalt

1 Einleitung

Im Rahmen dieser Diplomarbeit werden Methoden zur Erstellung von Umsatzprognosen für den Lebensmitteleinzelhandel analysiert und bewertet. Zu diesem Zweck bietet das erste Kapitel eine kurze Einführung in die Thematik.

Im zweiten Kapitel werden alle relevanten Methoden identifiziert und ihre Vorgehensweise erläutert. Insbesondere werden die theoretischen Grundlagen der Hauptwerkzeuge Regression und Assoziationsregeln erörtert. Des Weiteren werden zusätzliche WED[1] Werkzeuge vorgestellt, die im Rahmen dieser Untersuchung ebenfalls zum Einsatz kommen.

Das dritte Kapitel beschreibt das Vorgehen dieser einzelnen Methoden, und wie aus ihrer Anwendung Erkenntnisse gewonnen werden können. Dies beinhaltet die Modellbildung, Bewertung und Gegenüberstellung der Daten mittels Regression, Assoziationsregeln und CBR[2].

Das vierte Kapitel identifiziert alternative Vorgehensweisen und Erweiterungen, um die Prognosefähigkeit zu verbessern. Es werden Ansatzpunkte für weitergehende Untersuchungen aufgezeigt und die Nützlichkeit anderer WED-Werkzeuge für die Umsatzprognose getestet.

Die Ergebnisse dieser Untersuchungen werden im fünften Kapitel abschließend zusammengefasst. Außerdem bietet dieses Kapitel einen Ausblick auf die möglichen weiteren Thematiken, die auf der Grundlage dieser Diplomarbeit aufbauen.

Das folgende Kapitel soll die Thematik und die Vorgehensweise dieser Untersuchung erläutern.

1. WED: Wissensentdeckung in Datenbanken bzw. KDD: Knowledge Discovery in Databases
2. CBR: Case-Based-Reasoning

1.1 Problemstellung

Die Prognose des Umsatzes im Lebensmitteleinzelhandel ist ein signifikantes aber auch komplexes Problem.

Die Signifikanz einer Umsatzprognose ergibt sich aus der Tatsache, dass ein großer Teil der Personalkosten direkt aus den Umsatzdaten ableitbar ist. Dies gilt zum Beispiel für den Kassenbereich, den Bereich der Nachfüllarbeiten und für die Wurst-, Fleisch-, und Käsethekenbesetzungen. Eine gezielte Umsatzprognose eröffnet die Möglichkeit einen genaueren und feineren Einsatzplan zu erstellen.

Ein weiterer Grund für die Bedeutung der Umsatzprognose liegt in der Möglichkeit, hierdurch den Erfolg von Sonderaktionen zu bewerten, da der Umsatz höher liegen sollte als prognostiziert. Auf der anderen Seite fungiert die Umsatzprognose als Alarmsystem, falls der Umsatz sich niedriger als erwartet erweist.

Die Komplexität der Umsatzprognose begründet sich aus der Menge von Faktoren, die in dem vielschichtigen System des Erwerbs und Angebots von Lebensmitteln zusammenspielen. Beispielsweise seien hier die Faktoren Feiertage, saisonelle Schwankungen, Trends, Wetter, Urlaubszeiträume, Weihnachtszeiträume, Sonderaktionen, Erreichbarkeit des Betriebes und Liquidität der Kunden angegeben.

1.2 Ziel der Untersuchung

Das Ziel der Untersuchung besteht darin die Struktur und Zusammenhänge der Umsatzdaten eines Betriebes über einen definierten Zeitraum darzustellen. Damit sollen Anhaltspunkte aufgezeigt werden, wie eine erfolgreiche Umsatzprognose im Lebensmitteleinzelhandel durchgeführt werden kann. Das erarbeitete Modell soll sowohl mit den Daten eines Lebensmittelladens, als auch mit den Daten zweier weiterer Betriebe des gleichen Zeitraums optimale Ergebnisse liefern.

1.3 Grundlegende Methoden der Untersuchung

Als Grundlage für diese Diplomarbeit dienen die Methoden der WED. Dieses Gebiet beschäftigt sich mit dem Informationsgewinn aus Massendaten und enthält daher Werkzeuge, die für diese Untersuchung geeignet sind. Hierbei sind besonders die Methoden der Regression, der Assoziationsregeln und des CBR von Bedeutung und bilden daher die Hauptbestandteile dieser Untersuchung.

Die Regressionsanalyse ist eine der ältesten Methoden der WED und daher sehr ausgereift. Weiterhin ermöglicht sie die Prognose numerischer Werte. Die Ergebnisse des Regressionsmodells können deshalb direkt als Prognoseergebnis verwendet werden.

Das Gebiet der Assoziationsregeln ist ein junges Gebiet und setzt eine Klassifizierung der Umsatzdaten voraus. Allerdings sind die erzeugten Regelmodelle leichter verständlich und direkt mit den Erfahrungswerten der Einzelhändler vergleichbar.

Der CBR Ansatz ist sehr direkt und intuitiv. Die grundsätzliche Vorgehensweise hierbei stellt sich wie folgt dar: „Zur Bewertung eines unbekannten Vorfalls, benutze alle Ereignisse der Vergangenheit, die dem neuen Vorfall möglichst ähnlich sind.“ Zum einen ist dieser Ansatz intuitiv, weil er der menschlichen Vorgehensweise zur Bewertung unbekannter Daten ähnelt. Zum anderen ist das Ergebnis dieser Vorgehensweise leicht erklärbar, da Vergleichsfälle aus der Vergangenheit angegeben werden können, die die Bewertung unterstützen.

1.4 Datenbasis der Untersuchung

Die verwendeten Umsatzdaten für die Untersuchungen stammen aus den Datenbeständen dreier Lebensmittelläden der REWE Dortmund eG. Sie beziehen sich auf die Jahre 1997-1999. Die Daten konnten, mit der Genehmigung der Einzelhändler, anonymisiert verwendet werden. Ein Datensatz besteht aus einem Tagesdatum und einem Umsatzwert. Um die Datensätze für die Untersuchungen vorzubereiten, sind aus jedem Tagesdatum neue Merkmale extrahiert worden, die einerseits den auf das jeweilige Datum fallenden Wochentag wiedergeben und andererseits eine Angabe darüber machen, ob dieser Wochentag in der Nähe eines Feiertages liegt. Diese Merkmale werden anschließend in binärer Form kodiert.

Falls nicht explizit angegeben, sind alle statistischen Untersuchungen mit dem Softwareprogramm SPSS für Windows Version 10 durchgeführt worden.

2 Theoretische Betrachtung der Werkzeuge des Bereichs WED

Das folgende Kapitel liefert einen Überblick über das Gebiet der WED. Zunächst erfolgt eine Abgrenzung der Aufgabenbereiche voneinander, indem Probleme aufgezählt werden, die mit den Methoden der WED lösbar sind. Eine Übersicht der wichtigsten Methodenbereiche der WED und eine grobe Klassifizierung schließen sich an. Eine Zuordnung der Methoden zu den Aufgabenbereichen wird nicht durchgeführt, da zum Lösen eines Aufgabentyps im Allgemeinen ein Großteil der Algorithmen der WED zur Bearbeitung geeignet sind.

Die generelle Übersicht wird durch eine ausführliche und theoretische Darstellung der wichtigsten Werkzeuge in dieser Untersuchung komplettiert:

1. Regression, als ein Standardverfahren zur Erstellung von Prognosen.
2. Assoziationsregeln, da hierin ein angemessenes der regellernenden Verfahren für die Umsatzprognose gesehen wird.

Ergänzend werden drei Methodenbereiche dargestellt, die aufgrund des begrenzten Umfanges dieser Arbeit einen sekundären Rang einnehmen:

1. Neuronale Netze kommen in dieser Untersuchung zum Einsatz, um die Möglichkeit der Modellverbesserung zu überprüfen.
2. Mittels der WED-Methode Clustering wird die Struktur der Daten untersucht.
3. Statistische Tests sind Standardverfahren, die zum einen innerhalb der Regression zur Prüfung des Modells eingesetzt werden, zum anderen um eine Voruntersuchung der Daten durchzuführen.

2.1 Historische Einordnung

Der wissenschaftliche Bereich der WED beschäftigt sich mit dem Problem, dass heutige Datenbanken „[...] längst das Maß menschlicher Überschaubarkeit und manueller Analysetechniken überschritten [...]“([NRW98], S. 2) haben. Erst die computergestützte Analyse bietet eine Möglichkeit neue Muster in den Daten aufzufinden, die

möglichst nützlich und verständlich sind. Hierbei wird ein Teilbereich der WED, die direkte Anwendung von Algorithmen auf Datenmengen, mit dem Begriff „Data Mining“ bezeichnet.

Der Begriff Data Mining erscheint bereits 1978 in einer Arbeit von Leamer: „This book is about data mining. It describes how specification searches can be legitimately used to bring to the surface the nuggets of truth that may be buried in the data set.“ [Lea78]

Allerdings werden die klassischen statistischen Verfahren, die ebenfalls zum Bereich der WED gehören, wie zum Beispiel die Regression, bereits zu Beginn des 20. Jahrhunderts entwickelt.

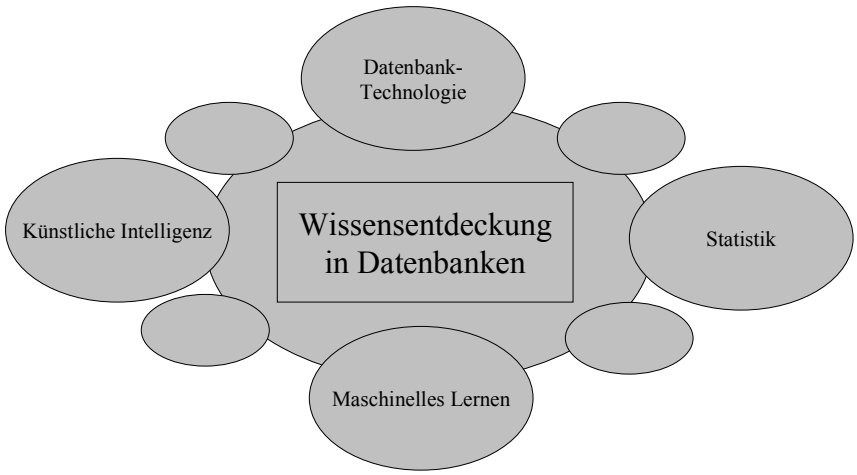

Abbildung 2.1: Das Gebiet der Wissensentdeckung in Datenbanken als interdisziplinäre Wissenschaft (entnommen aus [NRW98], S. 2)

Charakteristisch für den Bereich WED ist der interdisziplinäre Einfluss aus den verschiedensten Bereichen, der notwendig gewesen ist, um die Datenmengen heutiger Systeme effizient bewältigen zu können (siehe Abbildung 2.1).

Eine weitere Eigenschaft der WED ist die Vielfältigkeit der Verfahren und die Abhängigkeit vom Anwender. Zur Lösung eines WED-Problems sind verschiedene Schritte erforderlich, die zum Teil mehrfach zu durchlaufen sind, menschliche Entscheidungen erfordern und zum großen Teil vom Anwender unterstützt werden müssen.

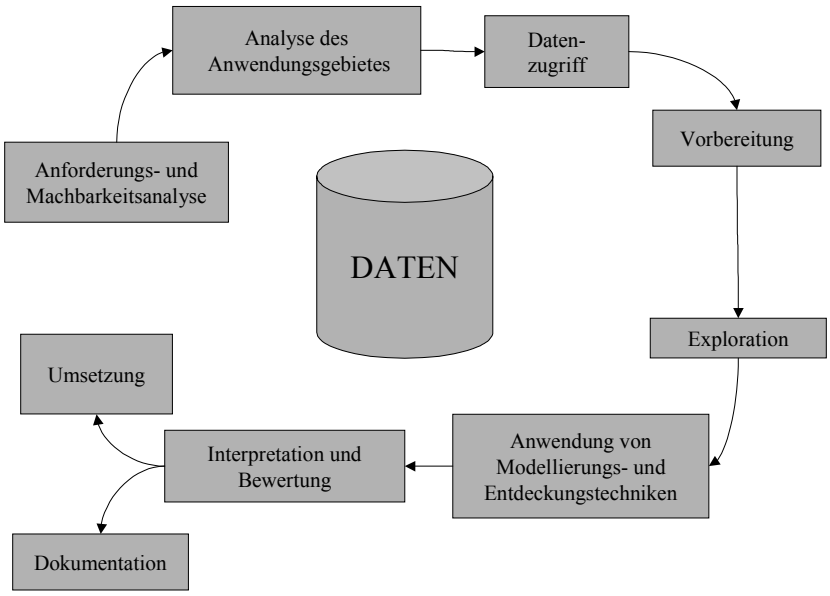

Abbildung 2.2: Überblick über den Prozess der Wissensentdeckung (entnommen aus [NRW98], S. 4)

Aus diesem Grund ist es von Vorteil solch ein Problem interaktiv zwischen Mensch und Maschine zu lösen. Der Anwender benötigt sowohl Kenntnisse aus dem Bereich WED als auch Kenntnisse über das Problemgebiet an sich. Die wichtigsten Schritte sind hierbei in der Abbildung 2.2 dargestellt. In Kapitel 3 und 4 wird zur Veranschaulichung auf diese Abbildung verwiesen, um aufzuzeigen, welcher Schritt zur Zeit durchgeführt wird.

2.2 Ein Überblick über den Bereich der WED

Um den Bereich WED zu umschreiben und zu klassifizieren, werden im Folgenden die verschiedenen Aufgaben der Wissensentdeckung aufgelistet. Die Einteilung ist aus dem Buch „Data Mining“ [NRW98] von Gholamreza Nakhaeizadeh übernommen worden. Da das Gebiet WED umfangreich und dynamisch ist, wird hier kein Anspruch auf Vollständigkeit erhoben.

Bei der Bearbeitung einer Aufgabe muss berücksichtigt werden, dass, während der Durchführung des WED Prozesses, verschiedene Fragestellungen aufgeworfen werden. Weiterhin ist zu berücksichtigen, dass die Beantwortung einer Fragestellung, die Lösung anderer neuer Fragestellungen voraussetzen kann und hierzu entweder eine Methode ausgewählt oder mehrere nacheinander ausprobiert werden müssen.

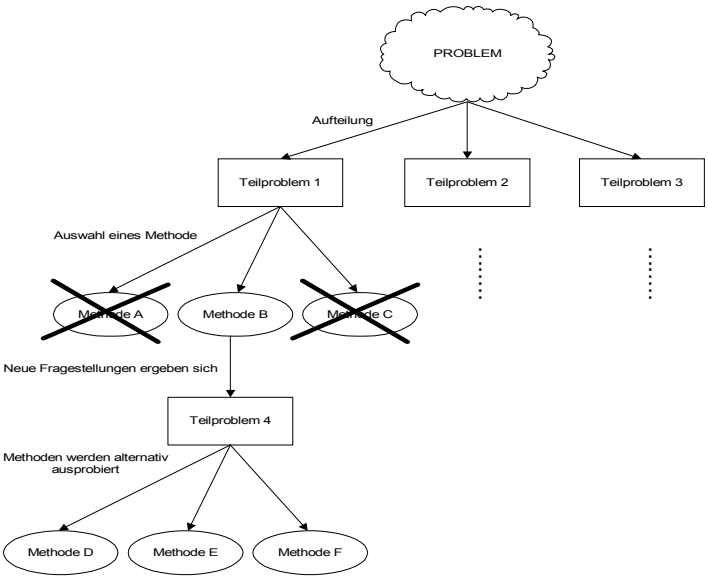

Abbildung 2.3: Darstellung eines WED-Prozesses

Hierbei ist es notwendig, dass der Benutzer während der gesamten Durchführung den Lösungsprozess aktiv unterstützt. Das methodische Vorgehen zur Lösungsfindung ist beispielhaft in Abbildung 2.3 dargestellt.

2.2.1 Aufgaben der WED

Aufgabe Segmentierung

Bei einer Segmentierung sollen die Daten in bestimmte Teilmengen bzw. Klassen eingeteilt werden. Dies kann das Hauptziel einer Untersuchung sein. Oft handelt es sich dabei aber nur um eine Vorbereitung, z. B. wenn die interessanten Daten von den uninteressanten zu trennen sind oder die Einteilung für weitere Untersuchung notwendig ist.

Aufgabe Konzeptbeschreibung

Die Konzeptbeschreibung verfolgt das Ziel aus den Daten verständliche Klassen bzw. Konzepte herauszuarbeiten. Die Verständlichkeit des Konzeptes steht, im Unterschied zur Segmentierung, im Vordergrund, da die Segmentierung zwar Objekte zu Klassen zusammenfasst, die Einteilung jedoch nicht notwendigerweise verständlich ist.

Aufgabe Klassifikation

Die Klassifikation hat die Aufgabe, unbekannte Objekte einer Klasse zuzuordnen. Das Klassifizierungsschema kann dabei vorher bestimmt oder mit Hilfe der Segmentierung ermittelt werden. Während bei der Klassifizierung einem unbekannten Objekt eine Klasse zugeteilt wird, erfolgt bei der Prognose die Zuordnung eines numerischen Wertes.

Aufgabe Prognose

Im Gegensatz zur Klassifikation werden bei der Prognose keine Klassen, sondern numerische Werte vorhergesagt. Dies ist auch bei dem Problem der Umsatzprognose der Fall, da hier für ein unbekanntes Objekt, in diesem Fall ein beliebiger Tag in der Zukunft, ein Umsatzwert prognostiziert werden soll.

Aufgabe Datenbeschreibung und Zusammenfassung

Bei dieser Aufgabe sollen wichtige Merkmale der Daten in kompakter Form beschrieben werden. In diesem Zusammenhang bietet die deskriptive Statistik wichtige Methoden, die zum Bereich der Datenbeschreibung und Zusammenfassung gezählt werden können. Allerdings zählen diese Methoden nicht zu dem Kernbereich der WED, können aber sehr gut zur Voranalyse und Präsentation eingesetzt werden. Als einfache Beispiele sind Summen und Durchschnittsbildung, sowie die Ermittlung der Verteilung zu nennen.

Aufgabe Erkennung von Abweichungen

Eine Abweichung ist gegeben, wenn ein Objekt Charakteristika besitzt, die einer vom Anwender festgelegten Norm nicht entspricht. Die Ermittlung solcher Objekte lassen zum einen fehlerhafte Eingaben oder Ausreißer erkennen. Zum anderen können solche Abweichungen auf unbekannte Phänomene hindeuten, die weiterer Untersuchung bedürfen.

Aufgabe Abhängigkeitsanalyse

Diese Analyse bestimmt die Abhängigkeiten der Merkmale von Objekten zueinander. Besteht zwischen zwei Merkmalen eine starke Abhängigkeit, so kann bei Bekanntheit eines Merkmals das andere vorausgesagt werden. Besitzen die Objekte eine Reihenfolge bzw. eine zeitliche Komponente, so können Abhängigkeiten ebenfalls zwischen räumlich oder zeitlich benachbarten Objekten festgestellt werden.

2.2.2 Beispielhafte Einordnung der Techniken der WED

In diesem Kapitel werden fünf Algorithmengruppen der WED vorgestellt und aufgrund ihrer Eigenschaften klassifiziert. Diese Vorgehensweise soll einen Anhaltspunkt bieten, wie eine vollständige Klassifikation des Gebietes WED aussehen könnte. Wie oben beschrieben, erweist sich dieses Gebiet als zu umfangreich und dynamisch, um in dieser Untersuchung eine vollständige Klassifizierung vornehmen zu können. Die Techniken, die im Folgenden untersucht werden, sind:

- Regression
- Assoziationsregeln
- Neuronale Netze
- Clustering
- Entscheidungsbäume
- CBR als Klassifikation bzw. Prognose

Einordnung bezüglich der Struktur der Eingabe

Die erste Einordnung erfolgt durch die Beantwortung der Frage: „Wie viel Informationen über die Problemstruktur brauchen die Algorithmen?“ Im Folgenden wird eine Auflistung der Verfahren vorgenommen, beginnend mit den Verfahren, die den geringsten Strukturinformationsbedarf aufweisen, bis zu denjenigen, die den größten Bedarf aufweisen.

1. Clustering benötigt als einzige Information eine Bewertungsfunktion, die aussagt, wie stark sich zwei Mengen von Elementen ähneln. Mit Hilfe dieser Funktion fügt der Algorithmus beispielsweise die beiden Mengen zusammen, die sich am ähnlichsten sind, bis eine zufriedenstellende Aufteilung der Elemente erfolgt ist. Darüber hinaus existieren weitere Methoden, um Cluster in Datenmengen zu entdecken.
2. Der Ansatz CBR benötigt einerseits eine Bewertungsfunktion, um in der Vergangenheit ähnliche Fälle zu finden. Andererseits ist eine Funktion nötig, die aus diesen Fällen die Prognose beziehungsweise die Klassifikation berechnet.
3. Entscheidungsbäume enthalten Klassifizierungsregeln in der Form eines Baumes. Allerdings kann diese Struktur jegliche Klassifizierung beinhalten. Der bekannteste Algorithmus aus diesem Bereich ist ID3 (siehe [Qui79]), der, im Gegensatz zu vergleichbaren Algorithmen, die Einschränkung besitzt, dass die Untersuchungseinheiten sich nicht widersprechen dürfen.

4. Assoziationsregeln enthalten als Grundstruktur eine Regelmenge. Eine Regel $A \rightarrow B$ besagt: Kommt in einer Untersuchungseinheit die Elementmenge A vor, kommt demzufolge mit einer angegebenen Wahrscheinlichkeit die Elementmenge B ebenfalls vor. Grundsätzlich können Regelmengen nicht jegliche Klassifizierung repräsentieren. Allerdings gibt es Erweiterungen (negative Assoziationen) des Basisalgorithmus Apriori (siehe [ZO98]), die eine, mit den Algorithmen der Entscheidungsbäume, vergleichbare Flexibilität ermöglichen.

5. Neuronale Netze ermitteln, vergleichbar mit der Regression, für eine vorgegebene Struktur die besten Parameter. Erreicht wird dies durch Approximation einer Funktion, welche die zu lernenden Werte möglichst genau abdeckt. Allerdings sind neuronale Netze ein heuristischer Ansatz, da diese nicht zwingend die optimalen Parameter, sondern nur suboptimale Parameter finden. Dies ist dadurch bedingt, dass die Struktur von neuronalen Netzen flexibler und vielschichtiger sein kann, als die Struktur von Regressionen.

6. Regressionen besitzen eine feste Struktur, die vorgegeben werden muss. Der Algorithmus ermittelt die optimalen Parameter für diese Struktur. Die Einschränkung, dass ein Modell vorgegeben werden muss, kann auch ein Vorteil sein, da hierdurch viel Wissen, das vielleicht über das Problem bereits vorhanden ist, in das Modell übernommen wird.

Einordnung bezüglich der Art der Zielvariable

Generell wird zwischen nominalen, ordinalen und metrischen Zielvariablen unterschieden. Ist die Zielvariable nominal oder ordinal, kann ein Klassifizierungsverfahren benutzt werden. Ansonsten kommt ein Prognoseverfahren zur Anwendung, das ebenfalls für Klassifikationsprobleme einsetzbar ist. Auf der anderen Seite ist ein Klassifikationsverfahren nicht in der Lage mit metrischen Werten zu arbeiten, es sei denn, diese werden zuvor klassifiziert. Regression und neuronale Netze können als Prognoseverfahren eingesetzt werden. Assoziationsregeln und Entscheidungsbäume sind dagegen Klassifikationsverfahren. Der Ansatz des CBR und des Clustering nehmen hier eine Sonderrolle ein, da diese Verfahren zwar ähnliche Fälle finden, die Verarbeitung dieser Fälle allerdings dem Anwender überlassen ist. Hiermit können prinzipiell sowohl Klassifikations- als auch Prognoseansätze erstellt werden.

Einordnung bezüglich der benötigten Informationen

Für die erste Einteilung lassen sich die Verfahren in strukturprüfende und struktursuchende Verfahren einteilen. Hierbei ist das Clustering und das CBR ein struktursuchendes Verfahren, alle anderen enthalten eine Struktur, die durch den Algorithmus geprüft und parametrisiert wird. Weiterhin können die strukturprüfenden Verfahren in Parameterverfahren und Regellernverfahren aufgeteilt werden. Regression und neuronale Netze sind Parameterverfahren, welche die Parameter der Struktur optimieren. Entscheidungsbäume und Assoziationsregeln sind Regellernverfahren, die versuchen, Regeln innerhalb der festgelegten Struktur zu entdecken und auszugeben. Eine graphische Darstellung dieser Einteilung ist in Abbildung 2.4 zu sehen.

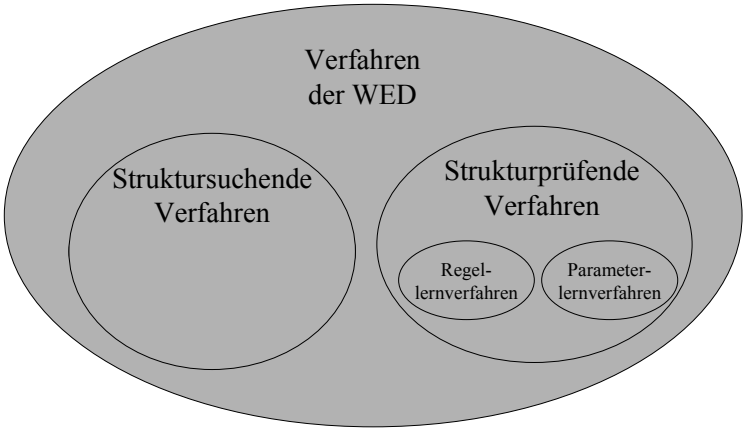

Abbildung 2.4: Verfahrenseinordnung bezüglich der benötigten Informationen

Einordnung bezüglich der Güte der Ausgabe

Wie oben erwähnt, liefern alle besprochenen Algorithmen im Rahmen ihrer Struktur optimale Ergebnisse, ausgenommen neuronale Netze, die aufgrund des heuristischen Ansatzes suboptimale Ergebnisse erreichen. Die Ausnahme bilden hierbei Clustering und CBR, da diese zu den struktursuchenden Verfahren zählen.

2.3 Regression

Das Verfahren der Regression ist eines der wichtigsten Prognoseverfahren im Bereich der WED. Zum einen, weil die Regression optimale Ergebnisse liefert und numerische Eingabewerte verarbeiten kann, zum anderen existiert das Verfahren schon seit über einhundert Jahren und zählt somit zu einem der am weitesten entwickelten Bereiche der WED. Für die Regression existieren mehrere Verfahren, welche die Güte und Vorhersagefähigkeit des Regressionsmodells bewerten.

In den folgenden Kapiteln werden nach einer historischen Einordnung des Verfahrens, die Algorithmen der linearen einfachen Regression und darauf aufbauend die Algorithmen der linearen multiplen Regression beschrieben und deren Korrektheit mathematisch bewiesen. Anschließend werden einige Tests zur Güte des Regressionsmodells erläutert.

Da die Erstellung einer Umsatzprognose ein Prognoseproblem darstellt, zeigt sich die Regression als ein wichtiges Werkzeug für diese Untersuchung.

Eine ausführliche Beschreibung der Regression findet sich in ([HEK82], S. 569).

2.3.1 Historische Einordnung

Geprägt wurde der Begriff „Regression“ durch Galton (1889), der in seinen Studien zur Vererbung das „Gesetz der universalen Regression“ formulierte: „Each pecularity in a man is shared by his kinsman but on the average in a less degree“([HEK82], S. 569).

Karl Pearson, ein Freund Galtons, untersuchte in diesem speziellen Zusammenhang, dass große Väter zwar dazu neigen große Söhne zu haben, die Größe der Söhne jedoch durchschnittlich geringer als die Größe der Väter ist. Auf der anderen Seite haben kleine Väter kleine Söhne, jedoch ist die Körperlänge der Söhne durchschnittlich größer als die der Väter.

Dieser Zusammenhang kann durch die Funktion $y = a + bx$, wobei y die Größe des Sohnes und x die Größe des Vaters bezeichnet, dargestellt werden (siehe Abbildung 2.5). Die „Regressionsanalyse“ der Galton-Pearson-Regression ergibt zum Beispiel einen mathematischen Zusammenhang von $y = 85,6742 + 0,561x$.

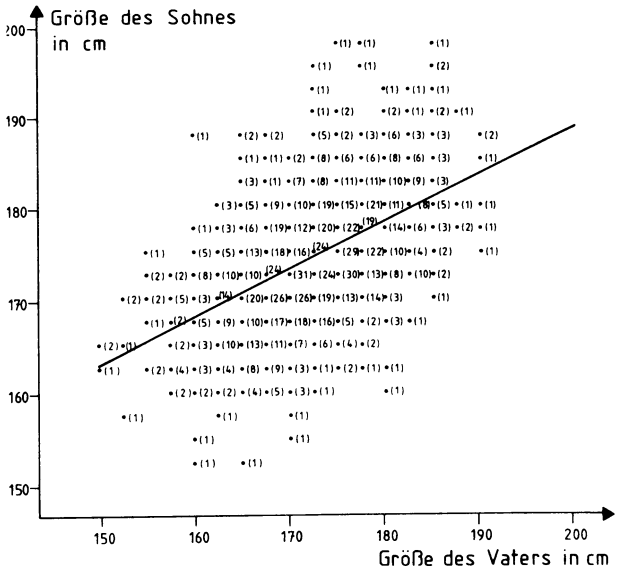

Abbildung 2.5: Galton-Pearson-Regression: Zusammenhang zwischen der Größe von Vätern (x) und ihren Söhnen (y) (Entnommen aus [HEK82], S. 570)

In der Regressionsanalyse wird der funktionale Zusammenhang zwischen einem Merkmal y (Regresssand) und einer Menge von Merkmalen $x_1,x_2,\ldots,x_n$ (Regressoren) $f(x_1,x_2\ldots,x_n) = y$ geschätzt. Bei der linearen Regression werden die Parameter $b_1,b_2,\ldots,b_n$ der Funktion $y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$ bezüglich der Fehlerfunktion: „Summe der Abweichungsquadrate“ optimal geschätzt.

$$S^2 = \sum_{i=1}^{n} (y_i - a - b_1x_{1i} - b_2x_{2i} - \ldots - b_kx_{ki})^2$$

Eine einfache Umbenennung von z. B. $x_1' = x_1^2$ ermöglicht es, beliebige polynominelle Zusammenhänge mit der Regression zu bestimmen.

Ein Spezialfall der linearen Regression ist hierbei die lineare einfache Regression, wenn genau ein Regressor vorhanden ist. Hierbei wird zwischen dem Regressanden und dem Regressor ein Zusammenhang $y = a + bx$ vermutet.

