Martin-Luther-Universität
Halle-Wittenberg
- Fachbereich Mathematik und Informatik -
Institut für Informatik

Seminararbeit im Fach Informatik
Sommersemester 2006

„On detecting differences between groups“

Name: Sascha Rüger

Matr.-Nr.: 199252267

Betreuer: Dr. Alexander Hinneburg

Vortragsdatum: 30.03.2006

Inhaltsverzeichnis

Tabellenverzeichnis III

1 Einleitung 1

1.1 Data Mining 1

1.2 Contrast-set Mining 1

2 Ausgewählte Data Mining Systeme und deren Eigenschaften 1

2.1 STUCCO (Search and Testing for Understandable Consistent Contrasts) 2

2.2 Magnum Opus 3

3 Auswertung 3

3.1 Ergebnisvergleich der beiden Systeme 3

3.2 Einschätzung der produzierten Regeln 5

4 Weitere Untersuchungen auf Datenmengen 6

5 Fazit 6

Quellenverzeichnis 7

Tabellenverzeichnis

Tabelle 1: Contrast-set ausgegeben von STUCCO 4

Tabelle 2: Sechs Regeln ausgegeben von Magnum Opus 4

Tabelle 3: Beispiel für eine aufbereitete Regel zur Präsentation beim Kunden 5

Tabelle 4: Auswertung der Regeln 5

1 Einleitung

Eine wichtige Aufgabe der Datenanalyse ist das Verstehen der Unterschiede zwischen gegensätzlichen Gruppen. Diese Herausforderung führte zu der neuen Data Mining Technologie Contrast-set Mining. Der Artikel „On detecting differences between groups"[1] basiert auf einer Studie mit Einzelhandelsmitarbeitern einer Warenhauskette, um Contrast-set Mining mit bereits existierenden Regelfindungstechniken zu vergleichen. Dabei werden drei unterschiedliche Regelfindungssysteme untersucht und überraschende Feststellungen gemacht. In dieser Seminararbeit wird der Artikel vorgestellt, ausgewertet und die Ergebnisse präsentiert. Außerdem werden weitere Untersuchungen der betrachteten Thematik benannt.

1.1 Data Mining

„Unter Data Mining versteht man das systematische (in der Regel automatisierte oder halbautomatische) Entdecken und Extrahieren unbekannter Informationen aus großen Mengen von Daten."[2]

Im Zuge immer schneller zunehmender Informationsmengen werden Techniken benötigt, die brauchbare Muster und Regeln auffinden, um spezielle und nützliche Informationen herauszufiltern. Data Mining ermöglicht das automatische Auswerten solcher Datenmengen mit Hilfe von statistischen Verfahren, genetischen Algorithmen, künstlich neuronalen Netzen oder Clustering-Verfahren.

1.2 Contrast-set Mining

Contrast-set Mining ist eine neue Data Mining Technologie zur Identifizierung von Unterschieden bei beobachteten mehrdimensionalen Daten. Das Problem, gegensätzliche Mengen zu analysieren, kann folgendermaßen formuliert werden: „Auffinden der Verbindungen von Attributen und Werten, welche sich vor allem in der Verteilung in Gruppen unterscheiden, was bedeutet, dass die Verbindungen von Attributen signifikant und umfangreich sind."[3] Offensichtlich ermöglicht das Contrast-set Mining, angewandt auf Datencluster, Unterschiede zu finden und unabhängige Entscheidungen für jedes Cluster zu treffen.

Wie oben festgestellt, kann Contrast-set Mining genutzt werden, um clusterdefinierende Regeln von Datengruppen zu finden. Somit fällt es leichter, die Unterschiede zwischen Datenclustern und der Ermöglichung von Einblicken in komplexe Beziehungen zwischen unterschiedlichen Clustern sowie den damit verbundenen Empfehlungen, die Verantwortliche der Datenanalyse treffen müssen, zu verstehen.

