Homework 5 - CS 422 Data Mining

Julen Ferro

October 31, 2022

1 Introduction

This will be the report for Assignment number 5.

2 Chapter 5 - Exercises

2.1 Exercise 15

Answer the following questions using the data sets shown in Figure 4.34. 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 percent (i.e., itemsets must be contained in at least 1000 transactions).

(a) Which data set(s) will produce the most number of frequent itemsets?
Data set (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?
Data set (d) which does not produce any frequent itemsets at 10

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

(d) Which data set(s) will produce frequent itemsets with highest maximum support?
Data set (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 percent to more than 70 percent)?
Data set (e). Because it contains items with a lot of support and also items that appear in a smaller number of transactions.

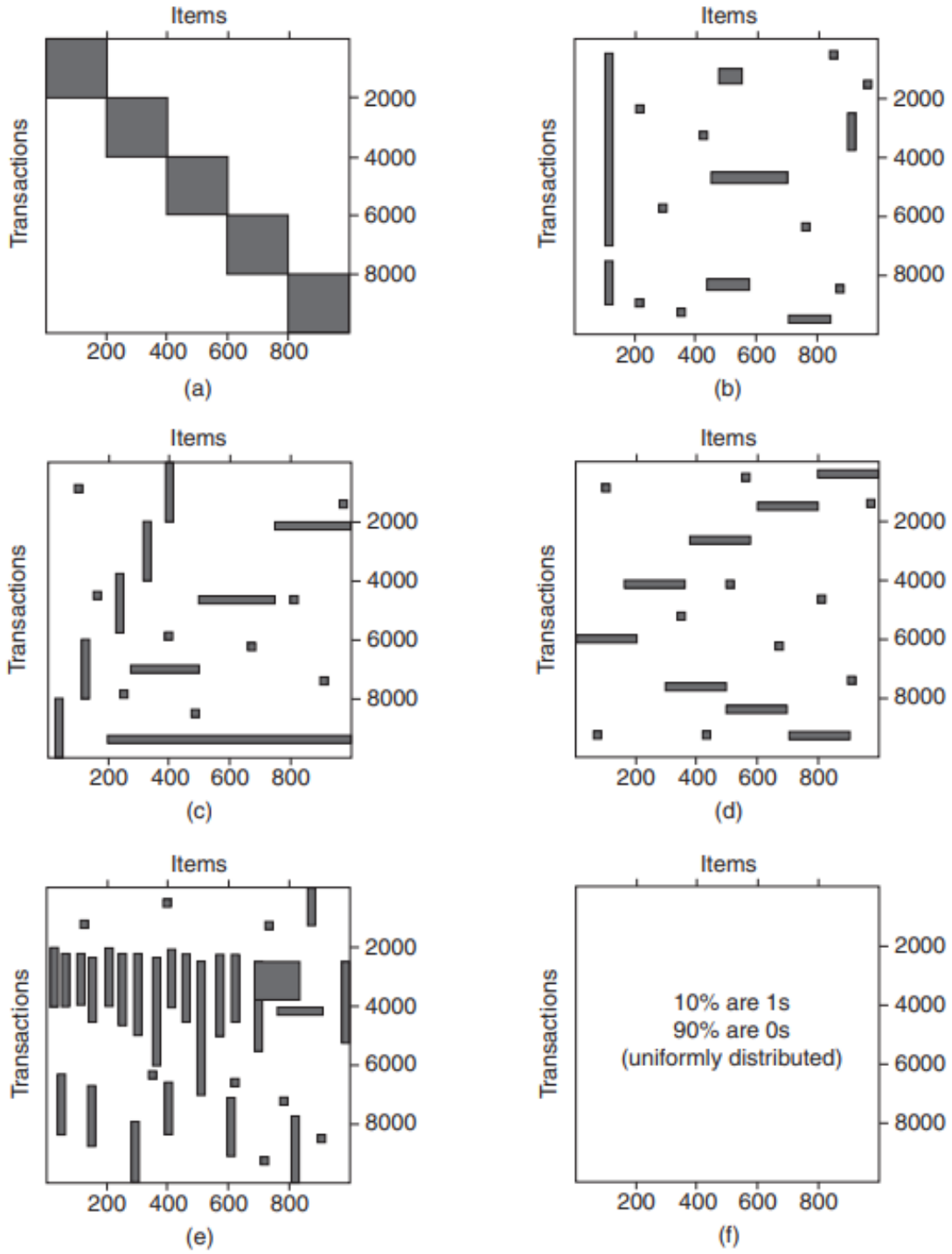

Figure 4.34. Figures for Exercise 18.

Figure for the exercise 18

3 Chapter 8 - Exercises

3.1 Exercise 1.a

Q1. Given the database in Table 8.2. (a) Using minsup = 3/8, show how the Apriori algorithm enumerates all frequent patterns from this dataset

Table 8.2. Transaction database for Q1

tid	itemset
t_1	$ABCD$
t_2	$ACDF$
t_3	$ACDEG$
t_4	$ABDF$
t_5	BCG
t_6	DFG
t_7	ABG
t_8	$CDFG$

Figure 1: Figure for the exercise 1 a

First, we will count how many times does each letter appear.
A = 5 B = 4 C = 5 D = 6 E = 1 F = 4 G = 5
AB = 3 AC = 3 AD = 4 AE = 1 AF = 2 AG = 2
BC = 2 BD = 2 BE = 0 BF = 1 BG = 2
CD = 4 CE = 1 CF = 2 CG = 3
DE = 1 DF = 4 DG = 3
EF = 0 EG = 1
FG = 2
The frequent patterns of this dataset are as follows:
Null (8), A (5), B(4), C(5), D(6), F(4), G(5), AB(3), AC(3), AD(4), CD(4), CG(3), DF(4), DG(3), ACD(3)

3.2 Exercise 4

Q4. Given the database in Table 8.4. Show all rules that one can generate from the set ABE.

For the itemset AB: AB support 3 Confidence: A – B 0,75; B – A 0,6
For the itemset AE: AE support 2 Confidence: A – E 0,5; E – A 0,5
For the itemset BE: BE support 4 Confidence: B – E 0,8; E – B 0,8
For the itemset ABE: ABE support 2 Confidence: AB – E 0,66; AE – B 1; A – BE 0,5; B – AE 0,4; E – AB 0,5

Table 8.4. Dataset for Q4

tid	itemset
t_1	ACD
t_2	BCE
t_3	$ABCE$
t_4	BDE
t_5	$ABCE$
t_6	$ABCD$

Figure 2: Figure for the exercise 4