Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 2

Due: 01/25/2024(100 points)

Dharun Suryaa Nagarajan

Frequent Item sets

Assume that there are only five items in the data set. This question was taken from Tan et al., which may help in reviewing Candidate Generation.

1. List all candidate 4- item sets obtained by a candidate generation procedure using the merging strategy. $F_{k-1} * F_1$

   Solution: Using the above strategy we can get the following : {1,2,3,4},{1,2,4,5},{1,3,4,5},{2,3,4,5}

2. List all candidate 4-item sets obtained by the candidate generation procedure in Apriori using $F_{k-1} * F_{k-1}$

   Solution: Using the above strategy we can get the following : {1,2,3,4},{1,2,3,5},{1,2,4,5},{2,3,4,5}

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

   Solution: {1,2,3,4}

Association Rules

4. Consider the following table for question 4:

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

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

  Solution: $3^n - 2^{n+1} + 1 = 3^7 + 2^8 + 1 = 1932$

- What is the confidence of the rule {Milk, Diapers} → {Butter}?

  Solution: $\frac{\sigma(Milk, Diapers, Butter)}{\sigma(Milk, Diaper)} = \frac{2}{4} = 0.5$

- What is the support for the rule {Milk, Diapers} → {Butter}?

  Solution: $\frac{\sigma(Milk, Diapers, Butter)}{N} = \frac{2}{10} = 0.2$

5. True or False with an explanation: Given that a,b,c,d is a frequent itemset, {a,b} is always a frequent itemset.

   Solution: True, this clearly states the property of the support known as support based pruning where sub sets will be a frequent sets if the parent set is one.

6. True or False with an explanation: Given that {a,b}, {b,c} and {a,c} are frequent itemsets, {a,b,c} is always frequent.

   Solution: False, although {a,b,c} is having a good chance of being a frequent sets, the support check at the end may cause it to be not part of the frequent set so {a,b,b} is not always frequent given the susbsets are frequent.

7. True or False with an explanation: Given that the support of {a,b} is 20 and the support of {b,c} is 30, the support of {b} is larger than 20 but smaller than 30.

   Solution: False, support of {a,b} is 20, support of {b,c} is 30 so support(b) $>$ 20 and support(b) $<$ 30 cannot be true because $f(b) \geq f(a, b)$ so $f(b) \geq 30$, support of b is greater than both 20 and 30

8. True or False with an explanation: In a dataset that has 5 items, the maximum number of size-2 frequent itemsets that can be extracted (assuming minsup $>$ 0) is 20.

   Solution: False, The maximum number of size-2 frequent itemsets that can be extracted from a dataset with 5 items is 10, not 20. This is because for a size-2 itemset, each item can form pairs with the other items in the dataset, resulting in combinations of 5 choose 2, which is equal to 10.

9. Draw the itemset lattice for the set of unique items I = {a, b, c}. **Solution**:

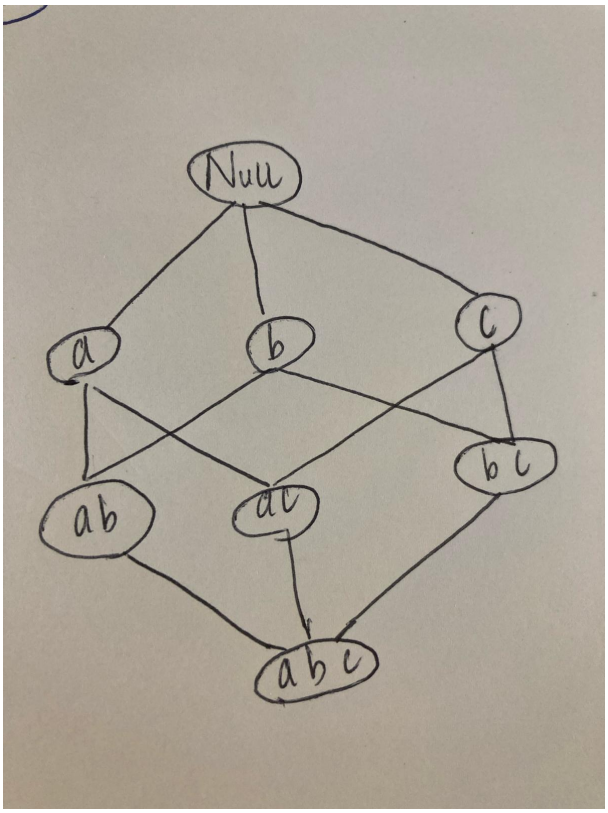

Figure 0.1: Answer