

Curso: Mestrado Integrado em Informática – Engenharia do Conhecimento
U.C.: Knowledge Discovery

Exercise Sheet FE03	
Teacher	Cristiana Neto
Theme	Exploring Weka
Class	PL
Year	2019-20 – 2nd Semester
Duration	2 hours

1. Statement

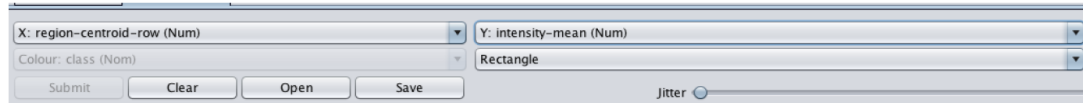
- [1] What are the differences between a database, a datawarehouse and a dataset?
- [2] What are some of the limitations of data mining and how can they be overcome?
- [3] What is the difference between datawarehouse and data mart?
- [4] Indicate some ethical constraints in the use and application of Data Mining.
- [5] What is database normalization and what are the impacts on OLTP and OLAP systems?
- [6] Design a relational database with at least three tables. Be sure to create the columns necessary within each table to relate the tables to one another.
- [7] Design a data warehouse table with some columns which would usually be normalized. Explain why it makes sense to denormalize in a data warehouse.
- [8] Do an online search and find 3 sites that contain information that can be applied to the Data Mining process.
- [9] Using the Internet, locate a data set which is available for download. Describe the data set (contents, purpose, size, age, etc.).

On the Weka home screen, open the "package manager" (Tools -> Package Manager). Install the "UserClassifier 1.0.3" package. After this step answer the following questions.

- [10] Open Weka / Explorer and load the data set "segment-challenge.arff". On the Classify tab, define data set "segment-test.arff" as test set.

[a] Use Trees -> UserClassifier;
Click on Start;

Select the Data Visualizer tab; and select the following options (another value can be used instead of the rectangle):



Select the groups you can define.
Determine the classification result.

[b] Compare the results obtained with this method of creating a decision tree with the results of the J48 algorithm.

[11] Open Weka/Explorer and load the data set “segment-challenge.arff”. With this data set loaded, answer the following questions:

[a] Use the J48 algorithm as a classifier; Use the data set “segment-test.arff” as a test set. What is the value of the classification?

[b] Using the “Use training set” option, determine the classification value. Why shouldn't this option be used to determine the quality and applicability of the algorithms to the data?

[c] Choose J48 as a classifier and change the percentages of split (“Percentage Split”) of the training and test groups by: 10%, 20%, 40%, 60% and 80%. What do you observe?

[d] Repeat the previous question using 90%, 95%, 98% and 99%. What happens to the number of correctly classified instances? And what happens to the percentage of instances correctly classified? Explain this variation.

[e] Although with a percentage of 98% for training and 2% for test gives a classification of 100%, does this mean that the model built is the most suitable for the problem presented?

[f] Based on the experiences above, what do you consider the best estimate of the true J48 accuracy for this data set?

[12] Open Weka/Explorer and load the data set “diabetes.arff”. With this data set loaded, answer the following questions:

[a] Selecting “Percentage Split” at 80% how many instances will be used for training and how many will be used for testing?

[b] Changing the “Random seed” (“More options”) between 1,2,3,4 and 5, keeping the “Percentage Split” at 80%, indicating the minimum and maximum value of instances incorrectly classified.

[c] What is the average percentage of instances correctly classified?

[d] If you repeated the exercise [12/b] with 10 “random seed” instead of 5, what would be the effect on the average?

COMPARE WITH “BASE LINE”

[13] Open Weka / Explorer and load the data set “iris.arff”. With this data set loaded, answer the following questions:

[a] The iris.arff dataset consists of three classes (Iris-setosa, Iris-versicolor, Iris-virginica), with 50 instances each. What is the accuracy of ZeroR on this dataset when testing on the training set?

[b] In practice, what is ZeroR's success rate on the iris dataset when evaluated using the default (66%) Percentage split?

[14] Open Weka/Explorer and load the data set "glass.arff". With this data set loaded, answer the following questions:

[a] What is the percentage of corrected classified instances of the ZeroR algorithm with 66% of "Percentage Split"?

[b] What is the result using J48 and the other default parameters?

[c] What is the accuracy of the NaiveBayes algorithm using the default parameters?

[15] Open Weka / Explorer and load the data set "segment-challenge.arff". Use the data set "segment-test.arff" for evaluation (test) dataset. With these loaded data sets answer the following questions:

[a] What is the accuracy of the ZeroR algorithm?

[b] What is the accuracy of the IBk's algorithm with all default parameters?

[c] What is the accuracy of the PART algorithm with all default parameters?