Assignment 2 - TDT4300

Hermann Owren Elton, Olaf Rosendahl

February 17, 2023

1 Apriori Algorithm

a)

First we create a 1-itemset by counting the amount of market basket transactions each item is in, this will be the support count:

Item	Support count σ
A	6
B	8
C	10
D	4
E	5
F	3
G	8
H	7

The assignment specifies that we should find all frequent itemsets with minimum support of 0.5 (50%), being a support count of 5 (since 10 * 0.5 = 5). Therefore, D and F can be removed. We continue and generate the 2-itemset.

Item		Support count σ
A	B	5
A	C	6
A	E	1
A	G	6
A	H	5
B	C	8
B	E	4
B	G	7
B	H	5
C	E	4
C	G	8
C	H	7
E	G	3
E	H	3
G	H	5

Once again, the itemsets with support count less than 5 are removed, and we then generate the 3-itemset.

Item			Support count σ
A	B	C	5
A	B	G	5
A	B	H	4
A	C	G	6
A	C	H	5
A	G	H	5
B	C	G	7
B	C	H	5
C	G	H	5

Once again, the itemsets with support count less than 5 are removed, and we then generate the 4-itemset.

Item				Support count σ
A	B	C	G	5
A	C	G	H	5
A	B	C	H	4
B	C	G	H	4

Here we see that the itemsets of 4 items is [{A,B,C,G}, {A,C,G,H}]

b)

The first step in the process is finding all the combinations of this set, before we calculate the confidence of each association rule. The following function is used to calculate confidence: $c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

Rule $X \rightarrow Y$	$\sigma(X)$	Confidence	Accepted
$\{A, B, C\} \rightarrow \{G\}$	5	$\frac{5}{5} = 1$	Yes
$\{A, G, B\} \rightarrow \{C\}$	5	$\frac{5}{5} = 1$	Yes
$\{A, C, G\} \rightarrow \{B\}$	6	$\frac{5}{6} = 0.83\overline{3}$	Yes
$\{B, C, G\} \rightarrow \{A\}$	7	$\frac{5}{7} = 0.714$	No
$\{G\} \rightarrow \{A, B, C\}$	8	$\frac{5}{8} = 0.625$	No
$\{C\} \rightarrow \{A, G, B\}$	10	$\frac{5}{10} = 0.5$	No
$\{B\} \rightarrow \{A, C, G\}$	8	$\frac{5}{8} = 0.625$	No
$\{A\} \rightarrow \{B, C, G\}$	6	$\frac{5}{6} = 0.83\overline{3}$	Yes
$\{A, B\} \rightarrow \{C, G\}$	5	$\frac{5}{5} = 1$	Yes
$\{C, G\} \rightarrow \{A, B\}$	8	$\frac{5}{8} = 0.625$	No
$\{A, C\} \rightarrow \{B, G\}$	6	$\frac{5}{6} = 0.83\overline{3}$	Yes
$\{B, G\} \rightarrow \{A, C\}$	7	$\frac{5}{7} = 0.714$	No
$\{A, G\} \rightarrow \{B, C\}$	6	$\frac{5}{6} = 0.83\overline{3}$	Yes
$\{B, C\} \rightarrow \{A, G\}$	8	$\frac{5}{8} = 0.625$	No

From this table we can se that there are 7 assicuation rules that will be generated.

Rule $X \to Y$	$\sigma(X)$	Confidence	Accepted
$\{A, C, G\} \to \{H\}$	6	$\frac{5}{6} = 0.83\bar{3}$	Yes
$\{A, H, C\} \to \{G\}$	5	$\frac{5}{5} = 1$	Yes
$\{A, G, H\} \to \{C\}$	5	$\frac{5}{5} = 1$	Yes
$\{C, G, H\} \to \{A\}$	5	$\frac{5}{5} = 1$	Yes
$\{H\} \to \{A, C, G\}$	7	$\frac{5}{7} = 0.714$	No
$\{G\} \to \{A, H, C\}$	8	$\frac{5}{8} = 0.625$	No
$\{C\} \to \{A, G, H\}$	10	$\frac{5}{10} = 0.5$	No
$\{A\} \to \{C, G, H\}$	6	$\frac{5}{6} = 0.83\bar{3}$	Yes
$\{A, C\} \to \{G, H\}$	6	$\frac{5}{6} = 0.83\bar{3}$	Yes
$\{G, H\} \to \{A, C\}$	5	$\frac{5}{5} = 1$	Yes
$\{A, H\} \to \{C, G\}$	5	$\frac{5}{5} = 1$	Yes
$\{C, G\} \to \{A, H\}$	8	$\frac{5}{8} = 0.625$	No
$\{A, G\} \to \{H, C\}$	6	$\frac{5}{6} = 0.83\bar{3}$	Yes
$\{H, C\} \to \{A, G\}$	7	$\frac{5}{7} = 0.714$	No

