1 Objectif

Méthodes « filtres » pour la sélection de prédicteurs discrets en apprentissage supervisé.

La sélection de variables est un dispositif crucial de l'apprentissage supervisé. On cherche à isoler le sous-ensemble de prédicteurs qui permet d'expliquer efficacement les valeurs de la variable cible. Trois approches sont généralement citées dans la littérature. Les méthodes « **embedded** » intègrent directement la sélection dans le processus d'apprentissage. Les arbres de décision en sont l'illustration la plus emblématique. Mais en réalité, on peut classer dans ce groupe toutes techniques qui évaluent l'importance d'une variable en cohérence avec le critère utilisé pour évaluer la pertinence globale du modèle (ex. STEPDISC basé sur le lambda de Wilks pour l'analyse discriminante[1] ; critère AIC pour la régression logistique[2]). Les méthodes « **wrapper** » optimisent explicitement un critère de précision, le plus souvent le taux d'erreur[3]. Elles ne s'appuient en rien sur les caractéristiques de l'algorithme d'apprentissage qui est utilisé comme une boîte noire. Si la performance en classement n'est pas en cause, on reproche souvent à cette approche la gourmandise en ressources de calcul. Enfin, troisième et dernière approche, les méthodes « **filter** » agissent en amont, avant la mise en œuvre de la technique d'apprentissage, et sans lien direct avec celui-ci. On présume donc qu'un processus indépendant basé sur un critère ad hoc permettrait de détecter les prédicteurs pertinents quel que soit l'algorithme d'apprentissage mis en œuvre en aval. Le pari est osé, voire hasardeux. Et pourtant, certaines expérimentations montrent que l'approche est viable même lorsque la méthode d'apprentissage utilise dans le même temps un dispositif intégré (embedded) de sélection de variables (les arbres de décision avec C4.5 par exemple[4]).

Nous nous intéressons aux méthodes de filtrage (filter) basées sur le principe suivant : le sous-ensemble de prédicteurs sélectionnés doit être composé de variables fortement liées avec la variable cible (pertinence) mais faiblement liées entre elles (absence de redondance). Deux idées sont à mettre en exergue dans ce schéma : (1) comment mesurer la liaison entre variables, sachant que nous nous restreignons aux cas des prédicteurs discrets ; (2) comment traduire la redondance dans un sous ensemble de variables.

Dans ce didacticiel, nous décrirons plusieurs méthodes de filtrage basées sur une mesure de corrélation pour variables discrètes. Nous les appliquerons sur un ensemble de données qui sera spécialement préparé pour mettre en évidence leur comportement. Nous évaluerons alors leurs performances en construisant le modèle bayesien naïf[5] à partir des sous-ensembles de variables sélectionnées. Nous mènerons l'expérimentation à l'aide du logiciel Tanagra ; par la suite, nous

[1] http://tutoriels-data-mining.blogspot.com/2008/03/stepdisc-analyse-discriminante.html

[2] http://tutoriels-data-mining.blogspot.com/2008/10/rgression-logistique-comparaison-de.html

[3] http://tutoriels-data-mining.blogspot.com/2009/05/strategie-wrapper-pour-la-selection-de.html ; http://tutoriels-data-mining.blogspot.com/2010/01/wrapper-pour-la-selection-de-variables.html

[4] L. Yu and H. Liu. "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution". In Proceedings of The Twentieth International Conference on Machine Leaning (ICML-03), pp 856-863, Washington, D.C., August 21-24, 2003.

[5] http://tutoriels-data-mining.blogspot.com/2010/03/le-classifieur-bayesien-naif-revisite.html

passerons en revue les méthodes filtres implémentées dans plusieurs logiciels libres de data mining (Weka 3.6.0, Orange 2.0b, RapidMiner 4.6.0, R 2.9.2 – package FSelector).

2 Méthodes filtres pour la sélection de variables

Tout le dispositif repose sur l'indicateur utilisé pour mesure le degré de liaison entre les variables. Il est donc tout à fait normal que nous le détaillions dans un premier temps. Pour ceux qui veulent approfondir l'étude des mesures d'association pour variables discrètes, un document est accessible en ligne (R. Rakotomalala, « Étude des dépendances - Variables qualitatives – *Tableau de contingence et mesures d'association* »[6]).

2.1 Corrélation entre variables discrètes

2.1.1 Mesurer l'association

Parler de corrélation s'agissant de variables discrètes peut paraître étrange. Pourtant, c'est sous cet angle que la mesure que nous décrivons dans cette section est souvent présentée dans la littérature anglo-saxonne (« correlation based measure »). Le point de départ est le tableau de contingence croisant la variable cible Y définie dans $\{y_1,\ldots,y_k,\ldots,y_K\}$ et $X \in \{x_1,\ldots,x_l,\ldots,x_L\}$ un des prédicteurs. Nous utiliserons les notations suivantes :

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

Nous pouvons en extraire différents profils : $p_{kl} = \frac{n_{kl}}{n}$; $p_{k.} = \frac{n_{k.}}{n}$; $p_{.l} = \frac{n_{.l}}{n}$.

En théorie de l'information, l'information mutuelle mesure le degré de liaison entre deux variables aléatoires. Elle s'écrit :

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

[6] http://eric.univ-lyon2.fr/~ricco/cours/cours/Dependance_Variables_Qualitatives.pdf

Notons H(Y) [resp. H(X)] l'entropie de la variable Y (resp. X). Elle mesure la quantité d'information nécessaire pour connaître les valeurs prises par la variable. Elle traduit une notion d'incertitude, mais nous pouvons également la voir sous l'angle de la dispersion. Nous écrivons :

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

On appelle alors « incertitude symétrique » (symmetrical uncertainty) la quantité suivante :

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

Elle prend ses valeurs dans [0 ; 1]. Au numérateur, nous avons quelque chose qui ressemble à une covariance, normalisée au dénominateur par les dispersions respectives des deux variables : cela n'est pas sans rappeler le coefficient de corrélation de Pearson, d'où le rapprochement souvent effectué dans la littérature.

La mesure est symétrique c.-à-d. ρ(Y,X) = ρ(X,Y). Peut être aurait-il été avantageux de privilégier un indicateur asymétrique, car nous souhaitons prédire les valeurs de Y en fonction de celles de X. Mais n'oublions pas que cette même mesure ρ sera utilisée pour calculer les corrélations croisées entre les prédicteurs. Ils ont un rôle équivalent. Un indicateur symétrique est tout à fait justifié.

Application numérique. Pour illustrer notre propos, nous prenons deux variables du fichier « vote au congrès » que nous traiterons dans ce didacticiel. Nous croisons la variable « group » (Y) avec « adoption-of-the-budget » (X). Nous obtenons le tableau de contingence suivant.

Nombre de group	tion-of-the-budget			
group	n	other	y	Total
democrat	29	7	231	267
republican	142	4	22	168
Total	171	11	253	435

En divisant toutes les valeurs par l'effectif total n = 435,

Nombre de group	tion-of-the-budget			
group	n	other	y	Total
democrat	0.0667	0.0161	0.5310	0.6138
republican	0.3264	0.0092	0.0506	0.3862
Total	0.3931	0.0253	0.5816	1.0000

Détaillons les calculs. Pour l'information mutuelle, nous avons

$$I(Y,X) = 0.0667 \times \log_2 \frac{0.0667}{0.3931 \times 0.6138} + \cdots + 0.0506 \times \log_2 \frac{0.0506}{0.5816 \times 0.3862}$$
$$= 0.4323$$

Pour l'entropie marginale de Y

$$H(Y) = -\left[0.6138 \times \log_2 0.6138 + 0.3862 \times \log_2 0.3862\right] = 0.9623$$

Et de X

$$H(Y) = -\left[0.3931 \times \log_2 0.3931 + 0.0253 \times \log_2 0.0253 + 0.5816 \times \log_2 0.5816\right] = 1.1184$$

Nous obtenons ainsi

$$\rho = 2 \times \left[\frac{I(Y,X)}{H(Y)+H(X)}\right] = 2 \times \left[\frac{0.4323}{0.9623+1.1184}\right]$$
$$= 0.4155$$

2.1.2 Tester la significativité de l'association

Lorsque les deux variables sont indépendantes, ρ = 0. Plus ρ est proche de 1, plus forte est la liaison. Nous ne savons pas en revanche à partir de quel stade nous pouvons considérer qu'elle est *suffisamment élevée*. Pour y répondre, nous utilisons un test de significativité : nous vérifions que la valeur mesurée permet de rejeter l'hypothèse d'indépendance (on parle d'hypothèse nulle Ho).

On sait (RR, page 19) que la quantité

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

Suit une loi du χ^2 à $(K-1)\times(L-1)$ degrés de libertés sous Ho entre Y et X. Au risque α, nous considérons que la liaison est statistiquement significative si $G > \chi^2_{1-\alpha}$, où $\chi^2_{1-\alpha}$ est le quantile d'ordre $(1-\alpha)$ de la loi du χ^2. Une autre manière de procéder est de calculer la probabilité critique (p-value) et de la comparer avec le risque α. Cette procédure est appelée « test du rapport de vraisemblance » dans certains logiciels.

Exemple numérique. Dans notre exemple ci-dessus, nous avons

$$G = 2 \times n \times \ln(2) \times I(Y,X)$$
$$= 2 \times 435 \times \ln(2) \times 0.4323$$
$$= 260.7046$$

Avec une loi du χ^2 à (3-1) x (2-1) = 2 degrés de liberté, nous obtenons une probabilité critique $p_c = 3.16 \times 10^{-56} \approx 0$. L'association semble très significative.

***Remarque** : De toute manière, la liaison mesurée est presque toujours significative dès que les effectifs augmentent. Il ne faut pas trop se focaliser sur ce test dans le cadre du data mining. Ce qui importe avant tout est de pouvoir nous appuyer sur l'indicateur ρ pour détecter les variables prédictives pertinentes.*

2.1.3 Tester la significativité de l'association (bis)

Il y a une autre manière de tester la significativité du coefficient. En utilisant l'approximation normale, nous pouvons à la fois reproduire l'analyse ci-dessus, sachant que l'approche est nettement plus conservatrice (elle favorise l'hypothèse nulle), et produire un intervalle de variation pour un niveau de confiance donné.

Posons

$$H(Y,X) = -\sum_k \sum_l p_{kl} \times \log_2 p_{kl}$$

L'expression générique de la variance de ρ s'écrit[7]

$$\sigma_\rho^2 = 4 \times \sum_k \sum_l \frac{n_{kl}\left[H(Y,X) \times \log_2\left(\frac{n_{k.} \times n_{.l}}{n^2}\right) - [H(Y)+H(X)] \times \log_2\left(\frac{n_{kl}}{n}\right)\right]^2}{n^2 \times [H(Y)+H(X)]^4}$$

Sous l'hypothèse nulle Ho (indépendance), elle devient

$$\sigma_\rho^2(0) = 4 \times \frac{\sum_k \sum_l n_{kl} \times \left[\log_2\left(\frac{n_{k.} \times n_{.l}}{n \times n_{kl}}\right)\right]^2 - \frac{[H(Y)+H(X)-H(Y,X)]^2}{n}}{n^2 \times [H(Y)+H(X)]^2}$$

On rejettera l'hypothèse nulle au risque α si

$$\frac{\rho}{\sigma_\rho(0)} > u_{1-\alpha}$$

$u_{1-\alpha}$ est le quantile d'ordre (1- α) de la loi normale centrée réduite.

Exemple numérique. Reprenons notre exemple « groups vs. adoption-of-budget ». Nous avons H(Y,X) = 1.6484 ; nous formons la variance asymptotique,

$$\sigma_\rho^2 = \frac{4 \times 1626.6563}{435^2 \times [0.9623+1.1184]^4} = \frac{6506.6252}{3546865.1306} = 0.0018$$

L'intervalle de variation au niveau de confiance (1- α)=95% est obtenu avec

$$\left[\rho \pm u_{1-\alpha/2} \times \sigma_\rho\right]$$
$$[0.415544 - 1.96 \times 0.0428 \; ; \; 0.415544 + 1.96 \times 0.0428]$$
$$[0.3316 \; ; \; 0.4995]$$

Pour tester la significativité, nous calculons la variance sous Ho,

$$\sigma_\rho^2(0) = 4 \times \frac{450.7425 - \frac{(0.9623+1.1184-1.6484)^2}{435}}{435^2 \times [0.9623+1.1184]^2} = 0.0022$$

Nous rejetons l'hypothèse nulle au risque α car

$$\frac{\rho}{\sigma_\rho(0)} = \frac{0.415544}{0.0469} = 8.8579 > u_{1-\alpha} = 1.6449$$

A l'instar du test du χ^2 basée sur la statistique G, le lien entre « group » et « adoption-of-budget » est jugé très fortement significatif.

2.2 Méthodes de « ranking » basées sur la corrélation

Les méthodes de « ranking » s'appuient uniquement sur la notion de pertinence, elles ignorent totalement la redondance. En pratique, il s'agit de calculer la corrélation de chaque prédicteur avec

[7] http://v8doc.sas.com/sashtml/stat/chap28/sect20.htm

la variable cible, puis de le classer par ordre décroissant. Pour déterminer le nombre de variables à retenir, nous pouvons utiliser le test de significativité décrit précédemment. Mais on se rend compte qu'il a tendance à beaucoup trop en accepter surtout lorsque les effectifs augmentent, sans compter que le test est lui-même biaisé par le fait que nous avons trié les variables selon leur importance et que multiplions les comparaisons pour déterminer le nombre adéquat de variables. Le risque réel n'est pas celui utilisé pour le test. D'autres approches, plus empiriques sont possibles : fixer un seuil directement sur l'indicateur ρ, pourquoi pas après tout ; détecter le « décrochage » dans les valeurs de ρ, en considérant que la liaison d'une variable doit être évaluée en fonction de celle des autres ; enfin, spécifier directement (arbitrairement) le nombre de variables à retenir.

