

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

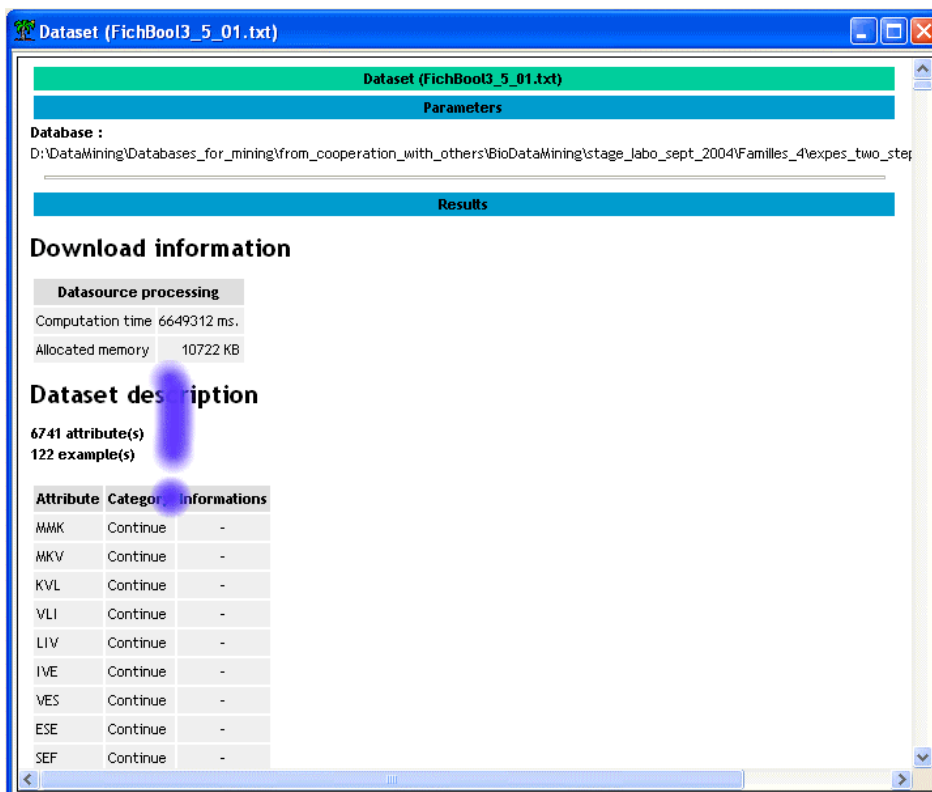
Proteins classification from their primary structures (Mhamdi et al., 2004).

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.



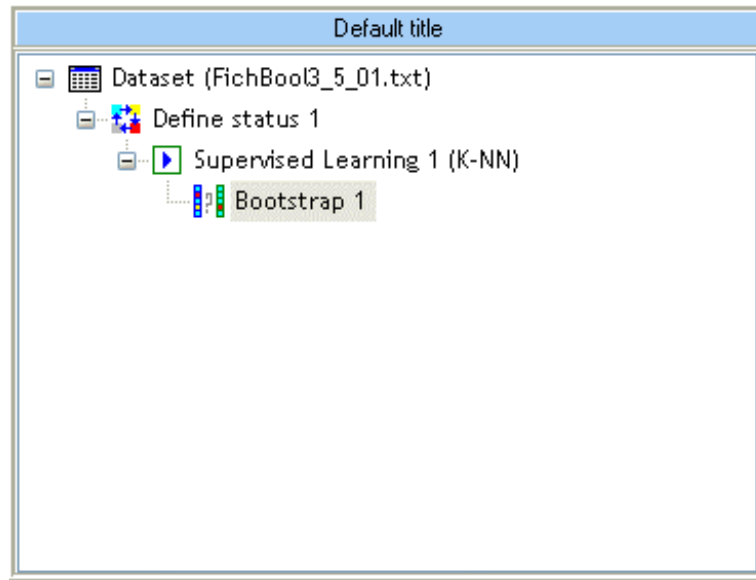
The screenshot shows the TANAGRA software interface for a dataset named 'FichBool3_5_01.txt'. The interface is divided into several sections:

- Dataset (FichBool3_5_01.txt)**: The main title of the window.
- Parameters**: A section for configuring the analysis.
- Database**: Shows the file path: 'D:\DataMining\Databases_for_mining\from_cooperation_with_others\BioDataMining\stage_labo_sept_2004\Familles_4\expes_two_step'.
- Results**: A section for displaying the output of the analysis.
- Download information**:
 - Datasource processing**:
 - Computation time: 6649312 ms.
 - Allocated memory: 10722 KB.
- Dataset description**:
 - 6741 attribute(s)
 - 122 example(s)
- Attribute Information Table**:

Attribute	Category	Information
MMK	Continue	-
MKV	Continue	-
KVL	Continue	-
VLI	Continue	-
LIV	Continue	-
IVE	Continue	-
VES	Continue	-
ESE	Continue	-
SEF	Continue	-

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).

