Lab-Week8: Data Pre-Processing Using Weka

Learning Objectives:

The exercises in this lab, through manipulating a dataset, will help you to

- ✓ get started with the data mining software *Weka*;
- ✓ how to save `.csv` file into the default `.arff` format;
- ✓ get to know how to use *Weka* to discretise values for certain attributes;
- ✓ practise missing data handling in *Weka*.

Data Cleaning Reference and Dataset

The Weka data cleaning tutorial from Depaul University provides a very practical introduction to data cleaning using *Weka*.

- You can download and install *Weka* on your own machine.
  - https://waikato.github.io/weka-wiki/downloading_weka/
- The following two files are available through LMS under the same lab.
  - PDF version of DePaul University tutorial (weka-data-preprocessing.pdf)
  - The bank-data.csv dataset

Exploratory study of the data

Download and load the bank-data.csv into *Weka*, find out

- the attributes that are of numerical type;
- the attributes that are of nominal/categorical type; and
- compare the data summary view with these two different data types.

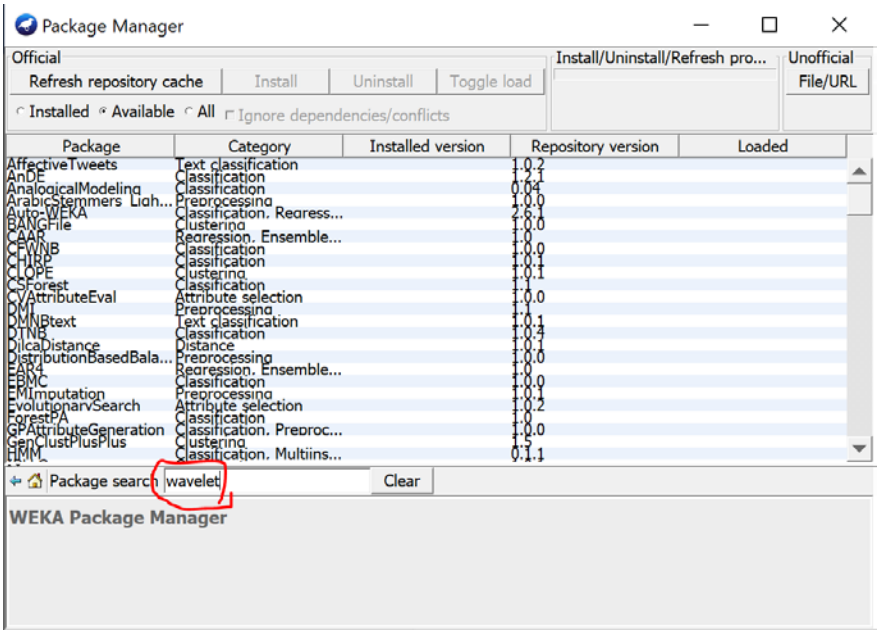
You will find reading and following the explanations in the tutorial section on "Loading the Data" helpful.

Attribute Selection

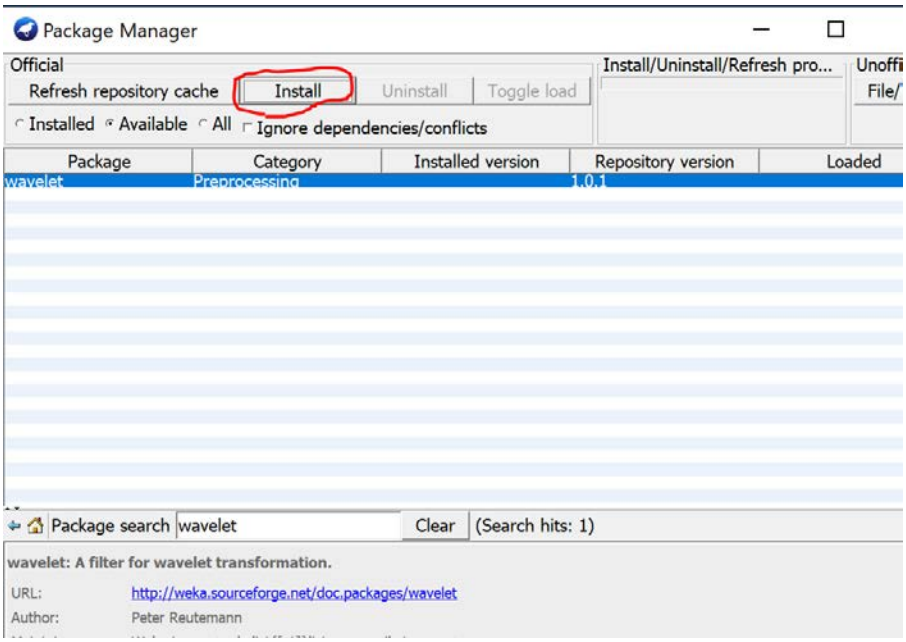
Unique record IDs useful for transactional data queries, are not useful in data mining and data warehousing. So very often the first task to do is to remove the `ID` column.

- Remove the `id` column using *Weka*; and
- Save the newly obtained dataset into `.arff` format.
- Find the Principal Component and Discrete Wavelet Transform filters and observe how the attributes are transformed.

To install the Wavelet Package, you will need to select from the main **Weka GUI Chooser** -> **Tools** -> **Package Manager**, and type "*wavelet*" in the search box, then hit Enter or Return.



Select the package and click Install.



Then you will be able to find the Wavelet transform from the unsupervised attribute filter list.

You will find reading and following the explanations in the tutorial section on "Selecting and Filtering Attributes" helpful.

Discretisation or Binning to transform numerical attributes

Convert `age`, `income` and `children` into discrete ranges, by following instructions from the tutorial section on "Discretization" (cf. weka-data-preprocessing.pdf). Note you will need to use both a text editor (e.g. PyCharm, *Notepad*++) and *Weka* to complete this task.

Dealing with Missing Data

Follow the tutorial section on "Missing Data" to introduce some missing data into some attributes, apply `ReplaceMissingValue` filter, and observe what strategy that this filter used to replace missing data.