Calculations details

N°	Attribute	Values	Statistic	Statistic (Histogram)	p-value
1	physician-fee-freeze	3	0.708862		0.000000
2	corr_physician-fee-freeze	3	0.540679		0.000000
3	adoption-of-the-budget	3	0.415544		0.000000
4	el-salvador-aid	3	0.394048		0.000000
5	corr_adoption-of-the-budget-re	3	0.371640		0.000000
6	corr_el-salvador-aid	3	0.366040		0.000000
7	education-spending	3	0.333286		0.000000
8	aid-to-nicaraguan-contras	3	0.319763		0.000000
9	crime	3	0.313788		0.000000
10	corr_aid-to-nicaraguan-contras	3	0.288226		0.000000
11	corr_crime	3	0.287527		0.000000
12	mx-missile	3	0.282252		0.000000
13	corr_education-spending	3	0.273481		0.000000
14	corr_mx-missile	3	0.269558		0.000000
15	superfund-right-to-sue	3	0.205050		0.000000
16	duty-free-exports	3	0.197825		0.000000
17	corr_duty-free-exports	3	0.194450		0.000000
18	anti-satellite-test-ban	3	0.186272		0.000000
19	corr_superfund-right-to-sue	3	0.179718		0.000000
20	corr_anti-satellite-test-ban	3	0.160502		0.000000

Dans cet exemple (que nous détaillerons plus loin dans ce didacticiel), nous affichons les corrélations des 20 meilleures variables prédictives sur 48. Elles sont toutes significatives apparemment. On note néanmoins des décrochages dans les valeurs de ρ : après la première variable ; après les 2 premières ; puis après la 14ème variable. Ce sont autant d'indications qui peuvent nous aider dans la détermination du nombre de variables à retenir. On se rend compte que la tâche n'est pas aisée quoiqu'il en soit.

Mais le principal reproche que l'on peut adresser à cette approche est l'absence totale de prise en compte de la redondance. De fait, si un prédicteur significatif est dupliqué 10 fois, tous seront sélectionnés. C'est un sacré problème, surtout lorsque nous devons faire face à des jeux de données comportant un grand nombre de variables bruitées et redondantes. Il est évident qu'une expertise manuelle des résultats est impossible.

2.3 Méthodes de filtrage basées sur la corrélation

Il faut prendre en compte la pertinence des variables prédictives, mais il faut aussi tenir compte de leur redondance. Le sous-ensemble de variables sélectionnées doit être le plus parcimonieux possible. Il nous faut donc combiner de manière appropriée la corrélation des prédicteurs avec la variable cible d'une part, et les corrélations croisées entre les variables sélectionnées d'autre part. Toutes les méthodes présentées dans cette section reposent sur cette idée.

2.3.1 La méthode CFS

La méthode CFS[8] repose sur une mesure globale de « mérite » d'un sous-ensemble M de *m* variables qui tient compte à la fois de leur pertinence et de leur redondance. Elle s'écrit

$$merit = \frac{m \times \overline{\rho}_{y,x}}{\sqrt{m + m \times (m-1) \times \overline{\rho}_{x,x}}}$$

Où $\overline{\rho}_{y,x}$ est la moyenne des corrélations entre les variables prédictives et la variable cible ; $\overline{\rho}_{x,x}$, la moyenne des corrélations croisées entre les variables prédictives.

Ainsi, le problème de sélection devient un problème d'optimisation. Nous devons maximiser la quantité MERIT à partir de l'ensemble des variables candidates. Nous pouvons utiliser des stratégies gloutonnes simples (méthodes pas à pas, forward ou backward) ou sophistiquées (ex. algorithmes génétiques, recuit simulé, etc.). En pratique, une technique simple, en lissant l'exploration de l'espace des solutions, suffit amplement. Elle évite l'écueil du sur apprentissage.

Complexité de calcul. L'algorithme (sélection gloutonne « forward ») est linéaire par rapport au nombre d'observations. Toutes les corrélations peuvent être pré calculées en une seule passe sur les données. En revanche, il est quadratique par rapport au nombre de descripteurs. Il est donc surtout avantageux sur les très grandes bases de données comportant un nombre important d'observations mais relativement peu de descripteurs.

Dans le cas contraire, lorsque les descripteurs sont très nombreux, le calcul et le stockage en mémoire de toutes les corrélations croisées deviennent un problème. Il devient plus avantageux de calculer (et recalculer) les corrélations à la volée. Les expérimentations montrent que le nombre de variables finalement sélectionnées est souvent faible.

[8] M. Hall, S. Lloyd, « Feature subset selection : a correlation based filter approach », in 1997 Int. Conf. On Neural Information Processing and Intelligent Information Systems, pp/ 855-858, Springer, 1997.

2.3.2 La méthode MIFS

La méthode MIFS (Mutual Information Feature Selection) est due à Battiti (1994)[9]. Elle s'appuie sur un algorithme pas à pas « forward ». Le critère d'évaluation de l'adjonction d'une variable supplémentaire X à l'ensemble M (de cardinal *m*) des variables déjà sélectionnées s'écrit

$$I(Y,X/M) = I(Y,X) - \beta \times \sum_{Z \in M} \frac{I(X,Z)}{m}$$

A chaque étape, on choisit la variable qui maximise la quantité I(Y, X / M), *qui est une sorte d'**information mutuelle partielle***. Une variable est jugée intéressante si sa liaison avec la cible Y surpasse sa liaison moyenne avec les prédicteurs déjà sélectionnés. On tient compte à la fois de la pertinence et de la redondance. La recherche s'arrête lorsque la meilleure variable X* est telle que $I(Y,X*/M) \leq 0$.

A vrai dire, MIFS n'apporte rien de plus par rapport à CFS. Il a à la fois l'avantage et l'inconvénient d'être paramétrable avec β. « L'avantage » parce que nous pouvons le piloter en fonction des caractéristiques des données et des objectifs de l'étude. Nous augmenterons β par exemple lorsque nous souhaitons que les prédicteurs soient très peu redondants. « L'inconvénient » parce que dans la pratique, le comportement de l'algorithme est trop fortement dépendant de la valeur de β. Et il est particulièrement difficile à manipuler. Nous sommes obligés de multiplier les essais pour déterminer la valeur adéquate, avec tous les risques de sur dépendance (sur apprentissage) au fichier de données utilisé.

Enfin, l'algorithme de sélection est lui aussi quadratique par rapport au nombre de variables de la base.

2.3.3 La méthode FCBF

FCBF (Yu et Liu, ICML 2003) est également basé sur le critère « symmetrical uncertainty -- ρ». Il se distingue en revanche par la stratégie de recherche mise en œuvre, basée sur la notion de « prédominance ». On dit que la corrélation entre une variable X* et la cible Y est prédominante si et seulement si

$$\rho_{y,x*} \geq \delta \text{ et } \forall X (X \neq X*), \rho_{x,x*} < \rho_{y,x*}$$

Concrètement, une variable est jugée intéressant si : (1) sa corrélation avec la cible est suffisamment élevée, δ est le paramètre qui permet de moduler cela ; (2) il n'existe pas dans la base une variable qui lui soit plus fortement corrélée.

Partant de cette définition, les auteurs ont élaboré un algorithme ad hoc qui est de **complexité quasi-linéaire** :

```
1. S l'ensemble des prédicteurs candidats, M = ∅ les prédicteurs sélectionnés
```

[9] R. Battiti, « Using mutual information for selecting features in supervised neural net learning », IEEE Transactions on Neural Networks, 5(4) : 537-550, 1994.

2. Chercher la variable X* (parmi S) maximisant sa corrélation avec Y → ρ_{y,x^*}
3. Si $\rho_{y,x^*} \geq \delta$, rajouter X* dans M et retirer X* de S
4. Retirer également de S toutes les variables X telles que $\rho_{x^*,x} \geq \rho_{y,x}$
5. Si S ≠∅ alors retourner en (2), sinon FIN de l'algorithme

En termes de temps de calcul, l'approche est particulièrement intéressante, surtout lorsque que nous avons à traiter des bases intégrant des milliers de prédicteurs candidats ; en termes de capacité à détecter les « bonnes » variables, les expérimentations montrent qu'elle se comporte tout à fait honorablement par rapport aux autres approches décrites dans ce didacticiel.

2.3.4 La méthode MODTREE

La méthode MODTREE (Rakotomalala et Lallich, 2002)[10] repose également sur les notions de pertinence et de redondance. Mais elle n'utilise pas la même mesure de la corrélation. En s'appuyant sur le principe de la comparaison par paires, les auteurs définissent la corrélation entre deux variables discrètes de la manière suivante :

$$r_{y,x} = \frac{g_{11}g_{22} - g_{12}g_{21}}{\sqrt{g_{1.} \times g_{2.} \times g_{.1} \times g_{.2}}}, \text{ avec}$$

- $g_{11} = \frac{1}{2}\sum_k \sum_l n_{kl}^2$
- $g_{12} = \frac{1}{2}\sum_k \sum_l n_{kl}(n_{k.} - n_{kl})$
- $g_{1.} = g_{11} + g_{12}$
- $g_{21} = \frac{1}{2}\sum_k \sum_l n_{kl}(n_{.l} - n_{kl})$
- $g_{22} = \frac{1}{2}\sum_k \sum_l n_{kl}(n - n_{k.} - n_{.l} + n_{kl})$
- $g_{2.} = g_{21} + g_{22}$
- $g_{.1} = g_{11} + g_{21}$
- $g_{.2} = g_{12} + g_{22}$

Le calcul est linéaire en nombre d'observations n, même si le critère est basé sur le principe des comparaisons par paires. C'est ce qui le rend opérationnel pour le traitement des bases comprenant un grand nombre de lignes.

Pour réaliser la sélection pas à pas « forward », nous utilisons la corrélation partielle. Elle mesure le degré de liaison entre deux variables Y et X en retranchant l'effet d'une tierce variable Z (http://en.wikipedia.org/wiki/Partial_correlation). Elle s'écrit

[10] Rakotomalala R., Lallich S., "Construction d'arbres de décision par optimisation", Revue Extraction des Connaissances et Apprentissage, vol. 16, n°6/2002, pp.685-703, 2002.

$$r_{y,x/z} = \frac{r_{y,x} - r_{y,z} \times r_{x,z}}{\sqrt{\left(1 - r_{y,z}^2\right)\left(1 - r_{x,z}^2\right)}}$$

Lorsque *m* variables M = {Z_1,..., Z_m} sont déjà sélectionnées, la corrélation partielle s'écrit

$$r_{y,x/z_1 \ldots z_m} = \frac{r_{y,x/z_1 \ldots z_{m-1}} - r_{y,z_m/z_1 \ldots z_{m-1}} \times r_{x,z_m/z_1 \ldots z_{m-1}}}{\sqrt{\left(1 - r_{y,z_m/z_1 \ldots z_{m-1}}^2\right)\left(1 - r_{x,z_m/z_1 \ldots z_{m-1}}^2\right)}}$$

A l'instar de CFS et MIFS, l'algorithme est quadratique en nombre de prédicteurs candidats. On note surtout qu'il nous oblige à calculer la table des corrélations partielles croisées (au départ, il s'agit simplement des corrélations brutes croisées), que l'on doit remettre à jour à chaque fois que l'on ajoute une nouvelle variable dans l'ensemble M. L'encombrement mémoire et le temps de calcul deviennent des contraintes fortes lorsque nous traitons des bases contenant un grand nombre de descripteurs.

Comme pour tout algorithme glouton, il faut définir une règle d'arrêt. A l'étape *m+1* c.-à-d. *m* variables ont déjà été sélectionnées, nous stoppons le processus si la meilleure variable X* maximisant la corrélation partielle est telle que

$$r_{y,x^*/z_1 \ldots z_m} < \frac{1}{\sqrt{n-m}}$$

Par rapport aux méthodes CFS et FCBF, les expérimentations montrent que MODTREE sait également détecter les prédicteurs les plus intéressants.

Exemple numérique. Étudions le mécanisme de la corrélation partielle à l'aide des 4 variables de la base « vote » que nous présenterons de manière détaillée dans la section suivante : Y = « group », X1 = « physician fee freeze », X2 = « adoption-of-budget » ; X3 = « education spending ».

Nous calculons les corrélations brutes (ou corrélations d'ordre 0) entre les variables.

Y	X	r	r²
adoption-of-the-budget	physician-fee-freeze	0.5328	0.2838
adoption-of-the-budget	education-spending	0.4005	0.1604
adoption-of-the-budget	group	0.5464	0.2986
physician-fee-freeze	education-spending	0.4529	0.2051
physician-fee-freeze	group	0.8097	0.6556
education-spending	group	0.4545	0.2066

(Y, X1) Dans un premier temps, $\boxed{r_{y,x_1} = 0.8097}$ semble indiquer une corrélation significative.

(Y, X2 / X1) Voyons l'information additionnelle apportée par X2, sachant que X1 est entérinée. Nous extrayons $r_{x_2,x_1} = 0.5328$ et $r_{y,x_2} = 0.5464$ du tableau ci-dessus, nous formons alors

$$r_{y,x_2/x_1} = \frac{0.5464 - 0.8097 \times 0.5328}{\sqrt{\left(1-0.8097^2\right)\left(1-0.5328^2\right)}} = 0.2316$$

La liaison est moins marquée parce que la variable X2 est corrélée à la fois avec Y et X1.

