

Subject

In order to evaluate a supervised learning algorithm, we often split the dataset into training set, which is used in the training process, and test set, which is used to obtain an unbiased error rate evaluation. There are sampling components in TANAGRA, which enable to subdivide randomly the dataset, but in some circumstances, the user want use a predefined test set for their comparisons.

Dataset

We use the SONAR dataset from the UCI Repository (<http://www.ics.uci.edu/~mlern/MLRepository.html> -- [sonar_with_test_set.xls](#)). The dataset contains 208 examples with 60 descriptors (ATTRIBUTE_1 to ATTRIBUTE_60) and a binary class attribute (CLASS). Rather than to use two distinct files for the learning and the test set, we prefer join them together in a single file and create an additional column indicating the role that each observation must play (EXAMPLESTATUS).

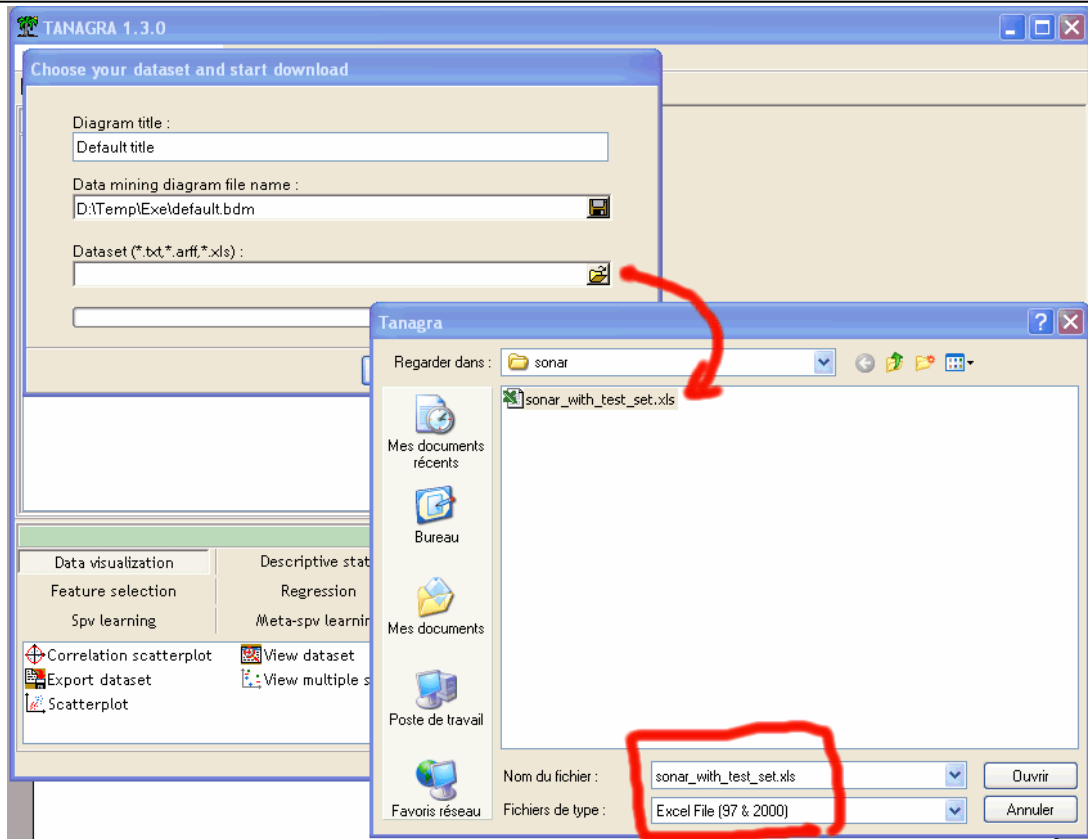
The screenshot shows a Microsoft Excel spreadsheet titled 'sonar_with_test_set.xls'. The spreadsheet contains data for the SONAR dataset. The columns are labeled as follows: BE (attribute 57), BF (attribute 58), BG (attribute 59), BH (attribute 60), BI (Class), and BJ (ExampleSta). The data rows are numbered 101 to 117. The 'Class' column contains values 'Rock' and 'Mine'. The 'ExampleSta' column contains values 'Learning' and 'Test'. A red dashed circle highlights the 'ExampleSta' column.

