

Faculty of Science

**Course**: CSCI 4030U Big Data Analytics

**Lab:** #6-10

**Topic:** Data Mining with Weka

**Objective**

The objective of this project is to gain hands-on experience in using data mining software to build models from real world data sets and to conduct evaluations of different data mining algorithms.

**Software**

The software to be used in this assignment is Weka [1]. You can download and install it on your own machine. A guide on how to use Weka's Explorer can be found in attached *ExplorerGuide.pdf*. A page documenting the ARFF data format used by Weka can be found in attached *Arff.pdf*.

[1] <http://www.cs.waikato.ac.nz/ml/weka/>

You can use "java -Xmx256m -jar weka.jar" to modify the heap size when you invoke weka. You can increase the value of 256m if it is not enough.

**Task 1 (Lab 6)**

Read attached documentation about WEKA

Consider an attached [lymphography data set](file:///C:\Users\100569705\Desktop\Project\UOIT\Courses\CSCI%204030U\Weka\CSE6412%20-%20Data%20Mining\CSE6412%20-%20Data%20Mining\Assignments\Assignment%202\Assignment%202_files\lymph.arff) (lymph.arff) that describes 148 patients with 19 attributes. The last attribute is the class attribute that labels a patient with one of the four categories (normal, metastases, malign\_lymph, and fibrosis). Detailed information about the attributes is attached (lymph.txt). The data set is in the ARFF format used by Weka.  
  
Use the following learning methods provided in Weka to learn a classification model from the data set with all the attributes:

* C4.5 (weka.classifier.trees.J48)
* RIPPER (weka.classifier.rules.JRip)

Note that C4.5 can dynamically discretize the numeric attribute during the learning process, but Id3 requires a pre-discretization before learning is performed. Use the supervised discretization method provided in Weka to discretize the numeric attribute before using Id3.

For each learning method, report the classification model learned from the data set.

Do a research how both of the algorithms work and provide one paragraph description for each of them.

**Task 2 (Lab 6)**

You are given a [training data (monks-3.train.arff) set](file:///C:\Users\100569705\Desktop\Project\UOIT\Courses\CSCI%204030U\Weka\CSE6412%20-%20Data%20Mining\CSE6412%20-%20Data%20Mining\Assignments\Assignment%202\Assignment%202_files\monks-3.train) and a [test data set](file:///C:\Users\100569705\Desktop\Project\UOIT\Courses\CSCI%204030U\Weka\CSE6412%20-%20Data%20Mining\CSE6412%20-%20Data%20Mining\Assignments\Assignment%202\Assignment%202_files\monks-3.test) (monks-3.test.arff) in which each training example is represented by seven **nominal** attributes. The last attribute is the class attribute that labels a training example with one of the two classes (0 and 1). The attribute information is given below:

|  |  |
| --- | --- |
| Attribute | Values |
| a1 | 1, 2, 3 |
| a2 | 1, 2, 3 |
| a3 | 1, 2 |
| a4 | 1, 2, 3 |
| a5 | 1, 2, 3, 4 |
| a6 | 1, 2 |
| class | 0, 1 |

Use the following learning methods provided in Weka to learn a classification model from the training data set and test the model on the test data set:

* C4.5 (weka.classifier.trees.J48)
* Id3 (weka.classifier.trees.Id3)
* RIPPER (weka.classifier.rules.JRip)
* k-Nearest Neighbor (weka.classifiers.lazy.IBk)
* Naive Bayesian Classification (weka.classifiers.bayes.NaiveBayes)
* Neural Networks (weka.classifiers.functions.MultilayerPerceptron)

Report the classification accuracy and confusion matrix of each algorithm on the test data set. Briefly discuss your results.

Do a research how each of the algorithms work and provide one paragraph description for each of them.

**Task 3 (Lab 7)**

You are given a data set on credit card application approval (credit-a.arff) in the ARFF format used by Weka. The data set describes 690 customers with 16 attributes. The last attribute is the class attribute describing whether the customer's application was approved. The data set contains both symbolic and continuous attributes. Several of the condition attributes contain missing values (which are marked by "?"). All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data.   
  
Randomly split the data set into a training set (70%) and a test set (30%), which can be done by using the "percentage split" test option in Weka's "Classify" section. Apply each of the following algorithms to learn a classification model from the training set and classify the examples in the test set.