(Y, X3 / X1, X2) Si nous voulons mesurer le lien entre Y et X3 en neutralisant les variables X1 et X2, il nous faut d'abord calculer

$$r_{y,x_3/x_1} = \frac{r_{y,x_3} - r_{y,x_1} \times r_{x_3,x_{11}}}{\sqrt{\left(1-r_{y,x_1}^2\right)\left(1-r_{x_3,x_1}^2\right)}} = \frac{0.4545 - 0.8097 \times 0.4529}{\sqrt{\left(1-0.8097^2\right)\left(1-0.4529^2\right)}} = 0.1678$$

$$r_{x_2,x_3/x_1} = \frac{r_{x_2,x_3} - r_{x_2,x_1} \times r_{x_3,x_1}}{\sqrt{\left(1-r_{x_2,x_1}^2\right)\left(1-r_{x_3,x_1}^2\right)}} = \frac{0.4005 - 0.5328 \times 0.4529}{\sqrt{\left(1-0.5328^2\right)\left(1-0.4529^2\right)}} = 0.2110$$

Nous pouvons alors former la corrélation partielle d'ordre 2.

$$r_{y,x_3/x_1,x_2} = \frac{r_{y,x_3/x_1} - r_{y,x_2/x_1} \times r_{x_3,x_2/x_1}}{\sqrt{\left(1-r_{y,x_2/x_1}^2\right)\left(1-r_{x_3,x_2/x_1}^2\right)}} = \frac{0.1678 - 0.2316 \times 0.2110}{\sqrt{\left(1-0.2316^2\right)\left(1-0.2110^2\right)}} = 0.1251$$

Dans l'explication de Y, l'apport additionnel de X3 par rapport à X1 et X2 est positif.

3 Données

Pour étudier le comportement des méthodes décrites dans ce tutoriel, nous utilisons le fichier des « Votes au congrès » (http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records). L'objectif est de prédire le groupe d'appartenance politique de parlementaires à partir des votes qu'ils ont effectué sur différents sujets. ***Mais nous avons quelque peu modifié la base***. Nous avons dupliqué les 16 descripteurs originels de deux manières :

1. Pour chaque variable « X », nous avons généré aléatoirement « noise_X » définie dans le même ensemble de valeurs et en respectant (approximativement) les fréquences. « X » et « noise _X » n'ont absolument aucun lien.

Prenons le cas de la variable « adoption-of-budget ». Les distributions de fréquences sont proches,

Nombre de adoption-of-the-budget	
adoption-of-the-budget	Total
n	39.31%
other	2.53%
y	58.16%
Total	100.00%

Nombre de noise_adoption-of-the-budget-re	
noise_adoption-of-the-budget-re	Total
n	39.77%
other	4.60%
y	55.63%
Total	100.00%

Mais les variables ne sont absolument pas liées comme le montre le test d'indépendance du KHI-2.

Results

Row (Y)	Column (X)	Statistical indicator
adoption-of-the-budget	noise_adoption-of-the-budget-re	

Stat	Value
d.f.	4
Tschuprow's t	0.054835
Cramer's v	0.054835
Phi²	0.006014
Chi² (p-value)	2.62 (0.6240)
Lambda	0.000000
Tau (p-value)	0.0030 (0.6249)
U(R/C) (p-value)	0.0045 (0.5465)

Cross-tab

	y	other	n	Sum
n	90	9	72	171
y	147	11	95	253
other	5	0	6	11
Sum	242	20	173	435

2. Toujours pour chaque variable « X », nous avons généré « corr_X » qui lui est liée, indépendamment des valeurs prises par la cible Y. Concrètement, nous avons construit corr_X de manière à ce qu'elle ait les mêmes valeurs que X dans (approximativement) 97% des cas.

Voyons ce qu'il en est pour « adoption-of-budget », nous constatons que les deux variables sont effectivement fortement liées.

Results

Row (Y)	Column (X)	Statistical indicator
adoption-of-the-budget	corr_adoption-of-the-budget-re	

Stat	Value
d.f.	4
Tschuprow's t	0.978187
Cramer's v	0.978187
Phi²	1.913699
Chi² (p-value)	832.46 (0.0000)
Lambda	0.950549
Tau (p-value)	0.9201 (0.0000)
U(R/C) (p-value)	0.8710 (0.0000)

Cross-tab

	n	y	other	Sum
n	165	6	0	171
y	3	250	0	253
other	0	0	11	11
Sum	168	256	11	435

L'objectif de ce didacticiel est de montrer la capacité des algorithmes de sélection de prédicteurs : (1) à discerner ceux qui sont les plus efficaces parmi les descripteurs originels ; (2) à écarter les attributs correspondant simplement à du bruit (« noise_ ») ; (3) à ne pas être induit en erreur par les leurres que constituent les variables redondantes (« corr_ ») c.-à-d. à préférer X à corr_X durant la sélection.

Bref, une méthode sera d'autant meilleure qu'elle intègre le plus petit nombre de prédicteurs, en excluant notamment à la fois les variables « noise » et « corr ». **Néanmoins, pour éviter une sélection trop restrictive, il faudra par la suite que ces prédicteurs permettent la construction d'un classifieur performant. Nous utiliserons le modèle « bayesien naïf » dans ce didacticiel**.

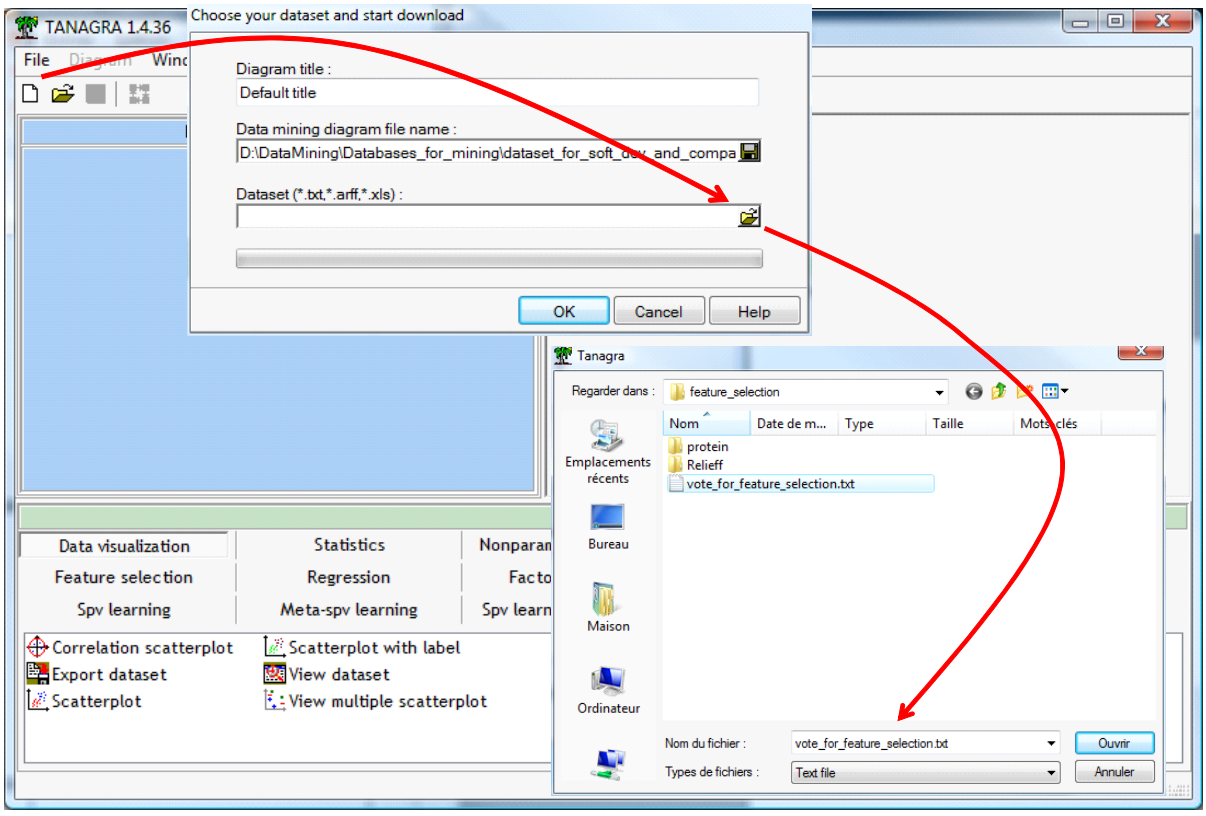
4 Filtrage des prédicteurs discrets avec Tanagra

4.1 Construction et évaluation du modèle complet

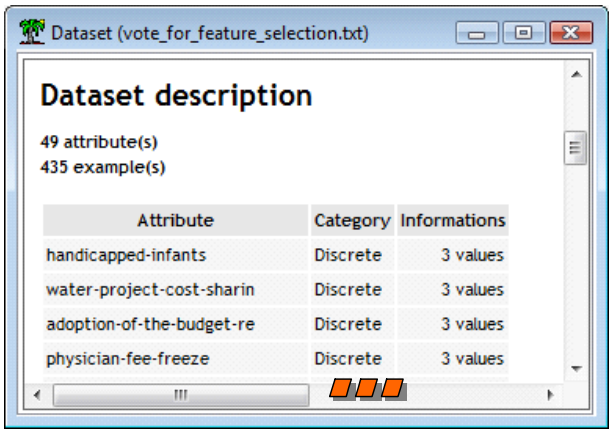
Dans un premier temps, nous allons construire le modèle d'indépendance conditionnelle (modèle bayesien naïf – voir http://tutoriels-data-mining.blogspot.com/2010/03/le-classifieur-bayesien-naif-revisite.html) à partir de la totalité des (16 x 3 = 48) variables disponibles. Nous évaluerons les performances en généralisation via le taux d'erreur bootstrap.

4.1.1 Importation des données

Après avoir démarré Tanagra, nous créons un nouveau diagramme avec FILE / NEW. Nous sélectionnons le fichier « vote_for_feature_selection.txt ».

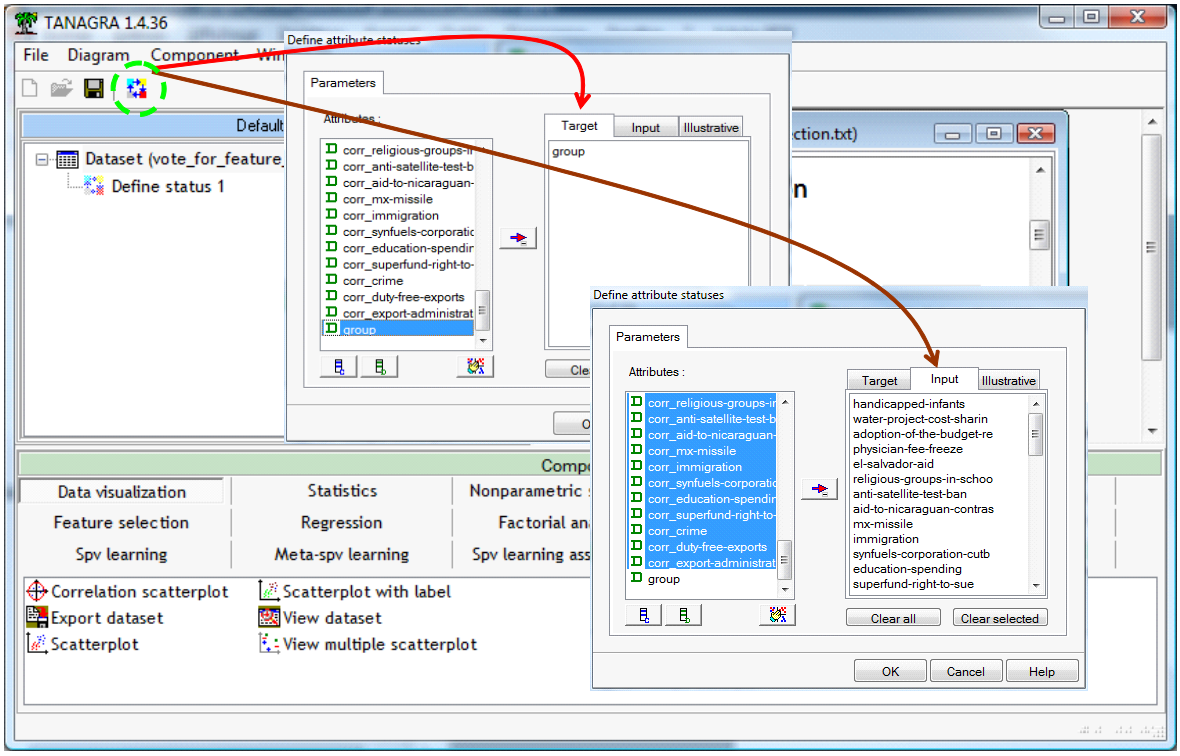


Nous validons. L'importation est lancée : 435 observations et 49 variables (cible + prédicteurs potentiels) sont disponibles.



