Data Mining and Decision Systems

08338

Assessed Coursework

Data Mining of Legacy Data

Student Number: STUDENTID

Stage5: Due • 2pm 14 December 2015 Report (PDF File with TurnItIn Report)

Date: Monday, 05 October 2015

Table of Contents

[1. Technique Selection 1](#_Toc431799508)

[2. Final Data Description 1](#_Toc431799509)

[3. Classifier Decision Rules 1](#_Toc431799510)

[4. Deployment. 2](#_Toc431799511)

[5. References 2](#_Toc431799512)

[Appendix A 3](#_Toc431799513)

[Appendix B 3](#_Toc431799514)

# Technique Selection

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.

See DMDS-X03-Selection-and-Comparison for help.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Criteria | Task | Classifier1 | Classifier2 | Classifier3 |
| *Criteria1* | value | value | value | value |
| *CriteriaN* | value | value | value | value |
| *Summary* | value | value | value | value |

1. This is an Example Table for this section Using Styles for Caption and Table

(1 page) 10%

# Final Data Description

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.

(1-2 page) **10%**

# Classifier Decision Rules

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 used in section 4**.** 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.If you have done extra conflict sets for all rules mention it and highlight where they can be found (e.g. in the submitted Data Warehouse or as an Appendix to this document)

(1-2 pages) **15%**

# Deployment.

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.



1. This is an Example Table for Deployment

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%**

# References

Dunham, M. H., Data Mining, Prentice-Hall, 2002.

Fox, J., Glasspool, D., Patkar, V., Austin, M., Black, L., South, M., Robertson, D. and Vincent, C. (2010) 'Delivering clinical decision support services: there is nothing as practical as a good theory' in J Biomed Inform, United States: 831-43.

Hall, M., Frank, E., Holmes, G., Pfahringer,B., Reutemann, P. and Witten, I.H. (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Haykin, S., Neural Networks : A Comprehensive Foundation, Prentice-Hall, 1999.

Inc, S. (2000) CRISP -DM 1.0, Chicago, Ill.: SPSS Inc.

Mitchell, T.M., Machine Learning, McGraw-Hill, 1997

Shearer C., The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13—22.

Weka Home Page, http://www.cs.waikato.ac.nz/ml/weka/ Last Accessed 20 November 2014.

Witten, I.H., Frank, E. and Hall, M.A. Data Mining: Practical Machine Learning Tools and Techniques (3/e), Morgan Kaufmann, 2011

Appendix A

This appendix shows some extra detail



Appendix B

This appendix shows some extra detail on data cleaning and classifier performance