Subject

Use NIPALS algorithm for dimensionality reduction in a proteins discrimination problem.

NIPALS is a possible implementation of singular value decomposition (SVD); it enables to compute factors (latent variable) of principal component analysis (PCA) without a correlation matrix diagonalization. The computing time is dramatically reduced on a huge dataset.

Dataset

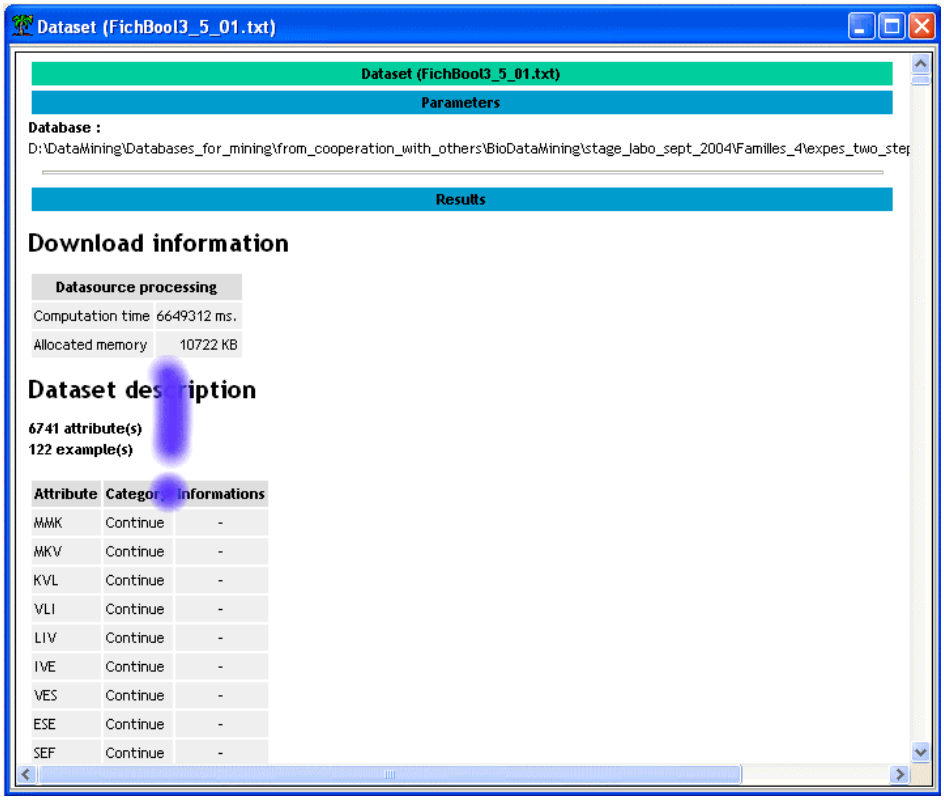
Proteins classification from their primary structures (Mhamdi et al., 2004).

There are 122 examples of 2 families {C1, C2}, and 6740 Boolean (1/0) descriptors (3-grams).

NIPALS

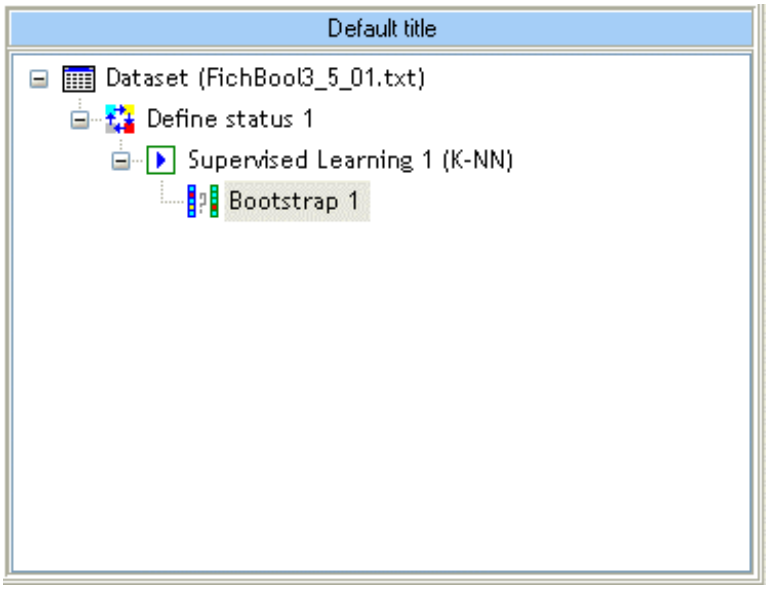
Download the dataset

Download TANAGRA_NIPALS.BDM.