Graphisch betrachtet, ermittelt die lineare einfache Regression eine Gerade durch den Punkteraum, so dass die Summe der Quadrate der vertikalen Abstände der Punkte zur Gerade möglichst gering ist (siehe Abbildung 2.6).

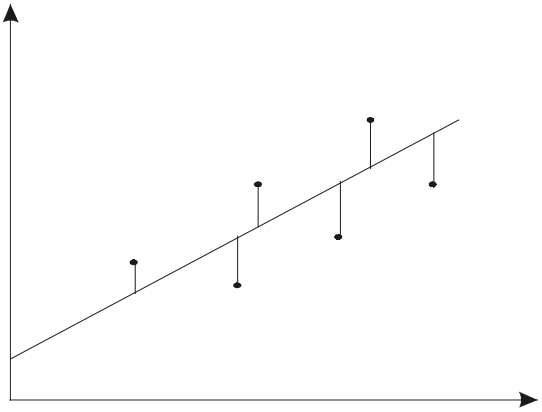

Abbildung 2.6: Graphische Veranschaulichung der Methode der kleinsten Quadrate (Entnommen aus [HEK82], S. 575)

2.3.2 Lineare einfache Regression

In diesem Kapitel wird der mathematische Hintergrund des Algorithmus erläutert und seine Korrektheit bewiesen.

Die Idee der linearen einfachen Regression besteht darin, eine Gerade zu finden, so dass die Summe der Abweichungsquadrate zwischen der Gerade und den Messpunkten minimal wird. Mathematisch wird dies dadurch erreicht, dass das Minimum der Fehlerfunktion über die Nullsetzung der Ableitung der Fehlerfunktion ermittelt wird.

$$\text{Fehlerfunktion} : S^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Falls zwischen zwei Merkmalen ein linearer Zusammenhang besteht oder vermutet wird, kann dieser mittels der linearen Regression näher bestimmt werden. Hierzu werden die Merkmale n-mal gemessen, so dass sich eine Messreihe $x_1, x_2, \ldots, x_n$ und $y_1, y_2, \ldots, y_n$ ergibt. Aufgrund der linearen Vermutung ist davon auszugehen, dass $y_i = \alpha + \beta x_i + e_i$ gilt. Hierbei bezeichnet α das Absolutglied, β den Steigungsparameter und $e_1, e_2 \ldots, e_n$ zufällige Fehler.

Aufgrund der Tatsache, dass bei einem Minimum einer Funktion die Ableitung der Funktion an derselben Stelle Null beträgt, ergeben sich folgende Gleichungen für a und b (siehe Anhang „Formel 1“ auf Seite 86 und „Formel 2“ auf Seite 87):

$$a = \bar{y} - b\bar{x}$$

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

Diese Parameter garantieren, dass die Summe der Abweichungsquadrate minimal ist, da die quadratische Fehlerfunktion ein Minimum und kein Maximum besitzt.

2.3.3 Lineare multiple Regression

Die lineare multiple Regression ist eine Verallgemeinerung der linearen einfachen Regression. Bei der multiplen Regression geht man vom folgenden Modellansatz aus:

$$f(x_1,x_2,\ldots,x_k) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k = \alpha + \sum_{j=1}^{n} \beta_j x_j = y$$

Das Merkmal y (Regressand) ist abhängig von den Regressoren $(x_1,x_2,\ldots,x_k)$.

Hierzu werden die Merkmale n-mal gemessen, so dass sich für den Regressand y eine Messreihe $y_1,y_2,\ldots,y_n$ und für jeden Regressor x_j sich eine Messreihe $x_{j1},x_{j2},\ldots,x_{jn}$ ergibt.

Um die optimalen Parameter $a,b_1,b_2,\ldots,b_k$ zu ermitteln, wird die bei der linearen einfachen Regression definierte Summe der Abweichungsquadrate auf multiple Regressoren erweitert:

$$S^2 = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki})^2$$

Da diese Summe minimiert werden soll, wird für die verschiedenen Parameter $a,b_1,b_2,\ldots,b_k$ die Ableitung der Summe auf Null gesetzt. Man erhält das im Anhang unter „Formel 3“ auf Seite 88 und „Formel 4“ auf Seite 89 angegebene Gleichungssystem, welches allerdings einfacher mit folgender Matrizenschreibweise beschrieben wird:

$$\tilde{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \ldots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & \ldots & x_{k2} \\ 1 & x_{13} & x_{23} & \ldots & x_{k3} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{kn} \end{bmatrix}, e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \ldots \\ e_n \end{bmatrix}, \beta = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \ldots \\ \beta_k \end{bmatrix}$$

Somit lässt sich das grundlegende Modell schreiben als:

$$\tilde{Y} = X\beta + e$$

Hieraus ergibt sich:

$$X^T X b = X^T \tilde{Y}$$

Im Anhang „Formel 5“ auf Seite 91 wird die Gleichheit der hier angegebenen Formeln bewiesen.

Falls nun $X^T X$ invertierbar ist, demnach die Merkmale untereinander unabhängig sind, ergibt sich:

$$b = (X^T X)^{-1} X^T \tilde{Y}$$

als eindeutiger Schätzer für β.

Anhand dieser Formel können die Parameter $a, b_1, b_2, \ldots, b_k$ optimal geschätzt werden.

2.3.4 Güte des Regressionsmodells

Um die Güte eines Regressionsmodells festzustellen, werden vier verschiedene Methoden vorgestellt. Wie oben beschrieben, erzeugt die Regression zwar optimale Ergebnisse, es kann aber trotzdem sein, dass das Modell zur Erklärung der Daten nicht ausreicht oder die Anzahl der vorhandenen Daten zu gering ist. Weiterhin besteht die Möglichkeit, dass einzelne Regressanden nicht signifikant auf die Ergebnisse eingewirkt haben und somit im Regressionsmodell vernachlässigt werden können.

Bestimmtheitsmaß

Das Bestimmtheitsmaß beantwortet die Frage, inwiefern sich die Varianz des Regressors Y durch das Regressionsmodell erklären lässt oder andere statistisch signifikante Einflüsse bestehen, die nicht berücksichtigt sind?

Zur Beantwortung wird der Anteil der Varianz der geschätzten Werte zur Varianz der beobachteten Werte berechnet. Dies wird als das Bestimmtheitsmaß der Regression bezeichnet:

$$B_{Y,X} = \frac{S_{\hat{Y}}^2}{S_Y^2} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = r_{XY}^2$$

Ist $B_{Y,X}$ gleich 1, bedeutet dies, dass die Varianz des Merkmals Y vollkommen durch die Regression erklärt wird. Für eine ausführliche Beschreibung des Bestimmtheitsmaßes (siehe [HEK82], S. 578).

T-Test

Der T-Test berechnet für jeden Regressanden X_i die statistische Wahrscheinlichkeit, ob dieser Regressand einen Einfluss auf das Regressionsmodell hat oder nicht. Anhand eines statistischen Tests wird ermittelt, ob die Nullhypothese „Der Wert b_i ist gleich Null!“ verworfen werden muss oder nicht. Besagt der T-Test, dass die Hypothese verworfen werden muss, so ist dies mit einer Wahrscheinlichkeit von $1-\alpha$ richtig. Der untersuchte Wert ist wahrscheinlich ungleich Null und daher für das Regressionsmodell statistisch signifikant. Die Hypothese wird verworfen, wenn $\left|\frac{b_i}{s_{b_i}}\right| > t_{n-1;(1-\alpha/2)}$ gilt. Hierbei benennt b_i den zu untersuchenden Wert, s_{b_i} ist die Varianz des Wertes und $t_{n-1;(1-\alpha/2)}$ sind die Quantile der t_{n-1} Verteilung (siehe [HEK82], S. 179).

F -Test

Der F-Test prüft die Nullhypothese „Alle Werte sind gleich Null!“. Demzufolge wird die Wahrscheinlichkeit überprüft, ob die beobachteten Werte unabhängig von den Regressanden, demnach nur zufällig sind oder von anderen, als den bisher untersuchten Merkmalen, beeinflusst werden. Wie beim T-Test gilt auch hier, wird die Nullhypothese verworfen, ist die Entscheidung zu $1-\alpha$ richtig. Die Werte sind dann nicht zufällig.

Das Verhältnis der durchschnittlichen Abweichung der Ausgangswerte zu den durchschnittlichen Abweichungen der Fehlerwerte wird beim F-Test als Prüfgröße genommen. Ist das Verhältnis größer als $F_{p-1,N-p;1-\alpha}$, muss die Nullhypothese verworfen werden. Hierbei bezeichnet p die Anzahl der Werte und N die Anzahl der Messungen. Benutzt wird die entsprechende F-Verteilung (siehe [HEK82], S. 611).

Residualanalyse

Ein graphisches Verfahren zur Beurteilung einer Regression ist die Residualanalyse. Als Residuen werden die Differenzen zwischen berechneten und geschätzten Werten $\hat{e}_i = y_i - \hat{y}_i$ bezeichnet. In der Residualanalyse werden die Residuen mit dem Schätzer s

der Standardabweichung der Messwerte normiert $d_i = \frac{y_i - \hat{y}_i}{s}$ und graphisch dargestellt. Ist der Modellansatz richtig, sind die Residuen aufgrund der Normierung approximativ, unabhängig, identisch N(0,1)-verteilt.

Beispiel von Residualanalysen:

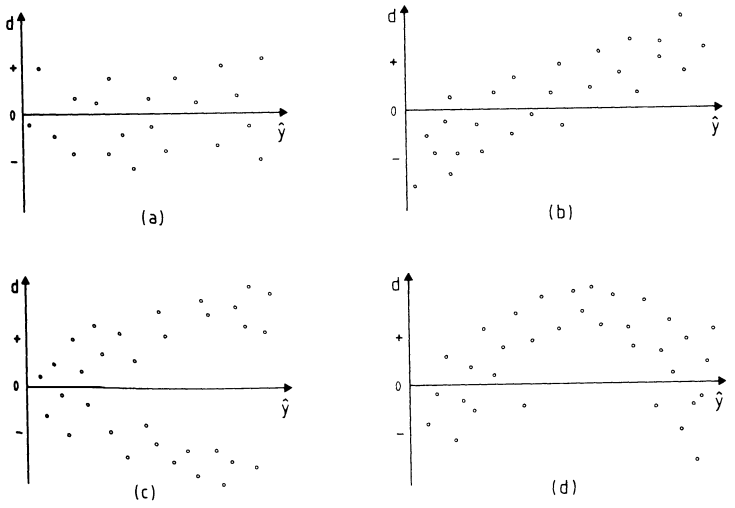

Abbildung 2.7: Graphische Darstellung der normierten Residuen in Abhängigkeit von den geschätzten Werten $\hat{y}$ des Regressanden (Entnommen aus [HEK82], S. 585)

Im Feld a) der Abbildung 2.7 lässt sich ein idealer Verlauf erkennen, b) deutet auf einen linearen Trend hin, der nicht berücksichtigt wurde, c) enthält ansteigende Varianzen, die auf ein anderes Modell hindeuten und d) zeigt einen nichtlinearen Verlauf, so dass die Daten innerhalb einer nichtlinearen Regression untersucht werden sollten (siehe [HEK82], S. 585).

2.3.5 Laufzeiten und Effizienz der Regression

Die Laufzeit der Regression ist abhängig von der Invertierung und Multiplikation der Matrix. Hierbei bezeichnet n die Anzahl der Datensätze, die bearbeitet werden und m die Anzahl der Merkmale. Die Multiplizierung der Matrix ist in $O\langle n+m^2\rangle$ und die Invertierung in $O\langle m^3\rangle$ möglich. Die Regression ermittelt für das gegebene Modell optimale Werte. Im Umkehrschluss bedeutet ein falsches Modell den Fehlschlag der gesamten Regression. Es gibt Methoden die Qualität des Modells zu testen und Verbesserungsansätze erkennbar zu machen. Allerdings gibt es keinen Algorithmus, der automatisch aus den Daten das optimale Modell berechnet.

2.4 Assoziationsregeln

Der Bereich Assoziationsregeln versucht in den Daten Regeln zu erkennen, die darauf verweisen, welche Merkmale häufig mit anderen Merkmalen auftreten. Aus diesem Grunde werden in dieser Diplomarbeit Assoziationsregeln verwendet. Der Umsatzbereich kann durch solche Regeln ermittelt werden. Zum Beispiel: „Ist heute Freitag und der morgige Tag ein Feiertag, dann ist der Umsatz sehr hoch!“. Allerdings sind Assoziationsregeln für den Bereich der Prognose bedingt geeignet, da es sich bei dem Algorithmus um einen Klassifikationsalgorithmus handelt. Um mit dem Algorithmus arbeiten zu können, muss daher das Merkmal Umsatz in Klassen eingeteilt werden. Dies ist allerdings bei den anderen Merkmalen (Wochentag und Feiertag) nicht nötig, da diese bereits nominal bzw. ordinal sind. Da sich die Erfahrungen der Einzelhändler ebenfalls leicht in diese Art von Regeln einpassen lassen, wird der oben genannte Ansatz in diese Untersuchung aufgenommen.

In den folgenden Kapiteln erfolgt eine kurze historische Einordnung der Assoziationsregeln. Der Basisalgorithmus wird erklärt und zum Schluss die neuesten Erweiterungen in diesem Bereich vorgestellt.

2.4.1 Historische Einordnung

Zum ersten Mal wird der Begriff „Assoziationsregeln“ im Bereich WED in der Arbeit „Mining Association Rules between Sets of Items in Large Databases“ von Rakesh Agrawal, Tomas Imielinski und Arun Swami geprägt [AIS93]. In dieser Arbeit wird das Problem aufgegriffen, dass die Speicherung von Daten sich von „global data“ zu „basket data“ entwickelt hat. So sind z. B. nicht mehr nur die Umsatzdaten über einen bestimmten Zeitraum verfügbar, sondern auch die Informationen, welche einzelnen Produkte innerhalb eines Einkaufs erworben werden. Diese moderne Datensammlung ermöglicht folgende Fragestellungen:

1. Welche Produkte werden zusammen mit einem anderen Produkt gekauft? Wie kann demnach der Verkauf von diesem Produkt durch die Verbilligung anderer Waren gesteigert werden?
2. Welche Produkte fördern den Verkauf eines anderen Produktes? Welche Auswirkungen hätte demzufolge die Streichung dieser Produkte?
3. Welche Produkte lassen sich gut mit anderen Produkten verkaufen, sollten daher räumlich eng platziert sein?

Diese gesamten Fragestellungen lassen sich auf das Prinzip der Assoziationsregeln zurückführen. Eine Assoziationsregel besteht aus zwei Warenmengen, die häufig zusammen gekauft werden. Die Güte der Regeln wird zum einen durch den Support (Wie statistisch signifikant ist die Regel?) und der Confidence (Wie stark ist die Gültigkeit der Regel?) angegeben.

Ein Beispiel:

Transaktionsdatenbank :

Transaktion	Gekaufte Ware
1	Milch, Butter, Wurst, Fisch
2	Butter, Käse, Fisch
3	Milch, Butter, Wurst, Fisch
4	Milch, Butter, Käse, Fisch
5	Milch, Butter, Käse, Wurst, Fisch
6	Butter, Käse, Wurst

Support aller Warenmengen ≥ 50% :

Support	Warenmengen
100 % (6 mal gekauft)	(Butter)
83 % (5 mal gekauft)	(Fisch), (Butter, Fisch)
67 % (4 mal gekauft)	(Milch), (Käse), (Wurst), (Fisch), (Milch, Butter), (Milch, Fisch), (Butter, Käse), (Butter, Wurst), (Milch, Butter, Fisch)
50 % (3 mal gekauft)	(Milch, Wurst), (Käse, Fisch), (Wurst, Fisch), (Milch, Butter, Wurst), (Milch, Wurst, Fisch), (Butter, Käse, Fisch), (Butter, Wurst, Fisch), (Milch, Butter, Wurst, Fisch)

Assoziationsregeln mit Confidence = 100%

Milch → Butter (4/4)	Milch,Butter → Fisch (4/4)	Wurst, Fisch → Butter (3/3)
Milch → Fisch (4/4)	Milch,Wurst → Butter (3/3)	Milch, Käse → Butter, Fisch (3/3)
Milch → Butter, Fisch (4/4)	Milch,Wurst → Fisch (3/3)	Wurst, Fisch → Milch, Butter (3/3)
Käse → Butter (4/4)	Milch,Fisch → Butter (4/4)	Milch, Butter, Wurst → Fisch (3/3)
Wurst → Butter (4/4)	Käse, Fisch → Butter (3/3)	Milch, Wurst, Fisch → Butter (3/3)
Fisch → Butter (5/5)	Wurst, Fisch → Milch (3/3)	Butter, Wurst, Fisch → Milch (3/3)

Assoziationsregeln mit Confidence zwischen 80% und 100%

Fisch → Milch (4/5)	Butter → Fisch (5/6)	Fisch → Milch, Butter (4/5)

Abbildung 2.8: Beispieldatenbank und Assoziationsregeln mit einer Mindest-Confidence von 80% (siehe [ZO98], S. 2)

In der oberen linken Tabelle der Abbildung 2.8 sind sechs beispielhafte Einkäufe von Lebensmitteln aufgelistet. In der oberen rechten Tabelle wird jede mögliche Menge angegeben, die einen minimalen Support von 50% besitzt, demnach Mengen von Lebensmitteln, die bei mindestens der Hälfte der Transaktionen (mindestens drei) zusammen gekauft wird. Im unteren Teil sind die Assoziationsregeln angegeben, die mindestens eine Confidence von 80% besitzen. Dies bedeutet, dass in mindestens 80% aller Fälle, wenn die Prämisse eintrat, d. h. die Lebensmittelmenge vor dem Pfeil gekauft wird, es zur Konklusion, folglich zum Erwerb der Lebensmittelmenge hinter dem Pfeil kommt. Die erste Regel besagt: Wann immer Milch gekauft wird, zu 100% auch Butter gekauft wird.

2.4.2 Beschreibung des Lösungsansatzes

Nachfolgend wird der Basisalgorithmus am Beispiel der Untersuchung von Supermarktwarenkörben erläutert. Es werden alle Warengruppen gesucht, die mit anderen Warengruppen zusammen gekauft werden. Hierbei muss der Kauf der Warengruppen häufig genug vorkommen, damit Aussagen signifikant getroffen werden können. Als Ergebnis wird von allen Einkäufen, in denen die erste Warengruppe vorkommt, prozentual angegeben, wie oft die zweite Warengruppe gekauft wird. Der Vorteil des Algorithmus besteht darin, dass auch Massendaten heutiger Datenbanksystem effizient analysiert werden können. Die formale Problemdefinition ist im Anhang „Formale Problemdefinition von Assoziationsregeln“ auf Seite 92 zu finden.

Die Aufgabe des Algorithmus lautet: Finde alle Assoziationsregeln, deren $Support \geq MinSupport$ und deren $Confidence \geq MinConfidence$ ist. Hierbei bezeichnet der Support einer Regel $A \rightarrow B$, die Anzahl der Transaktionen, bei denen A und B zusammen gekauft wurden. Die Confidence einer Regel bezeichnet das Verhältnis, wie oft A und B zusammen gekauft werden, zu den Transaktionen in denen A gekauft wird. Der Support liefert daher einen Ansatzpunkt über die Aussagefähigkeit der Regel, während die Confidence eine Aussage darüber trifft, wie groß die Wahrscheinlichkeit für das Einkaufsverhalten ist. Eine Menge wird als frequent bezeichnet, wenn der Support der Menge über dem Wert $MinSupport$ liegt.

Die Algorithmen zur Berechnung von Assoziationsregeln, beschäftigen sich hauptsächlich mit dem Problem der Ermittlung des Supports der Mengen, die frequent sind. Mit der Information über den Support lassen sich alle Assoziationsregeln und deren Confidence ohne größeren Aufwand ermitteln. Daher wird zuerst ein Algorithmus beschrieben, der alle frequenten Mengen und deren Support berechnet. Anschließend wird gezeigt, wie daraus Assoziationsregeln ermittelt werden.

Algorithmus Apriori

Der Algorithmus Apriori berechnet für alle möglichen Warenmengen der Reihe nach, wie oft diese in den Transaktionen enthalten sind. Hierbei wird zuerst die Anzahl der Transaktionen mit einer bestimmten Ware berechnet. Dies ist gleichbedeutend mit der Frage: „Wie oft wurde eine bestimmte Ware gekauft?“ Des Weiteren werden die Häufigkeiten von Transaktionen mit zwei bestimmten Waren für alle Kombinationen berechnet usw.

Der folgende Algorithmus nutzt die Tatsache, dass alle Untermengen einer Menge, die frequent ist, ebenfalls frequent sind. Zwei Optimierungen sind hierdurch realisierbar:

1. Neue Mengen mit n Elementen können aus den frequenten Mengen der Größe n-1 ermittelt werden.
2. Mengen, die eine Untermenge haben, die nicht frequent ist, können ebenfalls nicht frequent sein und werden daher gelöscht.

Für weitere Informationen siehe [ZO98].

Algorithmenaufbau zur Berechnung der Signifikanz der Warengruppen

Der Algorithmus berechnet aus den frequenten Mengen mit k-1 Elementen, die frequenten Mengen mit k Elementen. Zu Beginn werden daher die frequenten Mengen der Größe eins durch nachzählen ermittelt.

Sind alle frequenten Mengen der Größe k-1 bekannt, werden alle Kombinationen von Mengen gebildet, die k-2 Elemente gemeinsam haben. Aus diesen beiden Mengen werden neue Mengen gebildet, die genau k Elemente besitzen.

Weiterhin wird jeder dieser Kandidaten überprüft, ob jede seiner Untermengen zu den bekannten frequenten Mengen der Größe k-1 gehört. Gibt es eine Untermenge, die nicht frequent ist, kann der Kandidat selber nicht mehr frequent sein und wird nicht weiter betrachtet.

Nun werden alle verbleibenden Kandidaten durchgezählt, wobei die nicht frequenten gelöscht werden.

Algorithmus zur Berechnung der Kausalität der Regeln

1. Berechne für jede Menge $Y \in L_k$ für $k = 2,\ldots,n$ und jeder Untermenge $X \in Y$ die Confidence der Regel $X \rightarrow (Y - X) = \frac{Support(Y)}{Support(X)}$.
2. Falls $Support \geq MinSupport$ gilt, genügt die Regel den Anforderungen und kann ausgegeben werden.

2.4.3 Erweiterungen des Basisalgorithmus

In den folgenden Jahren sind mehrere Erweiterungen zu dem oben beschriebenen Basisalgorithmus veröffentlicht worden. Einige Beispiele von Erweiterungen folgen in den nächsten Abschnitten.

Integration von Gliederungsinformationen

Gliederungsinformationen, wie in Abbildung 2.9 dargestellt, helfen die Daten und Regeln aussagekräftiger zu gestalten, indem Aussagen über höhere Gliederungsebenen getroffen werden können. Z. B. ist die Aussage „50% der Leute, die Milch kaufen, kaufen auch Brot“ allgemeiner und daher wertvoller, als die Aussage „50% der Leute, die 500 ml H-Milch der Marke y kaufen, kaufen auch Sonnenblumenbrot 500 gr.“

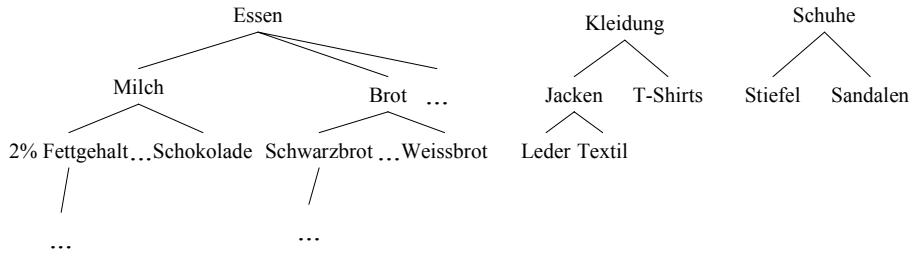

Abbildung 2.9: Beispielsgliederung (siehe [HF95a], S. 4)

Integration von Metaregeln

Metaregeln, wie in Abbildung 2.10 dargestellt, erlauben es Filter zu bestimmen, um nur die Regeln zu erkunden, die für den Anwender interessant sind. Zum Beispiel: Suche alle Regeln über bayrische Studenten, die vom Studienfach und einer anderen Eigenschaft auf bestimmte Eigenschaften schließen lassen. Hierbei sind die Eigenschaften Studienfach, Abschlussnote, Geburtsort, Vordiplomsnote und Adresse von Interesse (siehe [HF95b]).

Entdecke Regeln in der Form:
Studienfach(s : Student,x) $\wedge$ Q(s,y) $\rightarrow$ R(s,z)
From Students
Where Geburtsort = „Bayern“
In relevance to Studienfach, Abschlussnote, Geburtsort, Vordiplomsnote und Adresse
Beispielsregel:
Studienfach(s,“BWL“) $\wedge$ Vordiplomsnote(s,“1“) $\rightarrow$ Abschlussnote(s,“1“)

Abbildung 2.10: Beispiel von Metaregeln

Ausschließende Assoziationen

Das Konzept der ausschließenden Assoziationen, dargestellt in Abbildung 2.11, erweitert die Anzahl möglicher Regeln durch die Einbettung von Negationen. Hierdurch sind Erkenntnisse möglich wie: „50 % der Leute, die Milch und keine Cornflakes kaufen, kaufen Haferflocken“. Diese Erweiterung des Verfahrens bedeutet eine erhebliche Verschlechterung der Laufzeit, da die Menge möglicher Regeln exponentiell ansteigt (siehe [AFK97]).

$S_1 \wedge \quad S_2 \Rightarrow S_3$
Beispiel :
Milch $\wedge$ Cornflakes $\Rightarrow$ Haferflocken

Abbildung 2.11: Beispiel von ausschließenden Assoziationen

2.4.4 Laufzeiten und Effizienz von Assoziationsregeln

Obwohl der Algorithmus in der Lage ist viele Kandidaten zu eliminieren, ist die Worst-Case Laufzeit exponential, wie in Abbildung 2.12 zu sehen ist. Grund hierfür ist, dass prinzipiell alle Mengen von Elementen für eine Assoziationsregel geeignet sind. Eine prinzipiell geeignete Menge muss gezählt werden. Die Laufzeit ergibt sich aus dem Zusammenhang, dass die Anzahl der möglichen Mengen exponential zur Anzahl der Elemente ist.

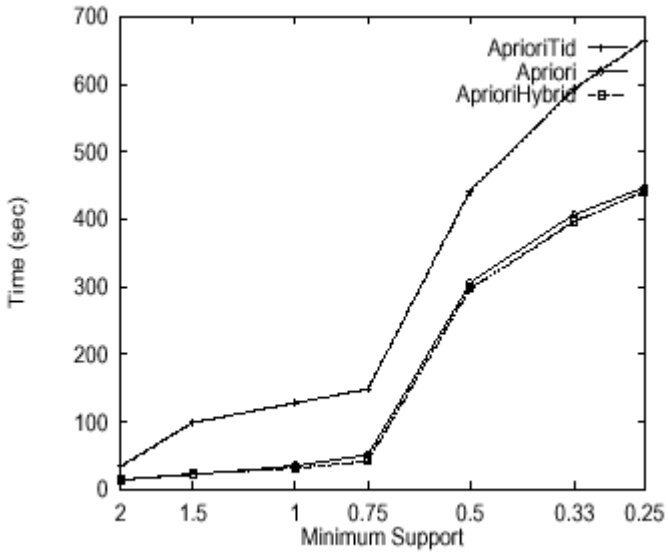

Abbildung 2.12: Beispiele von Laufzeiten im Verhältnis zum minimalen Support (siehe [HF95a])

Die Assoziationsregeln werden vollständig ermittelt, allerdings können durch diese Regeln nicht alle mathematischen Zusammenhänge dargestellt werden, wie zum Beispiel: $f(x_1, x_2) = x_1 \quad x_2$. Wird jedoch die Erweiterung der ausschließenden Assoziationen berücksichtigt, so kann jeder binäre mathematische Zusammenhang dargestellt werden.

2.5 Neuronale Netze

Neuronale Netze können beliebige funktionale Zusammenhänge modellieren und erlernen, falls genügend Neuronen vorhanden sind. Die Nachteile bei diesem Verfahren sind:

1. Der Lernalgorithmus kann suboptimale Ergebnisse liefern.
2. Das Modell des neuronalen Netzes lässt sich schwer untersuchen und analysieren.
3. Die richtige Anzahl und die Anordnung der Neuronen ist entscheidend für das Ergebnis. Zu wenig Neuronen können zur Folge haben, dass Zusammenhänge nicht erlernt werden. Bei zu vielen Neuronen besteht die Gefahr, dass das neuronale Netz die Werte nur „auswendig lernt“ und nicht genügend generalisiert.

Neuronale Netze werden in dieser Arbeit ausschließlich als prüfendes Verfahren eingesetzt. Durch die hohe Flexibilität können neuronale Netze Zusammenhänge repräsentieren, die mit anderen Modellen schwer zu modellieren sind. Hierbei ergeben sich durch eine Untersuchung mit neuronalen Netzen Anhaltspunkte, wie gut sich die Daten einordnen bzw. erklären lassen.

2.5.1 Historische Einordnung

Der Bereich „Neuronale Netze“ wurde schon seit der ersten Bearbeitung zu diesem Thema von McCulloch und Pitts [MP43] aus zwei Blickwinkeln betrachtet. Aus der mathematisch algorithmischen Sicht sind neuronale Netze eine Methode, um mathematische Funktionen zu repräsentieren, die aus Beispielen erlernt werden. Aus der biologischen Sicht sind neuronale Netze eine stark vereinfachte Repräsentation von Neuronen; den Zellen, welche einen Hauptanteil an der Informationsverarbeitung des Gehirns übernehmen. In den letzten Jahren ist der biologische Ansatz immer weiter in den Hintergrund getreten, da die stark vereinfachte Repräsentation mit den realistischen Abläufen im Gehirn zu wenig Gemeinsamkeiten aufweist.

2.5.2 Multilayer Feed-Forward Networks

1957 beschreibt Rosenblatt [Ros57] Netze, in denen Neuronen zu einer beliebigen Anzahl von Schichten miteinander verknüpft werden. Die erste Schicht (Eingabeschicht) empfängt Signale, wertet diese aus und gibt sie an die nächste Schicht weiter. Schicht für Schicht werden die Signale so bis zur letzten Ausgabeschicht weitergegeben. Allerdings konzentrierte sich die damalige Arbeit auf die Single-Layer Perceptrons, die nur aus einer Eingabe- und Ausgabeschicht bestehen, da gute Lernregeln für versteckte Ebenen schwierig zu finden sind. Erst 1969 wird von Bryson und Ho eine Lernmethode für Multilayer Feed-Forward Netzwerke gefunden, die unter dem Namen „Back-Propagation“ bekannt wird (siehe [BH69]).

