Filtrage des prédicteurs

Sélection de variables pour (préalable à) l'apprentissage supervisé

Les méthodes « FILTRE »

Ricco RAKOTOMALALA

Université Lumière Lyon 2

PLAN

1. Pourquoi la sélection de variables ? Les différentes approches.
2. Approches FILTRE pour prédicteurs qualitatifs
3. Approches FILTRE pour prédicteurs quantitatifs
4. Bilan
5. Bibliographie

Pourquoi la sélection de variables ?

Déploiement, interprétation, robustesse

Moins de variables… mais les plus pertinentes

2 aspects clés : (1) éliminer les variables qui n'ont rien à voir avec le problème que l'on traite (pertinence) ; (2) éliminer les variables qui font doublons c.-à-d. qui apportent le même type d'information (redondance).

Pourquoi ?

1. Faciliter l'interprétation des résultats (mieux situer l'impact des variables sur l'explication)

2. Faciliter le déploiement des modèles : moins de variables ➔ moins d'information à trouver (de questions à poser) pour appliquer le modèle

3. Robustesse. Principe du Rasoir d'Occam (principe de parcimonie) : à performances identiques (sur les données d'apprentissage), le modèle le plus simple sera plus robuste dans la population. Cf. les critères de type AIC ou BIC

3 approches de sélection de variables (1/2)

1. Approche intégrée (embedded). La sélection s'appuie sur un critère propre à la méthode, elle est intégrée dans le processus d'apprentissage (ex. arbres de décision, forward/backward pour la régression logistique, etc.).

Cohérence
Adaptée à la méthode

Pas optimale parce que critère parfois sans lien direct avec le taux d'erreur

2. Approche enveloppe (wrapper). La sélection utilise la méthode comme une boîte noire et cherche à optimiser explicitement un critère de performance (ex. taux d'erreur)

Optimalité au sens du critère de performance

Très lent
Danger de surapprentissage, même avec les précautions d'usage (ex. utilisation d'un fichier test mais, de facto, ce dernier participe à la sélection)

3 approches de sélection de variables (2/2)

3. **Approche filtre** (filter). La sélection vient en amont des méthodes d'apprentissage supervisé, traduit les notions de pertinence et de redondance par la « corrélation » (au sens large) entre les variables.

Rapidité

Défricher les grandes bases

Le filtrage serait efficace quelle que soit la méthode statistique utilisée en aval ? Hum !

La bonne solution serait un sous-ensemble de variables prédictives où…

Pertinence : Elles sont fortement corrélées avec la variable cible.

Redondance : Elles sont faiblement corrélées entre elles (dans l'idéal orthogonales deux à deux).

Approche filtre pour les prédicteurs qualitatifs

Méthodes de ranking et de sélection

Incertitude symétrique (symmetrical uncertainty)
Mesurer la « corrélation » entre variables qualitatives

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

Fréquences conjointes et marginales

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

Information mutuelle (~ covariance [liaison])

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

Entropie (~ écart-type [dispersion])

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

Incertitude symétrique
(~ corrélation)

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

Varie entre [0 ; 1]

Test de significativité

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

Sous H0 : indépendance entre X et Y, suit une loi du χ^2 à (K-1)*(L-1) degrés de liberté

Incertitude symétrique (symmetrical uncertainty)
Exemple de calcul

Par convention : 0 x $\log_2(0)$ =

Tableau de comptage

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**

Tableau des fréquences relatives conjointes et marginales

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%**

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

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

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

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

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

Base utilisée dans cette partie :

Vote au congrès (modifié) – n = 435 obs.