Default title

- Dataset (FichBool3_5_01.txt)
- Define status 1
- Supervised Learning 1 (K-NN)
- Bootstrap 1

Bootstrap 1

Parameters

Replications : 25

Results

Bootstrap error estimation

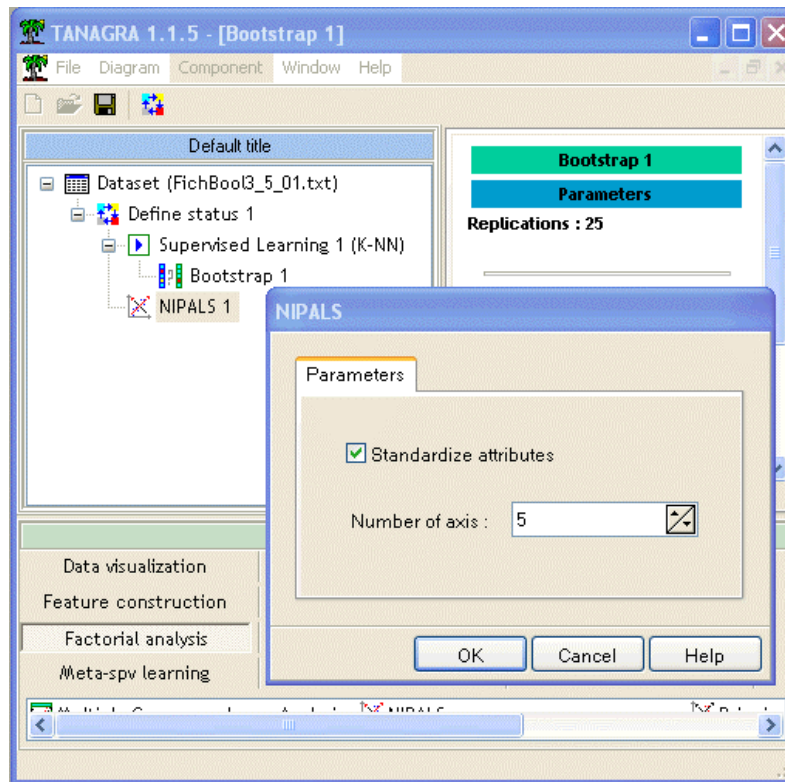
Error rate	
.632+ bootstrap	0.2706
.632 bootstrap	0.2130
Resubstitution	0.0246
Avg test set	0.3226

Computation time : 732609 ms.

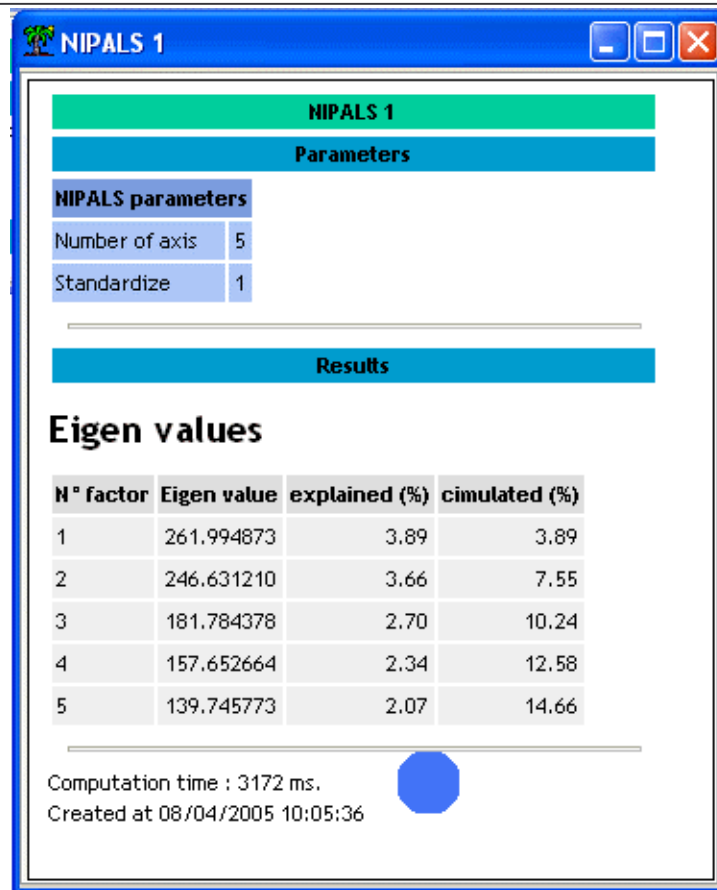
Created at 08/04/2005 10:00:00

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).