	BE	BF	BG	BH	BI	BJ
	attribute 57	attribute 58	attribute 59	attribute 60	Class	ExampleSta
101	0	0	0	0	Rock	Learning
102	0	0	0	0	Rock	Learning
103	0.01	0	0	0.01	Rock	Learning
104	0	0.01	0	0.01	Mine	Learning
105	0.01	0.01	0.01	0.01	Mine	Learning
106	0.01	0.01	0.01	0.01	Rock	Learning
107	0	0.01	0.01	0.01	Mine	Learning
108	0	0.01	0	0	Mine	Learning
109	0	0.01	0.01	0	Mine	Test
110	0	0	0.01	0	Mine	Test
111	0	0	0.01	0.02	Rock	Test
112	0.01	0.01	0.01	0	Mine	Test
113	0	0	0	0	Mine	Test
114	0.01	0	0.01	0.01	Rock	Test
115	0.01	0.01	0.01	0	Rock	Test
116	0	0.01	0.01	0.01	Mine	Test
117	0.01	0.01	0.01	0.01	Rock	Test

Comparing learning methods

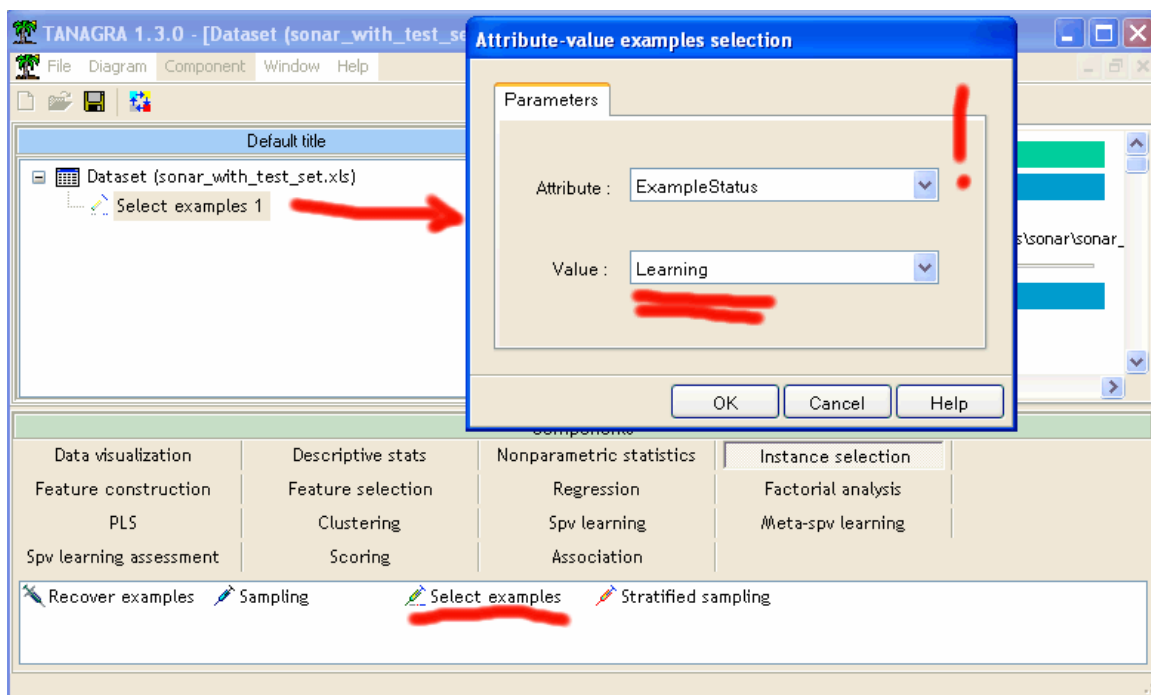
Download the dataset

First of all, we have to create a new stream diagram and select the dataset: SONAR_WITH_TEST_SET.XLS (FILE / NEW).

