Tutorial overview

This tutorial shows how to apply a supervised learning method on your data; for this example we're going to analyze the « breast.txt » dataset, using the ID3 method (decision tree).

This well-known file, from the medical domain, consists in the characteristics of cells sampled on women presenting (or not) a breast malignant tumor.

In this tutorial, you'll learn to use the following components:

Tab	Operator (Component)	Comment
Feature selection	Define status	Specify the attributes to use
Meta-spv learning	Supervised learning	A container for machine learning operators
Spv learning	ID3	A machine learning operator

Loading the data in TANAGRA

- Opening an existing diagram

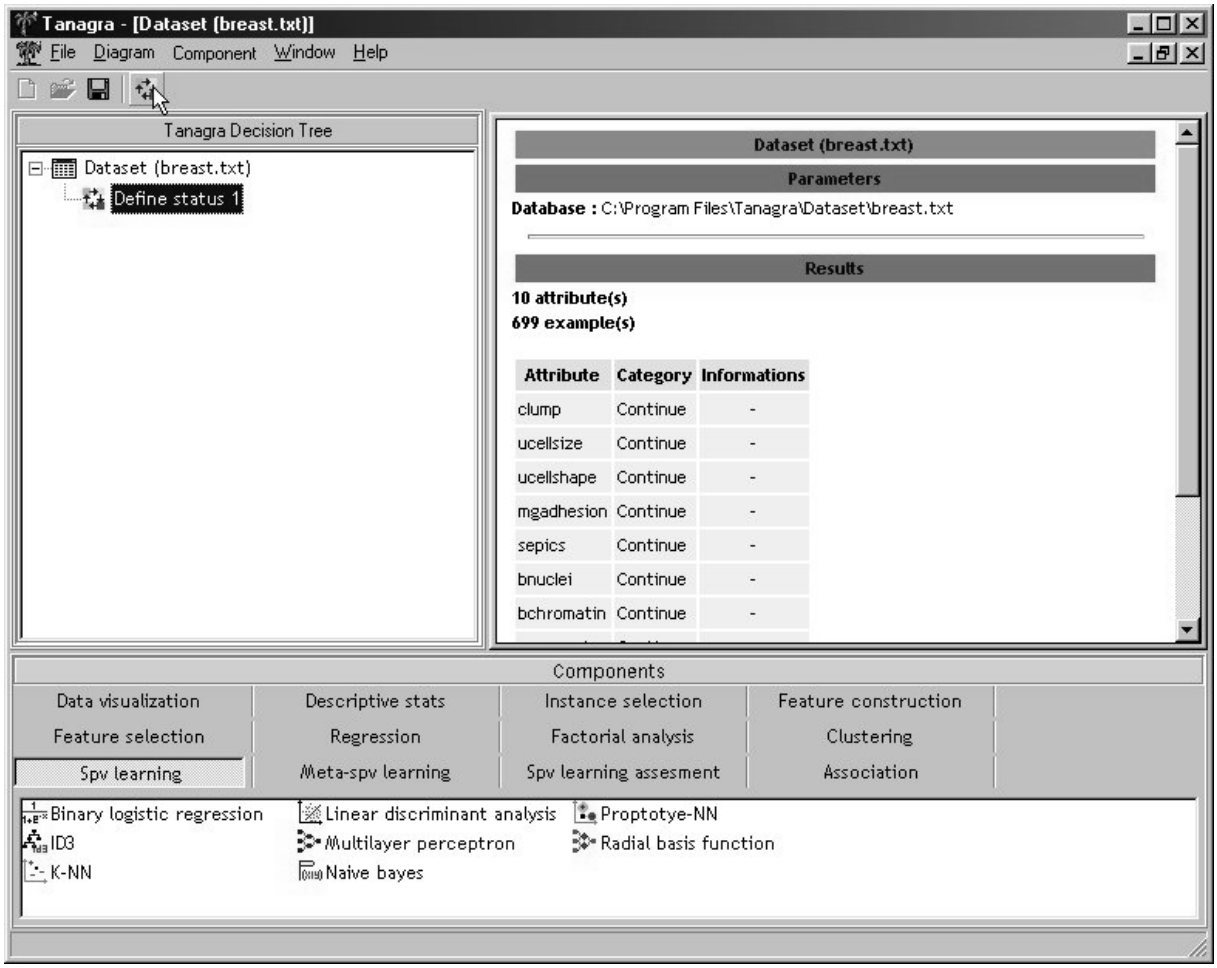
1 – Choose *File/Open…* in the main menu of TANAGRA.

