Subject

In this tutorial, we use the stepwise discriminant analysis (STEPDISC) in order to determine useful variables for a classification task.

Feature selection for supervised learning

Feature selection. The determination of relevant variables is an important step in a classification process. The model complexity is reduced; it is easier to interpret. Moreover, the diminution of the variables to collect is an advantage during the deployment of the model. In some cases, the variable selection enables to improve the model accuracy.

Manual selection by an expert domain is certainly the best approach. But because the number of candidate descriptors is often large, it is not always possible in practice. So, we must select automatically the best variables. We can also use the automatic process as a preliminary approach in order to filter out the really irrelevant attributes.

STEPDISC approach. STEPDISC (Stepwise Discriminant Analysis) is always associated to discriminant analysis because it relies on the same criterion i.e. the WILKS' LAMBDA. So it is often presented such as a method especially intended for the discriminant analysis. In effect, it could be useful for various linear models because they are based upon the same representation bias (e.g. logistic regression, linear SVM, etc.). However, it is not really adapted to non-linear model such as nearest neighbor or multi layer perceptron.

We implement the FORAWRD and the BACKWARD strategies in TANAGRA. In the FORWARD approach, at each step, we determine which is the variable that really contributes to the discrimination between the groups. We add this variable if its contribution is significant. The process stops when there is no attribute to add in the model. In the BACKWARD approach, we begin with the complete model with all descriptors. We search which is the less relevant variable. We remove this variable if the removing does not significantly damage the discrimination between groups. The process stops when there is no variable to remove.

Stopping rule. The stopping rule is a key point of the algorithm. Using the standard statistical framework of hypothesis testing is not adapted here. The interpretation of the computed p-value as a probability of rejecting the null hypothesis – if it is true – is erroneous. Indeed, the variable to test is selected among several variables, it maximizes the statistic of the test, and the variables used at one step are the resulting selected (or unselected) variables at the previous step. So the comparison of the p-value with a predefined significance level must be done with caution. It must be viewed mainly such as an adjustment tool that enables to direct the process to the desired solutions (a learning bias). In TANAGRA, we can compare the p-value with a predefined "significance level", or compare the statistic F with a predefined threshold value. We can also directly set the number of selected variables.

Dataset

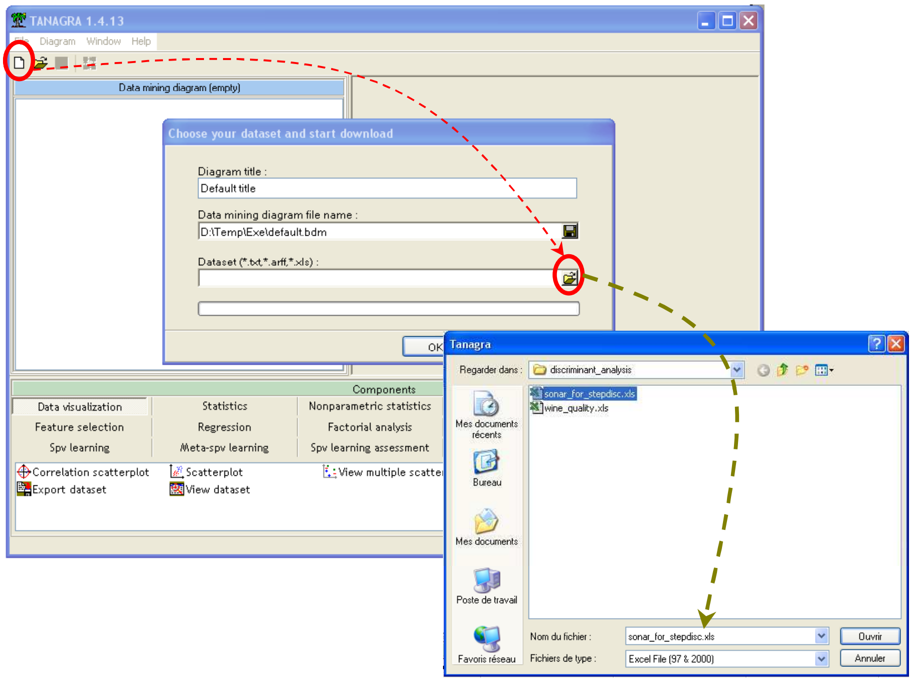
We use the SONAR dataset (SONAR_FOR_STEPDISC.XLS). The goal is to determine the kind of object (MINE or ROCK) starting from parameters gathered with a sensor. Variable selection process seems necessary here. The ratio between the number of candidate attributes (60) and the number of observations (208) seems unfavorable. The overfitting problem -- the model is too specific to the learning set -- often occurs in this situation.

