Kapitel DM:V

V. Association Analysis

- Assoziationsanalyse
- Frequent Itemset Mining
- Regel-Mining

Regel-Mining

Von häufigen Itemsets zu Regeln

Assoziationen zwischen Items ($A \Rightarrow B$)

- Gegeben einen k-Itemset I_A^k. Wie hoch ist die Wahrscheinlichkeit, dass Item B ebenfalls gekauft wird?

Regel-Mining

Von häufigen Itemsets zu Regeln

Assoziationen zwischen Items ($A \Rightarrow B$)

- Gegeben einen k-Itemset I_A^k. Wie hoch ist die Wahrscheinlichkeit, dass Item B ebenfalls gekauft wird?

Transaktion	Produkte
1	{ Milch, Butter }
2	{ Milch, Kaffee, Kuchen }
3	{ Milch, Kakao, Kuchen }
4	{ Kaffee, Zucker, Tee }
5	{ Milch, Kaffee, Zucker }
6	{ Tee, Zucker }

- $P(\text{Kaffee} \mid \text{Milch}) = 0.5$, d.h., 50% der Einkäufe von Milch führen auch zu einem Einkauf von Kaffee.

Regel-Mining

Assoziationsregeln

Definition 8 (Assoziationregel $I_A \Rightarrow I_B$)

Die Relation $I_A \Rightarrow I_B$ bezeichnet die Assoziation zweier Itemsets I_A und I_B, d.h. ein Auftreten von I_A lässt ein Auftreten von I_B vermuten.

Definition 9 (Starke Assoziationregel)

Ein Assoziationsregel $I_A \Rightarrow I_B$ wird als *stark assoziiert* bezeichnet, wenn gilt

- $\sigma_{I_A \cup I_B} \geq \sigma_{min}$ und
- $P(I_B \mid I_A) > \gamma_{min}$.

γ_{min} bezeichnet man als minimale Konfidenz in Regel $I_A \Rightarrow I_B$ und $P(I_B \mid I_A)$ bezeichnet die bedingte Wahrscheinlichkeit für Itemset I_A bei gegebenen Itemset I_B bezogen auf eine Instanzmenge X.

Regel-Mining

Extraktion von Assoziationsregeln

Beobachtung:

- Starke Assoziationsregeln können nur in häufigen Itemsets vorkommen, da $\sigma_{I_A \cup I_B} \geq \sigma_{min}$.
- Die bedingte Wahrscheinlichkeit zwischen zwei Itemsets lässt sich über den Support wie folgt abschätzen:

$$P(I_B \mid I_A) = \frac{P(I_B \cap I_A)}{P(I_A)} = \frac{\sigma_{I_A \cup I_B}}{\sigma_{I_A}}$$

Regel-Mining

Extraktion von Assoziationsregeln

Beobachtung:

- Starke Assoziationsregeln können nur in häufigen Itemsets vorkommen, da $\sigma_{I_A \cup I_B} \geq \sigma_{min}$.
- Die bedingte Wahrscheinlichkeit zwischen zwei Itemsets lässt sich über den Support wie folgt abschätzen:

$$P(I_B \mid I_A) = \frac{P(I_B \cap I_A)}{P(I_A)} = \frac{\sigma_{I_A \cup I_B}}{\sigma_{I_A}}$$

Algorithmus:

1. Ermittle die häufigsten Itemsets $\mathcal{L}$ ($\rightarrow$ A-Priori Algorithmus)
2. Ermittle starke Assoizationsregeln für jeden Itemset in den häufigsten Itemsets:
   (a) Für jedes $I \in \mathcal{L}$ erzeuge alle nicht-leeren Untermengen $s \subset I$
   (b) Für jede nicht-leere Untermenge $s \subset I$ ermittle die Konfidenz $\gamma_s = P(I \setminus s \mid s)$
   (c) Ausgabe aller Regeln $s \Rightarrow I \setminus s$ wenn $\gamma_s \geq \gamma_{min}$

Regel-Mining

Beispiel: Extraktion von Assoziationsregeln

Häufige Itemsets $\mathcal{L} = \{\{I1, I2, I5\}\}$

Assoziationsregel	γ
$\{I1, I2\} \Rightarrow \{I5\}$	2/4=0.5
$\{I1, I5\} \Rightarrow \{I2\}$	2/2=1
$\{I2, I5\} \Rightarrow \{I1\}$	2/2=1
$\{I1\} \Rightarrow \{I2, I5\}$	2/6=0.33
$\{I2\} \Rightarrow \{I1, I5\}$	2/7=0.29
$\{I5\} \Rightarrow \{I1, I2\}$	2/2=100

Regel-Mining

Evaluierung von Assoziationsregeln

Sind alle „starken Assoziationsregeln“ interessante Regeln?