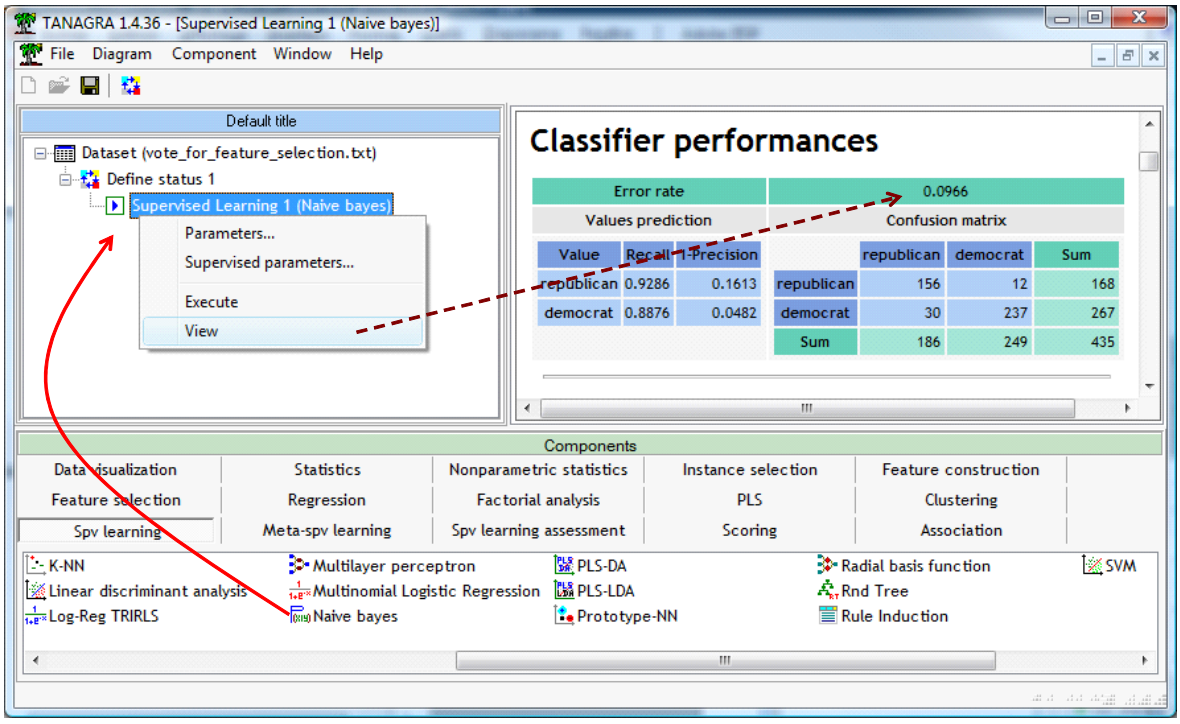
4.1.2 Modélisation

Nous souhaitons construire le modèle bayesien naïf en intégrant toutes les variables prédictives candidates. Nous insérons le composant DEFINE STATUS via le raccourci dans la barre d'outils.



Nous plaçons GROUP en TARGET, les autres variables en INPUT.

Nous ajoutons le composant NAIVE BAYES (onglet SUPERVISED LEARNING). Nous actionnons directement le menu VIEW. Nous obtenons le modèle suivant.

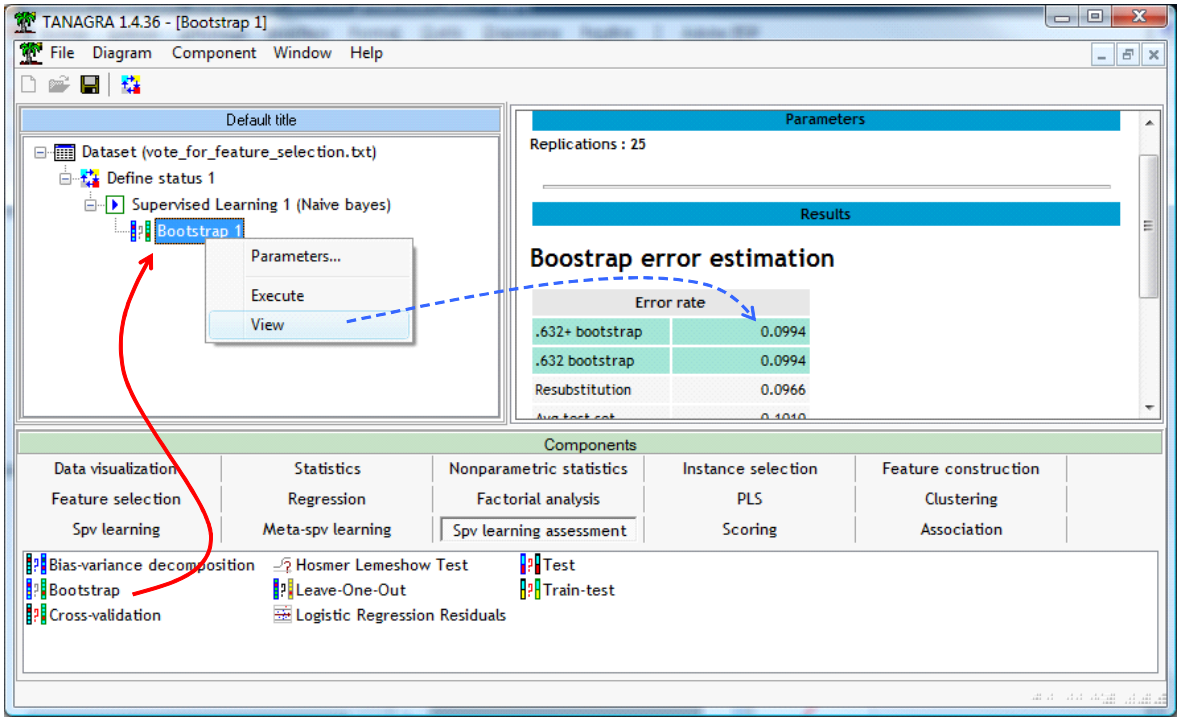


Le taux d'erreur en resubstitution est de 9.66%. Nous savons qu'il est souvent optimiste.

4.1.3 Évaluation des performances – Taux d'erreur bootstrap

Utilisons le bootstrap pour obtenir une évaluation fiable des performances en généralisation (voir http://tutoriels-data-mining.blogspot.com/2008/03/validation-croise-bootstrap-leave-one.html).

Nous ajoutons le composant BOOTSTRAP (SPV LEARNING ASSESSMENT). Nous actionnons le menu VIEW.

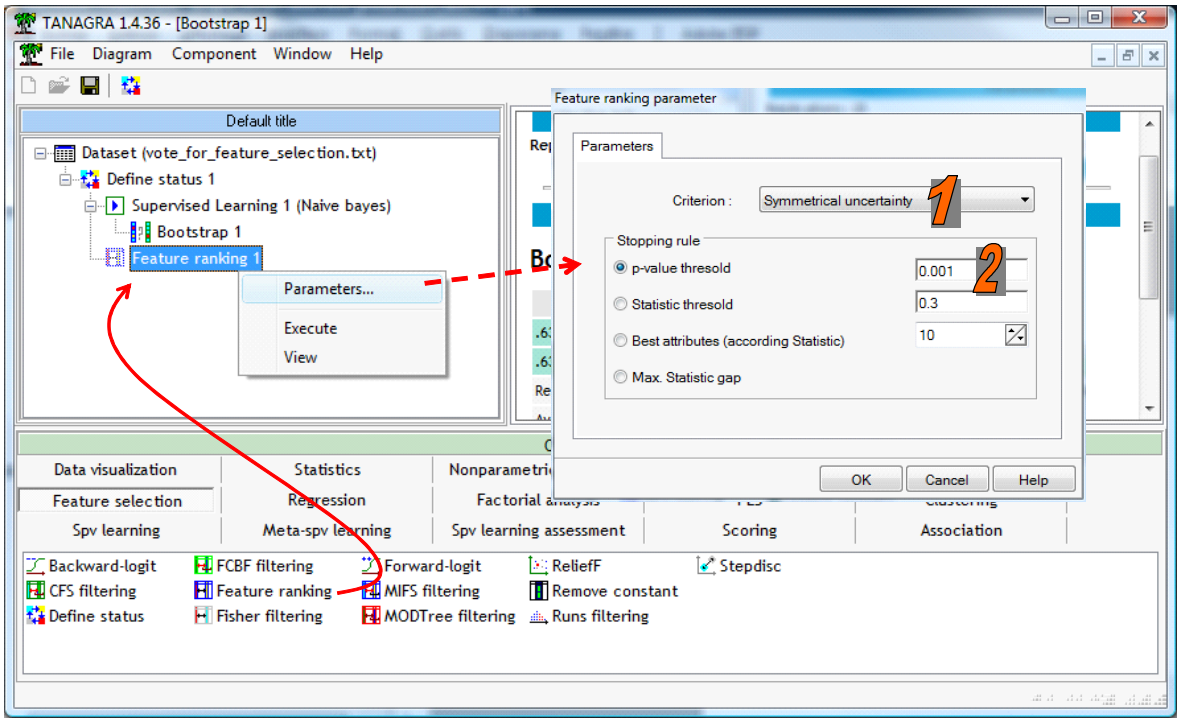


Le véritable taux d'erreur serait plutôt de 9.94%. Il nous servira de référence dans notre expérimentation. Nous cherchons à savoir si les méthodes de sélection sont capables de réduire le nombre de prédicteurs tout en préservant (au pire) ou en améliorant (au mieux) les performances en classement.

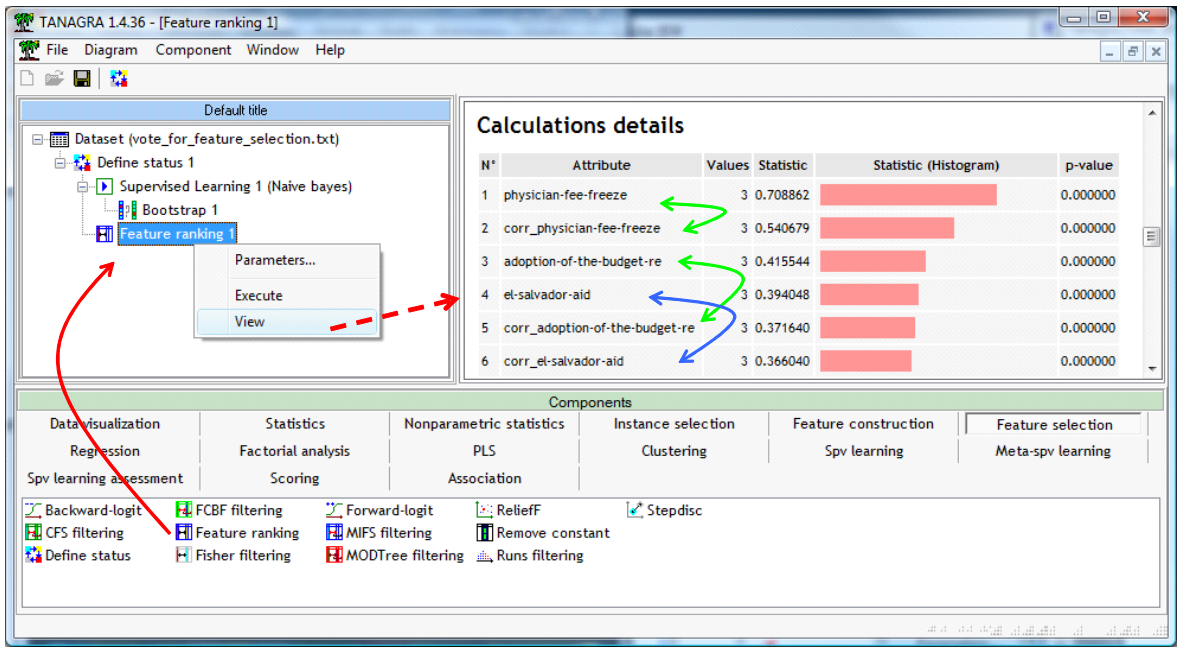
4.2 Comportement des méthodes de sélection de variables

4.2.1 Sélection basé sur le « ranking »

Nous introduisons le composant FAETURE RANKING (onglet FEATURE SELECTION) à la suite de DEFINE STATUS 1. Nous le paramétrons de manière (1) à utiliser le critère « symmetrical uncertainty » durant le processus de sélection. Seules les variables significatives au sens de la statistique G (section 2.1.2 – on l'appelle également test du rapport de vraisemblance dans certains logiciels) sont retenues. Nous utilisons (2) le niveau de signification $\alpha = 0.001$ pour restreindre la sélection.

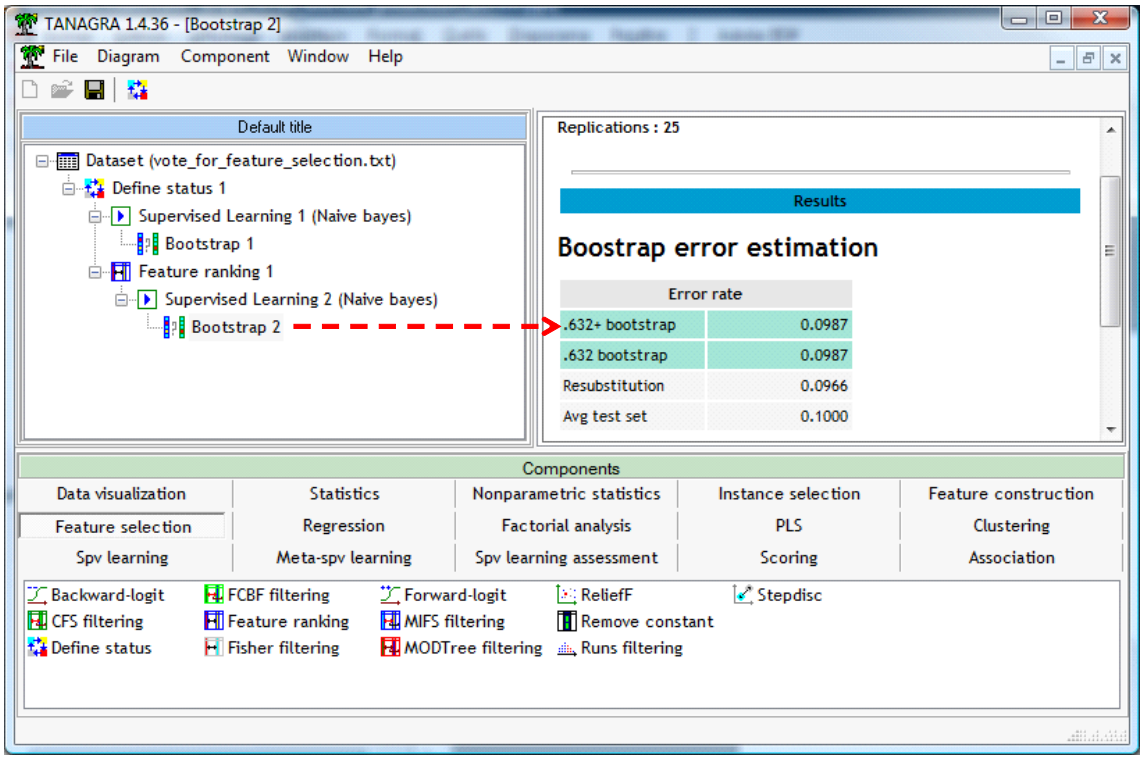


Nous validons puis nous cliquons sur le menu contextuel VIEW. Patatras, **28 variables ont été retenues**. Nous constatons que les variables « corr_ » accompagnent quasi systématiquement leurs homologues. La technique n'appréhende pas du tout le problème de la redondance.



