DSAA 5002 - Data Mining and Knowledge Discovery in Data Science

(Fall Semester 2023)

Mark:77/100

Homework 1

50015627 JIANG Zhuoyang

Q1 [15 Marks] 5/15

Given the transaction database below, set the minimum support count to 2 and the minimum confidence level to 60% to find the strong association rule. Generate the set C_3 of the candidate 3-itemset , using prunning on Apriori principle.

TID	Item
T1	A,C,D
T2	B,C,E
T3	A,B,C,E
T4	B,E
T5	A,C,E

With Apriori, get the frequent k-itemsets and generate the candidate k+1-itemsets as below:

1st Scan → C_1

Itemset	sup
{A}	3
{B}	3
{C}	4
{D}	1
{E}	4

prune → L_1

Itemset	sup
{A}	3
{B}	3
{C}	4
{E}	4

generate → C_2

Itemset
{A,B}
{A,C}
{A,E}
{B,C}
{B,E}
{C,E}

2nd Scan → C_2

Itemset	sup
{A,B}	1
{A,C}	3
{A,E}	2
{B,C}	2
{B,E}	3
{C,E}	3

prune → L_2

Itemset	sup
{A,C}	3
{A,E}	2
{B,C}	2
{B,E}	3
{C,E}	3

generate → C_3

Itemset
{A,C,E}
{B,C,E}

($C_3 = \{X \mid X \subset L_2 \bowtie L_2 \text{ with } AB \notin X\}$)

3rd Scan → C_3

Itemset	sup
{A,C,E}	2
{B,C,E}	2

$\Rightarrow C_3 = \{\{A,C,E\}, \{B,C,E\}\}$

Q2 [15 Marks] 15/15

Reducing the transactions using dynamic hashing and pruning(DHP) algorithm. Set the minimum support count to 2.

Hash function bucket #= h({x y}) = ((order of x)*10+(order of y)) % 7

TID	Item
T1	A,B,C
T2	B,D,E
T3	A,B,D,E
T4	B,E

1st scan → C_1

Itemset	sup
{A}	2
{B}	4
{C}	1
{D}	2
{E}	3

prune → L_1

Itemset	sup
{A}	2
{B}	4
{D}	2
{E}	3

Solution: $L_1 = \{\{A\},\{B\},\{D\},\{E\}\}$. for all 2-itemsets in transactions creat the hash table below:

bucket	0	1	2	3	4	5	6
count	1	1	1	4	3	2	1
itemset	{A,D}	{A,E}	{B,C}	{B,D} {D,E} {B,D} {D,E}	{B,E} {B,E} {B,E}	{A,B} {A,B}	{A,C}

⇒ Generate C_2:

2, M L_1 →

Itemset	# in bucket of the Itemset
{A B}	2
{A D}	1
{A E}	1
{B D}	4
{B E}	3
{D E}	4

prune → $C_2 = \{\{B,D\},\{D,E\},\{B,E\},\{A,B\}\}$

C_2

Itemset	Count
{A,B}	2
{B D}	2
{B E}	2
{D E}	2

⇒ $L_2 = \{\{B,D\},\{D,E\},\{B,E\},\{A,B\}\}$

so that we have:

D_2:

T_1: {A, B} ⇒ Discarded

T_2: {B, D}, {B, E}, {D, E} ⇒ keep {B, D, E}

T_3: {A, B}, {B, D}, {B, E}, {D, E} ⇒ keep {B, D, E}

T_4: {B, E} ⇒ Discarded

(item A occus just in one 2-itemset {A,B}. so that 3-itemset with A will be discarded.)

⇒ $D_3 = \{\{T_2, \{B, D, E\}\}, \{T_3, \{B, D, E\}\}\}$

Q3 [35 Marks]

0. An itemset X is said to be a frequent itemset if the frequency count of X is at least a given support threshold.
1. An itemset Y is a proper super-itemset of X if X $\subset$Y and X $\neq$Y.
2. An itemset X is said to be a closed frequent itemset if (1) X is frequent and (2) there exists no proper super itemset Y of X such that Y is frequent and Y has the same frequency count as X.
3. An itemset X is said to be a maximal frequent itemset if (1) X is frequent and (2) there exists no proper super-itemset Y of X such that Y is frequent.
4. Let F be the set of (traditional) frequent itemsets without specifying the frequency of itemsets.
5. Let F_c be the set of (traditional) frequent itemsets each of which is associated with a frequency in the dataset.

   For example, if there are three frequent itemsets, $\{I_1\}$ with frequency 4, $\{I_2\}$ with frequency 5, and $\{I_1, I_2\}$ with frequency 3, F = $\{\{I_1\}, \{I_2\}, \{I_1, I_{12}\}\}$ and Fc = $\{<\{I_1\}, 4>, <\{II_2\}, 5>, <\{I_1, I_2\}, 3>\}$.
6. Similarly, let C be the set of closed frequent itemsets without specifying the frequency of itemsets.