2 – In the "Dataset" subdirectory of TANAGRA, select the file named « breast.bdm ».

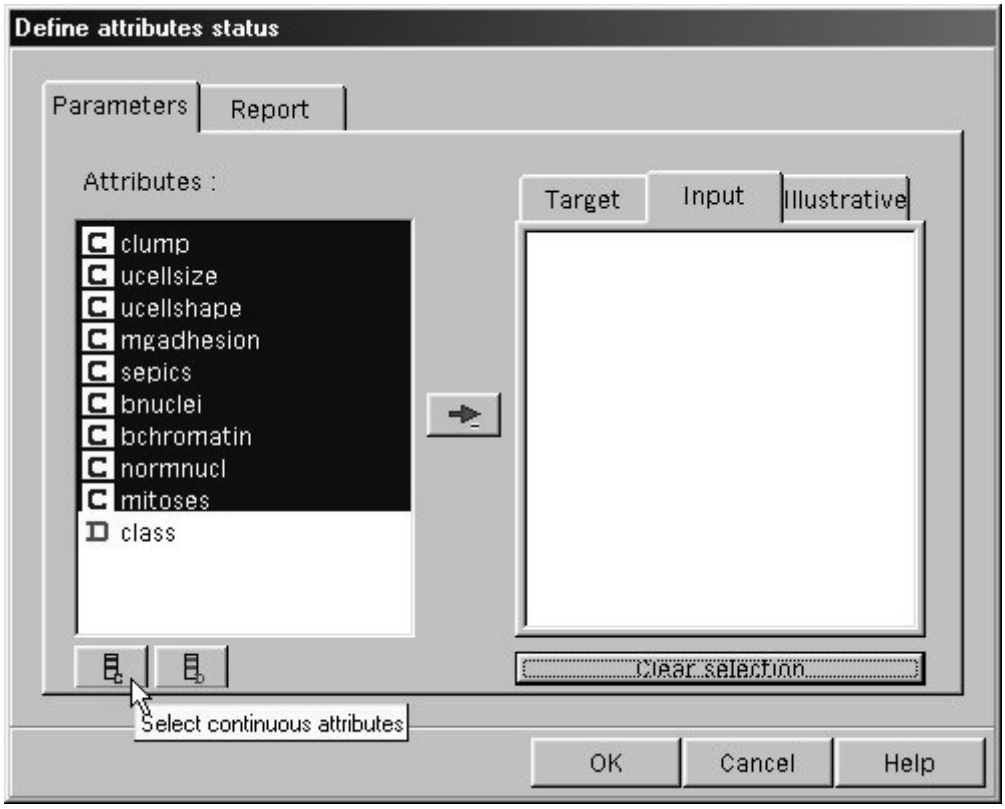
Using a supervised learning operator

- Specify the status of the attributes for the analysis

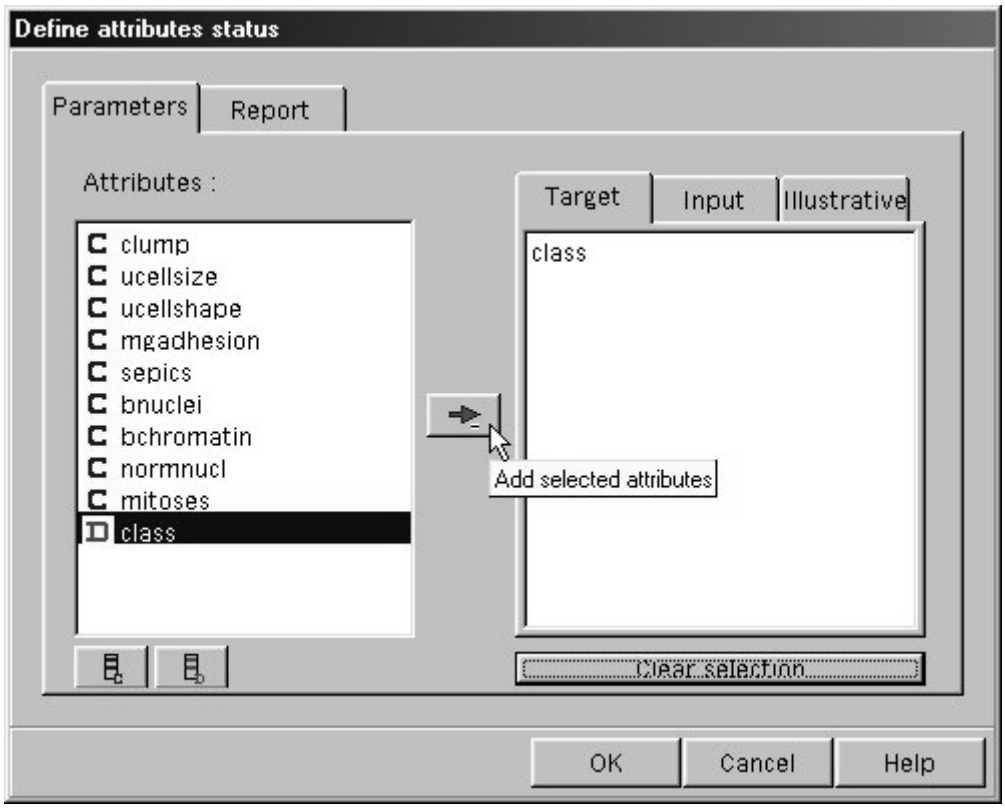
1 – Add a **Define Status** operator under the "Dataset" node, by clicking on its icon in the shortcuts toolbar (as shown below). A dialog box appears automatically, allowing the definition of the status of the attributes.



2 – Before all, be sure that the active tab in the dialog is the “Input” one. Then select the continuous attributes in the left list by clicking the corresponding button below the list (as shown in the following screenshot), and hit the arrow button to bring them in the Input list.



3 – In the same dialog box, activate the Target tab. Select the « class » attribute in the list and click the arrow button.



4 – Now you have defined the class attribute (« class » = Target), and the descriptors to do this (the others = Input). Click OK to validate and close this dialog box.

