Fachgruppe Informatik
Prof. Dr. A. Ittner, C. Menzer
Wintersemester 2020/21
IF18/MI18

HOCHSCHULE MITTWEIDA
University of Applied Sciences

Big Data / Data Mining

Assoziationsanalyse

Bei der Assoziationsanalyse (*pattern mining*) wird in (häufig sehr großen) Datenmengen nach Strukturen gesucht, die ein häufiges Auftreten aufweisen, z.B.:

häufige Artikelmengen (frequent itemsets) bei Darstellung der Daten als Menge (Warenkörbe, ...)

häufige Sequenzen (frequent sequences, frequent strings) bei Darstellung der Daten als Sequenz (natürlichsprachige Texte, Ausleihhistorie in Bibliothek, ...)

häufige Bäume/Graphen (frequent trees/graphs) bei Darstellung der Daten als Baum/Graph (Verbreitung von Informationen, ...)

Aus den identifizierten Mustern lassen sich daraufhin Assoziationsregeln ableiten. Diese können verwendet werden, um Vorhersagen zu treffen oder Empfehlungen zu generieren. Die Warenkorbanalyse als eine der ersten Anwendungen der Assoziationsanalyse hat die Terminologie maßgeblich geprägt (*item*, *transaction*, ...).

Beispiel:

$$\{Chips,\ Windeln\} \rightarrow \{Bier\}$$

Wer Chips und Windeln kauft, kauft häufig auch Bier.

Eine Assoziationsregel hat also die Form $A \rightarrow B$, wobei A Prämisse oder Regelrumpf und B Konsequenz oder Regelkopf genannt werden. Gegeben ist eine Menge von Items I und eine Menge von Transaktionen T, die als *Datenbasis* $D = (T_1, \ldots, T_n)$ mit $T_j \subset I$ zusammengefasst werden. Ein Item kann beispielsweise als ein Artikel und eine Transaktion als ein Einkauf/Warenkorb in einem Shop verstanden werden (vgl. Cleve und Lämmel 2014, S. 73 f., 223 f.; vgl. „Assoziationsanalyse“ 2008, S. 261).

Beispiel:

$$I = \{Bier,\ Chips,\ Cola,\ Wurst,\ Käse,\ Brot\}$$

$$D = \left\{ \begin{array}{l} (1, \{Bier,\ Cola\}), (2, \{Wurst,\ Chips\}), (3, \{Bier,\ Cola\}), (4, \{Brot,\ Wurst,\ Käse\}), \\ (5, \{Brot,\ Wurst,\ Käse,\ Bier\}), (6, \{Chips,\ Cola,\ Bier\}), (7, \{Brot,\ Wurst,\ Käse\}) \end{array} \right\}$$

Assoziationsregeln sind meist nur zu einem bestimmten Prozentsatz korrekt, da i.d.R nicht *alle* Kunden entsprechend dieser Regel einkaufen. Wir benötigen daher Interessantheitsmaße um die Qualität einer Regel zu bestimmen.

Definition: *Der* Support einer Itemmenge $X \subset I$ *ist definiert als die relative Häufigkeit dieser Itemmenge in der Datenbasis:*

$$supp(X) = \frac{|\{T \in D : X \subset T\}|}{|D|} \tag{1}$$

Wenn X unechte Teilmenge von T ist und T Element der Datenbasis ist

Definition: *Der* Support einer Assoziationsregel *ist gleich dem Support der Vereinigung von Prämisse und Konsequenz der Regel:*

$$supp(X \to Y) = supp(X \cup Y) = \frac{|\{T \in D : (X \cup Y) \subset T\}|}{|D|} \tag{2}$$

Somit ergibt sich der Support einer Assoziationsregel $X \to Y$ als Anteil der Transaktionen T der Datenbasis D, die diese Regel erfüllen. Damit wird zum Ausdruck gebracht, ob eine Kombination von Items zur Bildung einer Assoziationsregel überhaupt von Bedeutung ist. Die *Konfidenz* (confidence) einer Regel gibt dann Aufschluss über die Stärke des Zusammenhangs bzw. das Ausmaß der Gültigkeit einer Regel $X \to Y$.

