Thema:

Anwendung von Data-Miningmethoden in der Tourismusbranche: Analyse und Verfahren

Bachelorarbeit

Arbeitsgruppe Wirtschaftsinformatik

Themensteller: Dr.-Ing. Gamal Kassem
Betreuer: Dr.-Ing. Gamal Kassem

vorgelegt von: Marvin Bürkner

Abgabetermin: 8. März 2010

Eidesstattliche Erklärung

Hiermit versichere ich, die vorliegende Arbeit selbstständig und unter ausschließlicher Verwendung der angegebenen Literatur und Hilfsmittel erstellt zu haben.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Hamburg, den 28. Februar 2010

_____

Marvin Bürkner

„Wer nichts weiß, muss alles glauben."

Marie von Ebner-Eschenbach

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Data Ming umfasst den Prozess der Gewinnung neuer, valider und handlungsrelevanter Informationen aus großen Datenbeständen und die Nutzung dieser Informationen für betriebswirtschaftliche Entscheidungen. (Cabena, et al., 1998)
Um sich den Begriff „Data Ming“ bildhaft darzustellen kann man eine Analogie zum Bergbau(engl. Mining) herstellen. Im Bergbau wird mit Hilfe von großen technologischen Aufwand enorme Massen zu tage getragen und aufbereitet, um kostbare Rohstoffe für neue Produkte zu fördern. Analog dazu werden beim Data Ming große Datenbestände nach neuen, gesicherten und handlungsrelevanten Informationen durchsucht. Data Ming Verfahren sind dabei sehr rechenintensiv und werden auf immer größere Datenmengen angewendet (Adriaans P., 1997)
Data Mining Methoden sind der Oberbegriff einer Reihe von Ansätzen aus Statistik, Künstlicher Intelligenz, Maschinellem Lernen und Mustererkennung mit dem Ziel der autonomen Identifizierung von bedeutsamen und aussagekräftigen Mustern in großen Datenmengen. Die Ergebnisse des Data Mining sollen dem Anwender als interessantes Wissen präsentiert werden ohne vom Anwender a priori-Hypothesen und damit Aussagen über die gesuchten Inhalte zu fordern. Somit grenzt sich Data Mining deutlich von gängigen Abfragesprachen wie SQL ab, da der Anwender bei diesen Methoden vorher wissen muss nach welchen Informationen er sucht. "Data Ming ... lots uns zu nützlichen Antworten, bevor uns die passenden Fragen einfallen und fördert aus den Tiefendes Datenmeers Überraschendes zutage." (Janetzko ., 1997) Doch auch wenn heutige Data Mining Tools sehr mächtig Werkzeuge sind und eine Vielzahl von Algorithmen zu ihrem Funktionsumfang zählen, gibt es kein vollständig autonomes Data Mining Tool das in einem beliebigen Datenbestand Auffälligkeiten findet. Zu wichtig ist das domänenspezifische Fachwissen der Anwender, ohne das das Aufdecken nützlicher Erkenntnisse kaum möglich wäre.

Durch diese Tatsache ist die Anwendung von Data Mining-Methoden zur Gewinnung von Wissen ein umfassender Prozess bei dem viele Entscheidungen, hinsichtlich der Algorithmen Auswahl, der Parametrisierung, die Auswahl der geeigneten Datenquellen und deren Merkmalen, Validierung der Ergebnisse und Implementierung getroffen werden müssen.

1.1 Ziel

Ziel der Arbeit besteht darin sich mit den Verschieden Data Ming Methoden und den zugehörigen Algorithmen zu beschäftigen. Dabei werden die Verfahren der Klassifikations-, Cluster. und Assoziationsanalyse betrachtet. Zu jedem Verfahren werden die gängigen Algorithmen die angewendet werden erläutert. Diese Verfahren sollen dann verwendet werden um gegebene Problemstellung aus der Touristikbranche mit Hilfe von Data Ming Anwendungen zu lösen.

1.2 Knowledge Discovery in Databases

Oftmals werden die Begriffe „Data Ming“ und „Knowledge Discovery in Databases”(KDD) synonym verwendet. Da sich die beiden Begriff jedoch durchaus unterscheiden, soll hier eine Abgrenzung vorgenommen werden.
KDD unterteilt sich in eine Reihe von Prozessen die das Ziel haben Wissen aus Datenbanken zu extrahieren, das gültig (im statistischen Sinne), bisher unbekannt und für den Anwender nützlich sind (Ester, et al., 2000). Data Mining selbst ist dabei nur eine Phase des KDD Prozesses, wie im folgenden Abschnitt gezeigt wird.

1.2.1 Phasen des Knowledge Discovery in Databases

Die Phasen des Knowledge Discovery in Databases werden oft auch als Data Mining Prozess bezeichnet.
Im Folgenden werden diese vorgestellt (Ester, et al., 2000):

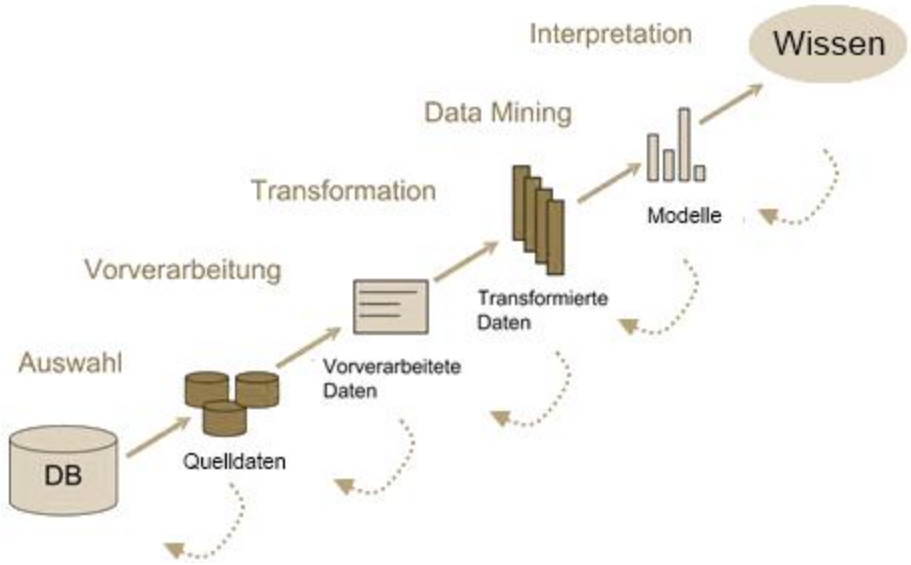


Abbildung 1: Prozess des Knowledge Discovery in Databases

• Auswahl: Im ersten Schritt wird, ausgehend von der betriebswirtschaftlichen Problemstellung und die analytischen Ziele des Knowledge Discovery Prozesses, die Datenbasis eingeschränkt.

• Vorverarbeitung: Danach werden die ausgewählten Daten vorbereitet. Hierbei geht es darum, die Daten konsistent zu machen, zu vervollständigen, fehlender zu behandeln. Insgesamt soll die Datenqualität erhöht werden.

• Transformation: Ziel der Transformation ist es, die Attribute in eine für das Data Mining bearbeitbare Form zu bringen. Eine Möglichkeit hierfür ist die Kodierung der Attribute.

• Data Mining: Nun erst erfolgt der eigentliche Schritt des Data Mining. In diesem Schritt werden die Algorithmen angewendet, die Muster und Auffälligkeiten in der Datenbasis finden sollen. Diese Verfahren werden im Rahmen der Bachelorarbeit vorgestellt.

• Interpretation: Im letzten Schritt des Knowledge Discovery Prozesses müssen die vom Data Mining gefundenen Modelle untersucht werden. Das geschieht zum einen durch Anwendung der Modelle auf Testdaten und der Bewertung der Ergebnisse, zum anderen durch die Untersuchung von Experten der Fachabteilung. Sind diese nicht zufriedenstellend, so kann ein weiterer Durchlauf von Teilprozessen notwendig sein.

1.3 Allgemeine Probleme beim Data Mining

1.3.1 Datenqualität

Data Mining darf nicht als eine isolierte Aufgabe – die einmalige Analyse eines gegebenen Datensatzes – angesehen werden. Wie im oberen Abschnitt beschrieben ist es nur ein Teilprozess, der in seiner Gesamtheit betrachtet werden muss, um nützliche Ergebnisse zu liefern. Eine wesentliche Voraussetzung für die effiziente Nutzung von Mining Algorithmen insbesondere in der Wirtschaft ist deren Integration mit Data Warehouse Lösungen. Data Warehousing entspricht im Kontext des Data Mining dem Prozess der Stichproben- bzw. Erhebungsplanung in der klassischen Statistik, da unternehmensweite Daten gesammelt und aufbereitet werden.
Einen wesentlichen Ansatzpunkt hierfür bieten die Konzepte der Metadaten-Modellierung, welche es erlauben Informationen über Inhalte und semantische Bedeutung der Daten eines Data Warehouse in Datenbankanwendungen zu integrieren und für Data Mining Zwecke verfügbar zu machen. Ein weiteres Merkmal für eine gute Data Warehouse - Lösung ist, dass sie integrierte, fehlerbereinigte Daten auf unterschiedlichen Aggregationsniveaus in Form einer historischen Datenbank zur Verfügung stellt (Hudec, 2003).
In der Realität jedoch, werden Data Mining Algorithmen häufig unmittelbar auf Datenbanken von operativen Systemen angewendet, welche ungeprüfte und keineswegs aggregierte Daten enthalten. Es ist kaum davon auszugehen, dass die unkritische Anwendung komplexer

Algorithmen auf schlecht strukturierte Datenbanken zu wahren und nützlichen Ergebnissen führen. Wodurch frühzeitig das Potenzial von Data Ming Verfahren unterschätzt wird.
Typische Schwachstellen der in Datenbanken vorzufindenden Daten, welche die Anwendbarkeit von Data Mining Konzepten einschränken und auf die im Sinne einer Qualitätssicherung hinzuweisen ist, sind:

- Mangelnde Repräsentativität (Fehlen relevanter historischer Daten)
- Systematisches Fehlen von Datenbeständen die für das operative Geschäft nicht von Bedeutung sind, für Data Mining dafür umso mehr(z.B.: die dauerhafte Speicherung von abgelehnten Krediten)
- Mangelnde Charakterisierung durch Attribute (Fehlen wichtiger Variablen)
- Komplexe Korrelationsstrukturen bedingt durch fehlende Versuchsplanung
- Laufende Veränderung von Strukturen und Mustern in den datengenerierenden Prozessen

1.3.2 Aussagekraft der Ergebnisse

Data Mining ist im Wesentlichen ein exploratives Vorgehensmodell, bei dem Muster, Strukturen, Klassifikationsregeln und Hypothesen bzw. Erklärungsmodelle auf semi-automatische Weise direkt aus Daten abgeleitet werden. In diesem Sinne handelt es sich also um eine hypothesengenerierende Vorgangsweise, deren Ergebnisse in Bezug auf die Anwendbarkeit auf andere Daten nur mit größter Vorsicht interpretiert werden dürfen. Bei Data Mining geht es in der Regel nicht darum, „wahre Gesetzmäßigkeiten“ über den datengenerierenden Prozess aufzuzeigen. Im Vordergrund steht, ob die Ergebnisse für den beabsichtigten Zweck brauchbar bzw. praxistauglich sind.
Ein weiteres Problem kann im Überschätzen der Fähigkeit des Algorithmus liegen. Es besteht zweifellos die Gefahr, dass der Anwender jegliche Beziehung zu den Daten und ihrer Zusammenhängen verliert. Der komplexe Algorithmus wird für den Anwender zur undurchschaubaren Black-Box, die ihm von den Daten trennt. Das Überprüfen von Modellannahmen entfällt, und das vom Algorithmus generierte Ergebnis, welches oft nur eines von vielen möglichen Ergebnissen eines Datensatzes darstellt, wird fälschlicherweise als Wahr angesehen.
Häufig erlauben empirische Daten keine eindeutige Untersuchung zwischen den in Bezug auf die den Algorithmus steuernden Parameter bei nahezu gleichwertigen Modellen, welche jedoch unterschiedliche Ergebnisse liefern. Sensibilitätsanalyse und Visualisierungstechniken werden zum unverzichtbaren Mittel, will man das Auffinden von Artefakten vermeiden.

1.3.3 Privacy und Datenschutz

Mit der wachsenden Verfügbarkeit von elektronisch gespeicherten Daten eröffnen sich nicht nur immer mehr vielversprechende Anwendungsmöglichkeiten des Data Mining, gleichzeitig wächst auch die Gefahr einer missbräuchlichen Verwendung von Daten.
Ein ethisch verantwortungsvoller Umgang ist hier, ebenso wie die Entwicklung von softwaretechnischen Schranken zur Vermeidung unerwünschter Datenverknüpfungen, auf der Ebene von Individualdaten gefordert.
Hieraus ergibt sich die Notwendigkeit gesetzlichen Rahmenbedingungen zur schaffen, die die Verwendung und Analyse von unternehmensweit gesammeltem Daten, regelt.

2 Data Ming-Verfahren

Um Data Ming erfolgreich anzuwenden, ist es erforderlich, aus der Menge der verschieden Data Ming Methoden die herauszufiltern, die zur Bearbeitung eines vorliegenden Problemtyps am geeignetsten ist. Dabei lassen sich drei grundlegende Gruppen zur Extraktion von Informationen aus Datenbeständen darstellen. Dies sind die Assoziations-, die Klassifikations- und Clusterverfahren.

Assoziationsverfahren: Diese finden Korrelationen zwischen verschiedenen Merkmalen einer Datenmenge, also der Art und Weise einer Verbindung zwischen unterschiedlichen Merkmalen.

Klassifikationsverfahren: Dienen zur Vorhersage einer oder mehrere diskreter Merkmale, basierend auf anderen Merkmalen der Inputdatenmenge.

Clusterverfahren: Teilen Datenmengen in Gruppen oder Cluster von Objekten, die ähnliche Eigenschaften haben.

Die Abbildung 2 zeigt die Unterteilung der Data Mining Verfahren und ordnet die bekanntesten Methoden den jeweiligen Verfahren zu, die in dieser Arbeit genauer betrachtet werden.

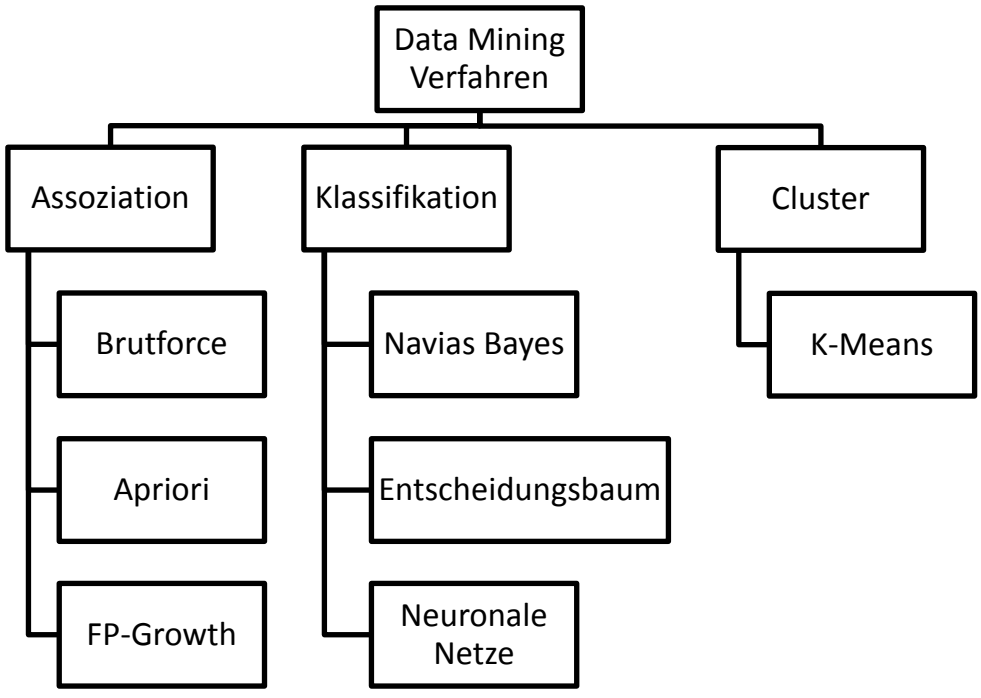


Abbildung 2: Die verschiedene Data Mining Verfahren und die zugehörigen Algorithmen

2.1 Assoziationsregeln

Assoziationsregeln versuchen aus einem Datenbestand Regeln zu extrahieren. Dabei wird nach Korrelationen zwischen verschiedenen Merkmalen einer Datenmenge gesucht, also die Art und Weise der Verbindung zwischen unterschiedlichen Merkmalen (Aggarwal, 1999).