Linear discriminant analysis

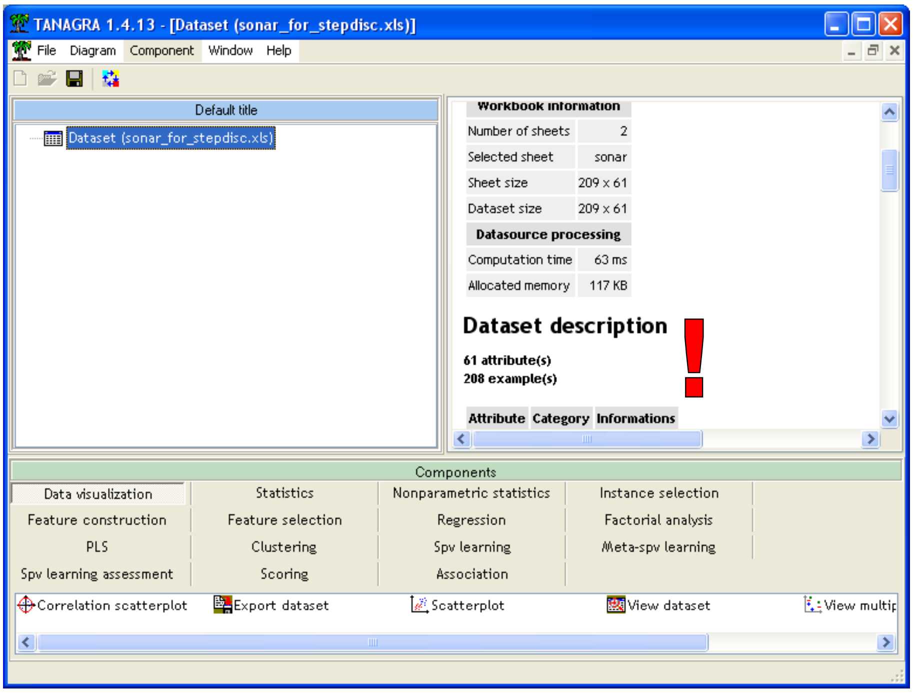
In a first step, we use all the descriptors for the construction of the prediction model. It is our reference situation; it will enable us to evaluate the efficiency of the STEPDISC method.

Data importation

In order to initialize a new diagram and import the dataset, we activate the FILE/NEW menu. We use the EXCEL file format in this tutorial; the dataset must be in the first sheet of the workbook.

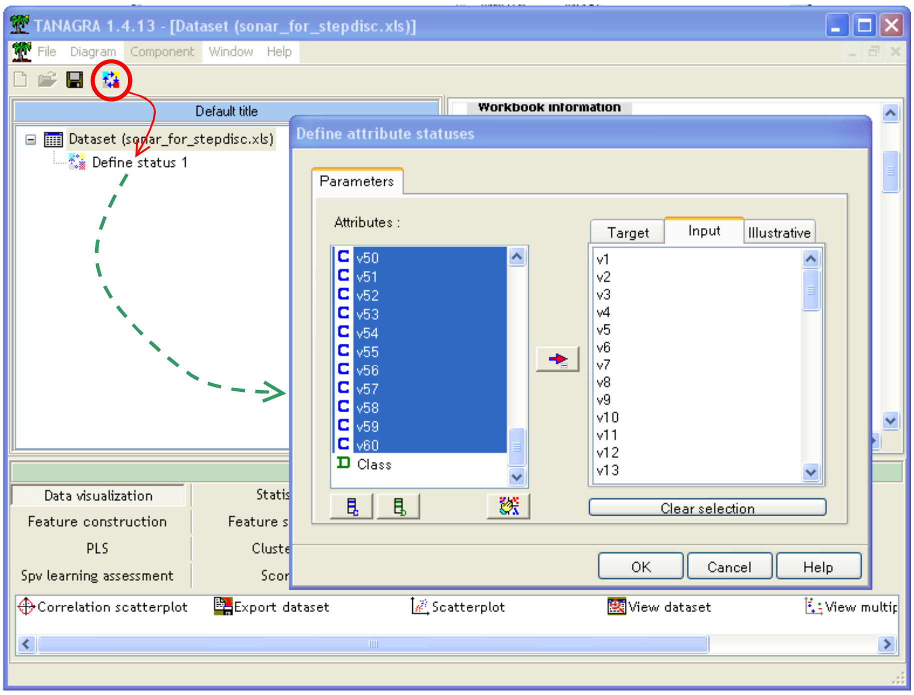


We check that there are really 61 variables (60 descriptors + the class attribute) and 208 instances in our file.

