CS-422 Fall 2018

Assignment 2

In this assignment you will use the Weka package and the data provided with the installation.

Page limit: 10 pages.

Weka Installation: http://www.cs.waikato.ac.nz/ml/weka/

Assignment

Use 3 data set provided with the package in the data folder:

Iris, Vote, Diabetes

1) Data Analysis

1.1 For each data set describe the number and type of attributes (no need to list the type of every attribute, just summarize in a table what attribute types are present and how many attributes of each type)

1.2 For the Iris data set also report

-Number of attributes

-Min/Max values, standard deviation for each attribute

- Discuss briefly the ranges of each of the attribute and explain *how that matters* for the classifier

-Use the examples that we covered at the end of lecture 3 to discuss the data, variables overlap, what performance we can expect by analyzing and visualizing the data.

2) For each data set use 3 decision tree algorithms: Decision Stump, J48, Random Forest

Note that we are using Decision Stump just to illustrate the steps of the decision tree algorithms, it is a very simple algorithm and unlikely to be used in practice.

- Look at the parameters for each classifier that you can set in Weka. Read in the Weka manual about what they mean. What do you think are the main parameters, based on our discussions in class? Different parameters may be important for different classifiers. Why do you think those parameters are important?

-Describe the class distribution for each dataset. How does it matter for the classifier?

-Use 10-fold cross-validation

--Discuss the size of training set and the test set for each iteration of cross-validation. What about the class distribution? What happens in the training/test set created in cross validation? Do we still have the same class distribution as in the full data set for them? Does it matter?

- In the test options tab, choose “more options” and explore the classifier evaluation metric using a cost function. How did you set the cost parameter and why? How does it affect the results and why?

-Describe the classification accuracy (on training and on test set, what is the difference? Does it matter?)

-Discuss the size of the tree, number of leaves.

-Analyze some of the most important parameters that you identified above (for maximum 3 of those parameters) - Use 2 different values from default for each of those parameters. How does accuracy/size of tree/number of leaves change if you change the parameters? Explain how each change affected the performance and why. Provide a plot or table with the results.

- For the Random forest classifier analyze the number of trees. Find the parameter that you can use to set the number of trees. Try different number of trees, from a few to a very large number (10, 20, 50, 100, 200, 500, 1000). For each option note the classifier performance and the time it takes to train. Have 2 plots for your report (Plot theperformance vs #trees, time vs #trees). Analyze the results.
- Set the max depth of trees to 3 different values (small, medium, large). Use 3 different option for the number of trees (10, 100, 500). Analyze the results –do the max depth and the number of trees work together?
- Use the option to print the classifier (for a small number of trees), can you see anything interesting in the base tree classifiers? (It's an open ended question, just report what you think)

3) Feature selection

We haven't discussed this subject in class yet, but you have enough background now to experiment with this approach. The idea is that not all features may be important for building the model. While many classifiers can "figure out" important features during the model training process, it is often helpful to remove unimportant features before training and also before testing. It improves classifier performance and also makes the feature space smaller adding an improvement in efficiency.

-For each data set use 2 ways to select features in Weka: CorrelationAttributeEval and InfoGainAttributeEval. What attributes are selected for each data set. Are the selected attributes what you would expect after the data analysis part? Why? How does the feature selection improve performance?

4) Noise and Missing values

Run the following experiments on one dataset only: Iris. Modify the data set and rerun the classification experiments. The data files are in the Weka installation folder, in the data subfolder. You can manually modify the data files using any editors.

-Introduce some missing values (insert question marks)

> -- Discuss what happens in terms of accuracy and training time if you add 5%, 10%, 50% of missing values

-Introduce noise (misclassify some of the examples). **Do NOT use the noise option in Weka**

> -- Discuss what happens in terms of accuracy and training time if you add 5%, 10%, 50% of noise examples.
>
> --Change the range of one attribute, make it 1000 times larger than the other features range. How does this affect the classifier's accuracy?

Describe all modifications in detail, run the three classifiers with the default options, analyze the new results and compare to what you had with the original data files.

In your report focus on:

What do you see

Why is it the case

Is this important

Does it help to understand the problem you are working on

Does it help to understand the results that you get with your approach

What did you learn

What do you want others to learn

Analysis and Conclusions are the most important parts of your report

Report tips

Do not include screen shots unless they illustrate a point that you discuss in your report. Shorter reports with good summaries and analysis are better than long reports with just listing results without evaluation.