Practice of Weka

Practice work #2 - Arnaud QUIRIN - May 2008

The goal of this practice is to get familiar with the Weka software. This software is one of the most known to classify, clusterize, analyze the data for machine learning.

1 Starting point

In the case in which Weka is not installed, you have to download it in :
http ://www.cs.waikato.ac.nz/ml/weka/ (section Download)

Actually, it is simply a ZIP you will have to uncompress in a folder. To execute it into Linux, you just have to type : `java -jar weka.jar`

When Weka is opening, you will get a little window with 4 buttons. Look for `Explorer`. Then you will get 6 tabs : `Preprocess`, `Classify`, `Cluster`, `Associate`, `Select attributes` and , `Visualize`.

- `Preprocess` : to get a quick look to a data file and apply to it some filters.
- `Classify`, `Cluster`, `Associate` : to launch a learning. `Classify` is for supervised learning, `Cluster` for unsupervised learning and `Associate` to generate association rules. This practice will be only focused on the `Classify` mode.
- `Select attributes` : to select only the most significant attributes, using supervised algorithms.
- `Visualize` : to visualize in 2D the data.

2 The data

Weka use the ARFF format to save the data. It is a list of examples, given with their attributes values.

Use any editor to open the file `data/contact-lenses.arff`. See how the lines can be commented (using %), and try to figure out the use of the keywords `@relation`, `@attribute` and `@data`. Weka can work with different *kinds* of attributes : they can be nominals (a set of values like : {myope, hypermetrope}), integrals (for instance, 156), or reals (for instance, 45.22).

The data are all the lines following the keyword `@data`. Each line is an example (or object, or instances), given by the values of all their attributes. For instance, the first object is :

age	spectacle-prescrip	astigmatism	tear-prod-rate	contact-lenses
young	myope	no	reduced	none

Launch Weka. Go in `Explorer`, `Open File` and load `data/iris.arff`.

1. How many objects contain the data ? How many attributes ?
2. What is the name of the attributes and their indexes ?
3. What is the class of an object ?
4. What is the goal of the `Selected attribute` function ? Try and observe what happen when you select different set of attributes. What is the meaning of the values `Name`, `Type`, `Missing`, `Distinct` and `Unique` ?
5. What is the goal of the small graph below ? What is the goal of the tab in the top left of the graph ? What is the goal of `Visualize all` ?

3 Filter the data

Many times, you will have to filter the data before using a learning algorithm. By filtering, you can remove the instances corresponding to measure errors or you can remove attributes that are equivalent. You can also normalize the data.

You can use a lot of filters with Weka. Just go into `Choose` in the `filter` area. You can use supervised filters (which will use the class of the data) or unsupervised filters. You can also choose between attribute filters (concerning a whole attribute) or instance filters (applied only on single instances).

Select the non supervised filters. Study them and try to figure out what they do. When you have an idea about the possible effect of a filter, try it : select it and press `Apply` to verify if it goes as you expected it.

1. What do the attribute filters `remove`, `normalize` and `standardize` ?
2. What do the instance filters `normalize`, `RemoveMisclassified` and `resample` ?
3. What do the supervised attribute filters `AttributeSelection` and `StratifiedRemoveFolds` ?
4. Explain the goal of all these filters.

4 Visualization of the data

Look for the `Visualize` tab. Each point is an object in the data. Try to figure out :

1. What is the meaning of the different colors ?
2. How we can get the values of the different attributes of an object ?
3. What is the meaning of the several graphs and why they are organized like a grid ?
4. Explain the organization of the data in the graphs located on the diagonal (y=x).
5. Using the visualization tool, select the attribute that is, according to you, useful to classify the data. Which one and why ?
6. It seems that some attributes are not well suited to be used for the classification. How many and which ones ? Try to click in a cross and observe what happen. All the objects corresponding to this cross will be described more precisely in another window. You have some details, for instance the indexes of the selected data.
7. What is the goal of the Jitter ?
8. Explain how we can simulate a tri-dimensional graph, by changing the color of the objects.

5 Classification

The classification or *supervised learning* take as an input labelled data (data in which at least one attribute is considered as a class), and compute a model of these data. This model can then be used to classify new data (non labelled, or with an unknown class value).

By default, the selected classifier is `ZeroR`. Use it on the IRIS dataset.

Let's do some research ! What exactly this classifier is doing ? You can try to figure out what happen using the visualization or the statistical tool shown in the other tabs.

Then, look the `Classifier output` area. This area contains the informations given by the classifier.

1. Explain what is the information given by `Sheme`, `Relation`, `Instances`, `Attributes` and `Test mode`.
2. Let's do some research again ! Try to imagine a way to score the classifier. A score is a value (called also an accuracy) that is 1 if the classifier is perfect, and 0 if the classifier is bad in all cases. Remember that to compute this score, you can use the result of the classifier on each instance, and you can also choose which proportion of the dataset the classifier can see for the learning.

3. What is a learning set ? What is a training set ?
4. More particularly for `Test mode`, try to figure out (or use Internet) the differences between using a test ensemble to compute this score, or the cross-validation procedure. These two way of computing the accuracy are really important in supervised learning.
5. Which information is given by `Classifier model` ?
6. Which information is given by `Summary` ?
7. What is a confusion matrix ? Look on the web about how it is computed. What a perfect confusion matrix looks like ?