Aufbau eines Neurons

Ein Neuron, wie in Abbildung 2.13 dargestellt, empfängt Signale von anderen Neuronen a_j, die durch die Verbindungen $w_{i,j}$ verstärkt oder abgeschwächt werden.

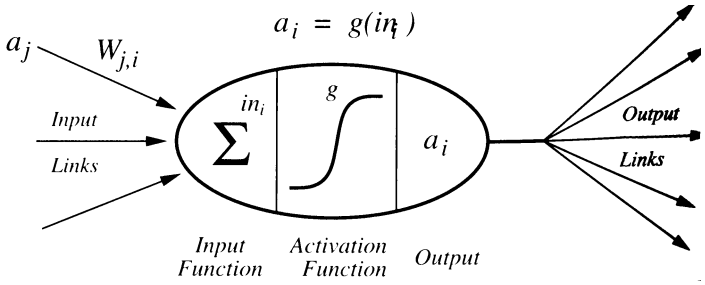

Abbildung 2.13: Aufbau eines Neurons (siehe [RN95], S. 568)

In der Eingangsfunktion werden diese berechnet, in der Aktivierungsfunktion ausgewertet und als Ausgabe an die nachfolgenden Neuronen weitergegeben.

Aufbau eines neuronalen Feed-Forward-Netzes

Die Abbildung 2.14 zeigt ein dreischichtiges Netz, wobei die Informationen von links nach rechts über die Verbindungen weitergegeben werden. Jedes Neuron verarbeitet die von links kommenden Eingaben, indem diese aufsummiert werden und über eine Schwellwertfunktion bestimmt wird, ob die Eingaben zur Aktivierung des Neurons ausreichen. Hierbei kann jeder Pfeil, sprich Verbindung, die Informationen verstärken oder abschwächen. Die formale Definition eines neuronalen Netzes findet sich im Anhang „Formale Definition eines neuronalen Netzes“ auf Seite 92.

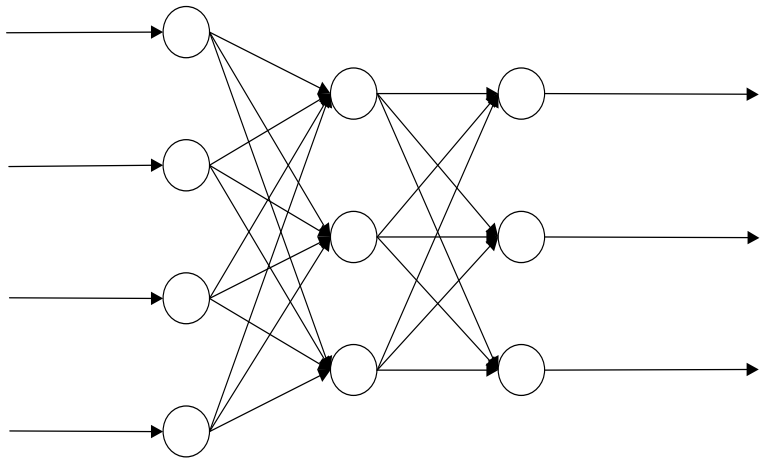

Abbildung 2.14: Beispiel des Aufbaus eines neuronalen Netzes

Ein vereinfachtes neuronales Netz als Beispiel für die XOR-Funktion.

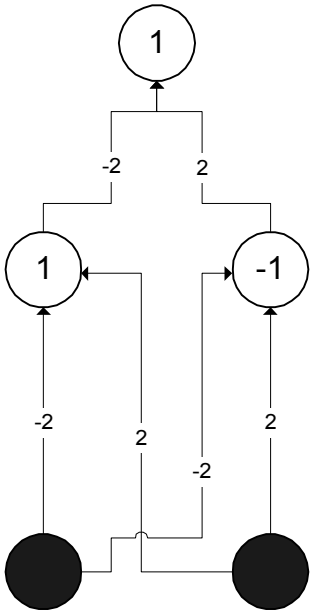

Abbildung 2.15: Beispiel eines neuronalen Netzes, welches die XOR-Funktion repräsentiert

Die Werte an den Pfeilen in der Abbildung 2.15 bezeichnen Gewichte, während diejenigen in den Kreisen Bias-Werte darstellen, die zu der jeweiligen Summe der Eingangswerte addiert werden. Ist die Summe der Eingangswerte größer Null, wird eine Eins weitergegeben sonst eine Null.

2.5.3 Backpropagation

Die Grundidee bei der Backpropagation liegt in der Minimierung der Fehlerfunktion.

$$E = \frac{1}{2} \sum_{n_i \in N_n} (t_i - o_i)^2$$

Hierbei bezeichnet t_i das aktuelle Testmuster und o_i die aktuelle Ausgabe des neuronalen Netzes.

Für den folgenden Lernalgorithmus gibt es zwei prinzipielle Erklärungsmethoden. Einerseits lässt sich zeigen, dass nach dem Prinzip der Ursache Fehler weitergegeben werden und je nachdem, welche Werte an dem Fehler beteiligt sind, diese geändert werden. Andererseits existiert eine mathematische Sichtweise, mit der nachgewiesen werden kann, dass sich der Algorithmus an der Fehlerfunktion nach unten „entlang hangelt“, bis ein Minimum erreicht wird.

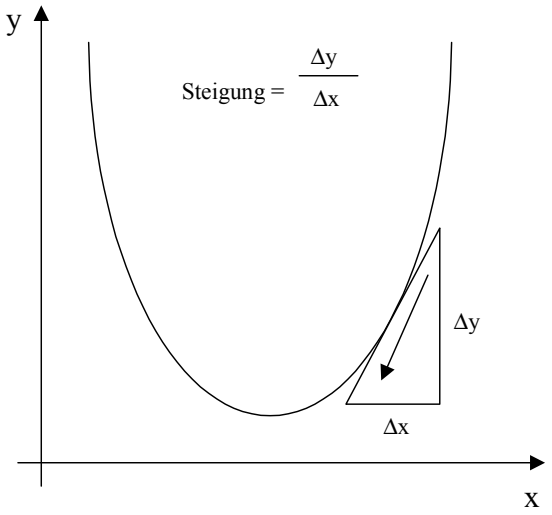

Abbildung 2.16: Minimierung des Fehlers durch „Gradient Descent“

Diese Vorgehensweise wird als „Gradient Descent“ bezeichnet (siehe Abbildung 2.16). Aus mathematischer Sicht wird klar, dass dies ein lokales Minimum sein kann und nicht zwingend ein globales ist.

Erklärung nach dem Prinzip der Ursache

Die folgenden Erklärungen zum Algorithmus stammen von [Gur97].

Es stellt sich die Frage: Inwieweit war das Gewicht beim aktuellen Durchgang verantwortlich für den Fehler?

Gewichtsänderung für Gewichte zwischen der vorletzten und letzten Schicht. Liegt das Gewicht zwischen einem Knoten der vorletzten Schicht und einem Knoten der Ausgabeschicht wird es um folgenden Wert korrigiert:

$$\Delta w_{ji} = \eta \delta_i o_j$$
$$\delta_i = (t_i - o_i)\Upsilon_i'(NET_i) \text{ falls } n_i \text{ ein Ausgabeneuron ist}$$

Die Gewichtsänderung hängt von den Faktoren η, $(t_i - o_i)$, $\Upsilon_i'(NET_i)$ und o_j ab. Sie lässt sich nach dem Prinzip der Ursache folgendermaßen erklären:

- η ist ein Faktor für die Stärke der Gewichtsveränderung in einem Schritt. Je größer η ist, umso größer fällt auch die Gewichtsveränderung aus. Dieser Faktor wird zu Beginn des Lernvorgangs vom Benutzer festgelegt.
- $(t_i - o_i)$ ist die Angabe über die Größe des Fehlers. Ist $(t_i - o_i) = 0$, gibt es für dieses Gewicht keinen Fehler. Demzufolge ist die Gewichtsveränderung ebenfalls gleich Null.
- $\Upsilon_i'(NET_i)$ liefert eine Angabe über den Teil der Aktivierungsfunktion (siehe Abbildung 2.17), in dem man sich innerhalb des Knotens befindet.

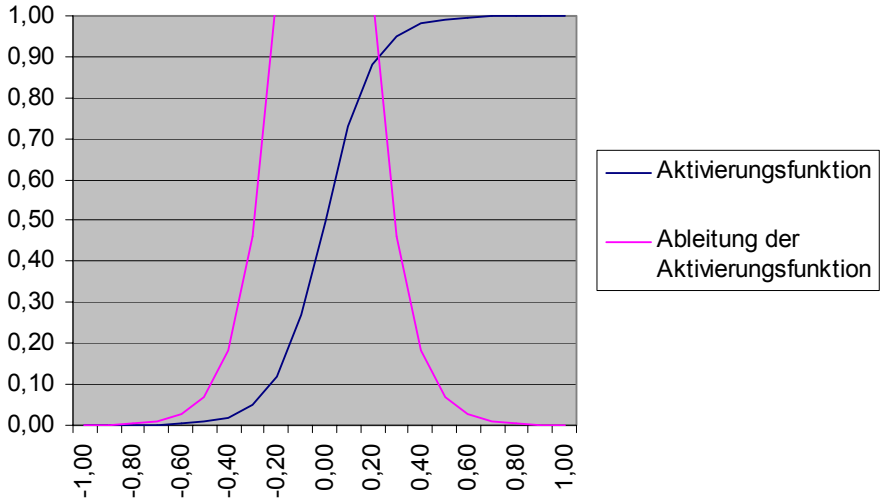

Abbildung 2.17: Beispiel einer Aktivierungsfunktion und ihrer Ableitung

Ist der Wert klein, befindet man sich am Rand und eine Änderung des Wertes o_j hat keinen großen Einfluss auf das Ergebnis. Ist der Wert hingegen groß, befindet man sich in der Mitte und kleine Änderungen von o_j haben große Auswirkungen. Dies spiegelt eine Rolle des Gewichtes wieder. Haben Änderungen in dem Gewicht große Auswirkungen, trägt es eine starke Verantwortung für den Fehler. Haben Änderun-

gen dagegen sehr geringe Auswirkungen, trägt das Gewicht kaum Verantwortung für den Fehler. Andererseits wird hierdurch der bisherige Lernerfolg mit einbezogen. Ein kleiner Wert und dadurch eine Position am Rand, lässt auf einen Lerneffekt bezüglich der zuvor gelernten Muster schließen. Um diesen Lerneffekt zu berücksichtigen, sollten die Änderungen gering ausfallen. Ein großer Wert und damit eine Position in der Mitte, lässt darauf schließen, dass noch nicht viele Muster erlernt worden sind bzw. das Neuron für die bisherigen Muster keine große Rolle spielt. Demgemäß sollte dieses stärker verändert werden.

- o_j ist der Ausgangswert des Knotens der vorletzten Schicht. Beträgt dieser Null, kann das Gewicht nicht zur Erklärung des Fehlers herangezogen werden, weil es nicht berücksichtigt worden ist. Daher ist eine Änderung nicht sinnvoll. Ist der Wert hingegen groß, steigt die Bedeutung des Gewichtes.

Gewichtsänderung für alle anderen Gewichte. Liegt das Gewicht zwischen zwei Knoten der versteckten Schichten, wird es um folgenden Wert korrigiert:

$$\Delta w_{ji} = \eta \delta_i o_j$$

$$\delta_i = \Upsilon_i'(NET_i) \sum_{n_j \in N_h} \delta_j \omega_{ij} \text{ mit } n_i \in N_k \text{ und } k < h, \text{ falls } n_i \text{ kein Ausgabeneuron ist.}$$

Die Gewichtsveränderung leitet sich zum großen Teil aus den oben begründeten Faktoren ab. Allein der Faktor $(t_i - o_i)$ als direkt messbarer Wert der Fehlergröße, wird für Neuronen der versteckten Schicht durch $\sum_{n_j \in N_h} \delta_j \omega_i$ ersetzt. Die zu beantwortende Frage lautet: Wie wirkt sich die Ausgabe des Neurons auf die Fehler aus, die in der Ausgabeschicht entstehen? Hierzu wird eine einzelne Verbindung des Neurons k mit seinem Nachfolgeneuron j betrachtet (siehe Abbildung 2.18).

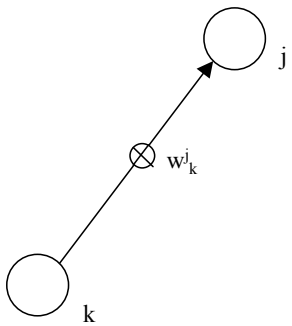

Abbildung 2.18: Verbindung zweier Neuronen

Um die obige Frage zu beantworten, sind zwei Faktoren zu berücksichtigen:

1. Wie stark kann das Neuron k das Neuron j beeinflussen? Ausgedrückt wird dies durch das Gewicht ω_{kj}.
2. Welche Auswirkungen hat das Neuron j auf den Gesamtfehler? Dies wurde bereits durch den Faktor δ_j berechnet.

Da allerdings das Neuron k mit vielen Neuronen verbunden ist, müssen für diese Verbindungen die Einflussgrößen addiert werden $\sum_{n_j \in N_h} \delta_j \omega_{ij}$.

Erklärung aus mathematischer Sicht

Das neuronale Netze benutzt zum Lernen das Verfahren „Gradient Descent“. Ziel beim „Gradient Descent“ ist die Minimierung der Fehlerfunktion, indem der Algorithmus sich an der Fehlerkurve entlang hangelt. Hierbei wird für jedes Neuron die Steigung der Fehlerfunktion in Abhängigkeit des zu korrigierenden Gewichtes betrachtet. Ist die Steigung positiv, muss das Gewicht verringert, ansonsten erhöht werden. Das Ausmaß der Veränderung wird durch den Faktor η bestimmt. Das bedeutet:

$$\Delta w_{ji} = -\eta \frac{\partial E}{\partial w_{ji}}$$

Fehlerfunktion:

$$E = \frac{1}{2} \sum_{n_i \in N_n} (t_i - o_i)^2$$

$$o_i = \Upsilon_i(NET_i)$$

$$NET_i = \sum_{n_j} \omega_{ij} o_j + \omega_i$$

Aufbauend ergibt sich für die Änderungen der Gewichte Δw_{ji} folgendes:

$$\Delta w_{ji} = \eta \delta_i o_j$$

$$\delta_i = (t_i - o_i)\Upsilon_i'(NET_i) \text{ falls } n_i \text{ ein Ausgabeneuron ist.}$$

$$\delta_i = \Upsilon_i'(NET_i) \sum_{n_j \in N_h} \delta_j \omega_{ij} \text{ mit } n_i \in N_k \text{ und } k < h, \text{ falls } n_i \text{ kein Ausgabeneuron ist.}$$

Der Beweis ist im Anhang „Formel 6“ auf Seite 93 aufgeführt. Für weitere Informationen siehe [RN95].

Bilder einer Lernkurve

Die Laufzeit von neuronalen Netzen, wie in Abbildung 2.19 dargestellt, hängt von der beabsichtigten Güte des Ergebnisses ab. Im Prinzip kann das Netz unendlich trainiert werden bzw. lässt sich zu jedem Zeitpunkt beenden. Normalerweise wird eine zu unterschreitende Fehlergrenze angegeben, damit das Training gestoppt wird. Es kann jedoch der Fall eintreten, dass dieser Wert in einem Training nicht unterschritten wird.

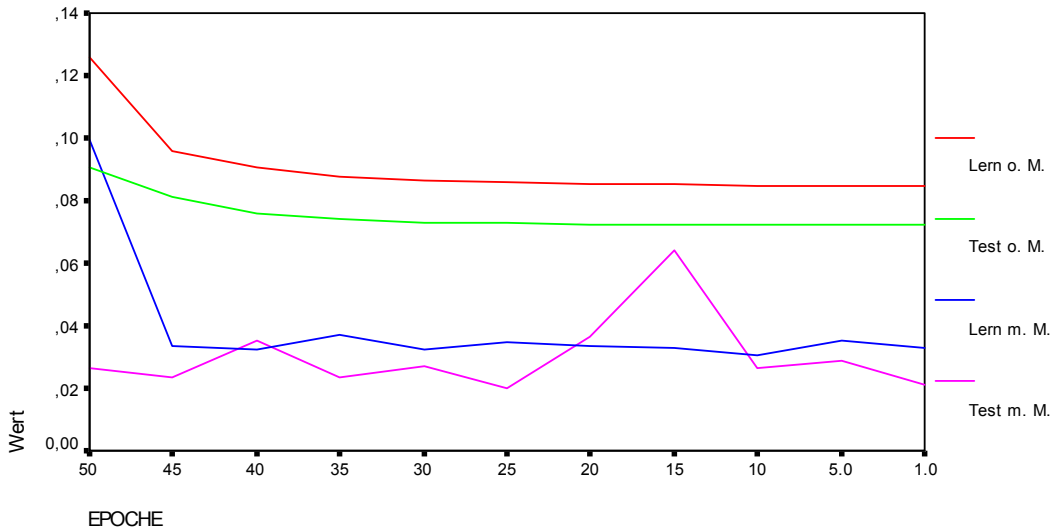

Abbildung 2.19: Beispielkurve aus dem in dieser Untersuchung benutzten neuronalen Netz

2.6 Clusteranalyse

Der Methodenbereich Clusteranalyse wird in dieser Untersuchung eingesetzt, um die Struktur der Daten besser erfassen zu können. Die Clusteranalyse ermittelt Mengen von Umsatztagen, so dass die Elemente verschiedener Mengen möglichst unterschiedlich sind und Elemente innerhalb einer Menge möglichst ähnlich. Für diese Untersuchung sind hierbei zwei Aspekte interessant:

- Haben die Elemente einzelner Cluster Merkmale, die bezeichnend für dieses Cluster sind?
- Erkennt das Verfahren Cluster in den Daten? Aufgrund der binären Merkmale sollten Cluster in den Daten erkennbar sein, wenn diese zum größten Teil durch diese Merkmale erklärt werden können.

Im folgenden Abschnitt werden die Anwendungsgebiete, Ziele und beispielhafte Algorithmen des Clustering angegeben. Dies soll einen kurzen Einblick in das Gebiet des Clustering gewähren. Die Informationen sind aus der Internetresource [HK01] entnommen.

2.6.1 Anwendungsgebiete des Clustering

Zwei Beispiele von Applikationen in denen die Clusteranalyse verwendet wird:

Marketing

Clustering wird eingesetzt, um unterschiedliche Käufergruppen zu unterscheiden und auf diese Gruppen zugeschnittene Marktstrategien zu entwickeln.

Versicherungen

Ermittlung der Gruppen von unterschiedlichen Versicherungsnehmern mit unterschiedlichen Risikofaktoren.

2.6.2 Ziele des Clustering

Die Anwendung einer Clusteranalyse soll Bereiche innerhalb der Daten ermitteln, für die folgendes gilt:

- Die Daten innerhalb der Gruppen sollten eine hohe Ähnlichkeit aufweisen.
- Die Daten unterschiedlicher Gruppen sollten eine geringe Ähnlichkeit aufweisen.
- Die Qualität der Daten ist sowohl abhängig von der Ähnlichkeitsbewertungsfunktion, als auch von der Implementierung der Algorithmen.

2.6.3 Beispielhafte Algorithmen

K-Means (Heuristischer Partitionierungsalgorithmus)

Hierbei wird eine bereits existierende Clusteraufteilung verbessert.

1. Zu jedem Cluster wird ein Mittelpunkt berechnet.
2. Für jeden Datensatz wird berechnet, zu welchem Mittelpunkt er am nächsten liegt.
3. Für jeden Mittelpunkt wird ein neues Cluster aus den ihm zugewiesenen Datensätzen berechnet.
4. Schritt 1-3 werden solange wiederholt, bis keine Veränderungen mehr auftreten.

Die graphische Vorgehensweise des Algorithmuses wird in Abbildung 2.20 dargestellt.

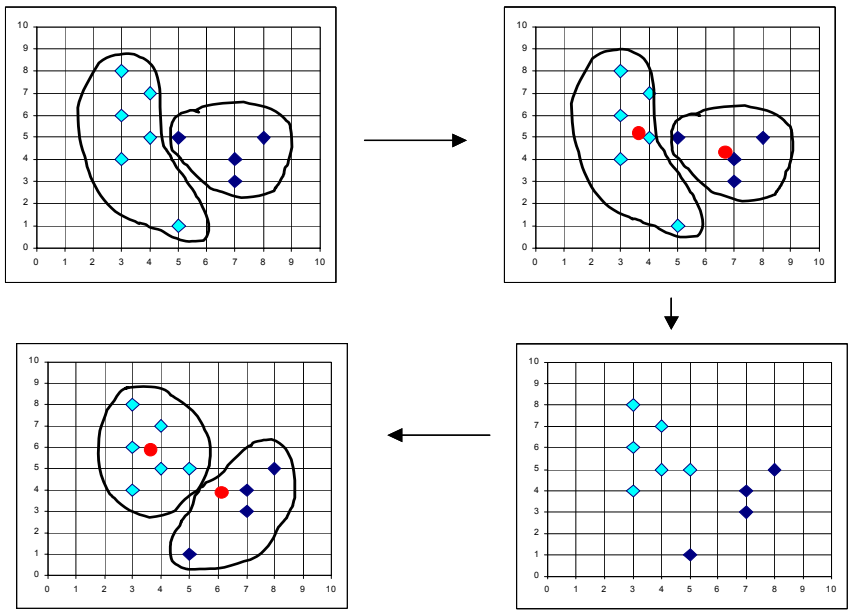

Abbildung 2.20: Beispiel des Algorithmus „K-Means“ (siehe [HK01])

Nearest Neighbour

Zu Beginn des Algorithmus ist jeder Datensatz ein Cluster für sich. Zu verbinden sind immer die beiden Cluster, die die größte Ähnlichkeit besitzen, bis die gewünschte Anzahl von Clustern erreicht wird (siehe Abbildung 2.21).

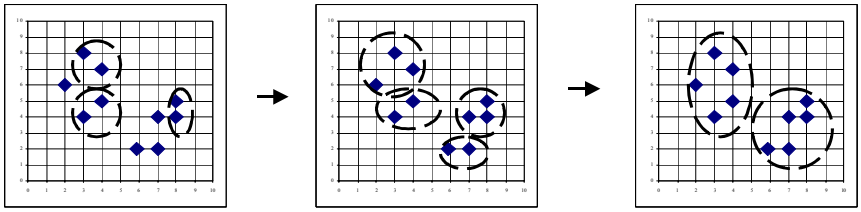

Abbildung 2.21: Beispiel des Algorithmus „Nearest Neighbour“ (siehe [HK01])

Divisives hierarchisches Clustering

Das divisive hierarchische Clustering stellt sich als umgekehrter Ansatz zum Nearest Neighbour dar. Zu Beginn existiert zunächst ein Cluster aus allen Datensätzen. Daraus werden in umgekehrter Reihenfolge zu Nearest Neighbour die Cluster weiter aufgespalten, bis die gewünschte Anzahl Cluster erreicht ist (siehe Abbildung 2.22).

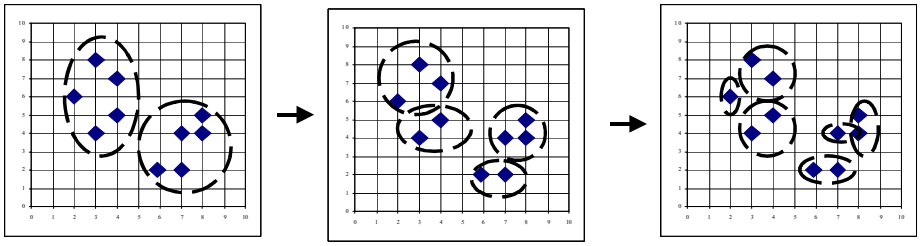

Abbildung 2.22: Beispiel des divisiven hierachischen Clustern (siehe [HK01])

2.7 Statistische Tests

Statistische Tests werden in dieser Untersuchung innerhalb der Regression eingesetzt, um die Güte des Regressionsmodells und die Aufbereitung der Daten zu testen. Folgend wird das Grundprinzip des statistischen Tests erklärt und zwei Beispiele von Tests angegeben. Für weitere Informationen über statistische Tests siehe ([HEK82], S. 133).

2.7.1 Überblick

Statistische Tests erlauben nach einer Stichprobe $x_1, ..., x_n$, Annahmen über dieselbe zu prüfen. Unterschieden wird zwischen der Nullhypothese H_0 „Die Annahme ist richtig“ und der Alternativhypothese H_1 „Die Annahme ist falsch“. Aufgrund der Beobachtung bezüglich der Stichprobe muss zwischen H_0 oder H_1 entschieden werden. Hierbei können zwei Arten von Fehlern unterlaufen. Fehler erster Art: Die Entscheidung wird zu Gunsten von H_1 getroffen, obwohl H_0 vorliegt. Fehler zweiter Art: Die Entscheidung wird zu Gunsten von H_0 getroffen, obwohl H_1 vorliegt.

Statistische Tests lassen sich in der Weise parametrisieren, dass Fehler erster Art ausschließlich mit einer festgelegten Wahrscheinlichkeit α vorkommen können. Die Wahrscheinlichkeit, dass ein Fehler zweiter Art vorkommt, wird nicht festgelegt. Die Entscheidungsmethode sollte in der Form gewählt werden, dass die Wahrscheinlichkeit für einen Fehler zweiter Art möglichst gering ausfällt. Die Tatsache, dass ein Fehler erster Art mit der Wahrscheinlichkeit α gemacht wird, bedeutet, wenn eine Entscheidung für die ursprüngliche Annahme getroffen wird, demnach für H_1, ist diese zu der vorher festgelegten Wahrscheinlichkeit α falsch und damit mit einer festgelegten Wahrscheinlichkeit von $1 - \alpha$ richtig. Ein derartiger Test wird als „Test zum Niveau α“ bezeichnet.

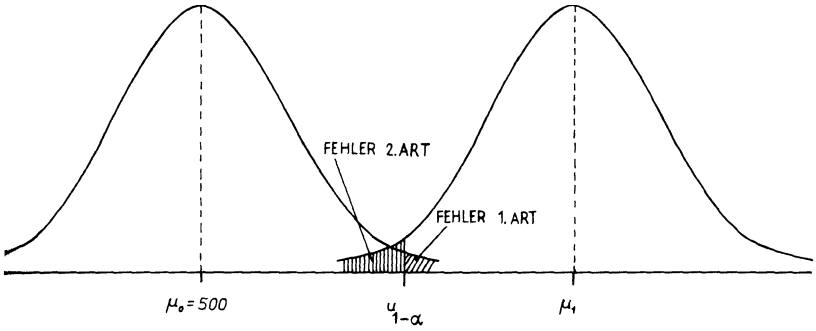

Abbildung 2.23: Einseitiger Einstichprobengaußtest (siehe [HEK82], S. 135)

Der Abbildung 2.23 ist zu entnehmen, dass zwei Verteilungskurven angegeben sind. Die linke Kurve zeigt eine Verteilung mit einem Mittelwert von 500. Wird nun ein Wert gemessen, der rechts von dem angegebenen Wert $u_{1-\alpha}$ liegt, muss entschieden werden, dass der Mittelwert nicht 500 beträgt. Diese Festlegung ist genau dann falsch, wenn der gemessene Wert innerhalb des schraffierten Bereichs „Fehler erster Art“ liegt (vgl. Abbildung 2.23). Die Wahrscheinlichkeit, dass ein gemessener Wert in diesem Bereich liegt, ist gleich α. Demzufolge kann ein Fehler erster Art ausschließlich mit der festgelegten Wahrscheinlichkeit α eintreten. Ein Fehler zweiter Art wird gemacht, wenn der tatsächliche Mittelwert nicht 500 beträgt, sondern z. B. $_1$ und entschieden wird, dass der Mittelwert 500 beträgt, weil der gemessene Wert links von $u_{1-\alpha}$ in dem schraffierten Bereich „Fehler zweiter Art“ (vgl. Abbildung 2.23) liegt. Da aber der tatsächliche Mittelwert $_1$ nicht bekannt ist, kann keine Aussage darüber getroffen werden, wie wahrscheinlich ein „Fehler zweiter Art“ ist.

2.7.2 Test auf Normalverteilung bei unbekanntem Mittelwert und Varianz

In den folgenden Tests wird aufgrund einer Beobachtungsreihe geschätzt, inwiefern die Werte einer Normalverteilung mit unbekanntem Mittelwert und unbekannter Varianz entsprechen. Die Nullhypothese lautet, dass die Beobachtungsreihe normalverteilt ist. Daher lassen diese Tests die Aussage zu: „Die Beobachtungsreihe ist mit einer Wahrscheinlichkeit von α nicht normalverteilt !“ Es ist jedoch keine Aussage darüber vorhanden, wie wahrscheinlich eine Normalverteilung ist. Eine Prüfgröße wird berechnet, die mit einer festen Verteilungskurve verglichen wird. Ist die Prüfgröße zu groß, dann ist die Normalverteilungshypothese entweder falsch oder der „Fehler erster Art“ liegt vor.

Es liegen n unabhängige Beobachtungen $x_1,\ldots,x_n$ vor.

Die Hypothesen sind:

H_0 : Die Grundgesamtheit ist $N(\ _0, \delta_0^2)$ verteilt.

H_1 : Die Grundgesamtheit ist nicht $N(\ _0, \delta_0^2)$ verteilt.

Die Parameter $\ _0$ und δ_0^2 werden aus der Stichprobe geschätzt.

Der χ^2-Anpassungstest:

(siehe [HEK82], S. 182).

1. Das Intervall $(-\infty, \infty)$ wird in k Klassen eingeteilt.
2. Für jede Klasse wird die Anzahl O_i der Messwerte ermittelt, die in dieser Klasse liegen.
3. Aufgrund der Nullhypothese H_0 wird berechnet, mit welcher Wahrscheinlichkeit p_i ein Element in dieser Klasse liegt; setze $E_i = np_i$.
4. Die Prüfgröße $T = \sum_{i=1}^{\kappa} \frac{1}{E_i}(O_i - E_i)^2$ wird berechnet.

Ist nun $T \leq \chi^2_{k-1,\,1-\alpha}$, kann die Nullhypothese H_0 nicht verworfen werden. Andernfalls muss sie zum Signifikanzniveau α verworfen werden.

3 Auswertung durch Regression und Assoziationsregeln

3.1 Vorbereitung der Daten

Im folgenden Kapitel wird die Vorbereitung der Umsatzdaten für die Analyse dargestellt. Als Ausgangspunkt werden die Umsätze eines Lebensmittelladens innerhalb des Zeitraums 1.1.1997 bis 31.12.1999 untersucht. Dieses Kapitel entspricht dem Schritt 'Vorbereitung' in Abbildung 2.2 auf Seite 10.

