CO832 Data Mining and Knowledge Discovery

Assessment 2: Practical Data Mining with the WEKA tool – 2019

Instructions and General Marking Scheme

Questions should be directed to Marek Grzes

Last updated: 07/03/19 19:11:24

You can do this assessment either individually or in a small group with just two people. In the latter case, the group must hand in a single assessment, and the two students in the group will get the same mark.

This assessment is worth 10% of the total marks for this module.

Software
This assessment involves the use of the freely-available data mining tool WEKA to build decision trees for a particular, real-life dataset. The dataset will be used as input to the J48 algorithm (a decision tree induction algorithm) available in WEKA.

In order to do this assessment in WEKA, you won't need to do any programming. However, if you wish, you can use the R programming language instead of WEKA. This way, you can have a chance to write some R code to execute the learning algorithm on the data that is provided with this assessment[1].

Decision Tree Algorithms
The algorithm that you will use is called J48 in WEKA. Its original implementation in C is called C4.5 and C5.0. This means that you can read books or research papers about J48 and C4.5/C5.0 because they are in practice the same algorithm. J48 is written in Java and is part of WEKA. C4.5/C5.0 are available in R. You could potentially run C4.5 and C5.0 using their original C code that is publicly available on the web. If you decide to use R, I would recommend the C50

1 I suggested R because it has implementations of the decision tree algorithms that we need for this assessment. In particular, it contains C4.5 and C5.0 which are algorithmically equivalent with WEKA's J48. I wanted to suggest Python, but I did not find any reliable implementations of C4.5/C5.0/J48 algorithms in Python.

package: https://cran.r-project.org/web/packages/C50/vignettes/C5.0.html and https://cran.r-project.org/web/packages/C50/C50.pdf

Dataset

The dataset is called Pima Indians Onset of Diabetes. Each instance represents medical details for one patient and the task is to predict whether the patient will have an onset of diabetes within the next five years. There are 8 numerical input variables all of which have varying scales. You can learn more about this dataset on the UCI Machine Learning Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html). Top results on this dataset are in the order of 77% accuracy.

The dataset is available in a file called diabetes.arff in the module's web page on Moodle. The file extension ".arff" means it is a file in a format specifically suitable for WEKA.

Data Format

After a few lines at the top of the file defining the dataset name, its attributes and corresponding values, each line in the file represents a data example/instance, and it consists of several attribute values. One of those attributes is then chosen as a class attribute, whose value is to be predicted by a classification algorithm. Missing values (if present in a dataset) are represented using the "?" symbol.

Note that the arff format is not recognised by some text editors as a plain text file, even though it effectively consists of plain text only. To see the contents of an ".arff" file, you may need either to load it into WEKA or to temporarily change its extension to ".txt", so that it would be recognised by most text editors. If you do that, of course do not forget to change the extension back to ".arff" before you try to run WEKA on that file.

Instructions About the Experiments To Be Done

In order to run the experiments specified below, first read the short tutorial on the use of WEKA available in the module's web page on Moodle. You can also find out more about the WEKA tool by reading the book chapter describing that tool (reference in the above tutorial), reading the online explanations about algorithms and parameters in WEKA, and maybe reading specific sections of the (very large) WEKA manual (available from the WEKA website: http://www.cs.waikato.ac.nz/

ml/weka/index.html - click on the "Documentation" link on the left-hand-side of the page).

Using WEKA, open the file diabetes.arff, and set the class attribute to "class". As mentioned in the comments in the diabetes.arff file, the value of this attribute for a given instance (patient) is 1 when the patient is "tested positive for diabetes". The class attribute has values 0 or 1.

Next, make sure the test mode for evaluating predictive accuracy is set to 10-fold cross-validation. This is the default for a just loaded dataset.

Then, use the J48 algorithm (a decision tree induction algorithm) to perform the following experiments.

First, using all the default parameters of J48, run that algorithm and analyse its results. This involves the measure of predictive accuracy obtained by J48 (mainly the % of correctly classified instances), as well as analysing the decision tree built by J48, i.e. trying to interpret the decision tree to understand how characteristics of a data example (represented by its attributes) can be used to predict the value of the class attribute.

Next, find out what is the meaning of the parameters of the J48 algorithm, and then run J48 several times, each time varying the value(s) of one or more J48 parameters. For each new run, analyse the results of that run, as explained above. It will make a lot of sense to change several parameters at a time.

