Subject

With the 1.4.8 version, we can save a part of the stream diagram. The goal is to perform some succession of analysis on several files.

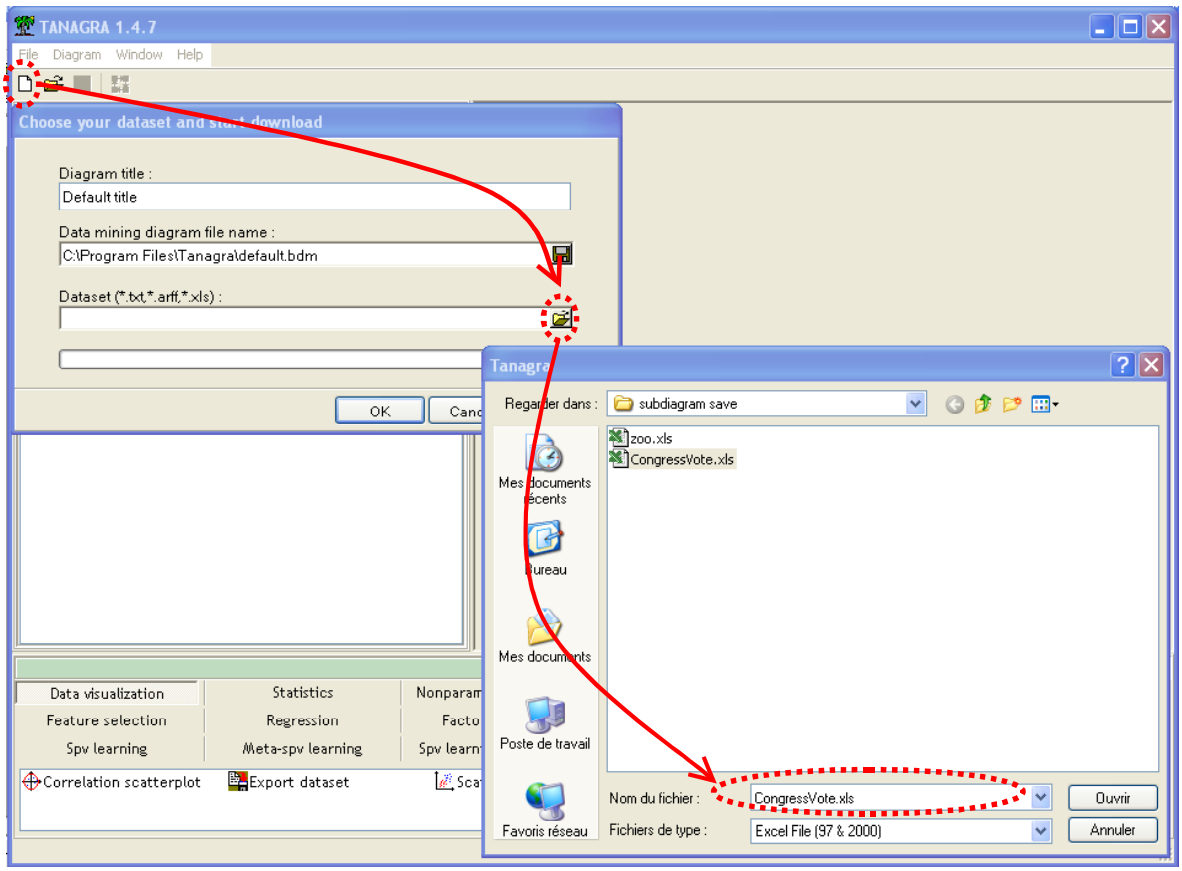
Dataset

We use CONGRESSVOTE.XLS and ZOO.XLS. We want to predict a class attribute from discrete descriptors with or without feature selection. We use the cross-validation in order to compare the error rate.

