

Subject

How to use SVM (Support Vector Machine) in TANAGRA ? Compare results with the linear discriminant analysis (LDA).

Our implementation is a port of "SMO.JAVA" (WEKA package -- version 3-4).

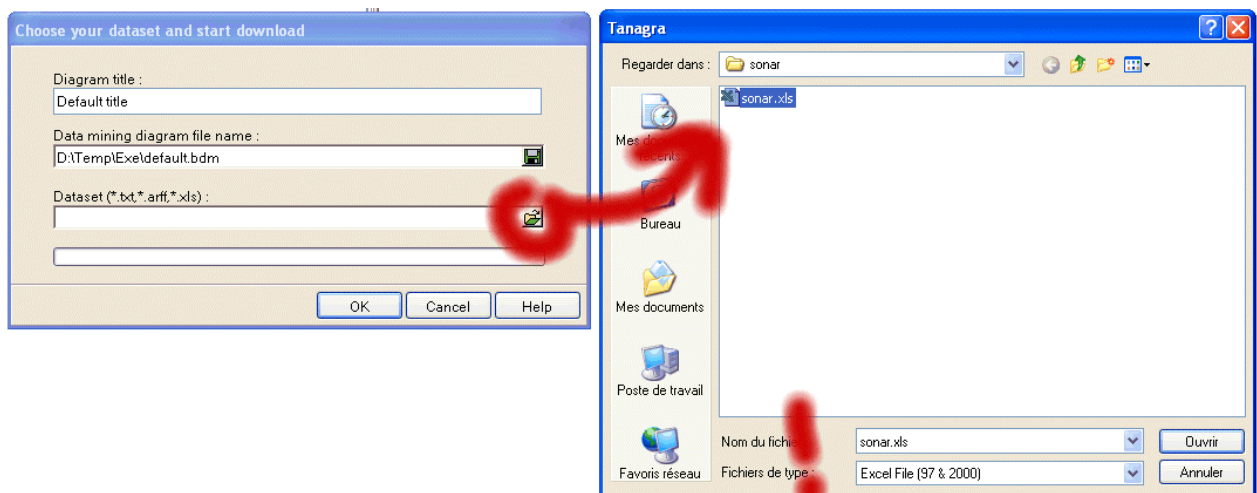
Dataset

SONAR dataset : predict an object (Rock or Mine) from a sonar (60 descriptors).

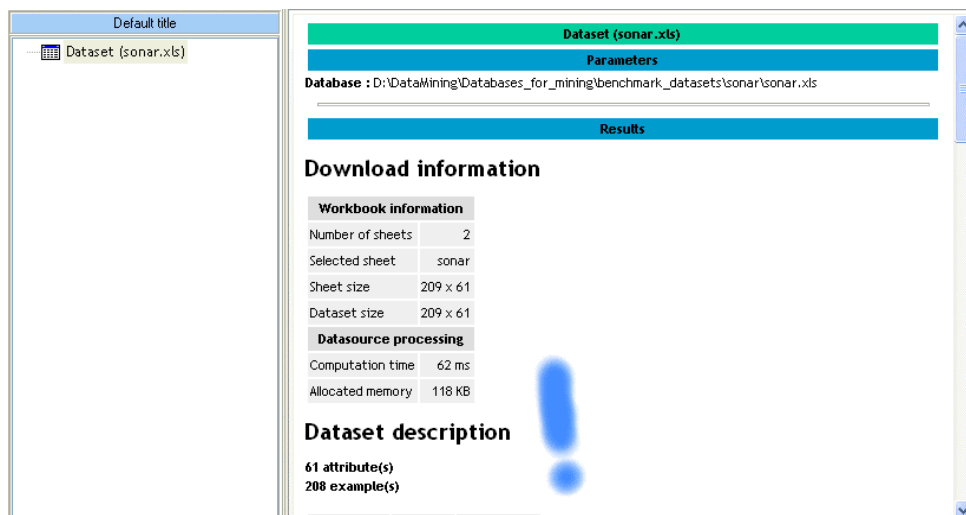
SVM

Download

Import the SONAR.XLS dataset (EXCEL file).