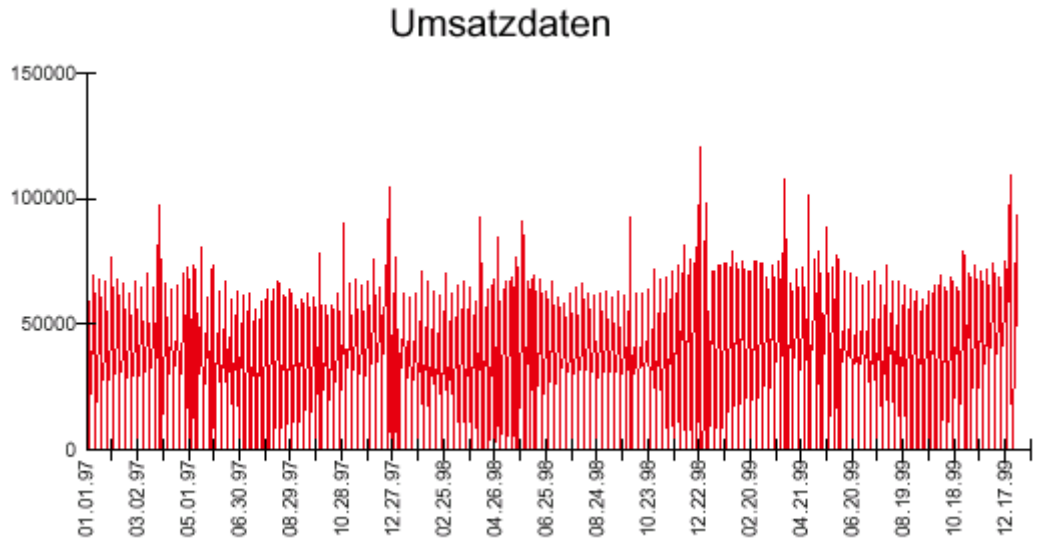

Abbildung 3.1: Rohdaten dieser Untersuchung

Das Ziel der Vorbereitung besteht darin, die Daten, dargestellt in Abbildung 3.1, von ungewünschten Einflussfaktoren zu bereinigen. Hierbei wird von der Hypothese ausgegangen, dass der Umsatz von folgenden Einflussfaktoren abhängt:

- Wochentag
- Nähe zu Feiertagen: Hierbei werden die fünf Einkaufstage vor und zwei Einkaufstage nach einem Tag berücksichtigt.

- Weihnachtszeitraum
- Trend: Hierunter zählt jeder Einfluss, der sich langfristig auf die Umsätze auswirken kann.

Der erste Teil der Untersuchung beschränkt sich auf die Merkmale **Wochentag** und die **Nähe zum Feiertag**.

3.1.1 Eliminierung nicht zu untersuchender Merkmale

Zunächst werden die unerwünschten hypothetischen Einflussfaktoren bezüglich einer augenscheinlichen Wirkung auf die Daten untersucht. Ist ein unerwünschter Einfluss beobachtbar, wird dieser aus den Daten möglichst herausgerechnet.

Weihnachtszeitraum

In diesem Zeitraum konnte eine merkliche Umsatzsteigerung festgestellt werden, so dass diese Daten in der Untersuchung nicht berücksichtigt werden. Hierzu erfolgt später noch eine statistische Untersuchung (siehe Kapitel 3.2.2 auf Seite 45).

Trend

Zu Beginn des Jahres 1999 lässt sich eine deutliche Umsatzsteigerung des untersuchten Geschäftes beobachten. Ein linearer Trend konnte hingegen innerhalb der Daten nicht festgestellt werden. Um den Sprung innerhalb der Umsätze zu eliminieren, ist für beide Zeiträume 1997-1998 und 1999 der durchschnittliche Umsatz ermittelt worden. In den folgenden Untersuchungen wird nun nicht mehr der Tagesumsatz als Merkmal betrachtet, sondern der relative Anteil des Tagesumsatzes zum durchschnittlichen Umsatz des zugehörigen Zeitraumes.

Sonn- und Feiertage

Da an diesen Tagen kein Umsatz erfolgt, sind diese Daten entfernt worden.

Die resultierende Datenmenge ist in Abbildung 3.2 graphisch dargestellt.

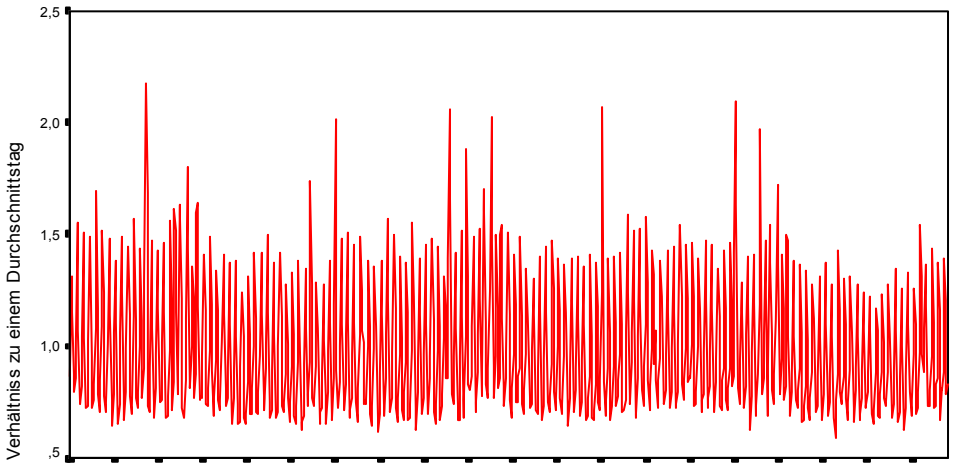

Abbildung 3.2: Aufbereitete Daten dieser Untersuchung

3.2 Exploration der Daten

Die Exploration der Daten erlaubt es, vermutete Zusammenhänge zu überprüfen und bestimmte Eigenschaften der Daten festzustellen, die wichtig sind für die späteren Untersuchungen. Dieses Kapitel entspricht dem Schritt ’Exploration’ in Abbildung 2.2 auf Seite 10.

3.2.1 Normalverteilungshypothese

Ist die oben angenommene Hypothese der Einflussfaktoren richtig, müssen die bereinigten Daten allein von dem Wochentag, der Nähe zum Feiertag und von einer zufälligen Komponente abhängen. Untersucht man demnach die Daten von Wochentagen, die nicht in der Nähe eines Feiertages liegen, müssen die Umsätze einer Normalverteilung folgen. Die Ergebnisse der Normalverteilungsuntersuchung sind im Anhang „Normalverteilungstests“ auf Seite 95 zu finden. Die Hypothese der Normalverteilung muss ausschließlich für Donnerstage verworfen werden, so dass dies insgesamt die Ausgangshypothese unterstützt. Außerdem ergibt die Voruntersuchung folgende arithmetische Mittel für die einzelnen Wochentage.

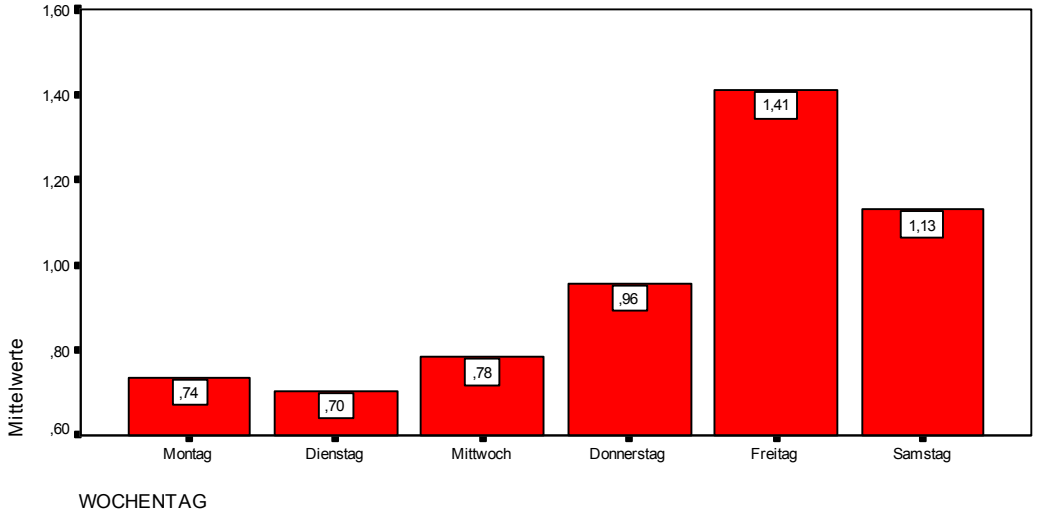

Abbildung 3.3: Mittelwerte der Wochentage, die sich nicht in der Nähe eines Feiertages befinden

3.2.2 Untersuchung des Einflusses des Dezemberzeitraums

In der folgenden Untersuchung wird der Dezember ausgeschlossen, da die Umsätze statistisch höher liegen, als im sonstigem Zeitraum. Dieser Einfluss soll hierbei durch eine statistische Untersuchung belegt werden. Als Methode kommt die arithmetische Mittelwertsbildung zur Anwendung und es zeigt sich, dass der durchschnittliche Wert der Wochentage ohne Feiertagseinfluss in den Monaten Januar bis November deutlich unter den Umsätzen des Dezembers liegt.

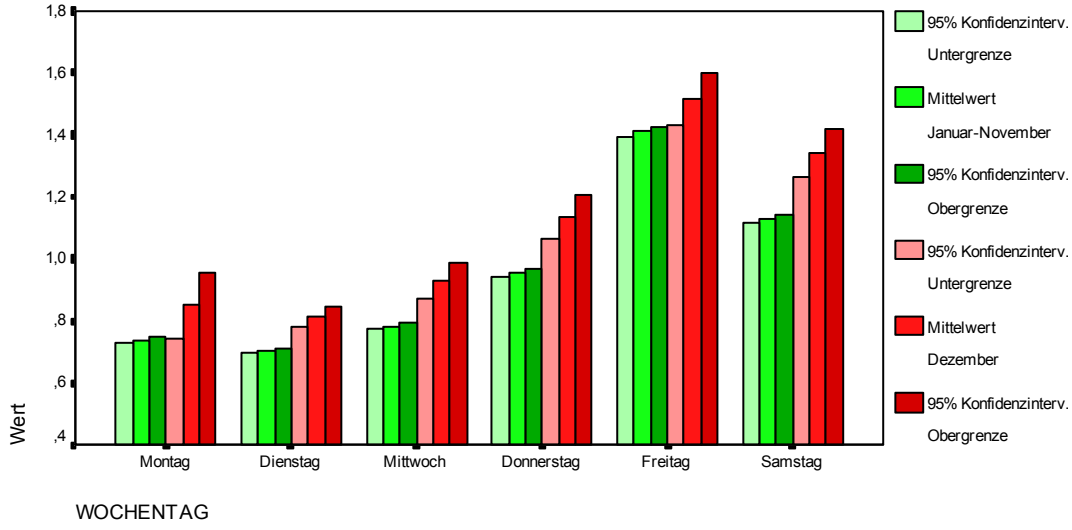

Abbildung 3.4: Statistische Auswertung des Dezembereinflusses

In Abbildung 3.4 ist ein deutlicher Unterschied der Umsatzdaten erkennbar. Bei der Untersuchung sind auch die 95% Konfidenzintervalle angegeben worden. Dabei wird deutlich, dass selbst diese sich bis auf den Montag nicht überschneiden. Daher lässt sich ein klarer Einfluss des Monats Dezember feststellen.

3.2.3 Klassifizierung der Umsatzdaten

Um den Einsatz von Klassifikationsalgorithmen zu ermöglichen, wie zum Beispiel das Verfahren der Assoziationsregeln, wird in diesem Kapitel eine Klassifikation der Umsatzdaten ermittelt. Hierbei sind mehrere Vorgehensweisen möglich:

- Ermittlung der Klassen durch graphische Veranschaulichung der Clusterbildung innerhalb der Daten.
- Ermittlung der Klassen durch Analyse des Problemumfeldes. Welche Klassen bzw. wie viele Klassen werden benötigt?
- Ausgehend von einer konstanten Klassengröße, ist zu untersuchen, bis zu welcher Klassenanzahl sinnvolle Ergebnisse von den Algorithmen geliefert werden.

Ermittlung der Klassen durch graphische Veranschaulichung

Als geeignete graphische Darstellung wird ein Histogramm gewählt. Das Histogramm besteht aus 100 Umsatzklassen, die über das Intervall von 50% bis zu 250% gleichverteilt sind. Jeder Tag wird, je nach seinem relativen Umsatz, einer dieser Klassen zugeordnet. Zu jeder Umsatzklasse wird gezählt, wie viele Tage dieser Klasse zugeordnet sind. Die Kardinalität jeder Klasse wird in einem Balkendiagramm dargestellt.

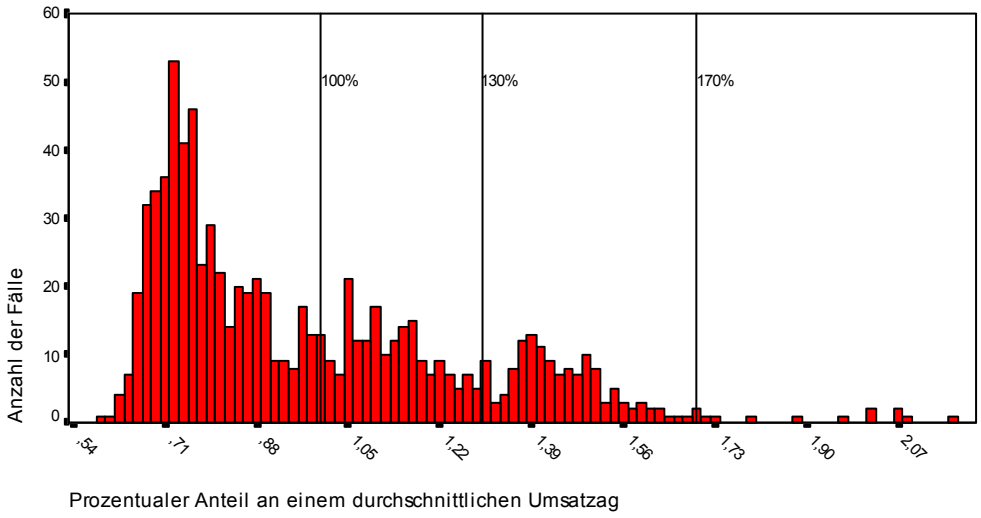

Abbildung 3.5: Häufigkeitsverteilung der Umsatzdaten

Anhand der Abbildung 3.5 lassen sich folgende Klassen identifizieren:

- Klasse 1: 0 bis 100 %
- Klasse 2: 100 % bis 130 %
- Klasse 3: 130 % bis 170 %
- Klasse 4: 170 % bis $+\infty$

3.3 Regression

Die folgenden Untersuchungen entsprechen dem Schritt Anwendung von ’Modellierungs- und Entdeckungstechniken’ in Abbildung 2.2 auf Seite 10.

In diesem Kapitel werden die Umsatzdaten mittels der Regression analysiert. Die folgende Formel dient als Regressionsmodell:

$$Um = a \cdot Mo + b \cdot Di + c \cdot Mi + d \cdot Do + e \cdot Fr + f \cdot Sa + g \cdot 5VF + h \cdot 4VF + i \cdot 3VF + j \cdot 2VF + k \cdot 1VF + l \cdot 1NF + m \cdot 2NF$$

Bedeutung der Variablen:

- Um = Umsatz
- $Mo \in \{0, 1\}$: beträgt 1, falls der Tag ein Montag ist, sonst 0
- Di, Mi, Do, Fr, Sa analog zum Montag
- $5VF \in \{0, 1\}$: beträgt 1, falls der Tag fünf Werktage von einem Feiertag entfernt ist, sonst 0
- $4VF$, $3VF$, $2VF$, $1VF$ analog zu $5VF$
- $1NF \in \{0, 1\}$: beträgt 1, falls der Tag einen Werktag nach einem Feiertagliegt, sonst 0
- $2NF$ analog zu $1NF$
- $a \ldots m$ sind reelle Zahlen

3.3.1 Präsentation der Ergebnisse der Regression über Umsatzdaten als Prognosesystem

Aufgrund dieser Beispielregression ergeben sich die folgenden Werte innerhalb der Spalte B der nachfolgenden Tabelle.

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	MONTAG	,733	,007	,289	106,915	,000
	DIENSTAG	,699	,007	,281	103,095	,000
	MITTWOCH	,792	,007	,318	116,957	,000
	DONNERST	,964	,007	,379	141,344	,000
	FREITAG	1,409	,007	,559	206,820	,000
	SAMSTAG	1,114	,007	,447	163,432	,000
	VF5	3,145E-02	,017	,005	1,830	,068
	VF4	5,115E-02	,017	,008	2,958	,003
	VF3	,107	,017	,017	6,193	,000
	VF2	,301	,017	,048	17,412	,000
	VF1	,644	,017	,102	37,187	,000
	NF1	7,005E-02	,016	,012	4,315	,000
	NF2	3,796E-03	,016	,001	,235	,814

Abbildung 3.6: Das Regressionsmodel – Daten des ersten Betriebes

Die Werte der Wochentage stimmen mit den, für die einzelnen Tage ermittelten, Durchschnittswerten überein (siehe Abbildung 3.7 auf Seite 48). Für die Variablen VF5 und NF2 ist keine Signifikanz abzulesen, daher finden sie in der folgenden Untersuchung keine weitere Verwendung. Zur Veranschaulichung eine graphische Darstellung der Werte des Modells.

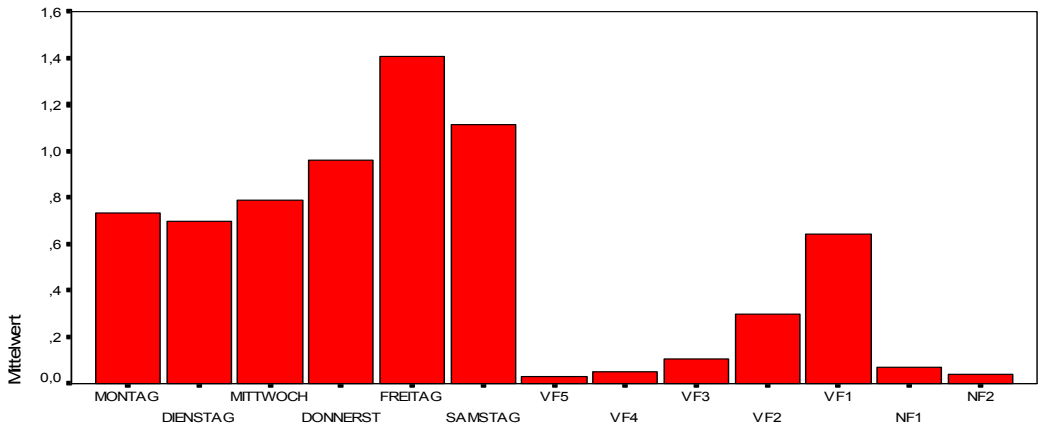

Abbildung 3.7: Graphische Darstellung des Regressionsmodells – Daten des ersten Betriebes

Bestimmtheitsmaß

Das Bestimmtheitsmaß der Regression wird durch den Wert R-Quadrat ausgedrückt. Je näher dieser Wert an der 1 liegt, desto geringer ist die Wahrscheinlichkeit, dass andere Einflüsse auf den Umsatz einwirken, außer den zu untersuchenden Werten.

Modell	R	R-Quadrat[a]	Korrigiertes R-Quadrat	Standardf ehler des Schätzers
1	,997[b]	,994	,994	7,832E-02

Abbildung 3.8: Bestimmheitsmaß des Regressionsmodells – Daten des ersten Betriebes

Der erreichte Wert von 0,994 zeigt, dass die Regression über gute Prognosefähigkeiten verfügt.

T-Test

Die Ergebnisse des T-Testes werden in der Modellbeschreibung in der Spalte T in Abbildung 3.6 angegeben. Die Untersuchung zeigt für die Variablen Mo-Sa, VF4, VF3, VF2, VF1, NF1 eine mehr als 99%-tige Sicherheit, dass diese einen Einfluss auf den Umsatz ausüben.

F-Test

Der F-Test ergibt ebenfalls eine mehr als 99%-tige Sicherheit, dass die gemessenen Werte nicht zufällig entstanden sind.

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	870,305	13	66,947	10914,651	,000[a]
	Residuen	5,030	820	6,134E-03		
	Gesamt	875,334[b]	833			

Abbildung 3.9: F-Test des Regressionsmodells – Daten des ersten Betriebes

Residualanalyse

Die Residualanalyse in Abbildung 3.10 zeigt einen annähernd optimalen Verlauf.

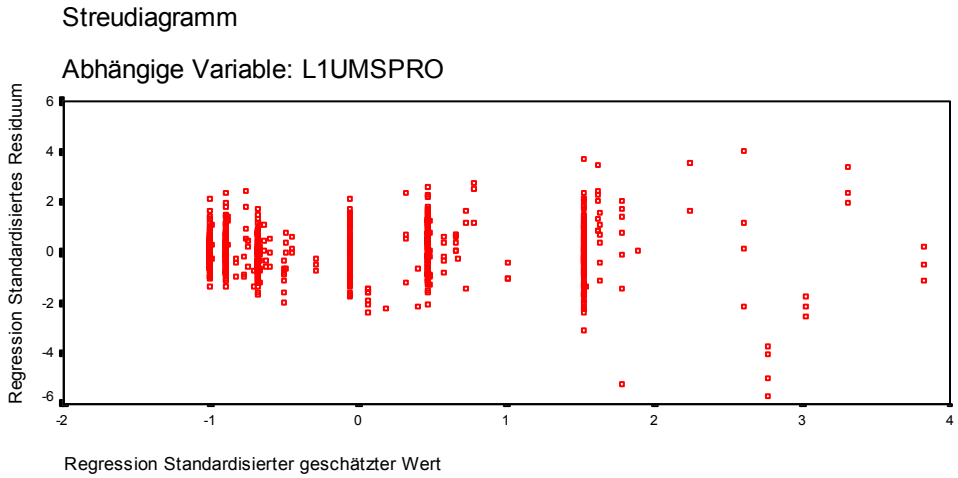

Abbildung 3.10: Residualanalyse des Regressionsmodells – Daten des ersten Betriebes

Test der Stabilität der Regression

Der Test läuft in drei Schritten ab:

1. Die Daten werden mittels einer Eigenimplementation aufbereitet und nach dem Zufallsprinzip in eine 2/3 große Lernmenge und eine 1/3 große Testmenge aufgeteilt und in verschiedenen Dateien für die weiteren Untersuchungen gespeichert.
2. Mit Hilfe eines externen Lisp-Programms wird anhand der Lernmenge die Regression durchgeführt und das Ergebnis gespeichert.
3. Das Regressionsmodell wird mittels einer Eigenimplementation eingelesen und sowohl die Lernmenge als auch die Testmenge mit den Prognosewerten des Regressionsmodells verglichen. Die Auswertung wird in einer Datei protokolliert.

Dieser Ablauf wird 100 mal hintereinander ausgeführt.

Folgend sind nun die Veränderungen der Modellvariablen des Regressionsmodells dargestellt. Zuerst werden die Veränderungen der Wochentagsvariablen und danach die Veränderungen der Feiertagsvariablen in den nachfolgenden Abbildungen dargestellt.

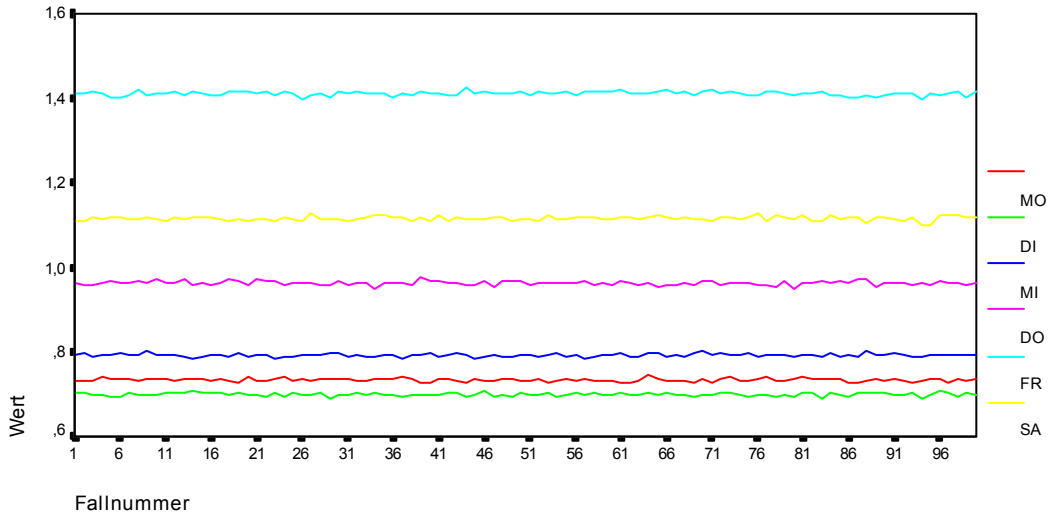

Abbildung 3.11: Stabilität der Wochentagsvariablen

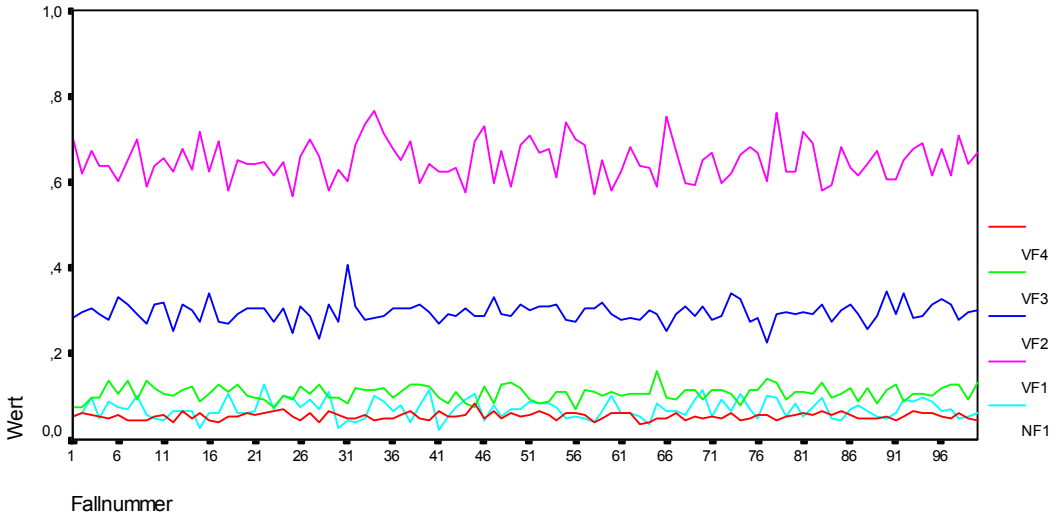

Abbildung 3.12: Stabilität der Feiertagsvariablen

Erkennbar ist, dass die Variablen der Feiertage einer größeren Varianz unterliegen, als die der Wochentage. Dies zeigt zum einen, dass die Modellwerte der Wochentagsvariablen bereits mit einem kleinen Teil der Lerndaten stabil bestimmt werden können. Zum anderen sind die Werte der Feiertage abhängiger von der Auswahl der Lerndaten. Dies liegt ebenfalls an der geringeren Zahl der Feiertage gegenüber den Werktagen.

Zusätzlich werden die durchschnittlichen quadratischen Fehler (Mean Square Error) der Lernmenge, der Testmenge und der Gesamtmenge angegeben.

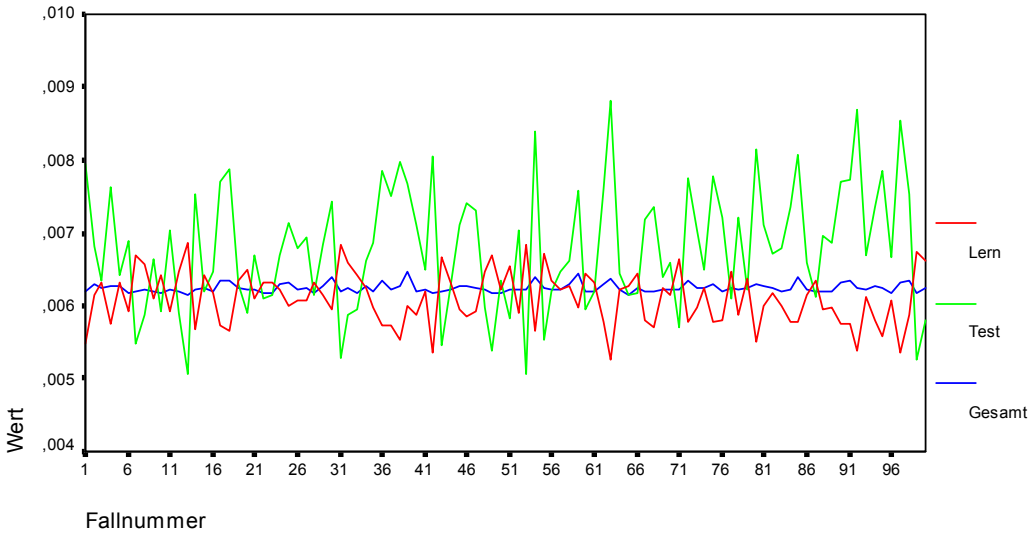

Abbildung 3.13: Prognosefehler des Regressionsmodells – Daten des ersten Betriebes

Es ergibt sich ein MSE[3] zwischen 0,006 und 0,007. Interessant ist an der Abbildung 3.13, dass das gelernte Modell sich stellenweise besser für die Testdaten als für die Lerndaten eignet. Dieses lässt vermuten, dass es Daten gibt, die nicht durch das Regressionsmodell vorhersagbar sind. Je nachdem, ob diese Daten hauptsächlich in den Test- bzw. Lerndaten vorkommen, ist der Fehler der Test- bzw. Lernmenge größer. Außerdem zeigt sich, dass die gelernten Modelle immer die gleiche Fehlergröße bei den Gesamtdaten aufweisen. Demzufolge kann mit einem Teil der Daten ein repräsentatives Modell der Gesamtdaten erstellt werden.

Deskriptive statistische Analyse der Auswertungsdaten

	N	*Minimum*	*Maximum*	*Mittelwert*	*Standard-abweichung*
MO	100	0,73	0,74	0,7344	0,0038
DI	100	0,69	0,71	0,6999	0,0036
MI	100	0,78	0,80	0,7925	0,0038
DO	100	0,95	0,98	0,9631	0,0049
FR	100	1,40	1,43	1,4111	0,0053
SA	100	1,10	1,13	1,1160	0,0052
VF 4	100	0,04	0,08	0,0543	0,0082
VF 3	100	0,07	0,16	0,1079	0,0173

Tabelle 3.1: Deskriptive Auswertung der Variablen des Regressionsmodells

3. MSE: Mean Square Error, dt: Mittlerer quadratischer Fehler

	N	*Minimum*	*Maximum*	*Mittelwert*	*Standard-abweichung*
VF 2	100	0,23	0,41	0,2964	0,0243
VF 1	100	0,56	0,76	0,6499	0,0453
NF 1	100	0,02	0,13	0,0706	0,0219

Tabelle 3.1: Deskriptive Auswertung der Variablen des Regressionsmodells

Interessant ist, dass die Wochentage eine um den Faktor 5 kleinere Standardabweichung aufweisen. Dies unterstützt ebenfalls die in Abbildung 3.12 dargestellte größere Varianz.