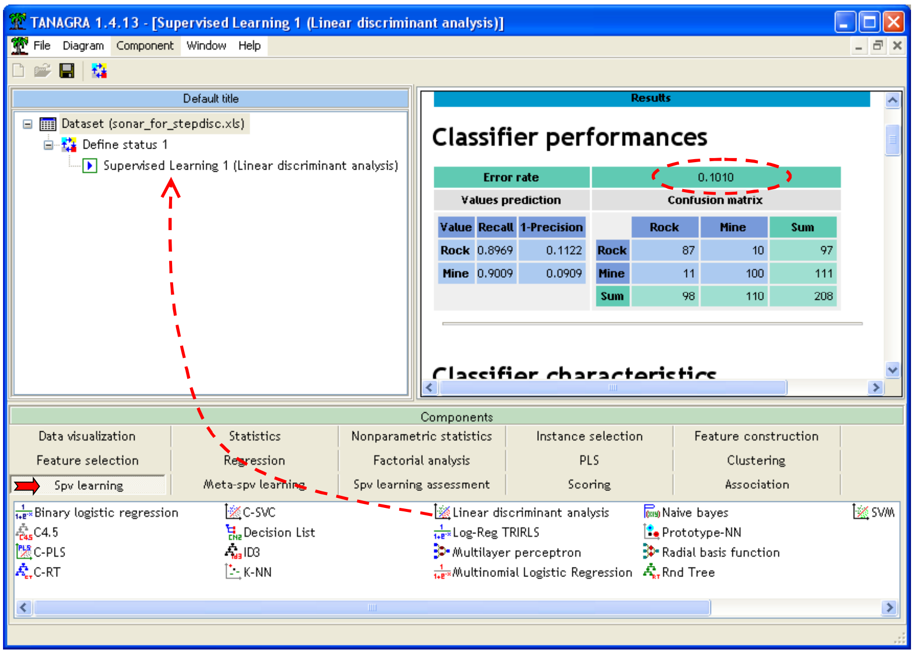


Linear discriminant analysis

We use the DEFINE STATUS component in order to set the INPUT attributes (predictive attributes, V1...V60) and the TARGET attribute (class attribute, CLASS).

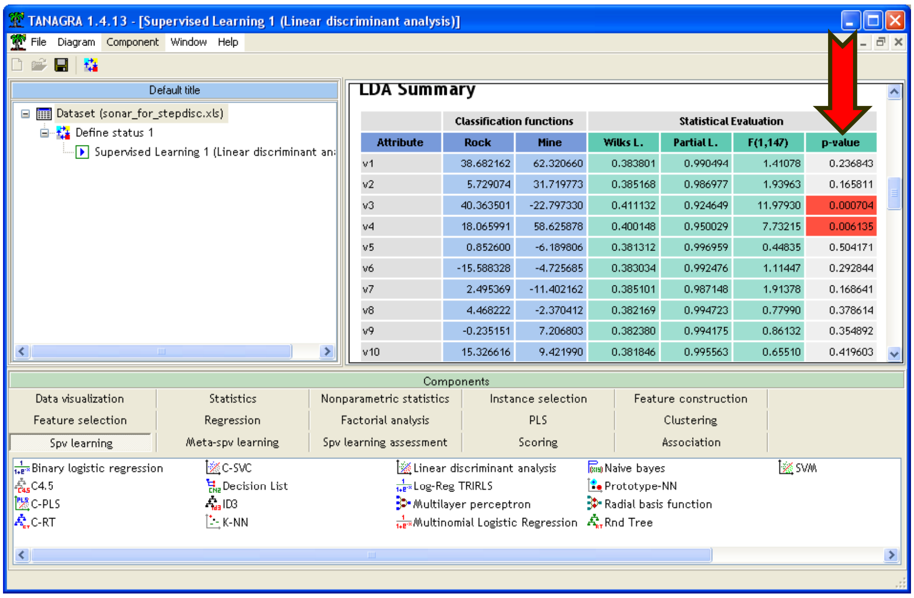


Then we add the LDA component into the diagram.