2 Ausgewählte Data Mining Systeme und deren Eigenschaften

Geoffrey I. Webb, Shane Butler und Douglas Newlands berichten in ihrem Artikel von einem Projekt zur Evaluierung von Unterschieden zwischen Contrast-set Mining und bereits auf dem Markt vorhandenen Formen von Regelfindungen. In diesem Projekt arbeiteten die Autoren mit der Marketing-Abteilung einer der größten australischen Einkaufsketten zusammen und testeten alternative Data Mining Techniken. Dabei stieß man auf unerwartete Ergebnisse.

[1] Webb, G./Butler, S./Newlands, D.: On detecting differences between groups.
[2] Wikipedia – Data Mining: http://de.wikipedia.org/wiki/Data_mining.
[3] Bay/Pazzani: Detecting group differences – Mining contrast sets.

Untersucht wurden Kaufaktivitäten von zwei verschiedenen Tagen und der Einfluss von speziellen Marketing-Maßnahmen auf das Kaufverhalten. Die Autoren wurden mit Transaktionsdaten dieser zwei Tage unterstützt. Die Daten enthielten alle Posten, die an einem Tag in sechs Geschäften gekauft wurden. Über Identifikationsnummern konnten sie dem Käufer und anderen Produkten, die auch am selben Tag gekauft wurden, zugeordnet werden. Die Aufgabenstellung war es, herauszufinden, wie sich die Einkaufskörbe der Abteilungen während des Zeitraumes unterschieden. Um die Praxistauglichkeit der zwei folgenden Systeme und deren Techniken zu untersuchen, wurden die Daten mit den Systemen ausgewertet und die resultierenden Regeln miteinander verglichen.

2.1 STUCCO (Search and Testing for Understandable Consistent Contrasts)

STUCCO wurde ausgewählt, weil es das einzige den Autoren bekannte System war, was speziell dafür entwickelt wurde, Gegensätze zwischen Gruppen zu finden. Es ist für Daten mit gruppierten Attributwert-Paaren vorgesehen.

Die Daten sind in eine Menge von Gruppen $G_1, G_2, ..., G_l$ gegliedert. Jede Gruppe ist eine Zusammenstellung von Objekten $O_1, O_2, ..., O_n$. Jedes Objekt wiederum besteht aus k Attributwert-Paaren (eins für jedes Attribut $A_1, A_2, ..., A_k$). Das Attribut A_j erhält die Werte von der Menge der $V_{j1} ... V_{jm}$. Ein Contrast-set ist eine Menge von Attributwert-Paaren, in der kein Attribut A_i mehr als einmal vorkommt. Wie bei der Anwendung der Item Mengen wurde nun die Überdeckung (*support*) der Contrast-sets gemessen. Die Überdeckung eines Contrast-set *cset* unter Berücksichtigung einer Gruppe G_i ist der Anteil der Objekte $o \in G_i$, so dass $cset \subseteq o$ ist. Diese Überdeckung wird $supp\ (cset, G_i)$ bezeichnet.

Contrast-set-Untersuchungen streben danach, alle Contrast-sets zu finden, deren Überdeckungen sich entscheidend über die Gruppen unterscheiden. Dieses Streben ist definiert als Suchen aller Contrast-sets cset, die folgende Ungleichungen erfüllen:

$$\exists ij\ P(cset \mid G_i) \neq P(cset \mid G_j) \quad (1)$$

und

$$\max_{ij} \mid supp(cset, G_i) - supp(cset, G_j) \mid \geq \delta, \quad (2)$$

wobei δ eine benutzerdefinierte Schwelle, genannt Minimum-Support-Differenz, ist. Contrast-sets, für die Gleichung 1 statistisch erfüllt ist, werden als signifikant bezeichnet. Contrast-sets, für die Gleichung 2 gültig ist, werden als umfangreich eingestuft.[4]