3.3.2 Präsentation der Ergebnisse der Regression als Klassifikationssystem – Variante Eins

In diesem Kapitel wird das Ergebnis der Regression und der tatsächliche Umsatzwert klassifiziert. Dies ermöglicht den direkten Vergleich zwischen Assoziationsregeln und Regression, da hierbei die Ausgaben gleich sind. Die Verteilung aus Kapitel 3.2.3 auf Seite 46 wird als Klassifikationssystem benutzt. Der Algorithmus zur Regression wird ohne Änderungen ausgeführt und sowohl das Ergebnis als auch der Vergleichswert klassifiziert. Unterscheiden sich die beiden Klassen, wird der Fall als Fehler gezählt, ansonsten nicht. Diese Prozedur wird ebenfalls, wie bei der ersten Regression (siehe Kapitel 3.3.1 auf Seite 47), 100 mal ausgeführt. Für jeden Fall wird mittels einer Eigenimplementation analysiert, wie viel Prozent der Fälle fehlerhaft klassifiziert worden sind.

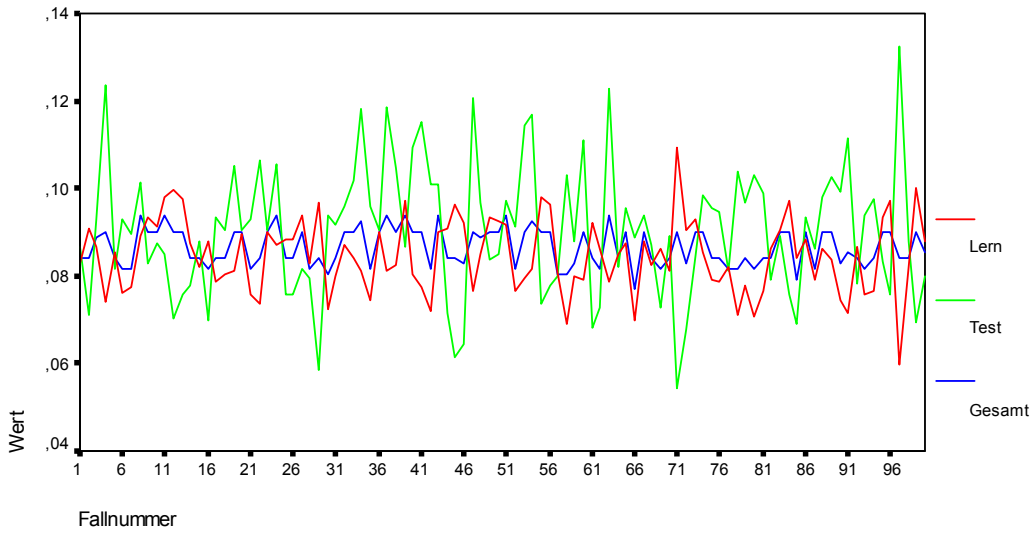

Abbildung 3.14: Klassifikationsfähigkeit des Regressionsmodells – Daten des ersten Betriebes

Die Abbildung 3.14 veranschaulicht, dass die Fehlerquote der Regression, bezogen auf diese Klassenverteilung, zwischen 8% und 9% liegt.

3.3.3 Präsentation der Ergebnisse der Regression als Klassifikationssystem – Variante Zwei

Eine weitere Variante der Analyse besteht darin, die Umsatzwerte, die in die Regression fließen, direkt in Klassen einzuteilen. Die Regression wird für jede Klasse durchgeführt. Es ergeben sich vier Regressionsmodelle, welche jeweils eine Klasse prognostizieren. Einem unbekannten Tag wird mittels der Regressionsmodelle die Klasse zugeordnet, die das maximalste Regressionsergebnis besitzt. Diese Vorgehensweise erlaubt es, die Ergebnisse der Regression und Assoziationsregeln direkt zu vergleichen, da in diesem Fall sowohl die Eingabewerte, als auch die Ausgabewerte vergleichbar sind. Wird hingegen das Ergebnis der Regression klassifiziert, wie in Kapitel 3.3.2 auf Seite 53, besitzt die Regression den Vorteil, die einzelnen genauen Umsatzwerte zu kennen und nicht, wie bei den Assoziationsregeln, nur die Klassen. Auf der anderen Seite bringt die Einteilung in Klassen vor der Analyse auch Vorteile, weil hierdurch ausschließlich die Informationen benötigt werden, die wichtig für die Klassifikation sind. Um die beiden Algorithmen vergleichen zu können, ist dieser Ansatz der Regression der geeignetere. Kein Algorithmus erhält mehr oder weniger Informationen als der andere.

Dieser Klassifizierungsalgorithmus wird wiederum mittels einer Eigenimplementation 100 mal hintereinander ausgeführt und die Klassifikationsfehler notiert.

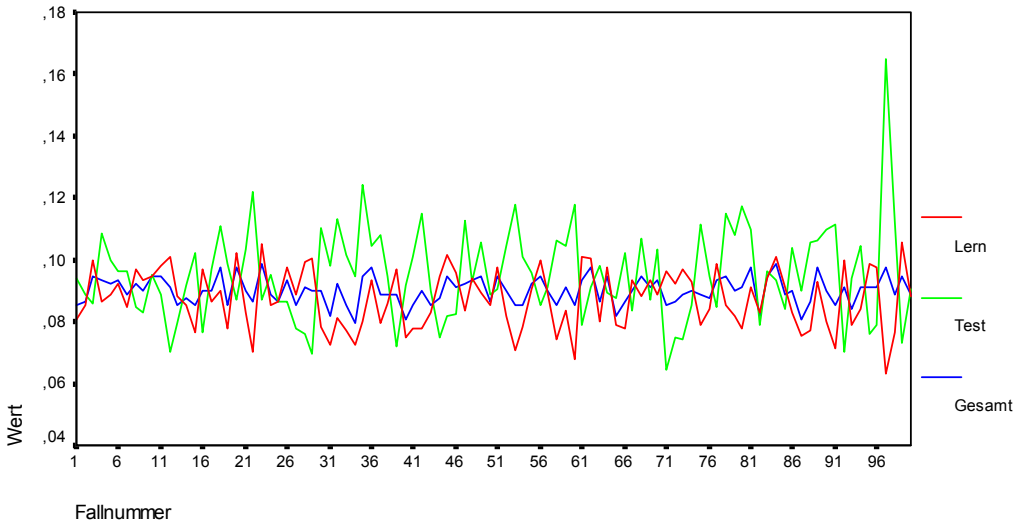

Abbildung 3.15: Klassifikationsfehler der Regression pro Klasse – Daten des ersten Betriebes

Diese Variante, dargestellt in Abbildung 3.15, erreicht bessere Klassifikationswerte als die Regression in Kapitel 3.3.2 auf Seite 53. Der Anteil der Fehler wurde von 9,5% auf 9% verbessert.

3.4 Assoziationsregeln

In diesem Kapitel wird die Umsatzanalyse mit Hilfe von Assoziationsregeln durchgeführt. Wie oben beschrieben, ist hierfür eine Klassifikation der Umsatzdaten erforderlich. In diesem Kapitel kommt die Klassifikation aus Kapitel 3.2.3 auf Seite 46 zur Anwendung. Zu Beginn werden einige problemspezifische Verbesserungen an dem Algorithmus vorgenommen. Der implementierte Algorithmus wird vorgestellt und eine beispielhafte Analyse durchgeführt. Anschließend wird analog zur Regression der Algorithmus 100 mal hintereinander ausgeführt und die Ergebnisse hieraus vorgestellt.

3.4.1 Modell- und Algorithmenbeschreibung

Die folgende Vereinfachung des Algorithmus beruht auf der Tatsache, dass nur bestimmte Regeln von Interesse sind. Uninteressant ist biespielsweise die Regel: „15 % aller Feiertage sind Montage“. Die für diese Untersuchung bedeutenden Regeln werden durch folgenden Ausdruck beschrieben:

$$(Mo|Di|Mi|Do|Fr|Sa)?,VF4?, VF3?, VF2?, VF1?, NF1?, NF2? = (U1|U2|U3|U4)$$

mit
$(x|y)$: x oder y
? : alternativ
Ui : Umsatzklasse i

Alle Regeln, bei denen maximal ein Element aus der Menge (Montag ... Samstag) in der Prämisse und genau ein Element aus der Menge (U1 ... U4) in der Konklusion vorhanden sind, werden ausgegeben.

Erweiterungen des Basisalgorithmus

- Erstelle ausschließlich Mengen, die aus der Gruppe Mo ... Sa maximal ein Element besitzen.
- Erstelle ausschließlich Mengen, die aus der Gruppe U1 ... U4 maximal ein Element besitzen.

- Erzeuge Regeln ausschließlich aus Mengen, welche aus der Gruppe U1 ... U4 genau ein Element besitzen.
- Erzeuge Regeln ausschließlich, indem das einzig vorhandene Umsatzelement gestrichen wird, so dass auf der rechten Seite der Regel als einziges ein Umsatzelement steht.

Der erweiterte Algorithmus

Der bestehende Algorithmus für Assoziationsregeln wird durch folgende Erweiterungen verbessert.

1. Es werden alle Kandidaten gelöscht, die mehr als einen Wochentag enthalten.
2. Es werden alle Kandidaten gelöscht, die nicht genau eine Umsatzklasse enthalten.
3. Die Regeln werden nur dahingehend erzeugt, dass die Umsatzklasse in der Konklusion liegt und alle restlichen Element in der Prämisse.

3.4.2 Beispieldurchlauf

Die Werte für den Algorithmus werden wie folgt festgelegt.

Es werden alle Regeln gesucht, die mindestens einen Support von 0,6% und mindestens eine Confidence von 12,5% aufweisen. Der Support bezeichnet die statistische Signifikanz, während die Confidence die Gültigkeit der Regel bestimmt.

Zu den wichtigen Regeln zählen jene, die einen Wochentag und eine Aussage über einen Feiertag enthalten. Demnach ist die Supportgrenze sehr niedrig angesetzt, um möglichst viele wichtige Regeln zu erhalten. Die angegebene Supportgrenze untersucht alle Vorkommen, die mindestens sechs mal gezählt werden. Trotz dieser niedrigen Supportgrenze ist die Regelmenge nicht vollständig, erweist sich daher als nicht optimal. Dieses Manko lässt sich durch eine größere Datenmenge eliminieren.

Die Abbildung 3.16 zeigt deutlich das Manko des Assoziationsregelmodells. Obwohl alle Wochentage ohne Feiertagsbezug berücksichtigt sind, fehlen die Spezialregeln, die sowohl einen Wochentag als auch einen Feiertag in der Prämisse beinhalten. Das Modell ist daher nicht vollständig. Die zugehörigen Assoziationsregeln werden im Anhang „Assoziationsregeln – Ein Betrieb“ auf Seite 99 aufgelistet.

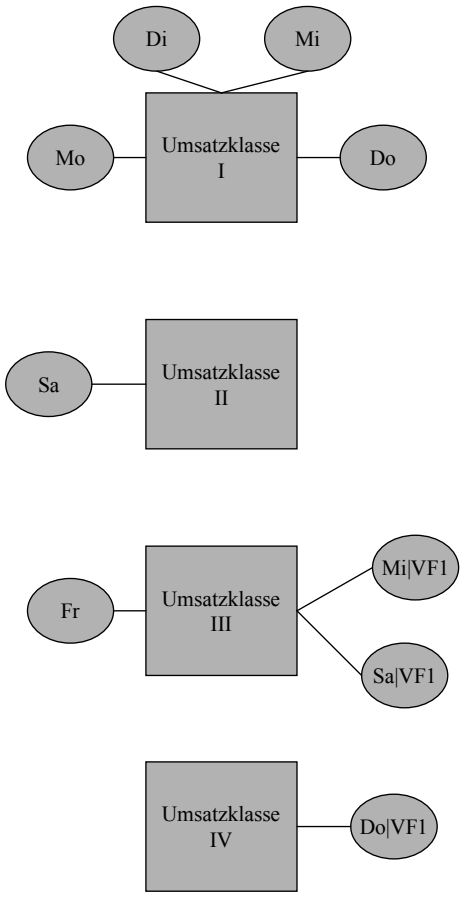

Abbildung 3.16: Graphische Veranschaulichung des Assoziationsmodells – Daten des ersten Betriebes

3.4.3 Anwendung des Ergebnisses

Bei dieser Untersuchung werden zwei Varianten der Anwendung von Regeln benutzt.

In der ersten Variante wird für jeden Tag jede Regel betrachtet, deren Prämisse mit dem Tag übereinstimmt. Wird diese Vorraussetzung von mehreren Regeln erfüllt, findet diejenige mit der größten Confidence Verwendung. Demnach wird die wahrscheinlichste Regel verwendet.

In der zweiten Variante wird hingegen die speziellste Regel verwendet. Dies ist diejenige, die die meisten Merkmale in der Prämisse besitzt.

Falls die angewandte Regel eine andere Umsatzklasse als die tatsächliche voraussagt, ist dies als Fehler zu betrachten.

3.4.4 Test der Stabilität der Assoziationsregeln als Klassifikationssystem

Der Algorithmus wird mittels einer Eigenimplementation 100 mal hintereinander ausgeführt. Der prozentuale Anteil der Lern- und Testfehler wird protokolliert.

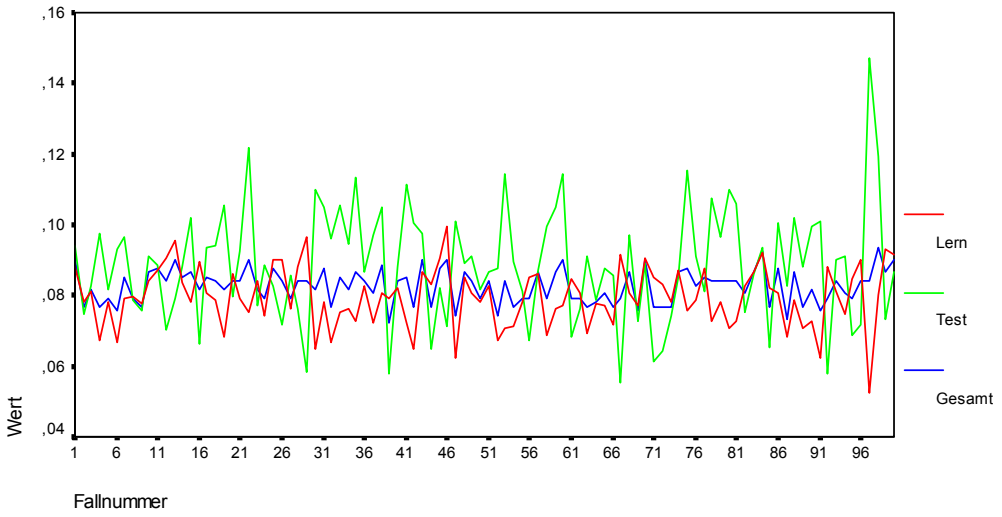

Abbildung 3.17: Klassifikationsfehler des Assoziationsmodells – Variante Eins (wahrscheinlichste Regel) – Daten des ersten Betriebes

Es ergibt sich bei der Variante Eins (wahrscheinlichste Regel) in Abbildung 3.17 ein ähnliches Bild wie bei der Regression (vgl. Kapitel 3.3.2 auf Seite 53 bzw. Kapitel 3.3.3 auf Seite 54). Die Klassifikationsfähigkeit dieser Variante der Assoziationsregeln ist hierbei geringfügig besser als die Fähigkeit der Regression und liegt bei 9%.

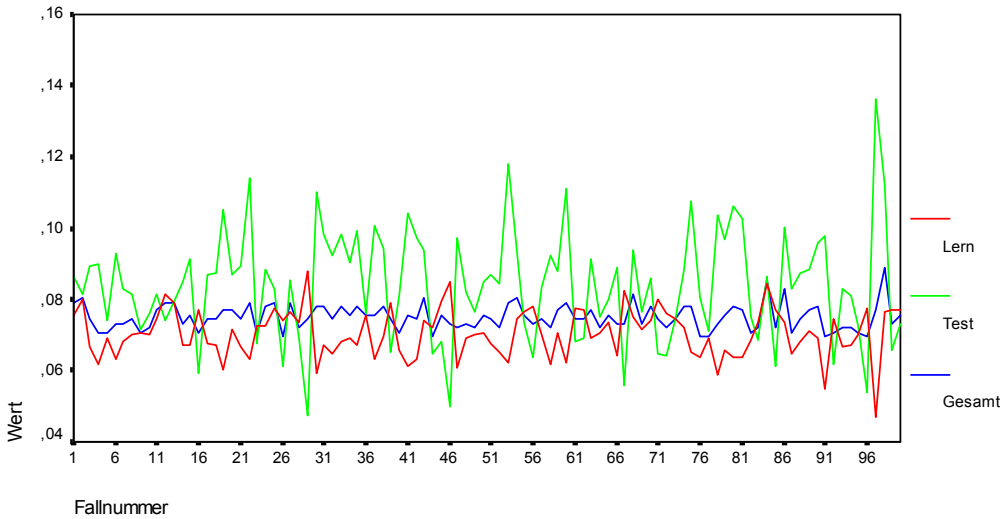

Abbildung 3.18: Klassifikationsfehler des Assoziationsmodells – Variante Zwei (speziellste Regel) – Daten des ersten Betriebes

Die zweite Variante der speziellsten Regel (siehe Abbildung 3.18) ergibt eine leichte Verbesserung der Klassifikation. Der Fehleranteil liegt bei 8% und ist deutlich besser als bei der Regression (siehe Kapitel 3.3.2 auf Seite 53).

3.4.5 Test der Stabilität der Assoziationsregeln als Prognosesystem

Es werden beide Varianten als Prognosesystem eingesetzt, indem statt der vorgegebenen Klasse, deren Durchschnittswert als Umsatzprognose benutzt wird.

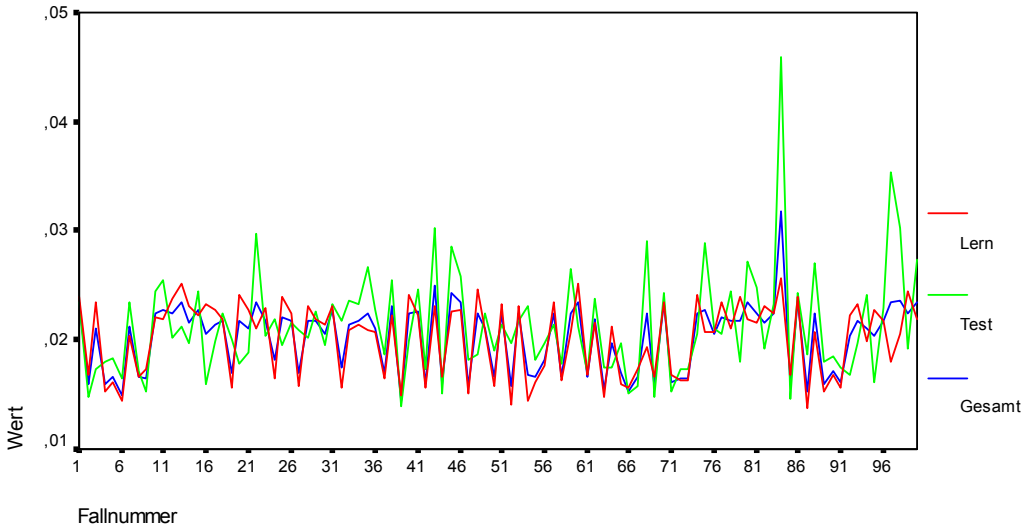

Abbildung 3.19: Prognosefehler des Assoziationsmodells – Variante Eins (wahrscheinlichste Regel) – Daten des ersten Betriebes

Es ergibt sich bei der Variante der wahrscheinlichsten Regel in Abbildung 3.19 ein MSE von 0,02-0,03. Gegenüber dem MSE der Regression von 0,007 zeigt sich hier eine deutlich verminderte Prognosefähigkeit der Assoziationsregeln gegenüber der Regression (vgl. Kapitel 3.3.1 auf Seite 47). Allerdings muss berücksichtig werden, dass, aufgrund der Zugehörigkeit der Assoziationsregeln zu den Klassifikationsalgorithmen, dieser Ansatz insgesamt nur vier verschiedene Werte prognostiziert.

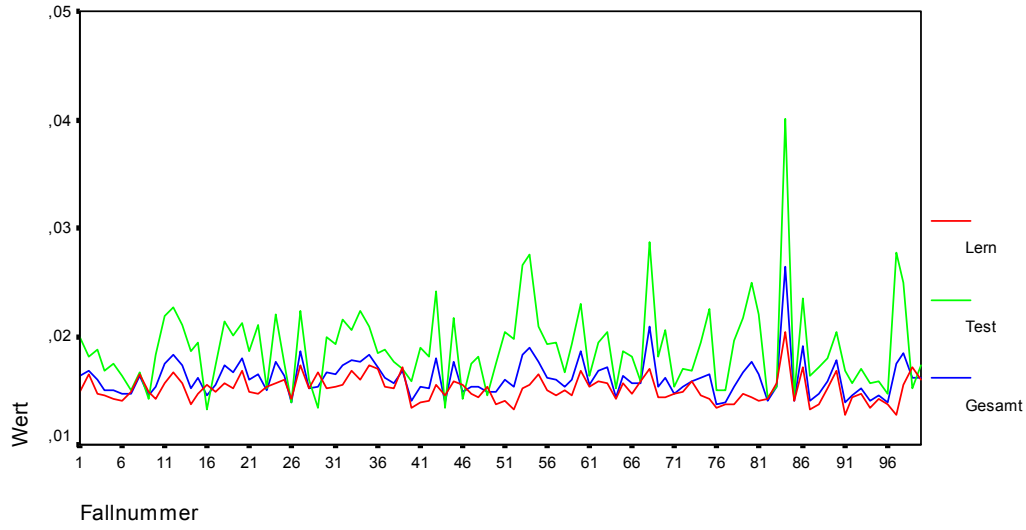

Abbildung 3.20: Prognosefehler des Assoziationmodells – Variante Zwei (speziellste Regel) – Daten des ersten Betriebes

Auch hierbei ist die Variante Zwei der speziellsten Regel (siehe Abbildung 3.20) erfolgreicher. Der MSE liegt zwischen 0,01 und 0,02 und ist damit ebenfalls deutlich schlechter als der Ansatz der Regression.

3.5 Case-Based-Reasoning

Die einfachste und intuitivste Vorgehensweise bei einer Prognose besteht darin, für den jeweiligen Tag einen möglichst ähnlichen Tag aus der Vergangenheit zu wählen und dessen Umsatz als Prognosewert zu übernehmen. Ein Verfahren, dass in diesem Kapitel übernommen und getestet wird.

3.5.1 Case-Based-Reasoning als Klassifikation

Zu jedem Tag aus der Lernmenge wird die Umsatzklasse bestimmt. Ein beliebiger Tag wird klassifiziert, indem alle möglichst ähnlichen Tage aus der Lernmenge gesucht werden. Dem Tag wird diejenige Klasse zugeordnet, die die meisten ähnlichen Tage enthält. Dieses wird mittels einer Eigenimplementation sowohl für die Lernmenge als auch für die Testmenge durchgeführt. Für Lern-, Test- und Gesamtfehler ergibt sich folgendes Bild.

Bemerkenswert in Abbildung 3.21 ist der durchschnittliche Gesamtfehlerwert um 7% der besser ist als Regression und Assoziationsregeln, die bei 8% bzw. 9% liegen.

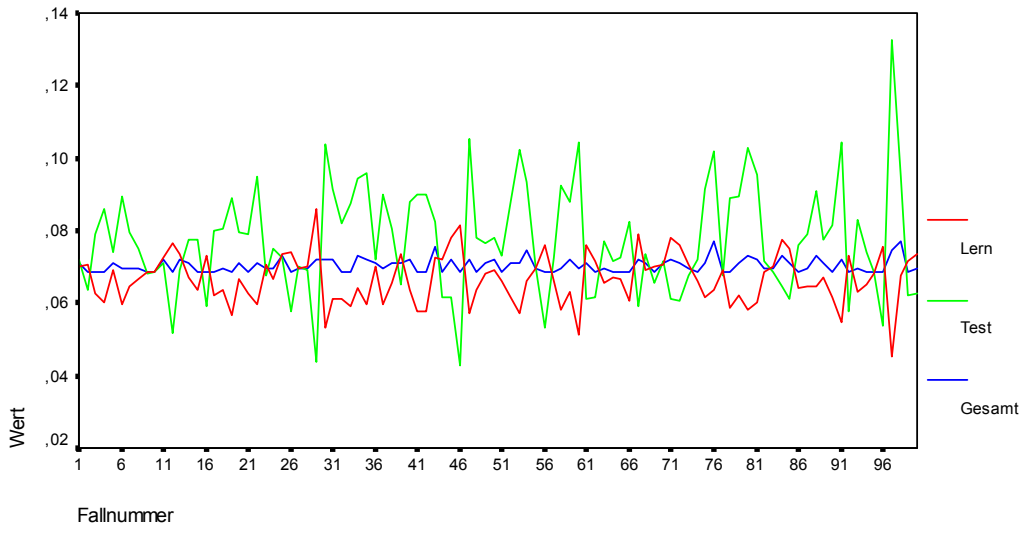

Abbildung 3.21: Klassifikationsfehler des fallbasierten Ansatz – Daten des ersten Betriebes

3.5.2 Case-Based-Reasoning als Prognose

Bei diesem Ansatz wird als Prognosewert der Durchschnitt aller Umsatzwerte der Tage gebildet, die in der Lernmenge dem zu prognostizierenden Tag am ähnlichsten sind.

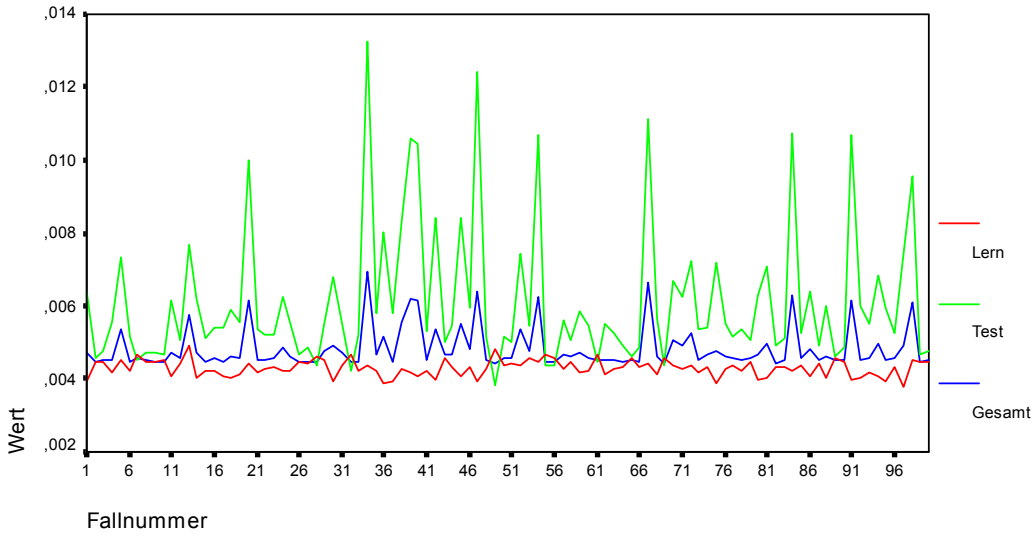

Abbildung 3.22: Prognosefähigkeit des fallbasierten Ansatz – Daten des ersten Betriebes

Es errechnet sich ein MSE von 0,006. Dieser Ansatz bringt auch in seiner Prognosefähigkeit die besten Ergebnisse.

3.6 Vergleich Regression, Assoziationsregeln und Case-Based-Reasoning

Der Vergleich der erarbeiteten Modelle entspricht dem Schritt 'Interpretation und Bewertung' in Abbildung 2.2 auf Seite 10.

In diesem Kapitel wird untersucht, welche Methode für die Modellierung der Daten geeignet sind. Hierbei wird analysiert in welchen Fällen die einzelnen Methoden Vor- bzw. Nachteile aufweisen.

3.6.1 Vergleich der Fehlerraten bei den Klassifikationsansätzen

Es ergibt sich die folgende Graphik bei den zu testenden Daten.

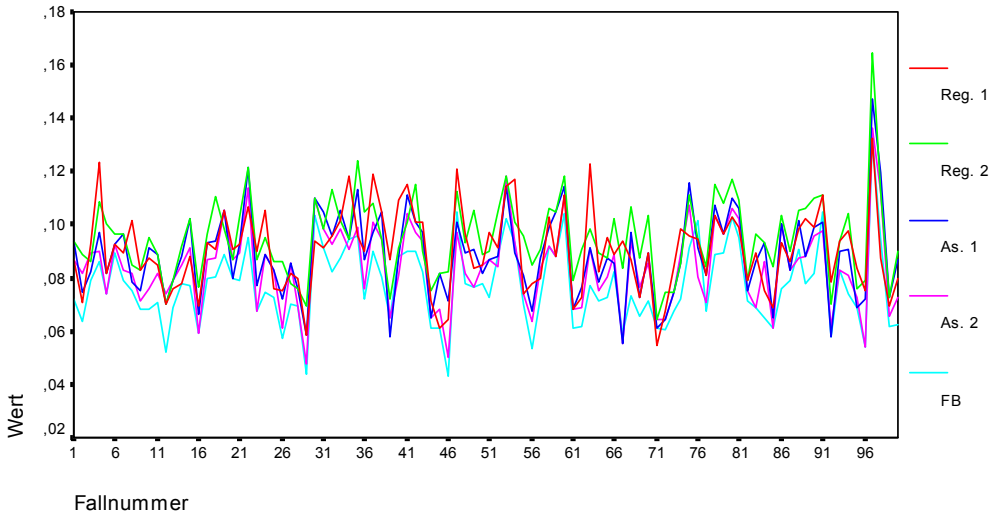

Abbildung 3.23: Vergleich der Klassifikationsfehler der Testdaten – Daten des ersten Betriebes

Die Abbildung 3.23 zeigt die Fehlerraten der einzelnen Algorithmen bezogen auf die Lerndaten. Es ist deutlich erkennbar, dass bestimmte Lernmengen allen Algorithmen Schwierigkeiten bereiten. Hierdurch bestätigt sich die schon genannte Hypothese, dass eine Teilmenge der Daten aussreicht um ein gutes Modell zu entwickeln. Trotzdem existieren Datensätze, die schwer in diese Modelle integriert werden können. Je nachdem, ob diese Daten zu den Lern- oder Testdaten gehören, erreichen die Ansätze gute oder schlechte Klassifikationswerte.