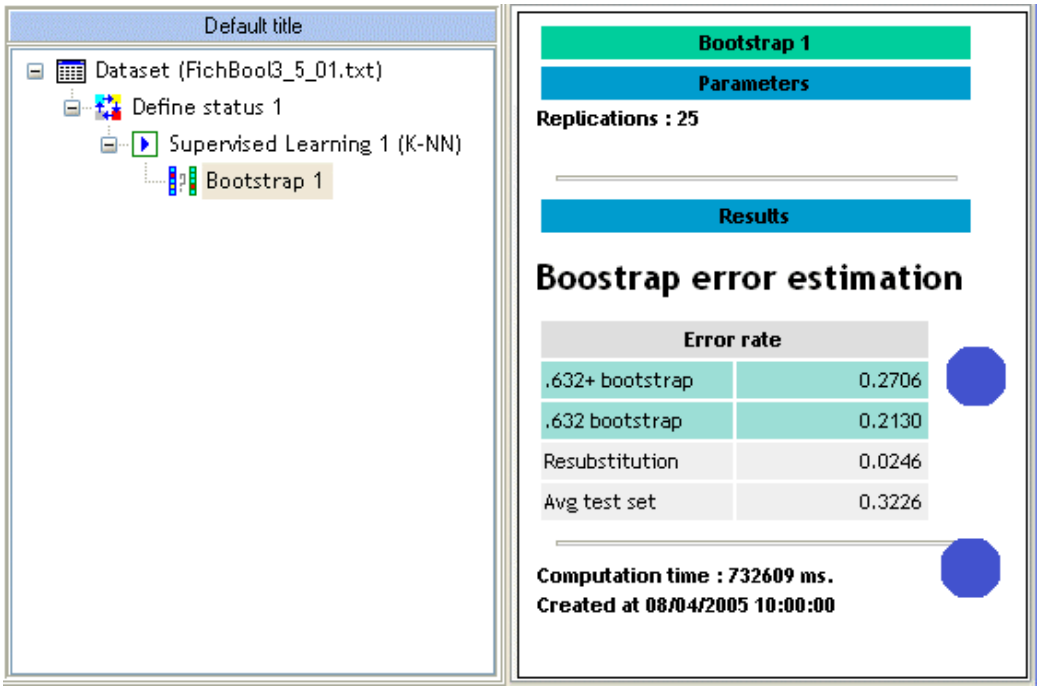


Supervised learning

Let us evaluate a 5-NN (nearest neighbor) on our dataset. Set as TARGET (*Classe*) and all the other descriptors as INPUT. The stream diagram is the following.