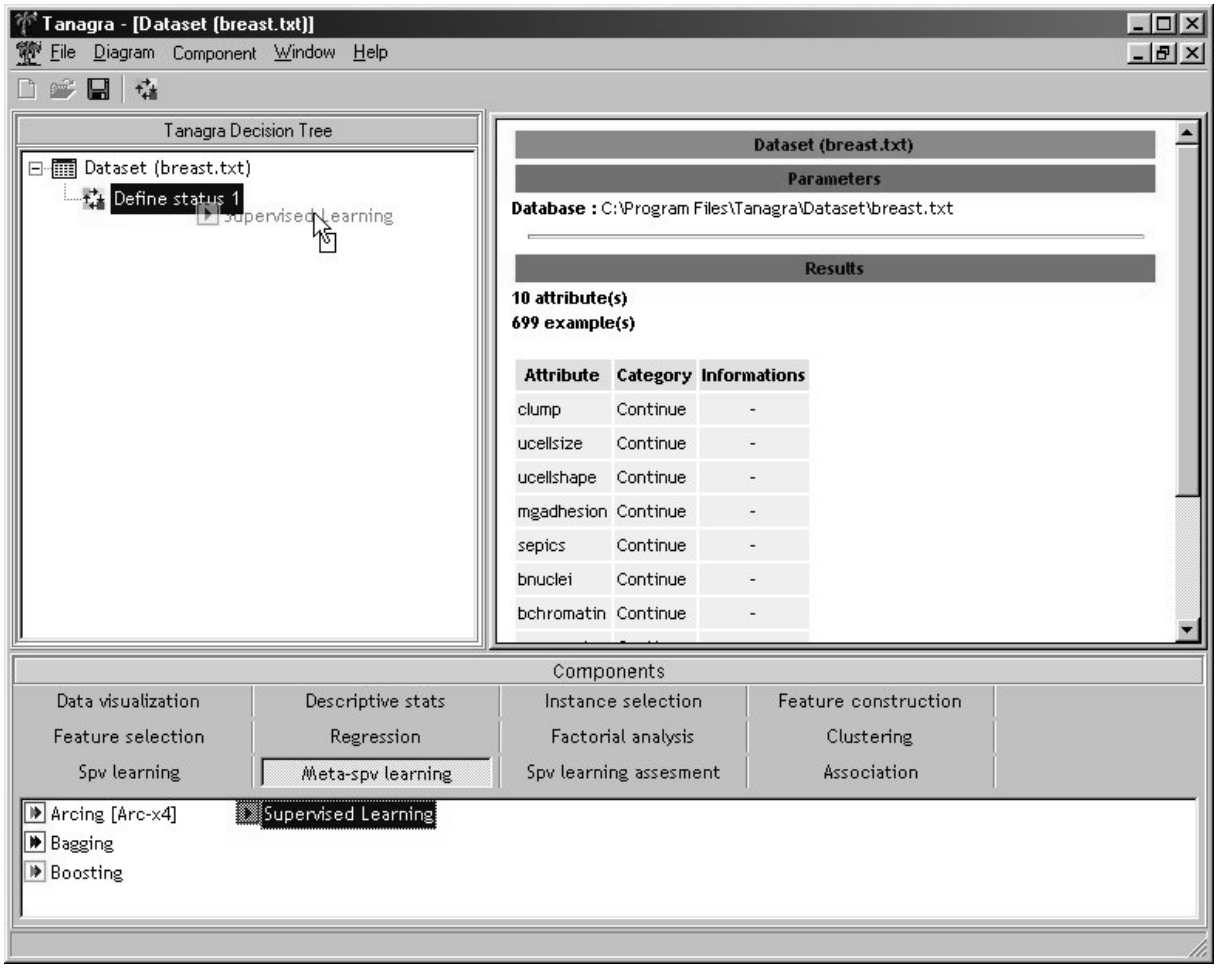
- Choosing the meta-operator

To obtain a better prediction model, you may execute a machine learning method many times on a weighted examples, and aggregate classifiers. This method, known as “boosting - arcing” can be applied to any learning algorithm.

TANAGRA implements this concept by encapsulating the machine learning operator in a meta-operator. It is the meta-operator that launches the method many times, taking some specific parameters into account.

In this quick start tutorial we won’t describe these advanced experimentations, we’ll simply launch the ID3 method once. But as TANAGRA always needs a meta-operator, a basic one exists that suits our case, launching only one time the encapsulated method: the **Supervised Learning** operator.

