Data Mining and Decision Systems 08338 Coursework

Data Mining of Legacy Data

Available Week 7 (3rd week of teaching)

Multi-Staged Submission via E-bridge

- ***2pm 2 November 2015 Data Description (PDF of Excel Worksheet with brief description)***
- ***2pm 16 November 2015 Data Cleaning (PDF of Excel Worksheet with brief description)***
- ***2pm 30 November 2015 Classifier Performance (PDF of Excel Worksheet with brief description)***
- ***2pm 14 December 2015 Data Warehouse Complete (Complete Excel Data Spreadsheet***
- ***2pm 14 December 2015 Report (PDF File with TurnItIn Report)***

Templates for all reports are available on the module SharePoint site. Feedback will be given during ACW oriented labs, annotation on the submitted reports and a module-level feedback summary sheet.

Aims

This coursework provides assessed practical experience in a simple data-mining project to aid decision support using legacy data (in this instance from cardio-vascular medicine). It requires the description and analysis of data for the knowledge domain, the manipulation of the given data and the generation and use of multiple forms of classifiers. The output from the classifiers is to be combined as decision rules to provide outcomes for unclassified data patterns.

It provides experience in applying data mining techniques to a real (if simplified) domain. All software required for the project is available in the labs and from the module web-page or from the internet. The raw data for the project is from the module SharePoint site as a zip file (**DMDS-ACW2015.zip**) containing a MS-Excel spreadsheet (***DMDS-2015-ACW-Data.xls***) and supplementary information. Support material will include ACW lectures, and lab sheets.

The given data is synthetic but derived from a real data set. The data worksheet (**BaseData-All**) contains duplicates, missing values, irrelevant

attributes and unknown outcomes. You are expected to filter, clean and transform the data so the data describing known outcome patients can be used to classify the unknown outcome patients. The Outcome attribute is the patient Risk.

Coursework Specification

You will be assessed on staged data mining project requiring three short reports plus a final report and the worked data. Templates are available for each report. The data file has been initialized with extra worksheets for each of the ACW stages.

Stage 1. Data Description. 5% of ACW

Due: 2pm 2 November 2015 via E-Bridge

Report: One page report (plus front page as PDF file), typically a table from the Excel Worksheet describing the given data (**BaseData-All**), based on ***GivenDataDescription*** in *DMDS-2015-16-ACW-Data*.xlsx, with a brief description.

Report Template: ''*08338-Stage1-Data Description-Template.docx*''

Stage 2. Data Cleaning. 10% of ACW

Due: 2pm 16 November 2015 via E-Bridge

Report: One page report (plus front page as PDF file), typically a table from the Excel Worksheet describing the data cleaning process (based on ***DataCleaning*** in *DMDS-2015-16-ACW-Data*.xlsx) with a brief description.

Report Template: ''08338-Stage2-Data Cleaning-Template.docx''

Stage 3. Classifier Performance. 10% of ACW

Due: 2pm 30 November 2015 via E-Bridge

Report: One page report (plus front page as PDF file), typically a table from the Excel Worksheet describing the classifier performance (based on ***Performance*** in *DMDS-2015-16-ACW-Data*.xlsx) with a brief description.

Report Template: ''08338-Stage3-ClassifierPerformance-Template.docx''

Stage 4. Data Warehouse. 25% of ACW

Due: 2pm 14 December 2015 via E-Bridge

Report: Your complete Data Warehouse for the ACW based the given data file *DMDS-2015-16-ACW-Data*.xlsx.

Stage 5 Data Mining Report. 50% of ACW

Due: 2pm 14 December 2015 via E-Bridge with TurnItIn Report

Report: Four to Six page (plus Front page and Contents page) report (PDF), as described below.

Template available as ''08338-Stage5-DataMiningReport-Template.docx''

Stage 5 requires the submission of a short report (about 2000 words - 4 to 6 pages plus Front Sheet, Table of Contents and any references and appendices). The report should include the ***clearly identified sections listed below (as given in the Template)***. Guidelines are given to the length of each section and their contribution to the overall mark of 50% for the report.

1. Technique Selection (10%). What classifiers in weka might be suitable for the domain data? Substantiate your reasoning through appropriate criteria. A Comparison Table based on lecture material with supporting text is suggested for pass marks. This table extended with weka ACW specific classifiers gets high marks. This plus a discussion of the implications of the classifiers for a Health Clinic receives very high marks.
2. Final Data Description (10%). Produce one table describing the final transformed data. This should describe the final clean data, type of data attributes and their value ranges, transformed attributes, and their new value ranges. You can make a spreadsheet in Excel based on the completed Description worksheet (***FinalDataDescription***) but edited to cover data transformations.

   Correctly completed table alone will receive 5 from the 10 marks. Full marks for this section should include text covering data analysis, for example: statistical analysis, clustering experiments, pattern frequency and expected classifications.
3. Classifier Decision Rules (15%). Use two classifiers (j48 and one non-tree other) plus an association rule generator (e.g. Tertius) in weka to produce decision rules for classifying patients as High or Low risk.

   Show the High Risk rules in a Decision Table using attribute-values that are consistent with the Patient test set.

   In a second table highlight any contradictory (conflict) rules - i.e. rules that disagree with those given in the first table.

   For full marks in this section, advanced informatics about the rules and their preferences should be given and briefly discussed
4. Deployment (15%). Using a table with supporting text, describe how your Classifier Decision Rules (from part 3) can be used with the data with unknown Risk (Hint: BaseData-All contains FIVE such

examples). This table should be based on your DSS worksheet in the Excel file.

What classifications for each of the patients do your rules produce? For full marks in this section, consider if you were building a decision support tool, and address the following two questions.

- What alternative ways of classifying the data would complement the deployed rules?
- What would be issues for Deployment in a Health Clinic?

(1 page) **15%**