Definition: *Die Konfidenz ist definiert als Anteil der Transaktionen, die sowohl X als auch Y beinhalten, an der Menge der Transaktionen, die die Prämisse X erfüllen.*

$$conf(X \to Y) = \frac{supp(X \to Y)}{supp(X)} \tag{3}$$

Um die wertvollen von den weniger wertvollen Assoziationsregeln zu trennen, werden Schwellwerte für Konfidenz und Support eingeführt. Eine Assoziationsregel ist i.d.R. umso wertvoller, je größer Support und Konfidenz sind. Die Schwellwerte für minimale Konfidenz und minimalen Support werden i.d.R. vom Nutzer festgelegt.

Beispiel (fort.)**:** *Wir legen den minimalen Support S_{min} auf 3 fest. Die folgenden* Frequent Itemsets *genügen S_{min}:*

0 Items	*1 Item*	*2 Items*	*3 Items*
∅: 7	*{Bier}: 4*	*{Brot, Käse}: 3*	*{Brot, Wurst, Käse}: 3*
	{Cola}: 3	*{Brot, Wurst}: 3*	
	{Wurst}: 4	*{Wurst, Käse}: 3*	
	{Käse}: 3	*{Bier, Cola}: 3*	
	{Brot}: 3		

Die Konfidenz der Assoziationsregel {Brot, Käse} → {Bier} z.B. beträgt:

$$conf(\{Brot, Käse\} \to \{Bier\}) = \frac{\frac{1}{7}}{\frac{3}{7}} = \frac{1}{3}$$

Aufgabe 1

a) Warum ist eine Assoziationsregel mit großem Support und großer Konfidenz nicht immer wertvoll bzw. interessant? Finden Sie ein Beispiel.

b) Welche weiteren Interessantheitsmaße können verwendet werden?

c) Welche Arten von Assoziationsregeln gibt es?

d) Welche wichtigen Algorithmen zum Erzeugen von Assoziationsregeln gibt es?

Aufgabe 2

Gegeben sind die folgenden 4 Items/Produkte (Milch, Bier, Cola, Saft) und die folgenden 8 Baskets/Warenkörbe:

B1 = {Saft}
B2 = {Saft, Milch, Bier}
B3 = {Bier, Cola}
B4 = {Cola, Saft, Milch, Bier}
B5 = {Milch, Saft, Bier}
B6 = {Bier, Milch}
B7 = {Milch, Saft}
B8 = {Bier, Saft}

a) Der minimale Support S_{min} beträgt 3 Baskets. Geben Sie alle Frequent Itemsets an, die der Bedingung von S_{min} genügen.

b) Wie groß ist die Konfidenz und der Interest für folgende Assoziationsregeln?:

R1 = {Saft} → {Bier}
R2 = {Milch, Bier} → {Cola}
R3 = {Milch, Saft} → {Bier}
R4 = {Cola} → {Bier}

Aufgabe 3

Assoziationsanalyse mit RapidMiner

Laden Sie den Datensatz *Transactions* aus den Beispieldaten des Repository.

a) Reichern Sie diesen Datensatz mit weiteren beschreibenden Informationen aus den Beispieldaten an und verbinden Sie entsprechende Tabellen.

b) Erzeugen Sie mit einem geeigneten Algorithmus Assoziationsregeln aus diesen Datensätzen.

Lösungen

1 a) [85% der Männer sind verheiratet = supp(B)
75% der verh. Männer sind regelm. Bierkonsumenten = conf(A→B)]

[26% der Männer sind schwere Raucher = supp(B)
38% der Männer, die schwere Raucher sind, sind regelm. Bierkons. = conf(A→B)]

↳ Frage ist, Risiko für Männer mittleren Alters für atherosclerosis → verheiratet sein wirklich entscheidend? → echter Zusammenhang?

b) Lift → $lift(x \to y) = \frac{supp(x \cup y)}{supp(x) \cdot supp(y)}$ noch mehr :→ https://michael.hahsler.net/research/association_rules/measures.html