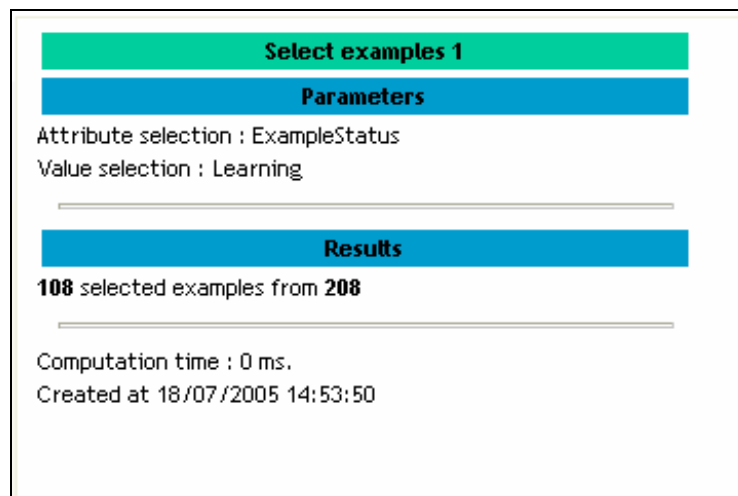


Select learning and test set

In the next step, we use the EXAMPLESTATUS column in order to define the learning and the test set. Add the SELECT EXAMPLES component in the diagram, and set the right parameters (Attribute: EXAMPLESTATUS, Value: LEARNING).