Zur Verdeutlichung wird die durchschnittliche Fehlerrate dargestellt.

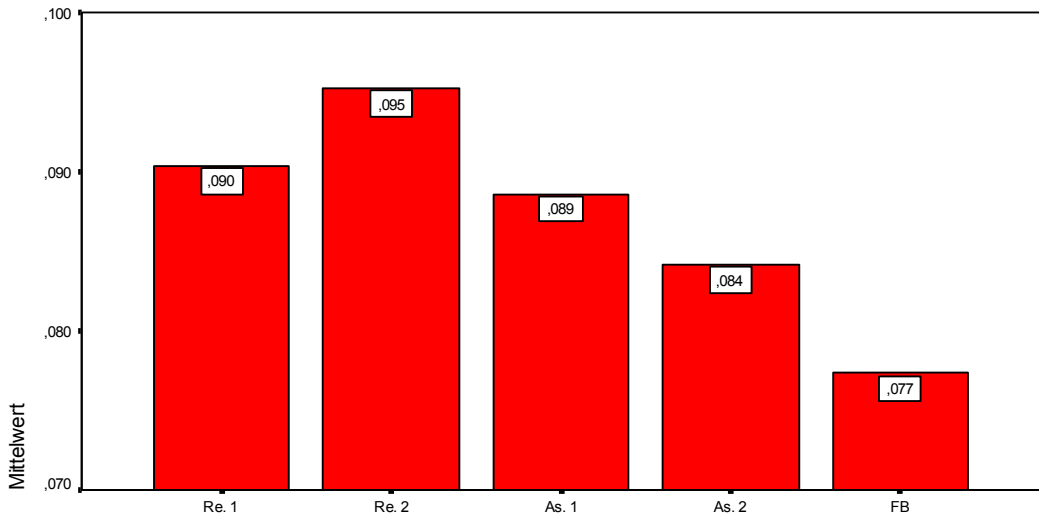

Abbildung 3.24: Graphische Veranschaulichung des Klassifikationsfehlers – Daten des ersten Betriebes

Die Abbildung 3.24 zeigt deutlich die gute Klassifikationsfähigkeit des Case-Based-Reasoning mit 7% Klassifikationsfehler. Es folgen die Assoziationsregeln mit 8% und die Regression mit 9% Fehler.

Es ergibt sich folgende Graphik bei den Gesamtdaten.

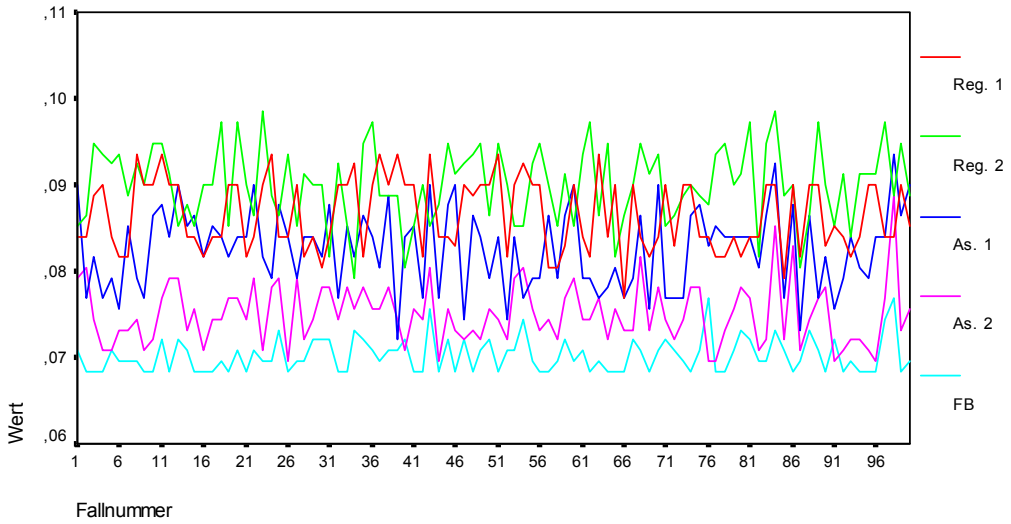

Abbildung 3.25: Vergleich der Klassifikationsfehler der Gesamtdaten – Daten des ersten Betriebes

In der Abbildung 3.25 ist deutlich die geringere Varianz der Fehler bezüglich der Gesamtdaten zu sehen. Hierbei muss berücksichtigt werden, dass hier ein geringer Bereich dargestellt wird als in Abbildung 3.23. Es ist nun deutlich die Rangfolge der einzelnen Algorithmen erkennbar.

3.6.2 Vergleich der MSE bei den Prognoseansätzen

Es ergibt sich folgende Graphik bei den zu testenden Daten.

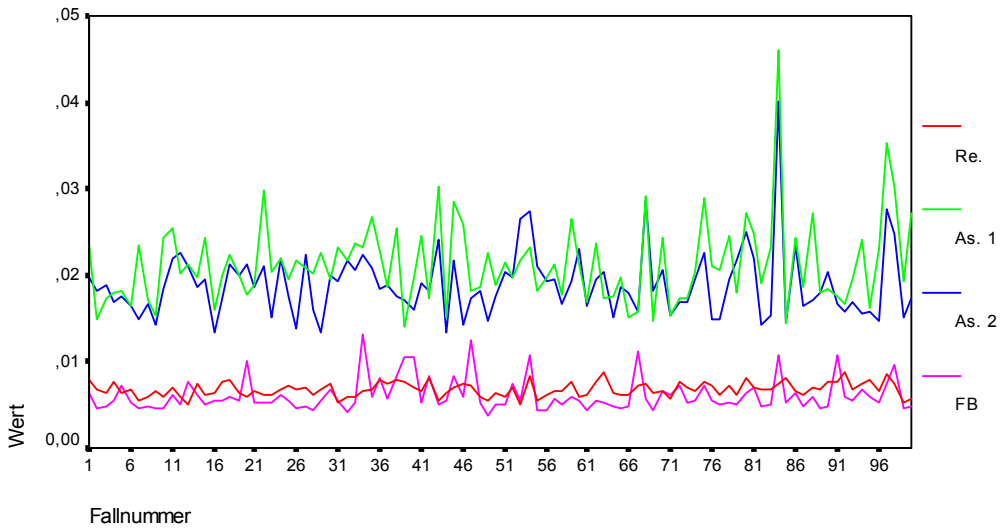

Abbildung 3.26: Vergleich der Prognosefehler der Testdaten – Daten des ersten Betriebes

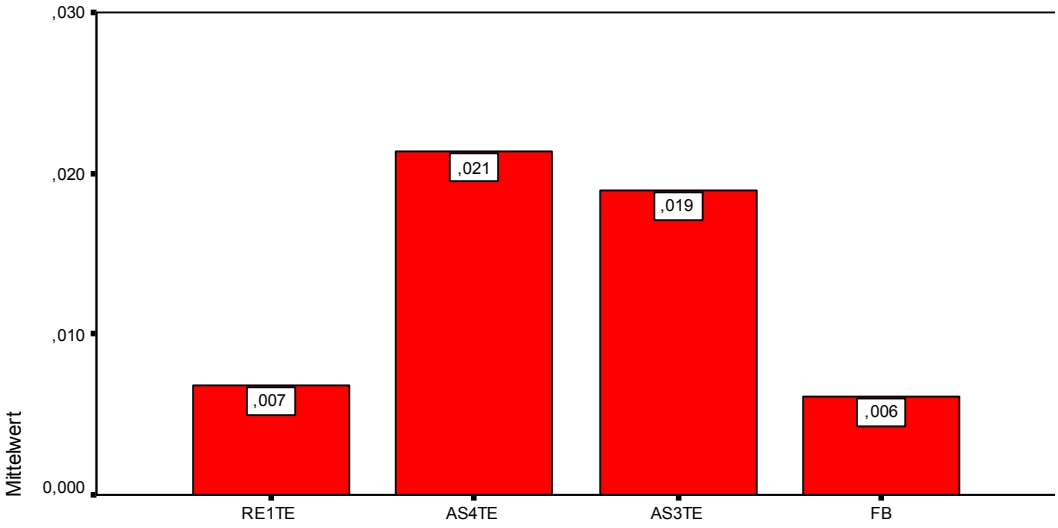

Abbildung 3.27: Graphische Veranschaulichung des Prognosefehlers – Daten des ersten Betriebes

In diesem Fall erweisen sich die Ansätze der Regression und des CBR gegenüber den Assoziationsregeln als überlegen. Allerdings ist dies bei dem Einsatz eines Klassifikationssystems als Prognosesystem nicht anders zu erwarten gewesen.

Es ergibt sich folgende Graphik bei den Gesamtdaten.

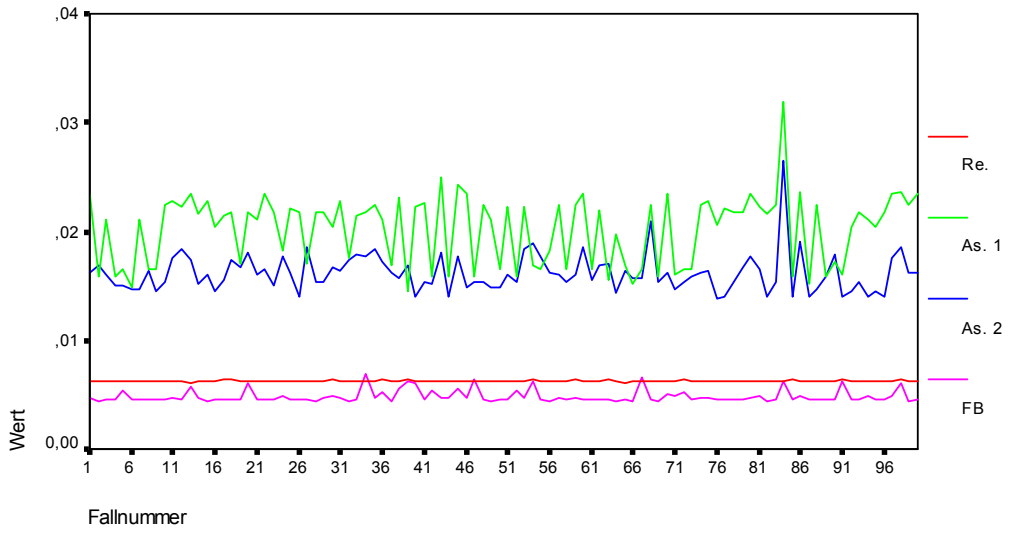

Abbildung 3.28: Vergleich der Prognosefehler der Gesamtdaten – Daten des ersten Betriebes

In Abbildung 3.28 ist nochmals der Effekt zu sehen, dass die Varianz des MSE bezüglich der Gesamtdaten geringer ist, als die Varianz des MSE hinsichtlich der Testdaten. Beachtenswert ist ebenfalls die geringe Varianz des MSE der Regression.

3.6.3 Vergleich der einzelnen Klassifikationsfehler in einem Durchlauf

In der folgenden Untersuchung wird für jeden Datensatz ermittelt, welche Methode bei welchen Daten Klassifikationsfehler aufweist. In diesem Zusammenhang wird die Frage beantwortet, ob die Methoden verstärkt an den gleichen Stellen Fehler erzeugen oder unterschiedliche Schwachstellen besitzen.

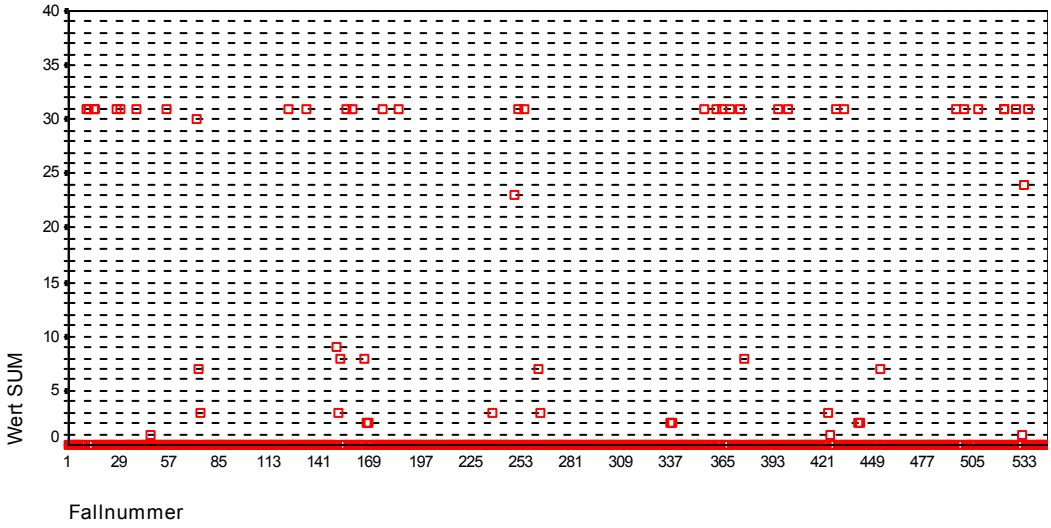

Abbildung 3.29: Graphische Veranschaulichung der einzelnen Klassifikationsfehler pro Methode der Lerndaten – Daten des ersten Betriebes

Die Abbildung 3.29 der Klassifikationsfehler bei den Lerndaten ist folgendermaßen zu interpretieren.

Je nachdem welcher Algorithmus einen Fehler erzeugt, wird dieser Fehler anders gewichtet.

Es werden die Gewichtungen RE1: 1, RE2: 2, AS1: 4, AS2: 8, CBR: 16 gewählt. Abbildung 3.29 zeigt die Summe der einzelnen gewichteten Fehler für jeden Testdatensatz.

Auswertung der Graphik:

- Eine klare Linie ist bei dem Wert 31 erkennbar. Dabei handelt es sich um Fehler die von allen Algorithmen gemacht werden.
- Weitere Linien liegen bei dem Wert 1 (Nur Re. 1), 2 (Nur Re. 2), 3 (Regressionsfehler), 7 (Regressionsfehler und As. 1) und 8 (nur As. 2).
- Es gibt keine Linie bei 16, also Fehler, die nur von dem CBR gemacht werden.
- Der Bereich 9 bis 29 ist nahezu leer. Es sind mehrheitlich gemeinsame Fehler von CBR, AS2 und AS1 begangen worden.

Fehlklassifikationen bei den Testdaten:

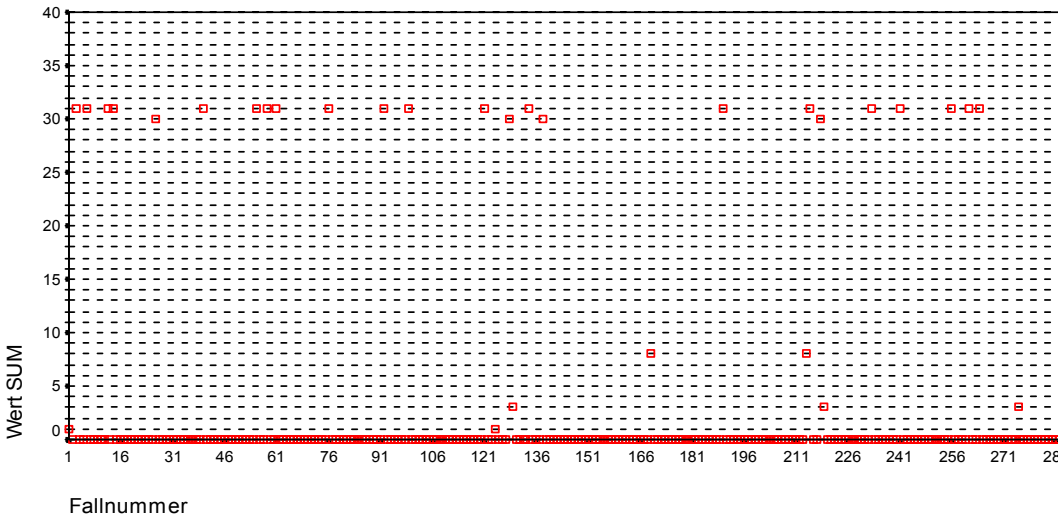

Abbildung 3.30: Graphische Veranschaulichung der einzelnen Klassifikationsfehler pro Methode der Testdaten – Daten des ersten Betriebes

Es ergibt sich für die Testdaten ein ähnliches Bild wie bei den Lerndaten.

4 Neue Ansätze und Erweiterungen zur Auswertung

Die folgenden Ansätze zeigen Möglichkeiten auf, um die durchgeführte Grunduntersuchung zu intensivieren. Es sollen Wege eröffnet werden, wie die Ergebnisse der bisherigen Methoden, Regression, Assoziationsregeln und Case-Based-Reasoning verbessert werden können. Des Weiteren sollen mögliche Ansätze durch neue Methoden aufgezeigt werden, die aufgrund des Umfangs dieser Arbeit sowie der geringen Datenmenge in dieser Untersuchung nur in Ansätzen bearbeitet werden konnten.

4.1 Clusteranalyse

Die Daten sind mittels eines Clusteralgorithmus analysiert und in die gefundenen Cluster einsortiert worden. Interessant ist die Frage, ob diese Klassen in einem Zusammenhang zu den Merkmalen Wochentag und Feiertagsnähe stehen. Hierzu wird eine Häufigkeitstabelle benutzt, die für jede Klasse deren Struktur analysiert. Für jede Klasse wird festgehalten, wie viel Montage, Dienstage usw. sich in dieser Klasse befinden. Weiterhin wird gezählt, wie viele Montage es vor einem Feiertag gibt usw. Besondere Aufmerksamkeit erhalten hierbei die Klassen, die von einem Merkmal (also Wochentag plus Feiertagszugehörigkeit) mehr als fünf Elemente enthalten. Die Untersuchung hat keine zufriedenstellenden Ergebnisse geliefert. Bei der Clusteranalyse sind 72 Klassen gefunden worden, von denen es nur bei den Wochentagen mehr als fünf Elemente gibt, die nicht in der Nähe eines Feiertages liegen. Die interessanteren Mengen von Wochentagen, die in der Nähe von Feiertagen liegen, sind vereinzelt bzw. zu zweit in einer Klasse zu finden gewesen. Es bildet sich zum Beispiel keine Klasse von Montagen vor einem Feiertag.

Trotzdem zeigt die Clusteranalyse, dass die Daten ausreichend strukturiert sind, um viele Cluster zu identifizieren. Demnach gibt es Klassen von Umsätzen, die identifiziert werden können.

4.2 Neuronale Netze

Um weitere Vergleichsmöglichkeiten zu erhalten, sind die Daten zusätzlich mit Hilfe eines neuronalen Netzes prognostiziert worden. Hierbei ist von Interesse, ob die neuronalen Netze bessere oder schlechtere Prognosen berechnen, als die übrigen Modelle. In diesem Zusammenhang ist sowohl die Variante mit als auch ohne Mittelschicht betrachtet worden.

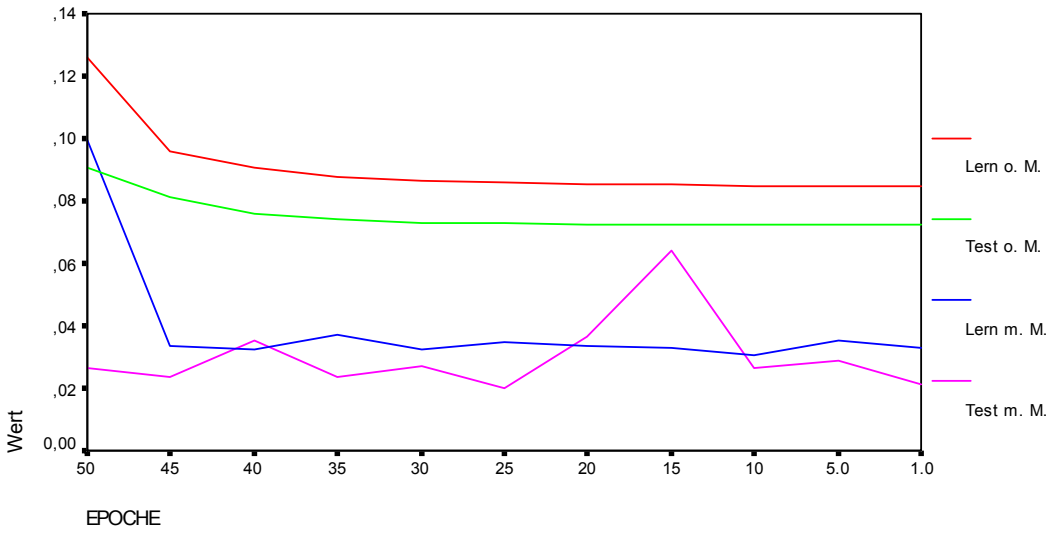

Abbildung 4.1: Lernverhalten des neuronalen Netzes mit und ohne Mittelschicht

Mit dem neuronalen Netz mit Mittelschicht werden deutlich bessere Werte als mit dem neuronale Netz ohne Mittelschicht berechnet. Dies lässt auf einen komplexen Zusammenhang zwischen den Umsatzdaten und Merkmalen schließen, der durch ein neuronales Netz ohne Mittelschicht nicht repräsentiert werden kann. Dargestellt werden in der Abbildung 4.1 der MSE beider neuronaler Netze bezogen sowohl auf die Lernmenge als auch auf die Testmenge. Im Vergleich zu den bisher präsentierten Algorithmen schneiden neuronale Netze eindeutig schlechter ab. Selbst die für die Prognose nicht geeigneten Assoziationsregeln haben, basierend auf den gleichen Daten, einen MSE von maximal 0,021. Die Prognosefähigkeit des neuronalen Netzes liegt somit im Bereich der Assoziationsregeln und ist deutlich schlechter als der Ansatz CBR und Regression (siehe Abbildung 3.27).

4.3 Feststellung der Defizite der berechneten Regressionsmodelle

Da die Regression im Gegensatz zu den Assoziationsregeln keine Klassifikationsfehler, sondern numerische Fehler aufweist, lassen sich genauere Aussagen darüber treffen, welche Tage am schlechtesten über- bzw. unterbewertet werden. Werden diese Falschbewertungen sortiert, können Ansatzpunkte darüber gewonnen werden, welche Problemfälle durch das Modell schlecht abgedeckt werden. Diese Untersuchung ist für die vorliegenden Daten mit einem den vorherigen Untersuchungen entsprechendem Regressionsmodell durchgeführt und ausgewertet worden.

Aus den Tabellen im Anhang „Darstellung der zwanzig größten Über- und Unterbewertungen des Regressionsmodells“ auf Seite 100 lässt sich ablesen, dass die auf den 24. und 31. Dezember fallenden Tage überbewertet werden, da die Geschäfte nur halbe Tage geöffnet sind. Folglich müssen diese Tage speziell berücksichtigt werden. Demgegenüber sind die anderen Tage im Dezember unterbewertet. Hier liegt also ein umsatzstarker Monat vor, der ebenfalls gesondert behandelt werden sollte. Die restlichen Daten ergeben keine weiteren deutlichen Zeiträume, die Besonderheiten aufweisen. Allerdings sind bestimmte Feiertags/Wochentagskombinationen unter- bzw. überbewertet. Zum Beispiel ist der Samstag überbewertet, wenn es sich bei dem Montag um einen Feiertag handelt. Dies bedeutet, dass der Einfluss des Feiertages auf den Samstag geringer ist als erwartet. Eine weitere Überbewertung liegt häufig an einem Dienstag vor, wenn der folgende Donnerstag ein Feiertag ist. Auf der anderen Seite wird der Mittwoch zu gering eingeschätzt, wenn es sich beim nächsten Tag um einen Feiertag handelt usw.

Diese Untersuchung lässt es sinnvoll erscheinen, keine generelle Regression der Gesamtdaten vorzunehmen, sondern getrennt für jeden Wochentag. Um genügend statistisch signifikante Aussagen zu treffen, ist der Datenbestand zu gering. Trotzdem wird diese Art der Untersuchung exemplarisch im nächsten Kapitel durchgeführt. Zuvor werden die Daten von zwei weiteren Betrieben, die den selben Zeitraum abdecken, untersucht, um festzustellen, ob die Struktur der Daten soviel Gemeinsamkeiten aufweist, dass eine gemeinsame Untersuchung gerechtfertigt ist. Damit wird der Datenbestand soweit aufgestockt, dass eine getrennte Untersuchung für jeden einzelnen Wochentag möglich ist, selbst wenn die notwendige Anzahl an Daten noch höher liegt.

4.4 Einbeziehung der Daten zweier weiterer Betriebe innerhalb des Untersuchungszeitraums

Ziel dieses Kapitels ist die Beantwortung folgender Fragestellungen:

- Ist die Struktur der neuen Daten vergleichbar mit den schon untersuchten Daten? Bestätigt sich das erforschte Modell der Regression und Assoziationsregeln?
- Können alle drei Betriebe zusammen analysiert werden oder muss eine getrennte Betrachtung erfolgen?

Zunächst werden die Daten der drei Betriebe mittels der Regression getrennt untersucht und die Ergebnisse miteinander verglichen. Weisen die Modelle genügend Übereinkünfte auf, werden die Daten der drei Betriebe mittels der Regression und Assoziationsregeln untersucht.

4.4.1 Untersuchung der Daten von Betrieb 1, 2 und 3

Die Daten werden mittels der Regression nacheinander untersucht und die Ergebnisse der Variablen tabellarisch dargestellt.

Folgende Ergebnisse der Regression der drei Datenmengen sind festzustellen:

Variable	*Betrieb 1*	*Betrieb 2*	*Betrieb 3*
Montag	0,734	0,809	0,875
Dienstag	0,700	0,755	0,882
Mittwoch	0,793	0,750	0,804
Donnerstag	0,964	0,933	0,937
Freitag	1,410	1,444	1,204
Samstag	1,115	0,986	1,040
VF4	0,055	0,056	0,050
VF3	0,106	0,105	0,082
VF2	0,301	0,293	0,220
VF1	0,644	0,669	0,656
NF1	0,071	0,093	0,055

Tabelle 4.1: Regressionsmodelle der einzelnen Betriebe

Die Ergebnisse stützen die Annahme, dass das untersuchte Basismodell nicht für einen, sondern für mehrere Betriebe eingesetzt werden kann. Nachfolgend werden die Daten mittels der Regression und Assoziationsregeln untersucht. Dies bringt den Vorteil, dass eine größere Menge von Daten für die Untersuchung vorhanden ist. Insbesondere trifft dies für die Assoziationsregeln zu, da hier der minimale Support der Regeln weit herab gesetzt werden muss, um Regeln zu erhalten.

4.4.2 Untersuchung der Gesamtdaten mittels Regression und Assoziationsregeln

Die Regression berechnet folgendes Modell:

Modell	R	R-Quadrat[a]	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,994[b]	,988	,988	,1101

Abbildung 4.2: Bestimmtheitsmaß des Regressionsmodells – Daten aller Betriebe

Aufgrund der Tatsache, dass Daten von drei verschiedenen Betrieben eingesetzt werden, nimm die Gesamtqualität des Regressionsmodells ab. Der Wert R-Quadrat beträgt bei diesem Modell nur 0,988 gegenüber 0,994 bei der ersten Regression (siehe Abbildung 3.8).

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	2550,634	11	231,876	19111,817	,000[a]
	Residuen	30,186	2488	1,213E-02		
	Gesamt	2580,820[b]	2499			

Abbildung 4.3: F-Test des Regressionsmodells – Daten aller Betriebe

Der F-Test in Abbildung 4.3 zeigt eine über 99%tige Sicherheit des Regressionsmodells an, wie bei der vorherigen Regression (vgl. Abbildung 3.9).

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	MONTAG	,806	,006	,320	145,794	,000
	DIENSTAG	,779	,005	,316	142,345	,000
	MITTWOCH	,782	,005	,317	143,378	,000
	DONNERST	,945	,006	,374	170,611	,000
	FREITAG	1,353	,005	,542	246,667	,000
	SAMSTAG	1,047	,005	,424	191,864	,000
	VF4	5,588E-02	,014	,009	3,918	,000
	VF3	9,735E-02	,014	,016	6,934	,000
	VF2	,271	,014	,043	19,316	,000
	VF1	,656	,014	,105	46,683	,000
	NF1	7,261E-02	,013	,012	5,536	,000

Abbildung 4.4: Regressionsmodell – Daten aller Betriebe

Es ergeben sich leichte Veränderungen der einzelnen Variablen. So steigen die Werte für Montag und Dienstag leicht an, während die Variablen für die anderen Tage leicht gefallen sind.

Die Assoziationsregeln werden durch folgendes Modell dargestellt.

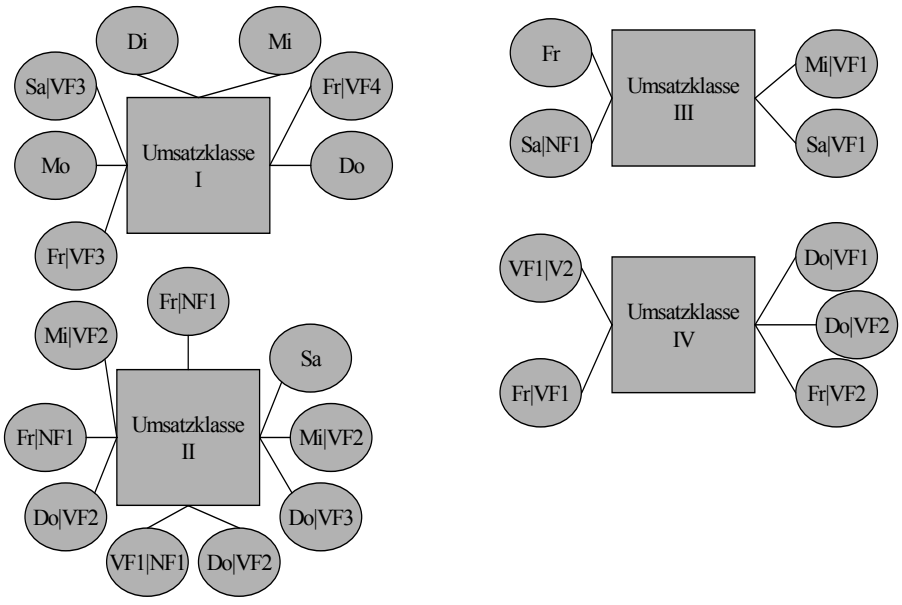

Abbildung 4.5: Assoziationsmodell bei Daten aller Betriebe

Bei der gleichen Supportgrenze ergeben sich ungefähr doppelt soviele Regeln als zuvor. In Abbildung 4.5 ist deutlich zu erkennen, dass das Modell optimiert werden konnte. Die Anzahl der Regeln insgesamt und die interessanten Regeln mit mehr als einem Merkmal in der Prämisse sind deutlich angestiegen. Die Regeln werden im Anhang „Assoziationsregeln – Alle Betriebe“ auf Seite 101 aufgelistet.

