1 Subject

Importing a data file in a Weka (.arff) file format into Sipina. Partitioning the dataset into train and test samples. Learning and evaluating a classification tree.

WEKA[1] is a very popular Data Mining tool. It supplies a very large of machine learning methods. WEKA can handle various files. But it has a native format (.ARFF) which is a text file with additional specifications. The text file format is very simple and very easy to manipulate. But, on the other hand, the processing of this kind of file is often slow, slower than binary file format. When we deal with a moderate size file, the text file is enough efficient. The differences between the time processing are not discernible.

There are 3 parts in the ARFF format. The upper part corresponds to eventual comments intended to describe the dataset. Each comment line must begin with the character "%".

```
%1. Title: Johns Hopkins University Ionosphere database
%
%2. Source Information:
%   -- Donor: Vince Sigillito (vgs@aplcen.apl.jhu.edu)
%   -- Date: 1989
%   -- Source: Space Physics Group
%              Applied Physics Laboratory
%              Johns Hopkins University
%              Johns Hopkins Road
%              Laurel, MD 20723
%
%3. Past Usage:
%   -- Sigillito, V. G., Wing, S. P., Hutton, L. V., \& Baker, K. B.
%      Classification of radar returns from the ionosphere using neural
%      networks. Johns Hopkins APL Technical Digest, 10, 262-266.
%...
```

The intermediate part begins with the "@relation" term. It corresponds to the data dictionary. In a supervised learning task, the last attribute is often the class attribute.

```
@relation ionosphere
@attribute a01 real
@attribute a02 real
@attribute a03 real
@attribute a04 real
@attribute a05 real
@attribute a06 real
@attribute a07 real
@attribute a08 real
@attribute a09 real
@attribute a10 real
```

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

Last, the third part of the file begins with the "@data" keyword. It corresponds to the description of the examples. The decimal point is always "." whatever the OS configuration.

```
@data
1,0,0.99539,-0.05889,0.85243,0.02306,0.83398,-0.37708,1,0.03760,0.85243,-
0.17755,0.59755,-0.44945,0.60536,-0.38223,0.84356,-0.38542,0.58212,-
0.32192,0.56971,-0.29674,0.36946,-0.47357,0.56811,-0.51171,0.41078,-
0.46168,0.21266,-0.34090,0.42267,-0.54487,0.18641,-0.45300,g
1,0,1,-0.18829,0.93035,-0.36156,-0.10868,-0.93597,1,-0.04549,0.50874,-
0.67743,0.34432,-0.69707,-0.51685,-0.97515,0.05499,-0.62237,0.33109,-1,-
0.13151,-0.45300,-0.18056,-0.35734,-0.20332,-0.26569,-0.20468,-0.18401,-
0.19040,-0.11593,-0.16626,-0.06288,-0.13738,-0.02447,b
…
```

Like the other file formats, ARFF has advantages and drawbacks. One of its success factors is that the authors of the tool have converted a large part of UCI server data file in the ARFF format (http://archive.ics.uci.edu/ml/datasets.html). They are widely used in the benchmarking of the learning methods.

In this tutorial, we show how to load a data file in the ARFF format into SIPINA. Then, we perform a classical supervised learning analysis where we subdivide the dataset in train and test samples. We learn a decision tree (C4.5 algorithm; Quinlan, 1993) on the first part; we evaluate its classification accuracy on the second part.

