UECS3213 / UECS3453 DATA MINING

SESSION: January 2019

TUTORIAL 6

Topic 6: Association Analysis

1. Consider the data set shown in Table 6.1.

Table 6.1. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

a) Compute the support for itemsets {e}, {b, d}, and {b, d, e} by treating each transaction ID as a market basket.

Answer:

$$s(\{e\}) = \frac{8}{10} = 0.8$$

$$s(\{b, d\}) = \frac{2}{10} = 0.2$$

$$s(\{b, d, e\}) = \frac{2}{10} = 0.2$$

b) Use the results in part (a) to compute the **confidence** for the association rules {b, d} → {e} and {e} → {b, d}. Is confidence a symmetric measure?

- c(bd --> e) = s({b, d, e})/s({b, d}) = 0.2 / 0.2 = 1.0 (100%)
- c(e --> bd) = s({b, d, e}) / s({e}) = 0.2 / 0.8 = 0.25 (25%)

$$c(bd \longrightarrow e) = \frac{0.2}{0.2} = 100\%$$
$$c(e \longrightarrow bd) = \frac{0.2}{0.8} = 25\%$$

No, confidence is not a symmetric measure.

c) Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

$$s(\{e\}) = \frac{4}{5} = 0.8$$
$$s(\{b,d\}) = \frac{5}{5} = 1$$
$$s(\{b,d,e\}) = \frac{4}{5} = 0.8$$

d) Use the results in part (c) to compute the **confidence** for the association rules {b, d} → {e} and {e} → {b, d}.

- c(bd --> e) = s({b, d, e})/s({b, d}) = 0.8 / 1.0 = 0.8 (80%)
- c(e --> bd) = s({b, d, e}) / s({e}) = 0.8 / 0.8 = 1.0 (100%)

$$c(bd \longrightarrow e) = \frac{0.8}{1} = 80\%$$
$$c(e \longrightarrow bd) = \frac{0.8}{0.8} = 100\%$$

2. Consider the market basket transactions shown in Table 6.2.

Table 6.2. Market basket transactions.

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

Total number of association rules:

$R = 3^d - 2^{d+1} + 1$

where d = number of items

https://people.revoledu.com/kardi/tutorial/MarketBasket/AssociationRules.htm

- There are six items in the data set.
- Therefore the total number of rules is $3^6 - 2^{(6+1)} + 1 = 602$.

b) What is the maximum size of the frequent itemsets that can be extracted (assuming *minsup* > *0*)?

- Because the longest transaction contains 4 items, the maximum size of frequent itemset is 4.

c) Write an expression for the maximum number of 3-itemsets that can be derived from this data set.

Answer: $\binom{6}{3} = 20$.

d) Find an itemset (of size 2 or larger) that has the largest support.

- {Bread, Butter}.
- s({Bread, Butter}) = 5/10 = 0.5

e) Find a pair of items, a and b, such that the rules {a} → {b} and {b} → {a} have the same confidence.

- (Beer, Cookies) or (Bread, Butter).
- c(beer --> cookie) = s({beer, cookie})/s({beer}) = 0.2 / 0.4 = 0.5 (50%)
- c(cookie --> beer) = s({cookie, beer}) / s({cookie}) = 0.2 / 0.4 = 0.5 (50%)
- c(bread --> butter) = s({bread, butter})/s({bread}) = 0.5 / 0.5 = 1.0 (100%)
- c(butter --> bread) = s({butter, bread}) / s({butter}) = 0.5 / 0.5 = 1.0 (100%)

3. The Apriori algorithm uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size k+1 are created by joining a pair of frequent itemsets of size k (this is known as the candidate generation step).

A candidate is discarded if any one of its subsets is found to be infrequent during the candidate pruning step. Suppose the Apriori algorithm is applied to the data set shown in Table 6.3 with minsup = 30%, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.

Table 6.3. Example of market basket transactions.

