Objectif

Proposer NIPALS pour réduire la dimensionnalité dans un problème de discrimination de familles de protéines.

NIPALS est une implémentation possible des techniques de décomposition en valeur singulières (SVD), elle permet de retrouver les axes factoriels de l'analyse en composantes principales (ACP) avec une amélioration considérable des temps de calculs lorsque le nombre de variables est très élevé.

Fichier

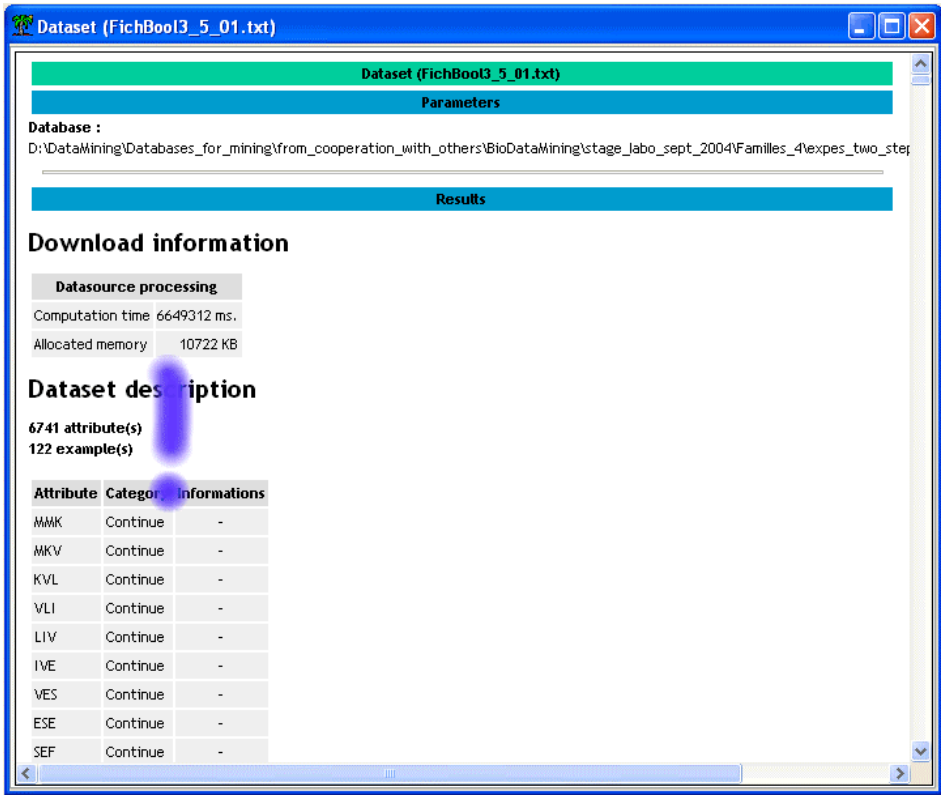
Un fichier issu de la discrimination de protéines à partir de leurs structures primaires (Mhamdi et al., 2004).

Le fichier contient 122 individus répartis en deux familles de protéines {C1, C2}, il y a 6740 descripteurs booléens (1/0) qui correspondent à la présence/absence de 3-grams extraits de la description « brute » des données.