Das bekannteste Anwendungsgebiet ist dabei die Warenkorbanalyse, die versucht heraus zu finden, welche Produkte Kunden gemeinsam einkaufen. Das klassische Beispiel ist die Wechselbeziehung zwischen gekauftem Bier und gekauften Windeln, das bei einer Analyse von Daten einer Supermarktkette entdeckt wurde.

Mithilfe der Kassen-Abrechnungen wurden Listen von Produkten erstellt, die sich bei jeweils einem Kunden im Einkaufswagen befanden. Danach stellte man den Zusammenhang beim Kauf eines Produktes A mit dem Kauf eines Produktes B in einem Schaubild durch eine Verbindungslinie zwischen A und B dar, und verfährt mit allen anderen Produkten genauso. So entstanden bei manchen Produkt-Kombinationen stärkere Linien, bei anderen dünnere. Durch Berechnungen und Vergleiche zwischen den Produkt-Kombinationen entstanden die so genannten Verknüpfungs-/Assoziationsregeln. Assoziationsregeln beschreiben also Korrelationen von gemeinsam auftretenden Dingen.

Um diese Verfahren in der Touristikbranche anzuwenden ersetzt man die Produkt-Kombinationen durch Kombinationen von Reisezielen eines Kunden. Da oft ein einzelner Kunde nicht genug Reisezielkombinationen aufweist um daraus aussagekräftige Regeln abzuleiten ist eine voran gegangene Clusteranalyse denkbar. Das Ergebnis der Clusteranalyse ist eine Reihe von Kundengruppen die wesentlich mehr Reisezielkombinationen enthalten und auf denen dann eine sinnvolle Assoziationsanalyse angewendet werden kann.

Eine zugehörige Assoziationsregel sieht dann z. B. wie folgt aus:
„12 % der Kunden aus Kundengruppe 1, die schon Urlaub auf Mallorca gemacht haben, war auch auf den Kanaren, diese beiden Reiseziele kommen in 2 % aller Kundengruppen vor“.

Jede Assoziationsregel enthält Informationen über die Stärke der Korrelation (nämlich in 12 % der Fälle), die sogenannte Konfidenz der Regel, wie auch Informationen über die Häufigkeit der darin gemeinsam vorkommenden Reiseziele (hier 2 %), die üblicherweise als Support bezeichnet wird.

Die Algorithmen, welche derartige Assoziationsregeln aufdecken, zeichnen sich dadurch aus,

dass sie alle Regeln finden, die eine gewisse Mindestkonfidenz und einen gewissen Mindestsupport besitzen. Regeln, die diese Anforderungen nicht erfüllen, werden dadurch nicht ausgewiesen.

Für den Benutzer ergibt sich hieraus die Tatsache, dass er nicht selber Hypothesen aufstellen muss, welche Dinge miteinander korrelieren könnten. Diese findet das Assoziationsverfahren selbständig heraus. Insbesondere, da es ein hoffnungsloses Unterfangen ist, alle möglichen Kombinationen Durchzuprobieren. Die Gefahr wäre sehr groß, dass dem Benutzer bestimmte Regeln gar nicht auffallen, weil er an diese gar nicht denken würde (A.).

Verfahren, um Assoziationsregeln aufzudecken, sind relativ einfach gebaut, vergleicht man sie mit den Verfahren der klassischen Statistik. Andererseits sollte ihr Nutzen nicht unterschätzt werden.

2.1.1 Assoziationsverfahren

Im Folgenden werden 3 Verfahren zur Bestimmung von Regeln mit Hilfe von Assoziationsverfahren vorgestellt. Dabei sind die Verfahren der Komplexität nach geordnet.

2.1.1.1 Brutforce

Der Brute-Force-Algorithmus stellt auch in der Assoziationsanalyse eine einfach zu realisierende, aber naive Möglichkeit zur Generierung von Regeln da. Das Grundprinzip besteht darin, alle möglichen Kombinationen von Elementen zu ermitteln. Diese werden mit den Transaktionen vergleichen und zu jeder Kombination werden die beiden Größen Support und Konfidenz ermittelt.
Durch die große Anzahl von Kombinationsmöglichkeiten, bei schon recht kleinen Mengen von Elementen, ist dieses Verfahren sehr rechenintensiv, insbesondere bei der Ermittlung mehrelementiger Regeln. Da jede mögliche Kombination aller Elemente des Datenbestand geprüft werden muss.

Ein erster Ansatz um diese Verfahren zu verbessern, besteht darin, im ersten Schritt Elemente zu suchen, die sehr oft vorkommen, also einen Mindest-Support und -Konfidenz erfüllen. Es werden nun Regeln bestimmt, die immer noch den Mindest-Support erfüllen, jedoch nur aus Elementen die im ersten Schritt gefunden wurden (8). Weitere Verbesserungen wie Reduktion der Anzahl der Transaktionen führen schließlich zum Apriori-Prinzip, dass im folgenden Abschnitt beschrieben wird.

2.1.1.2 Apriori-Algorithmus

Der Apriori-Algorithmus ist ein weiteres, jedoch deutlich verbessertes Verfahren der Assoziationsanalyse. Gegenüber dem Brute-Force-Algorithmus kommen die Beobachtungen bezüglich der Unabhängigkeit von Konfidenz und Support zum Einsatz. Der Name leitet sich von vorher bestehendem Wissen über häufig vorkommende Elemente ab. Tritt eine Item-Menge häufig auf, so wird davon ausgegangen, dass auch die Teilmengen, in denen dieses Item vorhanden ist, häufiger auftreten.

Die Arbeitsweise des Algorithmus besteht darin, aus einer *k*-elementigen Menge die (*k* + 1) Element-Menge zu erzeugen. Im ersten Schritt werden alle häufig vorkommenden ein-elementigen Itemsets erzeugt. Diese werden als Apriori-Gen bezeichnet. Alle 1-elementigen Itemsets, die nicht den Mindestsupport erfüllen, werden aus dem Apriori-Gen ausgeschlossen. Anschließend werden alle Itemsets des Apriori-Gens per Kreuzprodukt miteinander kombiniert und auf deren Support überprüft. Diese iterative Kombination von Item Sets, Festlegung dessen Supports und Ausschließung, der Itemsets, die nicht den minimale Support erfüllen, wird solange durchgeführt, bis kein sinnvolles Ergebnisse mehr zu erwarten ist oder eine maximale Anzahl von Items in einem Itemset erreicht wird. Die maximale Anzahl von Elementen in einem Itemset, sowie der minimale Support, werden vor Ausführung des Algorithmus festgelegt. Das Ergebnis des Apriori-Algorithmus ist eine Liste von *n* elementaren Itemsets die jeweils den minimalen Support erfüllen.

2.1.1.3 FP-Growth Algorithmus

Eine weitere Verbesserung gegenüber den zuvor betrachteten Algorithmen bietet der Frequent Pattern Algorithmus (FP-Growth). FP-Growth verwendet eine eigene Datenstruktur, den FB-Tree. Dieser komprimiert die originalen Transaktionsdaten und ermöglicht eine hohe Analyse-Effizienz. Der FP-Growth Algorithmus läuft in 2 Schritten ab. Im ersten wird der FB-Tree aufgebaut und im 2 wird dieser analysiert, um Itemsets mit hohem Support und Konfidenz zu finden. Das Vorgehen der einzelnen Schritte soll im Folgenden erklärt werden (Goethals, 2001).

1. Zur Erstellung der FB-Trees werden im ersten Schritt alle 1-elementigen Itemsets bestimmt, die den minimalen Support erfüllen. Diese werden absteigend nach ihrem Vorkommen sortiert. Dadurch wird die Reihenfolge der Elemente in der *Frequent Item Header* Tabelle festgelegt. Zur Konstruktion des Baumes wird die Transaktionsdatenbank ein zweites Mal durchlaufen. Nach Anlegen der Wurzel wird für die erste gefundene Transaktion ein Teilbaum angelegt. Jedes Element der Transaktion bekommt dabei einen Knoten, wobei der *count* auf 1 und die *nodelink*-Verweise auf null gesetzt werden. Bei den weiteren Durchläufen betrachtet man nun gemeinsam verlaufende Pfade. Enthält eine

weitere Transaktion die gleichen Elemente, wie eine zuvor eingefügte Transaktion, so werden die gleichen Knoten durchlaufen, *count* hochgezählt und beim ersten unstimmigen Knoten ein Zweig gebildet. Stimmt bereits der erste Knoten nicht überein, wird ein komplett neuer Pfad angelegt. Beim Anlegen von verschiedenen Pfaden kommt es vor, dass es trotzdem gleiche Knoten gibt. In diesem Fall muss der *node-link* Verweis vom ersten Vorkommen auf das nächste gesetzt werden, so das man über den Tabellenverweis *head of node link* Zugriff auf alle Knoten eines Namens hat. Nach dem Aufbau des Baumes folgt nun die Analyse der Pfade. Das Ergebnis des ersten Schrittes ist ein FP-Tree wie er in Abbildung 3 dargestellt. Dieser dient als Grundlage für den 2 Schritt.

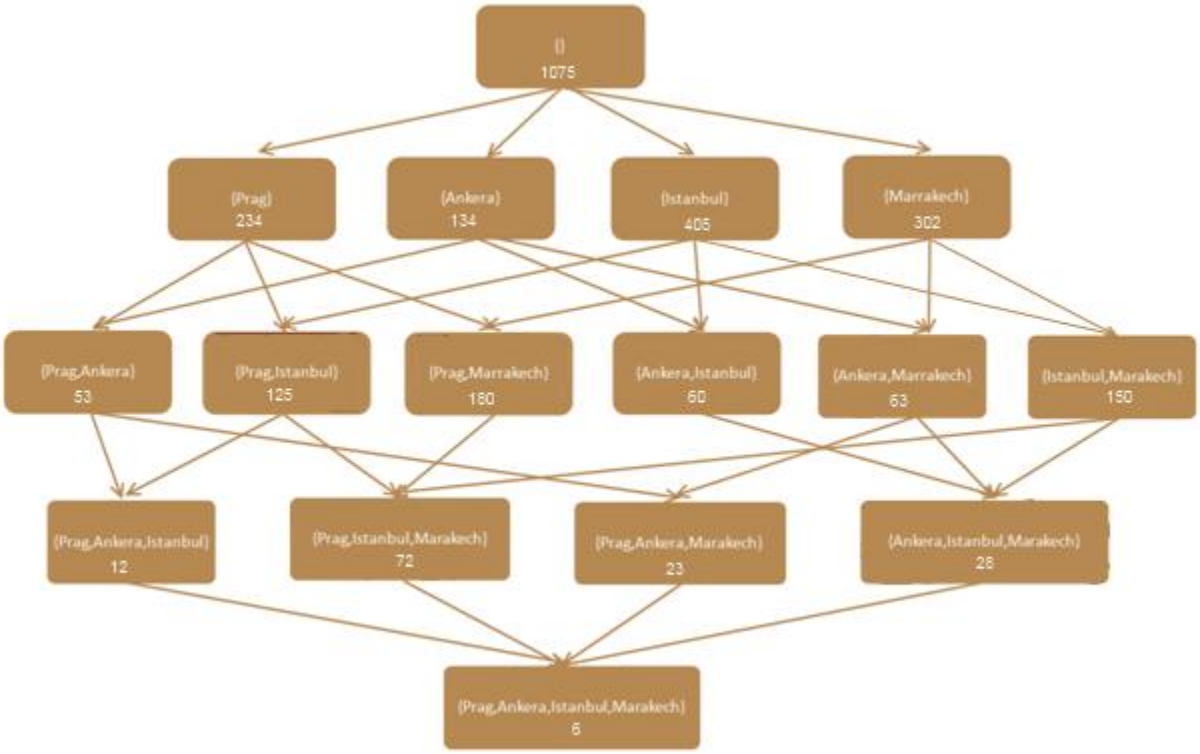


Abbildung 3: FP Growth Algorithmus

2. Im zweiten Schritt werden der Support jedes Itemsets und die Konfidenz jeder Regel, die abgeleitet werden kann, ermittelt. Der Support eines Itemsets (A, B) wird durch die Division des *count* Werts der jeweiligen Nodes durch den count des *Head Nodes* berechnet.
   Mit einem Itemset mit *n* Items lassen sich *n* Regeln ableiten. Die Konfidenz jeder Regel **(A->B)** wird berechnet indem der count Wert eines Nodes jeweils durch den Support des Implikates (A) geteilt wird.

Aus dem FP-Tree in Abbildung 3: FP Growth Algorithmus lassen sich die Supports und die Konfidenz, wie in Tabelle 1, ermitteln.
In den Tabellen ist nur ein Ausschnitt aller Itemsets und Regeln aufgelistet.
Aus den folgenden werten lassen sich nun die Assoziationsregeln ableiten. Ein Regel laute z.B.: „22.6 % der Kunden, die schon in Prag waren, waren auch in Ankara, diese beiden Reiseziele kommen in 4.9% aller Kundensätzen vor.“ Um nun die Qualität der Regel zu bestimmen wird der Lift berechnet. Dies ist die Konfidenz($\frac{count(A \rightarrow B)}{count(A)}$)durch die „erwartet Konfidenz“($\frac{count(A \rightarrow B)}{count(B)}$). Je höher dieser Lift ist desto besser ist die Aussagekraft der Assoziationsregel.

Support	Itemset
21.8%	Prag
12.5%	Ankara
37.7%	Istanbul
4.9%	Prag, Ankara
11.6%	Prag, Istanbul
5.6%	Ankara, Istanbul
1.1%	Prag, Ankara ,Istanbul

Lift	Konfidenz	Regel
1.8	22.6%	(Wenn) Prag ->(dann)Ankara
1.8	39.6%	Ankara->Prag
1.4	53.4%	Prag->Istanbul
1.4	30.9%	Istanbul->Prag
1.2	44.8%	Ankara->Istanbul
1.2	14.8%	Istanbul->Ankara
0.6	22.6%	Prag, Ankara->Istanbul
0.8	9.6%	Prag, Istanbul->Ankara
1.0	22.6%	Ankara, Istanbul->Prag

Tabelle 1: Ermittelte von Support und Konfidenz aus dem FB-Tree

2.1.2 Zusammenfassung Clusterverfahren

Abschließend lässt sich sagen, dass alle drei Assoziationsverfahren im Idealfall das gleiche Ergebnis liefern. Die Brut-Force Methode liefert das optimale Ergebnis, da alle möglichen Kombinationen, hinsichtlich Support und Konfidenz untersuche werden. Schlussfolgernd ist er sehr Rechenintensiv und bei großen Datenbeständen mit mehren Millionen Transaktionen und tausenden verschieden Items nicht realisierbar. Die beiden anderen Verfahren sind Heuristiken, die versuchen sich der optimalen Lösung zu nähern, ohne dabei alle Item Kombinationen zu untersuchen.
Aus diesem Grund wird in fast allen Fällen der FP Growth Algorithmus zur Assoziationsanalyse eingesetzt. Da die beiden anderen Algorithmen keine Vorteile im Vergleich zu diesem Algorithmus bieten.

2.2 Klassifikationsverfahren

Klassifikation ist das Einordnen von Objekten in vorgegebene Klassen. Dabei zeichnen sich die einzelnen Klassen durch ein eindeutiges Klassenmerkmal aus. Dem Klassifikationsverfahren sind alle Ausprägungen dieses Klassenmerkmals bekannt und alle vorhandenen Informationsobjekte können genau einer Klasse zugeordnet werden.
Die Frage des Klassifikationsverfahrens lautet nun: In welche Klasse passt ein neues Informationsobjekt anhand seiner individuellen Merkmalskombination und aufgrund der Analyse der vorhanden Informationsobjekte und den Merkmalskombinationen am besten?

Klassifikationsverfahren laufen in zwei wesentlichen Schritten ab:

1. Lernphase (Erstellung eines Klassifikators): Aus der Datenbasis werden zufällig einige Objekte ausgewählt und zu einer Trainingsmenge zusammengestellt. Zu jedem Trainingsobjekt muss in einem zusätzlichen Attribut die Klasse vorgegeben bzw. vermerkt werden, in die es gehört. Man spricht daher von überwachtem Lernen (engl.: supervised learning) (Aggarwal, 1999). Anhand der klassifizierten Trainingsdaten wird mittels eines Algorithmus ein Modell (z. B. ein Satz von Regeln) erstellt, dass zu jeder Merkmalskombinationen die zugehörige Klasse angeben kann. Dieses Modell bezeichnet man als Klassifikator.

