Kombinationen von Data Mining-Verfahren: Analyse und Automatisierung

(Masterarbeit bei Prof. U. Lipeck)

Oberseminar Datenbanksysteme

Universität Hannover

Ulf Mewe (ulf@mewenet.de)

29.04.2008

Kombiniertes Data Mining

- Die sequentielle Ausführung von Data Mining-Verfahren wird als „Kombiniertes Data Mining“ bezeichnet. [Hum04]
- Beim „Kombinierten Data Mining“ wird ein Data Mining-Verfahren A vor einem Data Mining-Verfahren B so ausgeführt, dass B von A profitiert.

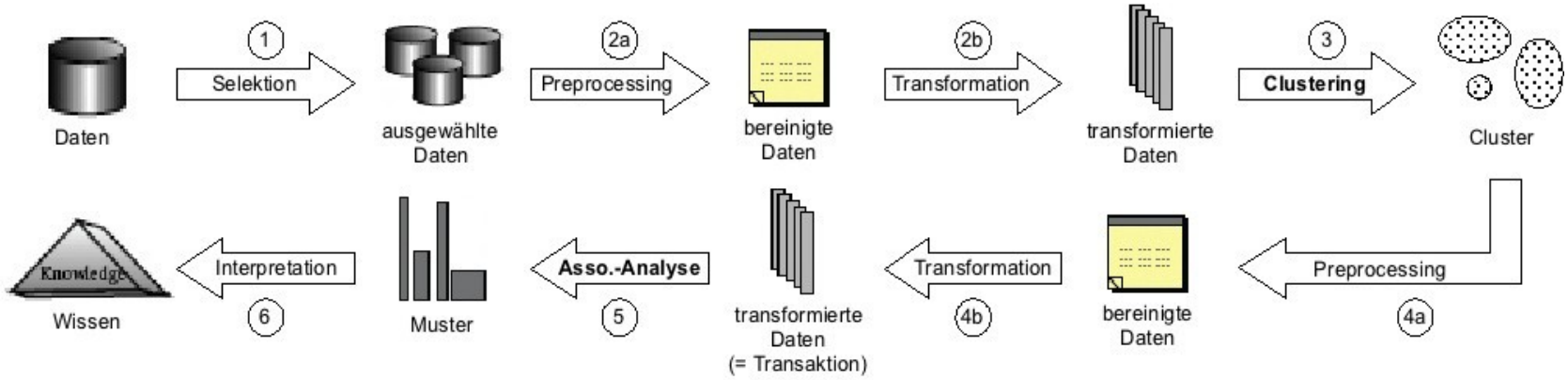
  B kann dazu das Ergebnis von A und/oder eigens ermittelte Hilfsinformationen von A für B nutzen.

  B profitiert dann von A, wenn das Ergebnis von B, gemäß einem geeigneten Gütemaß, besser ist [als ohne A], und/oder sich die Laufzeit von B verringert. [Stº 04]

Kombination
aus Clustering und Assoziationsanalyse:

- Auswahl der Daten (Selektion)
- Vorbereitung der Daten für das Clustering (Preprocessing / Transformation)
- **Clustering der Daten**
- Vorbereitung der Daten für die Assoziationsanalyse (Preprocessing / Transformation)
- **Assoziationsanalyse (für jeden Cluster)**
- Präsentation der Ergebnisse

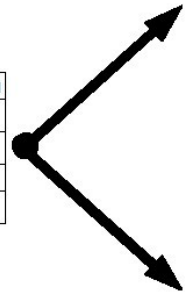
Ablauf


Daten
1
Selektion
ausgewählte Daten
2a
Preprocessing
bereinigte Daten
2b
Transformation
transformierte Daten
3
Clustering
Cluster
Knowledge
Wissen
Interpretation
6
Muster
Asso.-Analyse
5
transformierte Daten (= Transaktion)
Transformation
4b
bereinigte Daten
Preprocessing
4a

Auswahl der Daten

Basistabelle

ID	Items	Attribut 1	...	Attribut m
1	$item_a,...,item_b$	a_{11}	...	a_{m1}
2	$item_c,...,item_d$	a_{12}	...	a_{m2}
...	...	...	...	...
n	$item_x,...,item_y$	a_{1n}	...	a_{mn}



