Subject

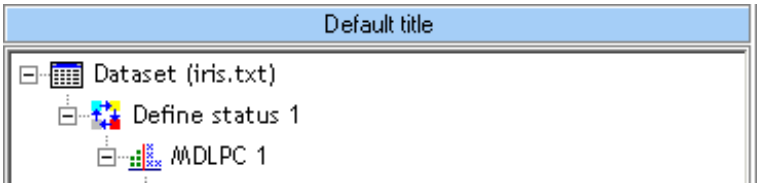
To show the impact of feature selection on the naive bayes classifier.

Dataset

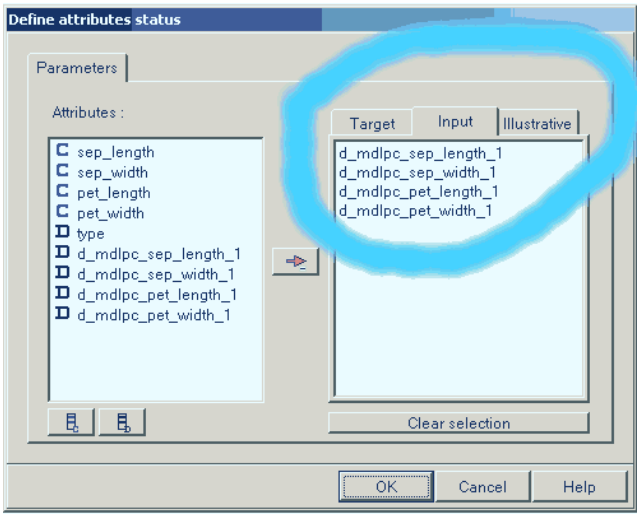
The well-known Fisher's IRIS dataset, its main interest is that we know the "right" feature subset.

Feature selection for naive bayes classifier

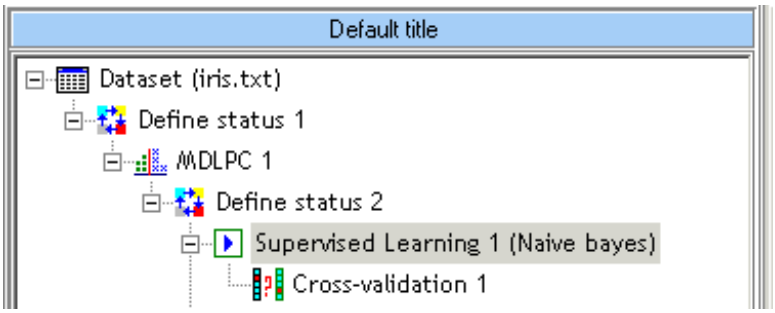
1. Download the IRIS.BDM dataset.
2. Define, with a "Define status": TYPE as TARGET; SEP_LENGTH, SEP_WIDTH, PET_LENGTH, PET_WIDTH, as INPUT.
3. Insert the MDLPC supervised discretization. We have the following diagram:



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



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



The cross-validation error rate is

Cross-validation 1

Parameters

Cross-validation parameters	
Folds	2
Trials	5

Results

CV error rate

Range	
MIN	0.0533
MAX	0.0933

Trial	Err rate
1	0.0667
2	0.0933
3	0.0600
4	0.0533
5	0.0733

Overral cross-validation error rate

Error rate	0.0693

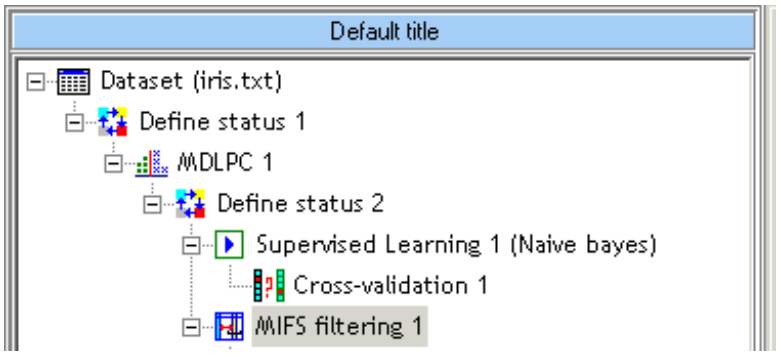
Values prediction

Value	Sensibility	Pred. error
Iris-setosa	0.9840	0.0000
Iris-versicolor	0.8840	0.0868
Iris-virginica	0.9240	0.1183

Confusion matrix

	Iris-setosa	Iris-versicolor	Iris-virginica	Sum
Iris-setosa	246	2	2	250
Iris-versicolor	0	221	29	250
Iris-virginica	0	19	231	250
Sum	246	242	262	750

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 Al, 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 filtering 1

Parameters

MIFS parameters	
Beta	1.50

Results

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

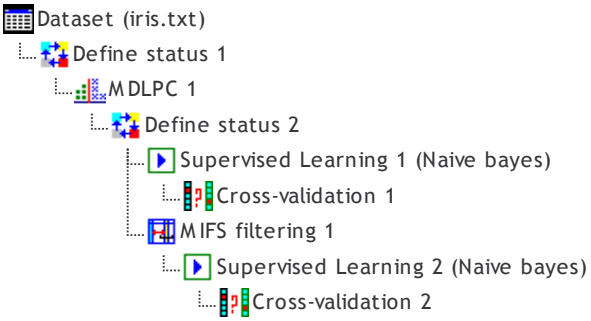
	Attributes
1	d_mdlpc_sep_length_1
2	d_mdlpc_sep_width_1

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-validation 2

Parameters

Cross-validation parameters	
Folds	2
Trials	5

Results

CV error rate

Range	
MIN	0.0333
MAX	0.0667
Trial	**Err rate**
1	0.0333
2	0.0400
3	0.0667
4	0.0533
5	0.0467

Overral cross-validation error rate

Error rate	0.0480

Values prediction

Value	Sensibility	Pred. error
Iris-setosa	1.0000	0.0000
Iris-versicolor	0.9200	0.0650
Iris-virginica	0.9360	0.0787

Confusion matrix

	Iris-setosa	Iris-versicolor	Iris-virginica	Sum
Iris-setosa	250	0	0	250
Iris-versicolor	0	230	20	250
Iris-virginica	0	16	234	250
Sum	250	246	254	750

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