Subject

Build a naive bayes classifier on continuous descriptors.

Naive bayes classifier implemented into TANAGRA handles only discrete attributes, we need to discretize continuous descriptors before use them.

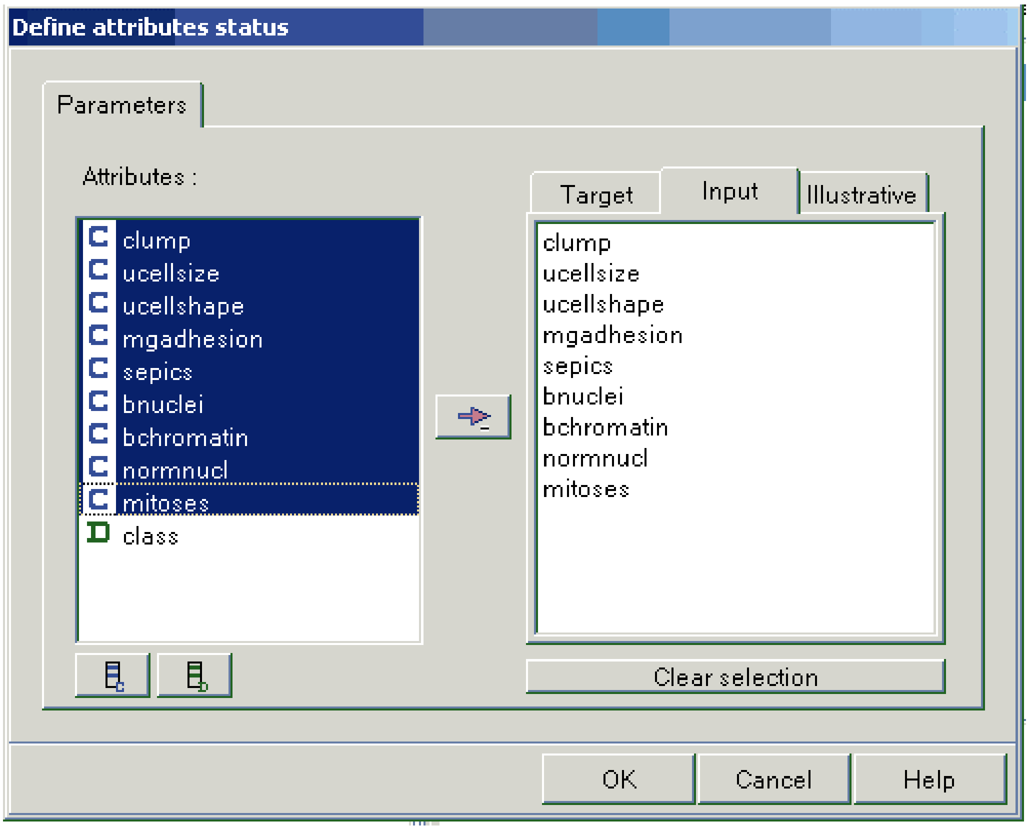
Because we are studying a supervised learning method, we must use a supervised discretization algorithm such as Fayyad and Irani's state-of-the-art MDLPC algorithm.

