2017 Fall Data Mining

Homework 1

Mining association rules

Due Date: 23:59, December 1st Friday, 2017

TA: 池昊頤 anarchih.tw@gmail.com

In this homework, the major task is to find the association rule, the most basic concept is “co-occurrence”, try to apply it with your imagination

Dataset:

- https://sites.google.com/site/cclljj/dataset-airbox (same as hw0)
- We will use “March 2017, Taiwan” dataset. (1 month)
- You can use the dataset before/after preprocessing.

Tasks:

- What is a transaction?
  Suppose I define a transaction is (pm2.5, humidity, temperature) for device_id=74DA3895C538, then the transaction probably are (20, 64, 35), (42, 76, 28) …

- What rules should be discover?
  But we only have limited data, co-occurrence may not happened in the real value, so we need to do discretization, that is, convert the real value to some symbols, maybe (20, 60, 30), (40, 70, 20) … (divided by 10)
  How many symbol should you use, depends on what’s your application.

  What you need to do is try to define two more your own transactions, and conduct some experiments on the three mining tasks. (including my example)

  - my example (change device and replace humidity or temperature with pm10 is allowed)
  - your1
  - your2

  For each task, you should try at least two discretization methods (ex: divided by 10, divided by 20 …) and two algorithms (apriori, fp-growth …) to find association rules.

Report:

- In this homework, you have three tasks to do. For each task, you need to clarify
    1. what transaction you define
    2. what discretization methods you use
    3. what algorithms you use
    4. what rules you discover
    5. what you have learned, and do some comparisons between different methods you use.

- Remember to mention what dataset you use in this homework. (dataset before/after preprocessing)

- Your code should be submitted with the report.

- Algorithm codes from github are allowed, but remember to cite the repo.

- Please hand in your report with pdf, odt, or html format (ipython notebook is allowed).

- If you hand in your report with doc(x) or other formats, you may lose some points.

- If you have any questions or suggestions, feel free to contact me:)