NIPALS

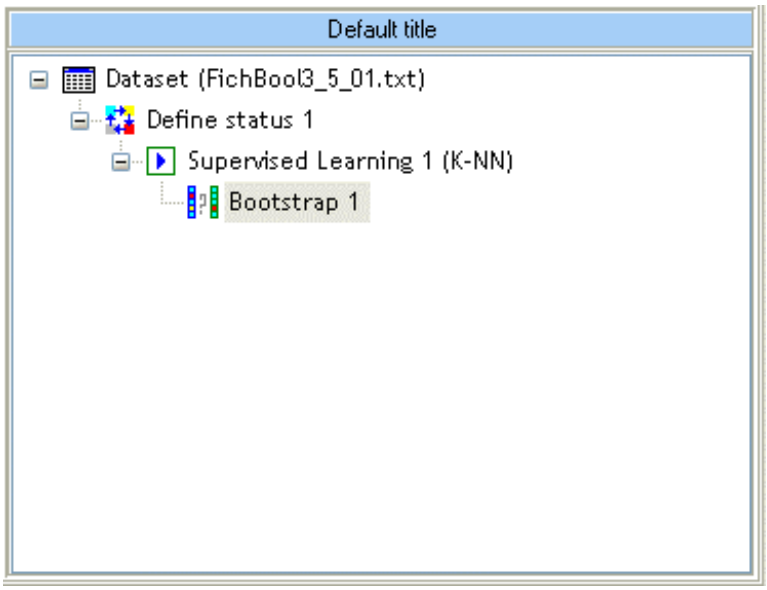
Charger le fichier de données

Charger le fichier TANAGRA_NIPALS.BDM.

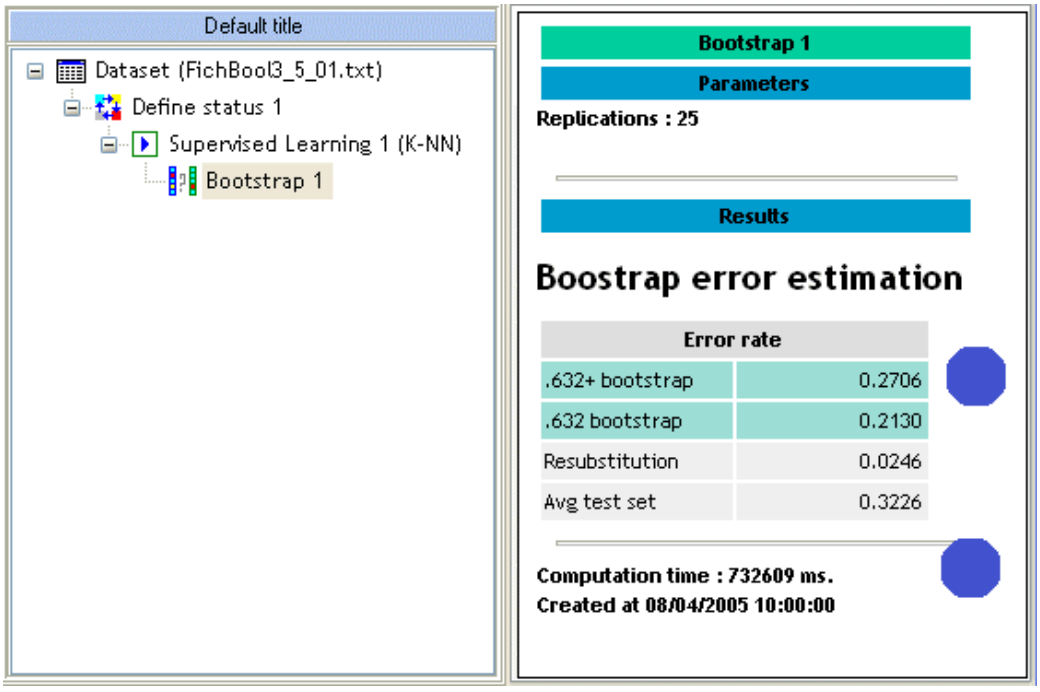


Apprentissage supervisé

Essayons de produire et d'évaluer un classifieur à l'aide de la méthode des K-PPV (K- plus proches voisins). Il faut pour ce faire : sélectionner les attributs TARGET (*Classe*) et INPUT (*tous les autres descripteurs*), placer les composants d'apprentissage et d'évaluation. Le diagramme de traitements est le suivant.

