Filter Approaches For Feature Selection

Selecting the relevant descriptors before a supervised learning process

Ricco RAKOTOMALALA

Université Lumière Lyon 2

OUTLINE

1. Various approaches for variable selection
2. FILTER approaches for discrete predictors
3. FILTER approaches for continuous predictors
4. Conclusion
5. References

Why feature selection is important?

Deployment, interpretation, robustness

Less input variables... but the most relevant ones

2 perspectives: (1) removing the variables which are not related to the target attribute (relevance) ; (2) removing the variables which provide the same information than the others (redundancy).

1. To make easier the interpretation of the results (to interpret the association between the target and the input attributes)

2. To make easier the deployment of the classifier: less input variables means less information to collect for classifying new instances

3. Robustness. Occam's Razor principle: with similar performance on a dataset, simpler models are more accurate in the population. E.g. See Akaike (AIC) or BIC critera

Three approaches for feature selection (1/2)

1. Embedded methods. The variable selection is a part and is specific to the learning procedure (e.g. decision tree learning)

Consistency
Adapted to the learning method

Sometimes not optimal because the criterion used is not related to the predictive performance (e.g. error rate)

2. Wrapper methods. The selection process uses the learning method as a black box to detect the most interesting predictors for optimizing a performance criterion (e.g. error rate)

The performance criterion is explicitly optimized.

High risk of overfitting (overdependence on the dataset) and highly computationally intensive.

Three approaches for feature selection (2/2)

3. **Filter methods**. The selection is performed **before** the construction of the classifier. It is based on the concepts of relevance and redundancy. They are quantified with a correlation measurement (in the broader sense).

Computationally simple and fast.
Handling very large database

The variables are not selected according to the classifier characteristics.

A good solution is a subset of variables for which…

Relevance They are highly correlated to the target attribute.

Redundancy They are weakly related each other (correlated).
Ideally, they are orthogonal each other.

Filter methods for discrete predictors

Ranking and selection methods

Symmetrical uncertainty
Correlation measure for discrete attribute

Y \ X	x_1	…	x_l	…	x_L	Σ
y_1						
⋮			⋮			
y_k		…	n_{kl}	…		$n_{k.}$
⋮			⋮			
y_K						
Σ			$n_{.l}$			n

Joint and marginal frequencies

$$p_{kl} = \frac{n_{kl}}{n} \qquad p_{k.} = \frac{n_{k.}}{n} \qquad p_{.l} = \frac{n_{.l}}{n}$$

Mutual information (~ covariance)

$$I(Y,X) = \sum_k \sum_l p_{kl} \times \log_2 \frac{p_{kl}}{p_{k.} \times p_{.l}}$$

Entropy (~ standard deviation)

$$H(Y) = -\sum_k p_{k.} \log_2 p_{k.}$$

Symmetrical uncertainty
(~ correlation)

$$\rho_{y,x} = 2 \times \left[\frac{I(Y,X)}{H(Y) + H(X)} \right]$$

Defined between [0 ; 1]

Test of significance

$$G = 2 \times n \times \ln(2) \times I(Y,X)$$

Under the null hypothesis (X and Y are independent), G follows a χ^2 distribution with (K-1)*(L-1) degrees of freedom.

Symmetrical uncertainty
Example of calculation

By convention: $0 \times \log_2(0) = 0$

Two-way frequency table

Nombre de Y / Étiquettes de lignes	A	B	C	D	Total général
absence	120	20	7	3	150
presence	40	38	26	16	120
Total général	**160**	**58**	**33**	**19**	**270**

Two-way relative frequency table

Nombre de Y / Étiquettes de lignes	A	B	C	D	Total général
absence	44.44%	7.41%	2.59%	1.11%	55.56%
presence	14.81%	14.07%	9.63%	5.93%	44.44%
Total général	**59.26%**	**21.48%**	**12.22%**	**7.04%**	**100.00%**

$$I(Y,X) = \sum_k \sum_l p_{kl} \times \log_2 \frac{p_{kl}}{p_{k.} \times p_{.l}} = 0.175278$$

$$H(Y) = -\sum_k p_{k.} \log_2 p_{k.} = 0.9911$$

$$H(X) = -\sum_l p_{.l} \log_2 p_{.l} = 1.5640$$

$$\rho_{y,x} = 2 \times \left[\frac{I(Y,X)}{H(Y)+H(X)} \right] = 2 \times \left[\frac{0.175278}{0.9911+1.5640} \right] = 0.137197$$

$$G = 2 \times n \times \ln(2) \times I(Y,X) = 2 \times 270 \times \ln(2) \times 0.175278$$
$$= 65.60655 \text{ (p - value} \approx 0)$$

Example data set : Congressional Voting Records (modified) – **n** = 435 obs.