2 Dataset

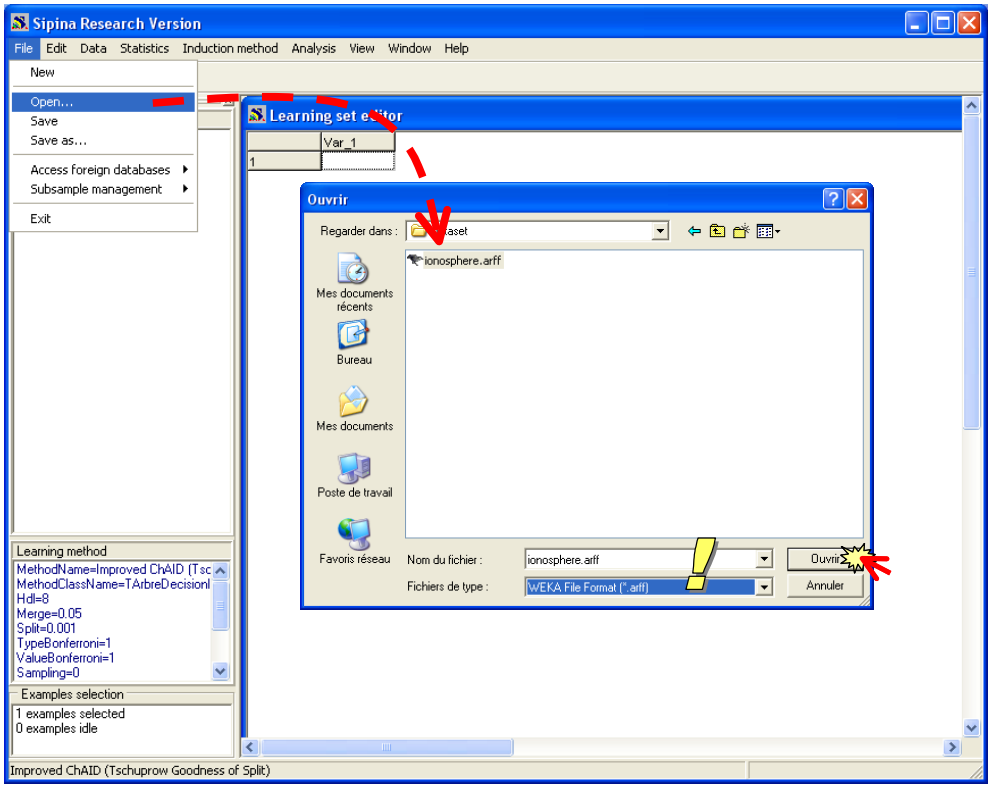
We use the IONOSPHERE.ARFF dataset (UCI server[2]). The class attribute is « CLASS » (« good » or « bad »); there are 34 continuous predictive variables.