Clusterdaten

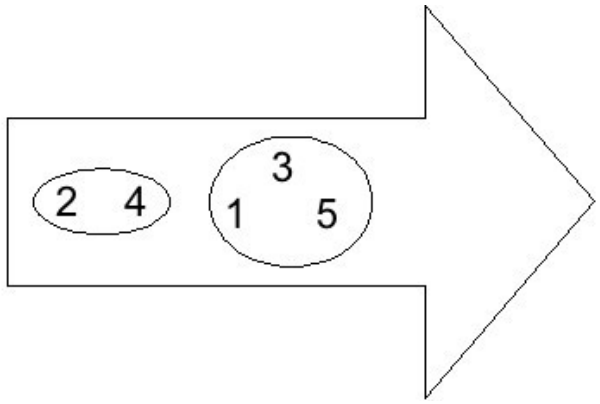
ID	Attribut 1	...	Attribut m
1	a_{11}	...	a_{m1}
2	a_{12}	...	a_{m2}
...	...	...	...
n	a_{1n}	...	a_{mn}

Transaktionen

ID	Items
1	$item_a,...,item_b$
2	$item_c,...,item_d$
...	...
n	$item_x,...,item_y$

Aufteilung der Transaktionen

1: $item_a,...,item_m$
2: $item_b,...,item_n$
3: $item_c,...,item_o$
4: $item_d,...,item_p$
5: $item_e,...,item_q$



2: $item_b,...,item_n$
4: $item_d,...,item_p$

1: $item_a,...,item_m$
3: $item_c,...,item_o$
5: $item_e,...,item_q$

Ziele

- Verbesserung der Ergebnisse
  - Verbesserung der Confidence durch Reduktion der Ausnahmen
  - ➔ durch zusätzliche Assoziationsregeln
  - ➔ durch verbesserte Assoziationsregeln
- Verbesserung der Laufzeit
- Verbesserung des Speicherplatzbedarfs

Wie können Mengen von Assoziationsregeln miteinander verglichen werden?

- keine Notwendigkeit für den Vergleich von Mengen von Assoziationsregeln beim herkömmlichen DM
- Support, Confidence, ... funktionieren nur für den Vergleich von einzelnen Regeln

➔ Einführung eines neuen Maßes

Ideen

- Bewertung einer Regelmenge anhand des enthaltenen Wissen (= inhärentes Wissen)
- Wissen ist gleichbedeutend mit der Fähigkeit, Daten zu komprimieren (vgl. MDL-Sicht)
- Komprimierung der Transaktionen mit Hilfe der gefundenen Assoziationsregeln zur Bestimmung des inhärenten Wissens

- **Assoziationsregeln**

  R_1: Chips $\Rightarrow$ Cola

  R_2: Babypuder $\Rightarrow$ Windeln

  R_3: Marmelade, Nutella $\Rightarrow$ Toast

  R_4: Schraubenzieher $\Rightarrow$ Muttern

- **Transaktionen**

  T_1: {Chips, Cola}

  T_2: {Babypuder, Windeln}

  T_3: {Chips, Cola, Marmelade, Nutella, Toast}

  T_4: {Cola}

  T_5: {Marmelade, Toast}

  T_6: {Chips}

  T_7: {Schraubenzieher, Schrauben}

- **Komprimierte Transaktionen**

$$T_1\text{: \{Chips, Cola\}} \xrightarrow{R_1'} T_1\text{: \{Chips\}} \quad (+1)$$

$$T_2\text{: \{Babypuder, Windeln\}} \xrightarrow{R_2'} T_2\text{: \{Babypuder\}} \quad (+1)$$

$$T_3\text{: \{Chips, Cola, Marmelade, Nutella, Toast\}} \xrightarrow{R_1', R_3'} T_3\text{: \{Chips, Marmelade, Nutella\}} \quad (+2)$$

$$T_4\text{: \{Cola\}} \xrightarrow{\emptyset} T_4\text{: \{Cola\}} \quad (0)$$

$$T_5\text{: \{Marmelade, Toast\}} \xrightarrow{\emptyset} T_5\text{: \{Marmelade, Toast\}} \quad (0)$$

- **Assoziationsregeln**

  R_1: Chips $\Rightarrow$ Cola

  R_2: Babypuder $\Rightarrow$ Windeln

  R_3: Marmelade, Nutella $\Rightarrow$ Toast

  R_4: Schraubenzieher $\Rightarrow$ Muttern

- **Transaktionen**

  T_1: {Chips, Cola}

  T_2: {Babypuder, Windeln}

  T_3: {Chips, Cola, Marmelade, Nutella, Toast}

  T_4: {Cola}

  T_5: {Marmelade, Toast}

  T_6: {Chips}

  T_7: {Schraubenzieher, Schrauben}

- **Komprimierte Transaktionen**

  T_6: {Chips} $\xrightarrow{R_1}$ T_6: {Chips, Cola}

  T_7: {Schraubenzieher, Schrauben} $\xrightarrow{R_4}$ T_7: {Schraubenzieher, Schrauben, Muttern}