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

a) Draw an itemset lattice representing the data set given in Table 6.3. Label each node in the lattice with the following letter(s):

- N: If the itemset is not considered to be a candidate itemset by the Apriori algorithm. There are two reasons for an itemset not to be considered as a candidate itemset: (1) it is not generated at all during the candidate generation step, or (2) it is generated during the candidate generation step but is subsequently removed during the candidate pruning step because one of its subsets is found to be infrequent.
- F: If the candidate itemset is found to be frequent by the Apriori algorithm.
- I: If the candidate itemset is found to be infrequent after support counting (candidate from F_{k-1} but not frequent after support counting)

- Level-wise algorithm:

1. Let k = 1
2. Generate frequent itemsets of length 1
3. Repeat until no new frequent itemsets are identified
   1. Generate length (k+1) candidate itemsets from length k frequent itemsets
   2. Prune candidate itemsets containing subsets of length k that are infrequent
      - *How many k-itemsets contained in a (k+1)-itemset?*
   3. Count the support of each candidate by scanning the DB
   4. Eliminate candidates that are infrequent, leaving only those that are frequent

Note: steps 3.2 and 3.4 prune itemsets that are infrequent

1. Generate frequent itemsets of length 1

itemset	support

a	0.5
b	0.7
c	0.5
d	0.9
e	0.6

Table 1

2. Generate candidate 2-itemsets from frequent 1-itemsets. Scan the database to compute support.

itemset	**support**
ab	>= 0.3
ac	0.1
ad	>= 0.3
ae	>= 0.3
bc	>= 0.3
bd	>= 0.3
be	>= 0.3
cd	>= 0.3
ce	0.2
de	>= 0.3

Table 2

3. Eliminate infrequent 2-itemsets.

itemset	**support**
ab	>= 0.3
ad	>= 0.3
ae	>= 0.3
bc	>= 0.3
bd	>= 0.3
be	>= 0.3
cd	>= 0.3
de	>= 0.3

4. Generate candidate 3-itemsets from frequent 2-itemsets. Scan the database to compute support.

itemset	**support**
abc	0.1
abd	0.2
abe	0.2

ade	>= 0.3
bcd	0.2
bde	>= 0.3
cde	>= 0.3

Table 3

5. Eliminate infrequent 3-itemsets.

itemset	**support**
ade	>= 0.3
bde	>= 0.3
cde	>= 0.3

6. Generate candidate 4-itemsets from frequent 3-itemsets. Scan the database to compute support

itemset	**support**
abde	0.2
acde	0.1
bcde	0.1

7. Eliminate infrequent 4-itemset. No more itemset left. End.

The lattice structure is shown below.