2. Klassifikationsphase (Anwendung des Klassifikators): Als Ergebnis wird zu jedem Objekt seine Klasse zugeordnet. Es ist darauf zu achten, dass das ermittelte Modell nicht zu genau an die Trainingsdaten angepasst ist, sondern flexibel genug bleibt, auch neue Daten korrekt zu klassifizieren (Problem des ’overfitting’). Daher sollte die Brauchbarkeit des Klassifikators vor der Anwendung überprüft werden, z. B. anhand von Testdaten. Neben der Vorhersagegenauigkeit sind auch die Geschwindigkeit, die Robustheit bei Ausreißern, die Eignung für große Datenmengen und die Interpretierbarkeit der Ergebnisse von Interesse.

Die mathematischen Methoden, die für das Aufstellen von Klassifikatoren hilfreich sind, entstammen sowohl der klassischen Statistik (K-Nächste-Nachbarn-Methode) als auch dem maschinellen Lernen. Symbolische Lernverfahren, beispielsweise Entscheidungsbaumverfahren, stellen Verfahren dar, welche für den Anwender verständliche Klassenbeschreibungen generieren. Subsymbolische Verfahren wie Künstliche Neuronale Netze arbeiten hingegen nach dem Black-Box-Prinzip, wodurch Klassenbeschreibungen nicht aus dem konstruierten Modell ableitbar sind.
Im Folgenden sollen werden die verschieden Methoden der Klassifikation beschrieben werden.

2.2.1 Neuronale Netze

Mit Hilfe von neuronalen Netze kann man nicht-lineare funktionale Beziehungen modellieren, indem man versucht, die Funktionsweise des menschlichen Gehirns nachzubilden, mit den Eigenschaften sich selbständig zu verbessern, sprich dazu lernen. Ein neuronales Netz besteht aus sogenannten Eingangs-, Inner- und Ausgangsneuronen, welche insgesamt die Werte eines Eingangs- bzw. Ausgangsvektor annehmen. Diese Neuronen werden über weitere Neuronen so verknüpft, dass über spezielle funktionale Zusammenhänge der Eingangsvektor in einen Ausgangsvektor transformiert wird (Palmaer, 2006).
Neben der Wahl einer geeigneten Netzwerktopologie, also dem Aufbau des Netzes, ist vor allem der Lernalgorithmus von Bedeutung, mit dem die Beziehungen zwischen den Neuronen über eine geschickte Wahl von Verknüpfungsgewichten hergestellt werden. Man verwendet zur Bestimmung ein Trainingsset aus Input- und Output-Vektoren. Abweichungen zwischen dem vom Netz ausgegebenen Output-Vektoren und dem tatsächlichen Output-Vektoren des Trainingssets resultieren aus bestimmten Änderungen der Gewichte. Der Lernalgorithmus läuft so lange, bis das Netz mit hinreichender Genauigkeit den funktionalen Zusammenhang zwischen Input- und Output-Vektoren wiedergibt. Danach ist das Netz im Idealfall in der Lage, für alle möglichen Eingangsvektoren die passenden Ausgangsvektoren zu ermitteln (Kudrass, 2007). Der größte Nachteil bei der Anwendung von neuronalen Netzen ist, dass das Lernen von Gewichtungen und Regeln der einzelnen Variablen nicht ersichtlich ist, weshalb man auch vom Black-Box-Prinzip spricht. Dadurch sind die Ergebnisse, die das Verfahren liefert, nicht nachvollziehbar. Aus diesem Grund sollten das Verfahren der neuronalen Netze nur dort zur Anwendung kommen wo folgende Eigenschaften erfüllt sind (Berry Michael J. A., 2004):

- der Datensatz ist verständlich
- die Ergebnisse können leicht interpretiert werden
- Genug ähnliche und vergleichbare Fälle, mit Eingabe- und Ausgabevektoren sind vorhanden.

Durch diese Tatsche, dass das Zustandekommen von Ergebnissen nicht nachvollzogen werden kann, machen sich neuronale Netzwerke besonders gut bei der Automatisierung von Aufgabenstellungen, die zwar einfach und nachvollziehbar sind, jedoch viel Zeit und die Analyse von vergleichbaren Fällen benötigen. Ein Anwendungsfeld ist die Ermittlung der Zugehörigkeit eines Kunden zu fest definierten Kundengruppen (Schneider, 2007)

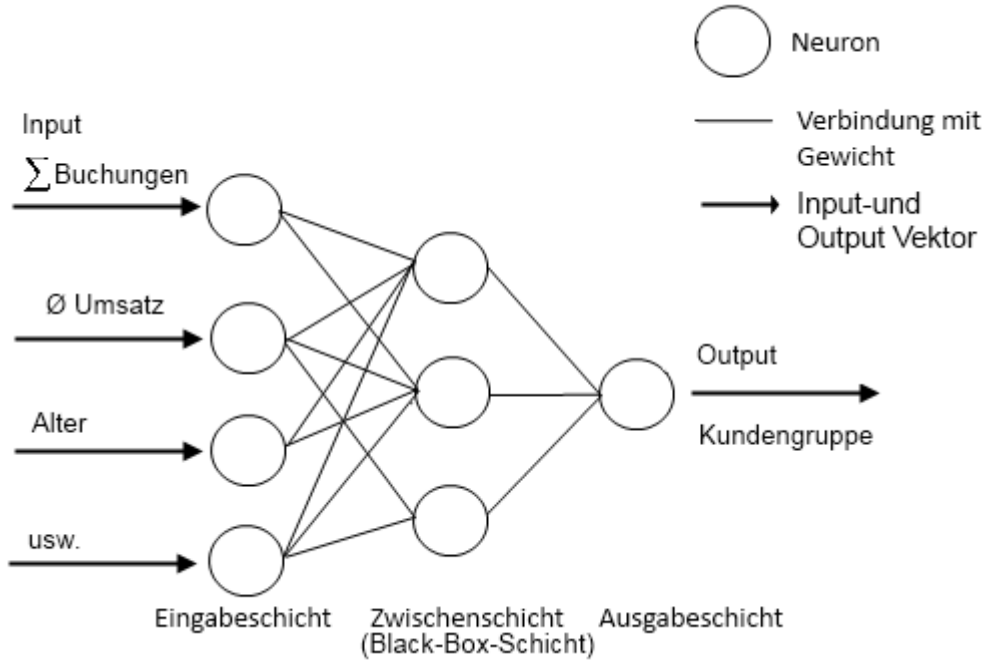


Abbildung 4: Neuronales Netz zur Bestimmung eines Kunden zur Zugehörigkeit zu einer Kundengruppe

2.2.2 Entscheidungsbaumverfahren

Entscheidungsbaumverfahren nutzen Baumstrukturen, um nicht klassifizierte Informationsobjekte einer Klasse zuzuordnen. Dabei repräsentiert jedes Blatt eines Entscheidungsbaumes eine der möglichen Klassen. Die Kanten des Baums entsprechen den möglichen Merkmalswerten des Informationsobjekts, die Knoten des Baums repräsentieren die Merkmale als solche.
Der Basis-Algorithmus arbeitet wie folgt: Anfangs gehören alle Trainingsdaten zur Wurzel des Baums. Im ersten Schritt wird ein Attribut, das so genannte Splitattribut ausgewählt und die Trainingsdaten werden durch die jeweils aktuellen Splitattribute partitioniert. In den folgenden Schritten werden iterativ alle Attribute der Trainingsmenge als Splitattribute zum Aufbau des Baumes verwendet. Dieser Ablauf wird beendet, wenn keine Splitattribute mehr vorhanden sind.

Abbildung 5 illustriert einen einfachen Entscheidungsbaum. Als erstes Splitattribut wurde das Attribut Alter gewählt. Die Kanten des ersten Knotens splitten die Trainingsdaten auf. Jedes Blatt des Baumes stellt eine mögliche Klasse dar, wobei die eine Klasse in mehreren Blättern vorkommen kann.

Identifikation von Kunden im Kundenstamm, die Potenzial für Dauercamper haben.

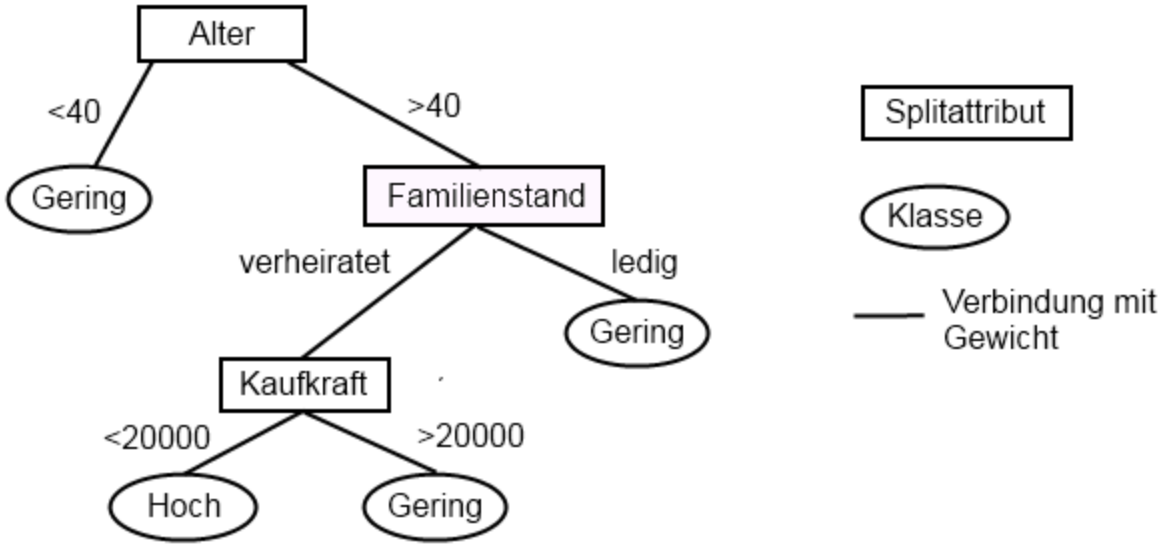


Abbildung 5: Beispiel eines Entscheidungsbaums

Die Vorteile des Entscheidungsbaum-Klassifikators liegen zum einen in der einfachen Interpretation des gefundenen Baumes, zum anderen können die Attribute implizit gewichtet werden. Er zählt zu den

leistungsfähigen Klassifikationen und wird häufig in der Praxis eingesetzt. Ein möglicher Nachteil des Entscheidungsbaum-Klassifikators ist, dass das Finden eines optimalen Entscheidungsbaums im ungünstigsten Fall exponentielle Zeit benötigt. Weiterhin können heuristische Methoden nur ein lokales Optimum im Baum finden.

2.2.3 Naives Bayes

Naives Bayes ist ein verbreitetes Klassifikationsverfahren, welches nach seinem Begründer, Thomes Bayes, einem britischen Philosophen des 18. Jahrhunderts, benannt ist. Es wird als „naiv" bezeichnet, da er davon ausgeht, dass die einzelnen Input Merkmale unabhängig untereinander sind, wodurch er einen geringen Rechenaufwand aufweist und zur schnellen Generierung von Miningmodellen dient. Diese einfachen Miningmodelle können verwendet werden, um sich einen Überblick über die Beziehungen zwischen Eingabespalten und vorhersagbaren Spalten zu verschaffen. Dieses Wissen, über die Größe des Einflusses einer Eingabespalte auf des Ergebnis der Ausgabespalte, kann man nutzen, um weitere Miningmodelle mit anderen Algorithmen, deren Rechenaufwand größer ist, zu erstellen. Mit diesen optimierten Miningmodell ist man in der Lage präzisere Aussagen zu erhalten.

Funktionswiese des Naives Bayes

Bei der Bayes-Klassifikation wird ein Objekt derjenigen Klasse zuordnet, die für seine individuelle Merkmalskombination am wahrscheinlichsten ist. Gegeben seien *k* Klassen C_1,C_2, . . . ,C_k und ein zu klassifizierendes Objekt mit *m* Merkmalen x_1, x_2, . . . , x_m, die im Merkmalsvektor X zusammengefasst werden. $P(C_i|X)$ gibt dann die Wahrscheinlichkeit an, dass das Objekt mit dem gegebenen Merkmalsvektor X zur Klasse C_i gehört. Gesucht ist diejenige Klasse C_i, für die diese Wahrscheinlichkeit am größten ist; in diese wird das Objekt eingeordnet. Nach dem Satz von Bayes gilt:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Da P(X) konstant ist, reicht es zu prüfen, für welche Klasse der Zähler $P(X|C_i)P(C_i)$ maximal wird. Die Auftrittswahrscheinlichkeiten der Klassen $P(C_i)$ werden anhand der relativen Häufigkeiten in den Trainingsdaten geschätzt oder auch vereinfacht als gleichverteilt angenommen. Um die Berechnung von $P(X|C_i)$ zu vereinfachen, wird angenommen, dass bei gegebener Klasse die Ausprägungen der Attribute unabhängig voneinander sind.
Unter der Annahme der Unabhängigkeit der Merkmale ist $P(X|C_i)$ das Produkt aus den bedingten Wahrscheinlichkeiten für die in X vorkommenden Ausprägungen x_j der einzelnen Attribute:

$$P(C_i|X) = \prod_{j=1}^{m} P(X|C_i)$$

Diese Einzelwahrscheinlichkeiten $P(x_j|C_i)$ können wiederum anhand der relativen Häufigkeiten in den Trainingsdaten geschätzt werden. Dazu betrachtet man die Trainingsdaten nach Klassen getrennt und setzt die Anzahl der Objekte mit der Ausprägung x_j für das j-te Attribut ins Verhältnis zur Anzahl aller Objekte dieser Klasse. Bei kontinuierlichen Merkmalen erfolgt die Schätzung anhand einer angenommenen Verteilungsfunktion (Han, 2000).
Naive Bayes-Klassifikation erzielt bei Anwendung auf großen Datenmengen eine hohe Genauigkeit und eine vergleichbare Geschwindigkeit wie Entscheidungsbaumverfahren und neuronale Netze. Voraussetzung dafür ist aber, dass die Annahmen über Verteilungen und der Unabhängigkeit der Attribute zutreffend ist. Dies festzustellen kann teilweise eine nichttriviale Aufgabe sein. Ist diese ungerechtfertigt, werden die Ergebnisse sehr ungenau.

2.2.4 Evaluierung und Vergleich von Klassifikationsverfahren

Im vorherigen Abschnitt wurden Verfahren vorgestellt, mit deren Hilfe man eine Klassifikation durchführen kann.
Klassifikationsverfahren ordnen nicht klassifizierten Objekten einer Klasse zu. Als Basis dient dabei das trainierte Modell. Allerdings muss das aufgestellte Modell seine Eignung für Praxisdaten noch beweisen, dazu existieren verschieden Testverfahren.

Die einfachste Variante wäre, wenn man das Modell mit den Trainingsdaten selbst testet. Das Ergebnis würde allerdings zu optimistisch sein, da das Modell auf diesen Daten basiert. Man benötigt also eine unabhängige Testmenge, um Aussagen über die Leistungsfähigkeit machen zu können. In den folgenden Abschnitten werden verschiedene Verfahren vorgestellt, mit denen man die Leistungsfähigkeit von Modellen vergleichen kann.

2.2.4.1 Aufteilung der Datenmenge

Falls man eine große Datenmenge zur Verfügung hat, so geht man bei der Aufteilung der Daten wie folgt vor. Die Datenmenge wird in drei Teile aufgeteilt:

1. eine Trainingsmenge, um damit passende Modelle zu erstellen
2. Testmenge, um den Vorhersagefehler des gewählten Modells zu bestimmen

3. eine Validierungsmenge, um mehre Modelle miteinander vergleichen zu können und sicher zu stellen das die Daten für alle Modelle unbekannt sind

Bei genügend großer Datenmenge ist aus Erfahrung 50% der Datenmenge
als Trainings-, und je 25% als Validierungs- bzw. Testmenge zu benutzen (T. Hastie, 2001). Ob eine Datenmenge genügend groß ist hängt von der Modelkomplexität ab. Desto komplexer ein Modell ist desto mehr Daten werden zum Trainieren -, Testen und Validieren benötigt. Bei Anwendungen im Data-Mining kann es jedoch vorkommen, dass nur sehr wenige Trainingsdaten zur Verfügung stehen und deshalb eine solche Partitionierung nicht möglich ist. Dann muss die Trainingsmenge sowohl zum Training als auch zum Bewerten der Modelle dienen. Um in diesem Fall dennoch verlässliche Vorhersagen über die Leistungsfähigkeit der Modelle zu bekommen, wurden Verfahren entwickelt, die eine automatisierte Einteilung der Datenmengen vornehmen. Zwei dieser Verfahren, die Kreuzvalidierung und Bootstrap, werden im folgenden Abschnitt vorgestellt.

