Solutions Frequent Pattern Mining 2018

Exercise 1: Frequent Item Set Mining

(a) Level 1:

candidate	support	frequent?
A	3	✓
B	5	✓
C	6	✓
D	5	✓
E	2	✓
F	1	✗

Level 2:

candidate	support	frequent?
AB	3	✓
AC	2	✓
AD	2	✓
AE	0	✗
BC	4	✓
BD	3	✓
BE	0	✗
CD	4	✓
CE	1	✗
DE	1	✗

Level 3:

candidate	support	frequent?
ABC	2	✓
ABD	2	✓
ACD	1	✗
BCD	2	✓

$ABCD$ is not a level-4 candidate because its level-3 subset ACD is not frequent.

(b) First we determine the generators, and then we obtain the closed frequent item sets by taking the closure of the generators. To determine the generators, we apply Apriori with an additional pruning step: an item set is pruned when it has a subset with the same support. The normal frequency pruning is indicated by ✗, the additional A-close pruning by ✂.

Level 1:

candidate	support	generator?
A	3	✓
B	5	✓
C	6	✓
D	5	✓
E	2	✓
F	1	✗

Level 2:

candidate	support	generator?
AB	3	✂
AC	2	✓
AD	2	✓
AE	0	✗
BC	4	✓
BD	3	✓
BE	0	✗
CD	4	✓
CE	1	✗
DE	1	✗

AB is pruned because its subset A has the same support.
Level 3:

candidate	support	generator?
ACD	1	✗
BCD	2	✓

Notice that ABC and ABD are no longer level-3 candidates, because AB is not a level-2 generator. There are no level-4 candidates.

Next we determine the closures of the generators. These closures together form the set of closed frequent item sets. A closure has the same support as its generator, and can be obtained by taking the intersection of all transactions in which the generator occurs.

generator	closure	support
A	AB	3
B	B	5
C	C	6
D	D	5
E	E	2
AC	ABC	2
AD	ABD	2
BC	BC	4
BD	BD	3
CD	CD	4
BCD	BCD	2

(c) The maximal frequent item sets are: E, ABC, ABD, and BCD.

(d)
$$\text{confidence}(A \to C) = \frac{s(AC)}{s(A)} = \frac{2}{3}$$
$$\text{lift}(A \to C) = \frac{s(AC) \times |db|}{s(A) \times s(C)} = \frac{2 \times 8}{3 \times 6} = \frac{16}{18} = \frac{8}{9},$$
where $|db|$ denotes the number of transactions in the data base. Since the lift is smaller than 1, the rule would not be considered interesting according to this measure.

The formula for lift used above can be derived from its definition as follows:
$$\text{lift}(X \to Y) = \frac{P(Y \mid X)}{P(Y)} = \frac{P(X,Y)}{P(X)P(Y)} = \frac{s(XY)/|db|}{s(X)/|db| \times s(Y)/|db|} = \frac{s(XY) \times |db|}{s(X) \times s(Y)}$$
Note that we write $s(XY)$ as a shorthand of $s(X \cup Y)$.

Exercise 2: Frequent Sequence Mining

(a) Level 1:

candidate	support	frequent?
A	5	✓
U	4	✓
H	3	✓
B	3	✓

Level 2:

candidate	support	frequent?
AA	2	✗
AU	4	✓
AH	3	✓
AB	2	✗
UA	2	✗
UU	3	✓
UH	3	✓
UB	2	✗
HA	2	✗
HU	3	✓
HH	0	✗
HB	1	✗
BA	1	✗
BU	0	✗
BH	0	✗
BB	0	✗

Level 3:

candidate	support	frequent?
AUU	3	✓
AUH	3	✓
AHU	3	✓
UUU	0	✗
UUH	0	✗
UHU	3	✓
HUU	0	✗

Level 4:

candidate	support	frequent?
AUHU	3	✓

There is no level 5 candidate. We can combine AUHU with itself to generate pre-candidate AUHUU but this contains the infrequent subsequence HUU.