3 Classification tree learning with SIPINA

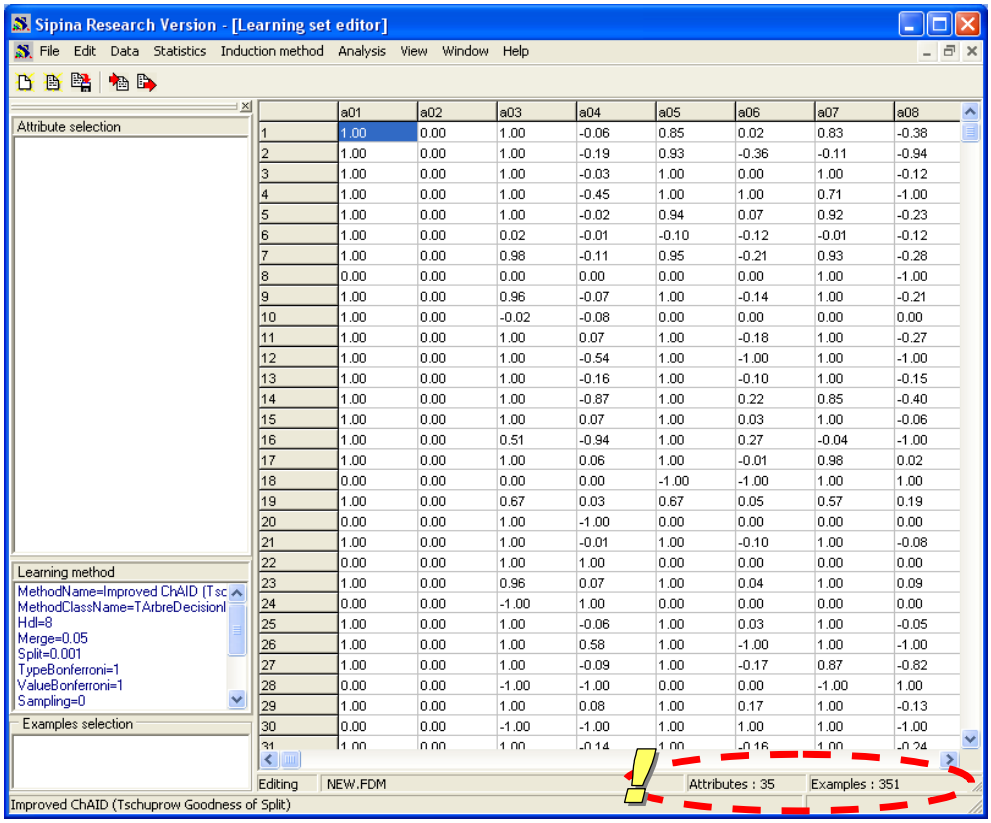
3.1 Importing the ARFF data file

After we launch SIPINA, we click on the FILE / OPEN menu. We choose the ARFF format and we select the IONOSPHERE.ARFF file.

[2] http://archive.ics.uci.edu/ml/datasets/Ionosphere. We can directly download the data file on the following URL: http://eric.univ-lyon2.fr/~ricco/dataset/ionosphere.arff

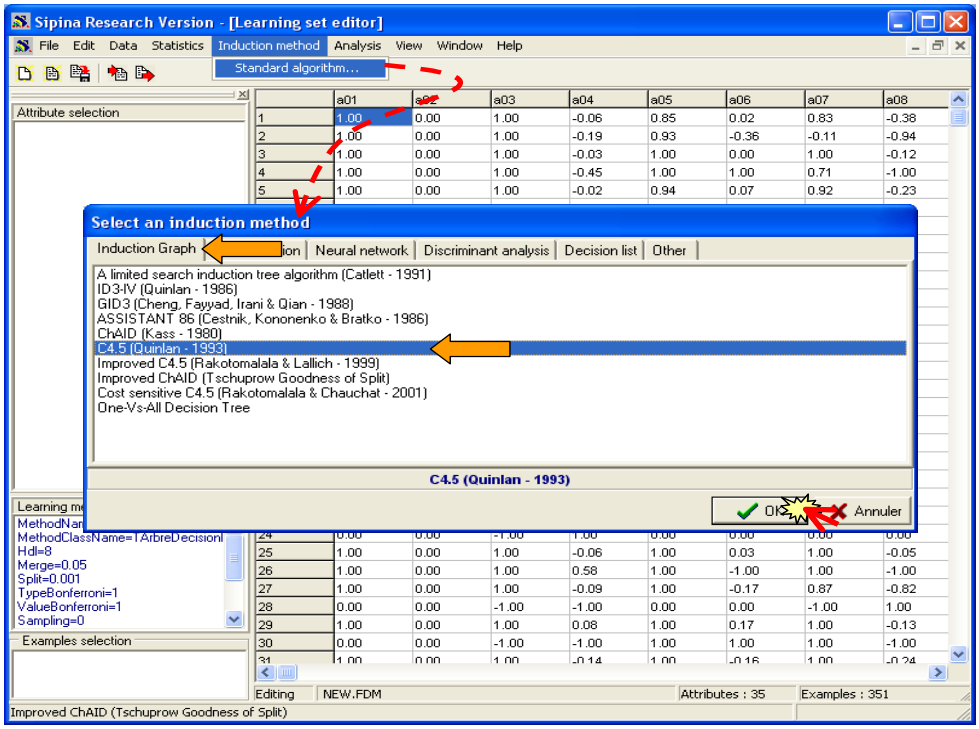


The dataset is displayed in the visualization grid. In the status bar, we see that 35 attributes and 351 examples are available for the subsequent analysis.

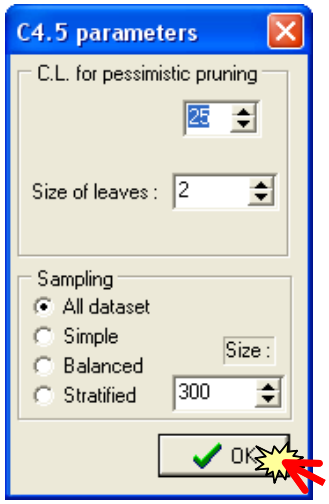


3.2 Selecting the learning method

We want to implement the C4.5 algorithm (Quinlan, 1993). In order to select the method, we click on the INDUCTION METHOD / STANDARD ALGORITHM menu. In the dialog box, we select C4.5 then we click on the OK button.