The algorithm `OneR` selects one attribute and classifies using it. Launch the `OneR` algorithm on the dataset.

What is the accuracy of the algorithm ? Which attribute did it select ? Was it the same than the one you selected before, using the visualization ?

There is 7 types of classifiers. You can access them by developing the corresponding branch in the selection tool.

1. `bayes` includes naive Bayes and the Bayesian networks.
2. `functions` includes the neural networks and linear regressions.
3. `lazy` includes IB1 (the nearest neighbor) and IBk (the k-nearest neighbours).
4. `meta` includes algorithms which run several iterations on the data, like Boosting (AdaBoost).
5. `misc` includes some exotic algorithms ;-)
6. `trees` includes decision trees algorithms. The famous C4.5 is included into another name (J48).
7. `rules` includes rules-based algorithms, such as `ZeroR` or `OneR`.

Try at least one algorithm of each category, and for each one, explore all the parameters they propose in order to select the good ones. Try naïve Bayes and J48.

Try the different test mode, that is : (1) Use the learning set as a test set. (2) Use a given set. (3) Use cross-validation. (4) Use a cut of the learning set.

1. Explain which can of problems could happen if you use the learning set as a test set in order to compute the accuracy of the classifier.
2. Explain why it could be better to use a cut of the learning set, than the cross-validation.
3. What is the meaning of doing cross-validation in which the number of `Folds` is equal to the number of instances ?

Once a classification is done, look in the `Result list` area. Using the right button, a list of options appear. Look for instance `Visualize classifier errors` which can be used to visualize the incorrect classified instances. You can also click on the `More options...` button. A window will open in which you can select the option `Output predictions`. Try it.

Try to see if the instances that are badly classified are the same or not, depending on the type of classifier. For instance, try to compare the badly classified instances for Naive Bayes and for J48. It could be better to use the `Percentage Split` option to perform this check.

6 Advanced experimentation with filters

Now, you will do your own experimentations. First of all, plan a protocol that will allow you to prove that a filter is useful for a given classifier. A protocol is a set of command you will have to do such as : cut the dataset in X parts, apply this classifier, etc.

Then, evaluate the different filters for the J48 classifier, using your protocol.

1. Which kind of filters are the most useful for J48 ? Why ?
2. What are the filters for which you will never launch a test ? Why ?
3. Which are the most efficient filters ?
4. The results you obtained, did you expect them ?

7 Attribute selection

Suppose that you will work with a dataset containing thousands of attributes. It is clear that two problems could arise. The first one is the computation time : some algorithms needs to process the values of all the attributes in order to take a decision. If your algorithm has a complexity of $O(n^2)$ or more, it could take a lot of time to complete ! The second problem is the quality of these attributes. This is what we will explore now.

A selection process is in fact really important. So, what could be done ?

Go into `Select attributes` and choose the attribute evaluator `CfsSubsetEval` and as a search method, `BestFirst`. Click on `Start`.

1. Interpret the content of the `Attribute selection output` pane. This result is it consistent with the one you choose during the visualization of the data ?
2. Learn the dataset using only these selected attributes. The quality of the learning is better or worst ? Look on the Web, how the C4.5 algorithm works. When you have some basic knowledge about this algorithm, try to explain the difference of accuracy.
3. Explain how the attribute selection can improve the accuracy of the learning.

8 Do your own investigation : looking for the best model

Given a dataset, several learning algorithms can have an accuracy sometimes really high. However, exposed to new data, the models generated can perform really poorly. In this exercise, you will have to make the best choice for the model.

First of all, download the `learn.arff` file in the website `http ://aquirin.ovh.org/datamining/`. You will have to learn the best possible model, provided the fact that it will be tested on a dataset you will not have immediately.

- Using the `Preprocess` window, you can have a first view of the data. These data are they easy to learn ? Why ?
- Now you will have to manipulate Weka, in order to select the best combination of filters and models to classify the data. The best way to proceed is the following. First, start by selecting some models by modifying slightly some parameters (use non aggressive filters, use the default parameters or slightly change their values). Then, compare these models between them. Check which can of modifications are the most useful and continue in this way. Notice that this way of doing could be called a meta-learning algorithm...
- **Note :** Save all the models you used. At the end, you will have to select 2 models which, according to you, have obtained the best accuracy. Then, and only then, you will have to test them on the test data (`test.arff`). Check if the models you have selected are really better than the ones you have discarded.
- **Note :** Notice that a classification rate slightly better is not significant. The choice you will accomplish in the selection of your models should be justified. Some measures seem to have a better quality that others.
- **Note :** Try also to use the visualization tool. Human intuition could be sometimes better than any algorithms (when the number of data to deal with is not so large) !
- Make a summary of the tests you have done so far. At which time you made some choices, and why ? Justify the way you have follow. For instance, you do not have to search randomly for the best model, but the outcome of a search protocol that you had clearly in mind...

The goal of this exercise was to become familiar with the behavior of the learning algorithms, the pre-processing filters, and the visualization. Please, do not loose time in small optimizations or small improvements if you think they will not let you use the full functionality of Weka.

Good luck !