Subject

To show the impact of feature selection on the naive bayes classifier.

Dataset

The well-known Fisher's IRIS dataset, its main interest is that we know the "right" feature subset.

Feature selection for naive bayes classifier

- 1. Download the IRIS.BDM dataset.
- 2. Define, with a "Define status": TYPE as TARGET; SEP_LENGTH, SEP_WIDTH, PET_LENGTH, PET_WIDTH, as INPUT.
- 3. Insert the MDLPC supervised discretization. We have the following diagram:



4. Insert again "Define Status" component; set TYPE as TARGET and the new discretized features as INPUT.

Define attributes status				
Define attributes status Parameters Attributes : C sep_length C sep_width C pet_length C pet_width Dype D d_mdlpc_sep_length_1 D d_mdlpc_pet_length_1 D d_mdlpc_pet_width_1	Target d_mdlpc_s d_mdlpc_i d_mdlpc_f	Input sep_length, sep_width_ set_length, set_width_1	Illustrativ	
E. E.		Clear select	ion	
	OK	Cano	cel	Help

5. We can insert now the naive bayes algorithm, and assess this one with a cross-validation.



The cross-validation error rate is

						Cross-validat	tion 1	
						Paramete	rs	
Cross-va	alidatio	n parameter	5					
Folds			2					
Trials			5					
						Results		
V error	r rate							
Rang	ge							
MIN 0	0.0533							
MAX 0	0.0933							
Trial Er	r rate							
1 0	0.0667							
2 0	0.0933							
3 0	0.0600							
4 0	0.0533							
5 0	0.0733						No.	
)verral (CLOSS-A	alidation eri	ror rate					
	I	Error rate				0.0693	1910 - 1910 - 1910 - 1910 - 1910 - 1910 - 1910 - 1910 - 1910 - 1910 - 1910 - 1910 - 1910 - 1910 - 1910 - 1910 -	
	Valu	ies predictio	n		C	onfusion matrix	A STATE	
Val	lue	Sensibility	Pred. error		Iris-setosa	Iris-versicolor	Iris-virginica	Sum
Iris-se	etosa	0.9840	0.0000	lris-setosa	246	2	2	2 50
	sicolor	0.8840	0.0868	Iris-versi color	0	221	29	2 50
lris-ver			And the second se		THE REAL PROPERTY AND INCOMES INCOMESTA INCOME	Contraction of the second	004	0.50
lris-ver Iris-vir	ginica	0.9240	0.1183	lris-virginica	U is a second	19	231	2.50

6. The idea now is to determine if it is possible to select a subset of the input attributes which would make it possible to obtain the same (even to improve) the error rate. We will use MIFS (Battiti and AI, 1994). We thus insert it after the component "Define Status 2", MIFS will have to filter the descriptors by selecting those which are most relevant for the supervised learning.



The results show that only the discretized PET_LENGTH and PET_WIDTH attributes are relevant.

MIFS para	ameters
Beta	1.50

INPUT attribute selection

INPUT selection					
Before filtering	4				
After filtering	2				

Keeped into INPUT selection

	Attributes
1	d_mdlpc_pet_length_1
2	d_mdlpc_pet_width_1

Removed from INPUT selection

	Attributes
1	d_mdlpc_sep_length_1
2	d_mdlpc_sep_width_1

Calculations details

 Selected attribute
 I(Y,X/S)

 d_mdlpc_pet_width_1
 1.378403

 d_mdlpc_pet_length_1
 0.293534

Execution time : 0 ms. Created at 18/05/2004 14:34:45

7. For unbiased evaluation of this feature selection, we must integrate it in the learning process with naïve bayes algorithm. The new diagram is the following,



The performances of the whole process show that this feature selection method is relevant on this dataset.

						Cross-validat	ion 2	
						Parameter	rs	
Cros	s-validatio	n parameter	'S					
Folds			2					
Trials	:		5					
						Results		
CV er	ror rate							
R	ange							
MIN	0.0333							
MAX	0.0667							
Trial	Err rate							
1	0.0333							
2	0.0400						88 AL	
3	0.0667							
4	0.0533							
5	0.0467					799		
Overr	al cross-va	alidation err	or rate					
	I	Error rate				0.0480		
					3°00	nfusion matrix		
	Valu	es predictio	n				Sec. 194	
	Valu Value	es predictio Sensibility	on Pred. error		Iris-setosy	iti. 	triv Eg <mark>inica</mark>	Sum
Iri	Valu Value s-setosa	es predictio Sensibility 1.0000	on Pred. error 0.0000	Iris-setosa	Iris-setos» 250	ter 	tris de <mark>ginica</mark> O	Sum 250
lri Iris-t	Valu Value s-setosa versicolor	es prediction Sensibility 1.0000 0.9200	on Pred. error 0.0000 0.0650	Iris-setosa Iris-versicolor	Iris-setos. 250 0	230	tri⊷ ∴r ginica 0 20	Sum 250 250
lri lris- lris-	Valu Value s-setosa versicolor •virginica	es prediction Sensibility 1.0000 0.9200 0.9360	Pred. error 0.0000 0.0650 0.0787	lris-setosa Iris-versicolor Iris-virginica	Iris-setos 250 0	res Services of the service of the s	Iri	Sum 250 250 250

Execution time : 360 ms. Created at 18/05/2004 14:49:58