08338 ACW Labs Week Sheet 1

All the remaining 08338 labs will focus on the ACW data mining step by step. The advice in these sheets will help you submit the ACW on time for each of the stages. Note these Lab Advice Sheets cover the data exploration and mining processes and not the submission of the ACW – read the ACW specifications for that.

Overview

You should tackle the ACW in the same way as described for the weka Tutorial. To refresh, you will be using the given data file (**DMDS-2015-16-ACW-Data.xlsx**) as a DataWarehouse. You should have ONE such FILE (keep backups). From this you will create csv files for weka input. The DataWarehouse will ultimately contain multiple appropriately named worksheets covering your data work and will be submitted as part of the ACW. As given the DataWarehouse contains the following EIGHT worksheets:

Methodology—first worksheet noting descriptions of all other worksheets and their derivation

GivenDataDescription—second worksheet noting descriptions of the given data

DataCleaning —third worksheet noting what data needs repairing or cleaning and how

FinalDataDescription—fourth worksheet noting descriptions of the final data

Performance—fifth worksheet noting descriptions of classifier performance

DSS - sixth worksheet with the data test set and the decision rules applied

BaseData-All– original given data containing 1025 data records plus header row

Risk-Unknown – header row only but will contain original test data set derived from BaseData-All

For the ACW you will complete these and generate many more sheets. The ACW requires multiple cleaning and transformation steps. In terms of working, if you want a current worksheet to be used in weka,

- you save it as an appropriately named CSV file
- Immediately resave the Excel file as your xlsx file

Warning: No matter what you do when you save a worksheet as a CSV file, Excel will rename the working DataWarehouse and be in CVS mode – **ALWAYS** revert back to 08338-DataWarehouse in Excel mode before you edit or close it. This way you maintain just the one Excel file plus the CSV files which should ONLY BE LOADED into weka (and otherwise left alone). You should build the data description and place all your weka outputs also in the Excel File.

ACW Labs: Step by Step Data Description : Exploration and Cleaning

Here is the guide to what will have been achieved. Your steps may vary slightly but the principles and process are as follows:

1. Download ACW Zip File, Understand what is required of ACW. Initially focus on Stage1 Data Description and Stage2 Data Cleaning. The advice in this sheet will help you complete these two parts of the ACW. Zip File at: http://intra.net.dcs.hull.ac.uk/student/modules/08338/ACW%20Material/Forms/AllItems.aspx
2. Save Excel File as **08338-DataWarehouse**
3. Make a version of the raw data that loads into weka
    a. Save **BaseData-All** as BaseData-All.csv (remember to revert to Excel mode!)
    b. Start weka
    c. Load **BaseData-All.csv** – it should load. Explore all attributes using visualisation. If it fails to load you need to clean below and right of the “visible” data

   d. Using the visualisation tools in weka explore each attribute and complete the Data Description worksheet **GivenDataDescription**.
   e. Using a global Sort in Excel on each attribute in turn you can confirm that the values gathered from weka are correct. The worksheet once complete can be pasted into the ACW Stage1 report template. Ensure you submit that report to E-bridge on time.

4. Now establish a base-line classifier performance using the given data and j48
   a. Using the data loaded in step3, go to Classifiers→Trees→j48. Run this with default parameters.
   b. Copy the output into a new (suitably named) worksheet in **08338-DataWarehouse** .
   c. Use the classification performance to complete the first entry in the worksheet **Performance**.
   d. Check when you complete the TP, FP, TN, FN entries that the Sum agrees with the number of instances given to j48. (The Sum formula is present for this record – you need to add it for all other entries)
5. Now you can start on data cleaning – stepA Dealing with Unknown Risk
   a. Copy the **BaseData-All** worksheet to New Worksheet and name it **Risk-Known**
   b. Save Excel File
   c. In **Risk-Known** Find all records with Unknown, Missing or Null **Risk**
   d. Cut these from **Risk-Known** and paste into **Risk-Unknown**
      Note the attribute names are used as the first row and freeze that row.
      Test that this is complete by loading csv version of **Risk-Known** into weka.
   e. Update Methodology and DataCleaning with these changes and Save the Excel file.
   f. Again in weka use j48 to see what the classifier performance is.
6. Data cleaning – stepB Dealing with duplicates
   a. Copy **Risk-Known** to new worksheet **Clean0** in Excel file
      – remove duplicate records (use *Conditional Formatting* on attribute **Id** to highlight them)
      - Remember to paste cut data into **Data-Dirty** and make note in **DataCleaning** and **GivenDataDescription** to reflect what you have found and done
   b. Save Excel File
   c. Load **Clean0**.csv into weka and look to descriptions of the data
      Note you may want to remove the attribute **Id** when in weka (BUT not in Excel)
   d. use j48 on **Clean0**, and copy result into a suitably named worksheet**,** and update **Performance**
7. Data cleaning – stepC Dealing with missing values, nulls and unknowns.
   a. Copy **Clean0** to new worksheet **Clean1** in Excel file
      –look for nulls, missing values and Unknowns. Use *Conditional Formatting* to highlight them, and SORT to bundle them together.
      - Remember to paste cut data into **Data-Dirty** and make note in **DataCleaning** and **GivenDataDescription**
   b. Save Excel File
   c. Load **Clean1**.csv into weka and look to descriptions of the data
      Note you may want to remove the attribute **Id** when in weka (BUT not in Excel)
   d. use j48 on **Clean1**, and copy result into a suitably named worksheet**,** and update **Performance**
8. Produce **Clean2** –This is **Clean1** with all data records with all categorical values are in the same case (eg lowercase text) and use in weka (use the Excel function LOWER). At this point you should have clean data. You can try alternative data cleaning processes – see advanced work
9. **Hints** on Completing the Data Description and Data Cleaning Worksheets
   a. Complete the *FIRST* data description worksheet (**GivenDataDescription**) and start on the **SECOND** data description table for the CLEAN data (**FinalDataDescription** in Excel). You will not know until

the second or third ACW lab which is your BEST data so the **SECOND** Data Table Description will be incomplete until then. However the process is useful for validating the data edits.

b. Although the Data Description table2 cannot be completed until the "final" Clean data is decided upon, the statistics required to complete the frequency for the attribute values is useful for every cleaning strategy. You should use the frequency values (from weka or Excel) to check that the data is consistent and clean as expected. The sum of the frequency for all values for each (and every) attribute should be the same as the number of data records. If this is not the case, there is a problem with the data.
c. Examples are given in the tutorial and lecture notes (look for ACW lecture)
    i. Note range of values including frequency for null and missing
    ii. Use Excel where weka fails to give data descriptions

10. **Advanced**: It is also possible to identify patterns and missing value replacements using Means, Modes, Medians and Classifiers within Weka
    a. This will require extending the data cleaning worksheet for every data repair technique used.
    b. You will also need a worksheet to monitor the performance of j48 on the data as you clean and transform it (similar to that on page 42 or 43 of **DMDS-6-ACW2015-16**)
    c. If you do any of this advanced data mining you can include in the ACW reports at any stage BUT ensure you highlight in the final report (**ACW-Stage5**)

Ideally you need to have done all this up to and including step6 by the start of the second ACW week lab. Further data cleaning and transformation hints will be provided prior to future labs.

MAKE SURE YOU UPDATE THE METHODOLOGY SHEET AT EVERY STEP. Questions via lab, lectures, module forum or email – answers to email on the forum