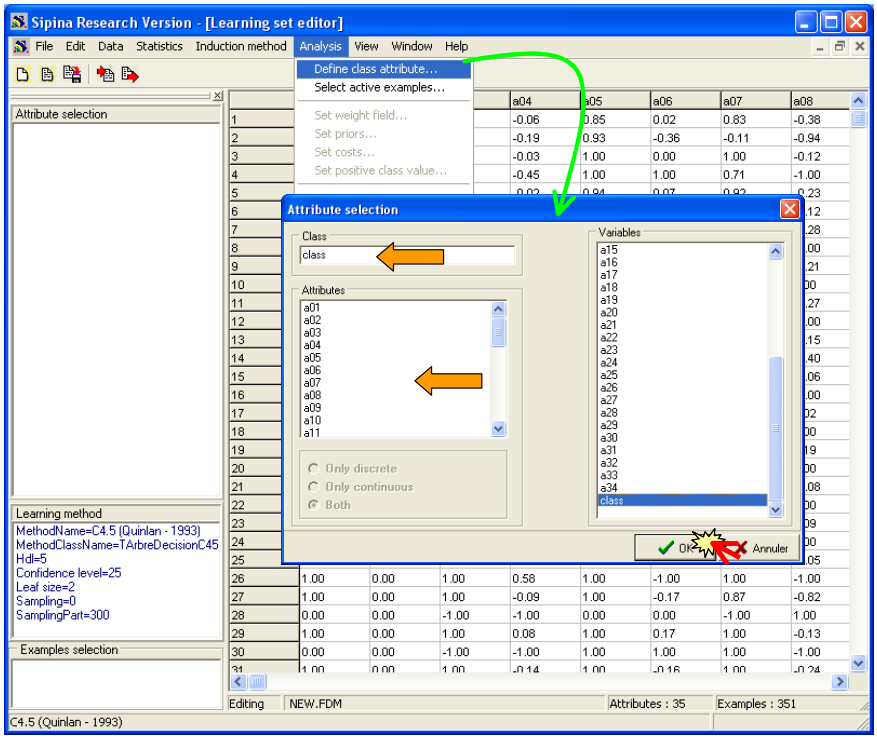


A second dialog box appears. We left the default settings.