STUCCO nutzt eine effiziente Suche zum Durchsehen der Contrast-sets, basierend auf dem Algorithmus Max-Miner von Bayardo[5]. Die statistische Signifikanz der Gleichung 1 wird geschätzt, indem der Chi-Quadrat-Test zur Schätzung der Nullhypothese, dass die Überdeckung der Contrast-sets unabhängig von der Gruppenzugehörigkeit ist, angewendet wird. In mehrfachen Vergleichen wird dann eine Korrektur von α vorgenommen, das systematisch je nach Ansteigen der Größe des Contrast-sets verringert wird. Dieses Vorgehen kontrolliert die Wahrscheinlichkeit eines Fehlers 1. Art (Inkorrekte Annahme der Existenz eines Contrast-sets). Auf jeder Stufe i dieser Suche, wobei die Stufe i die Anzahl i der Attributwert-Paare der Contrast-sets angibt, wird eine korrekte Signifikanz wie folgt angegeben:

$$\alpha_i = \min\left(\frac{\alpha}{2^i} / |C_i|, \alpha_{i-1}\right), \quad (3)$$

wobei α die Signifikanz-Stufe (0,05) und $|C_i|$ die Anzahl der Kandidaten von Contrast-sets auf Stufe i angeben. Pruning (Reduzieren) wird angewendet, um Contrast-sets, die nur Spe-

[4] Vgl. Webb, G./Butler, S./Newlands, D.: On detecting differences between groups.
[5] Vgl. Bayardo, J./Roberto J.: Efficiently mining long patterns from databases.

zialisierungen von allgemeineren Contrast-sets sind, zu entfernen. Abschließend wird der Suchraum unter einem Contrast-set *cset* gestrichen, wenn die Überdeckung für die Gruppe die höchste Überdeckung bleibt, ganz gleich welche zusätzlichen Ausdrücke zu *cset* hinzugefügt werden.

2.2 Magnum Opus

Magnum Opus ist eine kommerzielle Implementierung des OPUS_AR Regel-Findungs-Algorithmus. OPUS_AR erweitert den OPUS Such-Algorithmus um das Suchen nach Regeln der Form a $\rightarrow$ c. Dabei stehen a (Prämisse) für eine Menge (oder eine Konjunktion) von Attributwert-Paaren und c (Konklusion) für ein beliebiges Attribut aus der Menge erlaubter Attributwert-Paare. Für die Anwendung von Magnum Opus auf die geschilderte Aufgabenstellung der Contrast-set Suche wird die rechte Seite dieser Regeln auf Attributwerte beschränkt, die eine Gruppenzugehörigkeit repräsentieren.[6]

OPUS führt eine effiziente Suche über den Raum von möglichen Regeln für eine einzelne Konklusion durch, indem es systematisch den Raum möglicher Mengen von Attributwert-Paaren, die eine Prämisse bilden können, erweitert. Dabei wird das Reduzieren (pruning) der Kindknoten ausgeweitet und der Suchraum kann zur Verbesserung der Sucheffizienz dynamisch neu bestimmt werden. Magnum Opus nutzt diesen Such-Ansatz, um Assoziationsregeln zu finden. Dies unterscheidet sich von der bekannten Herangehensweise insofern, dass keine Finden-Häufiger-Item-Mengen-Strategie genutzt wird. Daher müssen auch keine Bedingungen an die minimale Überdeckung einer Regel geknüpft werden. Stattdessen muss bei einer Messung von Regelausprägungen die maximale Anzahl der Regeln *maxr* angegeben werden, welche die festgelegte Messung von Regelausprägungen optimiert und weitere nutzerspezifische Bedingungen erfüllt. Eine effiziente Suche wird erreicht, indem die Bereiche des Suchraumes weggenommen werden, die keine Zielregeln enthalten.

Die Ermittlung von Regelausprägungen die Magnum Opus unterstützt sind Support, Confidence (Strength), Lift, Leverage und Coverage. In dem dieser Arbeit zu Grunde liegenden Artikel wurde der Leverage untersucht. Der **Leverage** misst den Grad der Abweichung von der beobachteten gemeinsamen Häufigkeit von Prämisse und Konklusion sowie der gemeinsamen Häufigkeit, die erwartet werden kann, wenn Prämisse und Konklusion unabhängig voneinander sind. Der Leverage ist eine nützliche Methode für viele Regelermittlungsaufgaben. Magnum Opus unterstützt zahlreiche Hilfen die zu ermittelnden Regeln zu kontrollieren. So gibt es zum Beispiel eine Regel-Filter-Funktion.