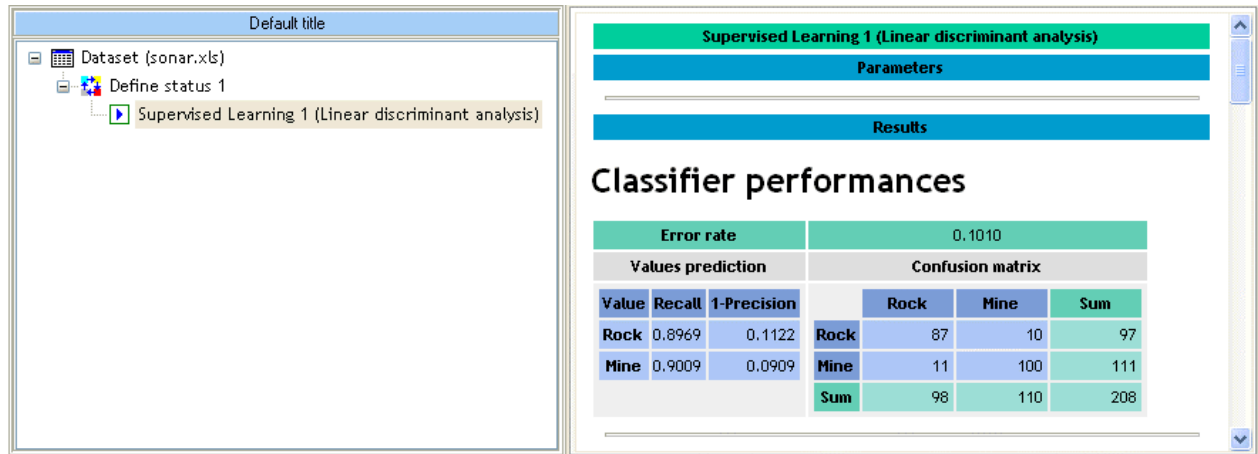


Check that you have the same data description.



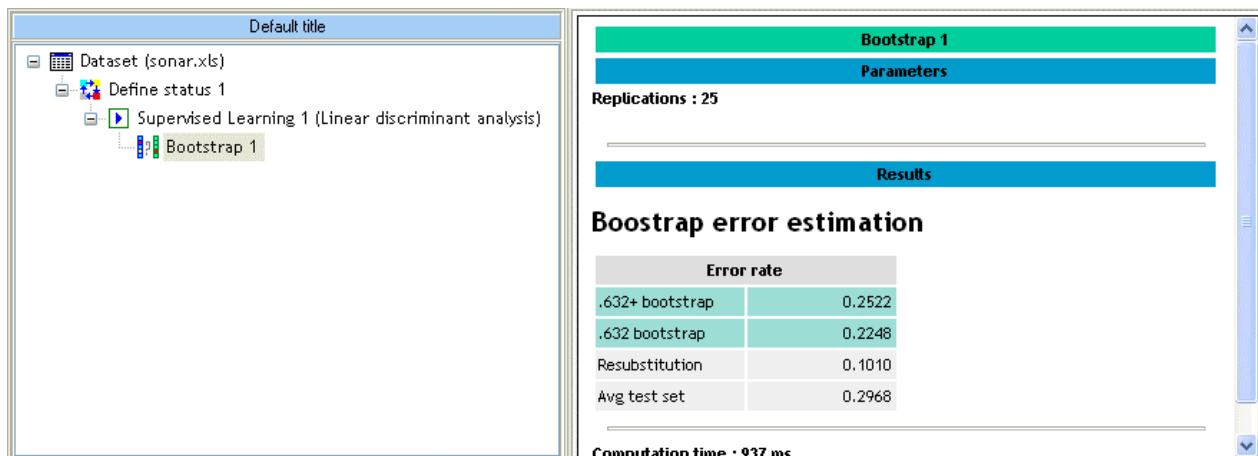
LDA

LDA induces a linear separator. The stream diagram is the following, TARGET attribute is "CLASS", the other continuous attributes are INPUT. The resubstitution error rate is 10% (default classifier error rate is 46% = 97/208).



We know that the resubstitution error rate is a biased estimator of the "true" error rate, especially when we have a lot of descriptors (60) compared to the number of examples (208).

To obtain an honest error rate estimate, we use the BOOTSTRAP component, the result is 25%.



Linear SVM

Use the SVM component with the default parameters which define a linear SVM.