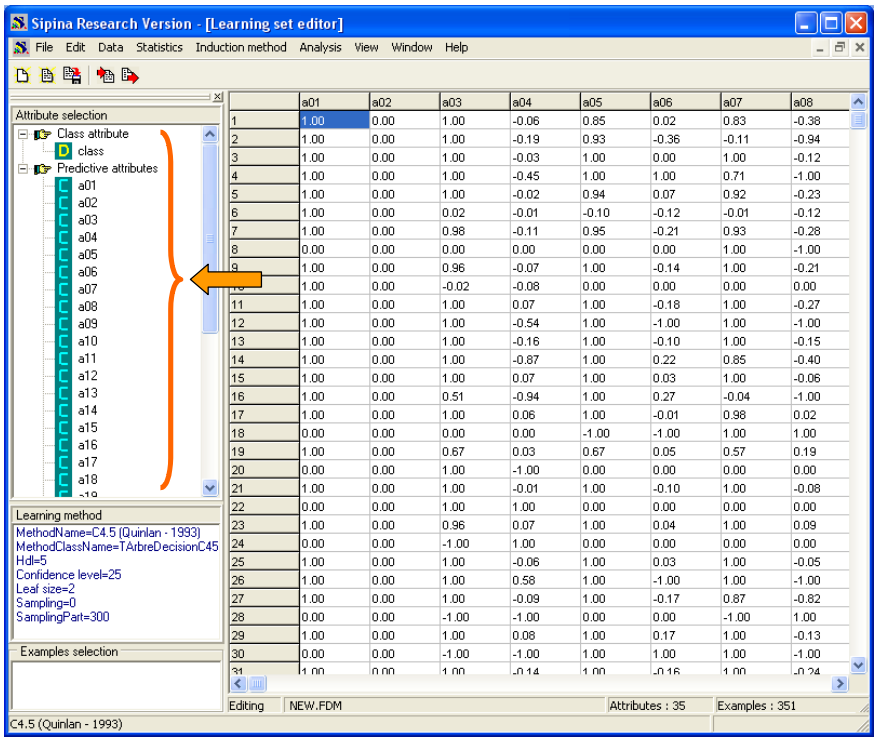


3.3 Specifying the types of the variables

We want to predict CLASS from the other descriptors. We click on the ANALYSIS / DEFINE CLASS ATTRIBUTE. We set CLASS as target attribute; the others (a01 to a34) as ATTRIBUTES (INPUT).



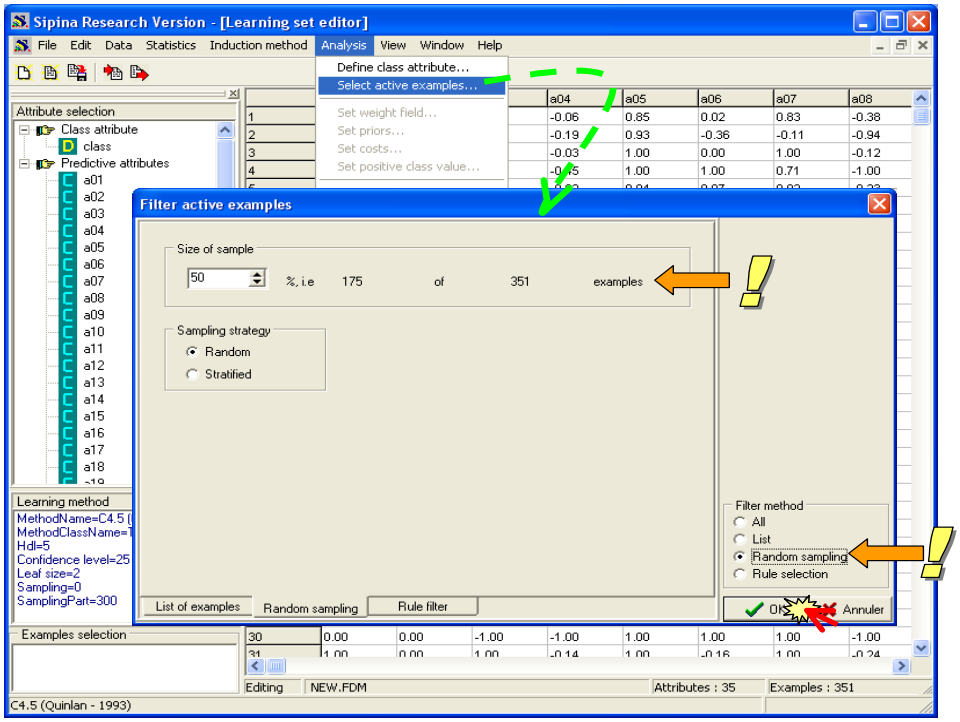
The statuses of the variables are displayed on the left part of the main window.