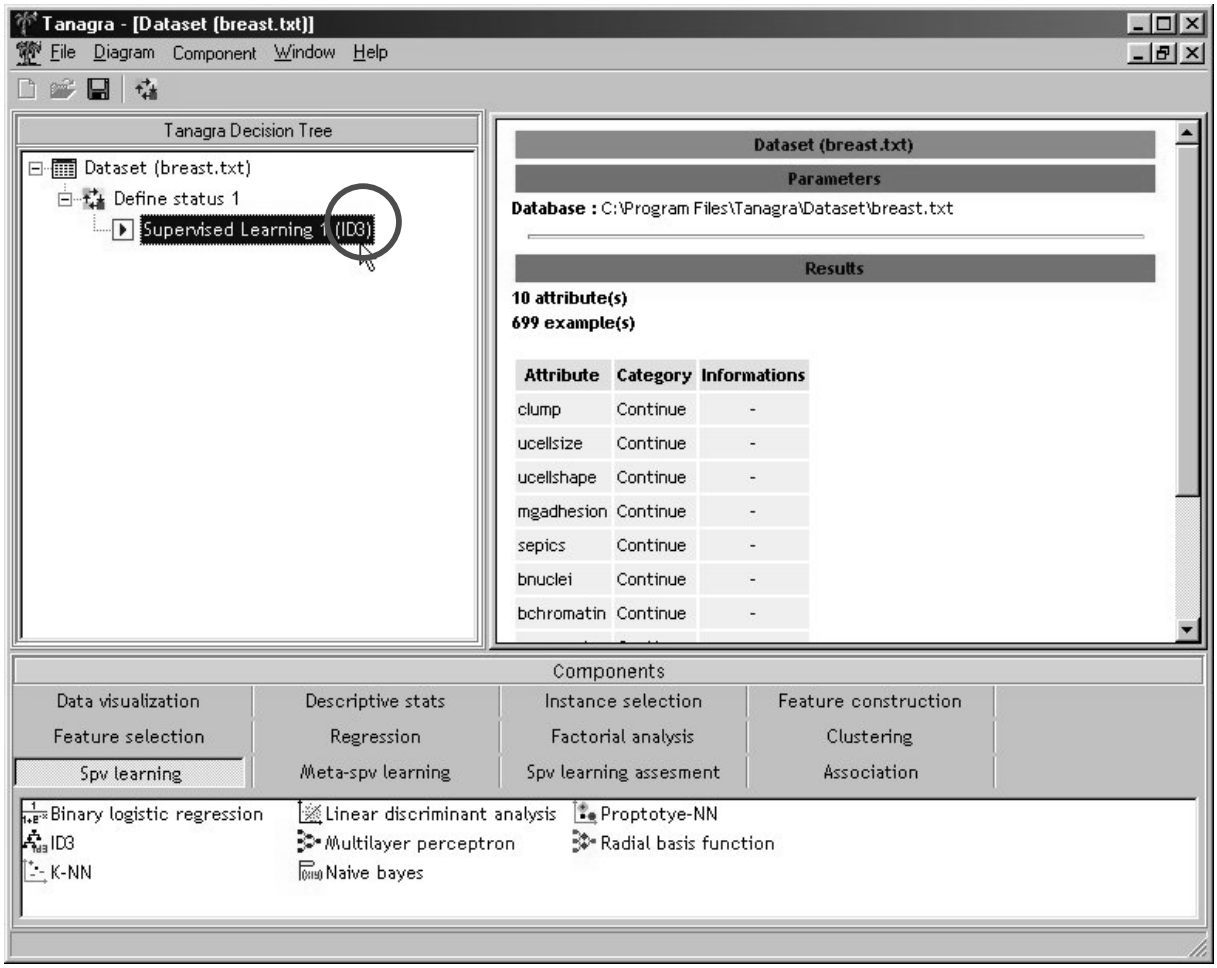
1 – Add a **Supervised learning** component (META-SPV LEARNING tab) to the diagram, under the «Define status 1» node.



