G52DMA (COMP4030) Coursework (2018/2019)

DATA ANALYSIS USING R AND WEKA

Overview

The coursework is organized into three parts, each one focusing on a different and important aspect of Data Pre-processing, Data Analysis and Data Mining. All parts involve the use of the same dataset. The first part focuses on describing and visualizing the data and preparing the data for subsequent treatment ('pre-processing'). The second part focuses on clustering and the third part focuses on classification. The main goal is to give you first-hand experience on working with a relatively large and real data set, from the earliest states of data description to the later stages of knowledge extraction and prediction.

Data Set

The data set is a slightly modified version of a real-world plant data set. The data concerns the classification of plant species. Each record consists of several attribute (input) columns, and one class (output) column corresponding to the information about the type of plant. The attributes are markers that have been determined by assessment of different features of each plant, and the class variable is a provisional labelling of the type of plant. The entire data set consists of 724 instances (plants studied). Some of the variables contain missing values, which are indicated by empty entries.

Software and Data Report Deliverable

Although it is possible to make use of many different software tools in order to answer the coursework questions, **you are required to only use R and Weka, as indicated in the details below**.

In order to complete this coursework you will need to submit a written report describing all the analyses conducted. The length for the report should be between **2000 and 3000 words and twenty sides of A4**, *excluding the cover page, but* ***including all*** *tables and figures*. **Number all of your pages and make sure to include your name and student ID on the front page.** The minimum font size allowed is 11pt (a full page of text in a similar style to this document would contain about 500 words, so the majority of the 20 sides will be tables and figures). The report should clearly explain **what you did with the data, how you did it and why you did it, and it should be well structured and illustrated**.

Your report should contain three sections in total as described below. You **cannot include any lengthy code, or raw output (e.g. the output of R commands) in the main body of your report**. Include a copy of your code in an appendix. Note that appendices will not contribute to the word count, and are not explicitly marked: they are for reference only.

Marks

Part 1 carries 35 marks, while Part 2 and Part 3 carry 20 marks each. In total, this coursework aggregates to 75 marks. Marks will only be awarded for the **first twenty pages** of the main body of your report.

Assessment Criteria

The main assessment criteria for the report are:

- Correctness – that is, do you apply techniques correctly; do you make correct assumptions; do you interpret the results in an appropriate manner; etc.?

- Completeness – that is, do you apply a technique only to small subsets of the data; do you apply only one technique, when there are multiple alternatives; do you consider all options; etc.?
- Originality – that is, do you combine techniques in new and interesting ways; do you make any new and/or interesting findings with the data?
- Argumentation – that is, do you explain and justify all of your choices?

Plagiarism and Collusion vs. Group Discussions

As you should know, plagiarism and collusion are completely unacceptable and will be dealt with according to the University's standard policies. Having said this, we do encourage students to have general discussions regarding the coursework with each other in order to promote the generation of new ideas and to enhance the learning experience. Please be very careful not to cross the boundary into plagiarism. The important part is that when you sit down to actually do the data analysis/mining and write about it, you do it individually. If you do this, and you truly understand what you have written, you will not be guilty of plagiarism. Do NOT, under any circumstances, share code or share figures, graphs or charts, etc. As examples, saying to someone, "I used a Pivot Table in Excel to do the cross tabulations" is completely fine; whereas Copying & Pasting the actual Pivot Table itself would be plagiarism.

Deadline and submission procedure

- The submission deadline is May 15th via Moodle at 3PM.
- Name your report DMA-Cwk-XXX.pdf, where XXX should be replaced with your student ID number (e.g. DMA-Cwk-4078181), and submit the single document via Moodle (see website for details).