Saving a sub-diagram

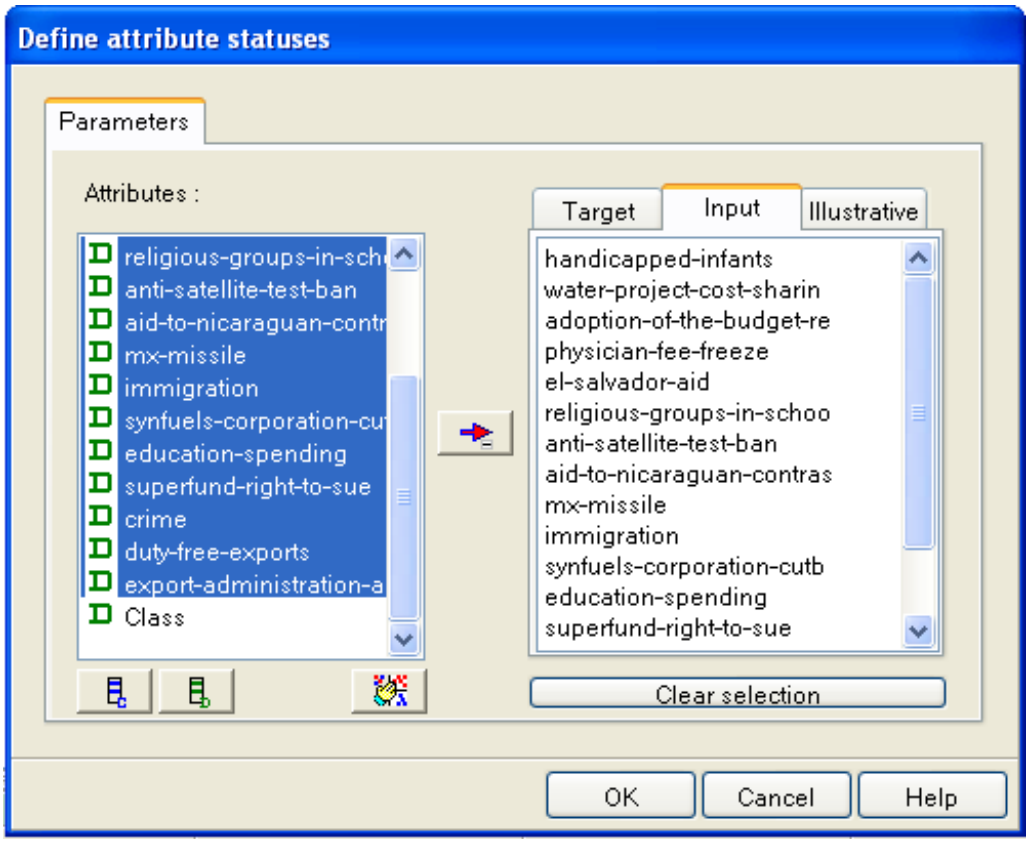
Dataset importation

We click on FILE/NEW in order to create a diagram and import the CONGRESSVOTE.XLS dataset.



Defining class and predictive attributes

We add the DEFINE STATUS component in the diagram: CLASS is the TARGET attribute; the others are INPUT attributes.

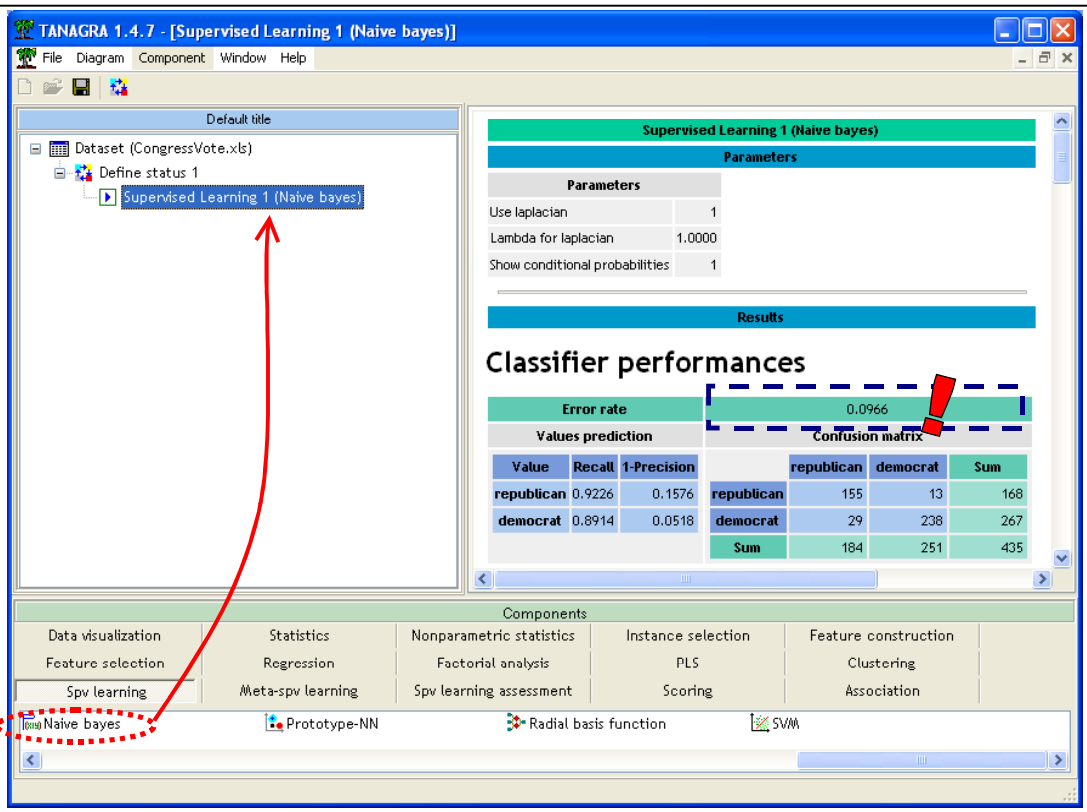


Learning algorithm and performance evaluation

From the 1.4.7 version, we can add directly a supervised learning algorithm in the diagram. TANAGRA inserts automatically the META SPV LEARNING component with implements one instance of this algorithm.