(b) AUHU and B are maximal.

(c) A, B, AU, AUHU.

Exercise 3: Frequent Tree Mining

(a)+(b) Let's denote the nodes of d_4 by v and the nodes of $T = aa \uparrow c$ by w. The nodes are numbered according to the pre-order (depth-first) traversal of the tree. T is an embedded subtree of d_4, and the corresponding matching function is:

$$\phi(w_1) = v_1 \quad \phi(w_2) = v_2 \quad \phi(w_3) = v_5$$

Verify that this matching function satisfies all the requirements:

1. The labeling is preserved: $L(w_1) = L(v_1) = a$, $L(w_2) = L(v_2) = a$, $L(w_3) = L(v_5) = c$.
2. The ancestor-descendant relation is preserved:
   - $w_1 \in \pi^*(w_2)$ and $v_1 \in \pi^*(v_2)$.
   - $w_1 \in \pi^*(w_3)$ and $v_1 \in \pi^*(v_5)$.
   - $w_2 \notin \pi^*(w_3)$ and $v_2 \notin \pi^*(v_5)$.
   - $w_3 \notin \pi^*(w_2)$ and $v_5 \notin \pi^*(v_2)$.
3. The ordering is preserved:
   - $w_1 < w_2$ and $v_1 < v_2$.
   - $w_1 < w_3$ and $v_1 < v_5$.
   - $w_2 < w_3$ and $v_2 < v_5$.

T is not an induced subtree of d_4. The matching function that worked for the embedded subtree relation does not work here, because it does not preserve the parent-child relationship: $w_1 = \pi(w_3)$ but $v_1 \neq \pi(v_5)$.

(c) No, it is not. For example, the matching function

$$\phi(w_1) = v_1 \quad \phi(w_2) = v_3 \quad \phi(w_3) = v_2$$

does not preserve the order relation, because $w_2 < w_3$, but $v_3 \not< v_2$.

(d) No.

(e) Yes. The matching function

$$\phi(w_1) = v_1 \quad \phi(w_2) = v_2 \quad \phi(w_3) = v_3$$

works.

(f) Yes. The matching functions

$$\phi(w_1) = v_1 \quad \phi(w_2) = v_2 \quad \phi(w_3) = v_3$$
$$\phi(w_1) = v_1 \quad \phi(w_2) = v_2 \quad \phi(w_3) = v_4$$

both work.

(g)+(h) It occurs 8 times as an embedded subtree. The distinct matching functions are: $\phi(w_1) = v_1$, and then

1. $\phi(w_2) = v_2$ and $\phi(w_3) = v_3$
2. $\phi(w_2) = v_2$ and $\phi(w_3) = v_5$
3. $\phi(w_2) = v_3$ and $\phi(w_3) = v_5$
4. $\phi(w_2) = v_2$ and $\phi(w_3) = v_4$
5. $\phi(w_2) = v_2$ and $\phi(w_3) = v_6$
6. $\phi(w_2) = v_3$ and $\phi(w_3) = v_6$
7. $\phi(w_2) = v_4$ and $\phi(w_3) = v_5$
8. $\phi(w_2) = v_4$ and $\phi(w_3) = v_6$

The first three matching functions are also induced subtrees.
The FREQT RMO-list is (v_3, v_5).

Exercise 4: Anti-monotonicity

No, for example A is a subsequence of AB, but AB occurs twice as a subsequence in ABB, and A just once.

Alternative definition of "distinct occurrence": $\phi_1(i) \neq \phi_2(i)$ for all positions i in the pattern sequence.

Exercise 5: Transitivity of the subsequence relation

If the subsequence relation is transitive, then from $\mathbf{r}_1 \subseteq \mathbf{r}_2$, and $\mathbf{r}_2 \subseteq \mathbf{s}$ (where $\mathbf{s} \in \mathbf{D}$) it follows that $\mathbf{r}_1 \subseteq \mathbf{s}$ and therefore that the support of $\mathbf{r}_1$ is at least as big as the support of $\mathbf{r}_2$.

