

COMP37332

WEKA: Introductory tutorial¹

The first step in a data mining project is getting to know your data. In this tutorial, you will examine two data sets using the Weka framework. Weka is one of the main tools we'll be using this semester. It provides a large number of machine learning algorithms and visualisations useful for exploratory data mining.

The first task is intended to provide an introduction to Weka with a data that is ARFF formatted. ARFF is the data format for Weka, so no data transformation is necessary. The second data set is used to provide an introduction to issues associated with data transformation, some of which we discussed in Data warehousing. The raw data must be transformed into an appropriated formatted ARFF file. Once transformed, the steps taken for the first data set are repeated.

Task A: Using WEKA

Task A1: Download the two data sets (labor.arff and contact-lenses.arff) from the course web site. Examine the format used in ARFF files. Use ArffViewer (in version 3.5) to study the two data sets (the number of instances, attributes, etc.; are there any missing values). ARFF limits the attribute types it supports in its data files. What are the attribute types ARFF supports? Speculate what this might imply when using Weka as a data mining tool.

Task A2: Start Weka and load labor.arff file. Browse each of the attributes in the data file. Notice that Weka provides summary information for each attribute. The data in labor.arff contains two classes, bad and good. Looking at the values for the different attributes, select three attributes that might be good predictors for the class (i.e. if you knew the value for that attribute, you could guess pretty well whether the class for the same instance was good or bad). Explain why you chose the three attributes you chose.

Task A3: Select the Visualization tab in Weka. The Visualization tab provides a scatter plot with two data attributes as the axes. You can change which attributes are along the axes using the drop down menus in the top portion of the visualizer or by clicking on the plots in the right portion of the visualizer. Explore different scatter plots.

Repeat the task with the contact-lenses.arff file.

¹ Tailored after: *Lab 1: Getting To Know Your Data* (MSCS 228: Data Mining) by Dr Craig A. Struble, Marquette University

Task B: Transforming Data Into ARFF

The second part of this tutorial is to transform a raw data set into ARFF. As part of this process, you need to identify the different attribute types as those are required by the ARFF format. Note that the ARFF format is more limited in the kinds of data attributes. When using a system with limited data types, it may be necessary to transform the raw data values in order for the tool to work properly. Keep this in mind as you work through the steps below.

B1. Visit the URL <http://lib.stat.cmu.edu/DASL/Stories/teacherpay.html> and read the information about the Teacher Pay by States data set.

B2. Identify the attributes of the data. Record the attributes and the type of attribute for the data.

B3. Select, download and save the raw data set in a file.

B4. Convert the raw data set into CSV (comma separated value) format. One easy way to do this is to load it into Excel and use "Save As..." to save the file in CSV format. You can also write a program to do this. Java, Perl, Python, and shell scripts all work well. When you do this conversion, it is considered good practice to replace any nominal values represented by numbers with their textual representation. This will make the data easier to interpret as you summarize and mine it.

B5. Edit the CSV file, and add the ARFF header information to the file. This involves creating the @relation line, one @attribute line per attribute, and @data to signify the start of data. It is also considered good practice to add comments at the top of the file describing where you obtained this data set, what its summary characteristics are, etc. A comment in the ARFF format is started with the percent character % and continues until the end of the line.

B6. Load your ARFF file into Weka and repeat the steps you performed in Task A. You may run into errors as you load your ARFF file - look at <http://weka.sourceforge.net/wekadoc/index.php/en:Troubleshooting> for tips how to solve the problem.

B7. Repeat the analysis of the data set as you did in Task A. In addition, there exists one data point that appears to be a clear outlier. What point is this? (Try to use Weka to identify this point). Do you see any natural groupings from the teacher pay vs. spending per student dataset? Do the natural groups correspond to the three regions in the dataset?