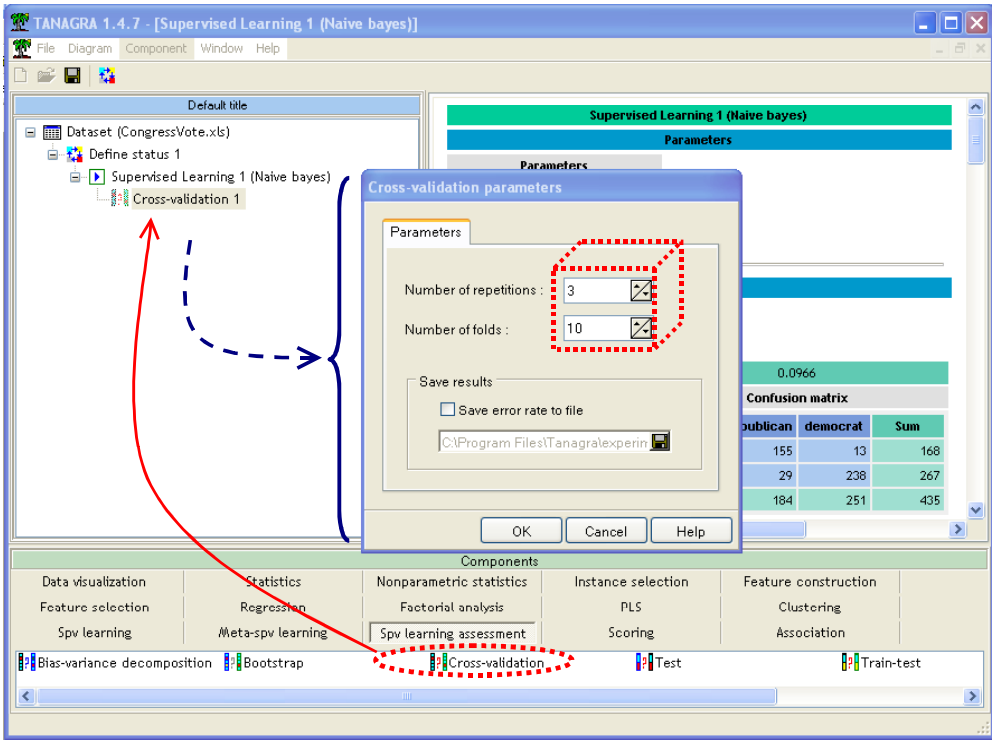
Caution: If you want use aggregation strategy such as BAGGING or BOOSTING, you must follow the old procedure: add in the first step the META SPV component (e.g. BAGGING) and embed in this component the supervised learning algorithm (e.g. NAÏVE BAYES CLASSIFIER).

We insert the NAÏVE BAYES algorithm in the diagram.



Resubstitution error rate is 9.66%.

In order to obtain an honest estimation of the "true" error rate, we add a cross-validation component in the diagram. We use a repeated (3 times) 10-cross validation.



The cross-validation error rate is 10.08%.