   Let C_c be the set of closed frequent itemsets each of which is associated with a frequency in the dataset.
7. Let M be the set of maximal frequent itemsets without specifying the frequency of itemsets.

   Let M_c be the set of maximal frequent itemsets each of which is associated with a frequency in the dataset.

The following shows six transactions with four items. Each row corresponds to a transaction where 1 corresponds to a presence of an item and 0 corresponds to an absence.

A	B	C	D
0	0	1	1
1	1	0	0
0	0	1	1
1	0	1	1
1	0	0	0
0	0	0	1

Suppose that the support threshold is 2.

Solution :

(a) (i) What is F_c? (ii) What is C_c? (iii) What is M_c? **(5 Marks)**

(i) $F_c = \{<\{A\},3>, <\{C\},3>, <\{D\},4>, <\{C,D\},3>\}$

(ii) $C_c = \{<\{A\},3>, <\{D\},4>, <\{C,D\},3>\}$

(iii) $M_c = \{<\{A\},3>, <\{C,D\},3>\}$

3/5

no process

(b) (i) What are the advantages and the disadvantages of using closed frequent itemsets compared with traditional frequent itemsets? **(5 Marks)**

5/5

1) Advantages: ① They provide a relatively compact representation of frequent itemsets. ② They contain the complete support information of their subsets. So that we can not only obtain the non-frequent itemsets but also obtain their support count without any additional pass.

2) Disadvantages: Their representation is not as compact as Maximal frequent itemsets.

(ii) What are the advantages and the disadvantages of using closed frequent itemsets compared with maximal frequent itemsets? **(5 Marks)**

4/5

1) Advantages: They can reach the most compact representation of frequent itemset in which all of the frequent itemsets can be obtained.

2) Disadvantages: They don't contain the complete support information of their subset, and an additional pass is needed.

(c) Please adapt algorithm FP-growth with the use of the FP-tree to find all closed frequent itemset. Please write down how to adapt algorithm FP-growth and illustrate the adapted algorithm with the above example. **(20 Marks)**

10/20

i) Construct the FP-Tree:

① Scan and generate L_1:

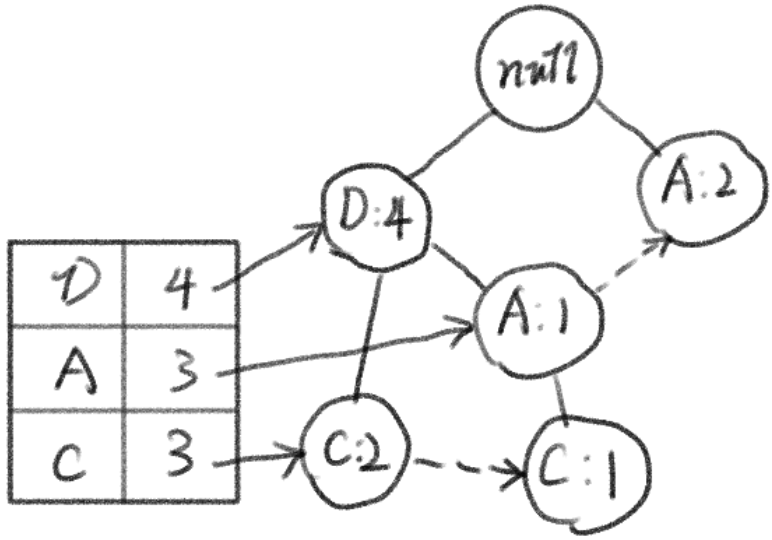
$\{<D:4>, <A:3>, <C:3>\}$

② Creat the root node:

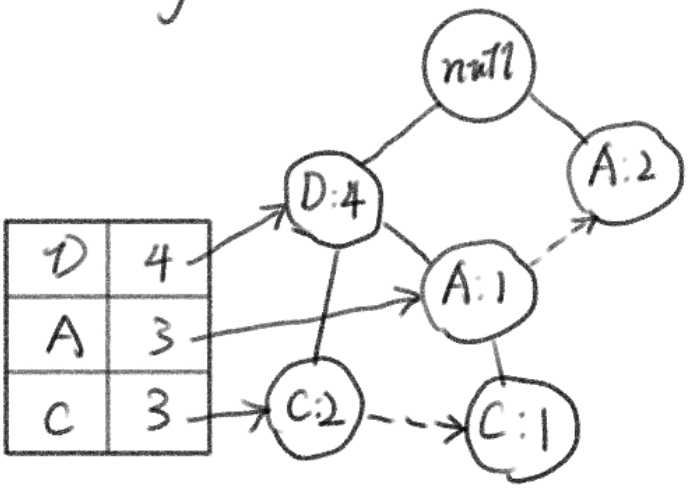
D	4
A	3
C	3

null

③ Scan and build the FP-Tree:



ii) FP-Growth for pattern mining:



① Start from the least frequent 1-pattern. here. we have C.