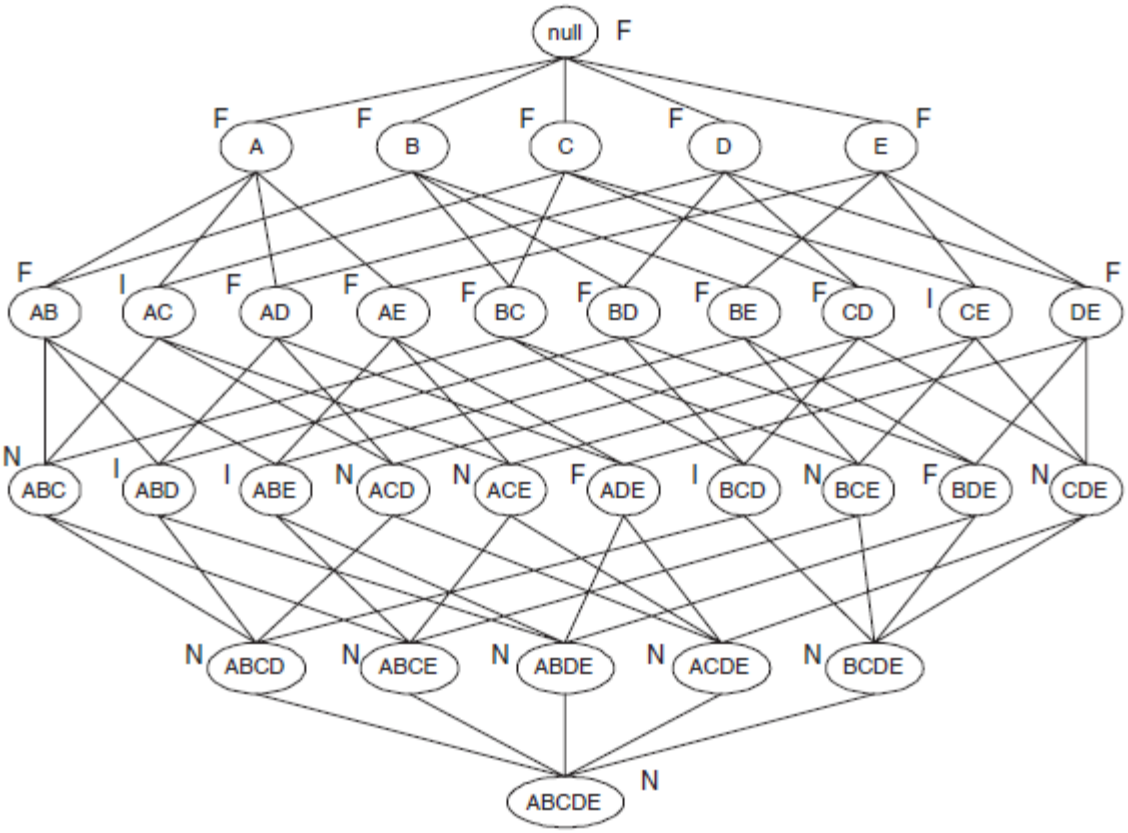

Figure 6.1. Solution.

b) What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?

- Percentage of frequent itemsets = 16/32 = 50.0% (including the null set).

c) What is the pruning ratio of the Apriori algorithm on this data set? (Pruning ratio is defined as the percentage of itemsets not considered to be a candidate because (1) they are not generated during candidate generation or (2) they are pruned during the candidate pruning step.)

- Pruning ratio is the ratio of N to the total number of itemsets. Since the count of N = 11, therefore pruning ratio is 11/32 = 34.4%.

d) What is the false alarm rate (i.e, percentage of candidate itemsets that are found to be infrequent after performing support counting)?

- False alarm rate is the ratio of I to the total number of itemsets. Since the count of I = 5, therefore the false alarm rate is 5/32 = 15.6%.

4. Consider the following set of frequent 3-itemsets:

{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5}.

Assume that there are only five items in the data set {1, 2, 3, 4, 5}.

a) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

List all candidate 4-itemsets obtained by the Fk−1 × F1 candidate generation method.

Supports for 1-itemsets:

Item	support
1	5
2	5
3	6
4	4
5	4

Candidate 4-itemsets (having lexicographical order pruning):

{1,2,3}: {1,2,3,4}, {1,2,3,5}

{1,2,4}: {1,2,4,5}

{1,2,5}: none, not possible to extend

{1,3,4}: {1,3,4,5}

{1,3,4} : no new ones

{2,3,4} : {2,3,4,5}

{2,3,5} : no new ones

{3,4,5}:none, not possible to extend

$F_1 = \{1, 2, 3, 4, 5\}$

$F_{(4-1)} = \{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 5\}, \{1, 3, 4\}, \{1, 3, 5\}, \{2, 3, 4\}, \{2, 3, 5\}, \{3, 4, 5\}$

b) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori.

5. Consider the same set of frequent 3-itemsets as above. List all candidate 4-itemsets obtained by the Fk−1 × Fk−1 candidate generation method.

When considering merging, only pairs that share first k-2 items are considered
(items in lexicographical order).

{1,2,3}x{1,2,4} : {1,2,3,4}

{1,2,3}x{1,2,5} : {1,2,3,5}

{1,2,4}x{1,2,5} : {1,2,4,5}

{1,3,4}x{1,3,5} : {1,3,4,5}

{2,3,4}x{2,3,5} : {2,3,4,5}

c) List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

Answer:
{1, 2, 3, 4}

The End