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

Define attributes status Parameters Attributes : C clump C ucellsize C ucellshape C mgadhesion C sepics C bnuclei C bchromatin C normnucl C mitoses C class	Þ	Target clump ucellsize ucellshape mgadhesio sepics bnuclei bchromatin normnucl mitoses	Input n	Illustrative	
	[с	lear select	ion	
		ОК	Can	cel H	lelp

• Set CLASS as TARGET

Tutorial Discretization for supervised learning

Define attribute <mark>s status</mark>				
Parameters Attributes : C clump C ucellsize C ucellshape C mgadhesion C sepics C bnuclei C bchromatin C normnucl C mitoses C class	->	Target class	Input Illus	trative
E. E.		Cle	ear selection	
		ОК	Cancel	Help

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.

Default	title
et (breast.txt)	
fine status 1	
MDLPC 1	
	Default et (breast.txt) efine status 1 MDLPC 1

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



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_mdipc_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.

Tutorial Discretization for supervised learning

Define attributes status Parameters Attributes : C normnucl mitoses class d_mdlpc_ucellsize_1 d_mdlpc_ucellshape_1 d_mdlpc_sepics_1	Target Input Illustrative d_mdlpc_clump_1 d_mdlpc_ucellsize_1 d_mdlpc_ucellshape_1 d_mdlpc_mgadhesion_1 d_mdlpc_sepics_1 d_mdlpc_bnuclei_1 d_mdlpc_bchromatin_1 d_mdlpc_normnucl_1
□ d_mdlpc_bnuclei_1 □ d_mdlpc_bchromatin_1 □ d_mdlpc_normnucl_1 □ d_mdlpc_mitoses 1 ■ d_mdlpc_mitoses 1	d_mdlpc_mitoses_1 Clear selection
	OK Cancel Help

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.

Default title
⊡-∰ Dataset (breast.txt)
🗄 🙀 Define status 1
🗄 🚠 Define status 2
🗄 🕞 Supervised Learning 1 (Naive bayes)
Cross-validation 1

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

						Cre	oss-validati
							Parameter
Cross	-valida	tion param	eters				
Folds			2				
Trials			5				
	_						Results
∶∨ ег	ror rate	e					
Trial	Err rat	e					
1	0.025	8					
2	0.027	2					
3	0.028	7					
3 4	0.028	5			- 14-1 - 18		
3 4 5	0.028 0.031 0.033	7 5 0			a an		
3 4 5 verr	0.028 0.031 0.033 al cros	7 5 0 s-validation	n error rate		e and a second		
3 4 5 9 verr	0.028 0.031 0.033 al cros	7 5 0 s-validation Error rate	n error rate		0.0	1292	
3 4 5)verr	0.028 0.031 0.033 al cros Val	7 5 90 s-validation Error rate ues predic	n error rate tion		0.0 sfusio	1292 on matrix	
3 4 5 Dverr Va	0.028 0.031 0.033 al cros Val	7 5 •0 •••••••••••••••••••••••••••••••••	n error rate tion Pred. error		0.0 Situsio Dega	1292 on matrix malige	Sum
3 4 5 Dverr Va be;	0.028 0.031 0.033 al cros Val	7 5 s-validation Error rate ues predic Sensibility 0.9646	tion Pred. error 0.0094	begnin	0.0 .::ifusio begin: 2260	1292 on matriz malig**** 7	Sum 2286
3 5 Overr Va be; mali;	0.028 0.031 0.033 al cros Val lue	7 5 s-validation Error rate ues predic Sensibility 0.9646 0.9826	tion Pred. error 0.0094 0.0641	begnin malignant	0.0 .:sfusio began 2201 21	1292 on matrix malig**** 7 81 1183	Sum 2286 1204

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 crossvalidation 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.