Practical II: Association Rules

Prof. M-Tahar Kechadi

School of Computer Science
University College Dublin

The aim of this practical is to use RapidMiner to generate association rules from data sets, using the algorithms discussed in the lectures. The data sets to be used can be found on Blackboard at the same location.

All files generated (5 in total) should be placed in a zipfile with the name **<student name>_<student number>_comp40370_practical03.zip**, and submitted via Blackboard.

General Hints: If you are unsure of a particular operator name in RapidMiner, you can enter search terms in the *[Filter]* input field in the *Operators* tab, this will then display a filtered list of operators containing those terms. The *Import Excel Sheet* wizard in the *Repositories* tab can be used to import spreadsheets into data sets, similar to what was done in the previous practical.

For this practical, the association rule operators can be found under *Association and Item Set Mining*. For details on the operator properties, see the associated help documentation.

Question1 Creating association rules with Apriori (1)

Using the gpa.xls data set, generate a process which does the following:

1. Filter out the *count* attribute as this won't be included in the rule generation (i.e. select the other attributes).
2. Add the W-Apriori operator, configured to use the default confidence value (0.9) and output the frequent itemsets generated. A value for support is not required at this time.
3. Run the process and copy the generated rules (if any) and frequent itemsets to your results document.
4. Change the confidence value to 0.7 and rerun the process. Copy the generated rules and frequent itemsets to your results document.

   The results document should be submitted, along with the RapidMiner process xml file (*File - Export Process* menu option), as explained at the start of this document.

Question2 Creating association rules with FP-Growth (1)

Using the transactions.xls data set, generate a process which does the following:

1. Filter out the *CAR* attribute as this won't be included in the rule generation (i.e. select the other attributes)
2. Add the FP-Growth operator, with a support value of 0.2, and uncheck *find min number of itemsets*.
3. Add the Create Association Rules operator, using the default min confidence value.
4. Run the process and copy the generated rules (if any) to your results document from the Text View.
5. Change the confidence value to 0.6 and rerun the process. Copy the generated rules to your results document.

   The results document should be submitted, along with the RapidMiner process xml file (*File - Export Process* menu option), as explained at the start of this document.

Question3 Creating association rules with Apriori (2)

The acorns.xls data set contains some data which will be used to find rules between various attributes of trees growing in the US. Generate a process which does the following:

1. Filter out the *Range* and *Species* attributes as these won't be included in the rule generation.
2. The Apriori algorithm will only work with nominal (discrete) attributes. Discretize the numeric attributes into 3 bins using a simple sorting method to store the values into the bins.
3. Add the W-Apriori operator, configured to use the default confidence value (0.9) and output the frequent itemsets generated. A value for support is not required at this time.
4. Run the process and copy the generated rules (if any) and frequent itemsets to your results document.
5. Change the confidence to a value which will generate rules, and rerun the process. Copy the generated rules and frequent itemsets to your results document.

   The results document should be submitted, along with the RapidMiner process xml file (*File - Export Process* menu option), as explained at the start of this document.

Question4 Creating association rules with FP-Growth (2)

The bank-data.xls spreadsheet contains customer records from the marketing department of a financial firm. The data contains the following fields:

id	a unique identification number
age	age of customer in years (numeric)
sex	MALE / FEMALE
region	inner_city/rural/suburban/town
income	income of customer (numeric)
married	is the customer married (YES/NO)
children	number of children (numeric)
car	does the customer own a car (YES/NO)
save_acct	does the customer have a saving account (YES/NO)
current_acct	does the customer have a current account (YES/NO)
mortgage	does the customer have a mortgage (YES/NO)
pep	did the customer buy a PEP (Personal Equity Plan) after the last mailing (YES/NO)

1. Filter out the *id* attribute as this won't be included in the rule generation.
2. Like Apriori, the FP-Growth algorithm requires nominal (discrete) attributes. Discretize the numeric attributes into 3 bins using a simple sorting method to store the values into the bins.
3. In addition, the FP-Growth algorithm requires discrete attributes to have only two values (binominal). Convert the nominal attributes into binominal attributes.
4. Add the FP-Growth operator, with a support value of 0.2, and uncheck *find min number of itemsets*.
5. Add the Create Association Rules operator.
6. Experiment with different confidence values, selecting a value which will produce more than 10 rules, copying them to your results document from the Text View. Select the top 2 most "interesting" rules and for each specify the following:

   - an explanation of the pattern and why you believe it is interesting based on the business objectives of the company;
   - any recommendations based on the discovered rule that might help the company to better understand behavior of its customers or in its marketing campaign.

   (Note: The top 2 most interesting rules may not be the top 2 rules in the result set. They are rules that provide some non-trivial, actionable knowledge based on the underlying business objectives.)

   The results document (with rules and discussion about the 2 most interesting rules) should be submitted, along with the RapidMiner process xml file (*File - Export Process* menu option), as explained at the start of this document.