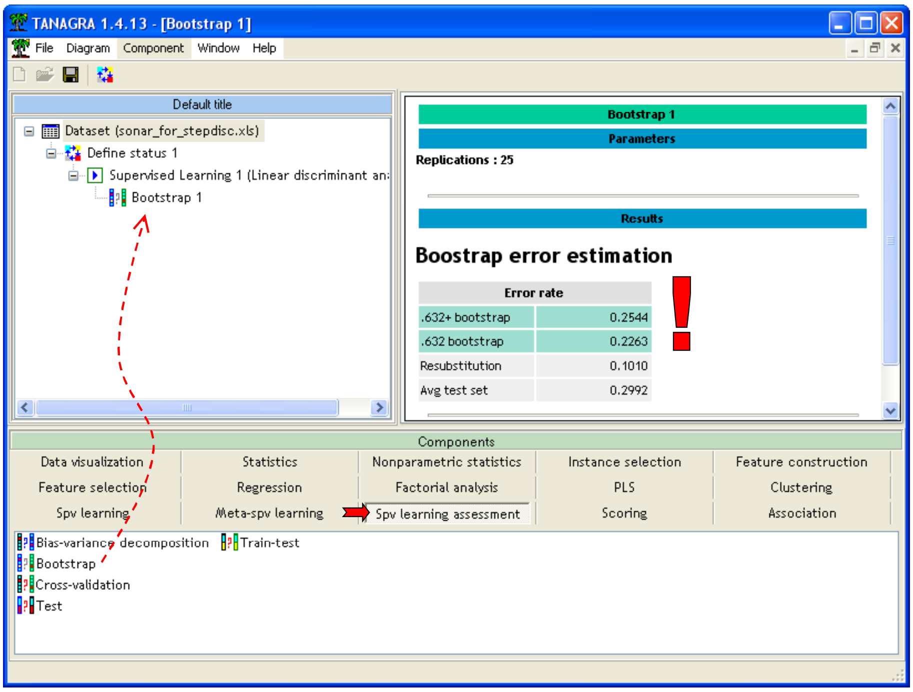


We see below the confusion matrix. The resubstitution error rate is **0.1010**. The detailed result of the learning process points out that several descriptors seem not relevant in the classification model.

Several questions can arise: "can we remove directly these variables?"; "all these variables or some of them?" ; "the remaining variables are all relevant in the resulting model?"... The feature selection process must provide responses to these questions.



We know that the resubstitution error rate is often optimistic. We use a resampling method -- bootstrap -- for obtaining an honest error estimate.



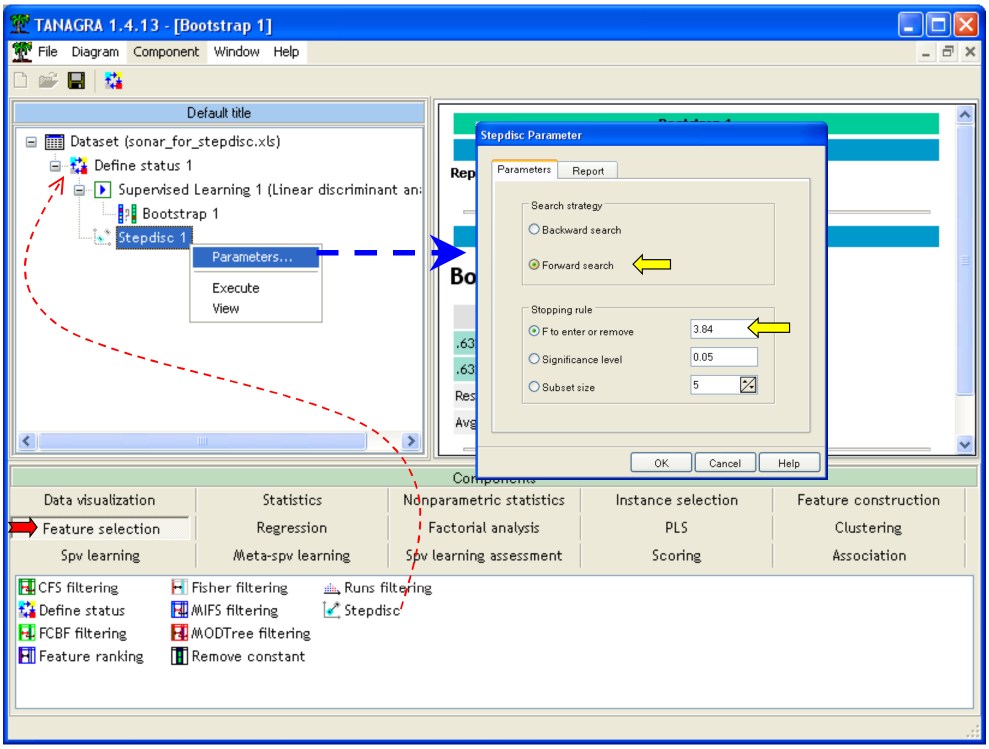
We observe that the true error rate is about **0.2544**, broadly higher than the indicated value by the resubstitution error rate.

