Objectif

Construire un modèle de prédiction à l'aide du bayesien naïf (NAIVE BAYES).

La méthode implémentée dans TANAGRA n'accepte que les descripteurs discrets. Or nos variables sont continues, il est nécessaire de les discrétiser avant de lancer les calculs.

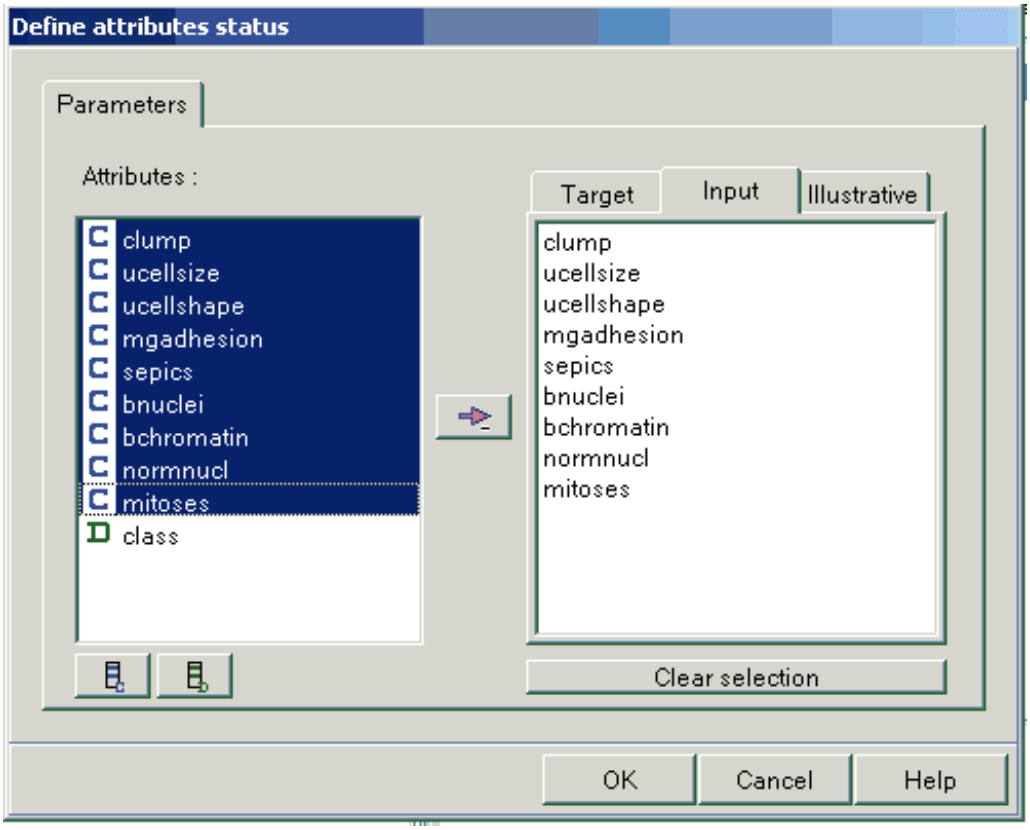
A partir du moment où nous sommes en train de traiter un problème d'apprentissage supervisé, il est naturel d'utiliser une méthode de discrétisation qui tienne compte de la variable à prédire. Dans ce contexte, nous utiliserons la méthode MDLPC (Fayyad et Irani, 1993) qui est la plus connue (la plus citée) en apprentissage automatique.