2. The selection method must select the relevant variables among these.

1. The selection method have to discard these variables.

Original input variables

Noise variables. The values of the original variables are randomly disturbed.

Attributes correlated with the original variables (97% of the values are identical).

Attribute	Category	Informations
handicapped.infants	Discrete	3 values
water.project.cost.sharin	Discrete	3 values
adoption.of.the.budget.re	Discrete	3 values
physician.fee.freeze	Discrete	3 values
el.salvador.aid	Discrete	3 values
religious.groups.in.schoo	Discrete	3 values
anti.satellite.test.ban	Discrete	3 values
aid.to.nicaraguan.contras	Discrete	3 values
mx.missile	Discrete	3 values
immigration	Discrete	3 values
synfuels.corporation.cutb	Discrete	3 values
education.spending	Discrete	3 values
superfund.right.to.sue	Discrete	3 values
crime	Discrete	3 values
duty.free.exports	Discrete	3 values
export.administration.act	Discrete	3 values
noise_handicapped.infants	Discrete	3 values
noise_water.project.cost.sharin	Discrete	3 values
noise_adoption.of.the.budget.re	Discrete	3 values
noise_physician.fee.freeze	Discrete	3 values
noise_el.salvador.aid	Discrete	3 values
noise_religious.groups.in.schoo	Discrete	3 values
noise_anti.satellite.test.ban	Discrete	3 values
noise_aid.to.nicaraguan.contras	Discrete	3 values
noise_mx.missile	Discrete	3 values
noise_immigration	Discrete	3 values
noise_synfuels.corporation.cutb	Discrete	3 values
noise_education.spending	Discrete	3 values
noise_superfund.right.to.sue	Discrete	3 values
noise_crime	Discrete	3 values
noise_duty.free.exports	Discrete	3 values
noise_export.administration.act	Discrete	3 values
corr_handicapped.infants	Discrete	3 values
corr_water.project.cost.sharin	Discrete	3 values
corr_adoption.of.the.budget.re	Discrete	3 values
corr_physician.fee.freeze	Discrete	3 values
corr_el.salvador.aid	Discrete	3 values
corr_religious.groups.in.schoo	Discrete	3 values
corr_anti.satellite.test.ban	Discrete	3 values
corr_aid.to.nicaraguan.contras	Discrete	3 values
corr_mx.missile	Discrete	3 values
corr_immigration	Discrete	3 values
corr_synfuels.corporation.cutb	Discrete	3 values
corr_education.spending	Discrete	3 values
corr_superfund.right.to.sue	Discrete	3 values
corr_crime	Discrete	3 values
corr_duty.free.exports	Discrete	3 values
corr_export.administration.act	Discrete	3 values
group	Discrete	2 values

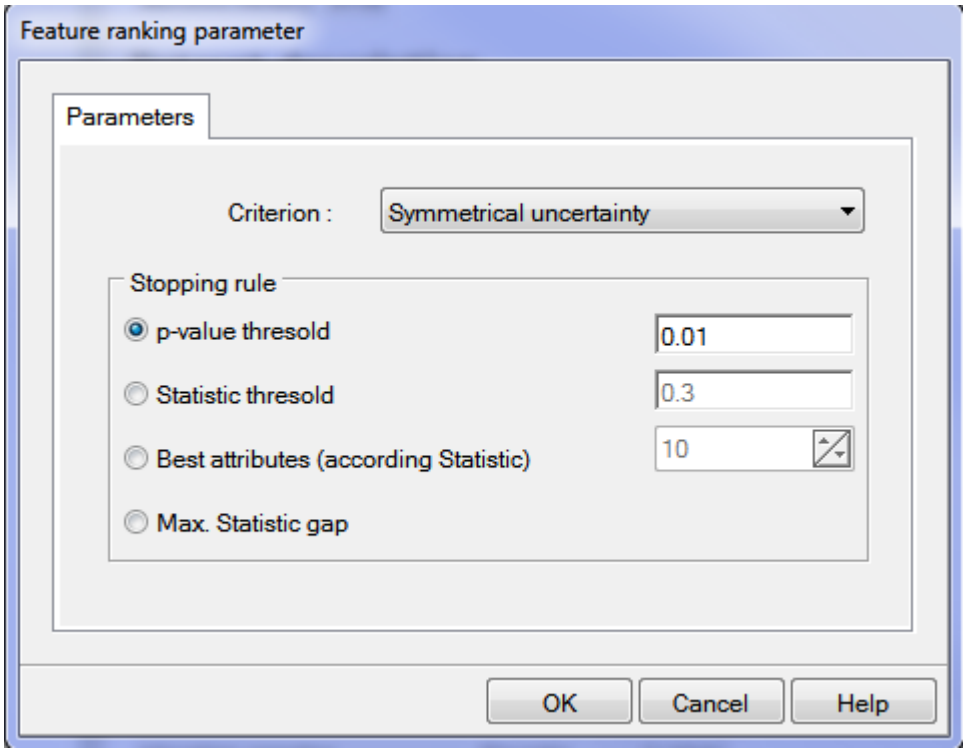
48 descriptors

← Target variable

Feature « ranking » method for discrete predictors

Steps:

