Practical 1: Data Pre-processing

Prof. M-Tahar Kechadi

School of Computer Science
University College Dublin.

The aim of this practical is to use a data mining tool to perform some data set preprocessing, using some of the concepts discussed in the lectures so far. RapidMiner or a similar tool should be used to complete this practical. The data sets to be used can be found on Moodle.

All files generated by the following questions should be placed in a zipfile with the name **<student number>_COMP40370_1.zip**, and submitted via Balckboard.

General Hint: 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.

Question1 Data Cleaning

Using the AutoMpg_question1.xls data set, generate a process which does the following

(Hint: see the operators in *Data Transformation - Data Cleansing* in RapidMiner):

1. Replace the missing *horsepower* values with the **average** value of this attribute
2. Replace the missing *origin* values with the **minimum** value of this attribute
3. Export the modified data to a new AutoMpg_question1_answer.xls spreadsheet

   The generated spreadsheet should be submitted, along with the RapidMiner process xml file if this tool has been used (with the *File - Export Process* menu option).

Question2 Data Integration

Using the AutoMpg_question2_a.xls and AutoMpg_question2_b data sets, generate a process which does the following:

(Hint: see the *Generate Attribute*, *Append* and type conversion operators in RapidMiner)

1. The first data set has a *car name* attribute, whereas the second data set has a *name* attribute. *name* should be renamed to *car name*.
2. The second data set has an additional *other* attribute which is not present in the first data set. Update the first data set to add an additional *other* attribute, with a default value of **1**. This attribute should also have a type of *Integer*.

3. Append the modified second data set to the modified first data set.

4. Export the new data to a new AutoMpg_question2_answer.xls spreadsheet

   The generated spreadsheet should be submitted, along with the RapidMiner process xml file if this tool has been used (with the *File - Export Process* menu option).

Question3 Data Transformation

(Hint: see the *Normalize* operator, and those under *Generation*)

Using the SensorData_question3ẋls data set, generate a process which does the following:

1. Generate a new attribute, *Original Input3*, which is a copy of the *Input3* attribute

2. Generate a new attribute, *Original Input12*, which is a copy of the *Input12* attribute

3. Normalize the *Input3* attribute, using the z-score transformation method

4. Normalize the *Input12* attribute to the range: [0.0, 1.0]

5. Generate a new attribute, *Average Input*, which is the average of all of the attributes, from the *Input1* ... *Input12* attributes. This average should include the normalized attribute values but not the copies that were made of these.

6. Export the new data to a new SensorData_question3_answer.xls spreadsheet

   The generated spreadsheet should be submitted, along with the RapidMiner process xml file if this tool has been used (with the *File - Export Process* menu option).

Question4 Data Reduction and Discretisation

(Hint: see the *Type Conversion* and *Attribute Set Reduction* operators)

Using the DNAData_question4.xls data set (which contains only numeric attributes), generate a process which does the following:

1. Reduce the number of attributes using Principal Component Analysis (PCA), keeping a variance of 0.95.

2. Discretise the PCA-generated attribute subset into 10 bins, using a simple sorting method to store the numerical values into the bins. Export this data to a new DNAData_question4_simplebins.xls spreadsheet.

3. Discretise the PCA-generated attribute subset into 10 bins, using the equal frequency binning method. Export this data to a new DNAData_question4_equalfrequency.xls spreadsheet.

   The generated spreadsheets should be submitted, along with the RapidMiner process xml file if this tool has been used (with the *File - Export Process* menu option).