Aucune variable « noise_ » en revanche ne s'est immiscée dans la solution. Les méthodes de « ranking » gèrent bien la pertinence. Avec le bémol du paramétrage compliqué à définir : comment fixer la valeur seuil α ? Les valeurs usuelles en statistique (5%, 1%,…) sont manifestement inadaptées.

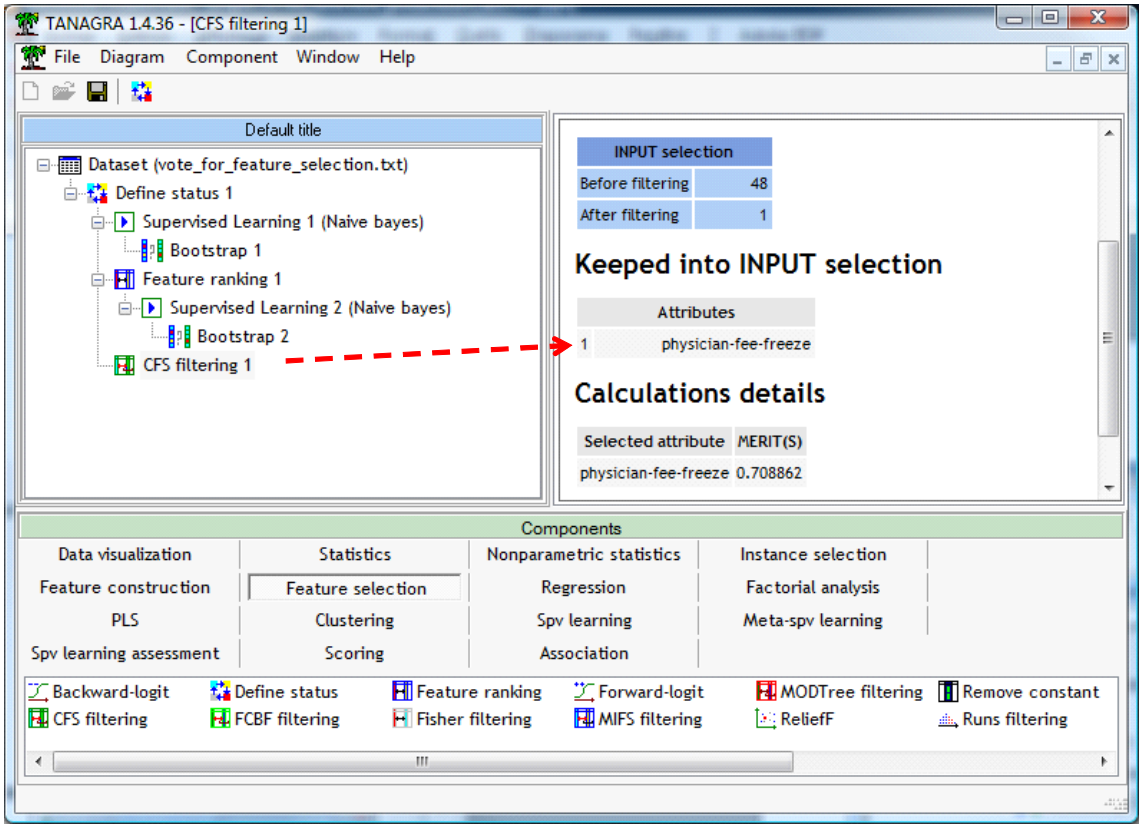
Nous insérons de nouveau les composants NAIVE BAYES et BOOTSTRAP. Le modèle est construit sur les variables sélectionnées par FEATURE RANKING à chaque étape du bootstrap, le taux d'erreur qui en résulte est 9.87%.



Nous reproduirons la même démarche expérimentale pour les autres techniques de sélection.

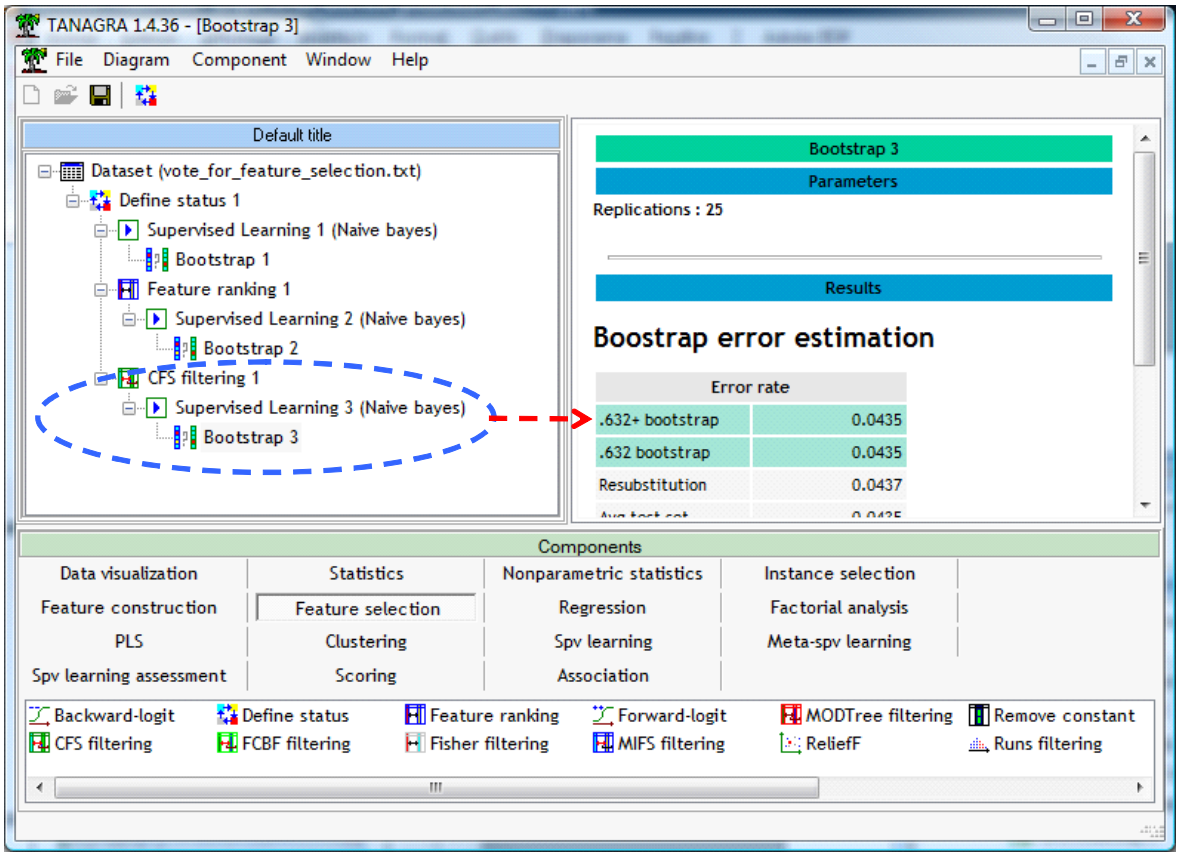
4.2.2 Sélection avec CFS

Nous insérons le composant CFS FILTERING (onglet FEATURE SELECTION) dans le diagramme. Il n'y a pas de paramétrage à effectuer. Nous actionnons directement le menu VIEW.



Une seule variable, « physician-fee-freeze », a été sélectionnée (MERIT = 0.709). L'adjonction d'une seconde variable aurait diminué le critère MERIT. Les variables « noise » et « corr » ont bien été évacuées.

Avec le modèle d'indépendance conditionnel, le taux d'erreur bootstrap est de 4.35%.

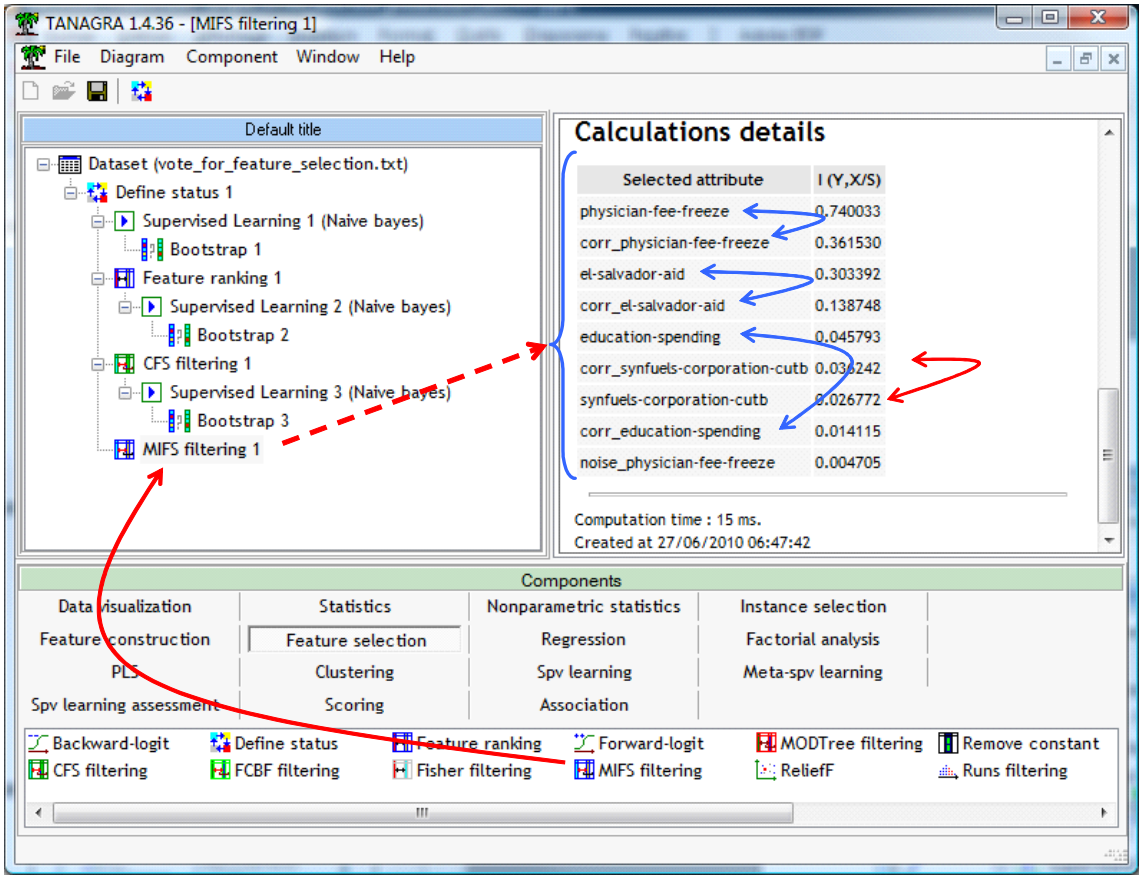


Non seulement la méthode a réduit drastiquement le nombre de prédicteurs, mais il améliore dans des proportions considérables les performances du classifieur. C'est le schéma idéal dans un processus de sélection de variables.

