Homework 6

1 Exercises

1.1.15 Answer the following questions using the data sets shown in Figure 5.34 . Note that each data set contains 1000 items and 10,000 transactions. Dark cells indicate the presence of items and white cells indicate the absence of items. We will apply the Apriori algorithm to extract frequent itemsets with minsup = 10% (i.e., itemsets must be contained in at least 1000 transactions).

a. Which data set(s) will produce the most number of frequent itemsets?

Dataset e, because it has to generate the longest frequent itemset along with its subsets.

b. Which data set(s) will produce the fewest number of frequent itemsets?

Dataset d, it does not produce any frequent itemsets at 10% support threshold.

c. Which data set(s) will produce the longest frequent itemset?

Dataset e.

d. Which data set(s) will produce frequent itemsets with highest maximum support?

Dataset b.

e. Which data set(s) will produce frequent itemsets containing items with wide-varying support levels (i.e., items with mixed support, ranging from less than 20% to more than 70%)?

Dataset e.

1.2.1.a Given the database in Table 8.2. Using minsup = 3/8, show how the Apriori algorithm enumerates all frequent patterns from this dataset.

Level one frequent items:

A = 5, B = 4, C = 5, D = 6, F = 4, G = 5

The frequent ones are: A, B, C, D, F, G

Level two frequent items:

AB = 3, AC = 3, AD = 4, AF = 2, AG = 2

BC = 2, BD = 2, BF = 0, BG = 2

CD = 4, CF = 3, CG = 3

DF = 4, DG = 3

FG = 3

The frequent ones are: AB, AC, AD, CD, CF, CG, DF, DG and FG.

Level three frequent items:

ABC = 1, ABD = 2, ACD = 3, CDF = 3, CDG = 2, CFG = 2, DFG = 3

The frequent ones are: ACD, CDF, and DFG.

Level four frequent items:

CDGF = 1

No more frequent itemsets.

1.2.4 Given the database in Table 8.4. Show all rules that one can generate from the set ABE.

All non-empty subsets of ABE:
A = 4, B = 5, E = 4, AB = 3, AE = 2, BE = 4, ABE = 2
Support of ABE = 2/6 = 1/3
Confidence of each rule:
A → BE, support of A = 4/6, confidence = 2/4 = 0.5
B → AE, support of B = 5/6, confidence = 2/5 = 0.4
E → AB, support of E = 4/6, confidence = 2/4 = 0.5
AB → E, support of AB = 3/6, confidence = 2/3 = 0.67
AE → B, support of AE = 2/6, confidence = 2/2 = 1.0
BE → A, support of BE = 4/6, confidence = 2/4 = 0.5

2 Practicum problems

Q2(b)(vi) Looking at the number of itemsets retrieved between b-(iv) and b-(v), why do you think there is such a discrepancy in the number of itemsets retrieved?

The significant discrepancy between the support thresholds of 0.01 which retrieved 124 itemsets and 0.10 which retrieved 2 itemsets can be attributed to the fact that lower support values allow for a wider range of itemsets to be considered frequent, including those that occur less often in the dataset. In contrast, a higher support threshold requires itemsets to be much more frequent, drastically reducing the number of itemsets that meet the criteria.

Q2(c)(ii) Comment on how good you think the rules are as a function of confidence and lift. If the rules are good, state so and provide a 1-sentence answer why you consider the rules to be good. If you think the rules are not good, state so and provide a 1-sentence answer why you consider the rules to be bad.

The rules generated with a confidence of 0.8 and support of 0.001 are likely not very good. While the high confidence indicates strong rule validity, the low support suggests the rules are based on rare patterns that may not generalize well and could be overfitting the data.