Data Mining Assignment 2

Zihao Xu zihxu@kth.se
Minchong Li mincli@kth.se

Date: November 17, 2023

1 Finding Frequent Itemsets

To find frequent itemssets from the provided sale transaction dataset, we implement a class A-Priori algorithm.

2 Generating Association Rules

Take the frequent itemsets found by A-Priori algorithm as input, we generally follow the method showing on the slides to generate association rules. We do the following things for all the frequent itemsets: we generate all the possible subsets of a single frequent itemset, then, we do single pass to compute the rule confidence. We also use the trick mentioned on the slides: If $A, B, C \rightarrow D$ is below confidence, so is $A, B \rightarrow C, D$, to avoid some unnecessary operations.

3 How to Build and Run

Firstly, download the dataset and unzip it into the assignment2 root.

Secondly, install the requirements:

```
pip install -r requirements.txt
```

Then, use the following command to run the program:

```
cd ./src/
python main.py ../dataset/T10I4D100K.dat -s 0.005 -c 0.5
```

Arguments explanation:

- required positional argument: path of dataset.

- `-s / --support`: Threshold of support
- `-c / --confidence`: Threshold of confidence

4 Results

4.1 Relationship between Time Cost vs. Threshold of Support

In this part, we did the experiment by exploring the relationship between time cost and threshold of support.

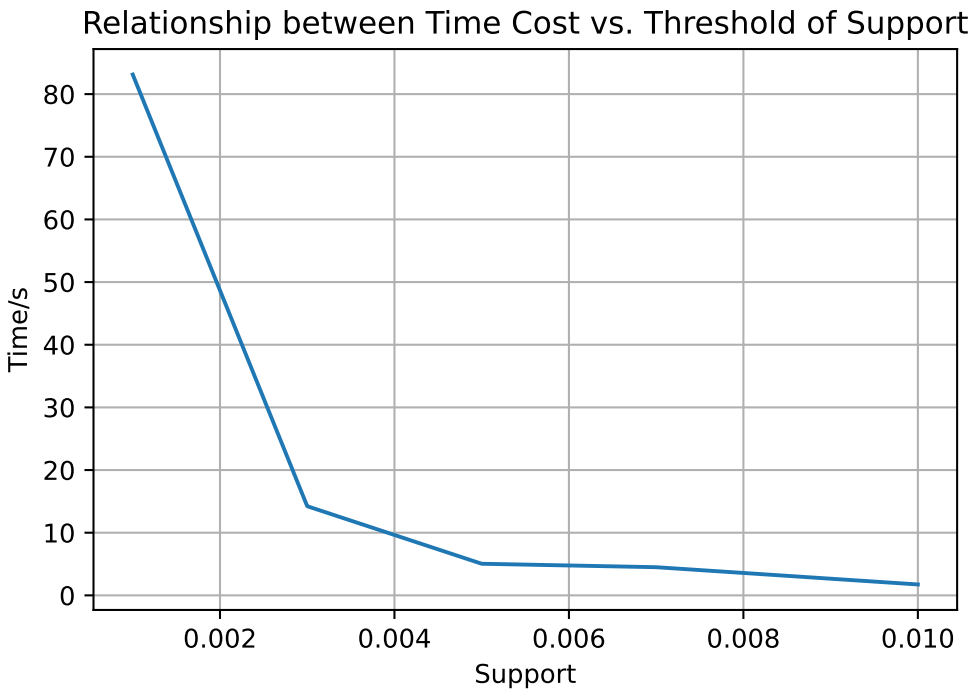

Figure 1: Relationship Between Time Cost vs. Threshold of Support

4.2 Relationship between Number of Association Rules vs. Threshold of Confidence.

In this part, we did the experiment by exploring the relationship between number of association rules and threshold of confidence.

The figure below shows part of our mined result of association rules.

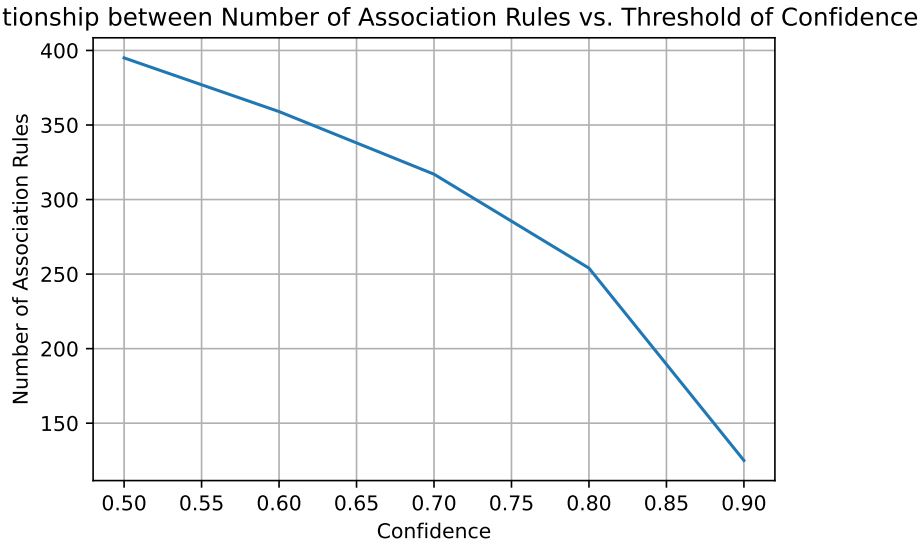

Figure 2: Relationship between Number of Association Rules vs. Threshold of Confidence.

```
rule 96: (336, 509) ===> (71,), score: 0.947
rule 97: (583, 617) ===> (354,), score: 0.904
rule 98: (112, 511) ===> (499,), score: 0.941
rule 99: (499, 511) ===> (112,), score: 0.956
rule 100: (392, 632) ===> (489,), score: 0.967
rule 101: (489, 632) ===> (392,), score: 0.967
rule 102: (580, 988) ===> (871,), score: 0.910
rule 103: (871, 988) ===> (832,), score: 0.909
rule 104: (385, 721) ===> (885,), score: 0.984
rule 105: (385, 885) ===> (721,), score: 0.916
rule 106: (192, 487) ===> (888,), score: 0.959
rule 107: (192, 638) ===> (487,), score: 0.936
rule 108: (487, 638) ===> (192,), score: 0.946
rule 109: (192, 888) ===> (487,), score: 0.929
rule 110: (487, 935) ===> (192,), score: 0.963
rule 118: (352, 978) ===> (708,), score: 0.959
```

Figure 3: Sample Result of Association Rules.