Fichier

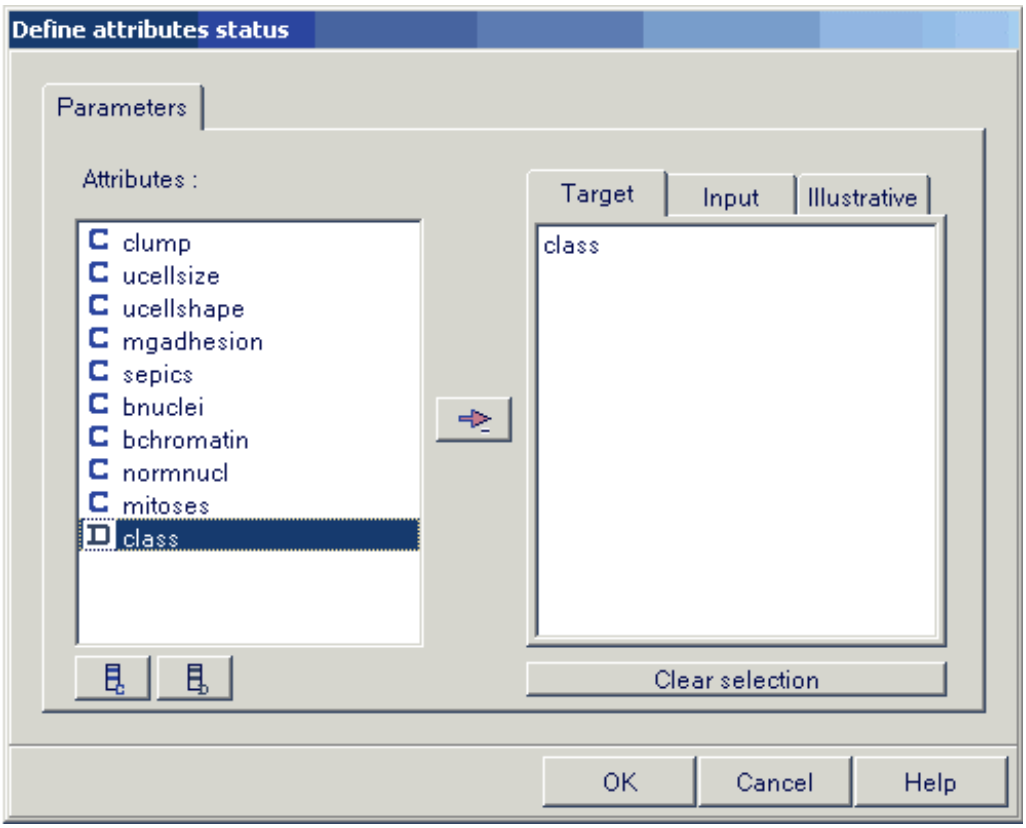
Nous utiliserons le fichier BREAST, la variable à prédire est CLASS (tumeur maligne ou bénigne), les descripteurs correspondent à des caractéristiques physiologiques des cellules.

Discrétiser pour l'apprentissage supervisé

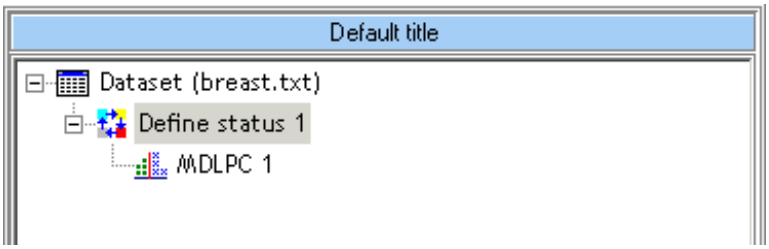
1. Charger le fichier BREAST.BDM
2. Il faut définir le statut des variables à discrétiser, insérer le composant « Define Status » et
    - Définissez comme INPUT les variables prédictives continues



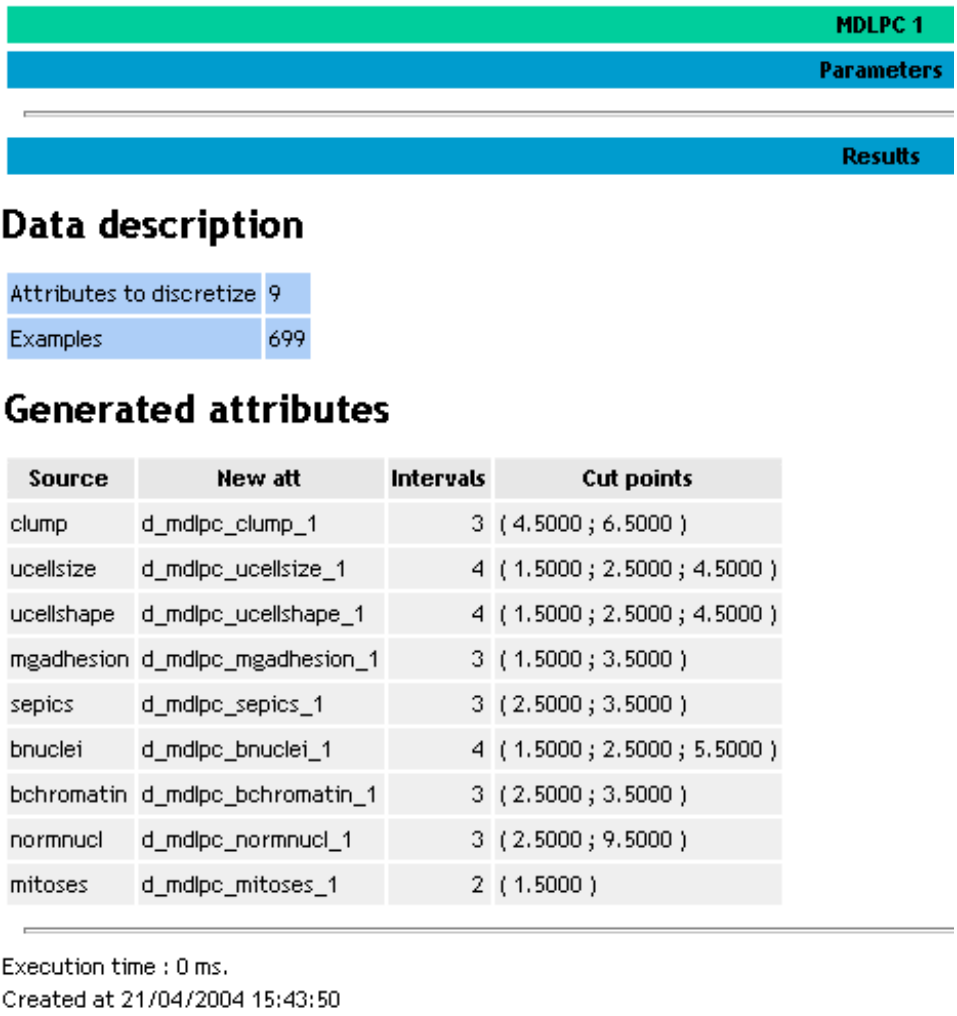
- Définissez comme TARGET la variable à prédire CLASS