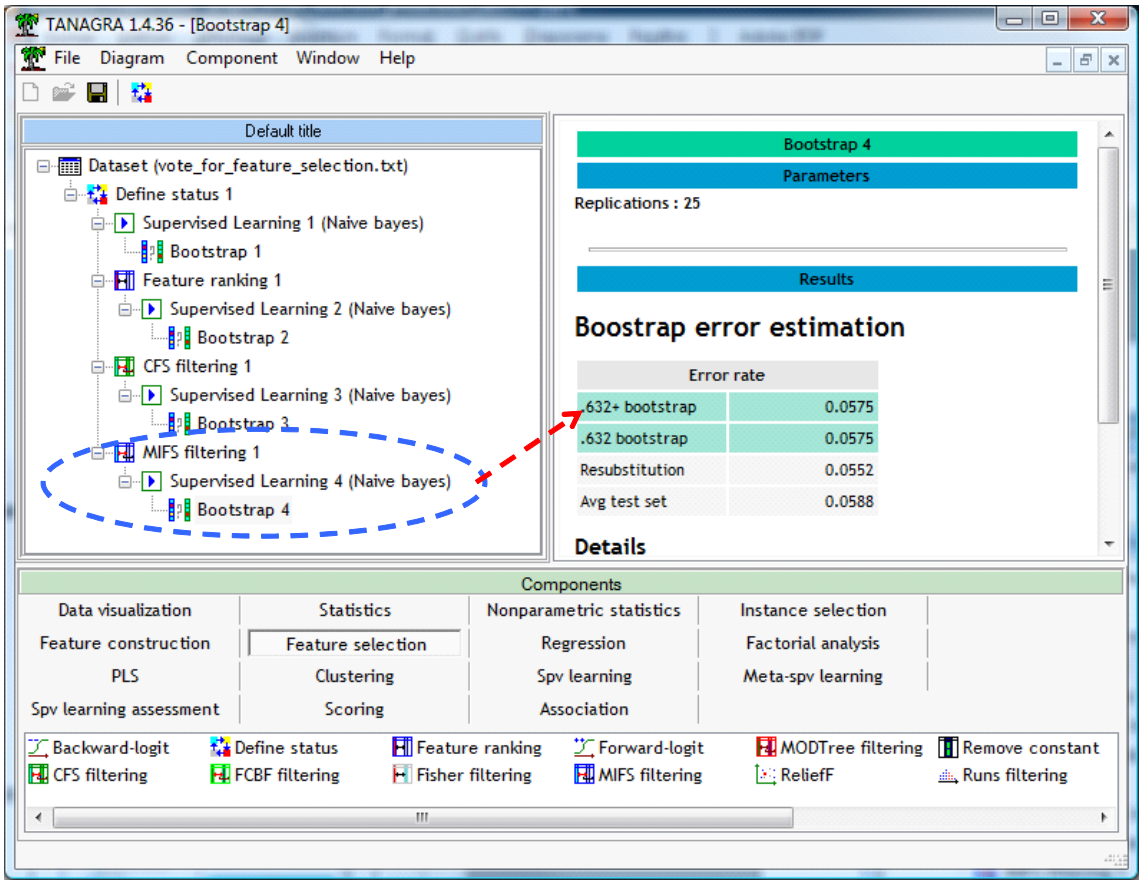
4.2.3 Sélection avec MIFS

Nous réitérons le même schéma pour la méthode MIFS avec le paramétrage par défaut (β = 1.5). Neuf (9) variables sont sélectionnées, parmi lesquelles les descripteurs corrélés (corr). C'est clairement une déception, surtout si l'on considère qu'ils s'intercalent entre les prédicteurs originels dans le processus de sélection (voir ***Calculations details***).

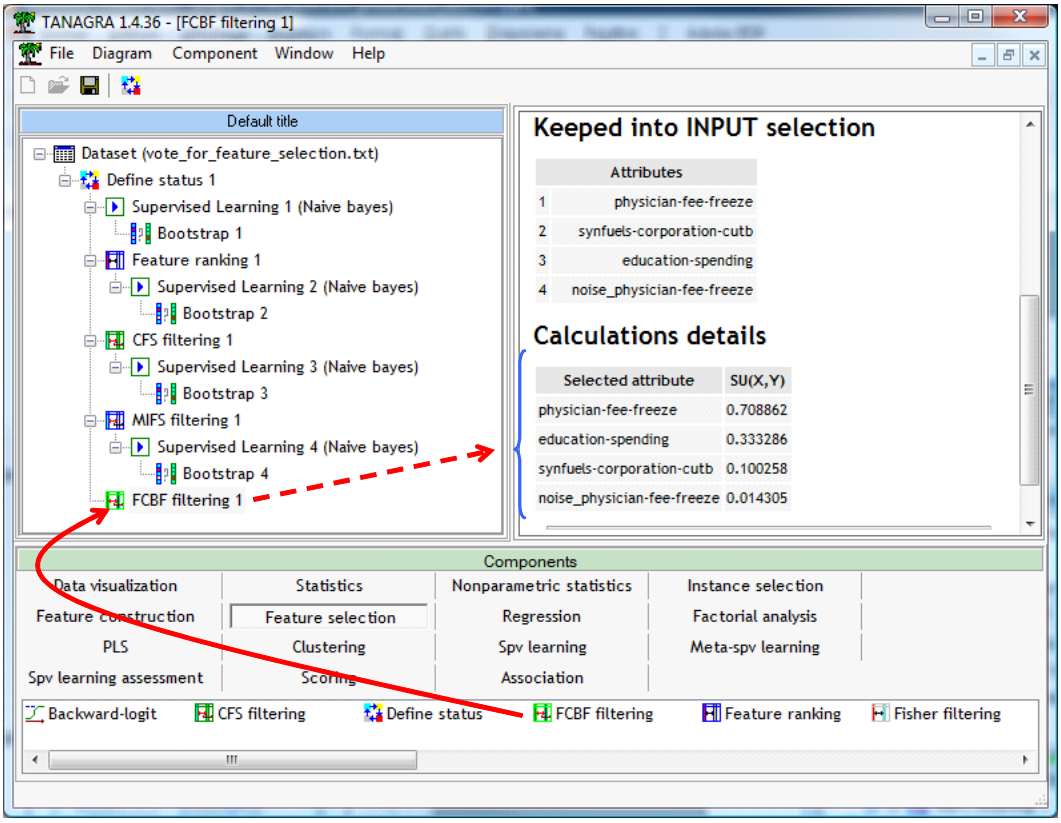
Manifestement, le paramètre β est mal réglé. Il faut littéralement « serrer la vis » pour pénaliser plus fortement la redondance. Plus on augmente β, moins il y aura de variables redondantes dans la sélection, et par conséquent, moins il y aura de variables *tout court*. Mais définir a priori la valeur de β est difficile car nous n'avons pas de référence explicite. Le tâtonnement semble être la seule voie possible.



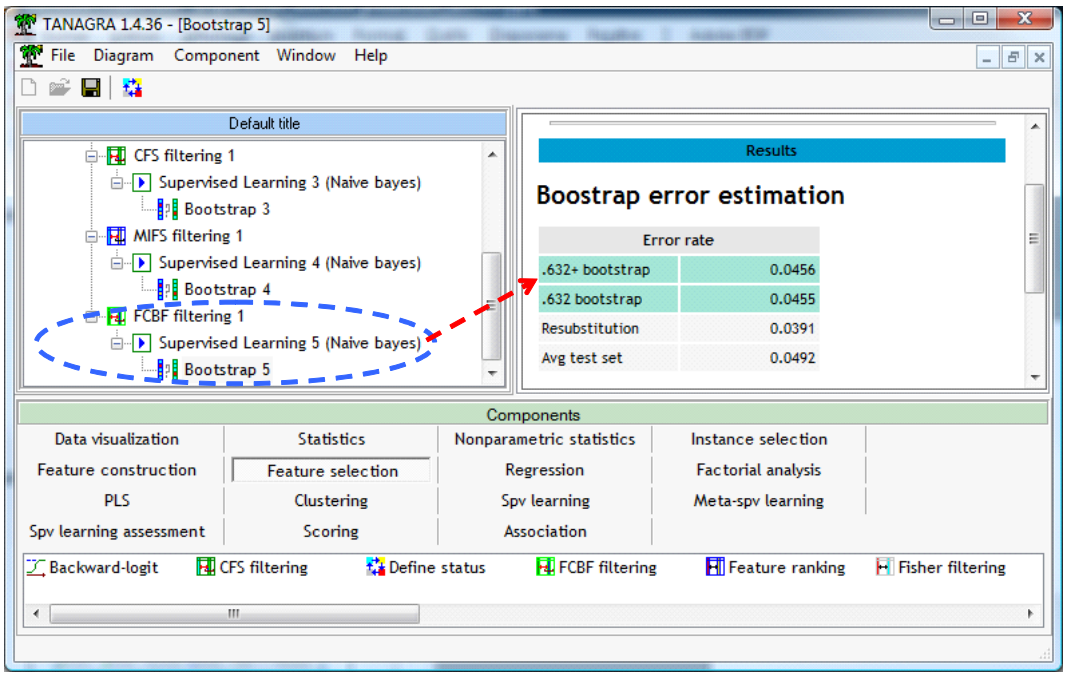
Le taux d'erreur du classifieur élaboré à partir des variables sélectionnées est de 5.75%.



4.2.4 Sélection avec FCBF

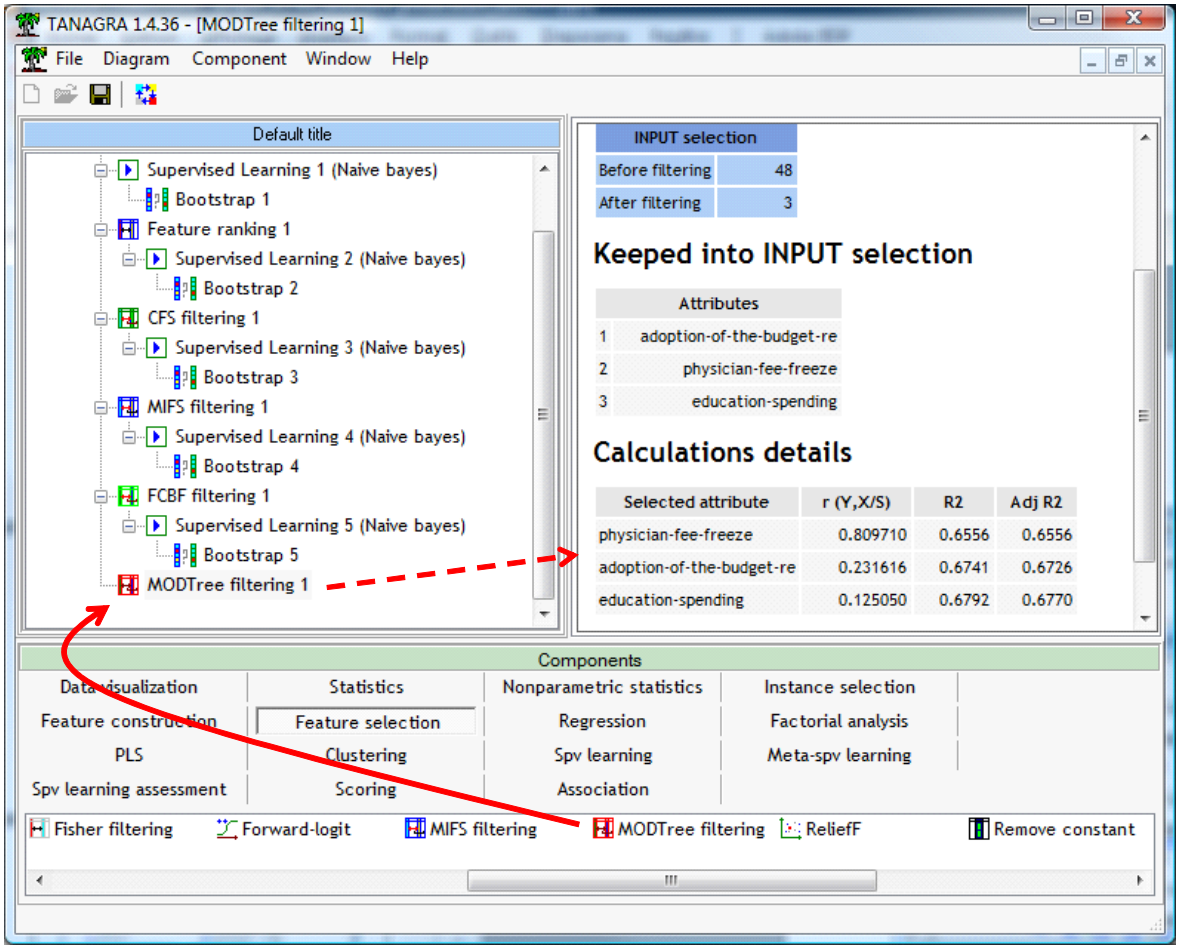


Nous utilisons $\delta = 0$ (valeur par défaut) pour la méthode FCBF. Quatre (4) variables sont sélectionnées. Nous remarquons que la variable bruitée a été introduite en dernière position. Contrairement à MIFS, nous savons à quelle entité nous référer pour fixer la valeur du paramètre δ. Il s'agit de la valeur seuil à partir de laquelle nous considérons que l'indicateur « symmetrical uncertainty – ρ » est suffisamment élevé. Clairement, si nous fixons $\delta = 0.1$, la dernière variable n'est pas introduite dans le modèle. En classement, nous obtenons un taux d'erreur de 4.56%.



4.2.5 Sélection avec MODTREE

Enfin, la méthode MODTREE sélectionne 3 variables issues de la base originelle.



Nous avons successivement les corrélations partielles d'ordre 0, 1 et 2 :

$$r_{group,\,physician-fee-freeze} = 0.809710$$

$$r_{group,\,adoption-of-budget\,/\,physician-fee-freeze} = 0.231616$$

$$r_{group,\,education-spending\,/\,physician-fee-freeze,\,adoption-of-budget} = 0.125050$$

Nous retrouvons les valeurs de l'exemple numérique décrit dans la section 2.3.4.