2.2.4.2 Kreuzvalidierung

Die einfachste und am meisten benutzte Methode zur Schätzung des Fehlers bei einer Klassifikation ist die Kreuzvalidierung. Bei der Kreuzvalidierung wird die Gesamtmenge der Daten in k etwa gleich große Mengen aufgeteilt. Daraus werden $k-1$ dieser Mengen als Trainingsmenge benutzt, die übriggebliebene Teilmenge bildet die Testmenge. Dieses Verfahren wird so lange wiederholt, bis jede der k Mengen einmal als Testmenge benutzt wurde. Schließlich werden die Ergebnisse der k verschiedene Durchgänge kombiniert um das endgültige Modell zu erhalten.
Bei der Wahl von k sind zwei Dinge zu beachten: Ein zu groß gewähltes k, im maximalen Fall besteht jedes k aus einem Informationsobjekt und die Anzahl der Teilmengen ist gleich der Anzahl den Informationsobjekte der gesamten Daten (wird als leave-one-out Methode bezeichnet), erhöht den Rechenaufwand und ein zu klein gewähltes k hingegen sorgt nur für eine zu kleine Anzahl von Trainingsdaten und führt zum Effekt des Underfitting, bei dem das Modell untertrainiert ist und schlechte Ergebnisse liefert. In der Praxis wird meistens ein k zwischen 5 und 10 gewählt. Dies stellt einen guten Kompromiss zwischen Trainings- und Testdatenmenge dar und liefert stabile Ergebnisse.

2.2.4.3 Bootstrap

Das Bootstrap-Verfahren ist eine geglättete Version der Kreuzvalidierung mit einigen Änderungen, um einen besseren Bias (dt.: statistische Verzerrung) zu erhalten (B. Elfron, 1993).
Bootstrap beruht darauf, zufällige Stichproben der Größe N (bootstrap Samples) mit Zurücklegen aus der Menge der Trainingsdaten zu ziehen, und diese dann als Trainingsdaten zu benutzen. Die so gewonnene Trainingsmenge ist also genauso groß wie die ursprünglichen Trainingsdaten, kann aber manche Datensätze der ursprünglichen Datenmenge enthalten. Das ganze Verfahren wird *B* mal wiederholt und man erhält *B* sogenannte bootstrap-Datensätze. Diese werden benutzt, um die Modelle zu trainieren und zu testen.
Bei Evaluierung des Modells muss darauf geachtet werden, dass sich die Teilmengen der bootstrap-Datensätze nicht überschneiden, was zu einer zu optimistischen Abschätzung der Leistungsfähigkeit des Modells führen würde, da die bootstrap-Datensätze zum Trainieren benutzt werden und gleichzeitig Teil der Testdaten wären.
Um diesen Effekt zu vermeiden, nutzt man zum Testen nur diejenige Datensätze, bei denen ein bestimmtes (x_n, y_n) nicht im bootstrap-Datensatz vorkommt (ähnlich der leave-one-out–Methode). Die Wahrscheinlichkeit, dass ein bestimmter Datensatz nicht Teil der bootstrap-Datensätze ist, beträgt

$$(1 - \frac{1}{N})^N \approx e^{-1} \approx 0{,}368$$

da für jede einzelne Auswahl die Wahrscheinlichkeit $1 - \frac{1}{N}$ beträgt, dass ein bestimmter Datensatz nicht gezogen wird. Bei hinreichend großer Datenmenge enthält die bootstrap-Menge im Mittel 63,2% aller Trainingsdaten, es bleiben als im Mittel 36,8% der Daten als Testdaten, die im Training nicht eingesetzt werden.

2.2.4.4 Receiver-Operating-Characteristic – Kurven

Eine Möglichkeit der Visualisierung der Sensitivität und Spezifität von Klassifikationsverfahren stellen die Receiver-Operating-Characteristic (ROC)–Kurven dar, die zur Leistungsbewertung von mehreren Modellen zur Trennung von zwei Klassen benutzt werden kann. Es werden also immer Ja-Nein–Entscheidungen betrachtet, beispielsweise relevant/irrelevant oder krank/gesund. Bei der Klassifikation wird jeweils ein Schwellenwert definiert, der die Zuordnung zu einer der beiden Klassen bestimmt; liegt $\hat{y}$ über diesem Schwellenwert, so erfolgt die Zuordnung zu der einen Klasse und sonst zur anderen. Zur Bestimmung der ROC-Kurve wird auf der x-Achse die relative Anzahl der falsch klassifizierten Werte (Spezifität) und auf der y-Achse die relative

Anzahl der richtig klassifizierten Werte(Sensitivität) gegeneinander aufgetragen. Der Bereich eines solchen Diagramms ist also auf beiden Achsen der Bereich [0,1]. Variiert man nun den bei der Klassifikation benutzten Schwellenwert und trägt jeweils das bei einem bestimmten Schwellenwert resultierende Verhältnis von Sensitivität und Spezifität gegeneinander auf, so erhält man eine Kurve. Diese Kurve beginnt immer im Punkt [0,0], endet in [1,1], ist degressiv steigend und liefert eine Möglichkeit, die Güte verschiedener Modelle zu vergleichen. Der optimale Punkt einer solchen Kurve liegt in der linken oberen Ecke: Dort ist die relative Anzahl der „false positives“ gleich null und gleichzeitig die relative Anzahl der „true negatives“ gleich eins, das Modell klassifiziert also immer korrekt und liefert eine optimale Vorhersage. Verläuft die Kurve unterhalb dieses Punkts, so kommt es zu Fehlklassifikationen und diese kann man mittels ROC-Kurven visualisieren.

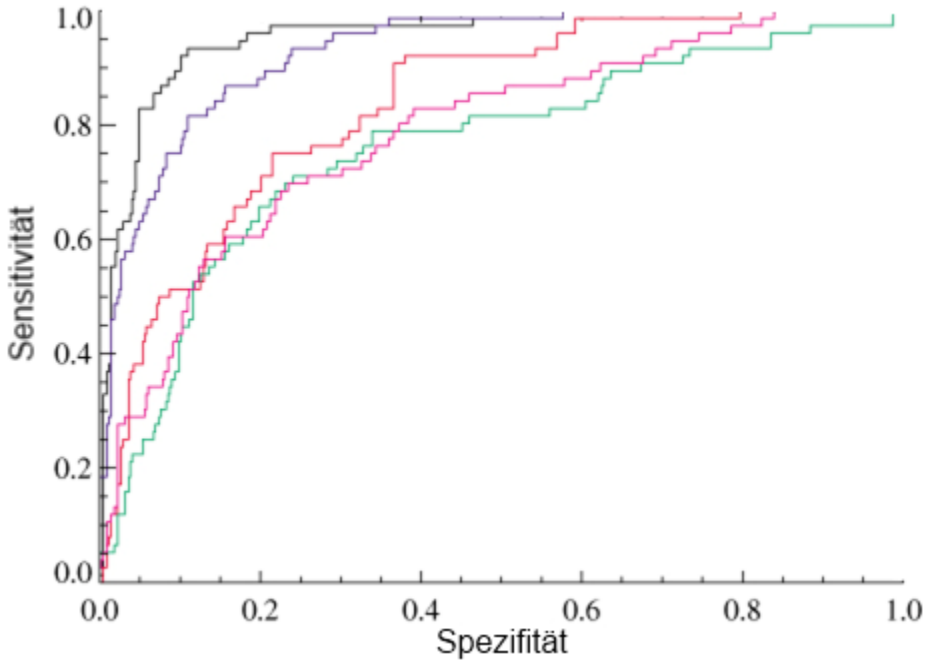


Abbildung 6: Receiver-Operating-Characteristic

2.3 Clusterverfahren

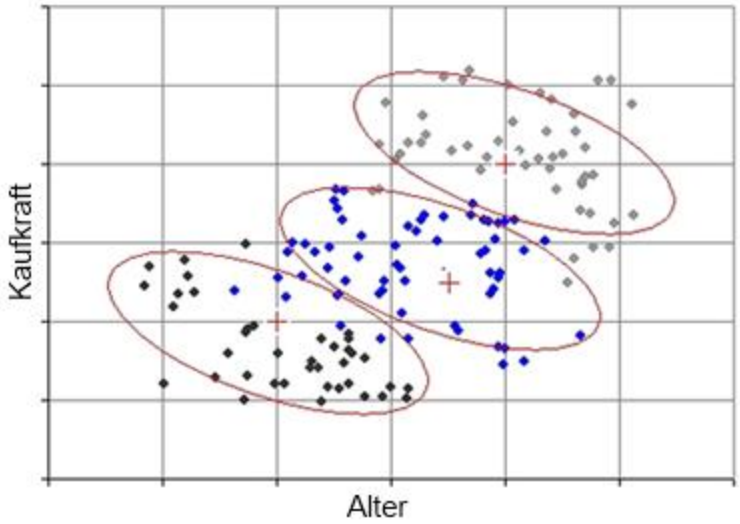


Abbildung 7: gebildete Cluster nach Kaufkraft und Alter

Mit Clusterverfahren werden Objekte anhand ihrer Merkmale zu Gruppen, sogenannten Clustern, zusammengestellt. Die Gruppierung soll dabei so erfolgen, dass die Merkmale der Objekte innerhalb eines Clusters sich möglichst ähnlich sind, die Cluster untereinander sich aber möglichst unähnlich sind.

Die Verwendung von Clusterverfahren erfolgt durch die Annahme, dass Objekte mit ähnlichen Merkmalen auch ähnliche Ursachen, Folgen oder Bedürfnisse haben. So werden Clusterverfahren im Marketing zur Kundensegmentierung verwendet, bei denen der Kundenstamm in Gruppen mit ähnlichen Merkmalen unterteilt wird. Diese Gruppen können dann individuell behandelt werden. Die Ähnlichkeit oder Unähnlichkeit wird mit Hilfe von verschiedenen Distanzfunktionen bestimmt.

Voraussetzung für die Anwendung von Distanzfunktionen ist die Möglichkeit, dass jedes Informationsobjekt durch einen Vektor von Merkmalen in einen bekannten *n*-dimensionalen Raum dargestellt werden kann.

Der wesentliche Unterschied zu Klassifikationsverfahren besteht darin, dass die einzelnen Klassen im Voraus nicht bekannt sind. Somit ist unklar, was die einzelnen Klassen auszeichnet und welche Merkmalskombinationen auftreten müssen, um ein Objekt einer Klasse zuzuordnen.

Dadurch ist im Anschluss eines Clusterverfahrens immer eine umfangreiche Interpretation und Evaluierung der gebildeten Cluster notwendig, um genau zu analysieren, wie sich die einzelnen Cluster unterscheiden.

Oft werden Clusterverfahren angewendet, um sehr große Datenmengen, im Rahmen des Prozesses der Vorverarbeitung des Knowledge Discovery in Databases, in kleine Teilmengen zu unterteilen, um auf diesen andere Data Mining Verfahren anzuwenden.

Bei der Anwendung des Clusterverfahren sollten im Voraus folgende Entscheidungen getroffen werden:

- Wahl der Merkmale:
  Welche Merkmale eines Objekts sollen verwendet werden. Dabei empfiehlt es sich Merkmale zu wählen, die unabhängig von einander sind und die keine Redundanten Informationen enthalten.

- Festlegung der Anzahl der zu bildenden Cluster.
  Ist gibt Clusteralgorithmen (*soft Cluster*), die im Rahmen der Clusterbildung die Anzahl der zu bildenden Cluster bestimmen. Bei anderen Verfahren (*hard Cluster*) muss vor der Berechnung festgelegt werden, wieviele Cluster abgeleitet werden sollen. Dabei sollte bedacht werden, dass es keine allgemeingültige Anzahl von Clustern gibt, die sich als optimal erweisen. Sondern je nach Problemstellung muss entschieden werden, wieviele Cluster gebildet werden sollen. Für einen Reiseveranstalter der z.B. Clusterverfahren zur Kundensegmentierung verwenden möchte, macht es keinen Sinn, 20 oder mehr Kundengruppen abzuleiten, da er nicht in der Lage ist so viele unterschiedliche Gruppen individuell zu betreuen. Es sollte eine max. Anzahl von Clustern gewählt werden die für einen bestimmten Anwendungsfall als sinnvoll erachtet werden.

- Wahl des Clusterverfahrens
  Bei der Beurteilung ist zu berücksichtigen, welche der generellen Anforderungen an das Clusterverfahren (Skalierbarkeit, Fähigkeit mit unterschiedlichen Datentypen umzugehen, Entdeckung von Clustern mit beliebiger Form, geringe Ansprüche an Inputparameter, Umgehen mit Ausreißern, Unabhängigkeit von der Ordnung der Daten, Behandlung hochdimensionaler Daten, Verwendbarkeit von Nebenbedingungen, Interpretierbarkeit und Anwendbarkeit der Ergebnisse) in der vorliegenden Analysesituation von besonderer Bedeutung ist.

Für die Aufteilung von Objekten anhand ihrer Merkmale gibt es zwei allgemeine Vorgehen von Clusterverfahren, denen alle Clusteralgorithmen jeweils zugeordnet werden können. Diese beiden Verfahren, das partitionierende Clusterverfahren und hierarchische Clusterverfahren, sollen im Folgenden beschrieben werden.

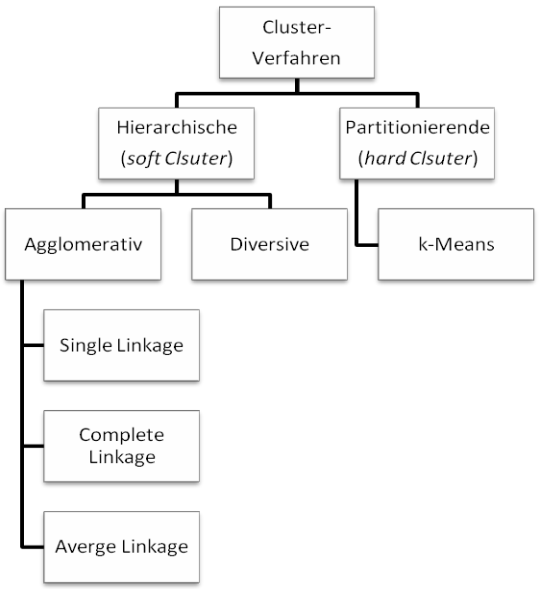


Abbildung 8: Clusterverfahren im Überblick

2.3.1 Partitionierende Clusterverfahren

Die partitionierenden Methoden suchen nach unbekannten Datenmustern. Ziel der Algorithmen ist es, die vorhandenen Daten möglichst trennscharf Klassen zuzuteilen (Aggarwal, 1999). Man spricht auch vom *hard Cluster* Verfahren, da jedes Objekt eindeutig einer Gruppe zugeordnet wird und diese Gruppen sich nicht überlappen.

Diese Verfahren gehen von einer gegebenen Gruppierung, der so genannten Startpartition, aus. Auf Basis von Austauschalgorithmen werden die Gruppen solange umgeordnet, bis ein vorgegebenes Optimum erreicht ist. Diese Verfahren sind von der Startpartition abhängig und sehr rechenintensiv.

Probleme, die im Zusammenhang mit partitionierten Clusterverfahren auftreten können:

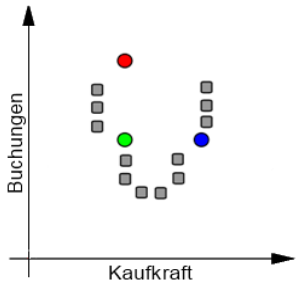
- Der Algorithmus konvergiert nicht, läuft nicht zusammen
- Wenn ein Startpunkt gewählt wird, dem keine Objekte zugeordnet werden, bleibt dieses Cluster in den folgenden Optimierungsschritten leer und wird nicht weiter angepasst

Beide Probleme können mit dem erneuten Ausführen des Algorithmus behoben werden, da bei jeder erneuten Ausführung neue zufällige Startpunkte gewählt werden und somit sich die Ergebnisse von partitionierende Clusterverfahren immer etwas unterscheiden.

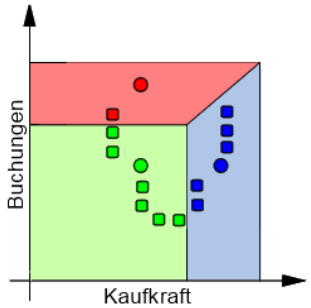
In der Literatur wird das partitionierte Clusterverfahren fast ausschließlich mit dem K-Means-Algorithmus in Verbindung gebracht.