* C4.5 (weka.classifier.trees.J48)
* Naive Bayes Classifier (weka.classifiers.bayes.NaiveBayes)
* Neural Networks (weka.classifiers.functions.MultilayerPerceptron)

Report the classification accuracy of each learning algorithm on the test data set.  
  
Please note that C4.5, naive Bayes and neural networks can handle missing values and continuous attributes automatically.

Note: some algorithms cannot handle these. You need to fill in the missing values and discretize the continuous attributes before using these algorithms. Use the global estimation method to replace missing values and the entropy-based discretization method to discretize all the continuous attributes before using some other algorithms. Both data preprocessing methods are provided by Weka.

**Task 4 (Lab 7)**

Conduct 10-fold cross validation to evaluate the following classification learning algorithms:

* C4.5 (weka.classifiers.trees.J48)
* RIPPER (weka.classifier.rules.JRip)
* Naive Bayesian Classification (weka.classifiers.bayes.NaiveBayes)
* k-Nearest Neighbor (weka.classifiers.lazy.IBk)
* Neural networks (weka.classifiers.functions.MultilayerPerceptron)

on the following data sets from the [UCI repository](http://www.ics.uci.edu/%7Emlearn/MLSummary.html):

* Ecoli database (database\_ecoli.arff and database\_ecoli.txt). There is an ID attribute in this data set, which should be removed before learning).
* Glass Identification Database (database\_glass.arff and database\_glass.txt). There is an ID attribute in this data set, which should be removed before learning)
* Image segmentation Database (database\_image.arff and database\_image.txt)

Report the misclassification rate and run time of each algorithm on each data set. Discuss the results.

Discuss the results regarding whether there is an overall winner and whether the misclassification rates for the algorithms are significantly different.

**Task 5 (Lab 7-8) COMPETITION (Student with best results will get bonus points! Your mark will be partially based on the rank.)**

You are given a training data set (risk-train.txt) that describes 30,000 online purchase orders for an online trader. Each example in the data set corresponds to an online purchase order and is described by 44 attributes. A detailed description of the attributes can be found in (risk-attributes.txt). The second attribute is the target (i.e., class) attribute that indicates whether an order has a high risk of default payment. The class attribute has two values, "yes" meaning high risk and "no" meaning low risk.

Randomly split the data set into a training set (70%) and a test set (30%), which can be done by using the "percentage split" test option in Weka's "Classify" section.

Your task is to help the online trader to recognize if a person who makes an order is a customer who will eventually pay the goods by using data mining techniques. You will use a classification algorithm to build a prediction model based on the training data. This prediction model shall then be used for classifying incoming orders into the high risk or low risk class. The continual prognosis of the expected probability of loss of payment for the incoming orders. You will need assign each order in the test set to one of the two classes based on the prediction model learned from the training data. Your classification results are to be submitted in a file that contains a list of the orders with predicted classification in the following form:

ORDER-ID CLASS  
23083   no  
23082   yes  
23078   no  
23076   no

For this learning and classification task, data preprocessing is very important. The data set is a raw data set and contains missing values and possibly irrelevant attributes. You may consider using feature selection (i.e., removing some attributes) and other data preprocessing techniques (filling NULL values, discretize values).   
  
In addition to the classification result file, you will also need to describe the data preprocessing methods (e.g., discretizing birth date into buckets) either from Weka or optionally written on your own and the classification method used in your solution (test different ones). If you write any programs for data processing, submit your programs as well together with a readme file describing how to use the programs.

Report over the test set the **classification accuracy** (as in tasks before) and **misclassification cost** computed for the two classes that do not have equal weights. Below is the cost matrix for this data set.

|  |  |  |
| --- | --- | --- |
|  | High risk | Low risk |
| High risk | 0 | 50 |
| Low risk | 5 | 0 |

Students who will have the higher classification accuracy and lowest misclassification cost will get better mark for this task.

**Task 6**

**Lab 8/9**

Implement in a chosen programming language the Perceptron algorithm (see lecture notes: Machine Learning) and report the same results as for TASK 5 with a single pass with both fixed and changing rate.

**Lab 10**

Finishing and polishing all the tasks and final report.

**What to hand in after lab 10.**

* A .pdf file that contains your anwsers to the six questions. Please write your name and student ID in the file that you submit.
* The file that contains your classification result in the required format for Question 5 and 6.
* The programs, if any, that you write for solving Question 5, and a readme file showing how to use these programs (and the code for solving Question 6).