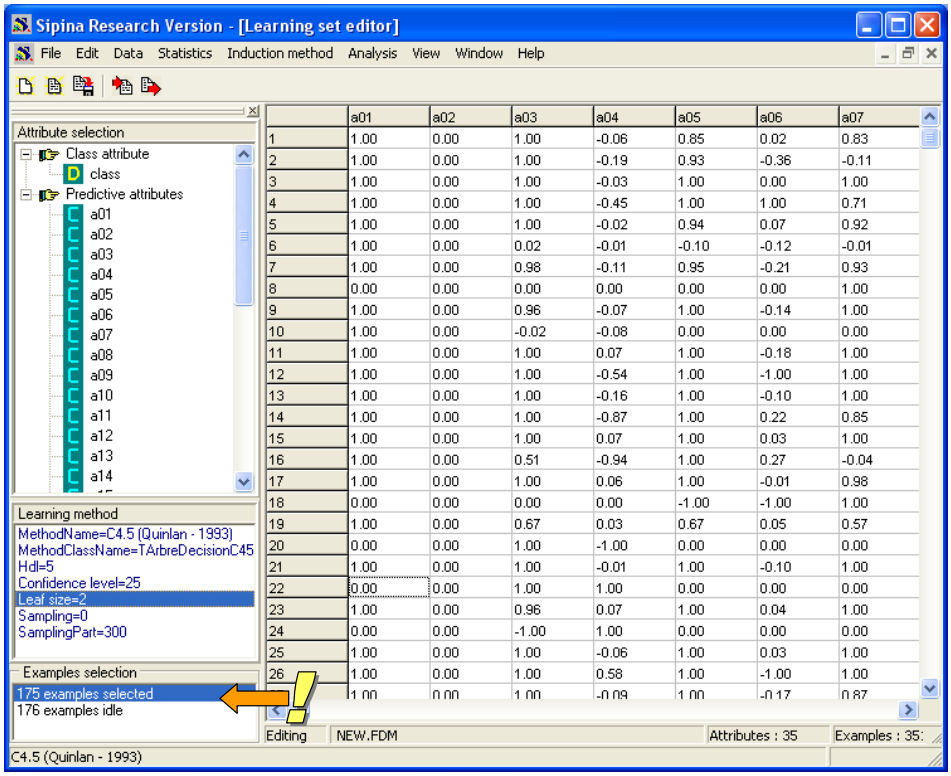
3.4 Defining the train and the test samples

Last step before the learning process, we want to subdivide the dataset into train and test samples. We learn the tree on the first part; we evaluate its accuracy on the second part.

We click on the ANALYSIS / SELECT ACTIVE EXAMPLES menu. We select the RANDOM SAMPLING: 50% of the instances are used for the learning process (175); the remainder is used for the assessment (351 – 175 = 176).

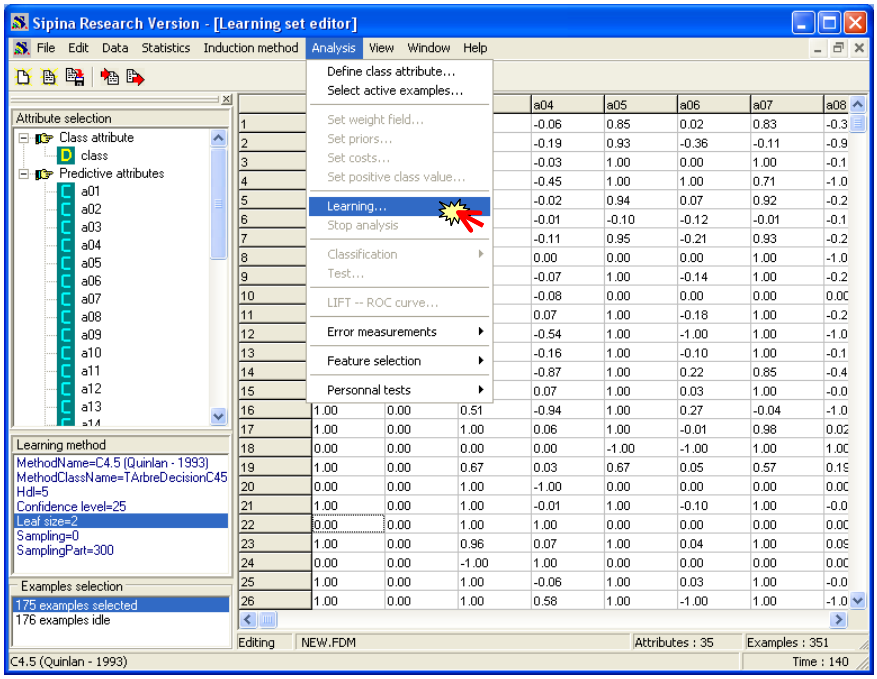


The samples are described in the low part of the main window.

