On detecting differences between groups

Seminar im Fach Informatik

Sommersemester 2006

Sascha Rüger

Gliederung

1. Einleitung
2. Data Mining Systeme
3. Auswertung
4. Weitere Untersuchungen
5. Fazit

1. Einleitung (1)

- wichtige Aufgabe der Datenanalyse: Verstehen der Unterschiede zwischen gegensätzlichen Gruppen

→ Contrast-set Mining

1. Einleitung (2)

1.1 Data Mining

1.2 Contrast-set Mining

1.1 Data Mining (1)

- systematisches Entdecken und Extrahieren unbekannter Informationen aus großen Mengen von Daten
- zunehmende Informationsmengen → Techniken, die brauchbare Muster und Regeln auffinden, um spezielle und nützliche Informationen herauszufiltern

1.1 Data Mining (2)

- automatische Auswerten solcher Datenmengen mit Hilfe von:
  - statistischen Verfahren,
  - genetischen Algorithmen,
  - künstlich neuronalen Netzen oder
  - Clustering-Verfahren.

1. Einleitung (3)

1.2 Contrast-set Mining (1)

- neue Data Mining Technologie zur Identifizierung von Unterschieden bei beobachteten mehrdimensionalen Daten

„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.“ (Bay, Pazzani)

1.2 Contrast-set Mining (2)

- Finden von clusterdefinierenden Regeln von Datengruppen
- Einblicke in komplexe Beziehungen zwischen unterschiedlichen Clustern sowie den damit verbundenen Empfehlungen, die Verantwortliche der Datenanalyse treffen

2. Data Mining Systeme (1)

- Projekt zur Evaluierung von Unterschieden zwischen Contrast-set Mining und auf dem Markt vorhandenen Formen von Regelfindungen
- Marketing-Abteilung einer großen australischen Einkaufskette

➢ testen alternativer Data Mining-Techniken und -Systeme

2. Data Mining Systeme (2)

- Kaufaktivitäten von zwei verschiedenen Tagen
- Einfluss von speziellen Marketing-Maßnahmen auf das Kaufverhalten
- Transaktionsdaten enthielten alle Posten, die an einem Tag in sechs Geschäften gekauft wurden

2. Data Mining Systeme (3)

2.1 STUCCO

2.2 Magnum Opus

2.1 STUCCO (1)

Search and **T**esting for **U**nderstandable **C**onsistent **Co**ntrasts

- das einzige den Autoren bekannte System, dass Gegensätze zwischen Gruppen finden kann
- ist für Daten mit gruppierten Attributwert-Paaren vorgesehen

2.1 STUCCO (2)

- 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 besteht aus k Attributwert-Paaren (eins für jedes Attribut A_1, A_2, ..., A_k)
- Attribut A_j erhält die Werte von der Menge der V_{j1} ... V_{jm}

2.1 STUCCO (3)

- Contrast-set ist eine Menge von Attributwert-Paaren, in der kein Attribut A_i mehr als einmal vorkommt
- Messung der Überdeckung (*support*) der Contrast-sets
- Überdeckung eines Contrast-set *cset*: *supp (cset, G_i):* Anteil der Objekte $o \in G_i$, so dass $cset \subseteq o$ ist

2.1 STUCCO (4)

- Contrast-set-Untersuchungen streben danach, alle Contrast-sets zu finden, deren Überdeckungen sich entscheidend über die Gruppen unterscheiden:

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

(2) $\max ij \left| supp(cset, G_i) - supp(cset, G_j) \right| \geq \delta$

δ: benutzerdefinierte Schwelle, genannt Minimum-Support-Differenz

2.1 STUCCO (5)

- statistische Signifikanz von (1): Chi-Quadrat-Test (Schätzung der Nullhypothese – Support der Contrast-sets unabhängig von der Gruppenzugehörigkeit)
- Korrektur von α: systematische Verringerung je nach Ansteigen der Größe des Contrast-sets

→kontrolliert die Wahrscheinlichkeit Fehlers 1. Art (Inkorrekte Annahme der Existenz eines Contrast-sets)

2.1 STUCCO (6)

- Anwenden von Pruning: Entfernen von Contrast-sets, die nur Spezialisierungen von allgemeineren Contrast-sets sind
- Streichen des Suchraums unter einem Contrast-set *cset*, wenn Überdeckung für die Gruppe die höchste Überdeckung bleibt (egal welche zusätzlichen Ausdrücke zu *cset* hinzugefügt werden)