3 Auswertung

Dieses Kapitel gibt einen Überblick über die Ergebnisse, die mit den zwei Systemen nach Auswertung bzw. Anwendung der Daten erzielt wurden. Dabei werden die ausgegebenen Regeln untersucht und ausgewertet.

3.1 Ergebnisvergleich der beiden Systeme

STUCCO produzierte 19 Contrast-sets und Magnum Opus 83 Regeln. Eine Ausgabe für ein Contrast-set von STUCCO ist in Tabelle 1 zu sehen.[7]

[6] Vgl. Webb, G./Butler, S./Newlands, D.: On detecting differences between groups.
[7] Vgl. Webb, G./Butler, S./Newlands, D.: On detecting differences between groups.

220 = 1			
434	257	| 0.0689327	0.037214
= =			
d.f.		chi^2	pvalue
1		66.80	3.00e-16

Tabelle 1: Contrast-set ausgegeben von STUCCO

Die erste Zeile enthält das Contrast-set. In diesem Fall ist das Abteilung 220. Die Angaben 434 und 257 sind die Anzahl der Transaktionen an den jeweiligen Tagen. Die nächsten zwei Werte geben den Anteil der Transaktionen in dieser Abteilung an den Gesamttransaktionen an den zwei Tagen an. Die übrigen Zahlen sind das Ergebnis des Chi-Quadrat-Tests der Signifikanz.

Magnum Opus erzeugte 56 Regeln, die einen einzelnen Wert in der Prämisse enthielten, 23 mit zwei und vier mit drei Werten. Alle Werte, die in den zwei- oder drei-elementigen Zustandsregeln enthalten sind, sind auch in einer ein-elementigen Zustandsregel enthalten. Tabelle 2 enthält dazu einige Beispielregeln.[8]

851 -> August-21st [Coverage = 0.049 (649); Support = 0.038 (500); Strength = 0,770; Lift = 1.47; Leverage = 0.0122 (160)]
855 -> August-21st [Coverage = 0.043 (574); Support = 0.033 (432); Strength = 0,753; Lift = 1.44; Leverage = 0.0100 (131)]
855 & 851 -> August-21st [Coverage = 0.009 (119); Support = 0.008 (104); Strength = 0,874; Lift = 1.67; Leverage = 0.0032 (41)]
220 -> August-14th [Coverage = 0.052 (691); Support = 0.033 (434); Strength = 0,628; Lift = 1.32; Leverage = 0.0079 (104)]
355 -> August-14th [Coverage = 0.007 (98); Support = 0.006 (74); Strength = 0,755; Lift = 1.58; Leverage = 0.0021 (27)]
220 & 355 -> August-21st [Coverage = 0.001 (15); Support = 0.001 (13); Strength = 0,867; Lift = 1.66; Leverage = 0.0004 (5)]

Tabelle 2: Sechs Regeln ausgegeben von Magnum Opus

Coverage ist der Anteil an Transaktionen über die zwei Tage, die Käufe der Menge von Abteilungen beinhalten, die in der Prämisse enthalten sind. ***Support*** ist der Anteil an allen Transaktionen, die Abteilungen von dem Tag beinhalten, welcher in der Konklusion steht. Der Wert in Klammern hinter jedem dieser gemessenen Größen ist die Anzahl der Transaktionen. Zu beachten ist, dass die spezielleren Regeln (die mit mehreren Konditionen) eine höhere *Confidence* (von Magnum Opus *Strength* genannt) haben, als die allgemeineren Regeln, welche die gleichen Konditionen enthalten.

Die ersten drei Regeln zeigen das typische Verhältnis zwischen zwei allgemeineren und einer spezielleren Regel, die jeweils mehrere Abteilungen enthalten. Die ersten zwei Regeln deuten darauf hin, dass der Anteil an Kunden, die am zweiten Tag in den Abteilungen 851 und 855 eingekauft haben, höher ist, als am ersten Tag. Regel drei zeigt, dass dieser Effekt vergrößert wird, wenn Kunden, die in einer einzigen Abwicklung in beiden Abteilungen eingekauft haben, berücksichtigt werden. Die letzten drei Regeln sind besonders interessant. Während am 14. August mehr Warenposten in den Abteilungen 220 und 355 gekauft worden als am 21.