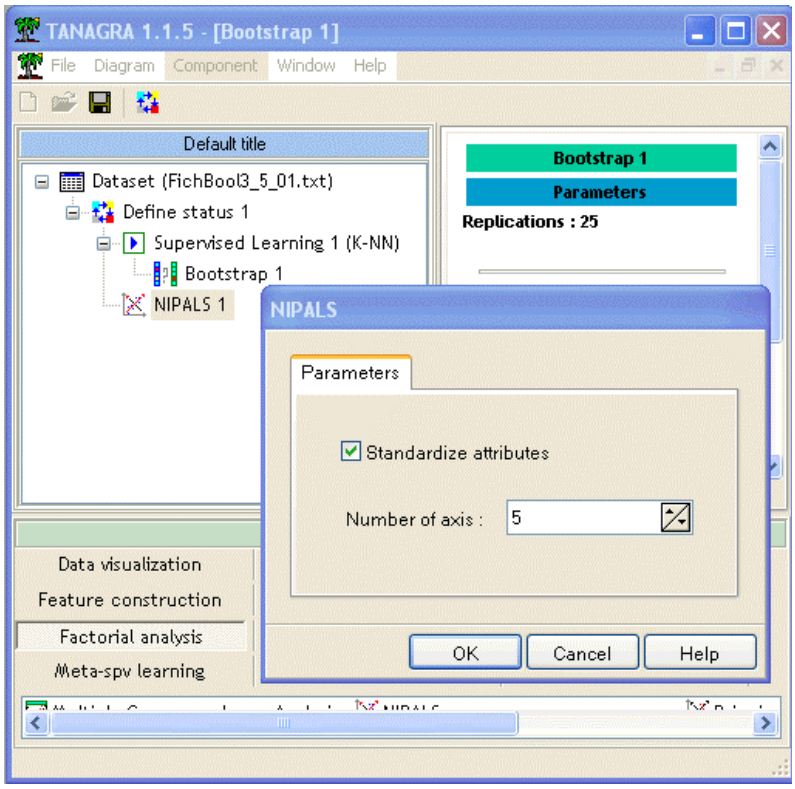


We have used the "Bootstrap plus" error rate measurement (Efron & Tibshirani, 1997). Two main results are available: generalization error rate is **0.2706**; the computing time is **732** sec (PIV – 3 Ghz – 1024 MB RAM).

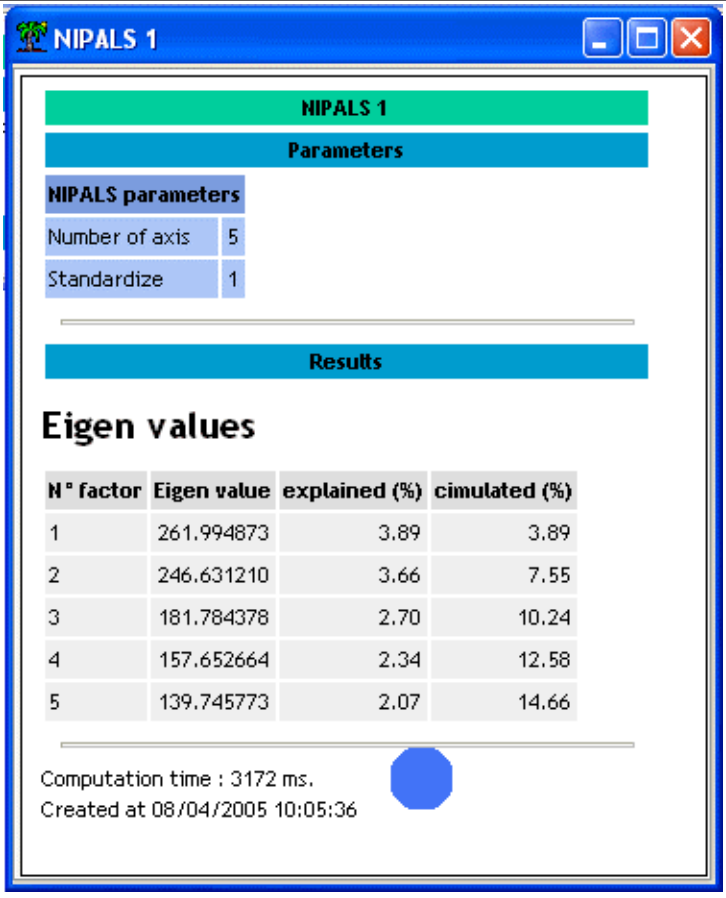


NIPALS

NIPALS calculates the p-first “latent variables “ without a correlation matrix diagonalization. There are two parameters: the number of factor (default value: p = 5); data normalization (default: standardize attributes).