2.3.1.1 K-Means

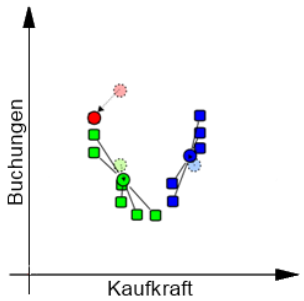
Der K-Means Clusteralgorithmus gruppiert Informationsobjekte nach deren Distanz zwischen den Objekten. Im Folgenden wird die Funktionsweise des Algorithmus erklärt (A.):

1. 

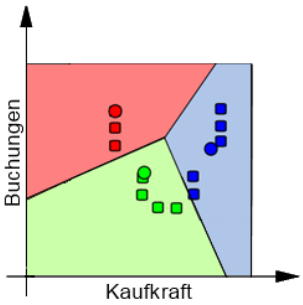
Zu Beginn liegen alle Objekte in einem Vektorraum, deren Position durch ihre Eigenschaftsvektor beschrieben wird. Die *k* Cluster-Schwerpunkte werden zufällig verteilt, dabei muss die Anzahl *k* der zu bilden Gruppen vor der Ausführung des Algorithmus festgelegt werden.

2. 

Jedes Objekt wird demjenigen Cluster zugeordnet, dessen Schwerpunkt ihm am nächsten liegt. Dazu muss eine Distanzfunktion, zum Beispiel die Euklidische Norm oder die Mahlanobis-Distanz, die bei nicht *gleichdimensionalen* Daten, verwendet wird.

3. 

Für jeden Cluster wird der Schwerpunkt neu berechnet, sodass er in der Mitte des Clusters liegt.

4. 

Basierend auf den neu berechneten Zentren werden die Objekte wieder wie in Schritt 2 auf die Cluster verteilt, bis sich die Schwerpunkte nicht mehr bewegen oder eine festgelegte maximale Iterationstiefe erreicht wurde.

Eigenheiten des K-Means-Algorithmus :

- Der K-Means-Algorithmus liefert für unterschiedliche Startpositionen der Clusterzentren möglicherweise unterschiedliche Ergebnisse
- Ein optimales Clustering zu finden gehört zur Komplexitätsklasse NP . Dadurch findet der K-Means Algorithmus findet nicht notwendigerweise die optimale Lösung, ist aber sehr schnell.

Um diese Probleme zu umgehen, startet man den K-Means-Algorithmus einfach neu in der Hoffnung, dass beim nächsten Lauf durch andere zufällige Clusterzentren ein anderes Ergebnis geliefert wird.

2.3.2 Hierarchische Clusterverfahren

Hierarchische Clusterverfahren werden unterteilt in *agglomerative(anhäufende)* und *divisive(teilende)* Methoden.
Man bezeichnet die hierarchischen Clusterverfahren auch als soft Cluster, da die Anzahl der zu bildenden Cluster nicht von Beginn an festgelegt wird, sondern vom Algorithmus automatisiert ermittelt wird.

Agglomerative Methode

Bei der agglomerativen Methode werden Cluster durch sukzessives Gruppieren von Objekten und im weiteren Verlauf durch Fusion von Gruppen zu größeren Gruppen konstruiert.

Die Aggregation beginnt mit den feinst möglichen Gruppen, die jeweils aus genau einem Objekt bestehen. Durch Zusammenfassen der zwei ähnlichsten Objekte wird eine erste zweielementige Gruppe gebildet. Zur Bestimmung der Ähnlichkeit zweier Objekte ist ein Maß zu bestimmen, dass in der Lage ist, die Ähnlichkeit in numerischer Form auszudrücken. Bei Metrisch skalierten Merkmalen können dabei Distanz- oder Ähnlichkeitsmaße eingesetzt werden. Ordinal und nominal skalierte Merkmale müssen dagegen als Dummy-Variablen (z.B. Geschlecht männlich 1 und weiblich 0) codiert werden. Dadurch wird jedes Objekt als Punkt in einem endlich-dimensionalen Raum repräsentiert. Seine Dimension stimmt mit der Anzahl der Analysevariablen überein. Als Maße für Unähnlichkeiten werden Metrikern in endlich dimensionalen reellen Räumen oder davon abgeleitete Größen wie die Euklidische Metrik oder deren quadrierter Wert verwendet.

Bereits nach der Bildung einer ersten Gruppe muss die ursprüngliche Definition des Abstands zwischen einzelnen Objekten zu einer Definition für Abstände zwischen Gruppen und Objekten bzw. allgemeiner zwischen verschiedenen Gruppen erweitert werden.

Dazu gibt es folgende Strategien zur Berechnung der Distanz zwischen mehrelementigen Gruppen

- Single-Linkage-Verfahren: Die kleinste Distanz zwischen Punkten der einen und Punkten der anderen Klasse wird berechnet.
- Complete-Linkage-Verfahren: Die größte Distanz zwischen Punkten der einen und Punkten der anderen Klasse wird berechnet.
- Average-Linkage-Verfahren: Die Distanz des Mittelwerts aller Abstände zwischen Punkten zweier verschiedener Klassen wird berechnet.

Die Abstände zwischen den Objekten, die im ersten Schritt des Verfahrens nicht aggregiert wurden, bleiben unverändert.
In den nachfolgenden Schritten werden Paare von Gruppen und/oder Objekten mit dem jeweils kleinsten Distanz zu neuen Gruppen zusammengefasst. An jeden Aggregationsschritt schließt sich wieder die Berechnung neuer Abstände analog zum vorhergehenden Schritt an. Das Verfahren erzeugt auf diese Weise in jedem Schritt eine neue Gruppe durch Vereinigung zweier bereits konstruierter Gruppen. Dadurch wird laufend eine etwas „gröbere“ Gruppierung generiert. Diese besteht nach dem *i*-ten Schritt aus *n*-*i* Gruppen, wobei *n* die Anzahl aller Objekte ist. Das Aggregationsschema besteht aus der Iteration von zwei aufeinanderfolgenden Operationen:
1. Berechnung von Distanzen zwischen den Gruppen der *i*-ten Stufe;
2. Vereinigung aller Gruppen der *i*-ten Stufe mit minimaler Distanz, wodurch die Gruppierung der (*i*+1)-ten Stufe erzeugt wird. Sie enthält eine Gruppe weniger als die Gruppierung der vorhergehenden *i*-ten Stufe und stimmt in (*n*-*i*-2) Gruppen mit denen der *i*-ten Stufe überein.

Der Prozess wird bis zur Aggregierung aller *n* Objekte in einer einzigen Gruppe fortgesetzt. Dies wird nach *n*-1 Schritten erreicht. Eine charakteristische Statistik des Prozesses ist die Folge der Abstände, die die sukzessiv vereinigten Paare von Gruppen besitzen. Der Verlauf dieser Abstände, die auch als Fusionswerte bezeichnet werden, ist ein Hilfsmittel bei der Bestimmung der Clusterlösung.

Divisiven Methode

Die divisiven Methode geht im Vergleich zur agglomerativen Methode in umgekehrt Reihenfolge vor. Hier wird bei der größten Gruppe, die alle Objekte enthält, gestartet und über das Verfahren wird diese in kleinere Gruppen aufgeteilt.
Nun werden in jedem Schritt jedes Cluster in zwei neue Cluster zerteilt. Das Verfahren kann beendet werden, wenn alle Cluster eine bestimmte Distanz zueinander unterschreiten oder wenn eine festgelegte Anzahl von Clustern erreicht wurde.

Bei beiden Verfahren des hierarchischen Clustern entstehen Baumstrukturen, die mit Hilfe von einem Dendrogramm visualisiert werden können. In diesem wird auch noch einmal die unterschiedliche Vorgehensweise beider Verfahren deutlich.

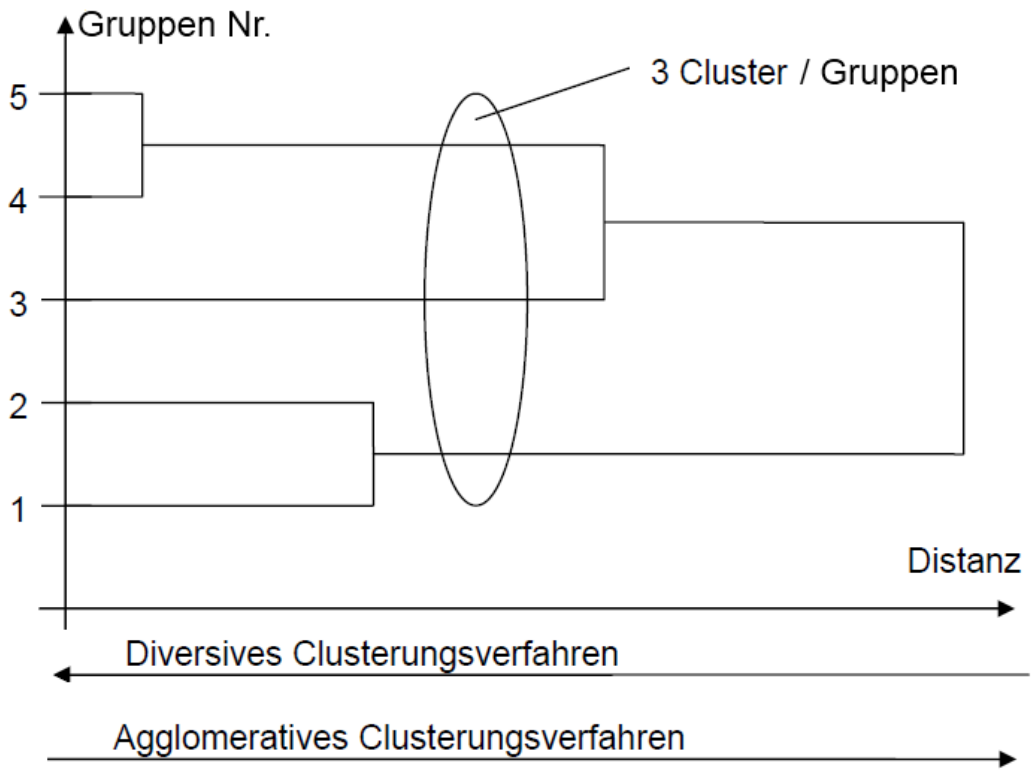


Abbildung 9: Darstellung der Vorgehensweise von diversiven- und agglomerativen Clusterverfahren in einen Dendrogramm

3 Anwendung der Data Mining Verfahren in der Touristikbranche

Allgemeines Vorgehen

Als Software für die Implementierung der Data Mininglösung wird das Open Source Tool Rapid Miner (http://www.rapid-i.com) verwendet.
Der RapidMiner ist eine Entwicklungsumgebung für das Erstellen von Modellen zum Data Ming und maschinellen Lernen. Es besitz eine graphische Benutzeroberfläche und mehr als 500 Operatoren die den gesamten Prozess des Data Minings abdecken. Vom Laden der Daten aus verschiedenen Quellen, über die Transformation der Daten bis hin zur Anwendung von verschiedenen Data Mining Algorithmen und deren Evaluation.

Die Touristikdaten, die für das Trainieren der Data Manning Modelle verwendet wurden, stammen von der Seaboard Business Intelligence GmbH aus Hamburg. Diese wurden aus rechtlichen Gründen so verändert, dass sie keinen originalen Buchungen oder Kundendatensätzen zugeordnet wurden können. Diese Veränderung hat einen gewissen Einfluss auf die Ergebnisse der einzelnen Data Mining Anwendungen gehabt, da durch die Veränderung Regeln und Muster erkannt wurden, die sich nicht mit realen Beobachtungen erklären lassen.

In den folgenden Abschnitten werden Problemstellungen aus der Touristikbranche mit Hilfe von Data Mining Verfahren bearbeitet.
Dabei werden die einzelnen Problemstellungen wie folgt abgearbeitet:

- Erklärung der Problemstellung
- Lösungsvorschlag
- Beschreibung der Daten und Merkmale die als Input für die Data Ming Verfahren verwendet werden
- Parametrisierung des Algorithmus der zur Bearbeitung der Problemstellung verwendet wird
- Beschreibung der Ergebnis die geliefert werden
- Ableitung des Nutzens der Ergebnisse

3.1 Assoziationsanalyse

3.1.1 Problemstellung

Ein Reiseveranstalter für Städtereisen zu Städten auf der ganzen Welt möchte seinen Kundenstamm und den dazugehörigen Buchungen auf Assoziationsregeln untersuchen. Das Ziel ist es die Korrelation der einzelnen Reiseziele zu untersuchen und dadurch gezielteres und kundenindividuelles Marketing durchzuführen.

3.1.2 Lösung

Mit Hilfe von Assoziationsverfahren sollen die Ermittlung von Items A (z.B. Berlin) gelingen, die das Auftreten anderer Items B(z.B. Paris) innerhalb einer Transaktion implizieren. Eine aufgedeckte Beziehung zwischen zwei oder mehr Items soll dann als Regel der Form „Wenn Item A, dann Item(menge) B“ bzw. $A \rightarrow B$ dargestellt werden.

Zur Assoziationsanalyse werden Daten benötigt, bei denen die einzelnen Transaktionen eine Reihe von Items enthalten, die auch in anderen Transaktionen vorhanden sind. Im vorliegenden Anwendungsfall wird dabei der einzelne Kunde als Transaktion betrachtet und seine jeweiligen Buchungen als Items. Dabei wird bei den einzelnen Buchungen nicht betrachtet in welcher Reihenfolge sie gebucht wurden, wie lange die Reisedauer war oder für wie viele Reisende gebucht wurde. Bei jeder Buchung sind ausschließlich das Reiseziel und die Zuordnung zu einem Kunden von Bedeutung.

3.1.3 Daten

Als Inputdaten dienen 52761 Buchungen, denen 17412 Kunden zugeordnet werden können. Dadurch hat jeder Kunde im Schnitt 3 Buchungen getätigt. Die Buchungen sind verteilt auf insgesamt 32 verschiedene Städte. Jede Stadt ist ein Item. Wenn man nun die maximale Anzahl von Items in einem Itemset auf 4 begrenzt, ergeben sich daraus 1.585.080 Kombinationsmöglichkeiten. Zu jeder müsste der Support und die Konfidenz ermitteln werden, um aussagekräftige Assoziationsregeln ableiten zu können. Da zur Assoziation der FP-Growth Algorithmus gewählt wird, wird nur ein Bruchteil dieser Kombinationen untersucht und dennoch ein gutes Ergebnis geliefert.

Damit das Ergebnis nicht durch bestimmte Datensätze verfälsch wird, wurden folgende Daten von der Analyse ausgeschlossen:

- Datensätze die eindeutig Geschäftsreisenden zugeordnet werden konnten, da diese meist andere Interessen bei der Wahl ihres Reiseziels verfolgenden als normale Urlauber.

- Kunden, die noch keine Buchung vorgenommen haben, aber im Kundestamm enthalten sind. Wenn man diese Kunden auch betrachten würde, hätte dies Einfluss auf den Support der einzelnen Reisekombinationen, der sinken würde. Dies mindert die Aussagekraft des Modells.

- Buchungen, die storniert wurden.

- Alle Datenätze, die auf Grund der Datenqualität Fehler enthalten. Dazu gehören Buchungen ohne Zuordnung zu einem Kunden oder ohne Angabe des Reiseziels.

Damit Kundendatensätze, die doppelt vorhanden sind oder von Kunden stammen, die zwischen den Buchungen ihre Adresse oder ihren Nachnamen geändert haben, nicht als unterschiedliche Datensätze behandelt werden, wurde eine Dublettensuche vorgenommen. Dabei wurde die erkannten Dubletten jeweils durch einen neuen Kundendatensatz ersetzt und alle Buchung der Dubletten diesem Datensatz zugeordnet.