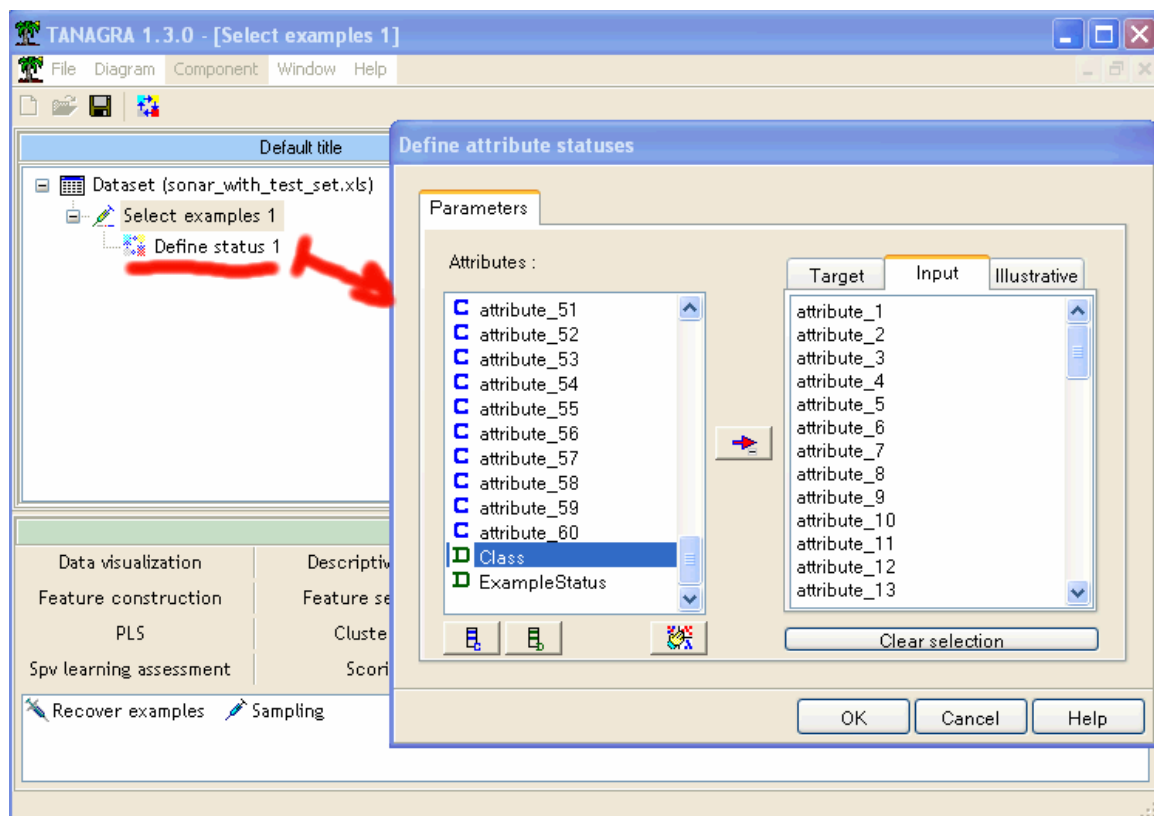


We see that the learning set size is 108 (thus, the test set size is 100).



SONAR prediction problem

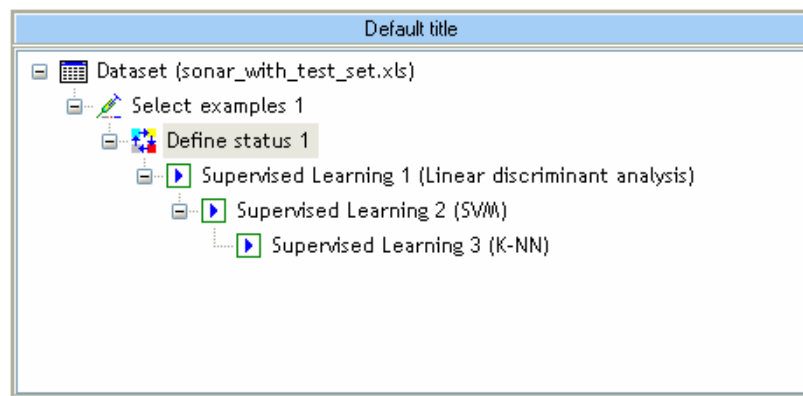
Add a DEFINE STATUS component in the stream diagram, set CLASS as TARGET, and the other descriptors as INPUT. Of course, we do not use the EXAMPLESTATUS column in the remaining.



Learning process

We want to compare three algorithms that have a very different representation and learning bias.

Add a linear discriminant analysis (LINEAR DISCRIMINANT ANALYSIS), a linear support vector machine (SVM) and a nearest neighbor algorithm (K-NN) in the diagram. Note that this kind of component must be embedded in meta-supervised component (META SPV LEARNING); we use a standard learning procedure (SUPERVISED LEARNING)¹.



The resubstitution error rates are:

- 2.78% for LDA;
- 12.04% for Linear SVM;
- 11.11% for K-NN (5-NN).