Cross-validation 1

Parameters

Cross-validation parameters	
Folds	10
Trials	3

Results

CV error rate

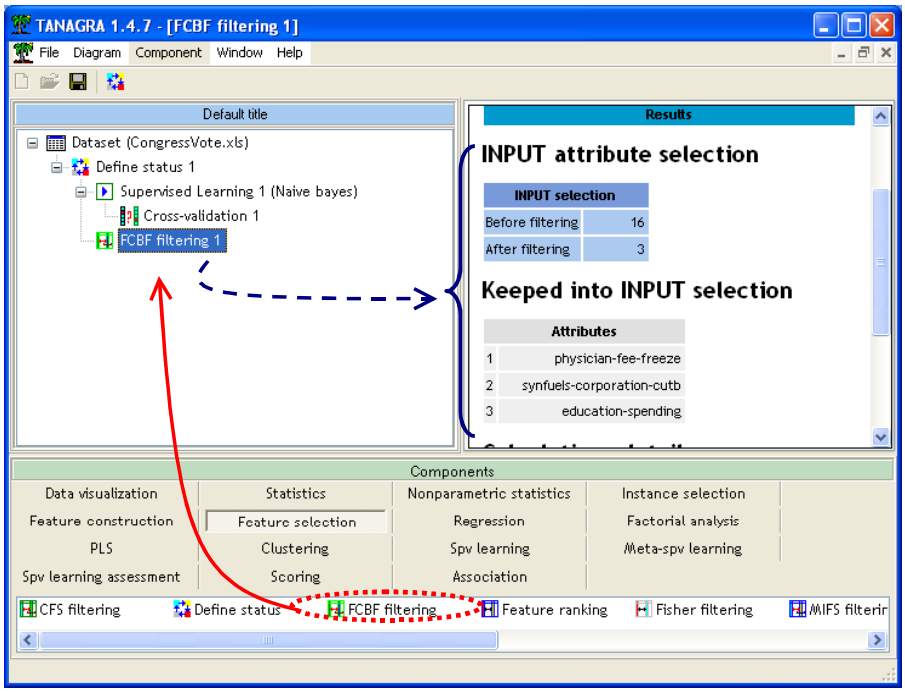
Range	
MIN	0.0977
MAX	0.1047
Trial	**Err rate**
1	0.1047
2	0.1000
3	0.0977

Overall cross-validation error rate

Error rate	0.1008
Values prediction	**Confusion matrix**

Feature selection

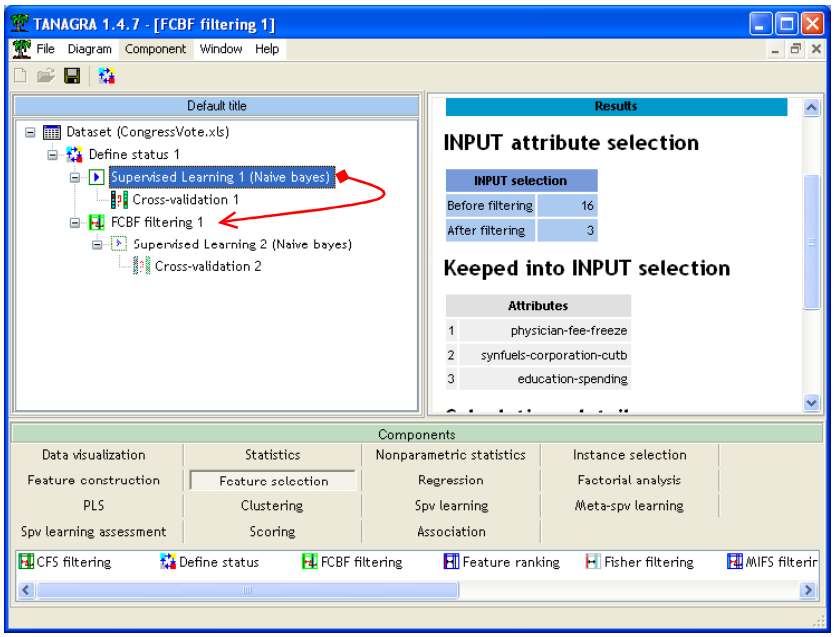
We want to perform a feature selection before the learning phase. We expect that selecting the relevant attributes improves the classifier performance. We insert in the diagram the FCBF (Liu et al.) feature selection component.



We note that only 3 descriptors among the 16 ones are selected.

In order to evaluate the efficiency of this feature selection, we add again in the diagram the naïve bayes classifier and the cross-validation error rate evaluation. Instead of adding the components manually, we can copy the corresponding sub-diagram (1.4.7 version and higher).

To do that, we select the "SUPERVISED LEARNING 1 (NAÏVE BAYES)" node in the diagram and drag this one on the "FCBF filtering 1" node.



We click on the VIEW menu of the "CROSS VALIDATION 2"; the error rate is 5.5%.

Cross-validation 2

Parameters

Cross-validation parameters	
Folds	10
Trials	3

Results

CV error rate

Range	
MIN	0.0465
MAX	0.0628

Trial	Err rate
1	0.0465
2	0.0628
3	0.0558

Overall cross-validation error rate

Error rate	0.0550

The FCBF feature selection improves significantly the naïve bayes performances on the VOTE dataset.