⇒ for C:

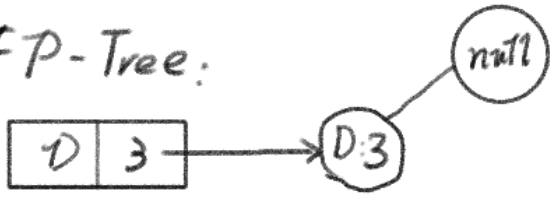
1° construct its conditional pattern base, for C. we have:

<D, C:2>, <D, A, C:1>

⇒ conditional pattern base:

<D:2>, <D, A:1>

2°. construct its conditional FP-Tree:



3°. Generate all frequent patterns for C: {<D, C:3>}

② for A: <D, A:1>, <A:2>

1° construct its conditional pattern base: null.

2° construct its conditional FP-Tree: null

③ for D: <D:4>

1° construct its conditional pattern base: null.

2° construct its conditional FP-Tree: null

iii) So that, the resulted frequent itemsets:

{<D:4>, <A:3>, <C:3>, <D, C:3>}

⇒ the resulted closed frequent itemsets:

{<D:4>, <A:3>, <D, C:3>}

Q4 [35 Marks] 35/35 good

A GSP Example: Suppose now we have 5 events: 'Upload Songs', 'Add Tags', 'Share', 'Listen' and 'Commet'. Let min-support be 40%. The sequence database of a Music Platform is shown in following table: (2/5)

Object	Sequence
A	<{ 'Upload Songs', 'Add Tags'}>
B	<{ 'Upload Songs', 'Share'}>
C	<{ 'Upload Songs'}, { 'Share', 'Listen'}>
D	<{ 'Upload Songs'}, { 'Upload Songs', 'Add Tags'}, {'Listen'}>
E	< {'Listen'}, { 'Add Tags', 'Comment'}, { 'Share', 'Listen'}>

Use the first letter to represent each event: 'Upload Songs' = U;
'Add Tags' = A; 'Share' = S; 'Listen' = L; 'Comment' = C

Please answer the following questions:

(a) Make the first pass over the sequence database to yield all the 1-element **frequent** sequences and what is the corresponding support? **(5 Marks)**

1- element frequent sequence & their support:

<{U}>: s=80%; <{A}>: s=60%; <{S}>: s=60%; <{L}>: s=60%

(b) Based on (a), do the 2-sequences Candidate Generation and Candidate Pruning. **(10 Marks)**

1) 2- element frequent sequence generation:

<{U,A}>, <{U,S}>, <{U,L}>, <{A,S}>, <{A,L}>, <{S,L}>,
<{U},{U}>, <{U},{A}>, <{U},{S}>, <{U},{L}>, <{S},{S}>, <{S},{U}>, <{S},{A}>, <{S},{L}>,
<{A},{A}>, <{A},{U}>, <{A},{S}>, <{A},{L}>, <{L},{L}>, <{L},{U}>, <{L},{A}>, <{L},{S}>

2) Candidate pruning:

for all of generated 2-sequences don't have infrequent 1-subsequence, so the pruned result shows below:

<{U,A}>, <{U,S}>, <{U,L}>, <{A,S}>, <{A,L}>, <{S,L}>,
<{U},{U}>, <{U},{A}>, <{U},{S}>, <{U},{L}>, <{S},{S}>, <{S},{U}>, <{S},{A}>, <{S},{L}>,
<{A},{A}>, <{A},{U}>, <{A},{S}>, <{A},{L}>, <{L},{L}>, <{L},{U}>, <{L},{A}>, <{L},{S}>

(c) What is the **frequent** 2-sequences based on the results of (b)? **(5 Marks)**

Candidate elimination: <{U,S}>, <{U,L}>, <{A,S}>, <{A,L}>, <{U},{A}>, <{U},{S}>, <{A},{U}>, <{A},{S}>, <{S},{U}>, <{S},{A}>, <{S},{L}>, <{L},{U}>, <{L},{A}>, <{L},{S}>, <{U},{U}>, <{A},{A}>, <{S},{S}> <{L},{L}> are eliminated for their support less than 40%

⇒ result: <{U,A}>, <{S,L}>, <{U},{L}>, <{A},{L}>

(d) Based on (c), do the 3-sequences Candidate Generation and Candidate Pruning. When a sequence should be pruned, you need to explain why. **(10 Marks)**

1) 3-sequences Candidate Generation by merging:

<{U,A},{L}>, <{U},{S,L}>, <{A},{S,L}>

2) Candidate Pruning:

for <{U,A},{L}>: <{U,A}>, <{U},{L}>, <{A},{L}> are all frequent

for <{U},{S,L}>: <{U},{S}> is infrequent, prunned.

for <{A},{S,L}>: <{A},{S}> is infrequent, prunned.

(e) What is the **frequent** 3-sequences based on the results of (d)? Please calculate the support. **(5 Marks)**

for <{U,A},{L}>: support = 20% < 40%. eliminated

There were not any Candidate 3-sequences have their support ≥ min-support.