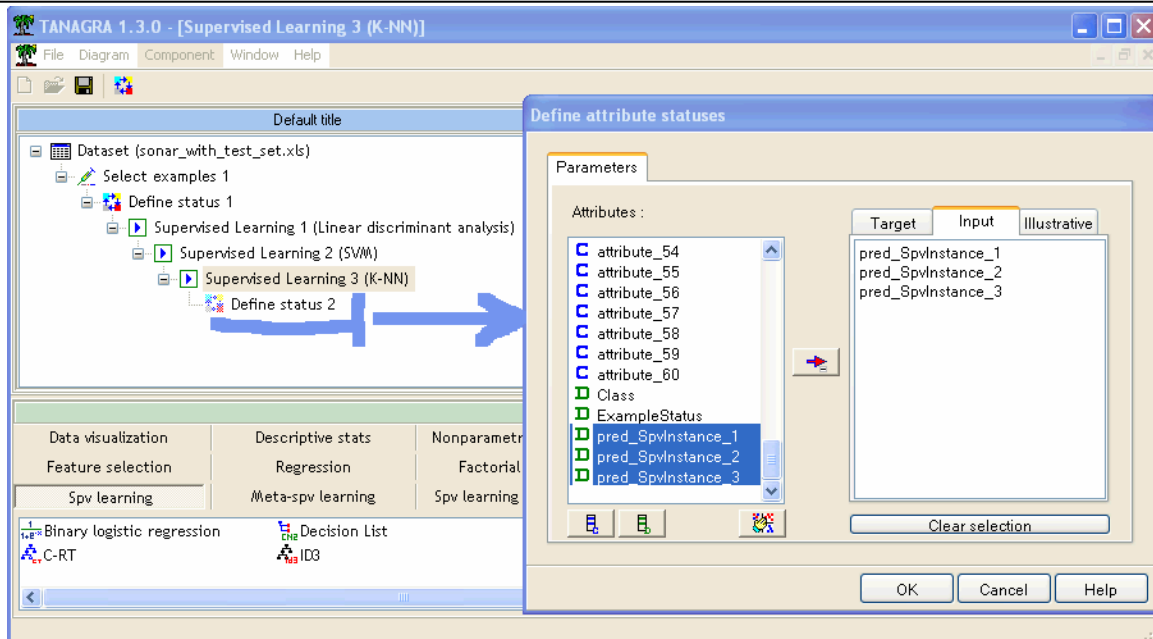
LDA seems the best learning algorithm on this dataset?

Test error rate

We must use the test set in order to obtain an unbiased error rate evaluation.

Add a DEFINE STATUS component in the diagram and set as TARGET the real class attribute (CLASS), the 3 projections of the methods are defined as INPUT.

¹ In concrete terms, we operate in two steps: first, we put the SUPERVISED LEARNING component (META SPV LEARNING tab) into the diagram; then we embed in this one the learning algorithm component (e.g. LINEAR DISCRIMINANT ANALYSIS) from the SPV LEARNING tab.



We add the TEST component in the diagram, the default parameter enables to compute the error rate on the unselected examples, that is the test set.²

The results illustrate the overfitting phenomenon, which is foreseeable if we consider the characteristics of our dataset: many descriptors (60) compared the learning set size (108 examples).

² It is possible to modify the parameter so that the error rate is computed on the selected examples (the learning set). In this case, we obtain again the resubstitution error rate.

Test 1

Parameters

Evaluation set : **unselected** examples

Results

pred_SplInstance_1

Error rate : 0.3600

Values prediction

Value	Recall	1-Precision	Mine	Rock	Sum	
Mine	0.6538	0.3462	Mine	34	18	52
Rock	0.6250	0.3750	Rock	18	30	48
Sum			52	48	100	

pred_SplInstance_2

Error rate : 0.2100

Values prediction

Value	Recall	1-Precision	Mine	Rock	Sum	
Mine	0.7500	0.1702	Mine	39	13	52
Rock	0.8333	0.2453	Rock	8	40	48
Sum			47	53	100	

pred_SplInstance_3

Error rate : 0.2200

Values prediction

Value	Recall	1-Precision	Mine	Rock	Sum	
Mine	0.9231	0.2727	Mine	48	4	52
Rock	0.6250	0.1176	Rock	18	30	48
Sum			66	34	100	

Computation time : 0 ms.
Created at 18/07/2005 15:22:48

Test error rate are:

- 36% for LDA (SUPERVISED LEARNING 1);
- 21% for SVM (SUPERVISED LEARNING 2);
- 22% for K-NN (SUPERVISED LEARNING 3).

It appears that the LDA particularly suffered on the SONAR problem, the SVM and K-NN present on the other hand comparable performances.