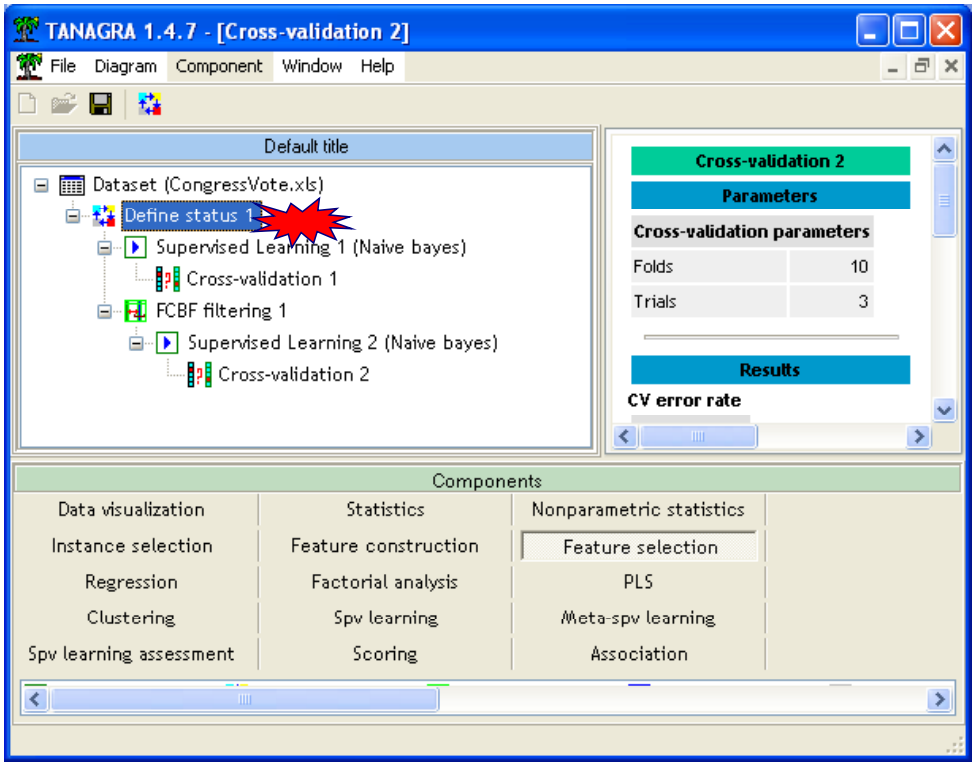
The same analysis on another dataset

We want to evaluate this framework (FCBF feature selection + naïve bayes classifier) on the ZOO dataset. To do that, we must define the same diagram on this dataset, in order to compare the performance of the classifier with or without feature selection.

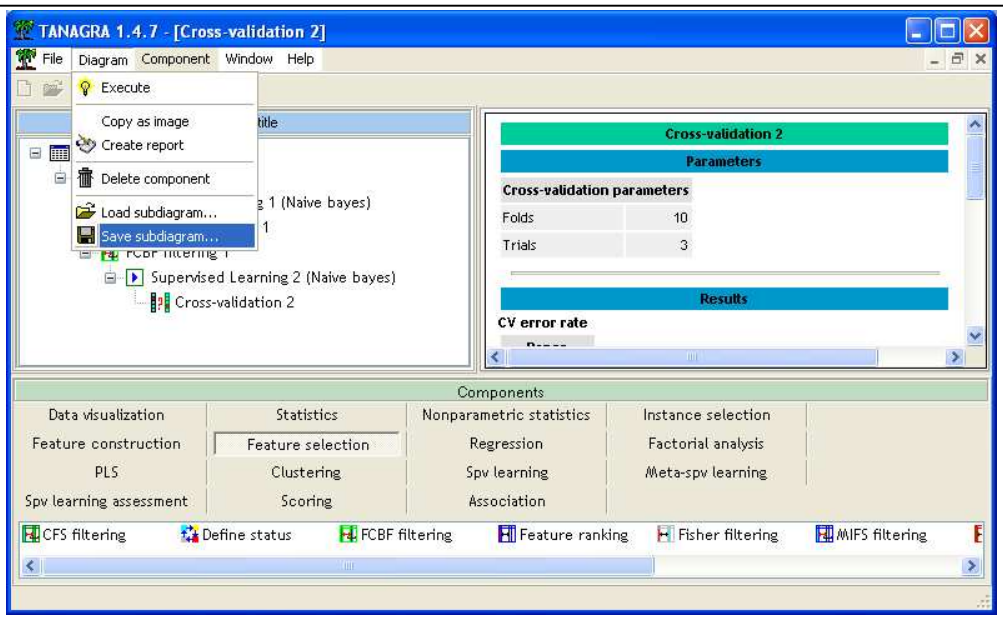
From the 1.4.8 version, we can save a part of the diagram in a file (SDM file extension) and insert this one in another diagram. We follow three steps: save the sub-diagram from the selected node; open or create a new diagram; insert the saved sub-diagram under the selected node in the new diagram.

