Objectif

Mettre en oeuvre le composant STEPDISC pour sélectionner automatiquement les variables pertinentes en apprentissage supervisé.

La sélection de variables

Nécessité de la sélection de variable. La sélection de variables est un processus très important en apprentissage supervisé. Nous disposons d'une série de variables candidates, nous cherchons les variables les plus pertinentes pour expliquer et prédire les valeurs prises par la variable à prédire. Les objectifs sont bien souvent multiples : nous réduisons le nombre de variables à recueillir pour le déploiement du système ; nous améliorons notre connaissance du phénomène de causalité entre les descripteurs et la variable à prédire, ce qui est fondamental si nous voulons interpréter les résultats pour en assurer la reproductibilité ; enfin, mais pas toujours, nous améliorons la qualité de la prédiction, le ratio nombre d'observations et dimension de représentation étant plus favorable.

La meilleure approche pour sélectionner les variables pertinentes est certainement la sélection experte. Seule la connaissance du domaine permet de bien comprendre les causalités sous-jacentes, discerner les vrais liens des simples artefacts, mettre en évidence les interactions, etc. Malheureusement, elle n'est pas toujours possible, notamment parce que le nombre de variables candidates est élevé, une sélection manuelle devient vite inextricable. De toute manière, dans une démarche exploratoire, il paraîtrait bien étrange finalement que tout soit connu à l'avance, on se demande alors à quoi servirait la fouille de données dans ce contexte.

Il nous faut nous tourner vers les méthodes automatiques. Ne serait-ce que pour défricher le terrain et proposer aux experts les groupes de variables les plus intéressantes, quitte à procéder manuellement dans une deuxième phase.

La méthode STEPDISC. Dans ce tutoriel, nous présentons la méthode STEPDISC (Stepwise Discriminant Analysis). Elle repose le critère du LAMBDA de WILKS. Géométriquement, il s'agit de trouver le sous-espace de représentation qui permet un écartement maximal entre les centres de gravité des nuages de points conditionnels c.-à-d. les nuages de points associés à chaque valeur de la variable à prédire. Elle est donc particulièrement bien adaptée à l'analyse discriminante linéaire qui utilise également le même critère, d'où son appellation. Elles sont systématiquement associées dans les logiciels.

En réalité, rien ne nous empêche de l'utiliser comme préalable à telle ou telle méthode d'apprentissage. Il faut simplement garder à l'esprit qu'elle est adaptée pour les méthodes induisant un séparateur linéaire (régression logistique, SVM linéaire, etc.). En revanche, il est totalement illusoire d'en attendre un quelconque bénéfice si nous souhaitons l'utiliser dans le cadre de la méthode des plus proches voisins par exemple[1]. Les idées qui les sous-tendent – le biais de représentation – ne sont pas du tout en adéquation.

Le LAMBDA de WILKS est la transposition multidimensionnelle de la décomposition de la variance, il représente le rapport entre l'inertie intra classes et l'inertie totale. Si les nuages sont totalement confondus, LAMBDA = 1 ; plus LAMBDA se rapproche de 0, plus les nuages conditionnels sont

[1] Mis à part peut-être qu'en réduisant la dimensionalité, nous améliorons la stabilité des estimations locales des probabilités. Il s'agit plus dans ce cas d'une conséquence de la réduction du nombre de variables plutôt que la définition d'un sous-espace adapté à la méthode.

distincts. TANAGRA implémente deux stratégies. L'approche FORWARD consiste à partir de l'ensemble vide, choisir la variable induisant la meilleure amélioration du LAMBDA, et la sélectionner si amélioration est statistiquement significative ; nous procédons itérativement en ajoutant unes à unes les variables jusqu'à ce que l'adjonction d'une variable n'apporte plus d'amélioration. A l'inverse, l'approche BACKWARD, part de l'ensemble des variables candidates, recherche la variable dont le retrait entraînerait la dégradation la plus faible du LAMBDA, et la retire effectivement si cette dégradation n'est pas statistiquement significative.

Il existe une variante dite STEPWISE qui mixe ces deux approches : après avoir ajouté une variable, nous regardons s'il n'est pas nécessaire de retirer certaines variables parmi celles qui ont déjà été introduites. Puis nous regardons à nouveau s'il n'est pas possible d'en ajouter de nouvelles, etc. Cette option n'est pas implémentée dans TANAGRA.