2. Et choisir les meilleures variables prédictives ici ?

1. Est-ce que la méthode va être capable d'évacuer ces variables ?

Variables originelles

Variables « bruitées », valeurs mélangées au hasard à l'intérieur des colonnes

Variables « corrélées »avec les variables originelles (partagent les mêmes valeurs dans 97% des cas)

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

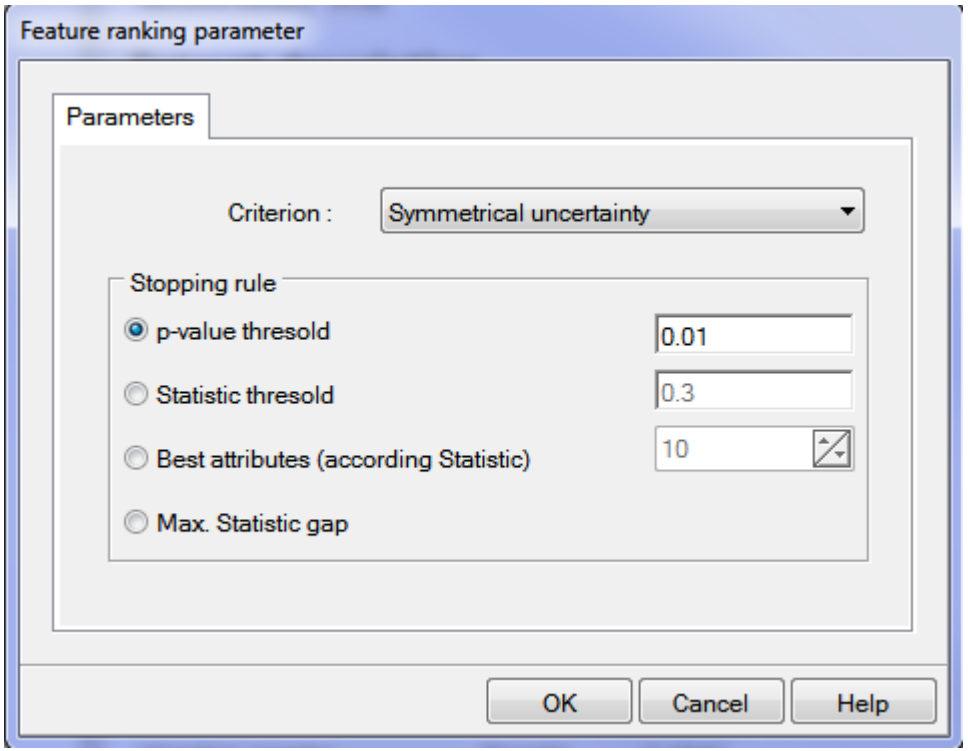
48 prédicteurs

← Variable cible

« Ranking » des prédicteurs qualitatifs

Etapes :

1. Calculer le critère ρ pour chaque variable prédictive
2. Les classer par ρ décroissant
3. Ne retenir que les variables significatives (ou autre règle, cf. bilan)



Paramétrage dans 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

Les « bonnes » variables dans les premières positions. Yes !

Les variables corrélées s'intercalent. Pas bon ça.

A 1 %, on évite de justesse les variables générées aléatoirement. Le choix de la règle d'arrêt est primordial !!!

Méthode de Ranking - Bilan

Avantages :

- Rapidité, traitement des très grandes bases (en nombre de variables)
- Permet d'évacuer les variables non pertinentes, réduction drastique
- Paramétrage délicat (choix du nombre de variables à retenir)

Inconvénients :

- Ne gère pas la redondance
- Quand « n » augmente, tout paraît significatif, et de toute manière la loi n'est plus valable si on cherche les « meilleures » variables (cf. ajustement de la p-value, ex. correction de Bonferroni). Mieux vaut utiliser des règles empiriques dans ce cas (décrochage, etc.)
- Considère les variables individuellement, ne gère pas les influences conjointes

Algorithme « CFS » de sélection de variables

(Correlation based Feature Selection)

Optimisation d'un critère réalisant un arbitrage entre la liaison des prédicteurs avec la cible et leurs liaisons croisées (« m » nombre de variables sélectionnées)