1. Compute ρ for each predictive variable.
2. Sort the variables according ρ (decreasing order)
3. Retain only the significant variables (other rules are possible)



Settings in TANAGRA

N	Attribute	Values	Statistic	p-value
1	physician.fee.freeze	3	0.708862	0
2	corr_physician.fee.freeze	3	0.540679	0
3	adoption.of.the.budget.re	3	0.415544	0
4	el.salvador.aid	3	0.394048	0
5	corr_adoption.of.the.budget.re	3	0.37164	0
6	corr_el.salvador.aid	3	0.36604	0
7	education.spending	3	0.333286	0
8	aid.to.nicaraguan.contras	3	0.319763	0
9	crime	3	0.313788	0
10	corr_aid.to.nicaraguan.contras	3	0.288226	0
11	corr_crime	3	0.287527	0
12	mx.missile	3	0.282252	0
13	corr_education.spending	3	0.273481	0
14	corr_mx.missile	3	0.269558	0
15	superfund.right.to.sue	3	0.20505	0
16	duty.free.exports	3	0.197825	0
17	corr_duty.free.exports	3	0.19445	0
18	anti.satellite.test.ban	3	0.186272	0
19	corr_superfund.right.to.sue	3	0.179718	0
20	corr_anti.satellite.test.ban	3	0.160502	0
21	religious.groups.in.schoo	3	0.143636	0
22	corr_religious.groups.in.schoo	3	0.132297	0
23	handicapped.infants	3	0.119647	0
24	corr_handicapped.infants	3	0.108347	0
25	synfuels.corporation.cutb	3	0.100258	0
26	corr_synfuels.corporation.cutb	3	0.096047	0
27	export.administration.act	3	0.089249	0
28	corr_export.administration.act	3	0.081269	0
29	noise_physician.fee.freeze	3	0.014305	0.010858
30	noise_export.administration.act	3	0.012427	0.014778

The "good" input variables are in the first places. This is a good news.

But the correlated variables (redundant variables) are also in top positions.

At the 1% level, the variables randomly generated are discarded. The settings of the stopping rule are essential.

Ranking approach - Outline

Pros:

- Quickness, ability to handle very high-dimensional datasets
- Allow to discard the irrelevant attributes
- Radical reduction of the number of predictors

Cons:

- Do not take account of the redundancy
- Hard to determine the right settings
- When the number instances increase, all the variables seem significant
- Ignore the interaction between the variables

CFS method (Correlation based Feature Selection)

The selection process tries to optimize a criterion which reflects a trade-off between relevance and redundancy. (« m » is the number of selected input variables)

MERIT criterion: $$\mu = \frac{m \times \bar{\rho}_{y,x}}{\sqrt{m + m \times (m-1) \times \bar{\rho}_{x,x}}}$$