Dataset

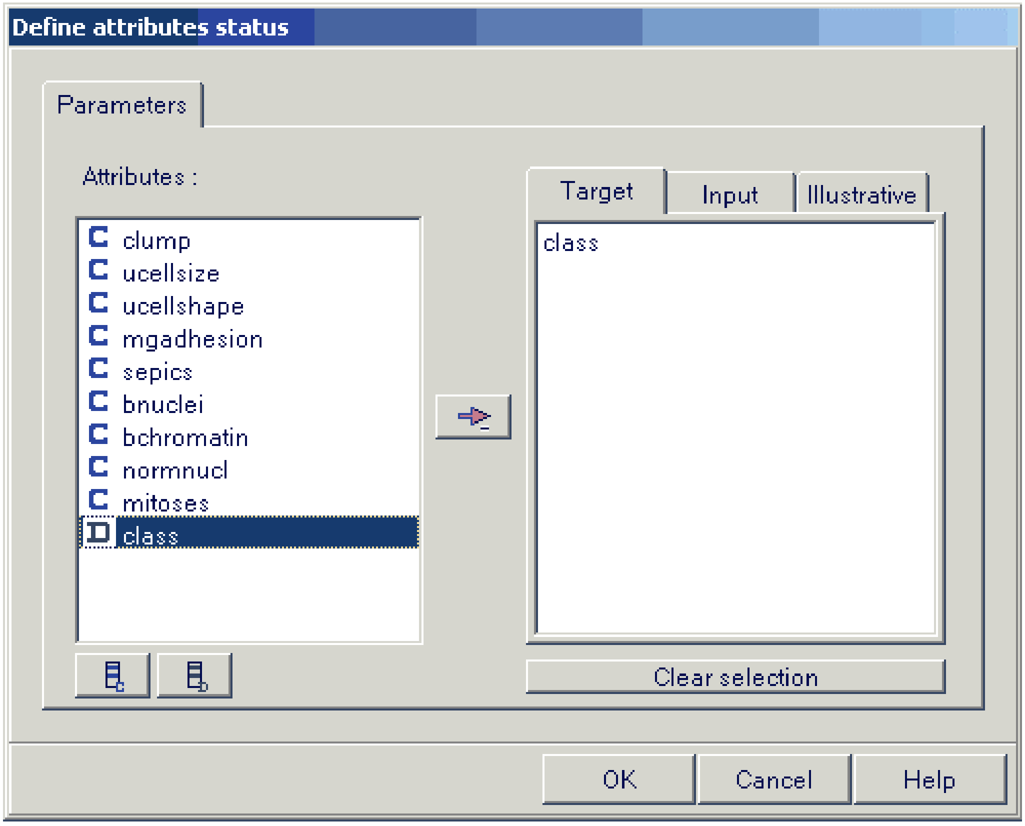
BREAST Cancer dataset. Class attribute is CLASS (malignant or benign tumor), descriptors are cells characteristics.

Discretization for supervised learning

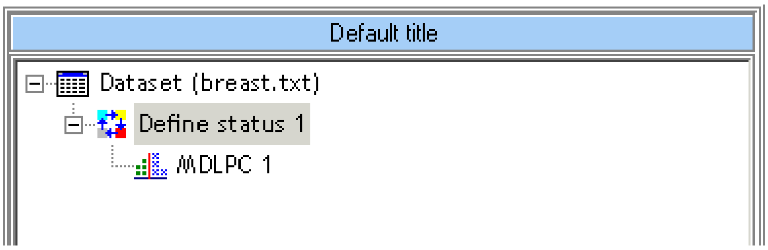
1. Load BREAST.BDM
2. You must select contextual attribute (class attribute) and continuous attributes to discretize.
    - Set continuous attributes as INPUT



    - Set CLASS as TARGET



3. Insert MDLPC component into the diagram. Be careful, this component works only if all inputs are continuous, and there is only one target discrete attribute.



4. ***You must execute the component to obtain discretized attributes.*** (Click on the VIEW menu, the cut points are displayed on the results frame).