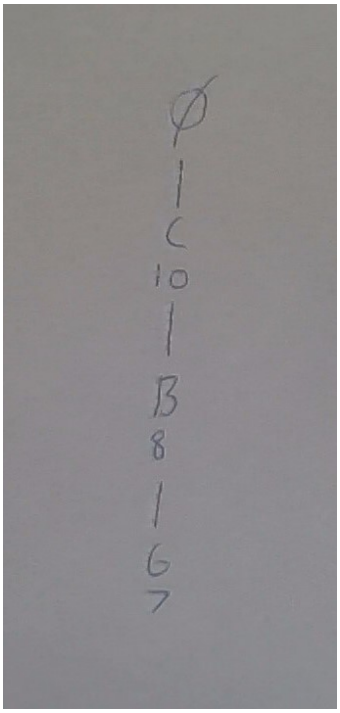
From this table we can se that there are 9 assicuation rules that will be generated.

2 FP-Growth Algorithm

The transactions sorted in to the order of the 1-itemset:

TID	Items
110	{C,G,H,A,F }
111	{C,B,G,E,D }
112	{C,B,H,E,F}
113	{C,B,G,A}
114	{C,H,E,D}
115	{C,B,G,H,A}
116	{C,B,G,H,A,D}
117	{C,B,G,E}
118	{C,B,G,H,A,F}
119	{C,B,G,H,A,E,D}

Next, we have to generate the tree. When we come to an item less frequent than the minsup count 0.5 (5), we prune it. For our case, we prune the items {A,D,E,F,H}.



Then we create a table for the frequent patterns:

Items	Conditional Pattern Base	Conditional FP-tree	Frequent Patterns
C	$\{Ø : 10\}$	$\{Ø : 10\}$	$\{C : 10\}$
B	$\{C : 8\}$	$\{C : 8\}$	$\{CB : 8\}$
G	$\{CB : 7\}$	$\{B : 7\}$	$\{CBG : 7\}$

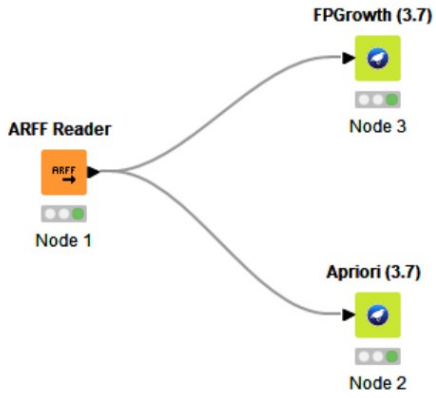
Conditional Pattern Base: the basepath of each of the occurences in the tree. In other words; the path to the correct item, with the corresponding support count for the item.

Conditional FP-tree: the support count for each item in the conditional pattern tree

Frequent Patterns: all patterns for a given item in the conditional pattern tree.

3 KNIME

First we created a new wokflow, and found the arff-reader and apriori and fpgrowth in the node repository. After adding these into the workflow, next we connected both algorithms to the arff reader, configured the path to the file containing the data, and configured the apriori-properties given in the assignment: lowerBoundMinSupport to 0.5 and minMetric to 0.8. The workflow and output are presented below:



Weka Node View - 3:2 - Apriori (3.7)

File

```
Apriori
=======

Minimum support: 0.75 (7 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 5

Generated sets of large itemsets:

Size of set of large itemsets L(1): 4

Size of set of large itemsets L(2): 4

Size of set of large itemsets L(3): 1

Best rules found:

 1. G=t 8 ==> C=t 8    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. B=t 7 ==> C=t 7    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. B=t 7 ==> G=t 7    <conf:(1)> lift:(1.25) lev:(0.14) [1] conv:(1.4)
 4. H=t 7 ==> C=t 7    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. B=t G=t 7 ==> C=t 7    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. B=t C=t 7 ==> G=t 7    <conf:(1)> lift:(1.25) lev:(0.14) [1] conv:(1.4)
 7. B=t 7 ==> C=t G=t 7    <conf:(1)> lift:(1.25) lev:(0.14) [1] conv:(1.4)
 8. G=t 8 ==> B=t 7    <conf:(0.88)> lift:(1.25) lev:(0.14) [1] conv:(1.2)
 9. C=t G=t 8 ==> B=t 7    <conf:(0.88)> lift:(1.25) lev:(0.14) [1] conv:(1.2)
10. G=t 8 ==> B=t C=t 7    <conf:(0.88)> lift:(1.25) lev:(0.14) [1] conv:(1.2)
```

Weka Node View - 3:3 - FPGrowth (3.7)