When cross-validation is used in WEKA, there is no information in the output about the error on the training data. The % of correctly classified instances that is printed for cross-validation is the average across k-folds of cross-validation. In order to obtain the % of correctly classified instances on the training data, you will need to repeat every experiment selecting "Use training set" instead of "Cross-validation" in the "Test options" panel in the "Classify" tab. This means that you will need to run every experiment (i.e., for every set of J48 parameters) twice to record accuracy on the training data and on cross-validation.

What to hand in

Your report should have the following structure: (1) a dedicated title page with your name and Kent login, (2) one page of technical analysis, (3) appendix. All

parts should be of the usual size A4 text. The technical analysis is limited to **one** page; if your technical analysis exceeds one page, only the first page of the technical analysis will be marked, and the remaining part of that section will be ignored. The appendix can include figures, tables, and bibliography, and it is unlimited in its size, i.e., you can have as many pages in your appendix as you want. If you do this assessment in R or any other programming language, or if you have to write any code for this assessment to enrich your analysis, include the listings of all your code in the appendix. All parts of your report should use font size of at least 11pt. Your tables and figures in an appendix should be labelled, and they should be referenced from your technical analysis; it would make sense to add reasonable captions to your tables and figures in the appendix. The ability to summarise your results and analysis is part of the assessment, and this is the reason why we have these strict requirements.

Your technical analysis should consist of two separate sections, as specified below. If your report is not clearly divided into those two sections, 5 marks will be subtracted from your overall mark for this assessment. Your appendix does not have to be divided into those two sections. We don't have any requirements with respect to the structure of the appendix.

Section (1) – Analysis of the effect of different J48 parameter settings in the predictive accuracy of the constructed decision trees

Report a table and a graph showing the % of correctly classified instances obtained by J48 (using both training data and 10-fold cross-validation) for each of the several unique settings of parameter values used in your experiments (including the default parameter setting). Also report an analysis of the table, explaining the main conclusions that you have drawn from those results – i.e., explaining which parameter setting(s) led to higher or lower predictive accuracy than other parameter setting(s), and explaining *why* that has occurred. Use technical terms to explain and justify your observations. Mentioning observations without explaining and justifying them won't give you the highest marks.

Section (2) – Analysis of the effect of different J48 parameter settings in the comprehensibility of the constructed decision trees

Compare and contrast several (not necessarily all) decision trees built by J48 in your experiments, discussing the pros and cons of different decision trees built using different settings of parameter values. That is, explain why changing the value of a parameter in one experiment produced a better or worse decision tree than other parameter values in other experiment(s), in terms of the

comprehensibility of the constructed trees. The comprehensibility of a tree is partly related to its size, since it is difficult to interpret a very large tree, but your analysis of a decision tree's comprehensibility also depends on your interpretation of the tree – i.e., your understanding of how characteristics of a data example can be used to predict the value of the class attribute.

In this section, you should illustrate your answer with one or two of the built decision trees, or at least the major part(s) of each tree, if a tree is too large to be shown as a whole. In particular, you must show (at least the major part of) the "best" decision tree that you obtained in your experiments, and clearly mention the J48 parameter settings used to build that tree. You can show the tree either in textual format or in a graphical format, but make sure that all the attribute names and other types of information in the tree are clearly legible – e.g. the attribute names in the tree should not be shown in a font size that is too small. Justify why you chose the tree that you are reporting here, out of all trees built across all runs of J48 in your experiments, as the "best" decision tree. Interpret that tree, discussing what it tells you about how some characteristics of a data example can be used to predict its class attribute. You can consider the shape of the trees or the location of different attributes in your trees. Your should argue which trees are comprehensible and which attributes are important. Your arguments should have technical justification.

Note that if you run J48 a number of times, you may not have space in the report to analyse all your results. In this case you have to be selective, and analyse only the most important results.

Deadline

The printed technical report has to be handed in to the Student Administration Office by the deadline for this assessment, which is specified in the Student Data System. It is your responsibility to find out what time the Student Administration Office closes on the day of the deadline. If you prefer to submit your report electronically, there will be a Turnitin link on the Moodle page that will allow you to so. For electronic submissions, the preferred format is PDF. If you submit ODT or DOCX, your file will be automatically converted to PDF using my bash script that will run libreoffice on your file.