[8] Vgl. Webb, G./Butler, S./Newlands, D.: On detecting differences between groups.

August, ist am 21. ein größerer Anteil an Kunden, die in beiden Abteilungen gekauft haben, festgestellt worden.

Der Hauptunterschied der beiden Systeme ist die Anwendung der Filter, die unbrauchbare und uninteressante Contrast-sets identifizieren und ausschließen. Magnum Opus nutzt einen Hypothesentest und STUCCO den Chi-Quadrat-Test. Der Filter von Magnum Opus ist etwas nachsichtiger als der von STUCCO, was auch die Anzahl der ausgegebenen Regeln zeigt. Die Anwendung statistischer Tests unterscheidet die beiden Systeme hauptsächlich von dem traditionellen Finden von Assoziationsregeln.

3.2 Einschätzung der produzierten Regeln

Die Analyse der von den untersuchten Systemen produzierten Regeln zeigt, dass der Filter entweder strenger oder weniger streng eingestellt werden sollte. Einerseits erhält man mit Magnum Opus Regeln, die scheinbar falsch sind, andererseits werden Regeln, die sehr repräsentativ sind, auf Grund des genau eingestellten Filters ausgesondert.

Um zu beurteilen, ob der genauere Filter von STUCCO berechtigt ist, gingen die Autoren des Artikels zu den Einzelhandelsmitarbeitern und baten sie um Mithilfe. Zu diesem Zweck mussten die ausgegebenen Regeln noch so aufbereitet werden, dass sie auch einen brauchbaren Inhalt bezüglich der Verkaufstage hatten. Eine Regel hatte dann folgendes Aussehen:

On August 21st customers were 7.6 times more likely to purchase items from department 445 (MENSWEAR; Mens Nightwear) than they were on August 14th. They were bought in 2.2% of transactions on August 21st and 0.3% of transactions on August 14th.

Tabelle 3: Beispiel für eine aufbereitete Regel zur Präsentation beim Kunden

Um den freiwilligen Helfern nicht unnötig Arbeit aufzubürden, wurden ihnen bezüglich der Regeln zwei einfache (wahr/falsch) Fragen gestellt:
F 1: Ist die Regel überraschend?
F 2: Ist die Regel möglicherweise nützlich für das Unternehmen?

Die Ergebnisse dieser Auswertung sind in Tabelle 4 abgebildet.[9] Da STUCCOs Filter die Regeln mit mehreren Abteilungen herausnimmt, ist es interessant zu sehen, dass die von Magnum Opus gefundenen Regeln mit mehreren Abteilungen weniger überraschend und auch weniger nützlich sind.

System	Total nr. rules	Surprising	Potentially Useful
Magnum Opus (1 Dept.)	56	12 (21%)	15 (27%)
Magnum Opus (2 Depts.)	23	10 (43%)	5 (22%)
Magnum Opus (3 Depts.)	4	1 (25%)	1 (25%)
Magnum Opus (All)	83	23 (28%)	21 (25%)
STUCCO	19	2 (11%)	5 (26%)

Tabelle 4: Auswertung der Regeln

Auffällig ist auch, dass trotz dem STUCCO einen strengeren Filter hat als Magnum Opus, der Anteil an überraschenden Regeln geringer ist, als der von Magnum Opus. Eine mögliche Erklärung dafür ist, dass Magnum Opus einige Regeln gefunden hat, die falsch sind und daher als überraschend von den Einzelhandelsmitarbeitern eingestuft wurden. Betrachtet man weiterhin die Magnum Opus Regeln mit unterschiedlicher Abteilungsanzahl, fällt auf, dass der Anteil nützlicher Regeln mit zwei und drei Abteilungen leicht abfällt.

[9] Vgl. Webb, G./Butler, S./Newlands, D.: On detecting differences between groups.