3. Placez à la suite le composant MDLPC (Feature Construction). Attention, ce composant ne fonctionne que s'il y a une seule variable TARGET discrète, et une ou plusieurs variables continues INPUT. Ce qui est le cas pour nous.



4. ***Vous devez alors lancer l'exécution pour rendre effectif la génération des nouvelles variables discrétisées*** (Cliquez sur le menu contextuel View de MDLPC). Les noms des variables générées et les bornes de discrétisation qui ont été calculées sont affichés.



MDLPC 1

Parameters

Results

Data description

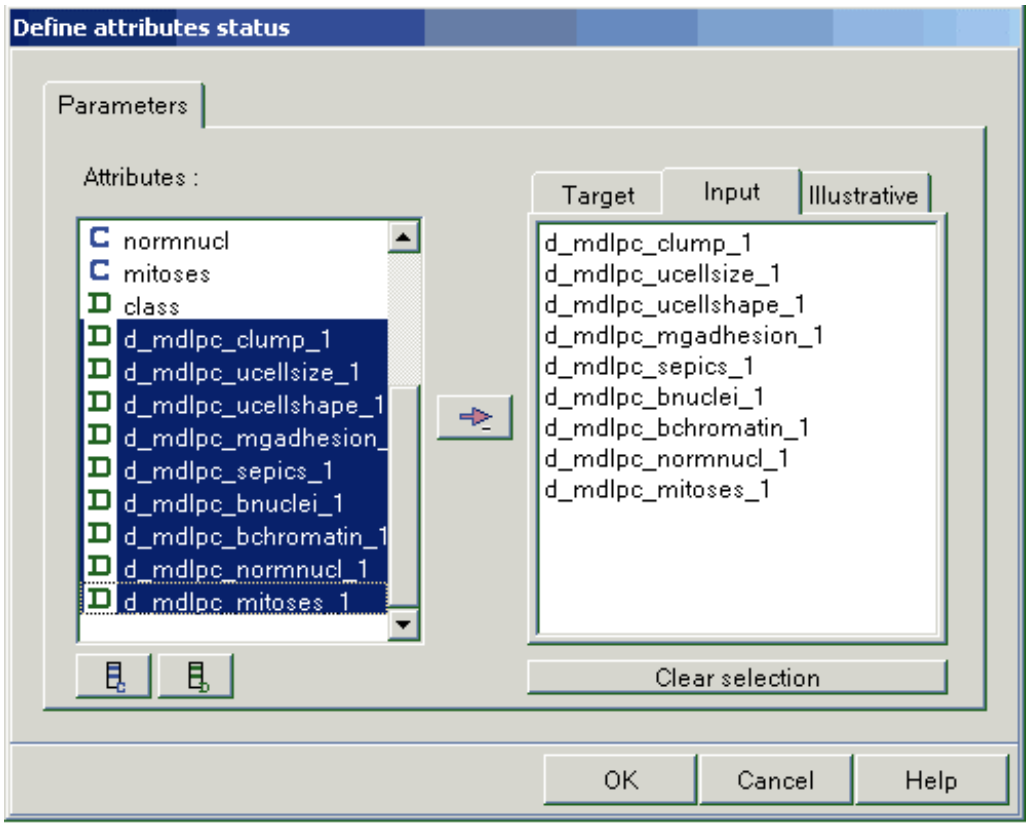
Attributes to discretize	9
Examples	699

Generated attributes

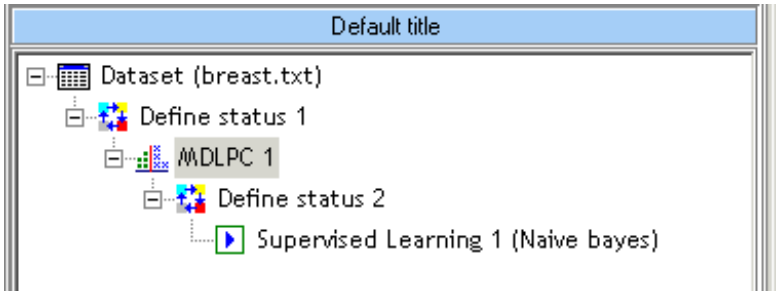
Source	New att	Intervals	Cut points
clump	d_mdlpc_clump_1	3	(4.5000 ; 6.5000)
ucellsize	d_mdlpc_ucellsize_1	4	(1.5000 ; 2.5000 ; 4.5000)
ucellshape	d_mdlpc_ucellshape_1	4	(1.5000 ; 2.5000 ; 4.5000)
mgadhesion	d_mdlpc_mgadhesion_1	3	(1.5000 ; 3.5000)
sepics	d_mdlpc_sepics_1	3	(2.5000 ; 3.5000)
bnuclei	d_mdlpc_bnuclei_1	4	(1.5000 ; 2.5000 ; 5.5000)
bchromatin	d_mdlpc_bchromatin_1	3	(2.5000 ; 3.5000)
normnucl	d_mdlpc_normnucl_1	3	(2.5000 ; 9.5000)
mitoses	d_mdlpc_mitoses_1	2	(1.5000)

Execution time : 0 ms.
Created at 21/04/2004 15:43:50

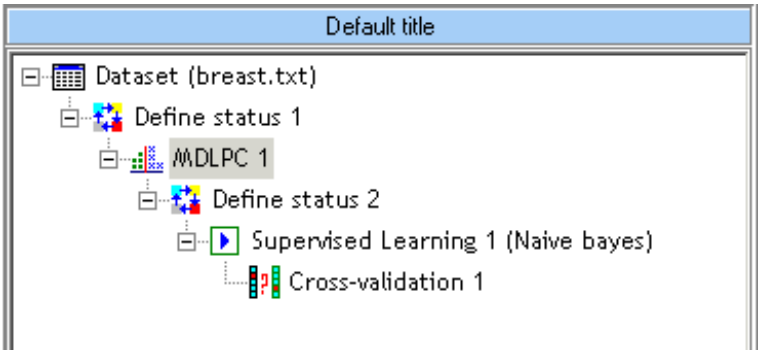
5. Pour lancer l'apprentissage avec le NAIVE BAYES, il faut de nouveau placer le composant « Define Status » en définissant toujours comme TARGET la variable CLASS, et en définissant comme INPUT les variables discrétisées.