STEPDISC feature selection method

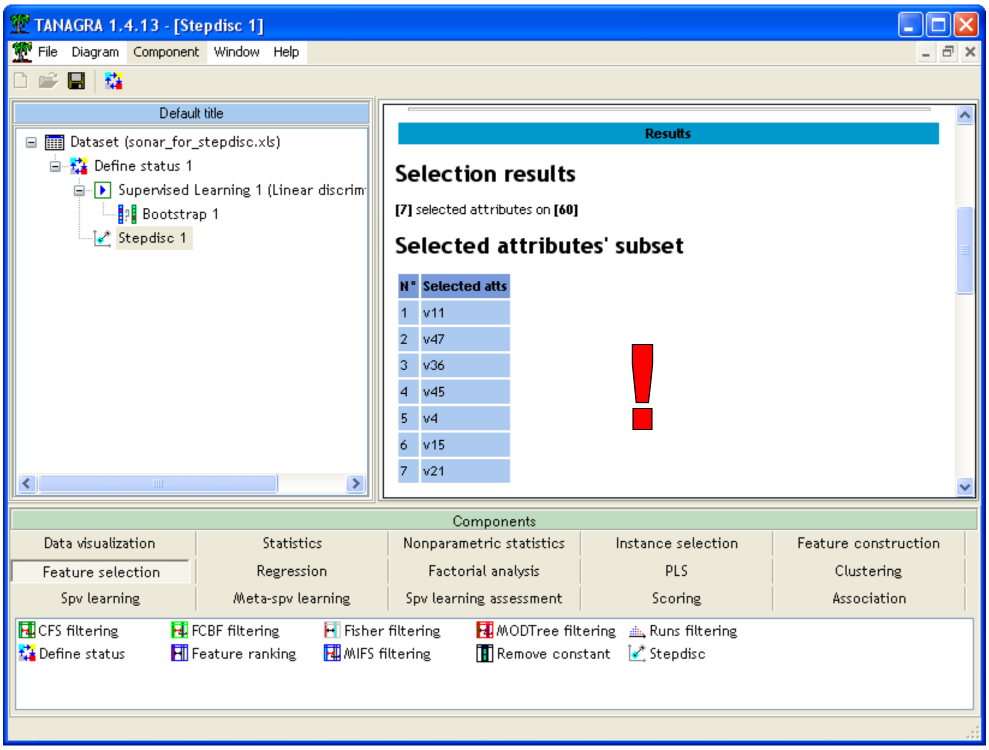
Now we insert the STEPDISC approach in the learning framework. The key questions are: "how many variables are sufficient for the classification?"; "what is the consequence of the selection on the subsequent model?"

STEPDISC

We add the STEPDISC component (FEATURE SELECTION tab) under the DEFINE STATUS 1 into the diagram. We activate the PARAMETERS menu.



We ask the FORWARD strategy and set the comparison of the computed statistic F to 3.84[1] as the stopping rule. We activate the VIEW menu in order to obtain the results. We observe that **7 descriptors** are selected. They are listed in a table.



[1] This is the default threshold value in SPSS.

The detailed results are displayed in the following table. Only the results for the 5 (can be modified) best attributes are displayed. At each step, we observe the best attributes. We can check especially if the best one has really a significant better F-value than the following attributes.