MDLPC 1

Parameters

Results

Data description

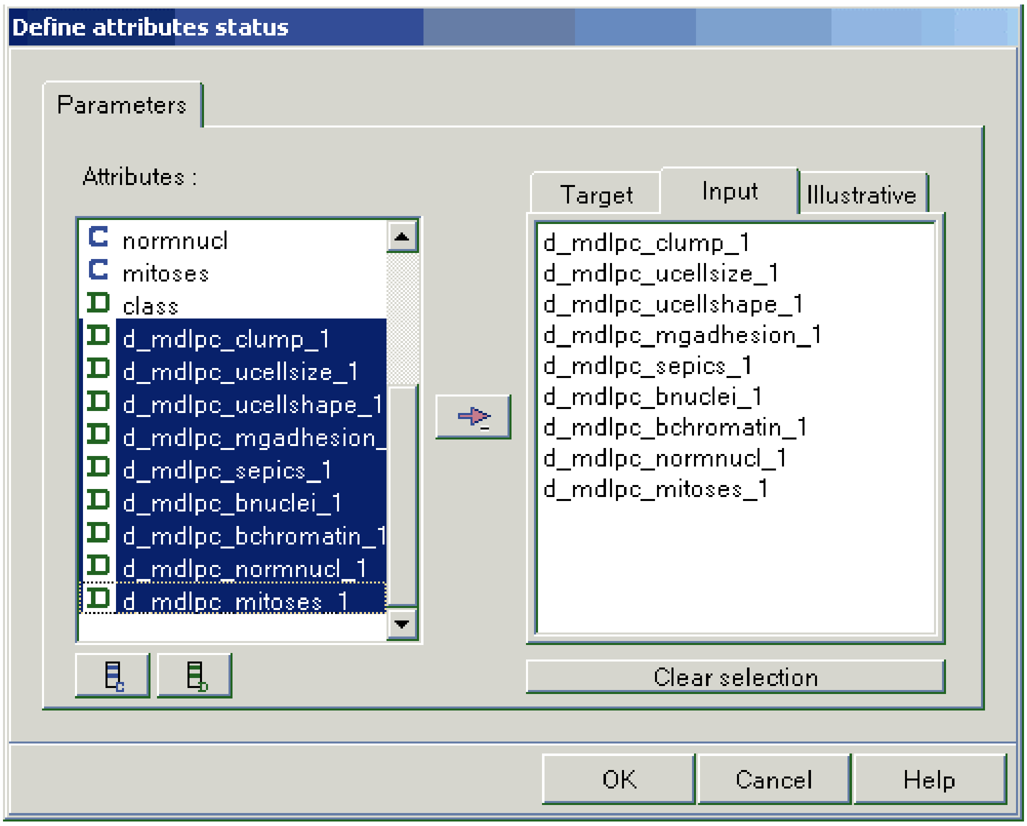
Attributes to discretize	9
Examples	699

Generated attributes

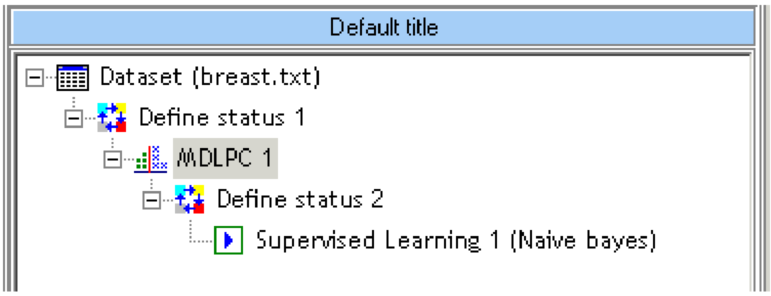
Source	New att	Intervals	Cut points
clump	d_mdlpc_clump_1	3	(4.5000 ; 6.5000)
ucellsize	d_mdlpc_ucellsize_1	4	(1.5000 ; 2.5000 ; 4.5000)
ucellshape	d_mdlpc_ucellshape_1	4	(1.5000 ; 2.5000 ; 4.5000)
mgadhesion	d_mdlpc_mgadhesion_1	3	(1.5000 ; 3.5000)
sepics	d_mdlpc_sepics_1	3	(2.5000 ; 3.5000)
bnuclei	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

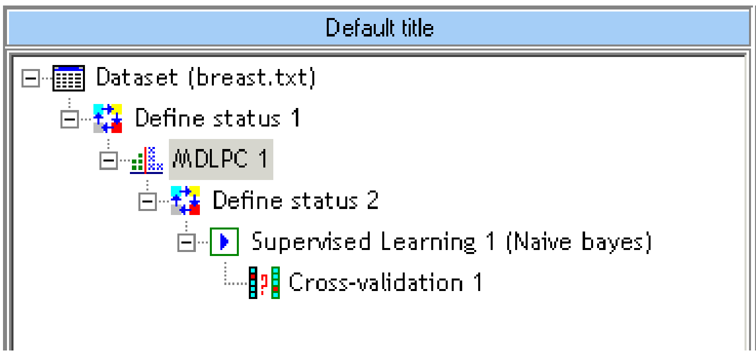
5. To run NAIVE BAYES, insert a new "Define status" component. Set CLASS as TARGET, but the new INPUT attributes are the discretized one.



6. You can now define the learning procedure with naive bayes classifier. Insert the « **Supervised Learning** » component (from *Meta-spv Learning*) in which you incorporate the « **Naive Bayes** » component (from *Spv Learning*). The data mining diagram is the following:



7. ***Resubstitution error rate is 0.0272***. To obtain a less biased error rate evaluation, we add a « Cross Validation » component (from *Spv Learning assessment*) with default parameters.



8. The results are comparable with those obtained with other supervised learning algorithm!

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: In the cross-validation, this is the path from the root of the diagram to the cross-validation component which is running for each learning session, especially the discretization. Then, the computed error rate corresponds to the performance of the whole process: the discretization "and" the naive bayes learning.