File

```
FPGrowth found 13 rules (displaying top 10)

 1. [G=t]: 8 ==> [C=t]: 8   <conf:(1)> lift:(1) lev:(0) conv:(0)
 2. [H=t]: 7 ==> [C=t]: 7   <conf:(1)> lift:(1) lev:(0) conv:(0)
 3. [B=t]: 7 ==> [C=t]: 7   <conf:(1)> lift:(1) lev:(0) conv:(0)
 4. [E=t]: 6 ==> [C=t]: 6   <conf:(1)> lift:(1) lev:(0) conv:(0)
 5. [A=t]: 6 ==> [C=t]: 6   <conf:(1)> lift:(1) lev:(0) conv:(0)
 6. [B=t]: 7 ==> [G=t]: 7   <conf:(1)> lift:(1.25) lev:(0.14) conv:(1.4)
 7. [A=t]: 6 ==> [G=t]: 6   <conf:(1)> lift:(1.25) lev:(0.12) conv:(1.2)
 8. [B=t]: 7 ==> [C=t, G=t]: 7   <conf:(1)> lift:(1.25) lev:(0.14) conv:(1.4)
 9. [C=t, B=t]: 7 ==> [G=t]: 7   <conf:(1)> lift:(1.25) lev:(0.14) conv:(1.4)
10. [G=t, B=t]: 7 ==> [C=t]: 7   <conf:(1)> lift:(1) lev:(0) conv:(0)
```

4 Compact Representation of Frequent Itemsets

In order to solve this, we'll use the algorithm 4.4 from page 244 in Tan et al. The algorithm is:

Algorithm 4.4 Support counting using closed frequent itemsets.

1: Let C denote the set of closed frequent itemsets and F denote the set of all frequent itemsets.
2: Let $k_{\max}$ denote the maximum size of closed frequent itemsets
3: $F_{k_{\max}} = \{f | f \in C,\ |f| = k_{\max}\}$ {Find all frequent itemsets of size $k_{\max}$.}
4: **for** $k = k_{\max} - 1$ down to 1 **do**
5: $F_k = \{f | f \in F,\ |f| = k\}$ {Find all frequent itemsets of size k.}
6: **for** each $f \in F_k$ **do**
7: **if** $f \notin C$ **then**
8: $f.support = \max\{f'.support | f' \in F_{k+1},\ f \subset f'\}$
9: **end if**
10: **end for**
11: **end for**

First we find that $k_{max} = 4$ as the maximum itemset is $\{a, c, d, e\}$.

Ten we loop through all k from $k_{max} - 1$ down to 1, and find all the frequent itemsets and find the maximum support count for a given itemset.

For $K = 3$, the frequent itemsets of size 3 are $\{a, c, e\}, \{a, d, c\}, \{c, d, e\}$

The calculated support counts are as follows:

$\{a, c, e\} = max\{acde.support\} = max\{5\} = 5$

$\{a, d, c\} = max\{acde.support\} = max\{5\} = 5$

$\{c, d, e\} = max\{acde.support\} = max\{5\} = 5$

For $K = 2$, the frequent itemsets of size 2 are $\{a, b\}, \{a, c\}, \{a, e\}, \{c, d\}, \{c, e\}$.

The calculated support counts are as follows:

$\{a, b\} = max\{abe.support\} = max7 = 7$

$\{a, c\} = max\{acd.support, ace.support\} = max6, 5 = 6$

$\{a, e\} = max\{abe.support, ace.support, ade.support\} = max7, 5, 5 = 7$

$\{c, d\} = max\{acd.support, cde.support\} = max\{6, 5\} = 6$

$\{c, e\} = max\{ace.support, cde.support\} = max\{5, 5\} = 5$

For $K = 1$, the frequent itemsets of size 1 are $\{a\}, \{c\}, \{e\}$.

The calculated support counts are as follows:

$\{a\} = max\{ab.support, ac.support, ad.support, ae.support\} = max\{7, 6, 11, 7\} = 11$

$\{c\} = max\{ac.support, cd.support, ce.support\} = max\{6, 6, 5\} = 6$

$\{e\} = max\{ae.support, be.support, ce.support, de.support\} = max\{7, 8, 5, 6\} = 8$

All the frequent items are then:

All frequent itemsets
$\{a\} : 11, \{b\} : 10, \{c\} : 6, \{d\} : 13, \{e\} : 8$
$\{ab\} : 7, \{ac\} : 6, \{ad\} : 11, \{ae\} : 7, \{bd\} : 7, \{be\} : 8, \{cd\} : 6, \{ce\} : 5, \{de\} : 6$
$\{abe\} : 7, \{acd\} : 6, \{ace\} : 5, \{ade\} : 5, \{bde\} : 4, \{cde\} : 5$
$\{acde\} : 5$