Detailed results

N°	d.f	Best	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	(1, 206)	**v11** L : 0.813 F : 47.23 p : 0.0000	**v11** L : 0.813 F : 47.23 p : 0.0000	**v12** L : 0.846 F : 37.36 p : 0.0000	**v49** L : 0.882 F : 27.57 p : 0.0000	**v45** L : 0.884 F : 27.11 p : 0.0000	**v10** L : 0.884 F : 26.94 p : 0.0000
2	(1, 205)	**v47** L : 0.731 F : 23.23 p : 0.0000	**v47** L : 0.731 F : 23.23 p : 0.0000	**v46** L : 0.738 F : 21.11 p : 0.0000	**v49** L : 0.743 F : 19.37 p : 0.0000	**v48** L : 0.749 F : 17.69 p : 0.0000	**v45** L : 0.750 F : 17.46 p : 0.0000
3	(1, 204)	**v36** L : 0.679 F : 15.67 p : 0.0001	**v36** L : 0.679 F : 15.67 p : 0.0001	**v37** L : 0.689 F : 12.45 p : 0.0005	**v21** L : 0.693 F : 11.01 p : 0.0011	**v35** L : 0.696 F : 10.09 p : 0.0017	**v22** L : 0.697 F : 9.78 p : 0.0020
4	(1, 203)	**v45** L : 0.653 F : 8.09 p : 0.0049	**v45** L : 0.653 F : 8.09 p : 0.0049	**v44** L : 0.655 F : 7.31 p : 0.0074	**v4** L : 0.657 F : 6.53 p : 0.0113	**v21** L : 0.658 F : 6.35 p : 0.0125	**v43** L : 0.660 F : 5.81 p : 0.0168
5	(1, 202)	**v4** L : 0.630 F : 7.19 p : 0.0079	**v4** L : 0.630 F : 7.19 p : 0.0079	**v21** L : 0.635 F : 5.58 p : 0.0191	**v31** L : 0.637 F : 4.91 p : 0.0279	**v54** L : 0.638 F : 4.55 p : 0.0342	**v23** L : 0.638 F : 4.46 p : 0.0359
6	(1, 201)	**v15** L : 0.611 F : 6.15 p : 0.0140	**v15** L : 0.611 F : 6.15 p : 0.0140	**v16** L : 0.612 F : 6.07 p : 0.0146	**v23** L : 0.616 F : 4.60 p : 0.0331	**v21** L : 0.616 F : 4.53 p : 0.0346	**v22** L : 0.617 F : 4.27 p : 0.0401
7	(1, 200)	**v21** L : 0.585 F : 9.03 p : 0.0030	**v21** L : 0.585 F : 9.03 p : 0.0030	**v20** L : 0.589 F : 7.57 p : 0.0065	**v22** L : 0.592 F : 6.58 p : 0.0111	**v31** L : 0.594 F : 5.86 p : 0.0164	**v23** L : 0.597 F : 4.85 p : 0.0288
8	(1, 199)	-	**v52** L : 0.577 F : 2.89 p : 0.0908	**v54** L : 0.578 F : 2.30 p : 0.1312	**v49** L : 0.579 F : 2.15 p : 0.1440	**v43** L : 0.579 F : 2.03 p : 0.1556	**v31** L : 0.580 F : 1.65 p : 0.2008

We can make several comments from this table. At the first step, V11 is the best attribute; the difference with the following variable (V12) seems large (+27%). At the second step, V47 is the best attribute, but the difference with the following variable is not obvious.... At the 6th step, the choice between V15 and V16 is not conclusive. Adding or removing a small number of instances in the dataset can invert the result.

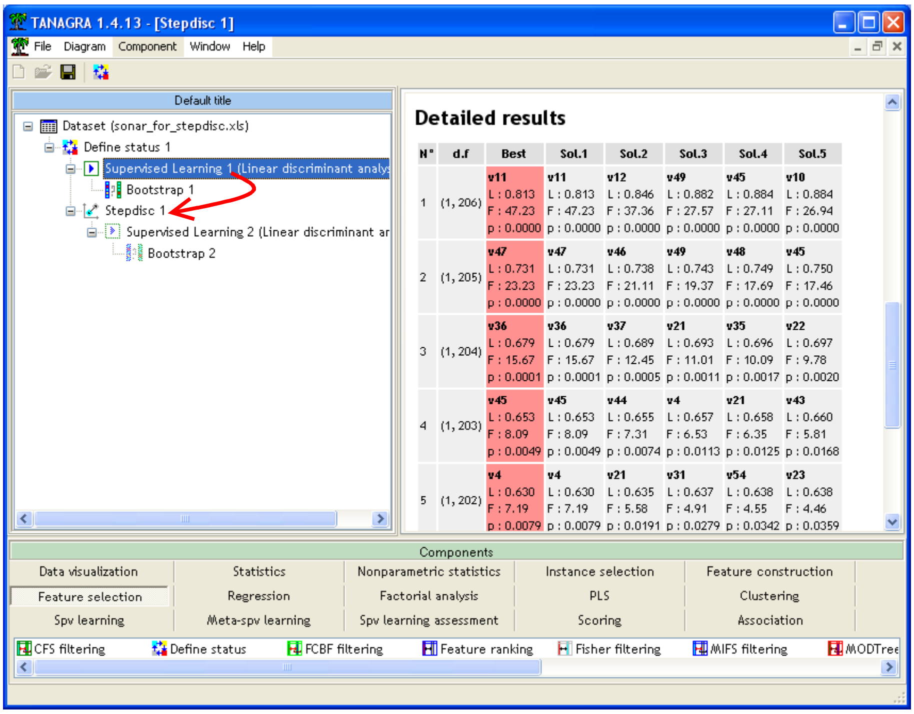
N.B. STATISTICA, with the same parameters, gives exactly the same succession of results.