  T_6: {Chips} $\xrightarrow{R_1'}$ T_6: {Chips, ¬Cola} (-1)

  T_7: {Schraubenzieher, Schrauben} $\xrightarrow{R_4'}$ T_7: {Schraubenzieher, Schrauben, ¬Muttern} (-1)

Definition des inhärenten Wissens

- Sei *k* der Speicherplatz, der zur Speicherung der mit den Assoziationsregeln aus *A* komprimierten Transaktionen benötigt wird,
  und *u* der Speicherplatz, der notwendig w^{a}re,
  um die unkomprimierten Transaktionen *T*
  dann ist das *inh^{a}rente Wissen IW(A,T)* definiert als:

$$IW(A,T) = 1 - \frac{k}{u}$$

Durchgeführte Experimente

- 6 unterschiedliche Ausschnitte aus einer Filmdatenbank (MovieDB)
- Clustering nach unterschiedlichen Informationen zu den Filmen (Budget, Genre, Rating, ...)
- Assoziationsanalyse mit Schauspielern

Blockbuster

- Kinofilme aus den USA mit einem Rating ≥ 7
- betrachtet werden die 4 erstgenannten Schauspieler
- 3691 Filme mit 3662 verschiedenen Schauspielern
- min. Support = 3 Filme
- min. Confidence = 50%

Blockbuster

Ref.	Quelle	Clusterattribute	#Cluster	#Regeln	*IW*	Δ*IW*
3	Blockbuster (USA)			53	0,79 (29/3650)	
3	Blockbuster (USA)	Dekade	3	23	0,49 (18/3650)	-38%
3	Blockbuster (USA)	Genre	3	52	1,04 (38/3650)	+31%
3	Blockbuster (USA)	Rating	3	53	0,85 (31/3650)	+7%
3	Blockbuster (USA)	Dekade, Genre	3	17	0,85 (31/3650)	+7%
3	Blockbuster (USA)	Genre, Rating	3	52	0,85 (31/3650)	+7%

Analyse des Ergebnises (Blockbuster:Genre)

- 3 Cluster
  - Cluster 1: unterschiedliche Genres
  - Cluster 2: Kriminalfilme
  - Cluster 3: Dramen
- Verlust einer Assoziationsregel

```
'Pesci, Joe'  =>  'De Niro, Robert'  (Confidence=0.75)
```

- keine neuen Assoziationsregeln

Analyse des Ergebnises (Blockbuster:Genre)

- Verbesserung von 7 Assoziationsregeln
  - herkömmliches Data Mining

```
'Shatner, William'  =>  'Doohan, James'  (Confidence=0.75)
'Shatner, William'  =>  'Kelley, DeForest'  (Confidence=0.75)
'Shatner, William'  =>  'Nimoy, Leonard'  (Confidence=0.75)
'Jackman, Hugh'  =>  'Stewart, Patrick'  (Confidence=0.75)
'Jackman, Hugh'  =>  'McKellen, Ian'  (Confidence=0.75)
'Farrow, Mia'  =>  'Allen, Woody'  (Confidence=0.57)
'Keaton, Diane'  =>  'Allen, Woody'  (Confidence=0.56)
```

  - „Kombiniertes Data Mining“

```
'Shatner, William'  =>  'Doohan, James'  (Confidence=1)
'Shatner, William'  =>  'Kelley, DeForest'  (Confidence=1)
'Shatner, William'  =>  'Nimoy, Leonard'  (Confidence=1)
'Jackman, Hugh'  =>  'Stewart, Patrick'  (Confidence=1)
'Jackman, Hugh'  =>  'McKellen, Ian'  (Confidence=1)
'Farrow, Mia'  =>  'Allen, Woody'  (Confidence=0.8)
'Keaton, Diane'  =>  'Allen, Woody'  (Confidence=1)
```

Plausibilit[a] t

- Aufteilung der Transaktionen, sodass
  - Ausnahmen (von den Regeln) reduziert werden
  - ➔ Erhöhung der Confidence der einzelnen Regeln
- Attributwerte bewegen sich in einem bestimmten Kontext
  - Schauspieler drehen Filme eines bestimmten Typ, eines bestimmten Genres, ...
  - dreht ein Schauspieler einen „ungewöhnlichen“ Film, dann tut er dies vermutlich mit anderen Schauspielern

Fazit

- Verbesserung / Verschlechterung der Ergebnisse durch die Kombination von Clustering und Assoziationsanalyse möglich
- deutliche Verbesserung der Laufzeit möglich