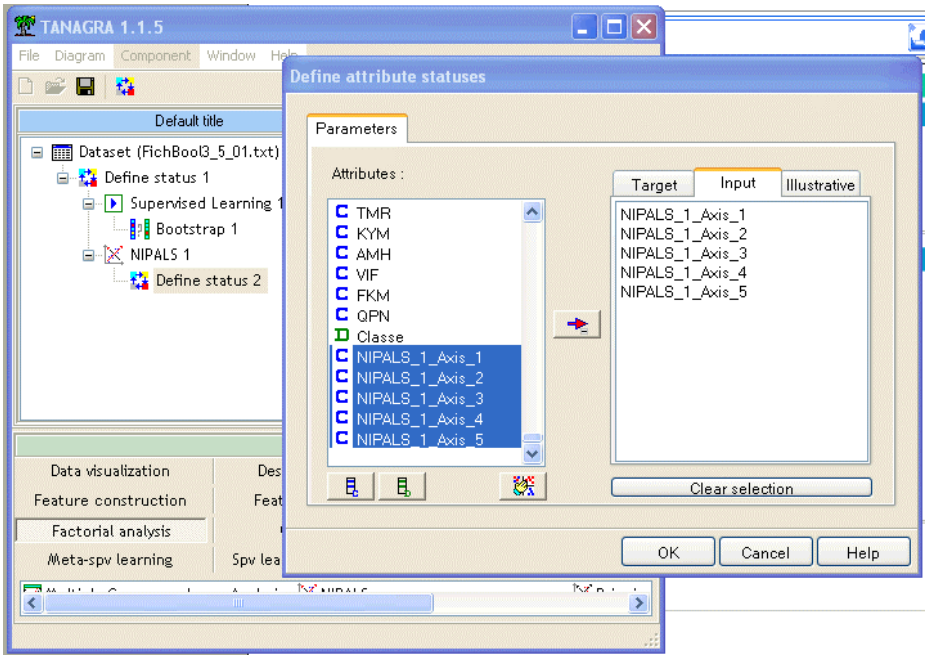


The factors are computed in 3 sec. We did not test the PCA on this dataset, but on all our benchmarks, PCA and NIPALS give very similar results.



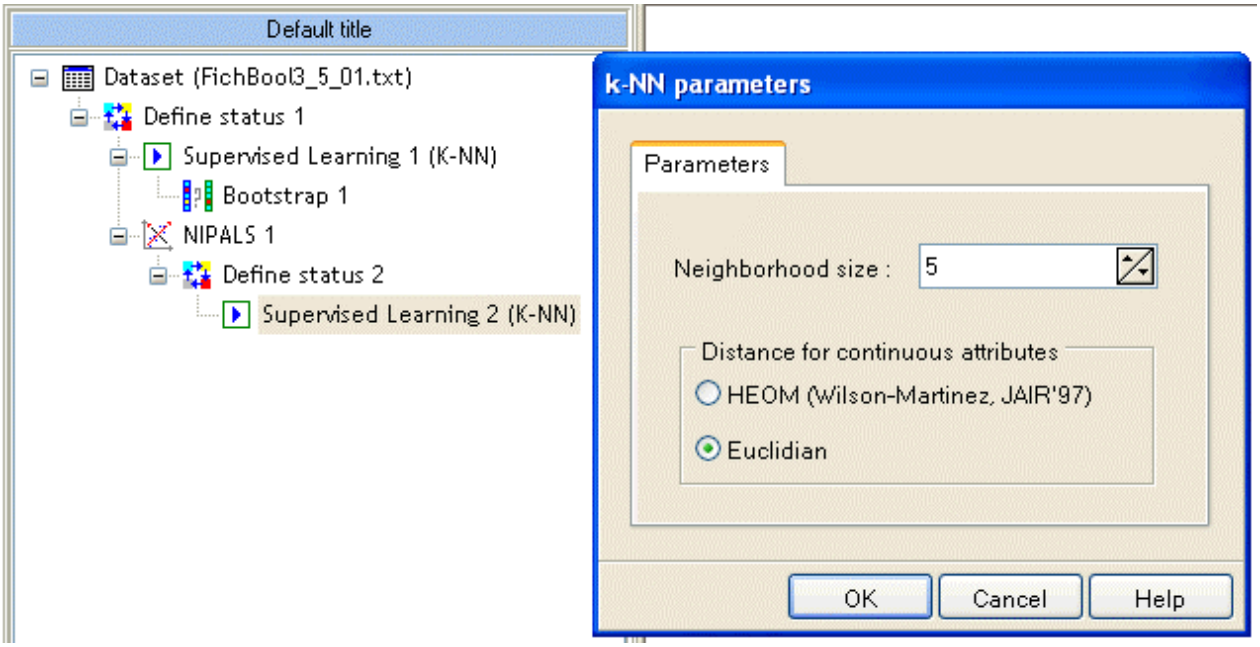
Learning on the reduced space

The next step is to run learning on the factors. Select the right TARGET (*Classe*) and INPUT (*5 factors*) attributes.



Evaluation

Add the supervised learning component. In this case, because the factors are weighted, we have to use the standard Euclidian distance for the nearest neighbor algorithm.



The bootstrap component enables us to evaluate the whole process (NIPALS + K-NN).

We see that the dimensionality reduction improves the error rate (**0.1342**) and reduces dramatically the computing time (**106** sec.).