→ $L() > 1$ ≙ Abhängigkeit

→ gibt ein Maß der Abhängigkeit zwischen den Größen an

Correlation → $Corr(A,B) = P(A \cup B) / P(A) \cdot P(B)$

Conviction → $Conv(x \to y) = \frac{1 - supp(y)}{1 - conf(x \to y)}$ ⇒ Wkt. x erscheint ohne y

c) single dimensional, multi dimensional, boolean, quantitativ

d) AIS, ↓ ineffizient 1) erzeugt frequent itemsets 2) erzeugt confidente und frequent assoziationsregeln

SETM, ↓ wie AIS, basiert auf SQL Basis

Apriori, ↓ erzeugt nur Assoziationsregeln, wenn diese über einem Threshhold (supp, conf) liegen

1) erzeugt frequent itemsets (supp > Thresh)
2) Erzeugung der Assoziationsregeln (conf > Thresh)

, AprioriTID, Apriorihybrid Quelle: http://www.ijsrp.org/research-paper-0513/ijsrp-p17133.pdf

2) a)

Items = {Saft, Milch, Bier, Cola} $D = \{B_1, \dots, B_8\}$ $|D| = 8$

$supp(Saft) = \frac{6}{8} = \frac{3}{4} > s_{min} = \frac{3}{8}$ ✓

$supp(Milch) = \frac{5}{8} > s_{min} = \frac{3}{8}$ ✓

$supp(Bier) = \frac{6}{8} = \frac{3}{4} > s_{min} = \frac{3}{8}$ ✓

$supp(Cola) = \frac{2}{8} = \frac{1}{4} < s_{min} = \frac{3}{8}$ ✗

b)

$conf(R_1) = \frac{supp(X \cup Y)}{supp(X)} = \frac{\frac{4}{8}}{\frac{6}{8}} = \frac{1}{2} \cdot \frac{4}{3} = \frac{1}{3} = 66,\overline{6}\,\%$

$conf(R_2) = \frac{\frac{1}{8}}{\frac{4}{8}} = \frac{1}{8_4} \cdot \frac{8 \times 1}{4_1} = \frac{1}{4} = 25\,\%$

$conf(R_3) = \frac{\frac{3}{8}}{\frac{4}{8}} = \frac{3}{8_4} \cdot \frac{2 1}{1} = \frac{3}{4} = 75\%$

$conf(R_4) = \frac{\frac{2}{8}}{\frac{2}{8}} = 1 = 100\,\%$

$lift(R_1) = \frac{supp(X \cup Y)}{supp(x) \cdot supp(y)} = \frac{\frac{4}{8}}{\frac{6}{8} \cdot \frac{6}{8}} = \frac{1}{2} < 1$ ⇒ geringe Abhängigkeit

$lift(R_2) = \frac{\frac{1}{8}}{\frac{4_1}{8_4} \cdot \frac{2 1}{8_2}} = \frac{\frac{1}{8}}{\frac{1}{8}} = 1 = 1$ ⇒ Abhängigkeit

$lift(R_3) = \frac{\frac{3}{8}}{\frac{4_1}{8_4} \cdot \frac{8_3}{8_4}} = \frac{3}{8_1} \cdot \frac{16^2}{8_1} = 2 > 1$ ⇒ große Abhängigkeit

$lift(R_4) = \frac{\frac{2}{8}}{\frac{1}{4} \cdot \frac{3}{4}} = \frac{1}{4_1} \cdot \frac{16^4}{3} = 1,\overline{3} > 1$ ⇒ große Abhängigkeit

Literatur

„Assoziationsanalyse“ (2008). In: *Datenanalyse und Statistik*. Wiesbaden: Gabler, S. 261–272. ISBN: 978-3-8349-0434-8 978-3-8349-9654-1. DOI: 10.1007/978-3-8349-9654-1_17. URL: http://link.springer.com/10.1007/978-3-8349-9654-1_17 (besucht am 12. 01. 2020).

Cleve, Jürgen und Uwe Lämmel (2014). *Data mining*. Studium. München: De Gruyter Oldenbourg. 306 S. ISBN: 978-3-486-72034-1 978-3-486-71391-6.