It is given that $\mathbf{q} \subseteq \mathbf{r}$, and $\mathbf{r} \subseteq \mathbf{s}$. So there exist one-to-one mappings ϕ_{qr} and ϕ_{rs} that preserve the labels and order. Now define the mapping $\phi_{qs}(i) = \phi_{rs}(\phi_{qr}(i))$. This mapping has the following properties:

1. $\mathbf{q}[i] = \mathbf{s}[\phi_{qs}(i)]$ because $\mathbf{q}[i] = \mathbf{r}[\phi_{qr}(i)]$ and $\mathbf{r}[\phi_{qr}(i)] = \mathbf{s}[\phi_{rs}(\phi_{qr}(i))]$.
2. $i < j \Rightarrow \phi_{qs}(i) < \phi_{qs}(j)$, because $i < j \Rightarrow \phi_{qr}(i) < \phi_{qr}(j)$ and $\phi_{qr}(i) < \phi_{qr}(j) \Rightarrow \phi_{rs}(\phi_{qr}(i)) < \phi_{rs}(\phi_{qr}(j))$.

Also, since ϕ_{qr} is one-to-one, and ϕ_{rs} is one-to-one, it follows that ϕ_{qs} is one-to-one as well. Hence, the mapping ϕ_{qs} satisfies all the required properties.

Note: Actually, it follows from $i < j \Rightarrow \phi(i) < \phi(j)$ that ϕ is one-to-one, so the latter restriction is superfluous. The "proof" above that ϕ_{qs} is one-to-one is therefore also superfluous: we have already shown that ϕ_{qs} preserves the order. Thanks to Max Hessey and Casper Hagenaars for noting this.

Exercise 6: Variations on a theme

We define the subsequence relation as follows: $S^1 = (X_1 X_2 \dots X_k)$ is a subsequence of $S^2 = (Y_1 Y_2 \dots Y_m)$ (denoted $S^1 \preceq S^2$) if there exists a mapping

$$\phi : [1, k] \to [1, m],$$

such that

1. $X_i \subseteq Y_{\phi(i)}$, and
2. $i < j \Rightarrow \phi(i) < \phi(j)$.

We define support as:

$$\sup(R) = \left|\{S^i \in D : R \preceq S^i\}\right|$$

To show that the anti-monotonicity property holds, it suffices to show that the subsequence relation is transitive, that is,

$$S^1 \preceq S^2 \text{ and } S^2 \preceq S^3 \Rightarrow S^1 \preceq S^3$$

So assume that $S^1 \preceq S^2$, and $S^2 \preceq S^3$. This means there exist mappings ϕ_{12} and ϕ_{23} that satisfy the subset and order constraints. Now define the mapping $\phi_{13}(i) = \phi_{23}(\phi_{12}(i))$. This mapping has the following properties:

1. $S^1_i \subseteq S^3_{\phi_{13}(i)}$ because $S^1_i \subseteq S^2_{\phi_{12}(i)}$ and $S^2_{\phi_{12}(i)} \subseteq S^3_{\phi_{23}(\phi_{12}(i))}$.
2. $i < j \Rightarrow \phi_{13}(i) < \phi_{13}(j)$, because $i < j \Rightarrow \phi_{12}(i) < \phi_{12}(j)$ and $\phi_{12}(i) < \phi_{12}(j) \Rightarrow \phi_{23}(\phi_{12}(i)) < \phi_{23}(\phi_{12}(j))$.

Therefore, S^1 is a subsequence of S^3. From this the anti-monotonicity property

$$S^1 \preceq S^2 \Rightarrow \sup(S^1) \geq \sup(S^2)$$

follows.

Notice that in this proof we left out the requirement that ϕ is a one-to-one mapping. We explained in the note at exercise 5 why it is superfluous.