Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Janina Sontheim, Maximilian Hünemörder

Knowledge Discovery in Databases

WS 2019/20

Exercise 11: Association Rules, Prefix Span, Interestingness

Exercise 11-1 Association Rules

Given the following frequent itemsets extract all strong association rules with a minimum confidence of $minConf = 80\%$. Which candidates can be pruned based on anti-monotonicity?

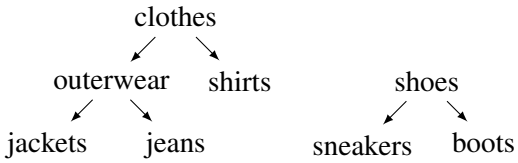
Itemset	Support
A	1.00
B	1.00
D	0.75
AB	1.00
AD	0.75
BD	0.75
ABD	0.75

#	Candidate Rule	Pruned?	Confidence	Strong
		from 2-itemsets		
1	A $\Rightarrow$ B		1.00	✓
2	A $\Rightarrow$ D		0.75	
3	B $\Rightarrow$ A		1.00	✓
4	B $\Rightarrow$ D		0.75	
5	D $\Rightarrow$ A		1.00	✓
6	D $\Rightarrow$ B		1.00	✓
		from 3-itemsets		
7	AB $\Rightarrow$ D		0.75	
8	AD $\Rightarrow$ B		1.00	✓
9	BD $\Rightarrow$ A		1.00	✓
10	A $\Rightarrow$ BD	with #2, #7		
11	B $\Rightarrow$ AD	with #4, #7		
12	D $\Rightarrow$ AB		1.00	✓

Exercise 11-2 R-Interestingness

Given the following item hierarchy and frequent itemsets decide whether the these association rules are R-interesting using $R = 1.6$ and explain why.

Itemset	Support
{clothes}	20
{outerwear}	10
{jackets}	4
{shoes}	15
{clothes, shoes}	10
{outerwear, shoes}	9
{jackets, shoes}	4



(a) clothes ⇒ shoes

(b) outerwear ⇒ shoes

(c) jackets ⇒ shoes

(a) clothes ⇒ shoes
Interesting! Rule has no ancestors.

(b) outerwear ⇒ shoes
Interesting! (wrt. support and rule (a)):

$$R \cdot \mathbb{E}(P(jackets \cup shoes)) = 1.6 \cdot 10 \cdot \frac{10}{20} = 8 < supp(outerwear \Rightarrow shoes) = 9$$

(c) jackets ⇒ shoes
Support:

$$R \cdot \mathbb{E}(P(jackets \cup shoes)) = 1.6 \cdot 9 \cdot \frac{4}{10} = 5.75 > 4 = supp(jackets \Rightarrow shoes)$$

Not Interesting!
Confidence:

$$R \cdot \mathbb{E}(P(shoes|jackets)) = 1.6 \cdot 0.9 \cdot \frac{4}{10} = 0.575 < 1.0 = conf(jackets \Rightarrow shoes)$$

Interesting! ⇒ Rule is 1.6-interesting!

Exercise 11-3 Sequential Pattern Mining

Let D be a database that contains the following five sequences.

SID	Sequence
1	ABBA
2	BBACA
3	CBAA
4	ACA
5	BAAB

In addition let *min_sup* = *40%*, i.e. there need to be 2 sequences supporting a pattern.

(a) Find all frequent sequence patterns using the *PrefixSpan* algorithm.

Start by constructing the project database for the empty prefix and count the support of 1-sequences.

SID	$D_\emptyset$ Sequence
1	ABBA
2	BBACA
3	CBAA
4	ACA
5	BAAB
	A(5)B(4)C(3)

Hence, all 1-sequences are frequent and none of those can be pruned (i.e. A, B, C are frequent). Next, create projected databases for all remaining items.

SID	D_A	D_B	D_C
1	BBA	BA	-
2	CA	BACA	A
3	A	AA	BAA
4	CA	-	A
5	AB	AAB	-
	A(5)B(2)C(2)	A(4)B(3)C(1)	A(3)B(1)C(0)

These yield the following frequent 2-sequences: AA, AB, AC, BA, BB, CA. Continue by constructing the projected databases for the 3-sequences.

SID	D_{AA}	D_{AB}	D_{AC}	D_{BA}	D_{BB}	D_{CA}
1	-	BA	-	-	A	-
2	-	-	A	A	AA	-
3	-	-	-	A	-	A
4	-	-	A	-	-	-
5	B	-	-	AB	-	-
	A(0)B(1)C(0)	A(1)B(1)C(0)	A(2)B(0)C(0)	A(3)B(1)C(0)	A(2)B(0)C(0)	A(1)B(0)C(0)

We can see that the frequent 3-sequences are ACA, BAA, BBA. Finally, the projections for the 4-sequences are given by

SID	$D_{\texttt{ACA}}$	$D_{\texttt{BAA}}$	$D_{\texttt{BBA}}$
1	-	-	-
2	-	-	A
3	-	-	-
4	-	-	-
5	-	B	-
	A(0)B(0)C(0)	A(0)B(1)C(0)	A(1)B(0)C(0)

In total, the frequent patterns are:

k	Pattern	Absolute Support	Closed	Maximal
0	-	5		
1	A	5		
	B	4		
	C	3		
2	AA	5	✓	
	AB	2	✓	✓
	AC	2		
	BA	4	✓	
	BB	3	✓	
	CA	3	✓	
3	ACA	2	✓	✓
	BAA	3	✓	✓
	BBA	2	✓	✓

(b) Which patterns are maximal? Which are closed?

c.f. (a)