4 Weitere Untersuchungen auf Datenmengen

Dong und Li[10] arbeiteten an dem Problem der Suche nach emergent patterns EP. Ein EP ist definiert als eine Item-Menge X, für die gilt:

$$\text{Wachstumsrate (X)} = \frac{\sup_{D_2}(X)}{\sup_{D_1}(X)} > g,$$

D_1, D_2 sind zwei verschiedene Datenmengen, g ist eine Wachstumsbegrenzung und die Abkürzung sup steht für support. Dieser Algorithmus ermittelt EPs unter Berücksichtigung von Grenzen. Ein Nachteil dieses Ansatzes ist, dass der Algorithmus die Daten bei einer gegebenen Wachstumsgrenze mehrere Male durchsuchen muss. Auch ist nicht klar, ob diese Methode mehr als zwei Mengen durchsuchen kann und welche Methoden zur Überprüfung der statistischen Signifikanz von gefundenen EPs angewendet werden können.

Ein anderer Ansatz von Chakrabarti, Sarawagi und Dom[11] behandelt das Finden von überraschenden temporalen Mustern in booleschen Daten von Warenkörben. Dabei werden Item-Mengen gesucht, deren Support über einen Zeitraum variiert. Unter Nutzung eines Minimum-Description-Length Ansatzes sind überraschende Muster gerade die, welche hohe Übersetzungskosten aufweisen. Die Daten werden in getrennte Zeitabschnitte eingeteilt, so dass ein Modell auf jeden Abschnitt angewandt werden kann und die Kosten berechenbar sind. Somit können Änderungen in einer Verteilung gefunden werden, die über einen Zeitraum variieren. Ein Vergleich zweier Gruppen ist daher nicht möglich, da nur eine Verteilung existiert.

5 Fazit

In dieser Arbeit wurden zwei Systeme auf Data Mining Techniken zur Bestimmung von Contrast-sets untersucht. Dabei wurde festgestellt, dass die Gleichung zur Bestimmung eines Contrast-set äquivalent zu der viel allgemeineren Gleichung eines Regel-Findungs-Systems ist. Das konnte daran festgemacht werden, da das System von Magnum Opus Regeln zusammengestellt hat, die mit allen von STUCCO gefundenen Contrast-sets übereinstimmen.

Es wurde außerdem herausgefunden, dass die Anwendung von Filtern auf die Daten, um unechte Korrelation auszuschließen, von großer Bedeutung ist. Dabei haben weder STUCCO noch Magnum Opus einen korrekt eingestellten Filter angewandt. STUCCO hat demnach brauchbare Daten ausgesondert, Magnum Opus hingegen hat auf Grund eines ungenaueren Filters zu viele Regeln produziert. Da in der Studie „On detecting differences between groups“ die Standardeinstellungen der Filter genutzt wurden, könnten mit Feineinstellung der Filtereigenschaften möglicherweise bessere Ergebnisse erzielt werden. Eine weitere Untersuchung zur Wahl und Einstellung des Filters könnte zusätzliche Erkenntnisse bringen.

Contrast-set Mining ist eine wertvolle Data Mining Technologie, die in vielen Belangen zum Einsatz kommen kann. Zwar können hierbei die traditionellen Assoziationsregel-Messungen von Support, Confidence und Lift angewendet werden, um die Kerneigenschaften von Contrast-sets Aufgaben zu erfüllen. Trotzdem es notwendig, neue Methoden zu finden, die Contrast-sets besser herausfiltern und somit für die Nutzer brauchbare und interessante Informationen hervorbringen.

[10] Dong, G./Li, J. (1999): Efficient mining of emerging patterns – Discovering trends and differences. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
[11] Chakrabarti, S./Sarawagi, S./Dom, B. (1998): Mining surprising patterns using temporal description length. In: Proceedings of the 24th International Conference on Very Large Databases.

Quellenverzeichnis

Bay/Pazzani: Detecting group differences – Mining contrast sets.

Bayardo, J./Roberto J.: Efficiently mining long patterns from databases.

Chakrabarti, S./Sarawagi, S./Dom, B. (1998): Mining surprising patterns using temporal description length. In: Proceedings of the 24th International Conference on Very Large Databases.

Dong, G./Li, J. (1999): Efficient mining of emerging patterns – Discovering trends and differences. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Webb, G./Butler, S./Newlands, D.: On detecting differences between groups.

Wikipedia – Data Mining: http://de.wikipedia.org/wiki/Data_mining.