3.1.4 Parametrisierung

Folgende Parametereinstellungen müssen beim FP-Growth Assoziationsalgorithmus vorgenommen werden:

Parameter	Beschreibung	Auswirkung	Fest-legung
Maximum Itemset Count	Anzahl der Maximalen Itemset (Gruppierungen von Städten) die erstellt werden soll	Grenzt die Ergebnismenge ein	1000
Maximum Itemset Size	Maximale Anzahl der Items, die in einem Itemset vorkommen sollen	Hat erheblichen Einfluss auf die Rechenzeit, da mit jeder Erhöhung der Anzahl *k* Items in einem Itemset die Anzahl der Kombinations-möglichkeiten *x* bei *n* Items um das *n-k* Fache steigt. $x = \frac{n!}{(n-k)!}$	3
Maximum Support	Maximaler Support, den ein einzelnes Itemset haben soll	Durch Erhöhung dieses Wertes kann die Ergebnismenge eingegrenzt werden und es werden Einelementige Itemsets mit sehr hohem Support ausgeschlossen.	1
Minimum Itemsize	Minimale Anzahl der Items, die in einem Itemset vorhanden sein sollen	Schränkt die Anzahl der Kombinationen in einem Itemset ein und minimiert damit die Rechenzeit	2,

Minimum Probabilty	Die Mindestwahrscheinlichkeit, die eine Regel aufweisen muss, als Regel ausgewiesen zu werden.	Eine zu geringe Wahrscheinlichkeit erhöht den Zufall einer Buchungskombination, wodurch diese nicht mehr entscheidungsrelevant ist.	0.7
Minimum Support	Gibt an, wie groß der Support eines Itemsets sein muss, damit die Konfidenz dieses Itemsets berechnet wird.	Durch Erhöhung dieses Wertes kann die Ergebnismenge eingegrenzt werden. Ein zu niedrig gewählter Wert liefert zu viele irrelevante Ergebnisse.	10

Abbildung 10: Parametereinstellungen des FP-Growth Algorithmus

3.1.5 Ergebnis

Das Ergebnis der Assoziationsanalyse kann in drei Sichten beurteilt werden. Dies sind die Sichten „Itemset“, „Regeln“ und „Abhängigkeitsnetzwerk“.

Itemset

Die Itemsets, die in Tabelle 2 dargestellt sind, listeten nur einen Ausschnitt aller Buchungskombinationen, die gefunden wurden und die Parametereinstellung, hinsichtlich max. Support und Itemsetgröße, erfüllen.

Support	Größe	Itemsets
1146	2	Marrakech, Prag
1069	2	Hamburg, Prag
708	2	Buenos Aires, Sydney
656	2	Rimini, Orlando
422	2	Buenos Aires, Prag
422	3	Marrakech, Buenos Aires, Prag
364	2	New York City, Barcelona
346	2	Amsterdam, Buenos Aires
339	2	Athen, Sydney
333	3	Istanbul, Hamburg, Prag

Tabelle 2: Itemset

Insgesamt wurden 739 Itemsets mit einem Support von 1146 bis 13 gefunden. Um dieses Ergebnis einzuschränken kann ein Minimum Supports festgelegt werden. Zur Bestimmung dieses Minimum Supports kann das Pareto-Prinzip angewendet werden. Dieses sagt aus, dass oft ein kleiner Teil der Beobachtungsmenge einen großen Einfluss auf das Ergebnis hat und umgekehrt. Dabei hat sich in einer Vielzahl von Situationen ein Verhältnis von 80:20 eingestellt, d.h. 20% der Beobachtungen tragen zu 80% des Ergebnisses bei. Durch die Forderung eines Minimum Supports von 95 wird die Anzahl der Itemsets auf 20% der Ausgangsmenge verkleinert. Diese restlichen 147 Ergebnisse könnten, nach dem Pareto-Prinzip, sehr lohnend sein, näher betrachtet zu werden.

Der Support gibt an, wie häufig die betreffende Buchungskombination in der Gesamtheit der Kundensätze gefunden wurde. Es haben also schon 1146 Kunden jeweils eine Reise nach Marrakech und Prag unternommen, wobei die Reihenfolge der Buchungen nicht entscheidend ist. Je höher dieser Wert ist desto relevanter ist diese Buchungskombination. Die Größe gibt an aus wie vielen Items das Itemset besteht. Dieser wert kann bei der Parametrisierung eingegrenzt werden. Ein Itemset der Größe „1“ steht für eine einzelne Buchung, die ein Kunde vorgenommen hat. Da es aber um die Suche von Gruppen geht, sind diese Ergebnisse oft nicht relevant. Andererseits sind zu groß gewählte Itemset auch irrelevant, da ihr Support sehr gering ausfällt.

Das Itemset selbst ist die Beschreibung der Einzelbuchungen, die in dem Itemset vorhanden sind. Diese Liste kann verwendet werden, um Marketingaktionen zu planen oder Reisekataloge zu gestalten, da sie Ausschluss darüber gibt welche Buchungen von Kunden oft im Zusammenhang mit anderen Buchungen des Kunden stehen.

Regeln

Neben der Gruppierung von Objekten zu Gruppen, den Itemsets, können Assoziationsverfahren auch Regeln Erarbeiten. Dies sind Abhängigkeitsregeln die als „Wenn-dann“-Regeln formuliert werden. In Tabelle 3 ist ein Ausschnitt der Regeln die das Assoziationsverfahren abgeleitet hat.

Wahrscheinlichkeit	Konfidenz	Regel
0.92	1.53	Berlin, Barcelona -> New York City
0.85	1.46	Side, Orlando -> Rimini
0.84	1.46	Berlin, Barcelona, Sydney -> New York City
0.82	1.45	Berlin, Barcelona, Peking -> New York City
0.82	1.45	Paris, San Francisco -> München
0.76	1.44	Venedig, Barcelona -> Rom
0.74	1.42	London, Orlando -> Rimini

Tabelle 3: Regel-Ansicht der Assoziationsanalyse

Die erste Spalte gibt die Wahrscheinlichkeit an mit der die jeweilige Regel zutrifft. Im vorliegenden Anwendungsfall bedeutet dies z.B., dass, wenn ein Kunde bereits eine Reise nach

Berlin und Barcelona unternommen hat, er auch mit der Wahrscheinlichkeit von 92 % eine Reise nach New York City gebucht hat. Die Zweite Spalte Konfidenz ist ein Maß für die Qualität der Regel. Je höher dieser Wert ist desto besser ist die Qualität der Regel. Diese Spalte relativiert das Maß der Wahrscheinlichkeit aus der ersten Spalte. Ideal ist daher eine Regel mit hoher Wahrscheinlichkeit und hoher Konfidenz. In der dritten Spalte Regel wird die „Wenn-dann"-Beziehungen dargestellt. Eine Regel kann wie folgt formuliert werden: „Wenn eine Reise nach Venedig und Barcelona gebucht wurde, dann wurde auch eine Reise nach Rom gebucht."

Insgesamt wurden 476 Regeln gefunden. Auch hier lohnt es sich wieder das Pareto-Prinzip anzuwenden, um die Ergebnismenge überschaubar einzugrenzen. Es wird also eine minimale Konfidenz gewählt bei der 20% der Ursprungsregeln überbleibt. Bei einer Konfidenz von min. 1.08 bleiben 97 Regeln übrig, die näher betrachtet werden sollten.

Dieses Wissen könnte nun verwendet werden, um automatisiert Prospekte an Kunden zu versenden, für die die entsprechende Regel zutrifft. Alle Kunden die z.B. schon in Berlin und Barcelona waren, aber noch nicht in New York, würden Prospekte für eine Reise nach New York bekommen. Da diesen Kunden ein hohes Potenzial haben auf dieses Prospekt zu reagieren.

Abhängigkeitsnetzwerk

Das Abhängigkeitsnetzwerk in Abbildung 11 soll die Stärke und den Zusammenhang zwischen verschieden Reisezielen darstellen.

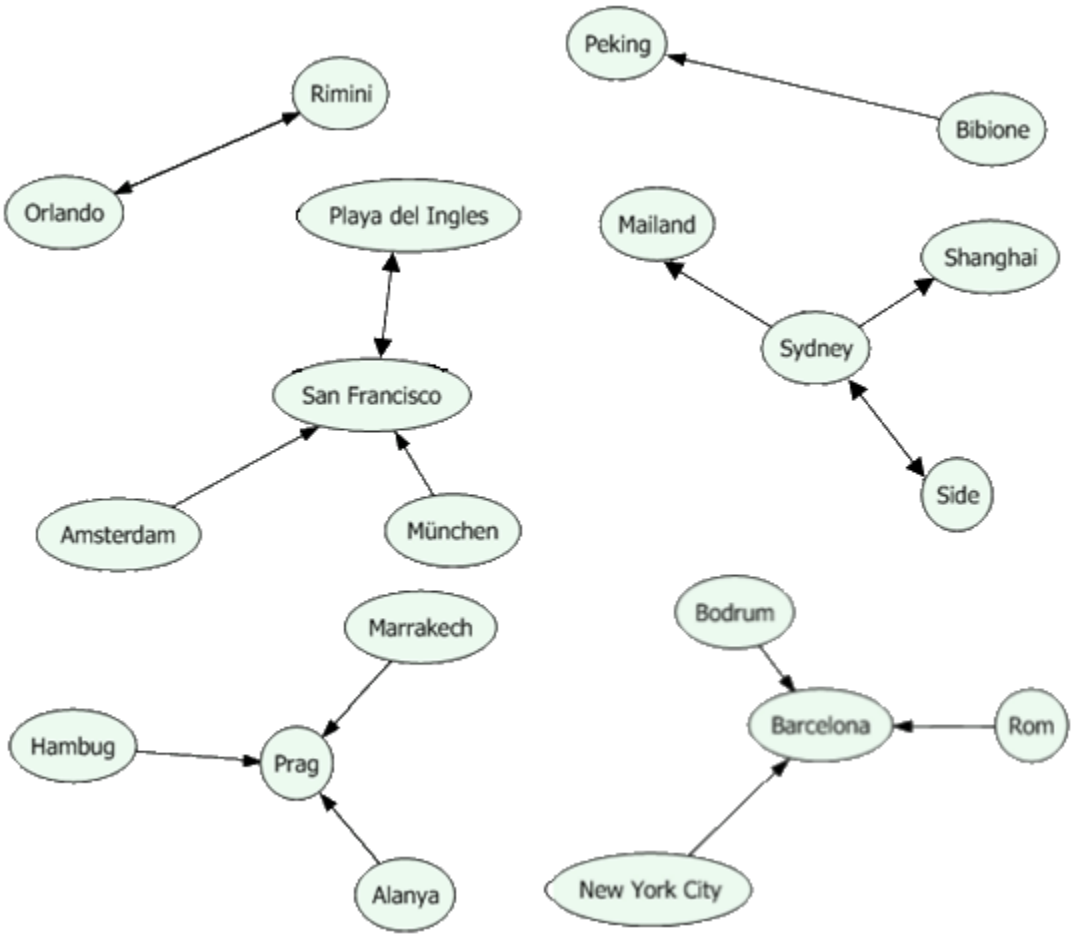


Abbildung 11: Abhängigkeitsnetzwerk

Es sind 6 Netzwerkgruppen zu erkennen, die isoliert voneinander sind. Die Verbindungslinien zwischen den einzelnen Elementen zeigt die Relation an. So besteht z.B. zwischen Rimini und Orlando eine bidirektionale Relation. Wer also eine Reise nach Rimini gebucht hat, hat auch eine Reise nach Orlando gebucht und umgekehrt. Zwischen Peking und Bibione hingen besteht nur eine einseitige Verbindung. Wer eine Reise nach Bibione gebucht hat, hat auch eine Reise nach Peking gebucht, dies gilt aber nicht umgekehrt.

Im Abhängigkeitsnetzwerk lassen sich auch wieder Regeln ableiten. So lassen sich einfache Regeln über eine Relation ableiten, wie „Wenn Rom gebucht wurde, dann wurde auch Barcelona gebucht“. Es können aber auch komplexere Regeln, über mehre Relationen beschrieben werden,

z.B. „ Wenn Side *und* Sydney gebucht wurden, dann wurde auch Mailand gebucht.“ In der Itemansicht würde diese Regel mit der Größe 3 erscheinen.

3.1.6 Nutzen der Assoziationsanalyse

Mit diesem Wissen und der Analyse der eigenen Kundendaten kann man nun gezielte und Kundenindividuelle Marketingkampagnen starten. Man ist in der Lage, seinen Kunden Reisen anzubieten, die er auf Grundlage seiner bisherigen Reisen und dem Wissen welche Reisen andere Kunden unternommen haben, bestimmt werden.

Zwar liefern sie dem Anwender durchaus auch Regeln, die er schon kennt, wie z.B. das Kunden, die schon nach Rom und Rimini unternommen haben, auch eine Reise nach Venedig unternehmen. Da muss aber kein Nachteil, sondern eher eine Art vertrauensbildende Maßnahme, da der Anwender hierdurch sieht, dass die Ergebnisse einer derartigen Analyse sich in manchen Fällen mit seinem Wissensstand decken. Kommen nun noch neue Erkenntnisse hinzu, dann kann dies zumindest nicht schaden. Man muss hier auch den eigentlichen Vorteil sehen, der darin besteht, dass nun Regeln automatisch erzeugt werden können, was einerseits den Anwender von aufwendigen Untersuchungen befreit, andererseits die Regeln weitaus schneller erzeugt werden können, als es ein Anwender jemals könnte. Mit Hilfe diese neuen Regeln können Entscheidungen besser und schneller getroffen werden. Da das vorhandene priori Wissen durch fundierte Regeln aus Assoziationsmethoden unterstützt wird.

3.2 Klassifikationsanalyse

3.2.1 Aufgabenstellung

Die Marketingabteilung eines Reiseveranstalters möchte potenzielle Kunden durch den Versand von Prospekten für eine Wellness-Reise anwerben. Dabei sollen die Kosten, die durch den Druck und Versand der Prospekte entstehen, reduziert werden. Deshalb sollen Prospekte nur an jene Kunden gesendet werden, die mit höherer Wahrscheinlichkeit auf diese reagieren werden.
Die einfachste Variante, ohne Anwendung von Data Mining, wäre das Versenden der Prospekte nur an Kunden, die bereits eine entsprechende Reise unternommen haben. Dadurch werden aber nicht alle potenziellen Kunden erfasst. Mit Data Mining wird nun versucht, auch die Kunden im vorhandenen Kundenstamm zu identifizieren, die noch keine Wellness-Reise unternommen haben, aber trotzdem mit höherer Wahrscheinlichkeit, als andere Kunden daran interessiert sein könnten.

3.2.2 Lösung

Zur Lösung der Aufgabenstellung werden die vorhandenen Kundendaten mit dem Entscheidungsbaum Klassifikationsverfahren untersucht. Dabei werden die Merkmale der potenziellen Kunden, die noch keinen Wellness-Reise unternommen haben, verglichen mit den Merkmalen der Kunden, die bereits eine Wellness-Reise gebucht haben. Dabei sollen die Ausprägungen der Merkmale identifiziert werden, die einen Kunden als möglichen Interessenten einer Wellness-Reise klassifizieren. Das Ergebnis des Klassifikationsverfahrens wird in einem Entscheidungsbaum mit den einzelnen Merkmalen als Knoten und der Ausprägung dieser als Gewicht an den Verbindungslinien dargestellt. Die Blätter stehen für die zugeordnete Klasse.

3.2.3 Daten

Tabelle 4 zeigt die Merkmale die zur Klassifikation der Datensätze verwendet wurden. Die ersten 6 Spalten dienen zur Klassifikation und die letzte Spalte ist die Zielspalte, die später verwendet wird um zu entscheiden, ob ein Prospekt an den Kunden versendet wird oder nicht. Diese Spalte muss im Voraus der Klassifikation für jeden vorhandenen Datensatz ermittelt werden.

Merkmal	Typ	Beschreibung
Kaufkraft	int	wird ermittelt durch die Anreicherung der Kundendaten mit externen Daten von Marktforschungsunternehmen, wie der Gesellschaft für Konsumforschung(GfK), über den Abgleich der PLZ
Gesamtaufenthalt	int	Reisedauer in Tagen
Anzahl Erwachsene	int	
Anzahl Kinder	int	
Durchschnittlicher Vorbuchungszeit	int	Ist die Differenz zwischen Buchungsdatum und Reisedatum
Bundesland	varchar	Demographischer Merkmalstyp
Wellness-Reise	bool	Wellness-Reise ist sowohl Eingabe- als auch als Vorhersagespalte

Tabelle 4: Merkmale die zur Klassifikation dienen sollen

Bevor die Merkmale als Input für den Algorithmus genutzt werden, werden alle Merkmale vom Typ int diskretisiert. Das heißt, dass die kontinuierlichen Werte in diskrete Teilmengen eingeteilt werden. Die diskreten Teilmengen dienen später als Gewichte im Entscheidungsbaum.
In Tabelle 5 wird ein Ausschnitt der Inputdaten dargestellt, der bereits diskretisiert wurden. Insgesamt umfasst der Datensatz 5527 einzelne Datensätze. Davon werden 3500 zum Trainieren des Data Mining Models verwendet und die restlichen Daten zur Evaluierung des Models anderen Modellen, die mit unterschiedlichen Parametern trainiert wurden.