- Adding a supervised learning component

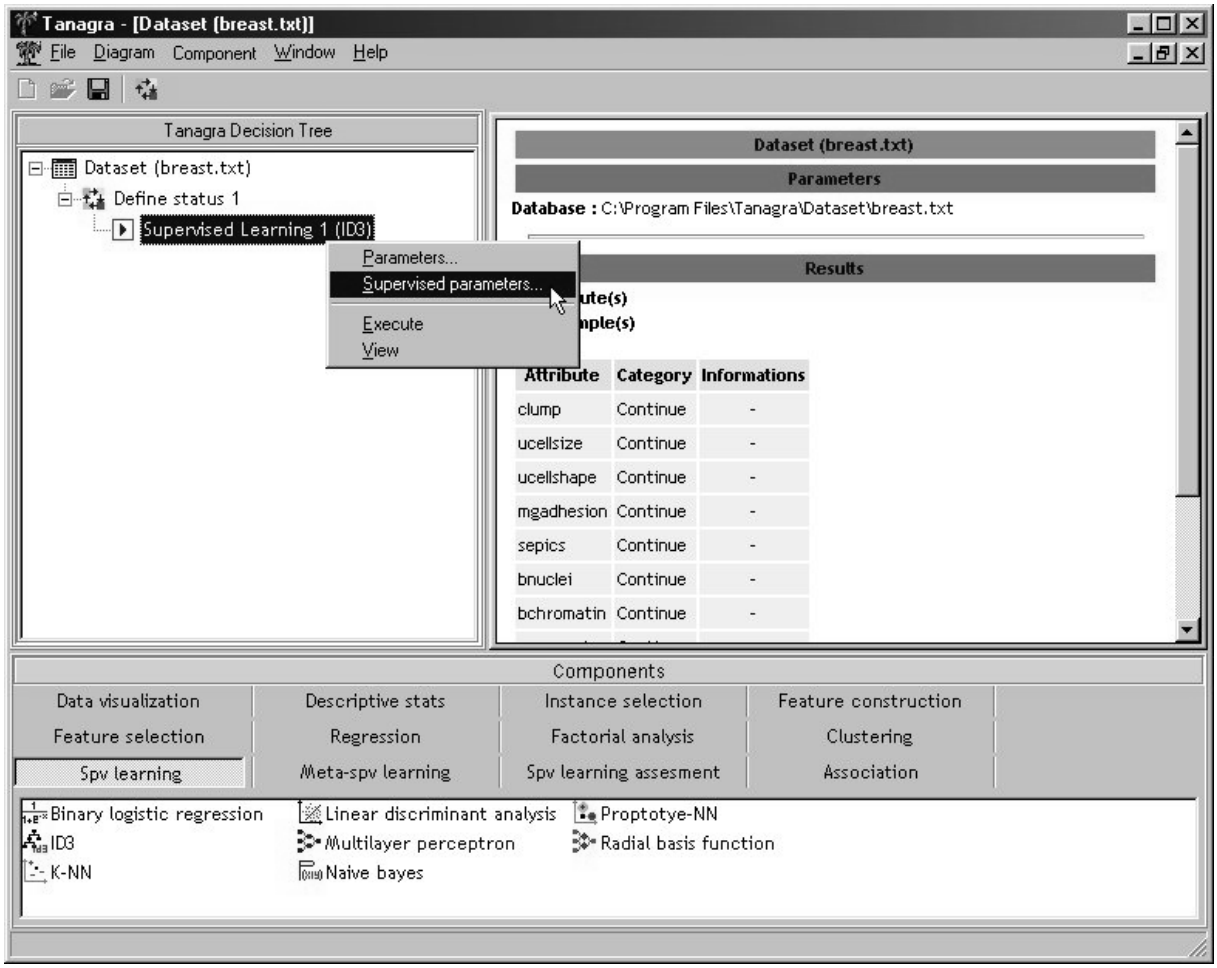
1 – In the components palette, SPV LEARNING tab, drag with the mouse a **ID3** component on the « Supervised Learning » node you've added just before.

The operator is included in the meta one, so we don't see another node under the meta node, but we notice that the name of the method appears in the meta node label.



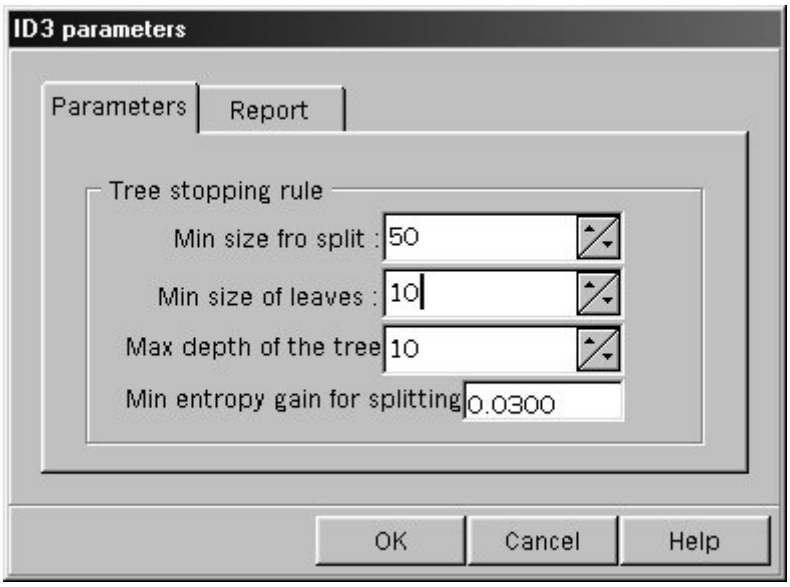
- Defining the method parameters (for ID3 operator)

1 – Activate the popup menu of the « Supervised learning (ID3) » node by right-clicking on it. There's a new command under the well-known *Parameters*: *Supervised parameters…*



The first (*Parameters…*) applies to the meta-operator, the second (*Supervised parameters…)* applies to the supervised learning operator. Choose this last one.