4.4.3 Vergleich Regression, Assoziationsregeln und Case-Based-Reasoning der Gesamtdaten

Ein weiterer interessanter Aspekt ist die Verhaltensweise der drei Methoden bei Erhöhung der Datenmenge. Basierend auf den Gesamt- und Lerndaten wird das Lernverhalten der drei Methoden wiederholt getestet.

Vergleich der Klassifikationsfähigkeit

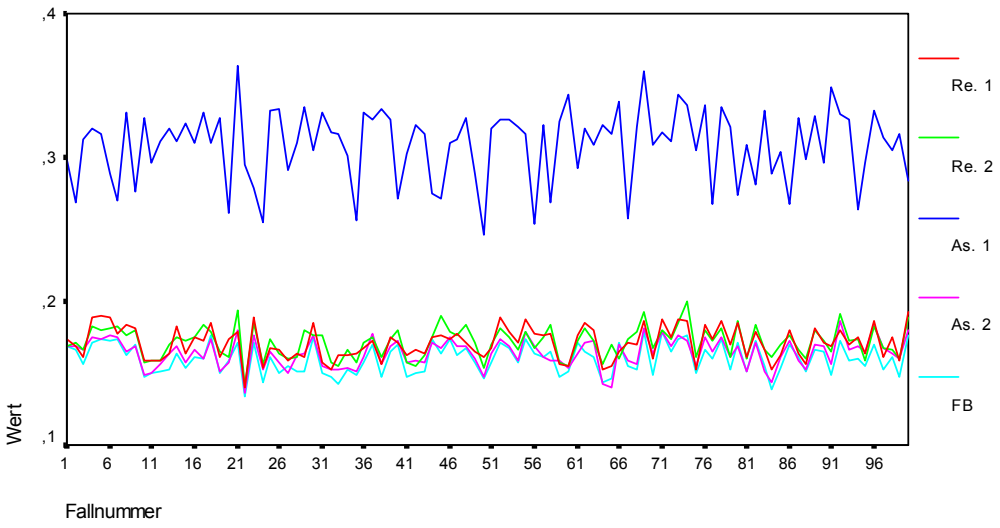

Abbildung 4.6: Vergleich der Klassifikationsfähigkeit bei den Testdaten – Daten aller Betriebe

Die Abbildung 4.6 verdeutlicht, wie die Klassifikationsfähigkeit der Assoziationsregeln Variante Eins (wahrscheinlichste Regel) nachlässt. Weiterhin verschlechtert sich die Klassifikationsfähigkeit aller Algorithmen von 8% auf 17%. Die Anzahl der Klassifikationsfehler haben sich demnach mehr als verdoppelt. Das liegt an der komplexen Prognose der Umsätze von drei verschiedenen Betrieben und der damit verbundenen Zunahme unerwünschter Merkmale, die Einfluss auf die Daten ausüben.

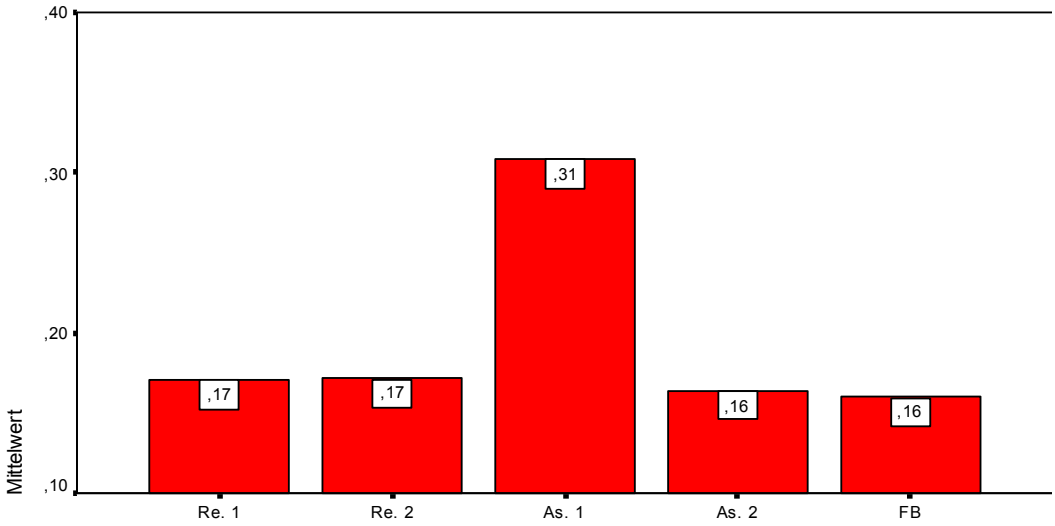

Abbildung 4.7: Graphische Veranschaulichung der Klassifikationsfähigkeit bei den Testdaten – Daten aller Betriebe

In Abbildung 4.7 ist deutlich die schlechte Performance des Assoziationsalgorithmus – Variante Eins (wahrscheinlichste Regel) zu erkennen. Alle anderen Algorithmen haben ähnliche Erfolgsquoten, die um die 16% liegen. Abbildung 4.8 zeigt die Performance bei den Testdaten.

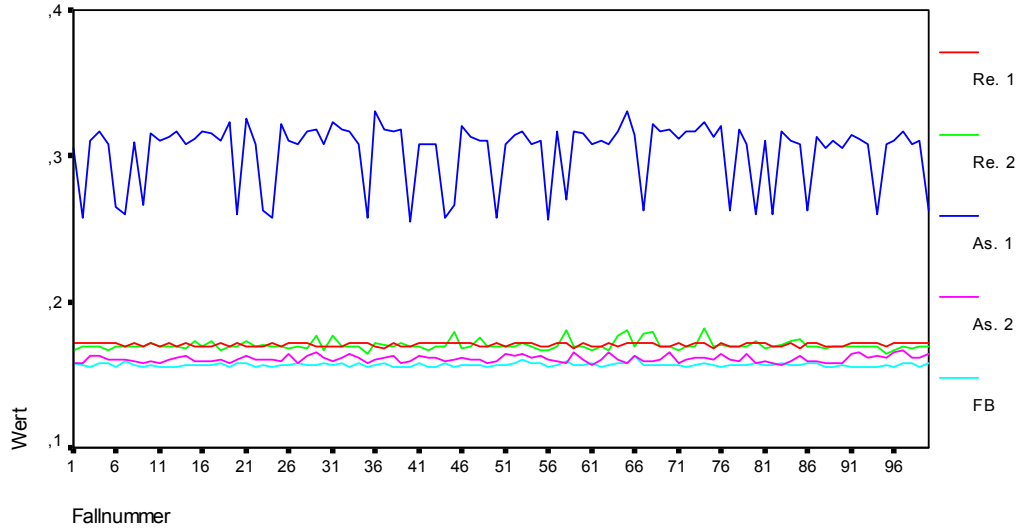

Abbildung 4.8: Vergleich der Klassifikationsfähigkeit bei den Gesamtdaten – Daten aller Betriebe

Vergleich der Prognosefähigkeit

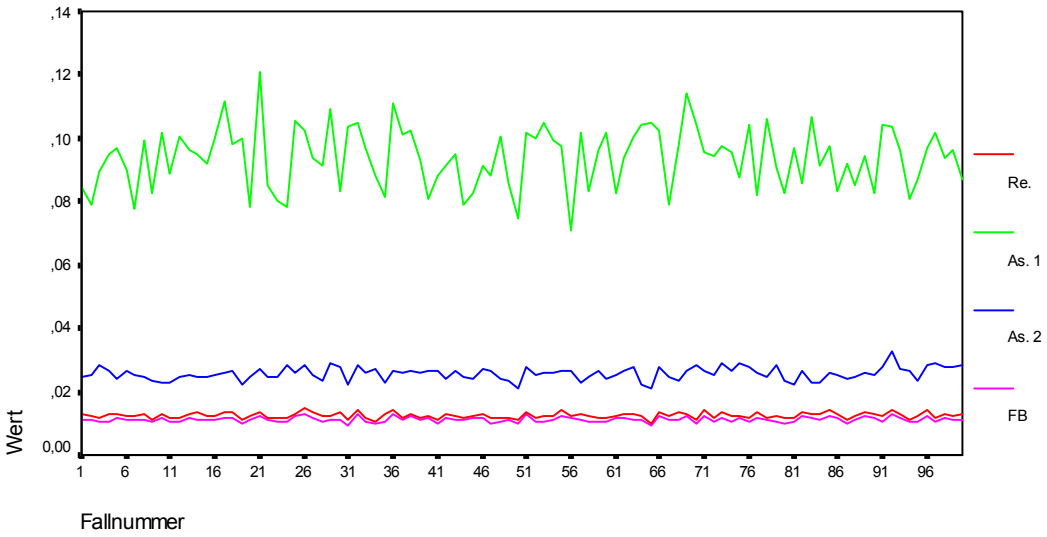

Abbildung 4.9: Vergleich der Prognosefähigkeit bei den Testdaten – Daten aller Betriebe

Abbildung 4.9 zeigt ebenfalls die schlechten Prognoseeigenschaften des Assoziationsalgorithmus – Variante Eins (wahrscheinlichste Regel). Deutlich besser, aber immer noch schlechter als die klassischen Prognosealgorithmen Regression und CBR, ist der Assoziationsalgorithmus – Variante Zwei (speziellste Regel). Eine genaue Übersicht der einzelnen Prognoseeigenschaften bietet folgende Graphik.

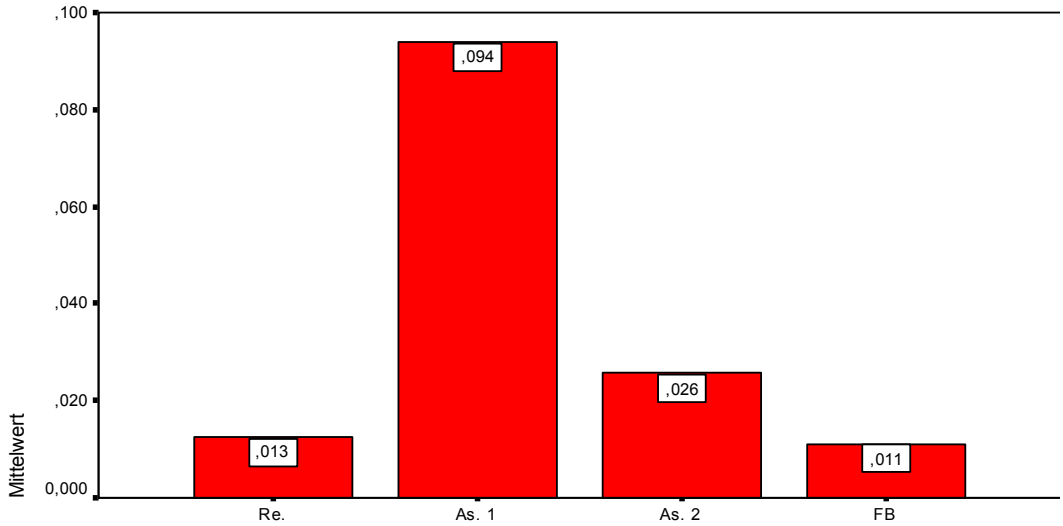

Abbildung 4.10: Graphische Veranschaulichung der Prognosefähigkeit bei den Testdaten – Daten aller Betriebe

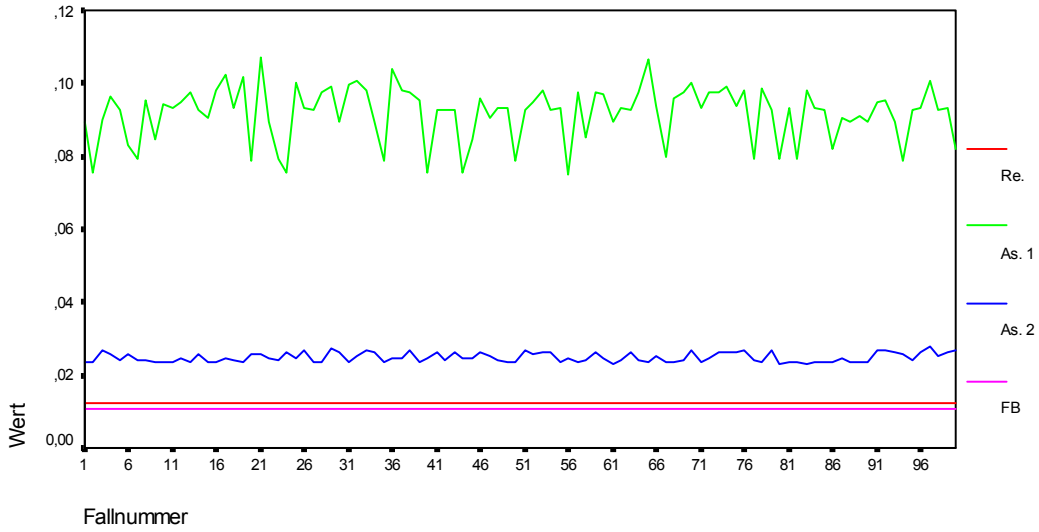

Abbildung 4.11: Vergleich der Prognosefähigkeit bei den Gesamtdaten – Daten aller Betriebe

Es lässt sich gleichfalls eine Verschlechterung der Prognosefähigkeit bezüglich aller Methoden feststellen. Deutlich schlechtere Ergebnisse liefert der Assoziationsalgorithmus – Variante Eins (wahrscheinlichste Regel).

4.5 Getrennte Regression für jeden Wochentag

In der Untersuchung der Modelldefizite, Kapitel 4.3 auf Seite 70, zeigt sich bei der Regression, dass das Merkmal „Feiertag am nächsten Werktag“ auf einige Wochentage eine größere Wirkung ausübt, als bei anderen Wochentagen. Daher erscheint es sinnvoll ein Regressionsmodell getrennt für jeden Wochentag zu berechnen:

	Tages-variable	*VF4*	*VF3*	*VF2*	*VF1*	*NF1*
Montag	0,809	0,047	0,026			0,120
Dienstag	0,783	0,038	0,117	0,127		0,129
Mittwoch	0,773	0,069	0,139	0,295	0,782	
Donnerstag	0,938		0,141	0,270	0,855	-0,177
Freitag	1,352			0,331	0,672	0,058
Samstag	1,058	0,051			0,336	0,190

Tabelle 4.2: Regressionsmodelle für jeden Wochentag

Die Schwankungen bei den Variablen VF2 und VF1 lassen sich deutlich ablesen. Aufgrund der Tatsache, dass in dem untersuchten Zeitraum nicht alle Wochentags-Feiertagskombinationen enthalten sind, sind die Regressionsmodelle nicht vollständig. Trotzdem zeigt dieser Ansatz, dass diese Art der Untersuchung bessere Ergebnisse verspricht, als die normale Regression aller Daten. Die Prognosefähigkeit ist bei einigen Wochentagen deutlich gesteigert worden. Der MSE verändert sich von vorher 0,012 auf 0,004. Allerdings weisen die Modelle für den Freitag und Samstag schlechtere MSE Werte auf.

4.6 Regression mit komplexen Features

In dieser Untersuchung werden die einzelnen Variablen sinnvoll miteinander verknüpft, um Beziehungen der Variablen untereinander zu analysieren. Für diese Untersuchung sind die Wochentagsvariablen mit jeder möglichen Feiertagsvariablen multiplikativ (Und) verknüpft worden. Basierend darauf werden zusätzliche Merkmale identifiziert, wie zum Beispiel: „Montag und der folgende Dienstag ist ein Feiertag“. Durch diese neuen Merkmale wird eine Regression für jeden Wochentag erreicht. Das Ergebnis:

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	2554,116	27	94,597	8756,821	,000[a]
	Residuen	26,704	2472	1,080E-02		
	Gesamt	2580,820[b]	2499			

Abbildung 4.12: Bestimmtheitsmaß der Regression mit komplexen Features

In der Abbildung 4.12 ist eine leichte Verbesserung von einem MSE von 0,0121 bei der normalen Regression zu 0,0108 bei der Regression mit komplexen Features abzulesen. Da diese Vorgehensweise einer Regression für jeden Wochentag entspricht, können analog zu dem vorangegangenen Kapitel 4.4 auf Seite 71 Verbesserungen erzielt werden.

5 Fazit und Ausblick

Im folgenden Kapitel wird eine abschließende Darstellung der Ergebnisse dieser Arbeit vorgestellt. Außerdem sollen hier die Gebiete benannt werden, die einer weiteren Untersuchung bedürfen. Abschließend erfolgt ein Ausblick hinsichtlich der Fragestellung, wie sich eine Umsatzprognose positiv in das jetzige Unternehmensmanagement eingliedern lassen kann.

5.1 Fazit

Die vorliegende Diplomarbeit liefert sowohl einen theoretischen Überblick des Gebietes WED, als auch eine praktische Anwendung und Bewertung der WED-Methoden bezogen auf die Umsatzprognose im Lebensmitteleinzelhandel.

Bei der Bewertung der Methoden zeigt sich, dass im Vergleich zur Wahl des Algorithmus, die Vorbereitung der Daten entscheidend ist. Die Durchführung einer Datenanalyse und die Bestimmung der korrekten Merkmale erweist sich als unvermeidlich für den Erfolg der angewendeten Algorithmen. Sowohl das Verständnis der Algorithmen und deren Vorraussetzungen, als auch das Verständnis der Daten und deren Restriktionen sind beim effizienten und erfolgreichen Einsatz der Algorithmen notwendig. Das Verständnis vorausgesetzt, können bei den einzelnen Methoden ähnliche Erfolgsquoten erreicht werden.

Daraus ergibt sich unter anderem die Überzeugung des Autors dieser Diplomarbeit, dass der relativ ergebnislose Einsatz des Clustering nicht darauf zurückzuführen ist, dass das Clustering zur Analyse von Umsatzdaten ungeeignet ist. Vielmehr fehlte im Bereich dieser Diplomarbeit die Zeit, um die Bedeutung der Klassen, die vom Clustering-Algorithmus erzeugt wurden, vollständig zu verstehen.

Grundsätzlich beweist diese Diplomarbeit, dass eine Prognose von Umsatzdaten mit sehr guten Ergebnissen und mit deutlichen Vorteilen für Einzelhändler möglich ist. Ein Vorschlag für eine zukünftige Implementierung wird im nächsten Kapitel gegeben.

5.2 Ausblick

Die folgenden Punkte bieten weitere potentielle Untersuchungen, die auf der Basis dieser Diplomarbeit aufbauen können.

- Wie genau setzt sich der Einfluss des in dieser Diplomarbeit nicht untersuchten Dezemberzeitraums auf die Umsatzdaten zusammen?
- Wie lassen sich die Ergebnisse der Methoden Clustering und neuronale Netze effektiver nutzen?
- Inwieweit lässt sich die Prognose verbessern, indem andere Merkmale (Wetter, Urlaubszeitraum...) hinzugenommen werden?
- Welche Ergebnisse können aus Massendaten gewonnen werden (10 Umsatzjahre eines Betriebes)?

Eine zukünftige Anwendung dieser Erkenntnisse liegt in der computerunterstützten Umsatzprognose, mit Anbindung an eine automatisierte Personaleinsatzplanung.

Eine computerberechnete Personaleinsatzplanung kann in der Regel nur ein Vorschlag für den Einzelhändler sein. Daher ist es unerlässlich, dass der Einzelhändler jeden Schritt der Umsatzprognose versteht und beeinflussen kann. Eine Vorraussetzung an die Umsatzprognose ist demzufolge die Verständlichkeit und Belegbarkeit der Ergebnisse. Der Methodenbereich des Cased-Based-Reasoning hat sich als besonders geeignet erwiesen. Dieser Ansatz liefert in dieser Untersuchung optimale Ergebnisse, ist leicht verständlich und zur Unterstützung des Ergebnisses können durch das Case-Based-Reasoning ähnliche Tage aus der Vergangenheit aufgezeigt werden.

Ein weitere Vorraussetzung für eine spätere automatisierte Umsatzprognose besteht in der Notwendigkeit langfristige Veränderungen, Sprünge und unvorhergesehene Ereignisse berücksichtigen zu können. Die zukünftige Implementation eines Systems zur Umsatzprognose sollte in der Lage sein, langfristige oder sprunghafte Trends zu erkennen und diese bei zukünftigen Prognosen, mit Unterstützung des Einzelhändlers, zu berücksichtigen.

Literaturliste

[NRW98] Nakhaeizadeh, Gholamreza/Reinartz, Thomas/Wirth, Rüdiger: Wissensentdeckung in Datenbanken und Data Mining: Ein Überblick, in: Gholamreza Nakhaeizadeh (Hrsg.): Data Mining: Theoretische Aspekte und Anwendungen, Heidelberg, Physica-Verlag 1998, S. 1-34.

[HEK82] Hartung, Joachim/Elpelt, Bärbel/Klösener, Karl-Heinz: Statistik: Lehr- und Handbuch der angewandten Statistik, München, Oldenbourg Verlag 1982.

[RN 95] Russell, Stuart/Norvig, Peter: Artificial Intelligence: A Modern Approach, New Jersey, Prentice Hall 1995.

[Lea78] Leamer, E.E.: Specification Searches: Ad Hoc Inference with Nonexperimental Data, New York, John Wiley & Sons 1978.

[Qui79] Quinlan, J.R.: Discovering Rules from large Collections of Examples: A case study, in: D. Michie (Hrsg.): Expert Systems in the Microelectronic Age, Edinburgh University Press, Edinburgh, Scotland 1979.

[AIS93] Agrawal, Rakesh/ Imielinski, Tomas/ Swami, Arun: Mining Association rules between Sets of Items in Large Databases, in: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington DC, Mai 26-28, 1993, S. 207-216.

[ZO98] Zaki, Mohammed Javeed/Ogihara, Mitsunori: Theoretical foundations of association rules, in: Proceedings of 3 rd SIGMOD'98 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'98), Seattle, Washington, USA, Juni 1998.

[AS94] Agrawal, Rakesh/Srikant, Ramakrishnan: Fast Algorithms for Mining Association Rules in Large Databases, in: Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, September 1994, S. 487-499.

[HF95a] Han, Jiawei/Fu, Yongjiang: Discovery of multiple-level association rules from large databases, in: Proceedings of the VLDB Conference, September 1995, S. 420-431.

[HF95b] Han, Jiawei/Fu, Yongjiang: Meta-rule-guided mining of association rules in relational databases, in: Proceedings of the 1995 International Workshop on Knowledge Discovery and Deductive and Object-Oriented Databases, Singapore, Dezember 1995.

[AFK97] Amir, A./Feldman, R./Kashi, R.: A New and Versatile Method for Association Generation, in: Principles of Data Mining and Knowledge Discovery, Proceedings of the First European Symposium, (PKDD'97), Trondheim Norwegen 1997, S. 221-231.

[MP43] McCulloch, W.S./Pitts, W.: A logical calculus of the ideas immanent in nervous activity, in: Bulletin of Mathematical Biophysics 5, 1943, S. 115-133.

[Ros57] Rosenblatt, F.: The perceptron: A perceiving and recognizing automaton. Technical Report 85-460-1, Project PARA, Cornell Aeronautical Lab, Itahca, New York, Januar 1957.

[BH69] Bryson, A.E./Ho, Y.-C.: Applied Optical Control, New York Blaisdell, 1969.

[Gur97] Gurney, Kevin: An Introduction to Neural networks, London, UCL Press 1997.

[HK01] Han, J./Kamber M.: Data Mining: Concepts and Techniques, http://www.cs.sfu.ca/~han/DM_Book.html, 5.1.2001.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang

Formel 1

$$S^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$$\frac{\partial S^2}{\partial a} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-1) = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\Leftrightarrow 0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - \sum_{i=1}^{n} bx_i$$

$$\Leftrightarrow 0 = \sum_{i=1}^{n} y_i - an - b \sum_{i=1}^{n} x_i$$

$$\Leftrightarrow an = \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i$$

$$\Leftrightarrow a = \bar{y} - b\bar{x}$$

q.e.d.

Formel 2

$$S^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$$\frac{\partial S^2}{\partial b} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-x_i) = -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i = -2\sum_{i=1}^{n} (y_i - (\bar{y} - b\bar{x}) - bx_i)x_i = 0$$

$$\Leftrightarrow 0 = \left(\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \bar{y}x_i + \sum_{i=1}^{n} b\bar{x}x_i - \sum_{i=1}^{n} bx_i^2 \right)$$

$$\Leftrightarrow 0 = \left(\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \bar{y}x_i \right) + b\left(\sum_{i=1}^{n} \bar{x}x_i - \sum_{i=1}^{n} x_i^2 \right)$$

$$\Leftrightarrow b\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} \bar{x}x_i \right) = \left(\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \bar{y}x_i \right)$$

$$\Leftrightarrow b = \frac{\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \bar{y}x_i}{\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} \bar{x}x_i \right)} = \frac{\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \bar{y}x_i - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\left(\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} \bar{x}x_i + n\bar{x}^2 \right)} =$$

$$\frac{\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \bar{y}x_i - \sum_{i=1}^{n} \bar{x}y_i + \sum_{i=1}^{n} \bar{x}\bar{y}}{\left(\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} \bar{x}x_i + \sum_{i=1}^{n} \bar{x}^2 \right)} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

q.e.d.

Formel 3

$$S^2 = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki})^2$$

$$\frac{\partial S^2}{\partial a} = \sum_{i=1}^{n} 2(y_i - a - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki})(-1) = -2 \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki}) = 0$$

$$\Leftrightarrow 0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - \sum_{i=1}^{n} b_1 x_{1i} - \sum_{i=1}^{n} b_2 x_{2i} - \ldots - \sum_{i=1}^{n} b_k x_{ki}$$

$$\Leftrightarrow 0 = \sum_{i=1}^{n} y_i - na - b_1 \sum_{i=1}^{n} x_{1i} - b_2 \sum_{i=1}^{n} x_{2i} - \ldots - b_k \sum_{i=1}^{n} x_{ki}$$

$$\Leftrightarrow na = \sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_{1i} - b_2 \sum_{i=1}^{n} x_{2i} - \ldots - b_k \sum_{i=1}^{n} x_{ki}$$

$$\Leftrightarrow a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - \ldots - b_k \bar{x}_k$$

q.e.d.

Formel 4

$$S^2 = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki})^2$$

für $j = 1, \ldots, n$

$$\frac{\partial S^2}{\partial b_j} = \sum_{i=1}^{n} 2(y_i - a - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_j x_{ji} - \ldots - b_k x_{ki})(-x_{ji})$$

$$= -2 \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_j x_{ji} - \ldots - b_k x_{ki}) x_{ji} = 0$$

$$\Leftrightarrow 0 = \sum_{i=1}^{n} y_i x_{ji} - \sum_{i=1}^{n} a x_{ji} - \sum_{i=1}^{n} b_1 x_{1i} x_{ji} - \sum_{i=1}^{n} b_2 x_{2i} x_{ji} - \ldots - \sum_{i=1}^{n} b_j x_{ji}^2 - \ldots - \sum_{i=1}^{n} b_k x_{ki} x_{ji}$$

$$\Leftrightarrow 0 = \sum_{i=1}^{n} y_i x_{ji} - a \sum_{i=1}^{n} x_{ji} - b_1 \sum_{i=1}^{n} x_{1i} x_{ji} - b_2 \sum_{i=1}^{n} x_{2i} x_{ji} - \ldots - b_j \sum_{i=1}^{n} x_{ji}^2 - \ldots - b_k \sum_{i=1}^{n} x_{ki} x_{ji}$$

$$\Leftrightarrow 0 = \sum_{i=1}^{n} y_i x_{ji} - (\bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - \ldots - b_j \bar{x}_j - \ldots - b_k \bar{x}_k) \sum_{i=1}^{n} x_{ji}$$

$$- b_1 \sum_{i=1}^{n} x_{1i} x_{ji} - b_2 \sum_{i=1}^{n} x_{2i} x_{ji} - \ldots - b_j \sum_{i=1}^{n} x_{ji}^2 - \ldots - b_k \sum_{i=1}^{n} x_{ki} x_{ji}$$

$$\Leftrightarrow \left(b_1 \sum_{i=1}^{n} x_{1i} x_{ji} - b_1 \bar{x}_1 \sum_{i=1}^{n} x_{ji} \right) + \left(b_2 \sum_{i=1}^{n} x_{2i} x_{ji} - b_2 \bar{x}_2 \sum_{i=1}^{n} x_{ji} \right)$$

$$+ \ldots + \left(b_j \sum_{i=1}^{n} x_{ji}^2 - b_j \bar{x}_j \sum_{i=1}^{n} x_{ji} \right) + \ldots + \left(b_k \sum_{i=1}^{n} x_{ki} x_{ji} - b_k \bar{x}_k \sum_{i=1}^{n} x_{ji} \right) = \sum_{i=1}^{n} y_i x_{ji} - \bar{y} \sum_{i=1}^{n} x_{ji}$$

$$\Leftrightarrow b_1 \left(\sum_{i=1}^{n} x_{1i} x_{ji} - \bar{x}_1 \sum_{i=1}^{n} x_{ji} \right) + b_2 \left(\sum_{i=1}^{n} x_{2i} x_{ji} - \bar{x}_2 \sum_{i=1}^{n} x_{ji} \right)$$

$$+ \ldots + b_j \left(\sum_{i=1}^{n} x_{ji}^2 - \bar{x}_j \sum_{i=1}^{n} x_{ji} \right) + \ldots + b_k \left(\sum_{i=1}^{n} x_{ki} x_{ji} - \bar{x}_k \sum_{i=1}^{n} x_{ji} \right) = \sum_{i=1}^{n} y_i x_{ji} - \bar{y} \sum_{i=1}^{n} x_{ji}$$

$$\Leftrightarrow b_1\left(\sum_{i=1}^{n} x_{1i}x_{ji} - \frac{\left(\sum_{i=1}^{n} x_{1i}\right)\left(\sum_{i=1}^{n} x_{ji}\right)}{n} - \frac{\left(\sum_{i=1}^{n} x_{1i}\right)\left(\sum_{i=1}^{n} x_{ji}\right)}{n} + \frac{\left(\sum_{i=1}^{n} x_{1i}\right)\left(\sum_{i=1}^{n} x_{ji}\right)}{n}\right)$$

$$+ b_2\left(\left(\sum_{i=1}^{n} x_{2i}x_{ji}\right) - \frac{\left(\sum_{i=1}^{n} x_{2i}\right)\left(\sum_{i=1}^{n} x_{ji}\right)}{n} - \frac{\left(\sum_{i=1}^{n} x_{2i}\right)\left(\sum_{i=1}^{n} x_{ji}\right)}{n} + \frac{\left(\sum_{i=1}^{n} x_{2i}\right)\left(\sum_{i=1}^{n} x_{ji}\right)}{n}\right)$$

$$+ \ldots + b_j\left(\sum_{i=1}^{n} x_{ji}^2 - \frac{\left(\sum_{i=1}^{n} x_{ji}\right)^2}{n} - \frac{\left(\sum_{i=1}^{n} x_{ji}\right)^2}{n} + \frac{\left(\sum_{i=1}^{n} x_{ji}\right)^2}{n}\right)$$

$$+ \ldots + b_k\left(\sum_{i=1}^{n} x_{ki}x_{ji} - \frac{\left(\sum_{i=1}^{n} x_{ki}\right)\left(\sum_{i=1}^{n} x_{ji}\right)}{n} - \frac{\left(\sum_{i=1}^{n} x_{ki}\right)\left(\sum_{i=1}^{n} x_{ji}\right)}{n} + \frac{\left(\sum_{i=1}^{n} x_{ki}\right)\left(\sum_{i=1}^{n} x_{ji}\right)}{n}\right)$$

$$= \sum_{i=1}^{n} y_i x_{ji} - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_{ji}\right)}{n} - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_{ji}\right)}{n} + \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_{ji}\right)}{n}$$

$$\Leftrightarrow b_1\left(\sum_{i=1}^{n} x_{1i}x_{ji} - \bar{x}_1\left(\sum_{i=1}^{n} x_{ji}\right) - \left(\sum_{i=1}^{n} x_{1i}\right)\bar{x}_j + n\bar{x}_1\bar{x}_j\right) + b_2\left(\sum_{i=1}^{n} x_{2i}x_{ji} - \bar{x}_2\left(\sum_{i=1}^{n} x_{ji}\right) - \left(\sum_{i=1}^{n} x_{2i}\right)\bar{x}_j + n\bar{x}_2\bar{x}_j\right)$$

$$+ \ldots + b_j\left(\sum_{i=1}^{n} x_{ji}^2 - 2\bar{x}_j\left(\sum_{i=1}^{n} x_{ji}\right) + n\bar{x}_j^2\right) + \ldots + b_k\left(\sum_{i=1}^{n} x_{ki}x_{ji} - \bar{x}_k\left(\sum_{i=1}^{n} x_{ji}\right) - \left(\sum_{i=1}^{n} x_{ki}\right)\bar{x}_j + n\bar{x}_k\bar{x}_j\right)$$

$$= \sum_{i=1}^{n} y_i x_{ji} - \bar{y}\left(\sum_{i=1}^{n} x_{ji}\right) - \left(\sum_{i=1}^{n} y_i\right)\bar{x}_j + n\bar{y}\bar{x}_j$$

$$\Leftrightarrow b_1\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(x_{ji}-\bar{x}_j) + b_2\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)(x_{ji}-\bar{x}_j) + \ldots + b_j\sum_{i=1}^{n}(x_{ji}-\bar{x}_j)^2 + \ldots + b_k\sum_{i=1}^{n}(x_{ki}-\bar{x}_k)(x_{ji}-\bar{x}_j)$$

$$= \sum_{i=1}^{n}(y_i-\bar{y})(x_{ij}-\bar{x}_j)$$

$$\Leftrightarrow b_1 SP_{x_1x_j} + b_2 SP_{x_2x_j} + \ldots + b_j SQ_{x_j} + \ldots + b_k SP_{x_kx_j} = SP_{yx_j}$$

q.e.d.