Saving the sub-diagram

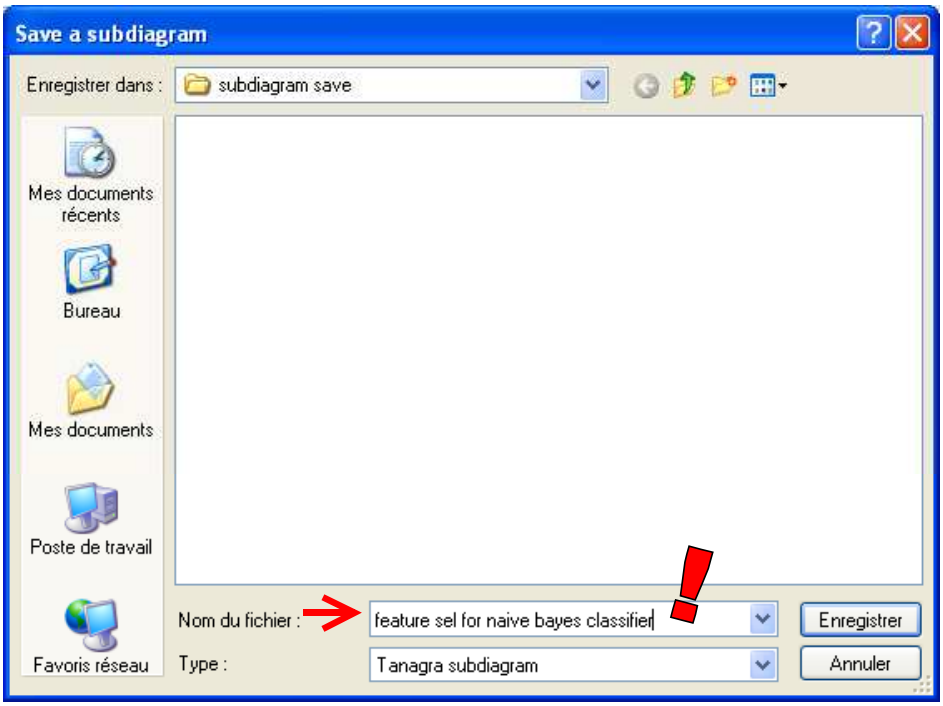
We select the "DEFINE STATUS 1".



We click on the DIAGRAM / SAVE SUBDIAGRAM item of the main menu.



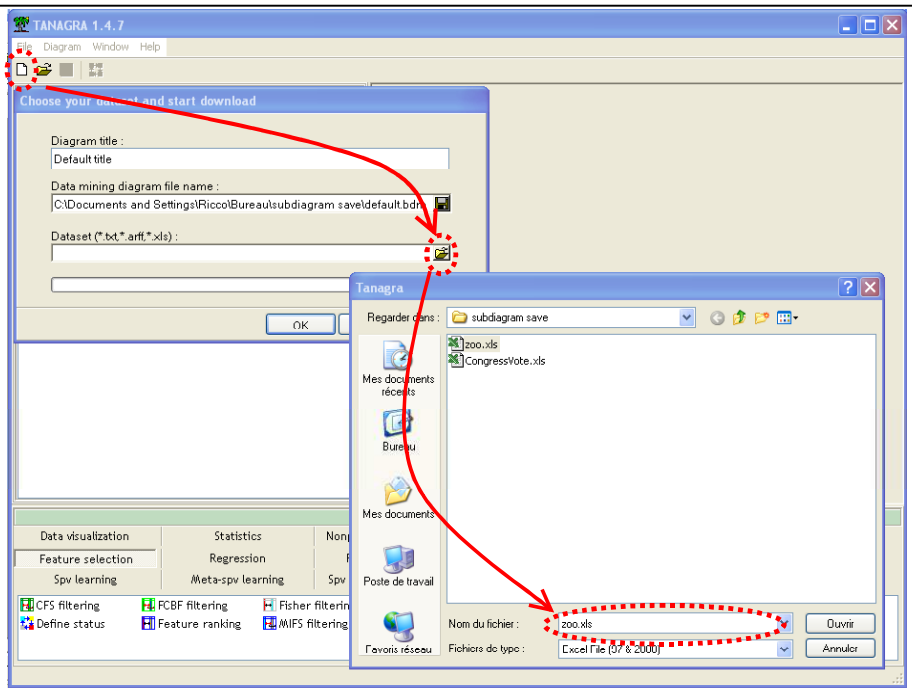
A dialog box appears, we set the sub-diagram file name.