The screenshot shows the Orange3 interface for a supervised learning task. On the left, a workflow is visible with 'Dataset (sonar.xls)', 'Define status 1', 'Supervised Learning 1 (Linear discriminant analysis)', 'Bootstrap 1', and 'Supervised Learning 2 (SVM)'. The right panel displays the 'Supervised Learning 2 (SVM)' parameters and results.

Supervised Learning 2 (SVM) Parameters

SVM Parameters	
Exponent	1
Filter type	NORMALIZE
Use polynom space normalization	0
Use RBF kernel	0
Gamma for RBF kernel	0.0100
Complexity	1.0000
Calculation parameter	
Epsilon for rounding	1.0E-012
Tolerance for accuracy	1.0E-003

Results

Classifier performances

Error rate		0.1202	
Values prediction		Confusion matrix	
Value	Recall	1-Precision	
Rock	0.8866	0.1400	
Mine	0.8739	0.1019	

	Rock	Mine	Sum
Rock	86	11	97
Mine	14	97	111
Sum	100	108	208

Resubstitution error rate is 12% and BOOTSTRAP error rate estimate is 20%, we significantly improve the LDA result.

The screenshot shows the Orange3 interface for a bootstrap analysis. On the left, a workflow is visible with 'Dataset (sonar.xls)', 'Define status 1', 'Supervised Learning 1 (Linear discriminant analysis)', 'Bootstrap 1', 'Supervised Learning 2 (SVM)', and 'Bootstrap 2'. The right panel displays the 'Bootstrap 2' parameters and results.

Bootstrap 2 Parameters

Replications : 25

Results

Bootstrap error estimation

Error rate	
.632+ bootstrap	0.2051
.632 bootstrap	0.1953
Resubstitution	0.1202
Avg test set	0.2390

Computation time : 4328 ms.
Created at 16/04/2005 09:13:48

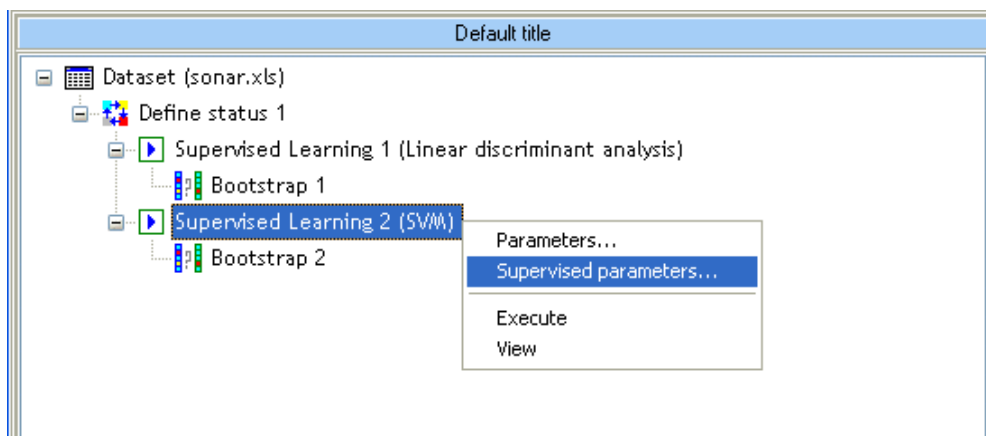
LDA and Linear SVM have the same representation bias, they induce a linear separator. But, the learning bias of SVM is more restrictive, it is more robust against the famous “curse of dimensionality”, when we have a lot of predictive attributes and numerous of them are irrelevant.