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

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

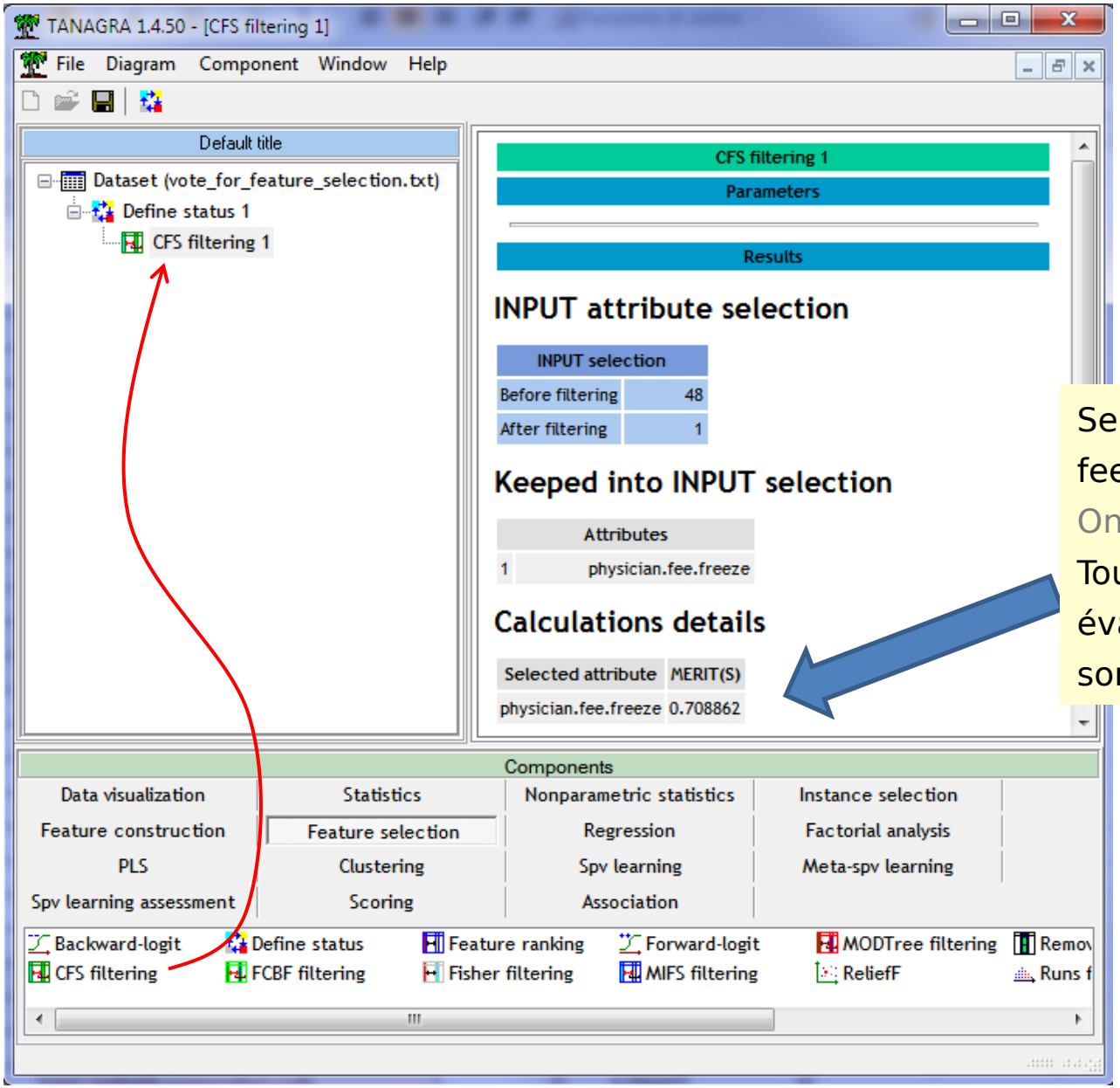
Moyenne des corrélations des prédicteurs avec la cible (pertinence)

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

Moyenne des corrélations croisées entre prédicteurs (redondance)

Toute stratégie d'agrégation est utilisable (FORWARD, BACKWARD, autres)

CFS sur la base « Vote »



Seule la variable « physician fee freeze » est sélectionnée. On sait que c'est la meilleure. Toutes les autres ont été évacuées, y compris celles qui sont redondantes.