$$\bar{\rho}_{y,x} = \frac{1}{m} \sum_{j=1}^{m} \rho_{y,x_j}$$

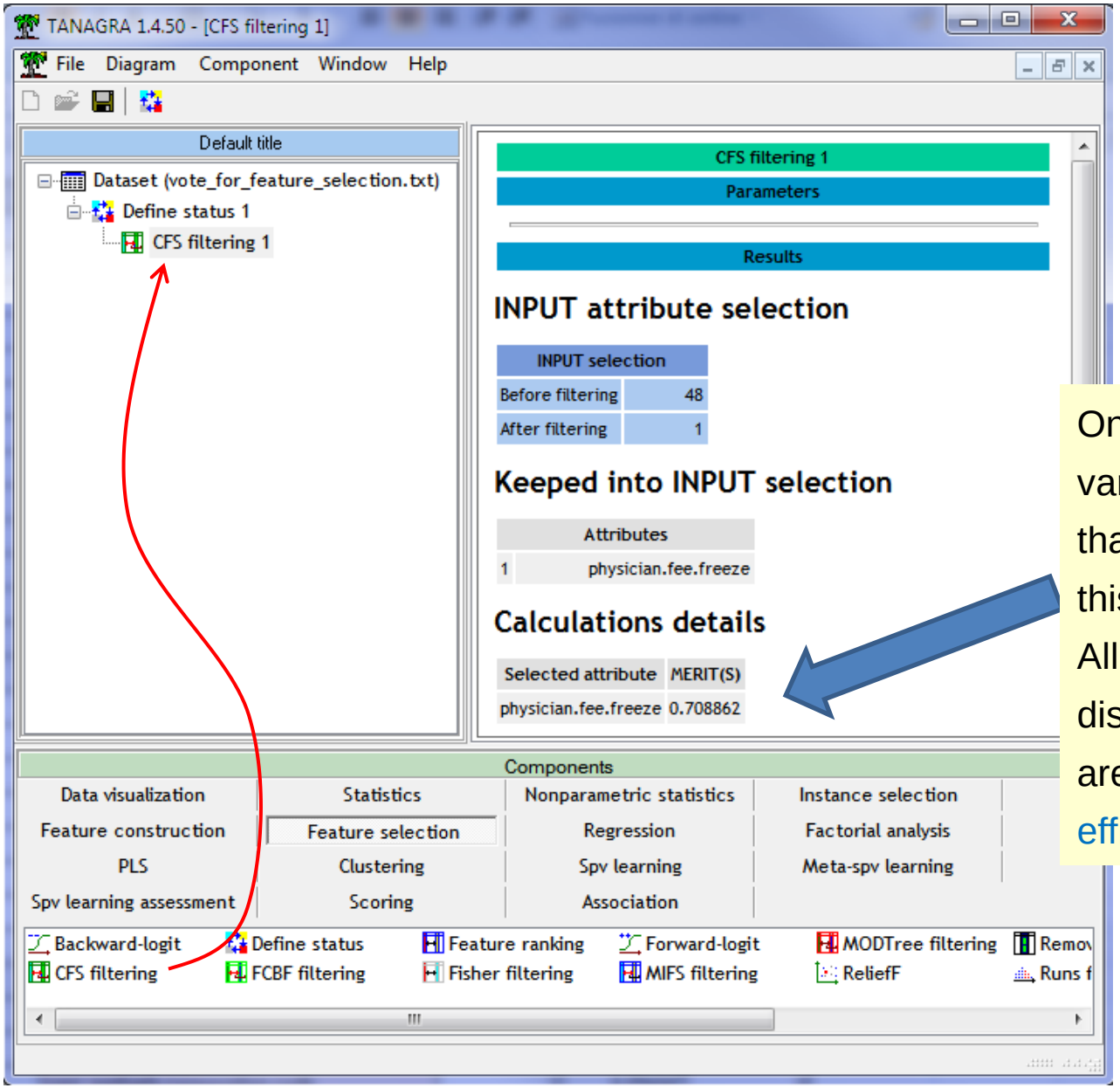
Average of the correlation between the target attribute and the selected predictors (relevance)

$$\bar{\rho}_{x,x} = \frac{2}{m \times (m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{x_i,x_j}$$

Average of the correlation between the selected predictors (redundancy)

Various optimization strategies may be used (FORWARD, BACKWARD, etc.)

CFS on the VOTE dataset



Only the "physician fee freeze" variable is selected. We know that this is the best (because this is a well-known dataset). All the others have been discarded, including those that are redundant. CFS is very efficient **on this dataset**.

CFS approach - Summary

Pros:

- Considering both the relevance **AND** redundancy
- No parameters
- Filtering of the high dimensional dataset…

Cons:

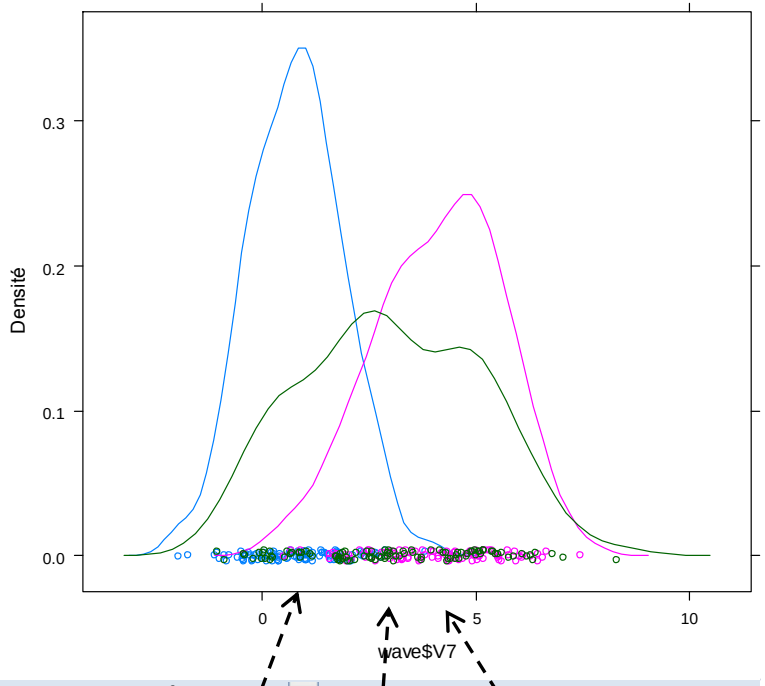
- …only up to a certain point because CFS is in quadratic complexity
- “No parameters” means also inability to adapt the algorithm to the characteristics of the data set
- The learning set size “n” does not influence the results. Yet, a correlation computed on 100 instances is not as important that if it was computed on 100000 instances.

