## Subject

The LIBSVM<sup>1</sup> library contains various support vector algorithms for classification, regression... The implementation is particularly efficient, especially about the processing time, as we will see below. Some documentations are available on the website of the authors.

We have compiled the C source code in a DLL on which we connect TANAGRA. In the first time, only C-SVC, multi-class support vector machine for classification, is available. We will add the other components in the near future.

## Dataset

We treat a protein classification problem from their primary structures (Mhamdi et al., 2004). There are 122 examples of 2 families {C1, C2}, and 6740 Boolean (1/0) descriptors (3-grams). We have already used this dataset in a previous tutorial (NIPALS). The subject was using a nearest neighbour classification method from latent variables computed with a singular value decomposition algorithm.

# C-SVC

### Download the dataset

Download TANAGRA\_NIPALS.BDM. This is a binary format; it is necessary to choose the « Binary Data Mining Diagram » option in the open dialog box.

TANAGRA 1.4.2							
File Diagram Window Help							
🗅 😅 🔳 🔛							
Data mining	) diagram (empty)						
	Tanagra					? 🗙	
	Regarder dans :	눱 nipals		O Ø	• 🗈 🍯		
	Mes documents récents	Tanagra_Nipals.b	dm				
	Bureau						
	Mes documents						
	Poste de travail						
		Nom du fichier :	Tanagra_Nip	als.bdm	~	Ouvrir	
Data visualization	Favoris réseau	Fichiers de type :	Binary data n	ining diagram (*.bdm)	<ul> <li>v</li> </ul>	Annuler	n
Feature selection	Nogr oppron	raoconara	101900		1	orestoring	
Spv learning	Meta-spv learning	Spv learning as	sessment	Scoring		Association	
⊕Correlation scatterplot	🚉 Export dataset	🔏 Scatterp	lot	🔛 View dataset	Į	🔆 View multiple	scatterplot
							.d

<sup>&</sup>lt;sup>1</sup> http://www.csie.ntu.edu.tw/~cjlin/libsvm/

#### Supervised learning

Build the following diagram, set "classe" as TARGET, and the other attributes as INPUT.



Default parameter of C-SVC in TANAGRA is linear kernel. We obtain a perfect separation; it is not surprising because the relationship between the number of descriptors and the number of observations is very high. The computation time is very impressive, 2.7 seconds on a P4 at 3Ghz, the library was built with much care.

Clas						
Error rate						
Values prediction				Confu		
∀alue	Recall	<b>1-Precision</b>		C1	C2	Sum
C1	1.0000	0.0000	C1	54	0	54
C2	1.0000	0.0000	C2	0	68	68
			Sum	54	68	122

### **Classifier characteristics**

#### Data description

Target attribute	Classe (2 values)
# descriptors	6740

#### SVM characteristics

Characteristic	∀alue
# classes	2
# support vectors	97
# support vectors for e	ach class
#sv. for C1	44
#sv. for C2	53
Computation time : 2672 Created at 09/01/2006	: ms.

The component shows the confusion matrix, the number of support vectors for each class. The parameters of the method are showed in the high part of the report.

Supervised Learning 1 (C-SVC)					
Parameters					
Parameters					
Kernel type	LINEAR				
Degree (poly)	1.00				
Gamma in kernel function (poly/rbf/sigmoid)		See LIBSVM website			
Coef0 in kernel function(ploy/sigmoid)	0.00				
Tolerance of termination criteria	0.0000	for more information			
C (Complexity Cost)					
Compute probability estimates					
Use shrinking heuristics		)			

#### Unbiased error rate estimate

To obtain an unbiased error rate estimate, we use the bootstrap method (Efron & Tibshirani, 1997) available in the SPV LEARNING ASSESMENT tab.

Default title		Bog	stetran 4		
🖃 🎹 Dataset (FichBool3_5_01.txt)	Parameters Replications : 25 Results Results				
arian and a status 1 arian arian ari arian arian ari					
	Boostrap error estimation				
	Error rate				
	.632+ bootstrap	0.0350	<b>(</b>		
	.632 bootstrap	0.0336	ē		
	Resubstitution	0.0000			
	Avg test set	0.0531			
	Computation time : 6078 Created at 09/01/2006 16	1 ms. :46:59			

The estimated error rate is 3.5% and the whole computation time is 60 seconds. These results are all the more interesting when we compare them with the performances of K-NN on the same dataset.