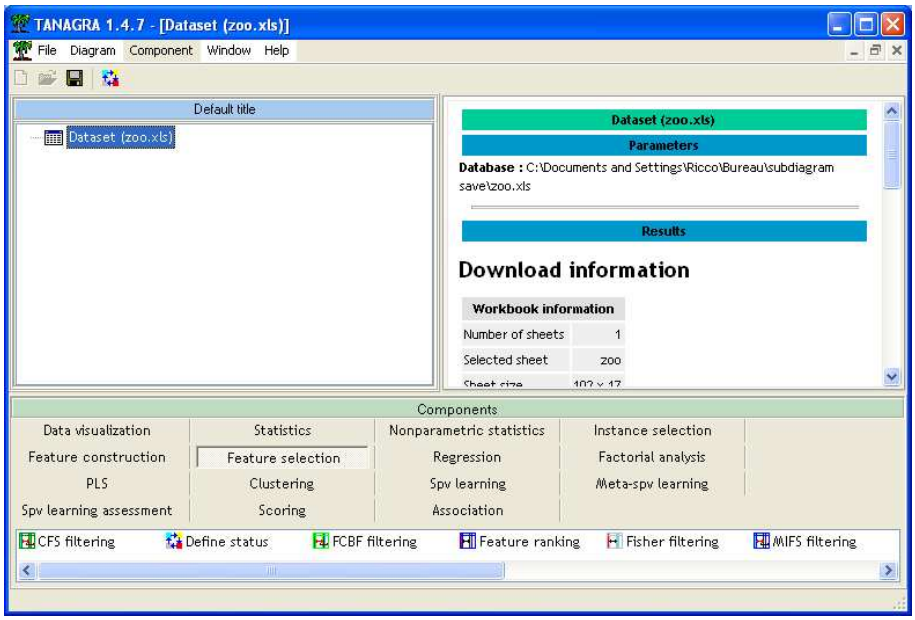
When the sub-diagram is saved, we can close the current diagram (FILE/CLOSE menu).

New diagram and data importation

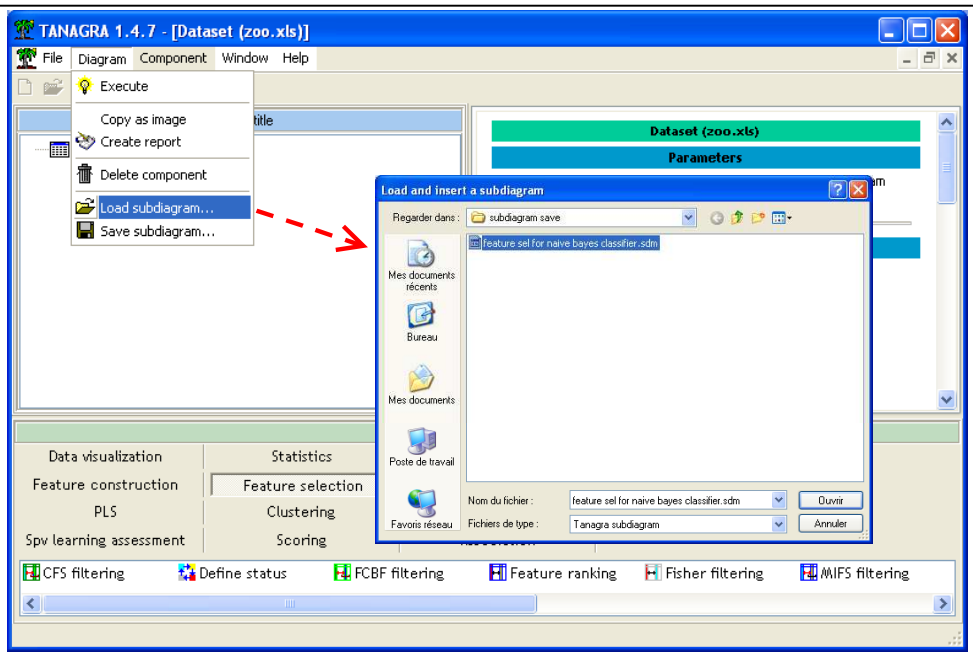
We select the FILE/NEW menu in order to create a new diagram and import the ZOO.XLS dataset.