Méthode CFS - Bilan

Avantages :

- Traitement de la pertinence ET de la redondance
- Filtrage des très grandes bases … jusqu'à un certain point, l'algorithme est en $O(m^2)$ [calcul des corrélations croisées, sélection…]
- Pas de paramétrage à faire

Inconvénients :

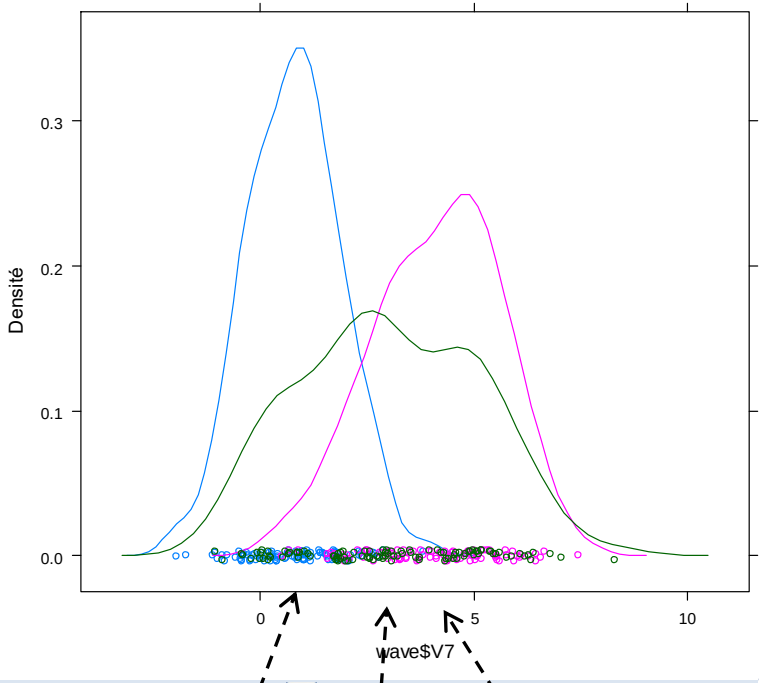
- Pas de paramétrage justement, on ne peut pas adapter aux données ou au cahier des charges
- La taille de l'échantillon « n » n'influe pas du tout ? Les corrélations n'ont pas la même portée sur n = 10 et sur n = 10.000 individus.

Approche filtre pour les prédicteurs quantitatifs

Méthodes de ranking et de sélection

Rapport de corrélation

Mesurer la liaison entre variables qualitatives (cible) et quantitatives (prédicteurs)



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

Moyenne conditionnelle

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

Equation d'analyse de variance

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

$$SCE = \sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^2$$ Dispersion expliquée par l'appartenance aux groupes

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

$$\rho^2_{Y/X} = \frac{SCE}{SCT}$$ Rapport de corrélation (varie entre [0 ; 1])

$$F = \frac{\frac{SCE}{K-1}}{\frac{SCR}{n-K}}$$ Pour tester la significativité [Fisher (K-1, N-K) sous H0] (F de l'ANOVA à 1 facteur)

Base utilisée dans cette partie : Waveform (modifié) – n = 300 obs.

2. Et choisir les meilleures variables prédictives ici ?

1. Est-ce que la méthode va être capable d'évacuer ces variables ?

Variables originelles

Variables « bruitées », valeurs mélangées au hasard à l'intérieur des colonnes

Variables « corrélées » avec les variables originelles (corrélation ~ 0.96 !!)

Variable cible

Hum, danger !