2. Data Mining Systeme (4)

2.1 STUCCO

2.2 Magnum Opus

2.2 Magnum Opus (1)

- kommerzielle Implementierung des OPUS_AR Regel-Findungs-Algorithmus
- OPUS_AR: Erweiterung des OPUS Such-Algorithmus um Suchen nach Regeln der Form a → c
  - a = Prämisse für eine Menge (oder eine Konjunktion) von Attributwert-Paaren
  - c = Konklusion für ein beliebiges Attribut aus der Menge erlaubter Attributwert-Paaren stehen

2.2 Magnum Opus (2)

- effiziente Suche über den Raum von möglichen Regeln für eine einzelne Konklusion:
  - systematische Erweiterung des Raums möglicher Mengen von Attributwert-Paaren, die eine Prämisse bilden können
- Ausweiten des Pruning der Kindknoten
  - der Suchraum kann zur Verbesserung der Sucheffizienz dynamisch neu bestimmt werden
- Angabe der maximalen Regelanzahl maxr

2.2 Magnum Opus (3)

- Ermittlung von Regelausprägungen:
  - Support
  - Confidence (Strength)
  - Lift
  - Leverage
    - Grad der Abweichung von beobachteter gemeinsamer Häufigkeit von Prämisse und Konklusion sowie gemeinsamer Häufigkeit, die erwartet werden kann, wenn Prämisse und Konklusion unabhängig voneinander sind
  - Coverage

3. Auswertung

3.1 Ergebnisvergleich

3.2 Einschätzung der produzierten Regeln

3.1 Ergebnisvergleich (1)

- STUCCO: 19 Contrast-sets
- Magnum Opus: 83 Regeln

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

Beispiel 1: Contrast-set ausgegeben von STUCCO

3.1 Ergebnisvergleich (2)

- Magnum Opus:
  - 56 Regeln (ein Wert in der Prämisse)
  - 23 Regeln (zwei Werte in der Prämisse)
  - 4 Regeln (drei Werte in der Prämisse)

3.1 Ergebnisvergleich (3)

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)]

Beispiel 2: Sechs Regeln ausgegeben von Magnum Opus

3.1 Ergebnisvergleich (4)

- Hauptunterschied der beiden Systeme: Anwendung der Filter
  - Magnum Opus: Hypothesentest
  - STUCCO: Chi-Quadrat-Test
- Anwendung statistischer Tests unterscheidet sich von dem traditionellen Finden von Assoziationsregeln

3. Auswertung

3.1 Ergebnisvergleich

3.2 Einschätzung der produzierten Regeln

3.2 Einschätzung der produzierten Regeln (1)

- Filtereinstellung: streng oder weniger streng
- Befragung der Einzelhandelsmitarbeiter:
  - F 1: Ist die Regel überraschend?
  - F 2: Ist die Regel möglicherweise nützlich für das Unternehmen?

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.

Beispiel 3: Beispiel für eine aufbereitete Regel

3.2 Einschätzung der produzierten Regeln (2)

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

Beispiel 4: Auswertung der Regeln

4. Weitere Untersuchungen (1)

- emergent patterns EP:
  - Item-Menge X, für die gilt:

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

  - D_1, D_2: verschiedene Datenmengen
  - g ist eine Wachstumsbegrenzung
  - Nachteile:
    - Algorithmus muss Daten bei einer gegebenen Wachstumsgrenze mehrere Male durchsuchen
    - mehr als zwei Mengen durchsuchbar?

4. Weitere Untersuchungen (2)

- Finden von überraschenden temporalen Mustern in booleschen Daten von Warenkörben
- Suchen von Item-Mengen, deren Support über einen Zeitraum variiert
- Nutzung Minimum-Description-Length Ansatzes:
  - überraschende Muster: hohe Übersetzungskosten
- getrennte Zeitabschnitte: Anwendung Modell auf jeden Abschnitt → Kosten berechenbar

5. Fazit

- Bestimmung Contrast-set äquivalent zu viel allgemeinerem Regel-Findungs-System
- Anwendung von Filtern auf die Daten von großer Bedeutung
- Contrast-set Mining: wertvolle Data Mining Technologie
- neue Methoden notwending, die Contrast-sets besser herausfiltern