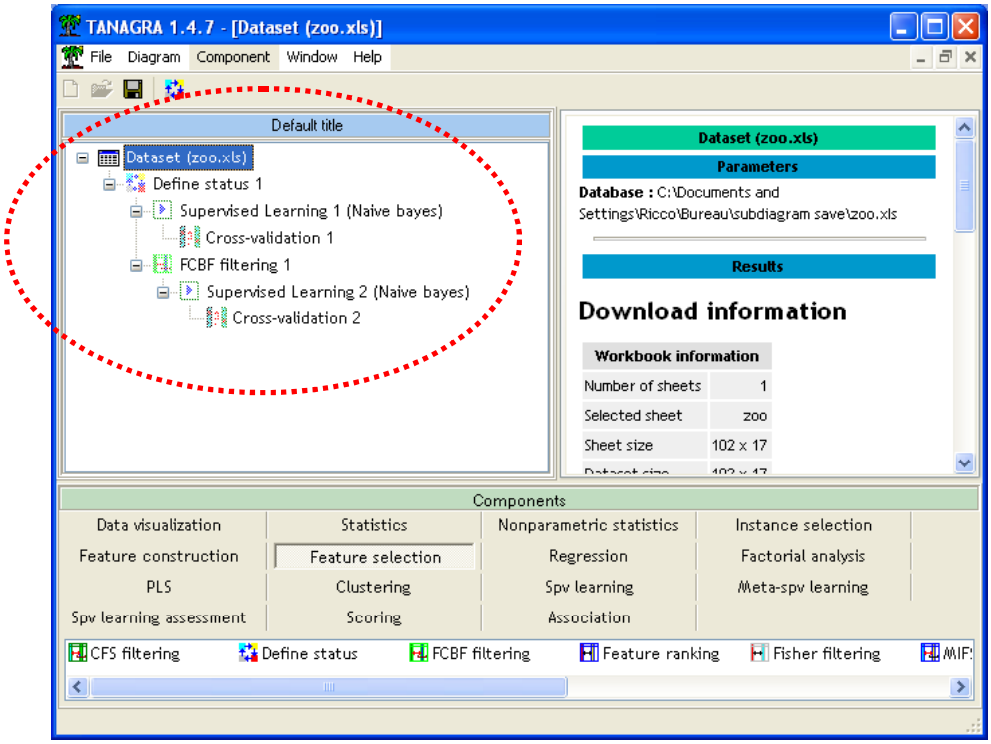
We obtain the following diagram.



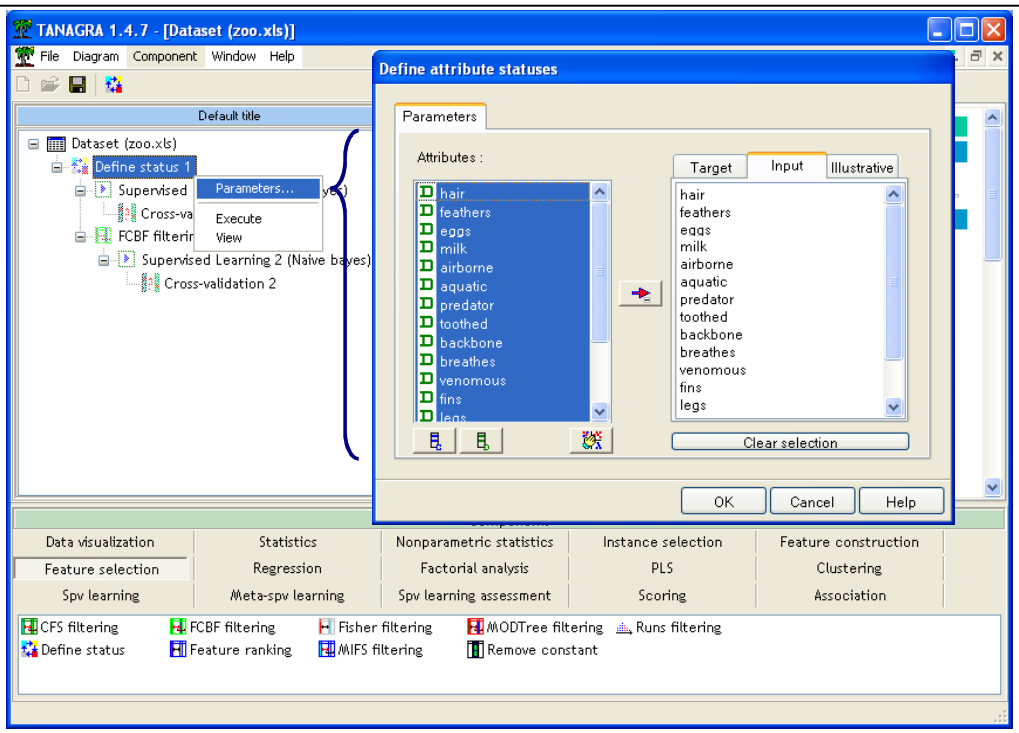
We want to implement the same analysis than the CONGRESSVOTE dataset. We click on the DIAGRAM / LOAD SUBDIAGRAM menu. We select the previous sub-diagram file.



The diagram is supplemented in the following way.



So, we must set the right predictive and class attributes. Of course, we must clear the previous selection before (use the CLEAR SELECTION button).



Then, we can execute the cross-validation when we use all the attributes (VIEW menu on CROSS-VALIDATION 1) and when we use only the selected attributes (VIEW menu of CROSS-VALIDATION 2).

The error rate without feature selection is 6%. When we insert the FCBF feature selection before the learning algorithm, the error rate becomes 10.67%. We note that this feature selection is not efficient on the ZOO dataset.

We note especially that this new functionality makes it possible to transpose very easily a succession of analysis on another dataset.