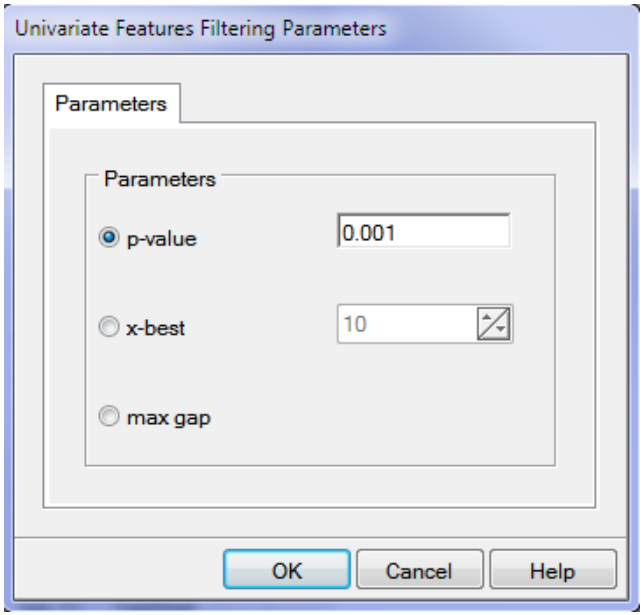
63 prédicteurs

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	-

« Ranking » des prédicteurs quantitatifs

Etapes :

1. Calculer le critère ρ^2 pour chaque variable prédictive
2. Les classer par ρ^2 décroissant
3. Ne retenir que les variables significatives (ou autre règle, cf. bilan)



Paramétrage dans TANAGRA

Le bilan identique au « ranking » pour prédicteurs qualitatifs, sauf…

Technique inopérante si distributions conditionnelles multimodales (*cf. les séquences*)

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

Les « bonnes » variables
dans les premières positions.

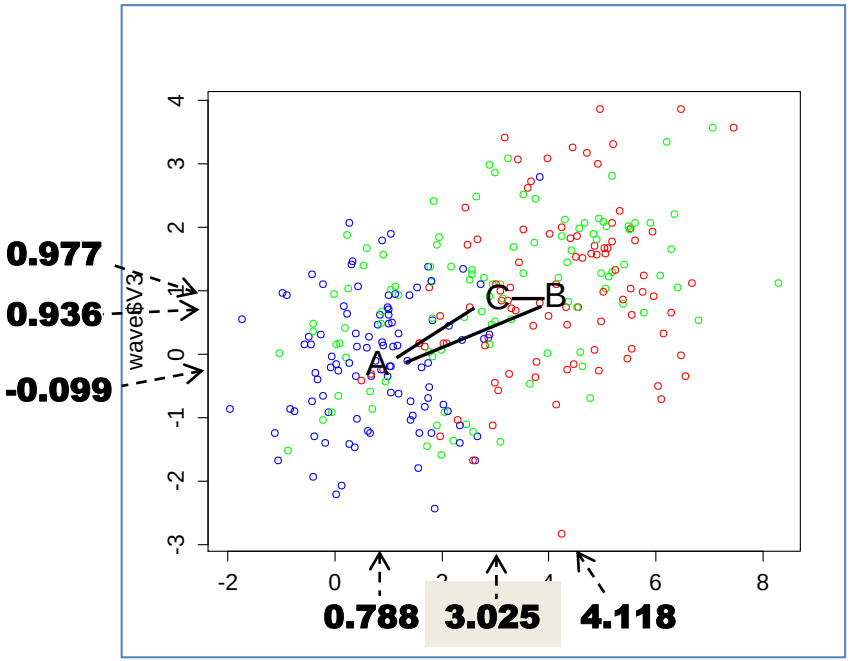
Les variables corrélées s'intercalent. Pas bon ça (*faut dire qu'on a fait fort avec une corrélation de ~0.96 en moyenne*)

Pas de variables bruitées significatives à 1%

Remarque : Tanagra trie selon F, ça revient au même puisque

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

Critère MANOVA pour la sélection de variables



Idée : ne conserver que les variables qui contribuent significativement à l'écartement des barycentres (des centres de classes)

Evaluation de l'écartement global : LAMBDA de Wilks

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

← Dispersion intra-classes

← Dispersion totale

(version multivariée de [**1 - ρ²**]) !