Wellness-Reise	Anzahl Erwachsene	Anzahl Kinder	Vorausbuchungs-zeit	Aufenthalts-dauer	Kaufkraft	Bundesland
false	r1 [-∞ - 2]	r2 [1 - 2]	r1 [-∞ - 235]	r3 [10 - 14]	r1 [-∞ - 17,986]	Thüringen
true	r1 [-∞ - 2]	r1 [-∞ - 1]	r1 [-∞ - 235]	r2 [5 - 10]	r2 [17,986 - 22,261]	Bayern
true	r1 [-∞ - 2]	r1 [-∞ - 1]	r1 [-∞ - 235]	r3 [10 - 14]	r1 [-∞ - 17,986]	Thüringen
false	r3 [3 - 4]	r4 [3 - ∞]	r1 [-∞ - 235]	r1 [-∞ - 5]	r1 [-∞ - 17,986]	Thüringen
true	r1 [-∞ - 2]	r1 [-∞ - 1]	r1 [-∞ - 235]	r3 [10 - 14]	r1 [-∞ - 17,986]	Sachsen-Anhalt
false	r1 [-∞ - 2]	r1 [-∞ - 1]	r1 [-∞ - 235]	r2 [5 - 10]	r2 [17,986 - 22,261]	Nieder-Sachsen
false	r1 [-∞ - 2]	r3 [2 - 3]	r1 [-∞ - 235]	r3 [10 - 14]	r1 [-∞ - 17,986]	Thüringen
false	r1 [-∞ - 2]	r1 [-∞ - 1]	r1 [-∞ - 235]	r2 [5 - 10]	r2 [17,986 - 22,261]	Bayern
false	r1 [-∞ - 2]	r1 [-∞ - 1]	r1 [-∞ - 235]	r1 [-∞ - 5]	r2 [17,986 - 22,261]	Bayern
false	r1 [-∞ - 2]	r1 [-∞ - 1]	r1 [-∞ - 235]	r2 [5 - 10]	r2 [17,986 - 22,261]	Bayern

Tabelle 5: Beispieldatensatz des Inputs

3.2.4 Parametrisierung

Folgende Parameter müssen beim Entscheidungsbaum Algorithmus festgelegt werden

Parameter	Beschreibung	Einfluss	Einstellung
criterion	Festlegung des Kriteriums, nach dem die Splitattribute bestimmt werden.		
minimal size for split	Die minimale Anzahl von Splitattributen im Baum	Erzwingt eine bestimmte Anzahl von Splits	4
minimal leaf size	Minimale Anzahl von Blättern, die ein Knoten haben kann	Erzwingt eine bestimmte Anzahl von Blättern	2

minimal gain	Minimaler Zuwachs der Genauigkeit die erreicht werden muss um einen Split zu erzeugen		0.1
maximum depth	Maximale Tiefe des Baums	Hat Einfluss auf die Größe des Baums und wie genau die einzelnen Klassifikationen abgebildet werden	20
confidence	Angabe des Konfidenzniveau für die Berechnung des Fehler der zum Optimieren des Baums genutzt wird		0.25
number of pruning alternatives	Anzahl der alternativen Blätter, die erzeugt werden, wenn durch das Prunning (Beschneidung) der Split unterbunden wird	Irrelevante Splitknoten werden durch zusätzliche Blätter ersetzt	3
no pruning	Festlegung, ob der aufgestellte Baum optimiert werden soll	Führt zur Entfernung von irrelevanten Ästen	false

Tabelle 6: Paramter des Entscheidungsbaum Algorithmuses

3.2.5 Ergebnisse

Im Entscheidungsbaum in Abbildung 12 wird dargestellt, welche Merkmale ein Kunde erfüllen muss, um zu einem Blatt zu gelangen, das als True deklariert ist und ihn als möglichen Interessenten an eine Wellness-Reise klassifiziert.

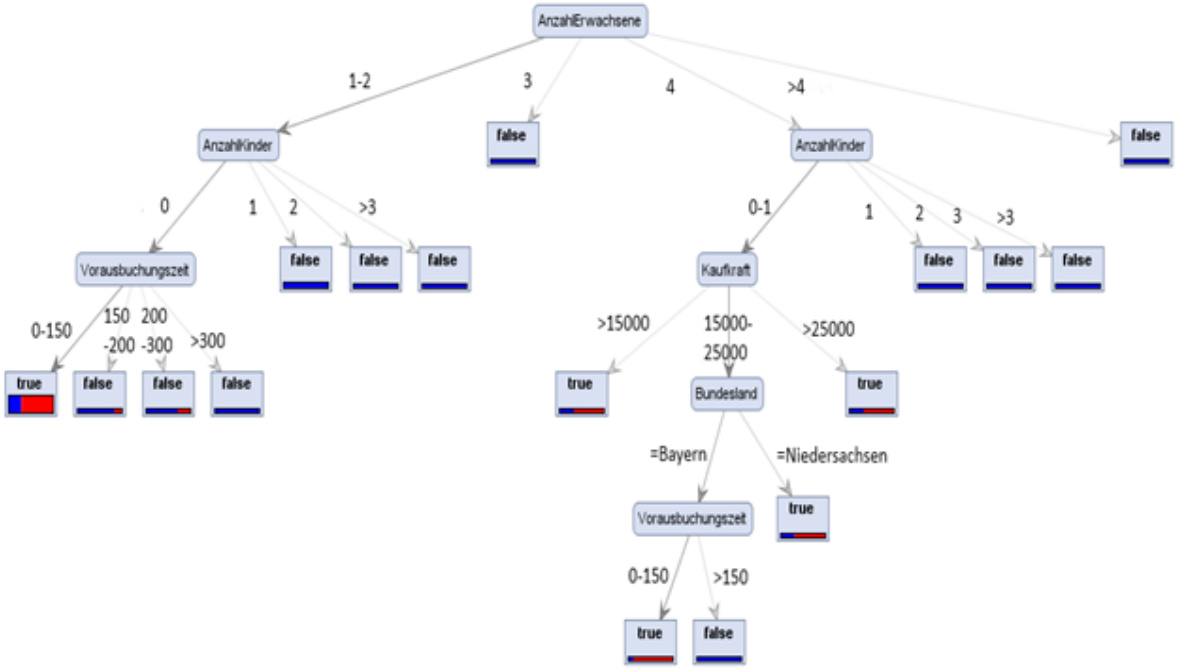


Abbildung 12: Entscheidungsbaum

Ein potenzieller Kunde wäre z.B. eine Kunde, der in der Vergangenheit mit 1-2 Erwachsenen, keinem Kind und eine Vorausbuchungszeit von <150 Tagen eine Reise gebucht hat, weil diese Merkmalskombinationen einen Pfad zwischen der Wurzel und einem Blatt mit dem Wert true darstellt.

Die einzelnen Merkmale sind nach dem Einfluss auf die Klassifizierung angeordnet. So hat das Merkmal „Anzahl Erwachsene“ den größten Einfluss und die „Vorausbuchungszeit“ den geringsten Einfluss.

Kreuzvalidierung

Die Performance des Entscheidungsbaums wird mit Hilfe der Kreuzvalidierung überprüft. Als Input werden 5000 Datensätze verwendet. Diese werden in 5 gleichgroße Teildatensätze geteilt. Nun wird jeweils 1 Datensatz zum Trainieren und 4 zum Testen des Models genutzt. Dieser Vorgang wird so oft wiederholt, bis alle Teildatensätze einmal zum Trainieren

genutzt wurden. Nach jedem Trainieren eines Modells wird dies mit den Testdaten untersucht. Dabei soll die Klassezuordnung, die in den Testdaten bekannt ist, vom Modell prognostiziert werden. Das Ergebnis jeder einzelnen Prognose kann einem der folgenden 4 Tupel zugeordnet werden:

- es wird prognostiziert das ein Kunde keine Wellness-Reise bucht, was mit der Angebe in den Testdatensatz übereinstimmt= pred. False/false (richtig negativ)
- es wird prognostiziert das ein Kunde keine Wellness-Reise bucht, obwohl er eine gebucht hat = pred. False/true (falsch negativ)
- es wird prognostiziert das ein Kunde eine Wellness-Reise bucht, obwohl er keine gebucht hat = pred. true /false(falsch positiv)
- es wird prognostiziert das ein Kunde eine Wellness-Reise bucht, was zutrifft = pred. true /true(richtig positiv)

Bei jeder Zuordnung zu einem Tupel wird dies um 1 erhöht. Das Ergebnis der Kreuzvalidierung ist in Tabelle 7 dargestellt. Insgesamt wurde der Entscheidungsbaum mit 20000 Datensäte getestet, wobei jeweils Test und Trainingsmenge unabhängig voneinander waren. Anhand dieser Daten kann die Genauigkeit (Precision), die Trefferquote (Recall) und Ausfallquote berechnet werden. Dabei haben die einzelnen Werte folgende Bedeutung:

- Die Genauigkeit entspricht der statistischen Relevanz der gefundenen Treffer
- Die Trefferquote entspricht der statistischen Sensitivität des Modells, gibt also den Anteil der richtig als positiv erkannten Objekte an der Gesamtheit der in Wirklichkeit positiven Sachverhalte an.
- Die Ausfallquote entspricht der statistischen *false positive rate* (FPR) und ist das Gegenteil der Spezifität ($1 - FPR$). Durch sie wird angegeben, wie hoch der Anteil der Objekte ist, die als wahr ausgegeben werden obwohl sie negativ sind.

	false false	true true
pred. false	3250	450
pred. true	3050	13250

Tabelle 7: Ergebnis der Kreuzvalidierung

Durch die Auswertung der Kreuzvalidierung können folgende Aussagen über die Performance des Entscheidungsbaums getroffen werden. Er besitzt eine Genauigkeit von 81% bei einer Trefferquote von 96% und einer Ausfallrate von 48%.

Receiver-Operating-Characteristic

In Abbildung 13 werden die Ergebnisse von mehreren Entscheidungsbaum-Algorithmen mit jeweils anderen Parametrisierungen visuell verglichen. Die y-Achse steht für die relative Anzahl der falsch klassifizierten Werte und die x-Achse für die relative Anzahl der richtig klassifizierten Werte. Das heißt, dass der optimale Punkt in der oberen linken Ecke liegt. Der Algorithmus, der sich diesem Punkt am weitesten nähert ist, am effizientesten bei der Klassifizierung von Objekten. Der in diesen Abwendungsfall verwendet Algorithmus wird rot dargestellt und macht nach Angabe des Receiver-Operating-Characteristic Diagramms bei der Klassifizierung weniger Fehler als die anderen 4 Algorithmen.

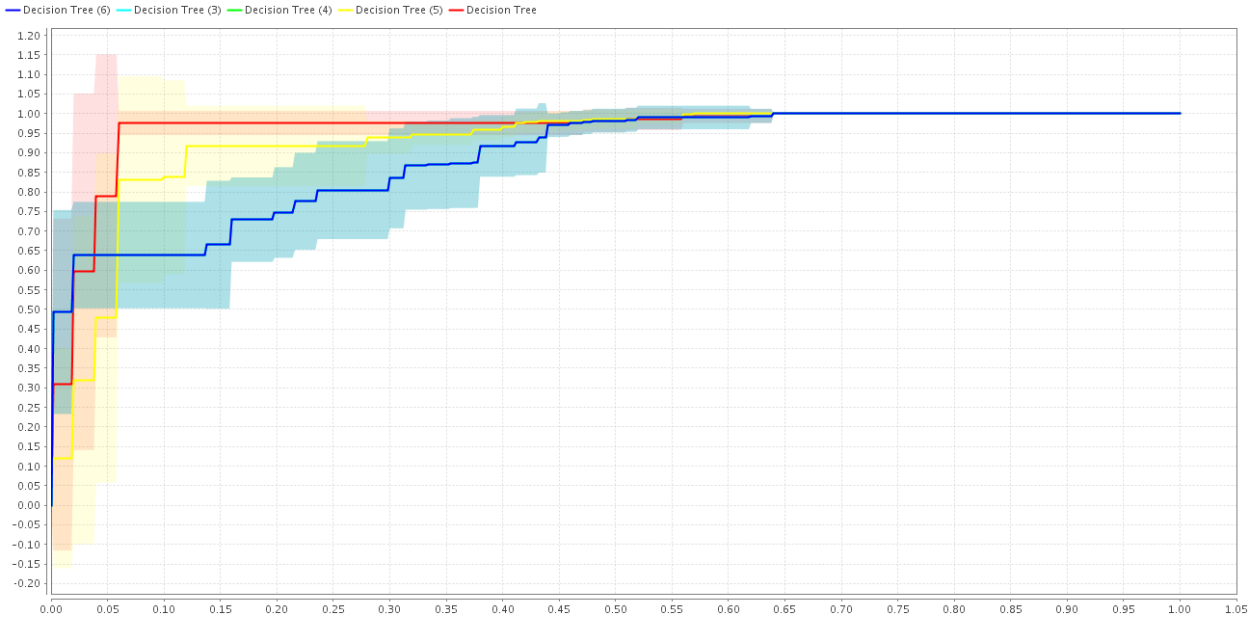


Abbildung 13: Vergleich mehrere Entscheidungsbaumalgorithmen in einem Receiver-Operating-Characteristic

3.2.6 Nutzen

Im vorliegenden Anwendungsfall wurde nach Kunden gesucht, die sich ideal für das Marketing von Wellness-Reisen eigenen. Man kann aber auch ohne großen Aufwand potenzielle Kunden für alle anderen möglichen Arten von Reisen finden. So wäre es in der Praxis möglich, in kürzester Zeit den Erfolg von möglichen Marketingkampanien zu steigern und gleichzeitig die Kosten zu senken. Erreicht wird dies durch gezieltes Marketing.

3.3 Clusteranalyse

3.3.1 Problemstellung

Um Marketingaktionen gezielter und kundenindividueller planen zu können, bedarf es der genauen Kenntnis über die Zielgruppen, die man ansprechen möchte.
Ein Reiseveranstalter möchte seinen vorhandenen Kundenstamm analysieren und dabei marketingrelevante Zielgruppen ableiten. Diese sollen systematisch bearbeitet werden, um jeder Zielgruppen individuelle Reisen anbieten zu können. Dadurch soll eine dauerhafte Kundenbindung hergestellt werden.

3.3.2 Lösung

Um Cluster oder Kundengruppen zu entdecken, wird der K-Means Cluster Algorithmus verwendet.
Das Ziel der Kundensegmentierung soll es sein, den vorhandenen Kundenbestand in marketingrelevante Zielgruppen zu unterteilen. Man geht dabei davon aus, dass Kunden mit ähnlichen Merkmalen auch ähnliche Bedürfnisse haben und sich ähnlich verhalten. Zu diesem Zweck wird der heterogene Kundenstamm, der traditionell nur in seiner Gesamtheit betrachtet wird, mit Hilfe von Clusterverfahren in Gruppen, den Clustern, unterteilt. Idealerweise sollten dabei die gebildeten Gruppen in sich möglichst homogen sein, aber untereinander wenig gemeinsame Merkmale aufweisen. Dadurch ist man in der Lage, für jede Kundengruppe unter Nutzung aller Marketing- und Vertriebsinstrumente – also Produktgestaltung, Kommunikation, Preisgestaltung sowie Service – ein spezielles Konzept zu entwerfen.

3.3.3 Daten

Im ersten Schritt der Kundensegmentierung ist zu überlegen, welche Merkmale zur Differenzierung der Kunden geeignet sind. Dabei sollten die einzelnen Attribute möglichst keine Redundanten Informationen enthalten und unabhängig sein. Folgende Merkmalstypen können verwendet werden:

- demografische Merkmale (Familienstand, Alter, Geschlecht,)
- soziografische Merkmale (Kaufkraft, Beruf, Wohnverhältnisse)
- psychografische Merkmale (Glaubenssätze, Religionsangehörigkeit, Verhalten, Lebensstil)
- regionale Merkmale (Wohnort, Wohngegend),

- verhaltensorientierte Merkmale (Kommunikationskanäle, Preisorientierung, Konsumverhalten)

Als Input für den k-Means Algorithmus werden 4300 Kundensätze verwendet. Jeder Kundendatensatz besitzt die Merkmale wie in Tabelle 8.