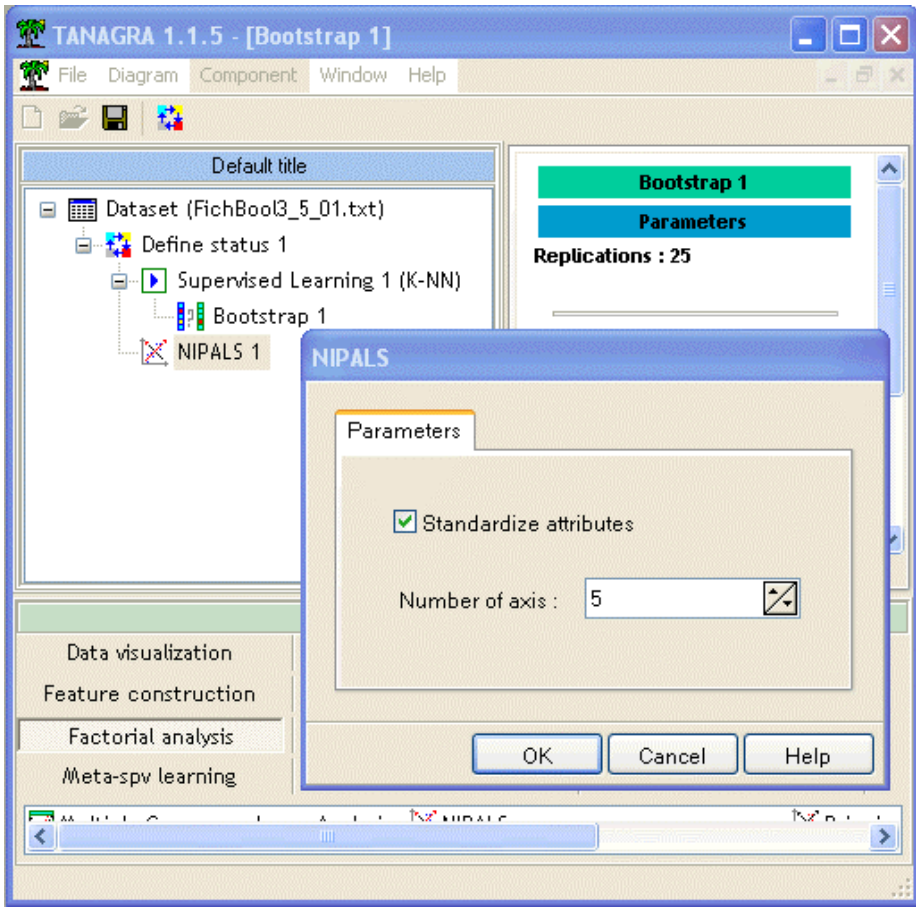


Nous avons utilisé la méthode « Bootstrap plus » (Efron & Tibshirani, 1997) pour évaluer l'erreur en généralisation. Deux résultats retiennent notre attention : le taux d'erreur estimé est de **0.2706** ; le temps de calcul du processus complet est de **732** secondes (PIV – 3 Ghz – 1024 MB RAM).