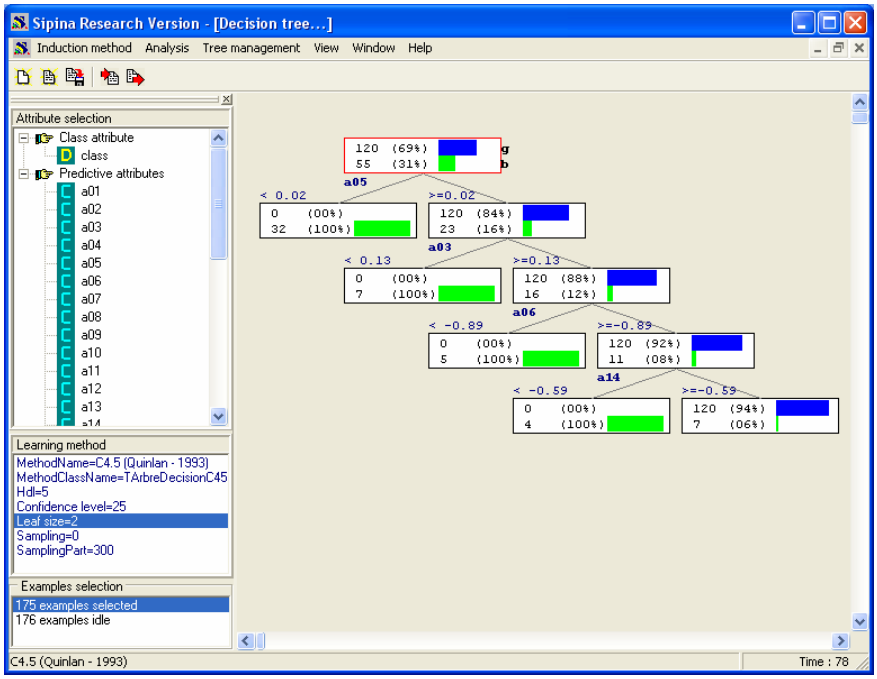


3.5 Learning the classification tree

We perform the analysis by clicking on the ANALYSIS / LEARNING menu.



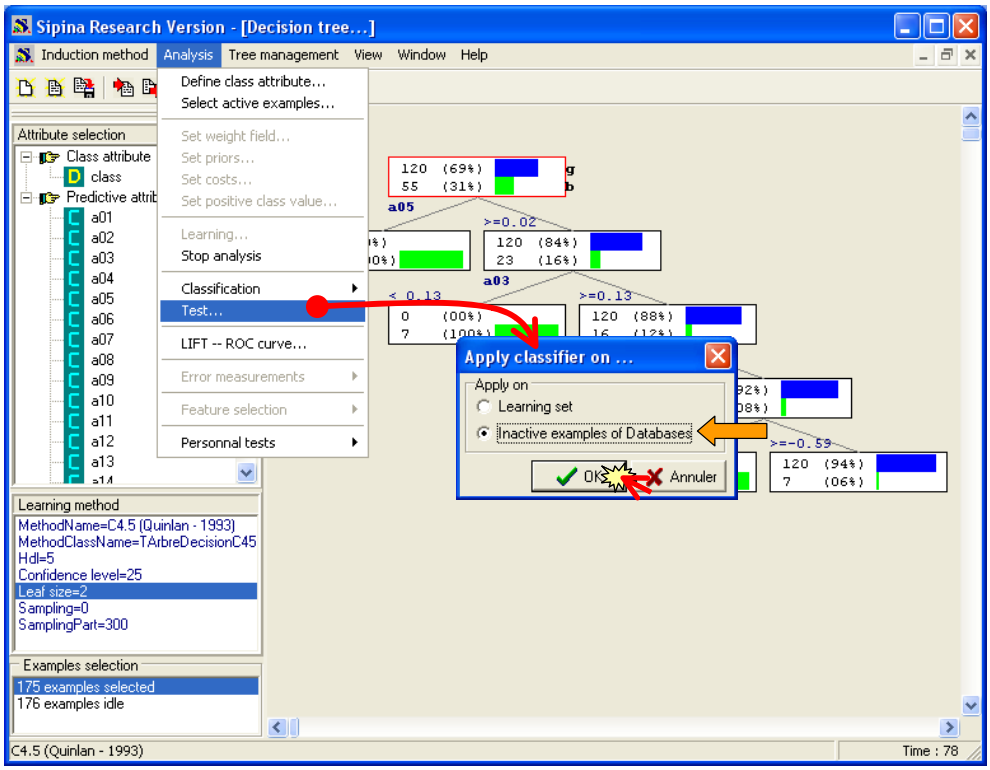
We obtain the tree. On each node, we observe the histogram of the class attribute.



