TD 1 – Frequent Pattern Mining

IMT Atlantique – FIL A3

UE Machine Learning

Exercise 1 – Appriori and Eclat algorithms

1. Given the database in Table 1a.
   i. What is the maximal number of frequent itemsets that can be extracted from this dataset ?
   ii. Draw the lattice of itemsets.
2. Using $\theta = 3/8$, show how the Appriori algorithm enumerates all frequent patterns from this dataset.

t	$\boldsymbol{i}(t)$
1	ADE
2	BCD
3	ACE
4	$ACDE$
5	AE
6	ACD
7	BC
8	$ACDE$
9	BCE
10	ADE

(a) Transaction database for question 1.

A	B	C	D	E
1	2	1	1	2
3	3	2	6	3
5	4	3		4
6	5	5		5
	6	6		

(b) Transaction database for question 2.

3. Consider the vertical database shown in Table 1b. Assuming that $\theta = 3$, enumerate all the frequent itemsets using the Eclat method.

Exercise 2 – Difference of Tidsets

One of the main bottleneck of the Eclat algorithm is the intersection operation performed to compute the support of itemsets. This algorithm can be improved by exploiting the concept of **Diffsets** or Difference of Tidsets. The variant of Eclat that uses the diffset optimization is called dEclat.

Consider two $k-$itemsets $X_a = \{x_1, \ldots, x_{k-1}, x_a\}$ and $X_b = \{x_1, \ldots, x_{k-1}, x_b\}$ that share the common $(k-1)-$itemset $X = \{x_1, \ldots, x_{k-1}\}$ as a prefix. The *diffset* of X_a (resp. X_b) is the set of tids that contain the prefix X, but not the item X_a (resp. X_a) :

$$\boldsymbol{d}(X_a) = \boldsymbol{t}(X) \setminus \boldsymbol{t}(X_a) \qquad \boldsymbol{d}(X_b) = \boldsymbol{t}(X) \setminus \boldsymbol{t}(X_b)$$

Now, consider the diffset of $X_{ab} = X_a \cup X_b = \{x_1, \ldots, x_{k-1}, x_a, x_b\}$, we have

$$\boldsymbol{d}(X_{ab}) = \boldsymbol{t}(X_a) \setminus \boldsymbol{t}(X_{ab}) = \boldsymbol{t}(X_a) \setminus \boldsymbol{t}(X_b) = \boldsymbol{d}(X_b \setminus \boldsymbol{d}(X_a)$$

Thus, the diffset of X_{ab} can be obtained from diffsets of its subsets X_a and X_b, which means that we can replace all intersection operations in Eclat with diffset operations. Using diffsets the support of a candidate itemset can be obtained by subtracting the diffset size from the support of the prefix itemset.

1. Define the frequency of X_a and X_b by mean of $\boldsymbol{d}(X_a)$ and $\boldsymbol{d}(X_b)$.
2. Using the diffset of X_{ab}, define the frequency of X_{ab}.
3. Show how the dEclat algorithm enumerates all frequent patterns from the dataset of Table 1b.

Exercise 3 – Association rules

Given the database in Table 1. Assume that $minconf = 0.9$. Show all rules that one can generate from the frequent itemset $ABDE$.

TABLE 1 – Dataset for Exercises 3 and 4.

t	$\boldsymbol{i}(t)$
1	$ABDE$
2	BCE
3	$ABDE$
4	$ABCE$
5	$ABCDE$
6	BCD

Exercise 4 – Closed and maximal itemsets

Given the database in Table 1.

1. Show the application of the closure operation on AD, that is, computes $\boldsymbol{c}(AD)$. Is AD closed ?
2. Find all frequent, closed and maximal itemsets using $\theta = 3/6$.

Exercise 5 – Closed itemset lattice

Consider the frequent closed itemset lattice shown if Figure 1. Assume that item space is $\mathcal{I} = \{A, B, C, D, E\}$. Answer the following questions :

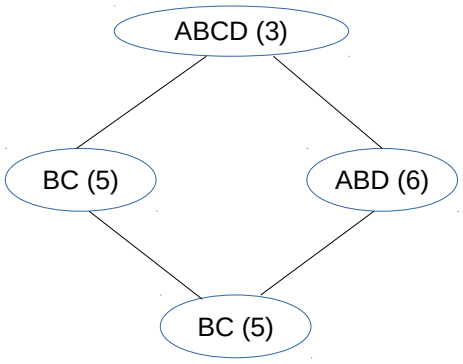

FIGURE 1 – Closed itemset lattice.

1. What is the frequency of CD ?
2. Find all frequent itemsets and their frequency, for itemsets in the subset interval [B, ABD].
3. Is ADE frequent ? If yes, show its frequency. If not, why ?