Dans le tableau décrivant les résultats, nous avons le coefficient de corrélation partiel $r_{y,x^*/z_1\ldots z_m} = r_m$; le coefficient de détermination R_m^2 obtenue à l'aide de

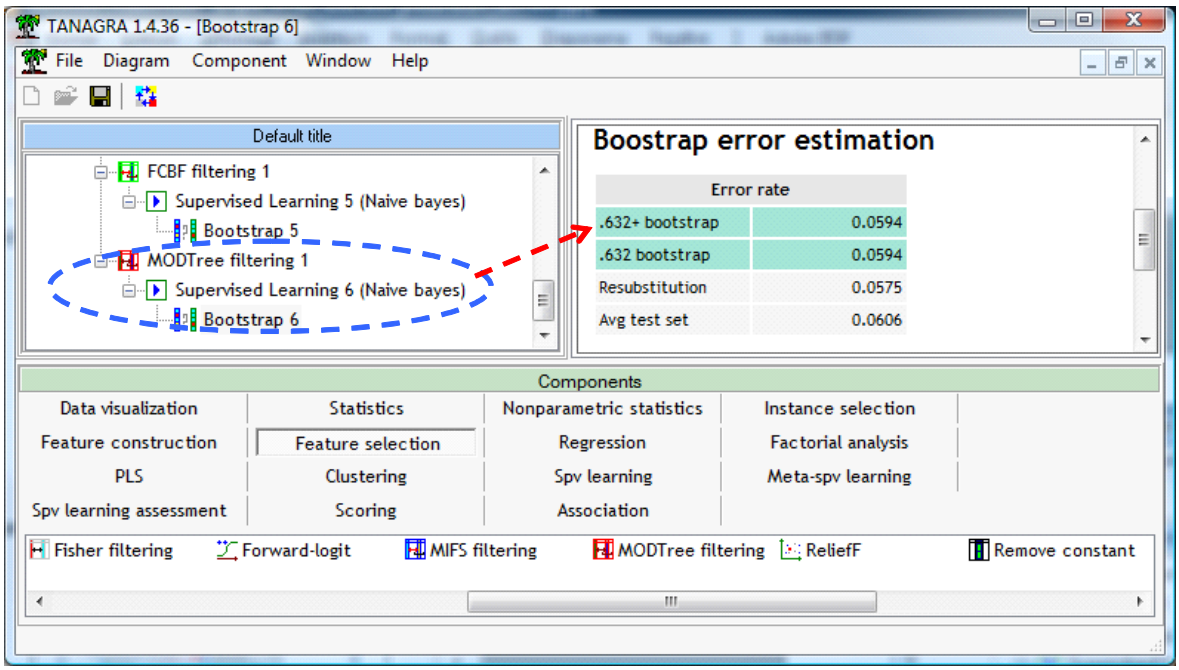
$$R_m^2 = 1 - \prod_{j=1}^{m}\left(1 - r_m^2\right)$$

Et le coefficient de détermination ajusté $\overline{R}_m^2$

$$\overline{R}_m^2 = 1 - \frac{n-1}{n-m-1}\left(1 - R_m^2\right)$$

On montre assez facilement que la sélection de MODTREE correspond à une maximisation du coefficient de détermination ajusté $\overline{R}_m^2$. Dans la procédure d'ajout pas à pas, le critère d'arrêt basé sur la corrélation partielle (section 2.3.4) revient à interrompre le processus dès que le critère $\overline{R}_m^2$ commence à diminuer.

Le taux d'erreur en bootstrap est de 5.94%.



4.3 Bilan

Nous récapitulons les résultats dans le tableau suivant.

Méthode	#Var. sélectionnées	#Var. « noise »	#Var. « corr »	Taux d'erreur en généralisation
Pas de sélection	48	16	16	9.94%
Ranking (α = 0.001)	28	0	14	9.87%
CFS	1	0	0	4.35%
MIFS (β = 1.5)	9	1	4	5.75%
FCBF (δ = 0)	4	1	0	4.56%
MODTREE	3	0	0	5.94%

Ne tombons pas dans le jeu puéril du « qui est le meilleur », surtout à partir d'une expérimentation sur un jeu de données. Essayons plutôt de dégager les principales tendances en exploitant au mieux les caractéristiques, créées pour la circonstance, de notre fichier :

- Le filtrage des variables peut améliorer les performances des classifieurs. Cela est d'autant plus vrai que la méthode est sensible aux descripteurs bruités. Des expérimentations rapportées dans la littérature montent que cela peut être également vrai même lorsque la méthode s'appuie sur un dispositif intégré de sélection (ex. les arbres de décision).
- Comme on pouvait s'y attendre, la méthode « ranking » gère la pertinence mais pas la redondance. Les descripteurs corrélés avec les prédicteurs originels, sans lien direct avec la cible, s'intercalent dans le classement. Ils viennent perturber l'interprétation des résultats si nous avions à le faire.
- Les méthodes CFS, FCBF et MODTREE ont un comportement similaire. Elles ont su détecter les bons prédicteurs, efficaces pour le classement. Dans le même temps, elles ont évacué à juste titre les variables bruitées (noise) et redondantes (corr).
- MIFS possède les mêmes qualités mais, comme cela est d'ailleurs souvent reporté dans la littérature, la gestion du paramètre β est délicate.

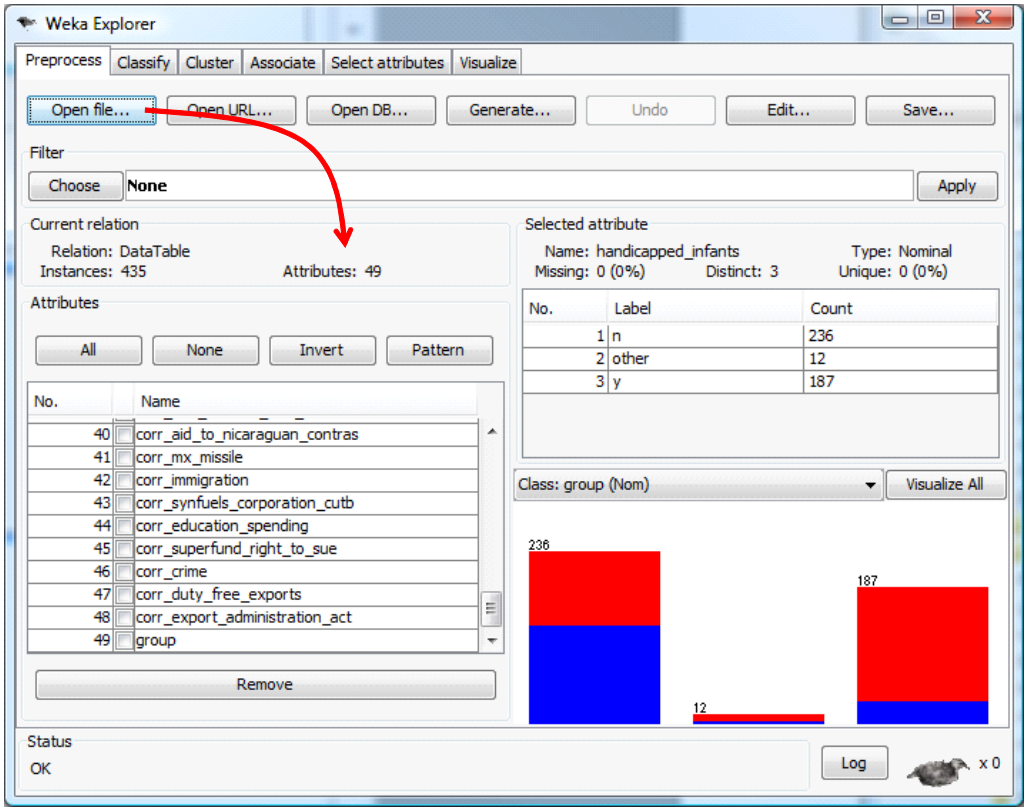
Nous en reparlerons dans la conclusion. Le même schéma expérimental a été mené sur d'autres fichiers avec des résultats très similaires.

5 Filtrage des prédicteurs discrets avec les autres logiciels

Dans cette section, nous décrivons les méthodes de filtrage implémentées dans quelques logiciels libres de Data Mining.

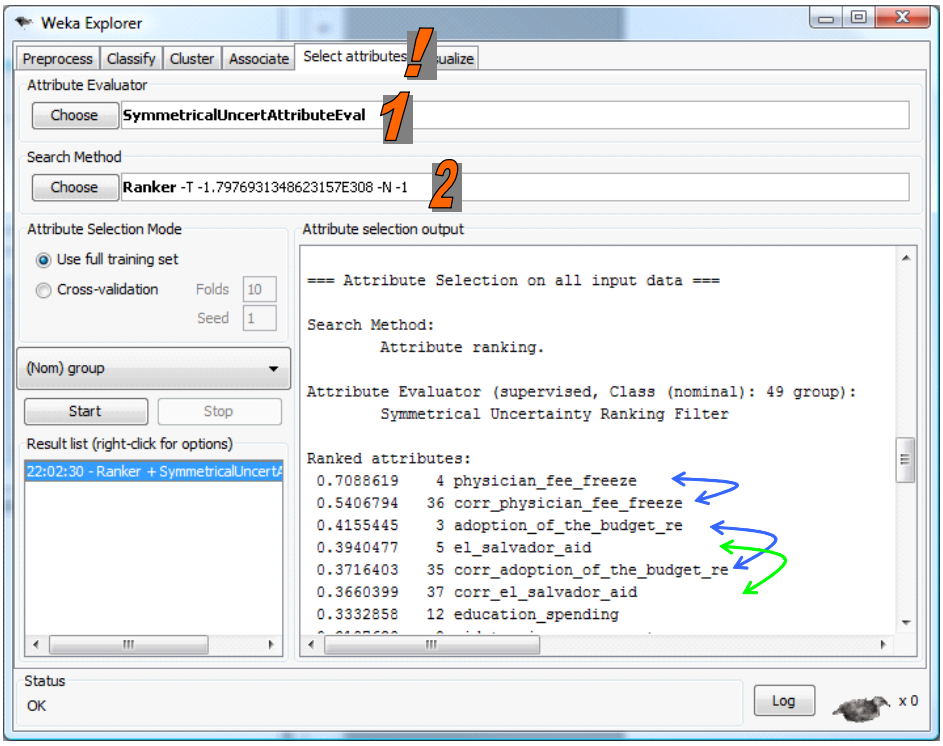
5.1 Filtrage dans Weka

Nous avons utilisé le mode EXPLORER. Après avoir charger les données (OPEN FILE… « vote_for_feature_selection.arff »),



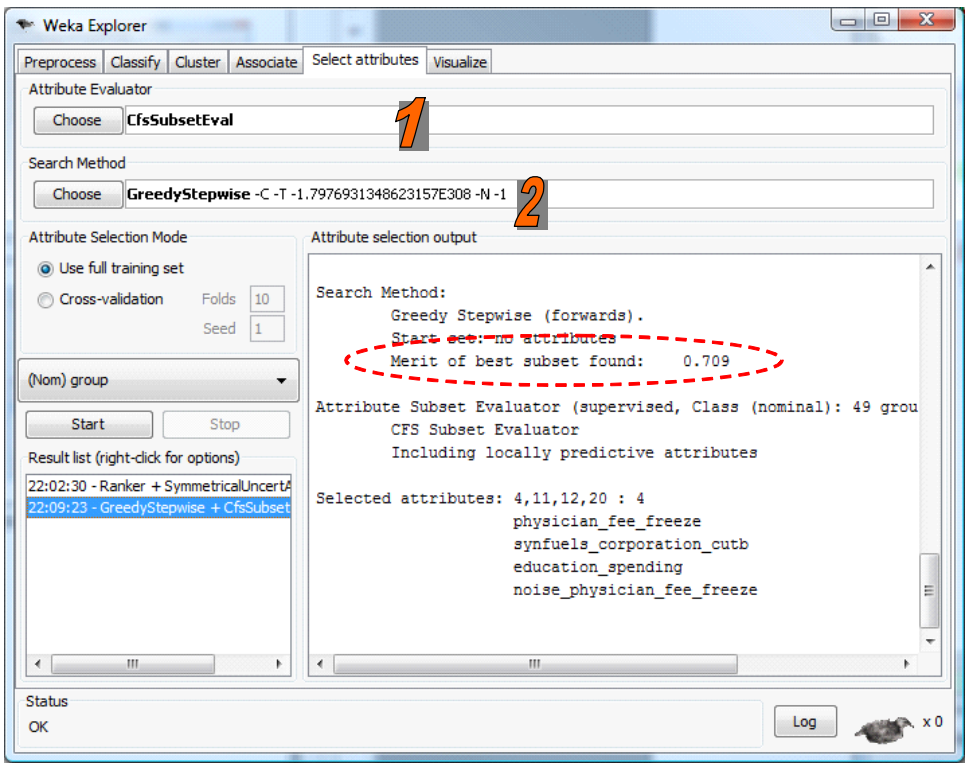
Nous actionnons l'onglet SELECT ATTRIBUTES. Les possibilités sont multiples : ATTRIBUTE EVALUATOR définit le mode d'évaluation des associations ; SEARCH METHOD définit la stratégie de recherche du sous-ensemble « optimal ».

Pour un simple « **ranking** », nous utilisons la combinaison suivante.



Il n'y a pas de règle d'arrêt. Les variables sont simplement ordonnées selon leur importance. Nous constatons que les variables « corr » s'intercalent dans la liste.