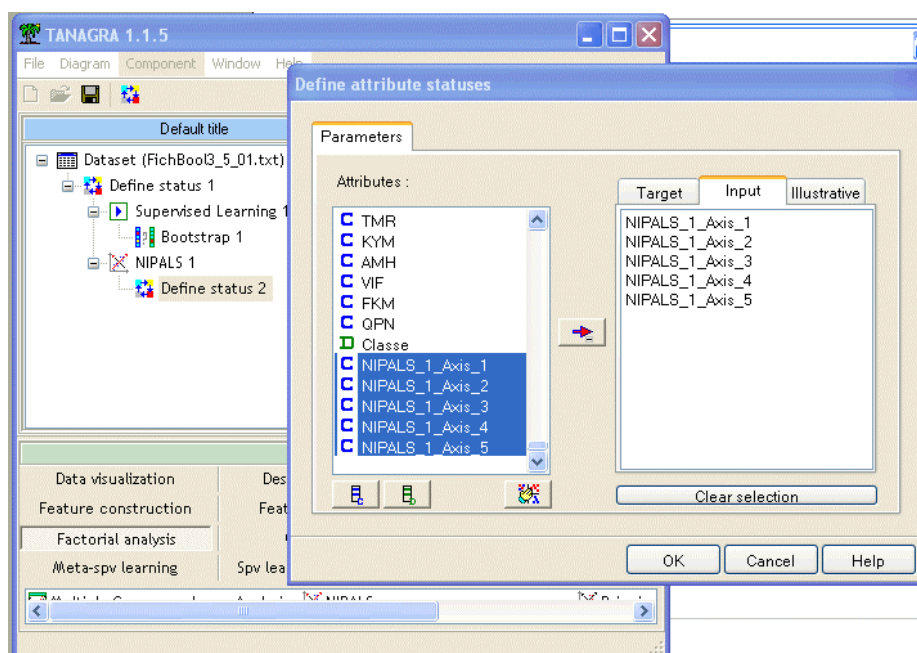


The factors are computed in 3 sec. We did not test the PCA on this dataset, but on all our benchmarks, PCA and NIPALS give very similar results.



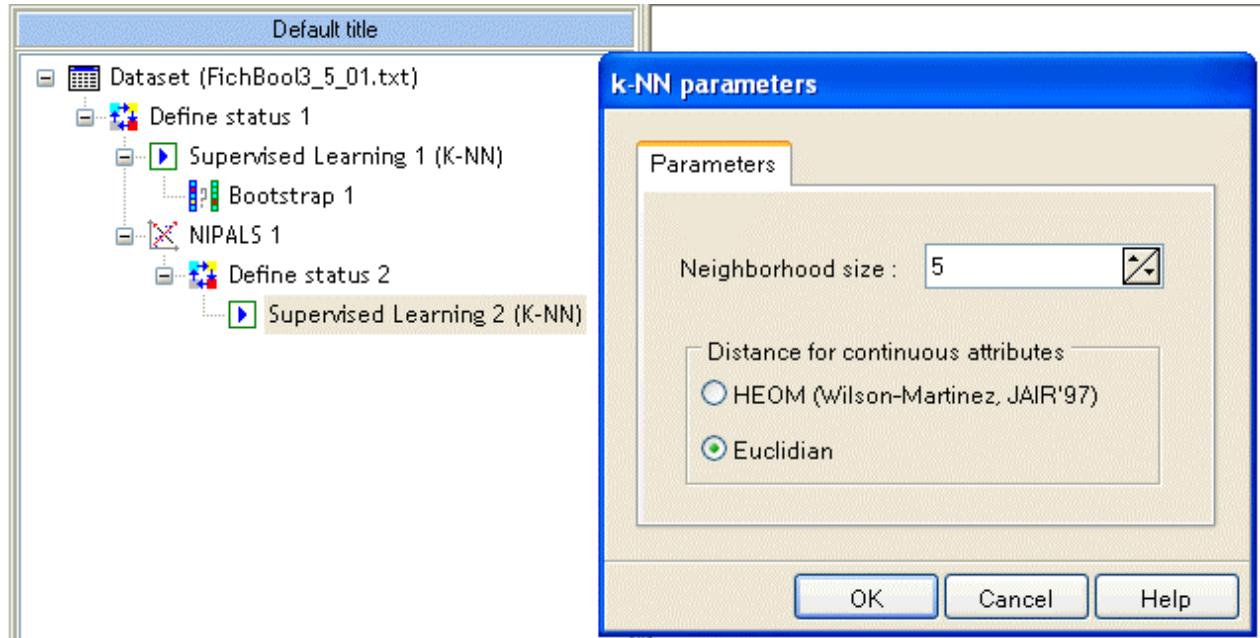
Learning on the reduced space

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.**).