If you submit electronically, please include your name and your Kent login on the title page of your document. It takes a lot of time for us to deal with anonymous submissions after we have printed them.

Time Estimated to Complete the Assessment

The time that students take to write a short report as required in this assessment varies significantly across students; but as a rough estimate, students can be expected to spend about 20 hours to do this assessment. This estimate refers to the total time to do the assessment, i.e., including the time to read documentation about the parameters of the J48 algorithm in WEKA, carry out the experiments and analyse the results, and write the technical report. Note that this time estimate assumes that the students have been learning the module material on a regular basis. If they did not engage in intensive self-study and reflection on the material provided in the lectures, they may need to spend considerably more time on this assessment.

Notes on Plagiarism

Senate has agreed the following definition of plagiarism:

"***Plagiarism is the act of repeating the ideas or discoveries of another as one's own. To copy sentences, phrases or even striking expressions without acknowledgement in a manner that may deceive the reader as to the source is plagiarism; to paraphrase in a manner that may deceive the reader is likewise plagiarism. Where such copying or close paraphrase has occurred the mere mention of the source in a bibliography will not be deemed sufficient acknowledgement; in each such instance it must be referred specifically to its source. Verbatim quotations must be directly acknowledged either in inverted commas or by indenting.***"

The work you submit must be your own, except where its original author is clearly referenced. We reserve the right to run checks on all submitted work in an effort to identify possible plagiarism, and take disciplinary action against anyone found to have committed plagiarism. When you use other peoples' material, you must clearly indicate the source of the material.

General Marking Scheme

Your technical report will be assessed based on two main criteria: (1) technical quality, and (2) the comprehensibility of the report. Technical quality, which is the most important criterion, involves the correct use of technical terms, concepts and arguments. In general, the more advanced (and correct) the technical concepts and arguments that you used in your report, the higher the mark. The comprehensibility of the report involves the use of well-written sentences, which are understandable and meaningful, as well as grammatically correct. It also

involves the use of clear figures to illustrate your arguments – for instance, you will lose marks if your figure includes text in a very small font size which is hard to read. The more clearly (and correctly) written your text is, and the clearer the figures are, the higher the mark.

Your report will be assigned a mark based on a *categorical marking scale* used by the University, which includes a range of a few discrete numerical marks for each categorical mark, as follows:

Mark range: 100, 95, 85, 78, 75, 72

Marks within that range are allocated based on the extent to which your technical report has the following characteristics:
The analyses in both section (1) and section (2) of the report are of excellent technical quality, reporting all the information required in the assessment's instruction and with many arguments that involve advanced technical concepts and are clearly and correctly explained – with no technical mistakes.

Mark range: 68, 65, 62

Marks within that range are allocated based on the extent to which your technical report has the following characteristics:
The analyses in both section (1) and section (2) of the report are of very good technical quality, reporting all the information required in the assessment's instructions and with several arguments that involve advanced technical concepts and are in general clearly and correctly explained – possibly with a few relatively minor technical mistakes or a few hard-to-understand sentences.

Mark range: 58, 55, 52

Marks within that range are allocated based on the extent to which your technical report has the following characteristics:
The analyses in both section (1) and section (2) of the report are not of good quality, in general, but at least the report contains most of the information required in the assessment's instructions. The technical arguments are not clearly and correctly explained – there are some significant technical mistakes and possibly many hard-to-understand sentences. If section (1) and (2) are of reasonable technical quality with arguments that show legitimate (although not advanced) technical knowledge, then a higher mark in this range (e.g. 58) will be allocated.

The marks below (i.e., marks < 50) correspond to a "fail" mark, since the pass

mark for this module is 50%.

Mark range: 48, 45, 42

Marks within that range are allocated based on the extent to which your technical report has the following characteristics:

The analyses in both section (1) and section (2) of the report are of poor quality, in general, and/or the report contains a relatively small part of the information required in the assessment's instructions. The technical arguments are not clearly and correctly explained – there are many significant technical mistakes and many hard-to-understand sentences, and/or too few technical arguments.

Mark range: 38, 35, 32, 20, 10, 0

Marks within that range are allocated based on the extent to which your technical report has the following characteristics:

The analyses in both sections (1) and section (2) of the report are of very poor quality, in general, and/or the report lacks most parts of the information required in the assessment's instructions. The technical arguments cannot be understood, and/or the exiting arguments are invalid.