Merkmal	Typ	Beschreibung
Kaufkraft	int	soziografische Merkmale wird ermittelt durch die Anreicherung der Kundendaten mit externen Daten von Marktforschungsunternehmen, wie der Gesellschaft für Konsumforschung(GfK), über den Abgleich der PLZ
Alter	int	demografische Merkmale
Geschlecht	bool	demografische Merkmale(true=weiblich, false=männlich)
Anzahl Kinder	int	demografische Merkmale
Familienstaus	bool	psychografische Merkmale (true=verheiratet false=nicht verheiratet)

Tabelle 8 Merkmale zur Clusteranalyse

Die Daten müssen folgende Voraussetzung erfüllen damit eine Clusteranalyse angewendet werden kann:

- Skalenqualität sollte möglichst hoch sein und bei alle Merkmalen gleich
- alle Merkmale, die analysiert werden, müssen messbar sein. Es muss für jedes Merkmale eine Distanz oder Ähnlichkeitsfunktion geben
- Ausreißer sollten im Voraus aus der Trainingsmenge gefiltert werden, da sie erheblichen Einfluss auf das Ergebnis haben können
- Fehlende Werte sollten wenn möglich ergänzt werden (z.B. Ableitung des Geschlecht vom Vornamen)

3.3.4 Parametrisierung

Parameter die bei der Anwendung des K-Means Algorithmus festgelegt werden:

Prater	Beschreibung	Einfluss	Festlegung
add cluster attribute	Festlegung, ob das Cluster zu dem ein Objekt zugeordnet wird, als neues Attribut angelegt werden soll		true
remove unlabeled	Automatisiertes Löschen von fehlerhaften Input Objekten	Fehlerhafte Input-Objekte können das Ergebnis stark beeinflussen und verfälschen	true
k	Anzahl der Cluster die gebildet werden sollen	Muss je nach Anwendungsfall entschieden werden.	4
max optimization steps	Maximale Anzahl der Wiederholungen, in denen der optimale Schwerpunkt der einzelnen Cluster bestimmt werden soll	Hat Einfluss auf die Rechenzeit und die Genauigkeit der Cluster	100
use local random seed	Bestimmt, ob die zufälligen Ausgangs-schwerpunkte(seeds) lokal oder global bestimmt werden sollen	false= höhere Rechenzeit- da Bestimmung der Schwer-punkte global erfolgt true=Verringerung der Rechenzeit, da Schwerpunkte lokal bestimmt werden	false
kernel type	Festlegung der Verteilungsfunktion der einzelnen Objekt zu den Clustern	Bestimmt, wie die Objekte auf die einzelnen Cluster verteil werden.	Gaußverteilung

Tabelle 9: Parameter des K-Means Algorithmus

3.3.5 Ergebnisse

In Abbildung 14 werden die Ergebnisse des Clusterverfahrens das mit dem k-Means Algorithmus durchgeführt wurde in einer Scatter Matrix dargestellt. Wellness-Reise

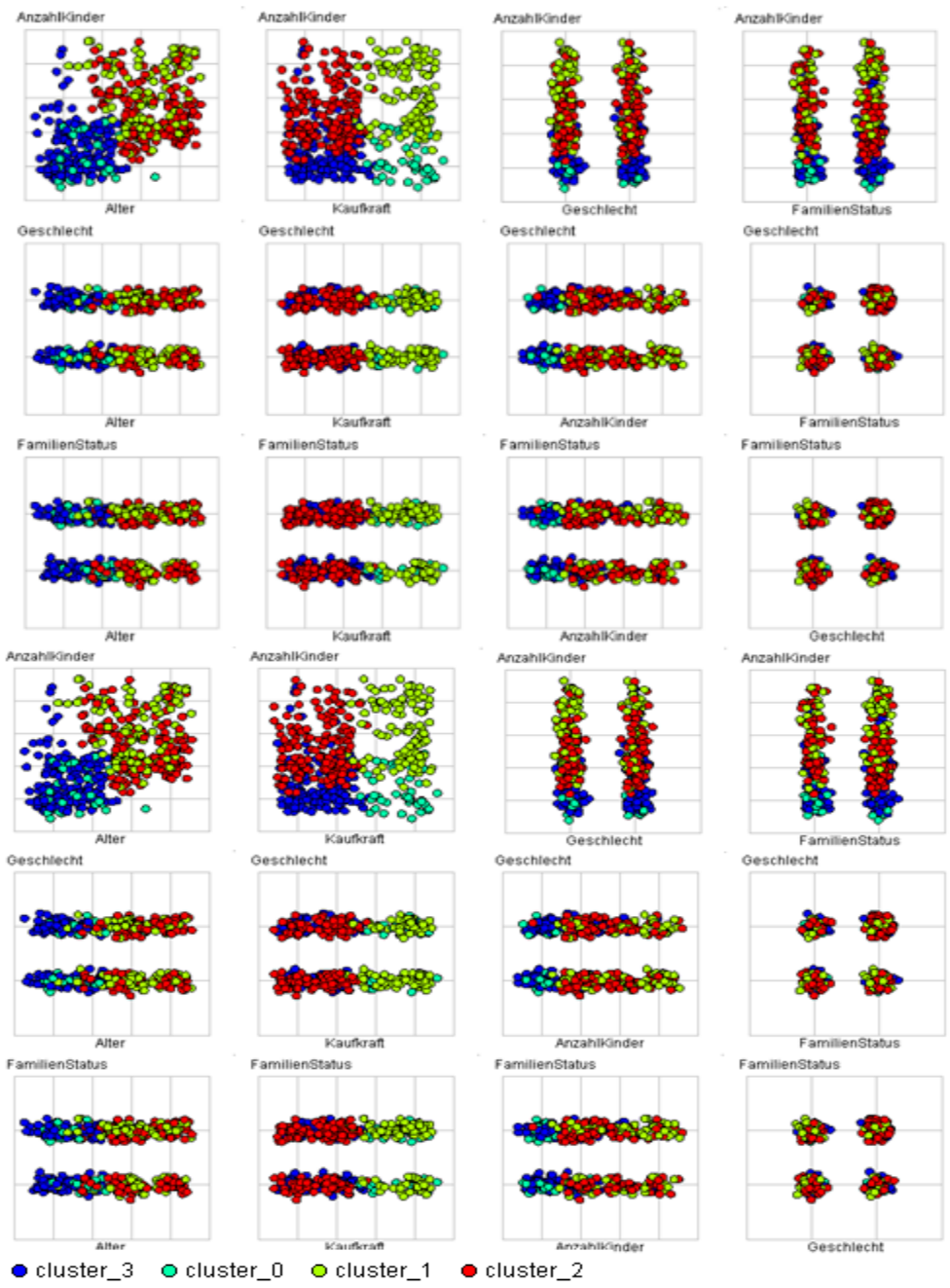


Abbildung 14: Visuelle Darstellung der Clusterergebnisse in einer Scatter Matrix

Jeder Punkt ist ein Informationsobjekt, im vorliegenden Anwendungsfall ein Kunde, der einer Klasse, die durch die Farbe symbolisiert wird, zugeordnet ist. In der Scatter Matrix werden immer 2 Merkmale gegenübergestellt, um deren Einfluss auf die Zuordnung zu einem Cluster zu visualisieren. So kann man erkennen welche Merkmale den größten Einfluss auf die Clusterzuordnung eines Objekts haben.

So kann man aus der Scatterdiagramm Nr.2, in der die Merkmale Anzahl und Kinder und Kaufkraft gegenüber gestellt werden, folgende Rückschlüsse zur Zuordnung eines Clusters ziehen:

- Kunden mit geringer bis mittlerer Kaufkraft und 0 bis 1 Kind werden Cluster 3 (blau) zugeordnet
- Kunden mit geringer bis mittlerer Kaufkraft und mehr als 1 Kind werden Cluster 2 (rot) zugeordnet
- Kunden mit mittlerer bis hoher Kaufkraft und mehr als 1 Kind werden Cluster 1 (grün) zugeordnet

- Kunden mit mittlerer bis hoher Kaufkraft und 0 bis1 Kind werden Cluster 2 (rot) zugeordnet

Diese Analyse kann für alle Scatterdiagramme durchgeführt werden mit dem Ziel, am Ende ein genaueres Verständnis darüber zu haben, was die einzelnen Cluster unterscheidet und was für Kunden den einzelnen Clustern zugeordnet werden.

In Tabelle 10 wurde die Zuordnung von 1000 Datensätzen zu den einzelnen Cluster ausgewertet. Dabei wird jeweils die Anzahl der Elemente erfasst, die in dem zugehören Cluster eine bestimmte Merkmalsausprägung besitzt. Danach wird zu jeder Merkmalsausprägung der prozentuale Anteil der Elemente in einem Cluster zur Gesamtheit aller Objekte, die diese Merkmalsausprägung besitzen, berechnet und farbig markiert. Damit ist man in der Lage, zu jedem Merkmal eine Aussage treffen zu können, wie groß der Einfluss der einzelnen Merkmalsausprägung auf die Zuordnung eines Kunden zu einem Clusters ist.

	Merkmals-ausprägung	Cluster 0		Cluster 1		Cluster 2		Cluster 3	
Merkmal		211		251		269		269	
Alter	[20 - 30]	88	31.43%	187	66.79%	0	0.00%	5	1.79%
	[30 - 37]	123	47.86%	64	24.90%	0	0.00%	70	27.24%
	[37 - 46]	0	0.00%	0	0.00%	141	57.32%	105	42.68%
	[< 46]	0	0.00%	0	0.00%	128	58.99%	89	41.01%
Kaufkraft	[0 - 24,500]	64	22.78%	100	35.59%	117	41.64%	0	0.00%
	[24,500 - 45,500]	137	35.13%	101	25.90%	152	38.97%	0	0.00%
	[45,500 - 52,500]	10	8.33%	25	20.83%	0	0.00%	85	70.83%
	[< 52,500]	0	0.00%	25	11.96%	0	0.00%	184	88.04%
anzahl Kinder	[0 - 1]	0	0.00%	248	92.19%	1	0.37%	20	7.43%
	[1 - 2]	146	35.18%	3	0.72%	164	39.52%	102	24.58%
	[2 - 3]	37	25.00%	0	0.00%	65	43.92%	46	31.08%
	[< 4]	0	0.00%	0	0.00%	39	27.86%	101	72.14%
Geschlecht	weiblich	88	17.60%	129	25.80%	136	27.20%	147	29.40%
	männlich	123	24.60%	122	24.40%	133	26.60%	122	24.40%
Familien-status	Nicht verheiratet	75	17.44%	148	34.42%	99	23.02%	108	25.12%
	verheiratet	136	23.86%	103	18.07%	170	29.82%	161	28.25%

Tabelle 10: Auswertung der Clusteranalyse

So sind z.B. alle Kunden die 0-1 Kind haben mit 92% Wahrscheinlichkeit dem Cluster 1 zugeordnet. Man kann diese Cluster umbenennen und als „Kunden ohne Kinder“ bezeichnen. Mit diesem Wissen kann bei späteren Marketingaktionen, bei denen es darum geht, kinderfreundliche Hotels zu bewerben diese Kundengruppe ausgeschlossen werden.

Die andern Cluster zeichnen sich besonders durch folgende Eigenschaften auf:

- Cluster 0=Kunden zwischen 30-37, mit mittlerer Kaufkraft und 1-2 Kindern
- Cluster 1=Kunden unter 30 ohne Kinder
- Cluster 2=Kunden über 37
- Cluster 4=Kunden mit hoher Kaufkraft und mehr als 4 Kindern

3.3.6 Nutzen der Clusteranalyse

Durch die Segmentierung des Kundenbestands und der individuellen Behandlung der einzelnen Gruppen werden folgende Effekte auf Unternehmens und Kundenseite erreicht:

- Reduzierung der Marketingkosten, durch das gezielte Versenden von Katalogen und Prospekten.
- Erhöhung der Kundenzufriedenheit, dadurch das jedem Kunden das Gefühl vermittelt wird, dass man auf seine spezifischen Eigenschaften und Bedürfnisse eingeht
- Es werden frühzeitig Änderungen in der Zusammensetzung des Kundenbestandes erkannt. Effekte der „Kundenwanderung“ können frühzeitig in die Planung des Reiseangebots einfließen.
- Der Erfolg von Marketingaktionen, die auf eine bestimmte Zielgruppe ausgerichtet sind, wird messbar, da der Zuwachs einer bestimmten Kundengruppe ermittelt werden kann.
- Ermöglicht das Erkennen der Gegensätzlichkeit von Kunden innerhalb eines Kundenbestandes.
- Auf Basis der Kundensegmentierung können differenzierte Kundenstrategien zur Kundenbetreuung entwickelt werden.

4 Schluss

4.1 Zusammenfassung

Ziel der Bachelorarbeit ist es, sich mit den verschiedenen Data Mining Verfahren zu beschäftigen und Anwendungsfälle für die Touristikbranche abzuleiten.

Im ersten Kapitel wurde kurz darauf eingegangen, welche Aufgaben im Zusammenhang mit Data Mining Verfahren betrachtet werden müssen um nützliche Ergebnisse zu gewinnen. Es wurden allgemeine Probleme bei Data Mining Verfahren genannt.

Im darauf folgenden Kapitel wurde auf die Methoden der Cluster-, Klassifikations und Assoziationsverfahren eingegangen. Zu jedem dieser Verfahren wurden die relevanten Algorithmen betrachtet und deren Funktionsweise erläutert.

Im abschließenden Kapitel wurden jeweils einzelne Algorithmen der 3 verschiedene Data Mining Verfahren angewendet, um Problemstellungen aus der Touristikbranche zu bearbeiten.

4.2 Ausblick

Der Anwendungsbereich von Data Mining Verfahren wird künftig weiter zunehmen und den Nutzen und Umgang mit operativen Daten maßgeblich verändern. Die Gründe dieser Bedeutungszunahme beruht auf folgenden Annahmen: Durch die schon in der Vergangenheit immer weiter sinkenden Kosten für Datensammlung und Speicherung, werden die Massen an Daten, die Unternehmen über Kunden und Transaktionen ansammeln, immer größer. Die Analyse dieser riesigen Datenmengen wird durch Methoden des Data Minings wesentlich erleichtert. Durch steigende Rechenleistung, Technologien wie Cloud Computing und einer Reihe von leistungsfähigen Data Mining Applikationen, werden sich Projekte in Zukunft mit geringem zeitlichem und finanziellem Risiko umsetzen lassen. Schon heute nehmen große Tourismusorganisationen eine Vorreiterrolle bei der gezielten Anwendung von Data Mining ein. Der Großteil der kleinen Tourismusbetriebe hingegen hat jedoch aufgrund relativ kleiner Budgets für Marketing und IT-Systeme dieses Potenzial noch nicht ausgeschöpft und kann aus den vorhergenannten Trends profitieren.

5 Literaturverzeichnis

A., Germont. *Mustererkennung mit Markov-Modellen: Theorie-Praxis-Anwendungsgebiete.* Wiesbaden : Vieweg+Teubner.

Adriaans P., Zantinge D. 1997. *Data Mining.* Harlow : Addison-Wesley, 1997.

Aggarwal, C. 1999. *Data Mining Techniques for Associations, Clusturing and Classifikation.* Berlin : Springer, 1999. pp. 13-23.

B. Elfron, R.J. Tibshirani. 1993. *An Introduction to the Bootstrap.* London : Chapman & Hall, 1993.

Berry Michael J. A., Linoff Gordon. 2004. *Data mining techniques for marketing, sales and customer relationship management.* Hoboken : Wiley, 2004.

Cabena, P., et al. 1998. *Discovering data mining - from concept to implementation.* Upper Saddle River, New Jersey : Prentice Hall , 1998. p. 12.

Ester, M. and Sander, J. 2000. *Knowledge Discovery in Databases. Techniken und Anwendungen.* Berlin : Springer, 2000.

Hajo Hippner, Klaus D. Wilde. 2001. *Handbuch Data Mining im Marketing.* Wiesbaden : Vieweg, 2001. p. 105.

Han, Jiawei,Kamber, Micheline. 2000. *Data Mining: Concepts and Techniques.* Massachusetts,California : Morgan Kaufmann, 2000.

Hudec, Marcus. 2003. *Data Mining – Ein neues Paradigma der angewandten Statitik.* Wien, Östereich : Institut für Statitik und Decision Support Systems Universität wien, 2003.

Janetzko ., Steinhöfer k. 1997. *Lotsen los! Data Mining: verborgene Zusammenhänge in datenbanken aufspüren.* Hannover : c't, 1997. pp. 294-300.

Kudrass, Thomas. 2007. *Taschenbuch Datenbanken.* München : Hanser Fachbuch, 2007. p. 166.

Küppers, B. 1999. *Data Mining in der Praxis - ein Ansatz zur Nutzung der Potentiale von Data Mining im betrieblichen Umfeld.* Berlin : Springer, 1999. p. 88.

Palmaer, A. Montano, J. J. 2006. *Tourismus Managment.* Berlin : ESV, 2006. pp. 781-790.

Schneider, Nadine Carmen. 2007. *Kundenwertbasierte Effizienzmessung: Der Beitrag von Marketingmaßnahmen zur Unternehmenswerterhöhung in der Automobilindustrie.* Wiesbaden : Gabler, 2007.

T. Hastie, R.J. Tibshirani, J. Friedman. 2001. *The Elements of Statistical Learning.* Heidelberg : Springer, 2001.