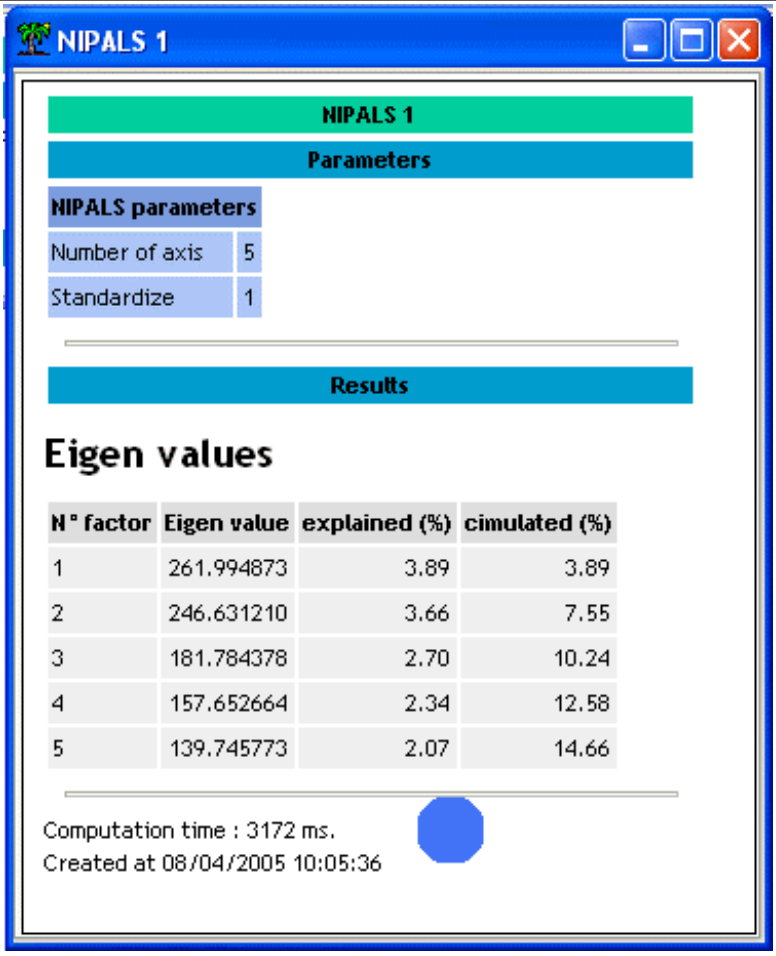


NIPALS

NIPALS produit les p-premiers axes factoriels de l'ACP sans avoir à diagonaliser explicitement la matrice de variance co-variance. L'encombrement mémoire est ainsi moindre et le temps de calcul réduit. NIPALS prend deux paramètres : le nombre de facteurs à construire (p = 5 par défaut) et le type de normalisation des données (centrées et réduites par défaut).

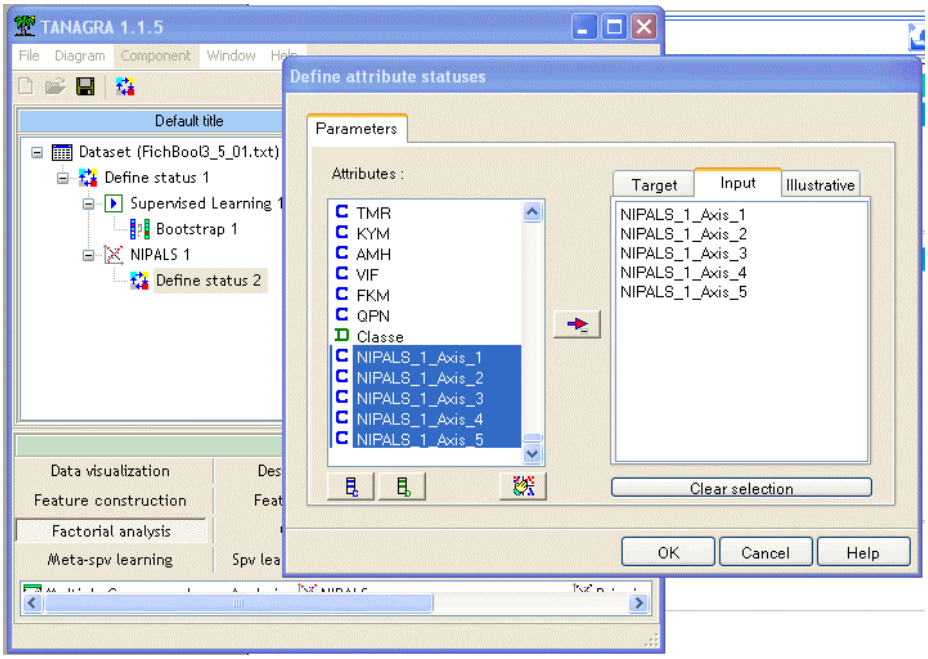


Notons que 3 secondes ont suffi pour produire les facteurs. Les résultats fournissent les valeurs propres des 5 premiers facteurs et le pourcentage d'inertie associé. Nous n'avons pas testé l'ACP sur ce fichier mais elle a fourni des valeurs très similaires à NIPALS sur tous les autres fichiers de tests que nous avons utilisés.