Test de significativité (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 et c sont obtenues à partir de n, m, K) !

Contribution d'une (m+1)ème variable additionnelle

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

Algorithme « STEPDISC » de sélection de variables

(Stepwise discriminant analysis)

FORWARD :

- Commencer par l'ensemble vide
- Ajouter la meilleure variable à chaque étape (F le plus élevé)
- S'arrêter quand la variable à ajouter n'est pas significative

BACKWARD :

- Commencer par la totalité des variables
- Retirer la pire variable à chaque étape (F le plus faible)
- S'arrêter quand la variable à retirer est significative

BIDIRECTIONNELLE:

Vérifier que chaque ajout ne provoque pas le retrait d'une variable précédemment sélectionnée

STEPDISC (FORWARD) pour WAVE

(Règle d'arrêt α = 1%)

BILAN

6 variables sélectionnées, 1 « corrélée » s'est immiscée (on est dans un contexte extrême ici, effectif faible par rapport au nombre de variables candidates, très forte corrélations)

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

Méthode STEPDISC - Bilan

Commentaires :

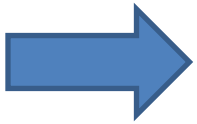
- Cette approche peut être considérée comme « intégrée » à l'analyse discriminante linéaire (séparabilité linéaire)
- Privilégier FORWARD lorsque le nombre de variables candidates est très élevé (plus rapide, moins de risque de plantage)

Avantages :

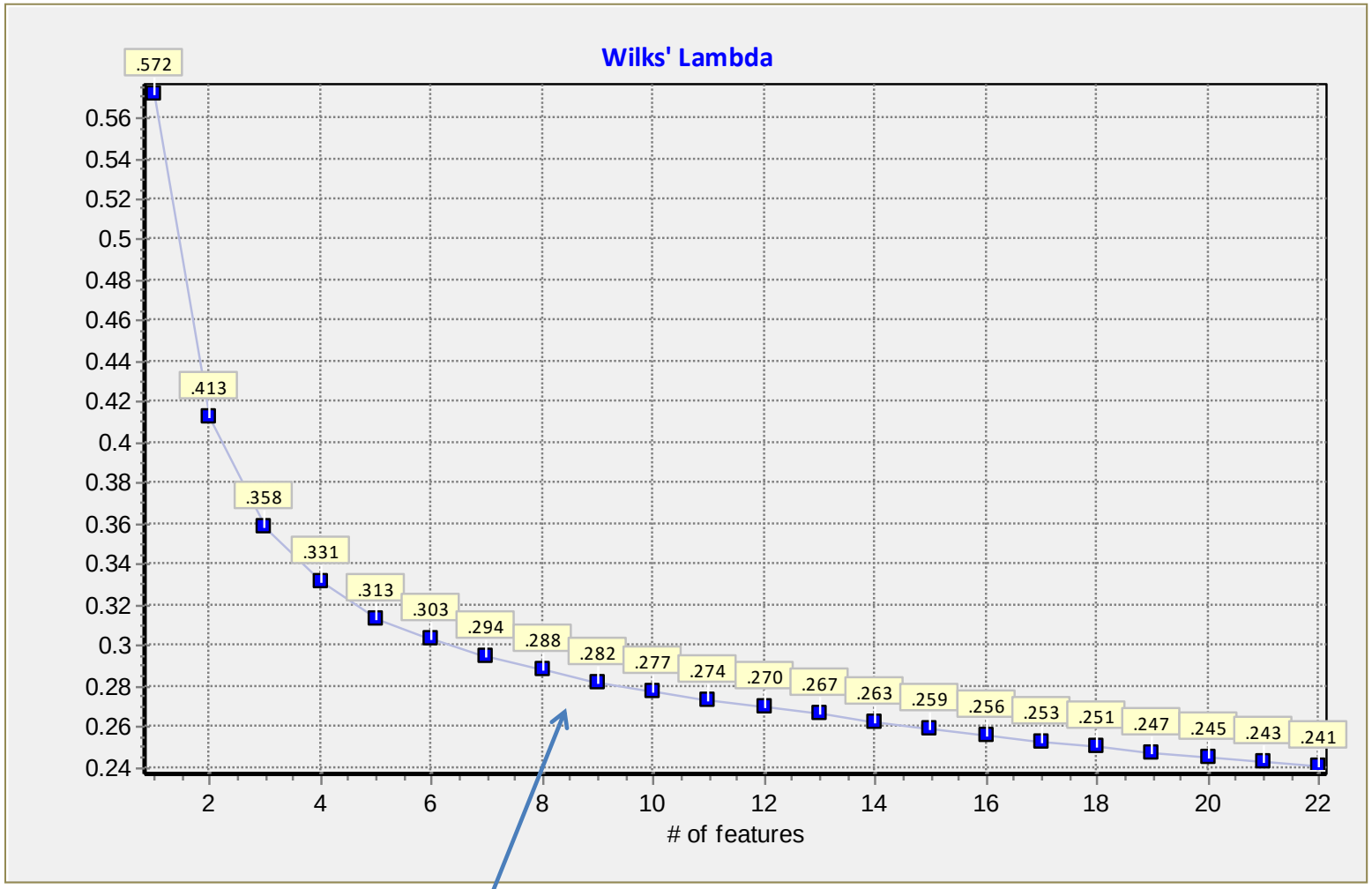
- Traitement de la pertinence ET de la redondance
- Filtrage des très grandes bases … jusqu'à un certain point, l'algorithme nécessite quand même beaucoup de calculs