PART 1 – DESCRIPTION, VISUALISATION AND PRE-PROCESSING [R ONLY: 35 MARKS]

a) Explore the data [5]
   i. Provide a table for all the attributes of the dataset including measures of centrality, dispersion, and how many missing values each attribute has.
   ii. Produce histograms for each attribute. Provide details on how you created the histograms and comment on the distribution of data. You may also use descriptive statistics to help you characterise the shape of the distribution.

b) Explore the relationships between the attributes, and between the class and the attributes [6]
   i. Calculate the following correlations. What do these correlations tell you about the relationships between these variables?
      i. ***orientation 1*** and ***orientation 7***
      ii. ***mass*** and ***orientation 0***
      iii. ***orientation 7*** and ***orientation 8***
   ii. What are the two most correlated variables? What does this tell you about the data?
   iii. What are the two least correlated varirables? What does this tell you about the data?
   iv. Produce scatterplots between the class variable and ***orientation 2***, ***depth*** and ***area*** variables (note: you may have to recode the class variable as numeric to produce scatterplots). What do these tell you about the relationships between these three variables and the class?

c) General Conclusions [5]

   Take into considerations all the descriptive statistics, the visualisations, the correlations you produced together with the missing values and comment on the importance of the attributes. Which of the attributes seem to hold significant information and which you can regard as insignificant? Provide an explanation for your choice.

d) Dealing with missing values in R [5]
   i. Write a script in R to find missing values and replace them using three strategies: replace missing values with 0, mean and median.
   ii. Compare and contrast these approaches.

e) Attribute transformation [6]

   Using the three datasets generated in d), explore the use of three transformation techniques (mean centering, normalisation and standardisation) to scale the attributes, and compare their various effects.

f) Attribute / instance selection [8]
   i. Starting again from the raw data, consider attribute and instance deletion strategies to deal with missing values. Choose a number of missing values per instance or per attribute and delete instances or attributes accordingly. Explain your choice. In the case in which missing values remain in the dataset, explain which next steps you would take to remove those.
   ii. Start from the raw data, use correlations between attributes to reduce the number of attributes. Try to reduce the dataset to contain only uncorrelated attributes and no missing values.
   iii. Use principal component analysis in R to create a data set with seven attributes. Explain the process and the result obtained.

As a result, you will end up with several different sets of data to be used in Part 2 & 3. Give each set of data a clear and distinct name, so that you can easily refer to again in the later stages.

PART 2 – CLUSTERING [R ONLY: 20 MARKS]

Using only R, explore the use of clustering to find natural groupings in the data, *without using the class variable* – i.e. use only the numeric (input) attributes to perform the clustering. Once the data is clustered, **use the class variable to evaluate or interpret the results** (how do the new clusters compare to the original classes?).

a) Use hierarchical, k-means, PAM as clustering algorithms to create classifications of five clusters and write the results. Which algorithm produces better results when compared to the class attribute? [10]

b) Each of these algorithms has adjustable parameters. Explore the automatic 'optimisation' or 'tuning' of these parameters. Which parameters produce the best results for each clustering algorithm? Provide the reasoning of the techniques you used to find the optimal parameters. [5]

c) Choose one clustering algorithm of the above and perform this clustering on these alternative datasets that you have produced as a result of Part 1: [5]
   i. The reduced data set featuring 10 Principal Components.
   ii. The dataset after deletion of instances and attributes.
   iii. The three datasets after you replaced missing values with the three techniques.
   iv. Which of these datasets had a positive impact on the quality of the clustering? Provide explanations using the results for each clustering of the alternative data set.

Part 3 – Classification [Weka and R: 20 marks]

You must use Weka to perform the classification, but you may use R to present results. Using Weka classification techniques to create models that predict the given class from the input attributes. Split the data (randomly) into a training set (2/3 of the data) and a test set (containing 1/3 of the data);

a) Use the following classification algorithms: *ZeroR*, *OneR*, *NaïveBayes*, *IBk* (k-NN) and *J48* (C4.5) algorithms. Which algorithm produces the best results? [10]

b) Choose one classification algorithm of the above and explore several parameters. What does each of the parameters you chose represent? Which parameters improve the predictive ability of the algorithm? [5]

c) Choose one classification algorithm of the above and use the data sets you created in Part 1 [5]:

   i. The reduced data set featuring 10 Principal Components.

   ii. The dataset after deletion of instances and attributes.

   iii. The three datasets after you replaced missing values with the three techniques.

   iv. Which of the datasets had a good impact on the predictive ability of the algorithm? Provide explanations using the results for each clustering of the alternative data set.