Using Kernel

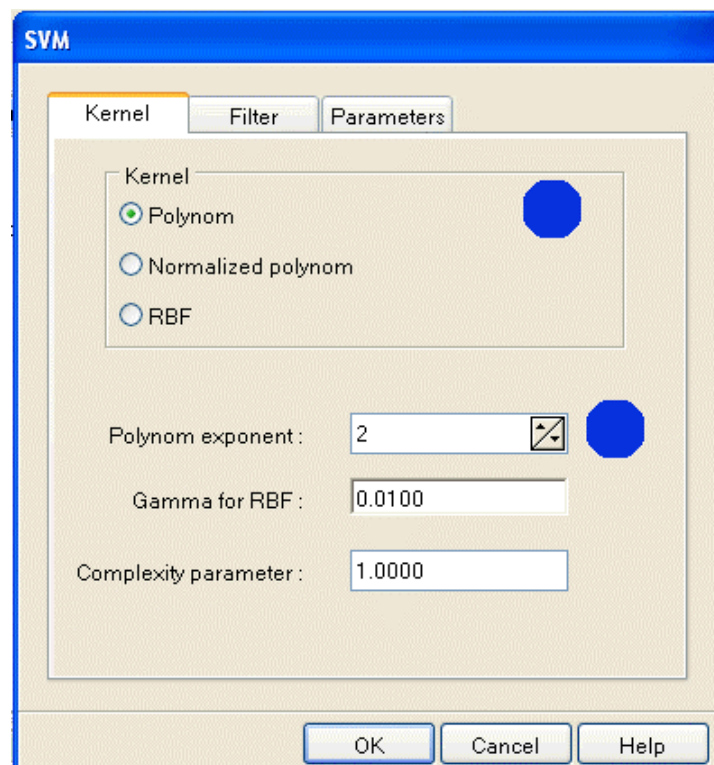
To obtain a more powerful classifier, we can generate a lot of new features, which are combinations of native descriptors: a linear classifier in the high dimensional description is a non-linear classifier in the native space.

The main contribution of SVM is that it is possible to make a projection in a new space through kernel functions, without explicitly generate the features.

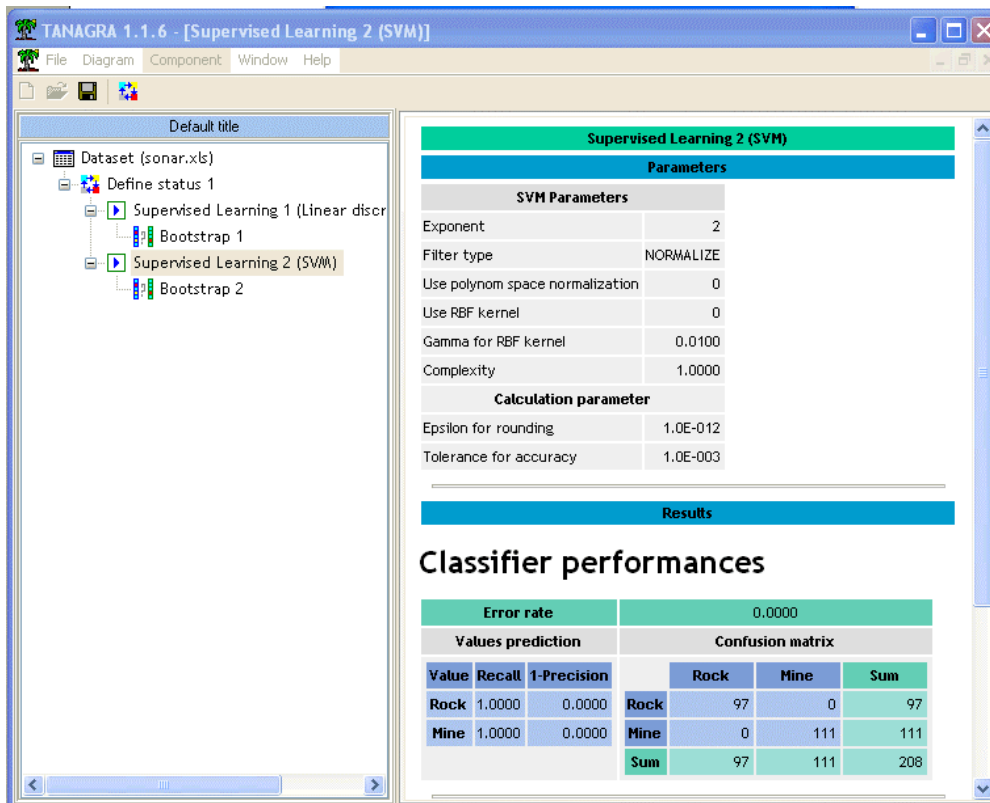
In this tutorial, we use a polynomial kernel of degree 2. Select the contextual menu.



And set the new parameters.



The resubstitution error rate is fallacious (0%).



The “true” error rate is actually 15%, this classifier seems definitely more powerful than linear classifiers (LDA and Linear SVM).