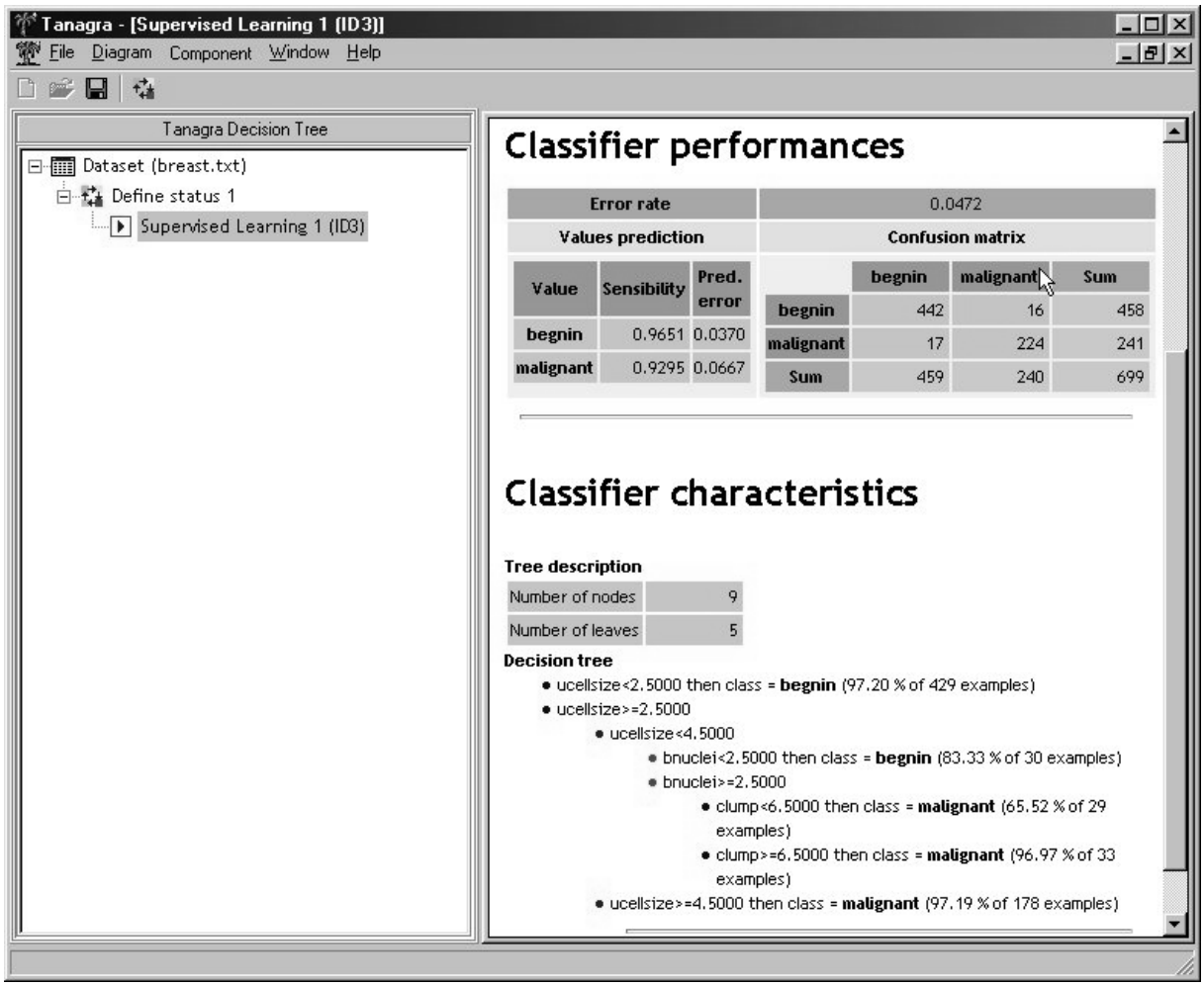
2 – In the opening dialog box, to take into account the size of the data file (699 records), we modify the ID3 parameters as follows:



3 – Click OK to validate.

Executing the learning process

1 – In the contextual menu of the meta node, choose the *View* command. Results are displayed in the right frame.



The resubstitution error rate seems to be good (4,72 %). We see in the confusion matrix that false positive and false negative examples are equivalent.

Finally we notice, looking at the results of our analyses, the importance of the “ucellsize” attribute in this problem.