Formel 5

$$X^T X b = X^T \tilde{Y}$$

$$\Leftrightarrow \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{k1} & x_{k2} & x_{k3} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ 1 & x_{13} & x_{23} & \dots & x_{k3} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ \dots \\ b_k \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{k1} & x_{k2} & x_{k3} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \dots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \dots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{2i}^2 & \dots & \sum_{i=1}^{n} x_{2i}x_{ki} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{1i}x_{ki} & \sum_{i=1}^{n} x_{2i}x_{ki} & \dots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ \dots \\ b_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_{1i} \\ \sum_{i=1}^{n} y_i x_{2i} \\ \dots \\ \sum_{i=1}^{n} y_i x_{ki} \end{bmatrix}$$

$$\Leftrightarrow an + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i} + \dots + b_k \sum_{i=1}^{n} x_{ki} = \sum_{i=1}^{n} y_i$$

$$a \sum_{i=1}^{n} x_{ji} + b_1 \sum_{i=1}^{n} x_{1i}x_{ji} + b_2 \sum_{i=1}^{n} x_{2i}x_{ji} + \dots + b_j \sum_{i=1}^{n} x_{ji}^2 + \dots + b_k \sum_{i=1}^{n} x_{ki}x_{ji} = \sum_{i=1}^{n} y_i x_{ji}$$

für $j = 1 \dots k$

q.e.d.

Formale Problemdefinition von Assoziationsregeln

Es sei A die Menge aller Waren $\{x_1,x_2,\ldots,x_n\}$.

Es sei T die Menge aller Transaktionen
$T = \{t_1,t_2,\ldots,t_m\}$ mit $t_i = \{x_{i1},x_{i2},\ldots,x_{in_i}\}$ Teilmengen von A für $i = 1\ldots m$.

Eine Warenmenge X ist eine Teilmenge von A $X \subseteq A$.

Der Support einer Warenmenge X ist $Support(X) = \#\langle t_i | X \subseteq t_i \rangle$.

Eine Warenmenge X ist frequent, wenn $Support(X) \geq minSupport$.

L_k ist die Menge aller Mengen der Größe k, die frequent sind.

$A \rightarrow B$ ist eine Assoziationsregel, wobei A und B Warenmengen sind.

$$Support(A \rightarrow B) = Support(A \cup B)$$

$$Confidence(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A)}$$

Formale Definition eines neuronalen Netzes

Menge von Neuronenschichten: $N = N_0 \cup N_1 \cup \ldots \cup N_n$ mit $n \geq 1$.

Es gilt: $\langle \forall i \in \{0, \ldots, n\} | N_i \neq \emptyset \rangle$ und $\langle \forall i,j \in \{0, \ldots, n\}, i \neq j | N_i \cap N_j = \emptyset \rangle$.

N_0 ist die Eingabe-, N_n die Ausgabeschicht, $N_1 \ldots N_{n-1}$ sind versteckte Schichten.

Menge von Schwellwerten B, mit $\omega_i \in B \subset R$, die jedem Neuron $n_i \in N_k, k > 0$ einen Schwellwert zuordnet.

Eine Abbildung $W : N \quad N \rightarrow R$, welche die Verbindungen zweier Neuronen gewichtet.

Verbindungen zweier Schichten existieren nur zwischen einer Schicht N_i und der nächsten Schicht N_{i+1}.

Menge von Aktivierungsfunktionen A, mit $\Upsilon_i : R \rightarrow R$, die jedem Neuron $n_i \in N$ eine Funktion zuordnet.

Menge von Propagierungsfunktionen NET, die jedem $n_i \in N$ eine Netzeingabefunktion NET_i zuordnet.

Formel 6

$$\frac{\partial E}{\partial \omega_{ji}} = \frac{\partial E}{\partial NET_i}\frac{\partial NET_i}{\partial \omega_{ji}} = \frac{\partial E}{\partial NET_i}\frac{\partial\left(\sum_{n_j} \omega_{ij} o_j + \omega_i\right)}{\partial \omega_{ji}} = \frac{\partial E}{\partial NET_i} o_j = -\delta_i o_j$$

mit $\delta_i = -\frac{\partial E}{\partial NET_i} = -\frac{\partial E}{\partial o_i}\frac{\partial o_i}{\partial NET_i} = -\frac{\partial E}{\partial o_i}\frac{\partial \Upsilon_i(NET_i)}{\partial NET_i} = -\frac{\partial E}{\partial o_i}\Upsilon_i'(NET_i)$

$\frac{\partial E}{\partial o_i} = \frac{\partial \frac{1}{2}\sum_{n_j \in N_n}(t_j - o_j)^2}{\partial o_i} = -(t_i - o_i)$, falls $n_i \in N_n$ also ein Ausgabeneuron ist.

$$\frac{\partial E}{\partial o_i} = \sum_{n_k \in N_h}\frac{\partial E}{\partial NET_k}\frac{\partial NET_k}{\partial o_i} = \sum_{n_k \in N_h} -\delta_k \frac{\partial NET_k}{\partial o_i} = \sum_{n_k \in N_h} -\delta_k \frac{\partial\left(\sum_{n_m} \omega_{km} o_m + \omega_k\right)}{\partial o_i} = \sum_{n_k \in N_h} -\delta_k \omega_{ki}$$

, mit $n_i \in N_l$ und $n_j \in N_h$ und $h > l$ n_i ist also kein Ausgabeneuron

(Beweis siehe nächste Seite)

q.e.d.

Beweis durch Induktion von

$$\frac{\partial E}{\partial o_i} = \sum_{n_j \in N_h}\frac{\partial E}{\partial NET_j}\frac{\partial NET_j}{\partial o_i}, \text{ mit } n_i \in N_k \text{ und } h > k \text{ für } k = 0 \ldots n-1$$

Induktionsanfang für $k = n - 1$

$$\frac{\partial\left(\frac{1}{2}\sum_{n_j \in N_n}(t_j - o_j)^2\right)}{\partial o_i} = \frac{\partial\left(\sum_{n_j \in N_n}\frac{1}{2}(t_j - o_j)^2\right)}{\partial o_i} = \sum_{n_j \in N_n}\frac{\partial\left(\frac{1}{2}(t_j - o_j)^2\right)}{\partial o_i}$$

$$= \sum_{n_j \in N_n}\frac{\partial\left(\frac{1}{2}(t_j - \gamma_j(NET_j))^2\right)}{\partial o_i} = \sum_{n_j \in N_n}\frac{\partial\left(\frac{1}{2}(t_j - \gamma_j(NET_j))^2\right)}{\partial NET_j}\frac{\partial NET_j}{\partial o_i} = \sum_{n_j \in N_n}\frac{\partial E}{\partial NET_j}\frac{\partial NET_j}{\partial o_i}$$

Induktionsvoraussetzung

$$\frac{\partial E}{\partial o_i} = \sum_{n_j \in N_h} \frac{\partial E}{\partial NET_j} \frac{\partial NET_j}{\partial o_i}, \text{ mit } n_i \in N_k \text{ und } h > k \text{ für } h = x \ldots n$$

Induktionsschluss zu zeigen

$$\frac{\partial E}{\partial o_i} = \sum_{n_j \in N_h} \frac{\partial E}{\partial NET_j} \frac{\partial NET_j}{\partial o_i}, \text{ mit } n_i \in N_k \text{ und } h > k \text{ für } h = x{-}1 \ldots n$$

$$\frac{\partial E}{\partial o_i} = \sum_{n_j \in N_{h+1}} \frac{\partial E}{\partial NET_j} \frac{\partial NET_j}{\partial o_i} = \sum_{n_j \in N_{h+1}} \frac{\partial E}{\partial NET_j} \frac{\partial \left(\sum_{n_k \in N_h} \omega_{jk} o_k + \omega_j \right)}{\partial o_i}$$

$$= \sum_{n_j \in N_{h+1}} \frac{\partial E}{\partial NET_j} \left(\sum_{n_k \in N_h} \frac{\partial (\omega_{jk} o_k)}{\partial o_i} \right) = \sum_{n_j \in N_{h+1}} \frac{\partial E}{\partial NET_j} \left(\sum_{n_k \in N_h} \frac{\partial (\omega_{jk} o_k)}{\partial NET_k} \frac{\partial NET_k}{\partial o_i} \right)$$

$$= \sum_{n_j \in N_{h+1}} \frac{\partial E}{\partial NET_j} \left(\sum_{n_k \in N_h} \frac{\partial NET_j}{\partial NET_k} \frac{\partial NET_k}{\partial o_i} \right) = \sum_{n_j \in N_{h+1}} \sum_{n_k \in N_h} \frac{\partial E}{\partial NET_j} \frac{\partial NET_j}{\partial NET_k} \frac{\partial NET_k}{\partial o_i}$$

$$= \sum_{n_k \in N_h} \frac{\partial NET_k}{\partial o_i} \sum_{n_j \in N_{h+1}} \frac{\partial E}{\partial NET_j} \frac{\partial NET_j}{\partial NET_k} = \sum_{n_k \in N_h} \frac{\partial NET_k}{\partial o_i} \sum_{n_j \in N_{h+1}} \frac{\partial E}{\partial NET_j} \frac{\partial NET_j}{\partial o_k} \frac{\partial o_k}{\partial NET_k}$$

$$= \sum_{n_k \in N_h} \frac{\partial NET_k}{\partial o_i} \frac{\partial o_k}{\partial NET_k} \sum_{n_j \in N_{h+1}} \frac{\partial E}{\partial NET_j} \frac{\partial NET_j}{\partial o_k} = \sum_{n_k \in N_h} \frac{\partial NET_k}{\partial o_i} \frac{\partial o_k}{\partial NET_k} \frac{\partial E}{\partial o_k} = \sum_{n_k \in N_h} \frac{\partial E}{\partial NET_k} \frac{\partial NET_k}{\partial o_i}$$

Normalverteilungstests

Tests auf Normalverteilung

		Kolmogorov-Smirnov[a]		
	WOCHENTA	Statistik	df	Signifikanz
L1UMSPRO	2,00	,059	118	,200*
	3,00	,050	117	,200*
	4,00	,064	115	,200*
	5,00	,088	122	,022
	6,00	,056	122	,200*
	7,00	,039	118	,200*

*. Dies ist eine untere Grenze der echten Signifikanz.

a. Signifikanzkorrektur nach Lilliefors

Abbildung A.1: Normalverteilungstest der Wochentage

Montag

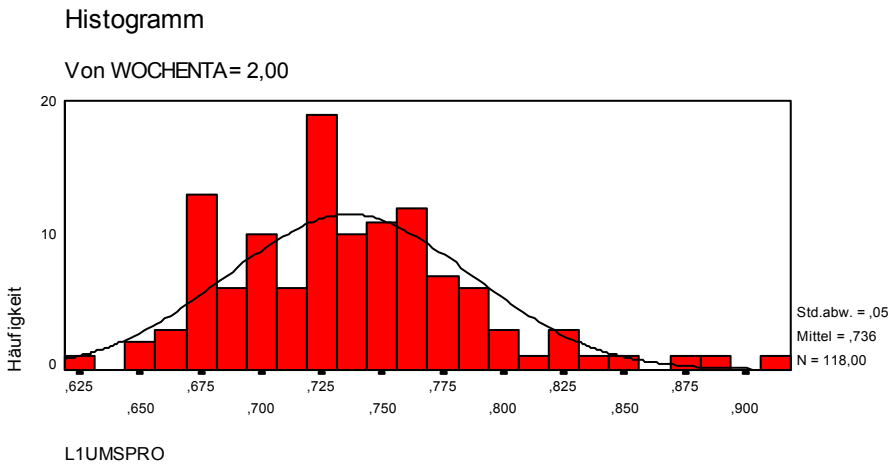

Abbildung A.2: Normalverteilungstest Montag

Dienstag

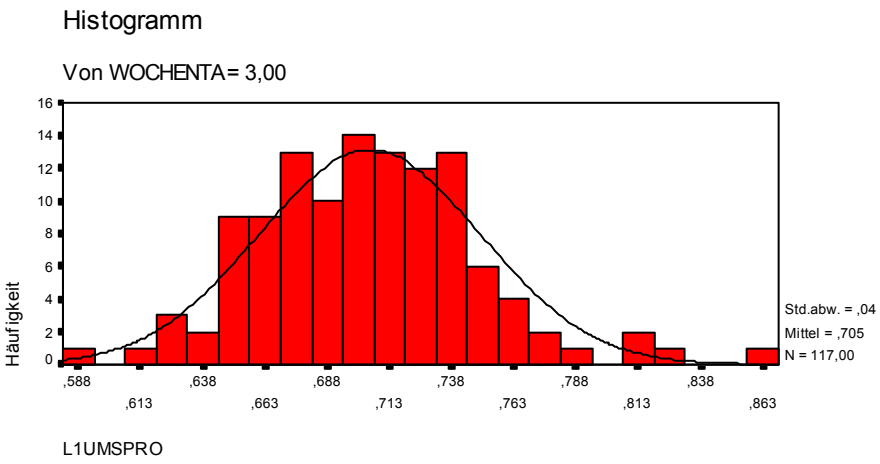

Abbildung A.3: Normalverteilungstest Dienstag

Mittwoch

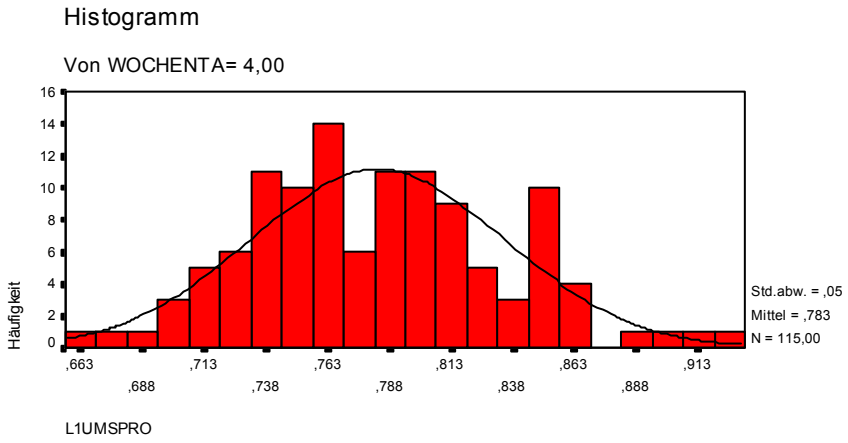

Abbildung A.4: Normalverteilungstest Mittwoch

Donnerstag

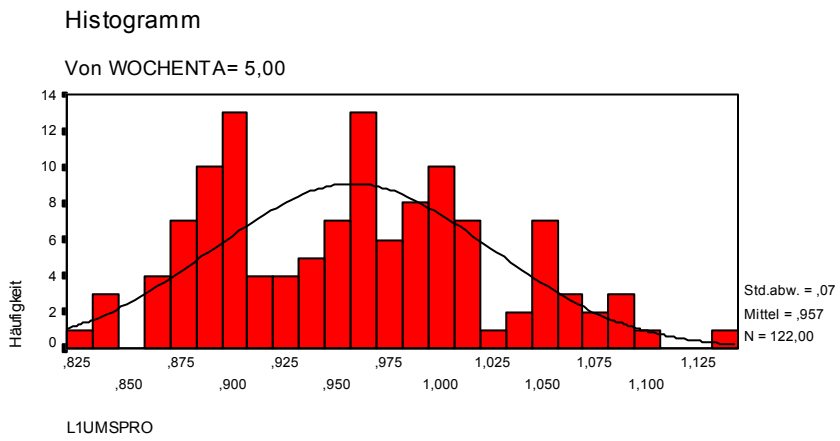

Abbildung A.5: Normalverteilungstest Donnerstag

Freitag

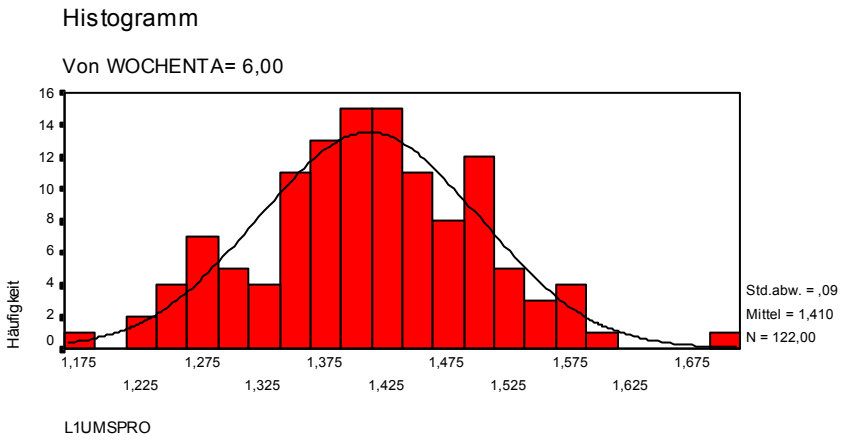

Abbildung A.6: Normalverteilungstest Freitag

Samstag

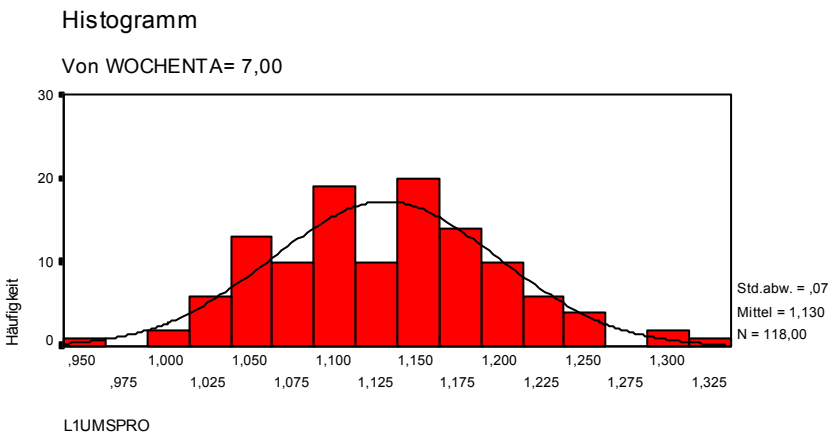

Abbildung A.7: Normalverteilungstest Samstag

Boxplots

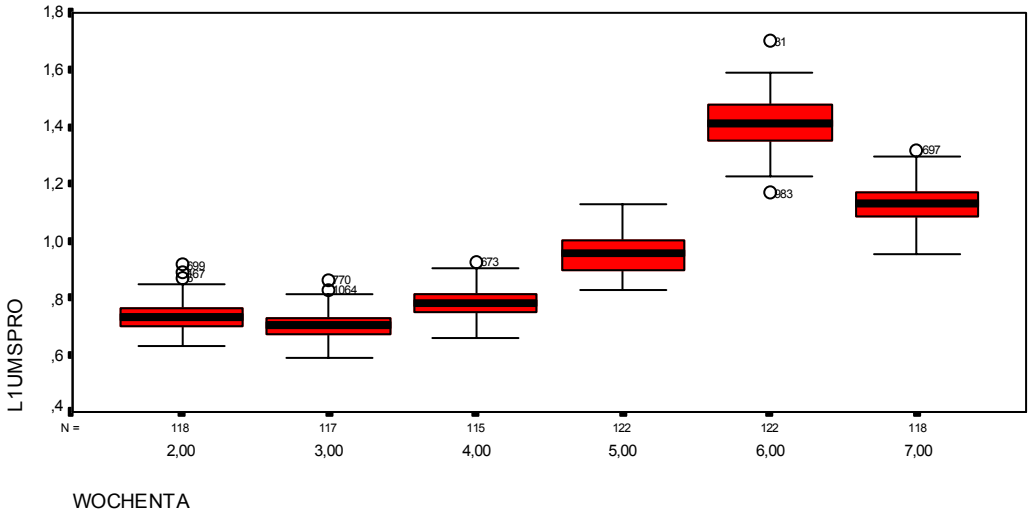

Abbildung A.8: Boxplot der Wochentagsverteilungen

Assoziationsregeln – Ein Betrieb

Regel	*Support*	*Confidence*
-> U1	61,01	61,01

Tabelle A.1: Assoziationsregeln – Ein Betrieb

Regel	*Support*	*Confidence*
-> U2	21,49	21,49
-> U3	15,94	15,94
Mo -> U1	17,68	100,00
Di -> U1	17,85	100,00
Mi -> U1	14,90	89,58
Do -> U1	10,40	63,83
VF4 -> U1	1,56	56,25
VF3 -> U1	1,56	60,00
VF2 -> U1	1,21	41,18
NF1 -> U1	1,21	53,85
Do -> U2	5,20	31,91
Sa -> U2	14,73	93,41
VF4 -> U2	1,21	43,75
VF3 -> U2	0,69	26,67
VF2 -> U2	0,87	29,41
Fr -> U3	13,86	87,91
VF1 -> U3	1,56	52,94
NF1 -> U3	0,87	38,46
VF1 -> U4	1,39	47,06
Mo,VF3 -> U1	0,69	100,00
Di,VF2 -> U1	1,04	100,00
Di,NF1 -> U1	0,87	100,00
Sa,VF4 -> U2	1,21	100,00
Mi,VF1 -> U3	0,87	83,33
Fr,NF1 -> U3	0,69	100,00
Sa,VF1 -> U3	0,69	100,00
Do,VF1 -> U4	0,69	100,00

Tabelle A.1: Assoziationsregeln – Ein Betrieb (Forts.)

Darstellung der zwanzig größten Über- und Unterbewertungen des Regressionsmodells

DATUM	UMSAPRO	REGRESS	DIFF
12/24/1999	,898765664123523	2,06	-1,161234336
12/31/1999	1,01028839867971	2,06	-1,049711601
12/24/1998	,989177413177683	1,62	-0,630822587
12/31/1998	1,05453488591572	1,62	-0,565465114
12/24/1997	1,01710617888535	1,43	-0,412893821
1/2/1998	1,11421383024584	1,52	-0,40578617
10/30/1999	1,36393779468381	1,76	-0,396062205
12/31/1997	1,05347075817189	1,43	-0,376529242
5/22/1999	1,42241231815125	1,76	-0,337587682
6/1/1999	,850941821187826	1,12	-0,269058179
5/17/1997	1,49607339964411	1,76	-0,2639266
10/1/1997	,962707020511393	1,22	-0,257292979
4/29/1997	,871322805907867	1,12	-0,248677194
10/29/1999	1,60522968567939	1,85	-0,244770314
9/10/1999	1,21714646517436	1,46	-0,242853535
5/30/1998	1,52279280376398	1,76	-0,237207196
6/9/1998	,8863031566996	1,12	-0,233696843
5/27/1997	,91075712541367	1,12	-0,209242875
12/27/1997	1,06278416339103	1,27	-0,207215837
5/11/1999	,91339982135547	1,12	-0,206600179

Tabelle A.2: Darstellung der 20 größten Überbewertungen des Regressionsmodells

DATUM	UMSAPRO	REGRESS	DIFF
12/5/1997	1,76155233144211	1,46	0,301552331
1/31/1997	1,7665749533832	1,46	0,306574953
12/5/1998	1,48055094443936	1,16	0,320550944
12/20/1997	1,53726774367264	1,21	0,327267744
4/30/1998	1,95868433883783	1,62	0,338684339
5/20/1998	1,77212173942839	1,43	0,342121739
12/18/1997	1,36373608615183	1,02	0,343736086
12/19/1998	1,56357397034322	1,16	0,40357397
12/20/1999	1,19988277086962	0,76	0,439882771
12/30/1999	1,90405031110573	1,41	0,494050311
12/21/1999	1,27739119722076	0,78	0,497391197
12/21/1998	1,34485152837276	0,81	0,534851528
12/30/1997	1,76442782533481	1,12	0,644427825
12/30/1998	1,98727203389091	1,22	0,767272034
12/23/1999	2,21150461031564	1,41	0,80150461
12/22/1999	1,81167412168976	0,96	0,851674122
12/22/1998	1,75159886042569	0,86	0,89159886
12/22/1997	1,82164598641711	0,89	0,931645986
12/23/1998	2,43805254324905	1,22	1,218052543
12/23/1997	2,42240704306156	1,12	1,302407043

Tabelle A.3: Darstellung der 20 größten Unterbewertungen der Regression

Assoziationsregeln – Alle Betriebe

Regeln	*Support*	*Confidence*
-> U1	65,86	65,86
-> U2	20,18	20,18
-> U3	12,49	12,49
Mo -> U1	15,09	95,86
Di -> U1	17,10	96,98
Mi -> U1	16,09	92,83
Do -> U1	12,43	77,78
Sa -> U1	5,15	31,07
VF4 -> U1	1,48	71,43
VF3 -> U1	1,60	67,50
VF2 -> U1	1,01	40,48
NF1 -> U1	1,12	40,43
Mo -> U2	0,65	4,14
Di -> U2	0,47	2,68
Do -> U2	2,78	17,41

Tabelle A.4: Assoziationsregeln – Alle Betriebe

Regeln	*Support*	*Confidence*
Fr -> U2	5,44	32,51
Sa -> U2	10,53	63,57
VF4 -> U2	0,59	28,57
VF3 -> U2	0,59	25,00
VF2 -> U2	0,53	21,43
NF1 -> U2	0,65	23,40
Mi -> U3	0,89	5,12
Fr -> U3	10,65	63,60
Sa -> U3	0,77	4,64
VF2 -> U3	0,30	11,90
VF1 -> U3	1,36	52,27
NF1 -> U3	0,95	34,04
Do -> U4	0,65	4,07
Fr -> U4	0,65	3,89
VF2 -> U4	0,65	26,19
VF1 -> U4	1,12	43,18
Mo,VF4 -> U1	0,41	87,50
Mo,VF3 -> U1	0,71	85,71
Mo,NF1 -> U1	0,24	80,00
Di,VF4 -> U1	0,65	91,67
Di,VF3 -> U1	0,30	83,33
Di,VF2 -> U1	0,89	93,75
Di,NF1 -> U1	0,71	92,31
Mi,VF4 -> U1	0,30	100,00
Mi,VF3 -> U1	0,41	53,85
Mi,VF2 -> U2	0,24	44,44
Do,VF3 -> U2	0,24	57,14
Do,VF2 -> U2	0,24	36,36
Fr,NF1 -> U2	0,36	37,50
Sa,VF4 -> U2	0,47	80,00
Mi,VF1 -> U3	0,71	92,31
Fr,NF1 -> U3	0,59	62,50
Sa,VF1 -> U3	0,53	75,00
Sa,NF1 -> U3	0,36	50,00
VF1,NF1 -> U3	0,30	83,33
Do,VF2 -> U4	0,41	63,64
Do,VF1 -> U4	0,65	84,62
Fr,VF2 -> U4	0,24	66,67

Tabelle A.4: Assoziationsregeln – Alle Betriebe (Forts.)

Regeln	*Support*	*Confidence*
Fr,VF1 -> U4	0,36	100,00
VF2,VF1 -> U4	0,41	100,00
Sa,VF1,NF1 -> U3	0,30	83,33
Do,VF2,VF1 -> U4	0,41	100,00
Di,VF4 -> U1	0,65	91,67
Di,VF3 -> U1	0,30	83,33
Di,VF2 -> U1	0,89	93,75
Di,NF1 -> U1	0,71	92,31
Mi,VF4 -> U1	0,30	100,00
Mi,VF3 -> U1	0,41	53,85
Mi,VF2 -> U2	0,24	44,44
Do,VF3 -> U2	0,24	57,14
Do,VF2 -> U2	0,24	36,36
Fr,NF1 -> U2	0,36	37,50
Sa,VF4 -> U2	0,47	80,00

Tabelle A.4: Assoziationsregeln – Alle Betriebe (Forts.)