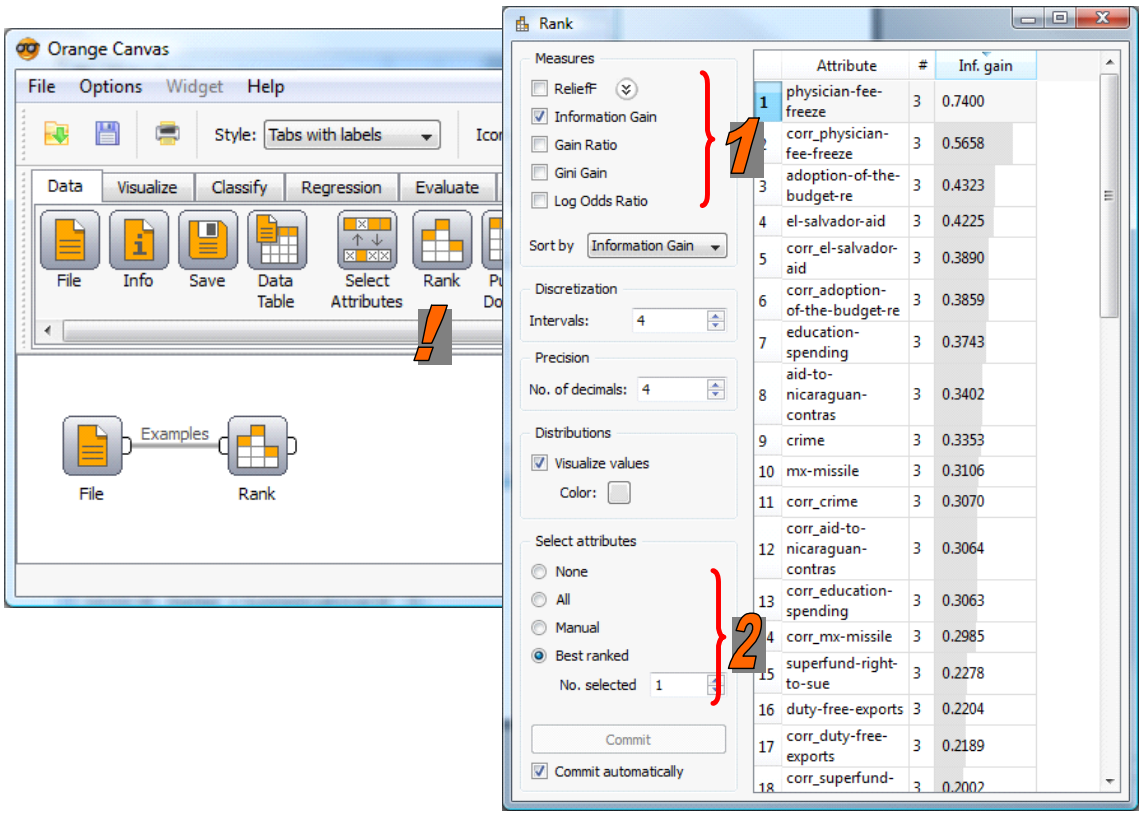
Pour la méthode CFS - forward, nous obtenons…



Curieusement, pour la même valeur de MERIT que Tanagra (merit = 0.709), Weka trouve 4 prédicteurs, les mêmes que ceux de FCBF d'ailleurs. Il faudrait se plonger dans le code source pour comprendre ces disparités.

5.2 Filtrage avec Orange

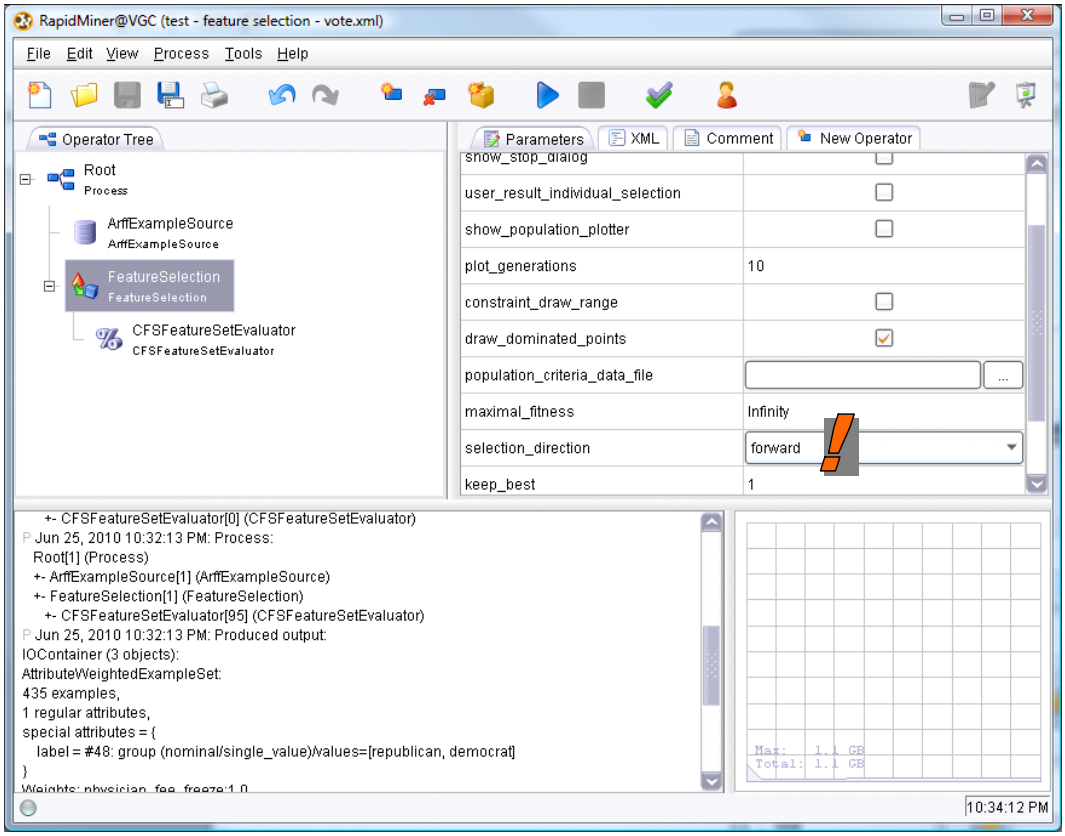
Nous utilisons le composant RANK dans Orange. Il trie simplement les prédicteurs selon leur degré de liaison avec la variable cible. Il ne tient absolument pas compte de la redondance.



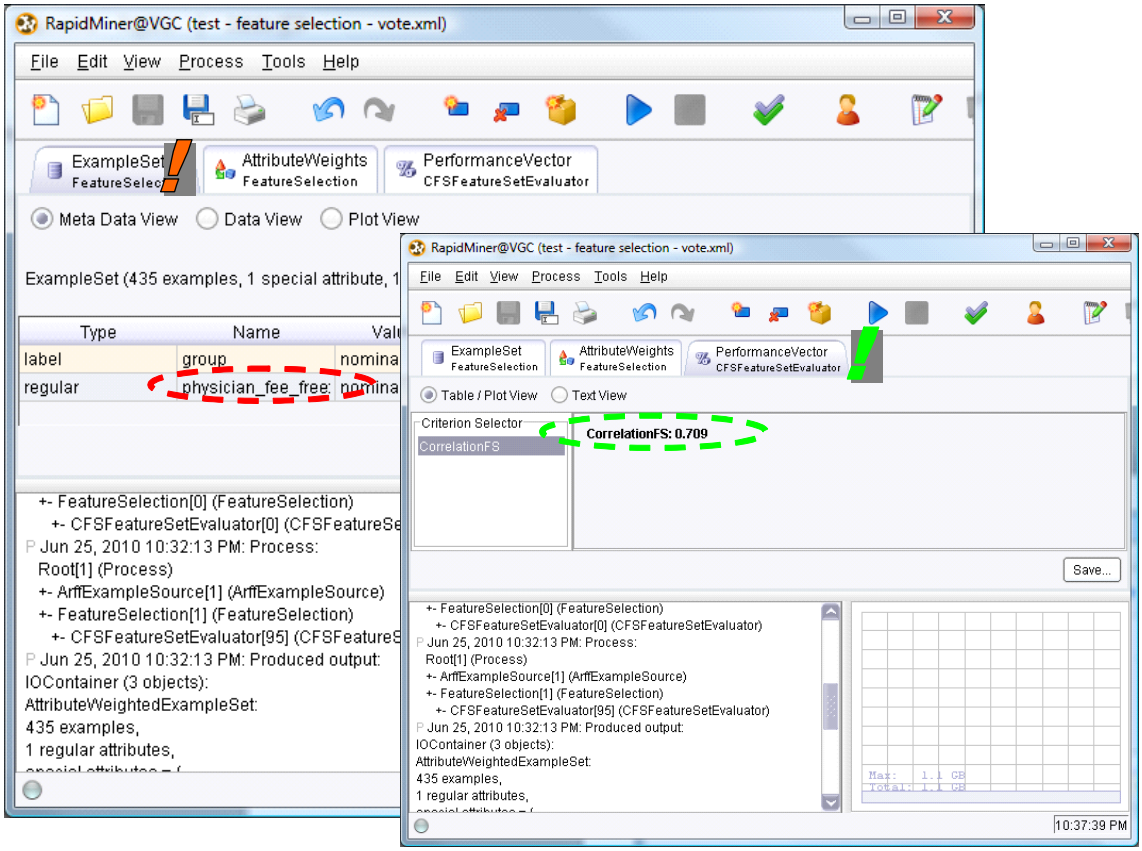
Plusieurs critères peuvent être utilisés pour le classement des variables (1). Le choix du nombre de variables à retenir est à la discrétion de l'utilisateur (2).

5.3 Filtrage avec RapidMiner

Nous utilisons le diagramme suivant pour implémenter la méthode CFS avec une stratégie de recherche *forward*.



RapidMiner fournit un résultat identique à celui de Tanagra. Seule la variable « physician-fee-freeze » est sélectionnée, avec un merit = 0.709.

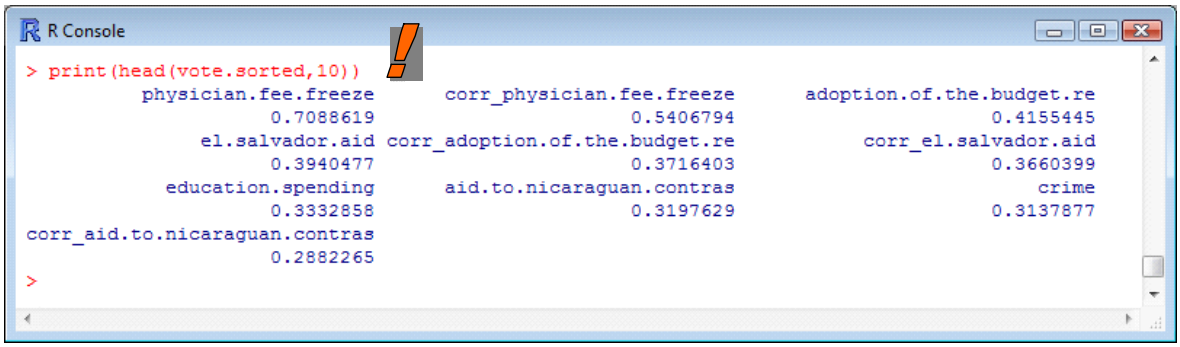


5.4 Filtrage avec R – package « FSelector »

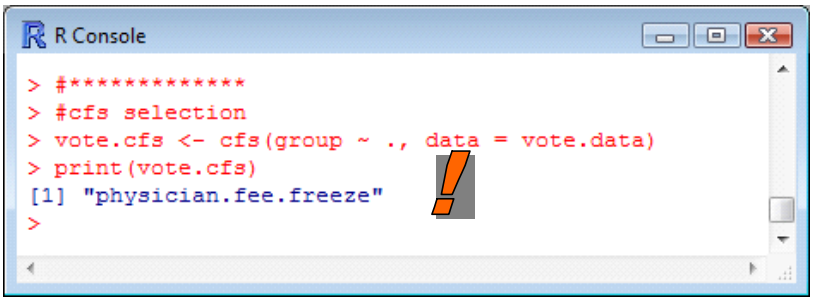
Nous utilisons le package « FSelector » avec R. Le code source utilisé est le suivant :

```
#clear the memory
rm (list=ls())

#load the dataset
vote.data <- read.table(file="vote_for_feature_selection.txt",header=T, sep="\t")

#loading the package
library(FSelector)

#*********************************
#ranking - Symmetrical uncertainty
#get the weight for each predictors
vote.ranking <- symmetrical.uncertainty(group ~ ., data = vote.data)
#sorting the result according the weight
index <- order(vote.ranking[[1]],decreasing=T)
vote.sorted <- vote.ranking[index,]
names(vote.sorted) <- rownames(vote.ranking)[index]
print(head(vote.sorted,10))

#*************
#cfs selection
vote.cfs <- cfs(group ~ ., data = vote.data)
print(vote.cfs)
```

La procédure **symmetrical.uncertainty()** calcule l'indicateur ρ pour chaque descripteur de la base. Nous trions alors les variables de manière décroissante. Nous affichons les 10 meilleures variables. L'ordonnancement et les valeurs de ρ sont cohérents avec ceux de Tanagra.



Pour la méthode **cfs()**, nous obtenons la seule variable « physician-fee-freeze » (comme Tanagra).



6 Conclusion

Nous avons décrit dans ce didacticiel quelques techniques de filtrage de prédicteurs discrets. Utiliser un fichier « pseudo artificiel »[11] nous a permis d'exacerber leur comportement. Les méthodes de ranking mettent en évidence les variables pertinentes, en relation avec la cible ; mais elles ne gèrent pas du tout la redondance. Chose à laquelle s'attellent les méthodes telles que CFS, MIFS, FCBF et MODTREE.

Le même schéma expérimental a été appliqué sur d'autres bases du serveur UCI : IRIS, OPTIDIGITS, WAVEFORM, KR-VS-KP, SPLICE (http://archive.ics.uci.edu/ml/ ; les éventuels prédicteurs continus ont été préalablement discrétisées avec la méthode MDLPC - http://tutoriels-data-mining.blogspot.com/2010/02/discretisation-comparaison-de-logiciels.html). Les résultats sont de la même nature.

[11] Un fichier « réel » ou tout du moins « réaliste » (de la base UCI en tous les cas, bien connu de notre communauté) auquel nous avons ajouté de nouvelles variables synthétiques avec les caractéristiques voulues.