- Geringer Minimum-Support liefert viele uninteressante Regeln
- Konfidenz ist hoch für Items die häufig vorkommen

Beispiel:

- 10.000 Kauftransaktionen
- 6.000 Kunden kauften Computer-Spiel, 7.500 kauften Videos und 4.000 kauften beides
- Starke Assoizationsregeln: $\{\text{Computer-Spiel}\} \Rightarrow \{\text{Videos}\}$ mit $\gamma = 0.66$ und $\sigma = 0.4$
- Aber: Die Wahrscheinlichkeit ein Video zu kaufen ist bereits 75%.
- Computer-Spiele und Videos sind negative korreliert

Regel-Mining

Evaluierung von Assoziationsregeln

Definition 10 (Korrelationsregeln)

Eine Assoziationsregeln $I_A \Rightarrow I_B$ wird als Korrelationsregel bezeichnet, wenn sie zusätzlich zu σ_{min} und γ_{min} einen Minimalen Korrelationsmaß κ_{min} erfüllt, d.h., $\kappa_{I_A \Rightarrow I_B} \geq \kappa_{min}$.

Regel-Mining

Evaluierung von Assoziationsregeln

Definition 10 (Korrelationsregeln)

Eine Assoziationsregeln $I_A \Rightarrow I_B$ wird als Korrelationsregel bezeichnet, wenn sie zusätzlich zu σ_{min} und γ_{min} einen Minimalen Korrelationsmaß κ_{min} erfüllt, d.h., $\kappa_{I_A \Rightarrow I_B} \geq \kappa_{min}$.

Korrelationsmaße:

- *Lift*: Gemeinsame Vorkommen bei stochastischer Unabhängigkeit.

$$\kappa_{I_A \Rightarrow I_B} = \frac{P(I_A \cap I_B)}{P(I_A) \cdot P(I_B)}$$

- *All_confidence:* Minimum der bedingten Wahrscheinlichkeiten.

$$\kappa_{I_A \Rightarrow I_B} = \min P(I_A \mid I_B), P(I_B \mid I_A)$$

Regel-Mining

Evaluierung von Assoziationsregeln

Korrelationsmaße (Fortsetzung):

- *Max_confidence:* Maximum der bedingten Wahrscheinlichkeiten.

$$\kappa_{I_A \Rightarrow I_B} = \max P(I_A \mid I_B), P(I_B \mid I_A)$$

- *Kulcynski Maß:* Mittelwert bedingten Wahrscheinlichkeiten.

$$\kappa_{I_A \Rightarrow I_B} = \frac{P(I_A \mid I_B) + P(I_B \mid I_A)}{2}$$

- Statistischer χ^2 Test, Kosinusmaß

Regel-Mining

Beispiel Lift

Kontingenztabelle:

	Spiel	$\overline{\text{Spiel}}$	$\sum$
Video	4.000	3.500	7.500
$\overline{\text{Video}}$	2.000	500	2.500
$\sum$	6.000	4.000	10.000

Regel-Mining

Beispiel Lift

Kontingenztabelle:

	Spiel	$\overline{\text{Spiel}}$	$\sum$
Video	4.000	3.500	7.500
$\overline{\text{Video}}$	2.000	500	2.500
$\sum$	6.000	4.000	10.000

- $P(\{\text{Spiel}\}) = 0.6$, $P(\{\text{Video}\}) = 0.75$, $P(\{\text{Spiel}, \text{Video}\}) = 0.4$
- Lift: $\frac{0.4}{0.75 \cdot 0.6} = 0.89$

$\Rightarrow$ Die Wahrscheinlichkeit das Videos und Computerspiele gemeinsam gekauft werden ist geringer als die zufällige Wahrscheinlichkeit eines gemeinsamen Kaufs.

$\Rightarrow$ Keine Korrelation

Regel-Mining

Zusammenfassung

- Identifikation häufiger Itemsets (Frequent Itemset Mining)
  - A-Priori Algorithmus (Breitensuche)
  - Pattern Growth/Eclat (Tiefensuche)
- Ableiten von Assoziationsregeln über minimale Konfidenz
- Ableiten von Korrelationsregeln über Korrelationsmaße
  - Lift, χ^2, all-support, max-support, Kosinus