Inconvénients :

- Paramétrage de la règle d'arrêt délicat (comparaisons multiples : loi modifiée ; taille d'échantillon : n grand, tout devient significatif)



Méthode STEPDISC
Règle de décision pour les grands effectifs



A partir d'ici (à peu près), l'adjonction d'une variable ne baisse plus significativement le Λ de Wilks

Bilan

Méthode FILTRE - Bilan

• Les techniques de filtrage permettent de réduire très fortement en amont (avant la modélisation) le nombre de variables candidates
• Deux notions clés sont mis en avant : pertinence, liaison du prédicteur avec la cible ; redondance, liaisons entre les prédicteurs
• La liaison est traduite par la « corrélation » (au sens large)

• Les techniques de ranking sont très rapides mais ne gèrent pas la redondance
• Les techniques de sélection appréhendent les deux notions mais sont moins rapides, problématiques quand m >> dizaine de milliers de prédicteurs (ça arrive quand ils sont générés automatiquement ex. traitement des données non structurées)

• Il reste un présupposé fort : le sous ensemble sélectionné est censé convenir quelle que soit la méthode statistique utilisée en aval (???)

Bibliographie

Tutoriel Tanagra, « Filtrage des prédicteurs discrets », 2010 ; http://tutoriels-data-mining.blogspot.fr/2010/06/filtrage-des-predicteurs-discrets.html (d'autres méthodes sont décrites : MIFS, FCBF, etc. ; mise en œuvre avec différents logiciels : Knime, RapidMiner, R, etc.)

Tutoriel Tanagra, « Stepdisc – Analyse discriminante », 2008 ; http://tutoriels-data-mining.blogspot.fr/2008/03/stepdisc-analyse-discriminante.html

Tutoriel Tanagra, « Stratégie Wrapper pour la sélection de variables », 2009 ; http://tutoriels-data-mining.blogspot.fr/2009/05/strategie-wrapper-pour-la-selection-de.html

Tutoriel Tanagra, « Wrapper pour la sélection de variables (suite) », 2010 ; http://tutoriels-data-mining.blogspot.fr/2010/01/wrapper-pour-la-selection-de-variables.html