Filter methods for continuous predictors

Ranking and selection methods

Correlation ratio

Measure of the relationship between discrete (target) and continuous (input) variables.



Étiquettes d	A	C	B	Total général
Moyenne de V7	0.788	3.025	4.118	2.626

Conditional mean

$$\bar{x}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik}$$

Partitioning the variance (ANOVA)

$$SCT = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Total deviation

$$SCE = \sum_{k=1}^{K} n_k(\bar{x}_k - \bar{x})^2$$

Between groups (explained)

$$SCR = SCT - SCE = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k)^2$$

Within groups (residual)

$$\rho^2_{Y/X} = \frac{SCE}{SCT}$$

Correlation ratio (ρ is defined between [0 ; 1])

$$F = \frac{\frac{SCE}{K-1}}{\frac{SCR}{n-K}}$$

F-test for testing the significance [Fisher (K-1, N-K) under H0]

(F of the one-way ANOVA)

Example data set: Waveform (modified) – n = 300 obs.

2. The selection method must select the relevant variables among these.

Original input variables

Noise variables. The values of the original variables are randomly disturbed.

1. The selection method have to discard these variables.

Attributes correlated with the original variables (correlation ≈ 0.96).

Target attribute

Attribute	Category	Informations
Onde	Discrete	3 values
V1	Continue	-
V2	Continue	-
V3	Continue	-
V4	Continue	-
V5	Continue	-
V6	Continue	-
V7	Continue	-
V8	Continue	-
V9	Continue	-
V10	Continue	-
V11	Continue	-
V12	Continue	-
V13	Continue	-
V14	Continue	-
V15	Continue	-
V16	Continue	-
V17	Continue	-
V18	Continue	-
V19	Continue	-
V20	Continue	-
V21	Continue	-
rnd_1	Continue	-
rnd_2	Continue	-
rnd_3	Continue	-
rnd_4	Continue	-
rnd_5	Continue	-
rnd_6	Continue	-
rnd_7	Continue	-
rnd_8	Continue	-
rnd_9	Continue	-
rnd_10	Continue	-
rnd_11	Continue	-
rnd_12	Continue	-
rnd_13	Continue	-
rnd_14	Continue	-
rnd_15	Continue	-
rnd_16	Continue	-
rnd_17	Continue	-
rnd_18	Continue	-
rnd_19	Continue	-
rnd_20	Continue	-
rnd_21	Continue	-
cor_1	Continue	-
cor_2	Continue	-
cor_3	Continue	-
cor_4	Continue	-
cor_5	Continue	-
cor_6	Continue	-
cor_7	Continue	-
cor_8	Continue	-
cor_9	Continue	-
cor_10	Continue	-
cor_11	Continue	-
cor_12	Continue	-
cor_13	Continue	-
cor_14	Continue	-
cor_15	Continue	-
cor_16	Continue	-
cor_17	Continue	-
cor_18	Continue	-
cor_19	Continue	-
cor_20	Continue	-
cor_21	Continue	-

n = 300

p = 63

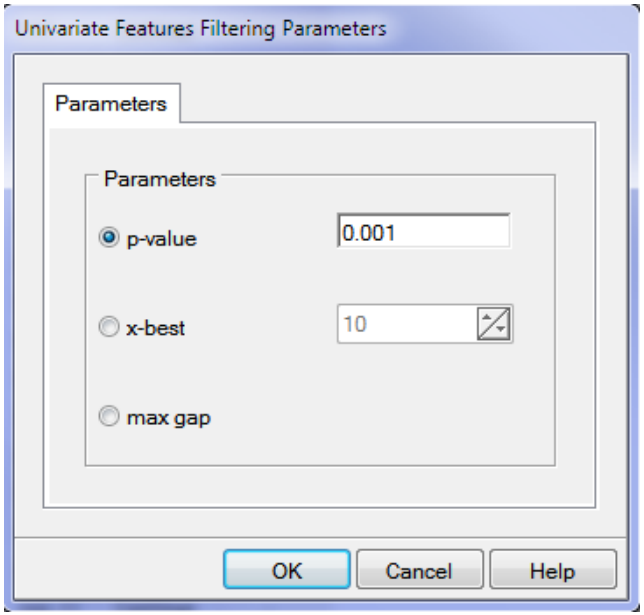
risk of unstable results (n / p weak)

63 candidate predictors

Feature « ranking » method for continuous predictors

Steps:

1. Compute ρ^2 for each predictive variable.
2. Rank the variables according ρ^2 (decreasing order)
3. Retain only the significant variables (other rules are possible)