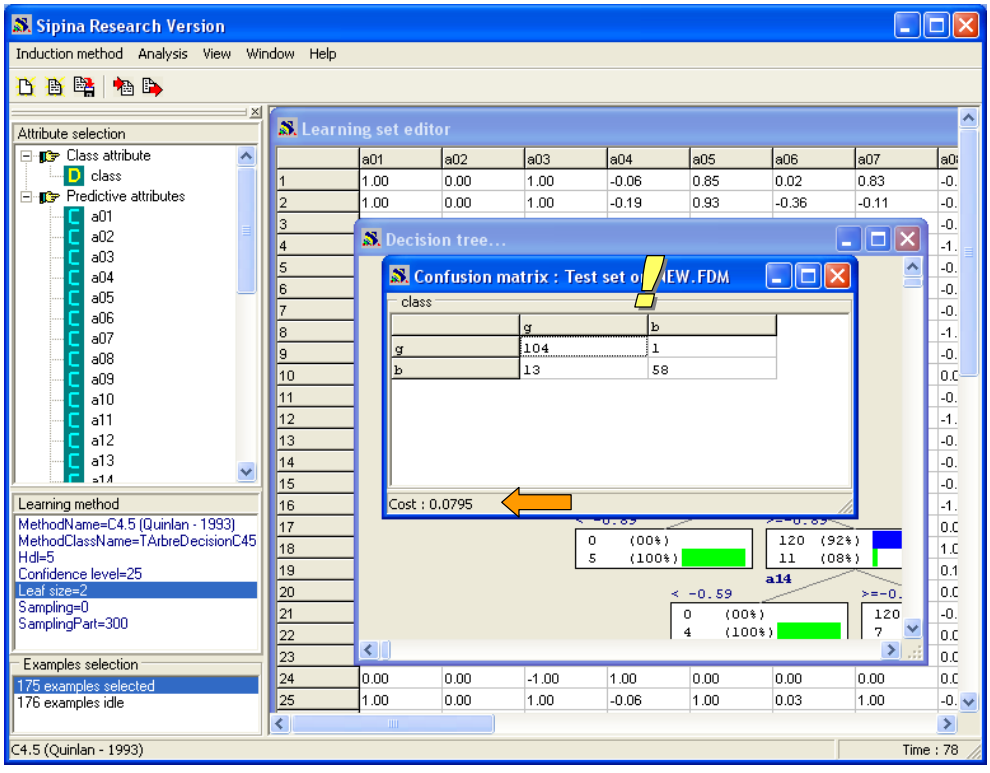
3.6 Evaluating the accuracy of the classification tree

Measuring the error rate is a crucial step. We want to estimate the probability of misclassification of the model when we apply it on an instance of the population.

In order to obtain an unbiased estimation, we use the test set. We click on the ANALYSIS / TEST. We select the inactive examples; those which are not used during the learning process.



The confusion matrix is displayed in a new window. We have also the error rate (7.95%).



4 Conclusion

In this tutorial, we show how to handle the ARFF file format into SIPINA. Because Weka is widely diffused, this potential uses is essential. The ability to handle the Excel spreadsheet file format is also an important functionality. Sipina can make the connexion with the Excel using an add-in (see http://eric.univ-lyon2.fr/~ricco/doc/sipina_xla_installation.htm and http://eric.univ-lyon2.fr/~ricco/doc/sipina_xla_processing.htm).