RMIT University
School of Science
COSC2110/COSC2111 Data Mining
Laboratory Week 5

Aims of this lab

- Learn how run the apriori algorithm and interpret the results.
- Learn about the importance of data representation for association finding.

---

1. You will need to have access to the WEKA package.
2. The data files for this lab can be found at `/KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data`
3. Load the file `arff/UCI/weather.nominal.arff`.
   (a) Inspect the file with an editor.
   (b) Run Apriori using the default settings of the options, but turn on the printing of item sets.
   (c) What is the confidence for rule 10. How was this confidence value computed?
   (d) How many instances form the support of rule 8?
   (e) Identify the itemset from which rule 1 was generated. What other rules could be generated from this itemset?
   (f) What does 'best rules' mean? What criteria are used to determine the best rules?
   (g) How many rules can be generated from this data? Experiment with the numRules parameter.
4. Load the file `arff/supermarket1-subset.arff`.
   (a) View the file with an editor.
   (b) Go to the Associate screen and run Apriori with the default values.
   (c) Are there any rules with high accuracy?
   (d) Are there any golden nuggets in the output?
   (e) Go to the Associate screen and run Apriori with the default values.
   (f) Are there any rules with high accuracy?
   (g) Are there any golden nuggets in the output?

(h) Experiment with different metrics. How do the generated rules change with the different metrics.

5. Load the file `arff/supermarket2-small.arff`.

   (a) Repeat (4) above on this data.

   (b) What do you conclude about the different representations?

6. Repeat (4) and (5) with FilteredAssociator and FPGrowth and compare the results with apriori.

7. The files `supermarket1-subset.arff` and `supermarket2-small.arff` are trimmed down versions of `supermarket1.arff` and `supermarket2.arff` in order to get quick execution times for apriori.

   (a) Run apriori on the larger files?

   (b) What problems arise and what can you do about them?

8. [Advanced] In apriori it is possible to designate an attribute as a class and generate only rules with class on the right hand side (CAR parameter). Load the file `arff/UCI/mushroom.arff` and run the apriori algorithm with `CAR=True` and `classIndex=23` and OneR and consider the rules as classifiers. Which do you think are more understandable?