Settings in TANAGRA

The behavior is similar to that the ranking for discrete predictors, but...

the approach is ineffective for multimodal conditional distributions (see "runs test" as a possible solution)

N	Attribute	F	p-value (2,297)
1	V7	111.13	0
2	cor_7	107.56	0
3	cor_15	101.32	0
4	V15	99.56	0
5	V16	90.1	0
6	cor_16	88.99	0
7	V8	83.96	0
8	V6	82.7	0
9	V9	82.32	0
10	V14	81.29	0
11	cor_6	79.99	0
12	cor_8	78.75	0
13	cor_9	78.15	0
14	cor_14	76.81	0
15	V17	74.47	0
16	cor_17	69.56	0
17	V13	66.26	0
18	V5	66.18	0
19	cor_13	64.45	0
20	cor_5	62.24	0
21	V11	59.13	0
22	cor_11	56.31	0
23	V12	52.82	0
24	cor_12	49.04	0
25	V4	48.5	0
26	cor_4	46.19	0
27	V10	46.08	0
28	cor_10	41.3	0
29	V18	36.24	0
30	cor_18	33.5	0

The "good" input variables are in the first places. This is a good news.

But the correlated variables (redundant variables) are also in top positions.

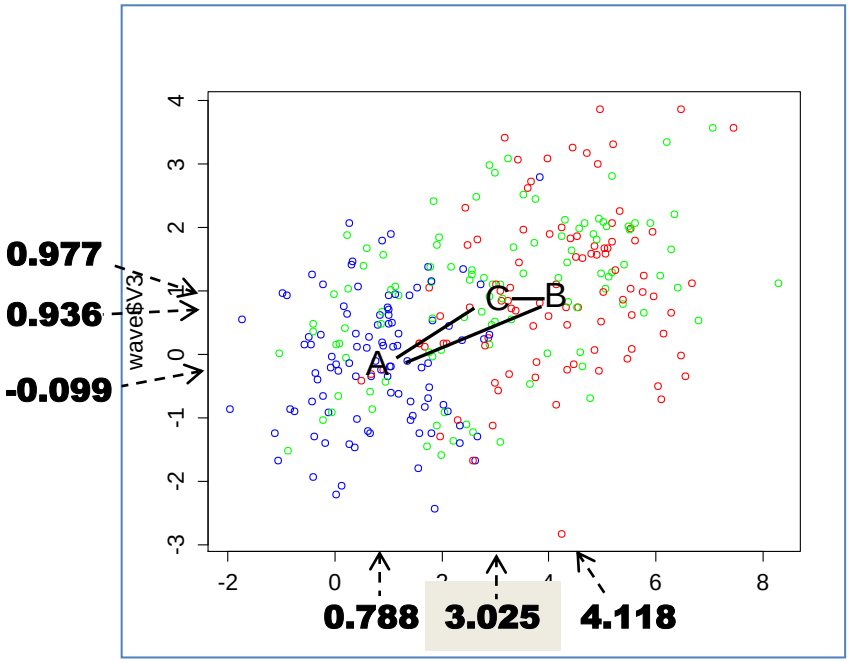
At the 1% level, the variables randomly generated are discarded. The settings of the stopping rule are essential.

Note: Tanagra sorts the variables according F, the results are the same because…

$$F = \frac{\rho^2 / (K-1)}{(1-\rho^2) / (n-K)}$$

MANOVA approach for variable selection

Multivariate Analysis of Variance



Idea: We retain only the variables which contribute significantly to the gap between the conditional centroids.

Evaluating the gap between the conditional centroids

Wilks' lambda criterion

$$\Lambda = \frac{\det(W)}{\det(V)}$$

← Within groups deviation

← Total dispersion

(multivariate version of [**1 - ρ²**]) !

Test for significance ("m" variables)

$$F_{\text{RAO}} = \left(\frac{1-\Lambda^{1/b}}{\Lambda^{1/b}}\right)\left(\frac{ab-c}{m(K-1)}\right) \cong Fisher\left(m(K-1), ab-c\right)$$

(a, b and c are obtained from n, m, K) !

Contribution of (m+1)th additional variable

$$F = \frac{n-K-m}{K-1}\left(\frac{\Lambda_m}{\Lambda_{m+1}} - 1\right) \cong Fisher\left(K-1, n-K-m\right)$$

« STEPDISC » algorithms for variable selection

(Stepwise discriminant analysis)

FORWARD :

- Start with the empty subset
- Add the best variable (which maximizes F) at each step
- Stop when the additional variable does not contribute significantly

BACKWARD :

- Start with all the variables
- Remove the worst variable, the one which minimizes F
- Stop when the variable that we want to remove contributes significantly

BIDIRECTIONAL:

Check if the adding of a variable does not imply the removing of an already selected variable.