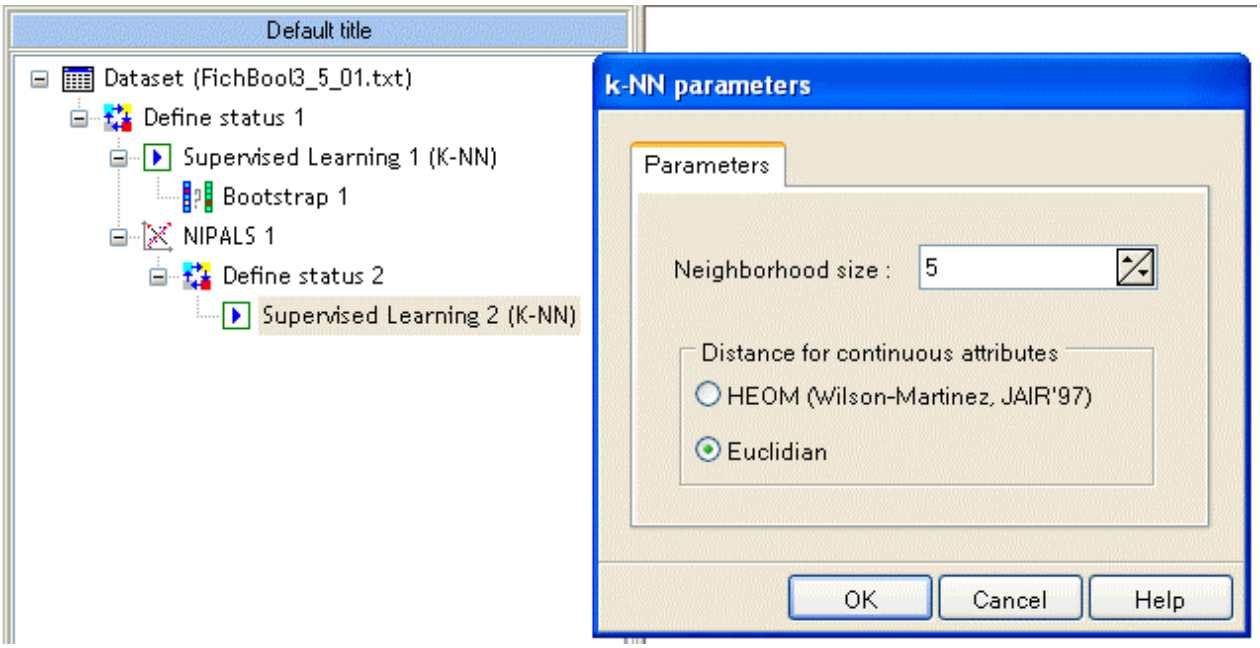
Apprentissage sur l'espace réduit

L'étape suivante consiste alors à lancer l'apprentissage sur les axes factoriels, il faut passer par une phase de sélection des attributs TARGET (*classe*) et INPUT (*les 5 facteurs*).



Evaluation

Il faut alors placer le composant d'apprentissage supervisé. Attention, dans ce cas, les facteurs étant pondérés, il faut veiller à utiliser une distance non normalisée pour les K-plus proches voisins. Le paramétrage du composant doit être comme suit :



Le composant Bootstrap permet de mesurer comme précédemment les performances de l'ensemble (NIPALS + K-PPV). Nous constatons que la réduction de la dimensionnalité divise par 2 le taux d'erreur (**0.1342**), le temps de calcul de tout le processus a été divisé par 7 (**106** secondes).