Synthèse Analyse Incrément. (so[illegible]_for_stepdisc.sta)

ANALYSE DISCRIM.	Etape	F inc./exc	dl 1	dl 2	niveau p	Nb. de var inc	Lambda	Valeur F	dl 1	dl 2	niveau p
V11 -(E)	1	47.23461	1	206	.000000	1.000000	.813475	47.23462	1	206	.000000
V47 -(E)	2	23.23266	1	205	.000003	2.000000	.730668	37.78254	2	205	.000000
V36 -(E)	3	15.67116	1	204	.000104	3.000000	.678543	32.21473	3	204	.000000
V45 -(E)	4	8.09142	1	203	.004903	4.000000	.652533	27.02379	4	203	.000000
V4 -(E)	5	7.19146	1	202	.007931	5.000000	.630101	23.71670	5	202	.000000
V15 -(E)	6	6.14801	1	201	.013979	6.000000	.611400	21.29227	6	201	.000000
V21 -(E)	7	9.03302	1	200	.002991	7.000000	.584979	20.27033	7	200	.000000

STEPDISC + Discriminant Analysis + Bootstrap

According to STEPDISC, 7 variables seem sufficient for the classification process. Now, we want to check the efficiency of the resulting model. We add the Linear Discriminant Analysis and the Bootstrap components into the diagram. The simplest way is to drag and drop the corresponding branch of the diagram under STEPDISC 1.



We observe that all the variables are relevant in the classification model (at the 5% level). The resubstitution error rate is **0.1875**.

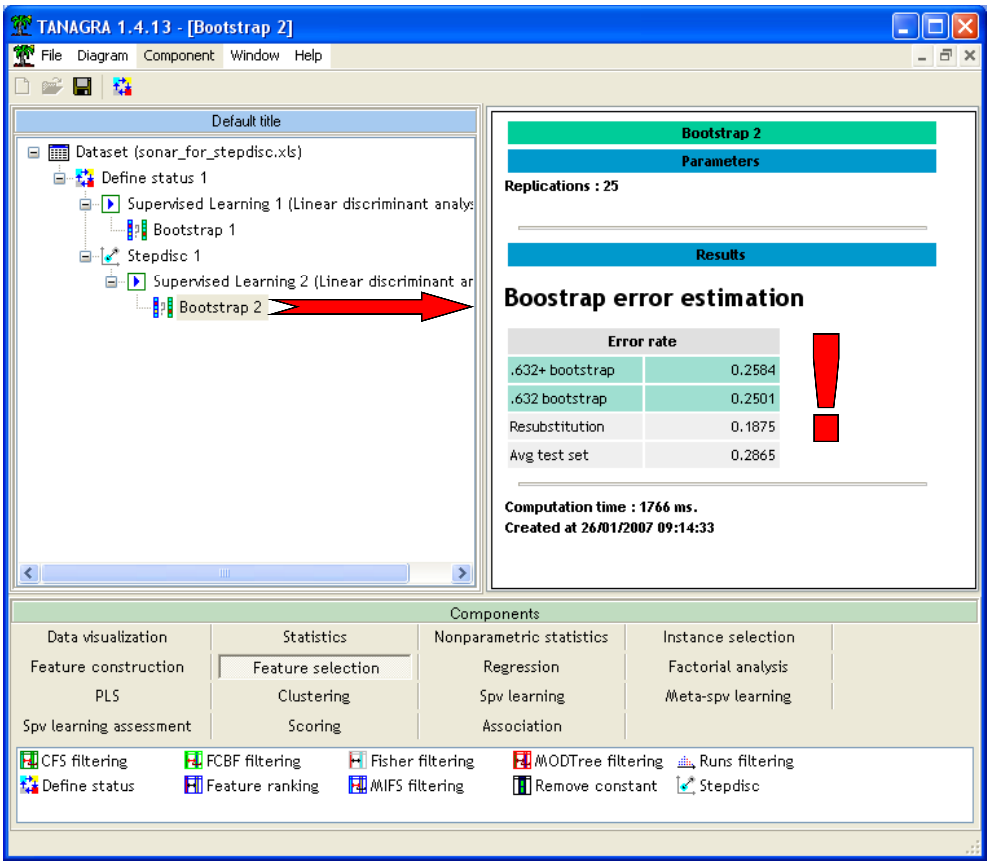
Classifier performances

Error rate			0.1875			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		Rock	Mine	Sum
Rock	0.8041	0.2041	Rock	78	19	97
Mine	0.8198	0.1727	Mine	20	91	111
			Sum	98	110	208