6. Nous pouvons dès lors appliquer sur ces variables l'apprentissage via le modèle bayesien naïf en plaçant successivement le composant « **Supervised Learning** » (Palette *Meta-spv Learning*) puis en y intégrant le composant « **Naive Bayes** » (Palette *Spv Learning*). Le diagramme est le suivant :



7. ***Le taux d'erreur en resubsitution est de 0.0272***. Pour obtenir une estimation moins biaisée des performances de l'algorithme, nous intégrons enfin le composant « Cross Validation » (Palette *Spv Learning assessment*) en laissant les paramètres par défaut.



8. Les résultats sont très bons, largement comparables aux autres méthodes supervisées !

Cross-validation 1

Parameters

Cross-validation parameters	
Folds	2
Trials	5

Results

CV error rate

Trial	Err rate
1	0.0258
2	0.0272
3	0.0287
4	0.0315
5	0.0330

Overral cross-validation error rate

Error rate	0.0292

Values prediction

Value	Sensibility	Pred. error
begnin	0.9646	0.0094
malignant	0.9826	0.0641

Confusion matrix

	begnin	malignant	Sum
begnin	2205	81	2286
malignant	21	1183	1204
Sum	2226	1264	3490

Execution time : 1152 ms.
Created at 21/04/2004 16:05:04

NB : Dans la validation croisée, c'est bien tout le chemin qui est exécuté à chaque session d'apprentissage sur les sous-ensembles de données, cela est vrai notamment pour le calcul des bornes de discrétisation. Le taux d'erreur calculé correspond aux performances de toute la filière : la discrétisation « et » l'apprentissage avec le modèle bayesien naïf.