STEPDISC (FORWARD) for WAVE dataset

(Stopping rule α = 1%)

N	d.f	Best	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	(2, 297)	V7	V7	cor_7	cor_15	V15	V16
		L : 0.5720	L : 0.5720	L : 0.5799	L : 0.5944	L : 0.5987	L : 0.6224
		F : 111.13	F : 111.13	F : 107.56	F : 101.32	F : 99.56	F : 90.10
		p : 0.0000	p : 0.0000	p : 0.0000	p : 0.0000	p : 0.0000	p : 0.0000
2	(2, 296)	V11	V11	cor_11	V17	cor_17	V10
		L : 0.4128	L : 0.4128	L : 0.4180	L : 0.4298	L : 0.4348	L : 0.4512
		F : 57.06	F : 57.06	F : 54.50	F : 48.96	F : 46.70	F : 39.61
		p : 0.0000	p : 0.0000	p : 0.0000	p : 0.0000	p : 0.0000	p : 0.0000
3	(2, 295)	V17	V17	cor_17	cor_16	V16	V9
		L : 0.3582	L : 0.3582	L : 0.3584	L : 0.3636	L : 0.3650	L : 0.3734
		F : 22.47	F : 22.47	F : 22.38	F : 19.96	F : 19.33	F : 15.59
		p : 0.0000	p : 0.0000	p : 0.0000	p : 0.0000	p : 0.0000	p : 0.0000
4	(2, 294)	cor_16	cor_16	V16	V9	cor_15	cor_9
		L : 0.3312	L : 0.3312	L : 0.3315	L : 0.3365	L : 0.3369	L : 0.3376
		F : 11.98	F : 11.98	F : 11.83	F : 9.49	F : 9.32	F : 8.98
		p : 0.0000	p : 0.0000	p : 0.0000	p : 0.0001	p : 0.0001	p : 0.0002
5	(2, 293)	V12	V12	cor_12	V5	cor_5	V14
		L : 0.3135	L : 0.3135	L : 0.3140	L : 0.3141	L : 0.3175	L : 0.3185
		F : 8.31	F : 8.31	F : 8.03	F : 7.99	F : 6.34	F : 5.88
		p : 0.0003	p : 0.0003	p : 0.0004	p : 0.0004	p : 0.0020	p : 0.0031
6	(2, 292)	V5	V5	V9	V14	cor_9	cor_15
		L : 0.3035	L : 0.3035	L : 0.3040	L : 0.3047	L : 0.3052	L : 0.3057
		F : 4.80	F : 4.80	F : 4.54	F : 4.19	F : 3.97	F : 3.70
		p : 0.0089	p : 0.0089	p : 0.0115	p : 0.0161	p : 0.0199	p : 0.0260
7	(2, 291)		V9	cor_9	cor_15	V14	rnd_12
			L : 0.2944	L : 0.2955	L : 0.2962	L : 0.2963	L : 0.2970
			F : 4.50	F : 3.95	F : 3.59	F : 3.54	F : 3.18
		-	p : 0.0119	p : 0.0204	p : 0.0289	p : 0.0303	p : 0.0430

RESULTS

6 variables are selected,

1 « correlated » is included

(we are in a challenging context here: a small sample size compared to the number of candidate variables; strong correlations between the variables)

STEPDISC approach - Outline

Comment:

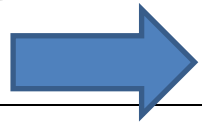
- This approach may be regarded as an embedded approach for linear discriminant analysis
- The FORWARD strategy is prefered when we handle a dataset with a large number of candidate variables (faster, less risk of error)

Pros:

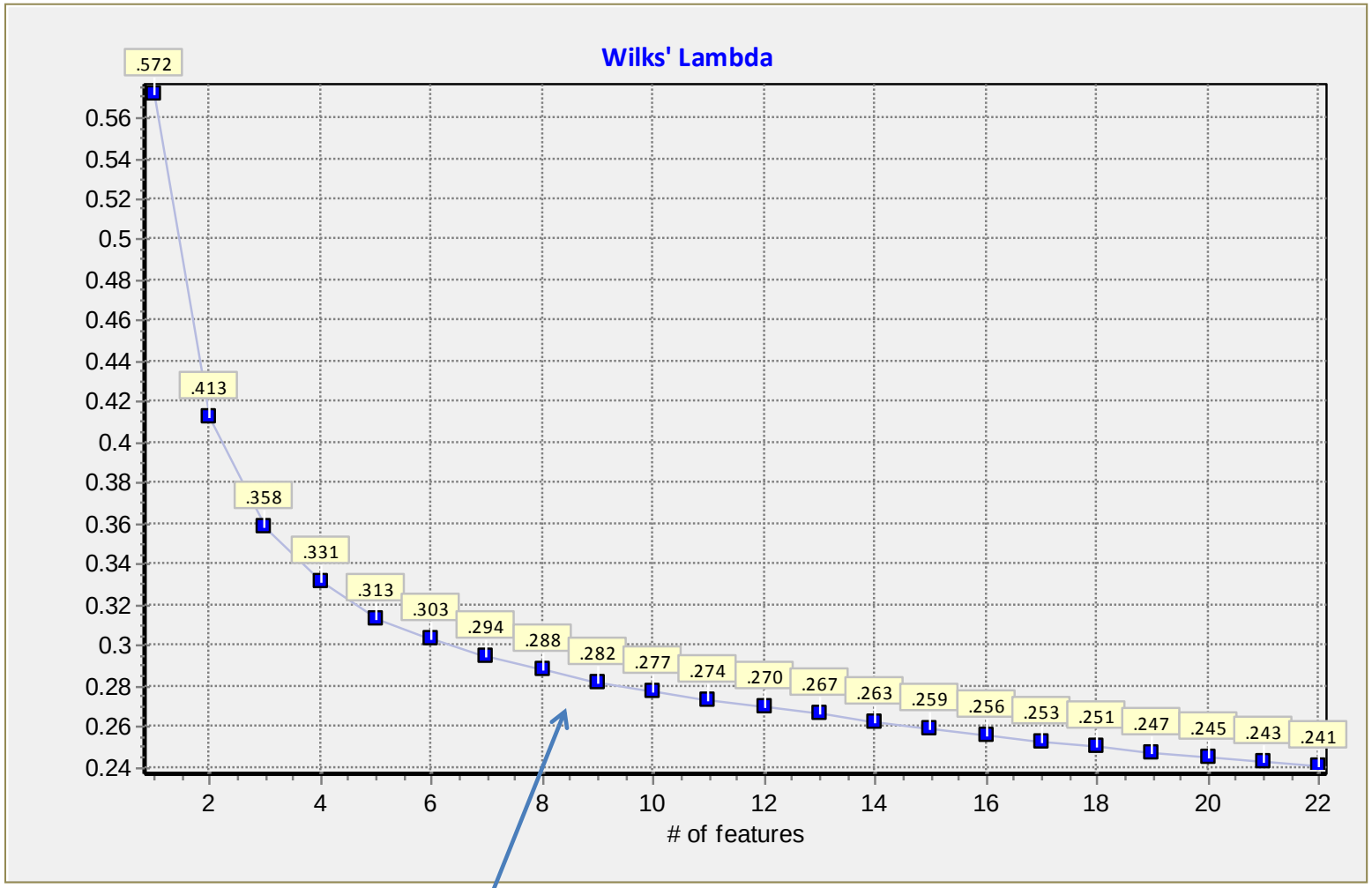
- Handle the relevance AND the redundancy
- Ability to deal with high dimensional datasets... but it is not faster than the ranking method

Cons:

- Difficulty to set a priori the stopping rule (e.g. the significance level may be disturbed by the multiple comparisons; for large sample size, all the variables seem significant)



STEPDISC approach
A possible rule of thumb for large datasets



From here, an additional selected variable does not decrease significantly the Wilks' Lambda

Conclusion

FILTERING APPROACH - Outline

• The filtering techniques enable to a substantial reduction of the number of candidate variables before the construction of the classifier

• Two key concepts are highlighted: relevance, the relationship between the predictors and the target variable; redundancy, the predictors are related with each other

• The relationship is measured with the correlation (in a large sense)

• The ranking methods are very fast but they do not handle the redundancy

• The approaches which handle the redundancy are slower and may be problematic in the context of very high dimensional dataset (when the number of variables is larger than the number of instances) (when we deal with unstructured dataset [e.g. text mining], a very large number of features is generated automatically)

• The variables selected with the filtering process would be the best for any subsequent machine learning algorithms. This is a strong assumption.

References

Tanagra tutorials, « Filter methods for feature selection », 2010 ; http://data-mining-tutorials.blogspot.fr/2010/10/filter-methods-for-feature-selection.html

Tanagra tutorials, « Feature selection using MIFS », 2008 ; http://data-mining-tutorials.blogspot.fr/2008/11/feature-selection-using-mifs-algorithm.html

Tanagra tutorials, « STEPDISC – Feature selection for LDA », 2008 ; http://data-mining-tutorials.blogspot.fr/2008/11/stepdisc-feature-selection-for.html

Tanagra tutorials, « ”Wrapper” for feature selection », 2010 ; http://data-mining-tutorials.blogspot.fr/2010/03/wrapper-for-feature-selection.html

Tanagra tutorials, « ”Wrapper” for feature selection (continuation) », 2010 ; http://data-mining-tutorials.blogspot.fr/2010/04/wrapper-for-feature-selection.html