LDA Summary

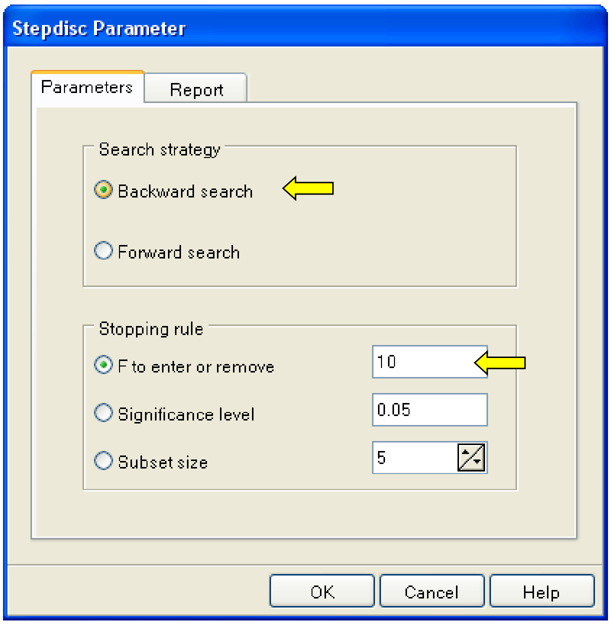
	Classification functions		Statistical Evaluation			
Attribute	Rock	Mine	Wilks L.	Partial L.	F(1,200)	p-value
v11	8.637494	16.695728	0.656707	0.890776	24.52325	0.000002
v47	29.753123	36.964342	0.600735	0.973772	5.38679	0.021298
v36	10.181178	6.494436	0.650366	0.899462	22.35513	0.000004
v45	-7.377507	-2.882663	0.601739	0.972147	5.73010	0.017601
v4	2.268385	15.468078	0.613508	0.953499	9.75378	0.002055
v15	3.120895	-0.313122	0.616228	0.949291	10.68357	0.001272
v21	11.106784	13.585777	0.611400	0.956787	9.03302	0.002991
constant	-8.278780	-11.421379	-			

And the bootstrap error rate estimate is **0.2584**. We note that this model with 7 variables is as accurate as the model with 60 variables.



BACKWARD Selection

In this section, we modify the parameter of STEPDISC. We select a backward strategy[2].



[2] The threshold value 10 is suggested by STATISTICA for this problem.

The process selects also 7 variables, but the selected subset is noticeably different of the subset selected with the FORWARD approach.

Selection results

[7] selected attributes on **[60]**

Selected attributes' subset

N°	Selected atts
1	v4
2	v12
3	v30
4	v31
5	v32
6	v36
7	v49

Detailed results

N°	d.f	Best	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	(1, 147)	**v26** L : 0.380 F : 0.00 p : 0.9929	**v26** L : 0.380 F : 0.00 p : 0.9929	**v45** L : 0.380 F : 0.00 p : 0.9846	**v33** L : 0.380 F : 0.01 p : 0.9210	**v46** L : 0.380 F : 0.01 p : 0.9048	**v38** L : 0.380 F : 0.03 p : 0.8683
2	(1, 148)	**v45** L : 0.380 F : 0.00 p : 0.9854	**v45** L : 0.380 F : 0.00 p : 0.9854	**v33** L : 0.380 F : 0.01 p : 0.9208	**v46** L : 0.380 F : 0.02 p : 0.9021	**v38** L : 0.380 F : 0.03 p : 0.8680	**v58** L : 0.380 F : 0.04 p : 0.8407
3	(1, 149)	**v33** L : 0.380 F : 0.01 p : 0.9212	**v33** L : 0.380 F : 0.01 p : 0.9212	**v46** L : 0.380 F : 0.02 p : 0.8784	**v38** L : 0.380 F : 0.03 p : 0.8682	**v37** L : 0.380 F : 0.04 p : 0.8403	**v58** L : 0.380 F : 0.04 p : 0.8394
4	(1, 150)	**v46** L : 0.380 F : 0.03 p : 0.8713	**v46** L : 0.380 F : 0.03 p : 0.8713	**v38** L : 0.380 F : 0.03 p : 0.8687	**v58** L : 0.380 F : 0.04 p : 0.8438	**v37** L : 0.380 F : 0.04 p : 0.8431	**v43** L : 0.380 F : 0.06 p : 0.8118

FORWARD – BACKWARD comparison

When we compare the results of the two strategies, we observe that two variables only are shared by the resulting subsets.

N°	Forward	Backward
1	v4	v4
2	v11	v12
3	v15	v30
4	v21	v31
5	v36	v32
6	v45	v36
7	v47	v49

This kind of result leaves often puzzled the users. What is the “true” best subset?

I think there is no numerical – statistical – satisfying response to this question. These tools give us some indications on the possible relevant attributes. Only the domain knowledge and the specification of the study enable us to select an appropriate result.