Définition de la règle d'arrêt. Pour évaluer le rôle significatif d'une variable, nous utilisons la statistique F qui, a priori, suit une loi de Fisher. Il suffirait donc de comparer la p-value calculée pour la variable à évaluer et la comparer avec le niveau de signification qu'a choisi l'utilisateur.

Cette démarche n'est malheureusement pas exempte de reproches. En effet, la loi calculée dans le contexte d'un test d'hypothèse statistique est mise en oeuvre dans un contexte exploratoire. Nous avons d'abord choisi la variable maximisant (forward) ou minimisant (backward) F, puis nous avons testé cette variable. A l'étape suivante, nous procédons de même, sachant que les évaluations actuelles sont dépendantes des calculs à l'étape précédente, etc. L'interprétation de la p-value en termes de risque n'est donc plus possible. La loi n'est pas adaptée. Et les techniques de correction usuelles plus ou moins restrictives du calcul de la p-value (Bonferroni, Sidak, etc.) posent problème du fait que les calculs sont en cascade, les évaluations à chaque étape sont tributaires des sélections réalisées à l'étape précédente. Pour cette raison, certains logiciels -- STATISTICA par exemple -- se refusent à comparer la p-value calculée dans un contexte de tests d'hypothèses avec un niveau de signification défini par l'utilisateur. D'autres – SPSS – la proposent quand même mais en concurrence avec d'autres règles d'arrêt.

Il est vrai que la p-value a un autre défaut, elle a tendance à prendre des valeurs très basses lorsque l'échantillon est grand, laissant à croire que toutes les variables sont pertinentes si nous la comparons avec les seuils usuels utilisés en statistique (5%, 1%, etc.)

Si la p-value ne convient pas, quel critère utiliser ? La plupart des logiciels contournent le problème en proposant de comparer la statistique F calculée avec un seuil défini de manière ad hoc par l'utilisateur. Reste alors à définir la valeur seuil. Il n'y a pas vraiment de repères. SPSS propose par défaut la valeur 3.84 qui ressemble à s'y méprendre à la valeur critique d'un test à 5% lorsque nous travaillons sur un échantillon de quelques milliers d'individus. Ce qui nous ramène au problème précédent. STATISTICA propose des valeurs seuils par défaut sans que l'on sache vraiment ce qu'il faut en penser.

Enfin, un seuil est toujours arbitraire. Si nous fixons le seuil F à 3.84, une variable avec un F calculé de 3.839 ne passera pas dans la sélection FORWARD. Ce n'est pas très satisfaisant. Surtout si l'information nous a échappé.

Si du point de vue statistique, tout cela semble bien compliqué, du point de vue exploratoire, c'est moins gênant. Il faut surtout voir ces seuils comme autant d'outils mis à la disposition du praticien pour guider la recherche vers les solutions qui répondent à son cahier des charges. S'il désire sélectionner un ensemble de variables plus large pour pouvoir choisir lui-même dans ce sous-ensemble réduit, il choisira un seuil plus permissif. S'il veut en revanche réduire de manière drastique le nombre de variables parce qu'il en a un très grand nombre au départ, il fixera le seuil en conséquence.

Pour cette raison, nous proposons dans **TANAGRA** les deux approches ci-dessus. Nous permettons également à l'utilisateur d'imposer directement le nombre de variables à sélectionner. Et surtout, **nous fournissons le détail des calculs** : nous avons la possibilité d'inspecter les variables qui ont été en compétition, celles qui ont été éliminées parce que redondantes (dans un processus FORWARD, elles sont en bonne position dans les premières étapes, puis n'apparaissent plus) ; celles qui sont complémentaires (toujours dans un FORWARD, elles ne sont pas présentes dans les premières étapes, puis sont très significatives par la suite). L'analyse fine des résultats permet comprendre les relations entre les variables.

Fichier de données

Nous utilisons le fichier SONAR (SONAR_FOR_STEPDISC.XLS). L'objectif est de prédire l'appartenance d'un objet (Roc ou Mine) à partir des relevés fournis par un détecteur.

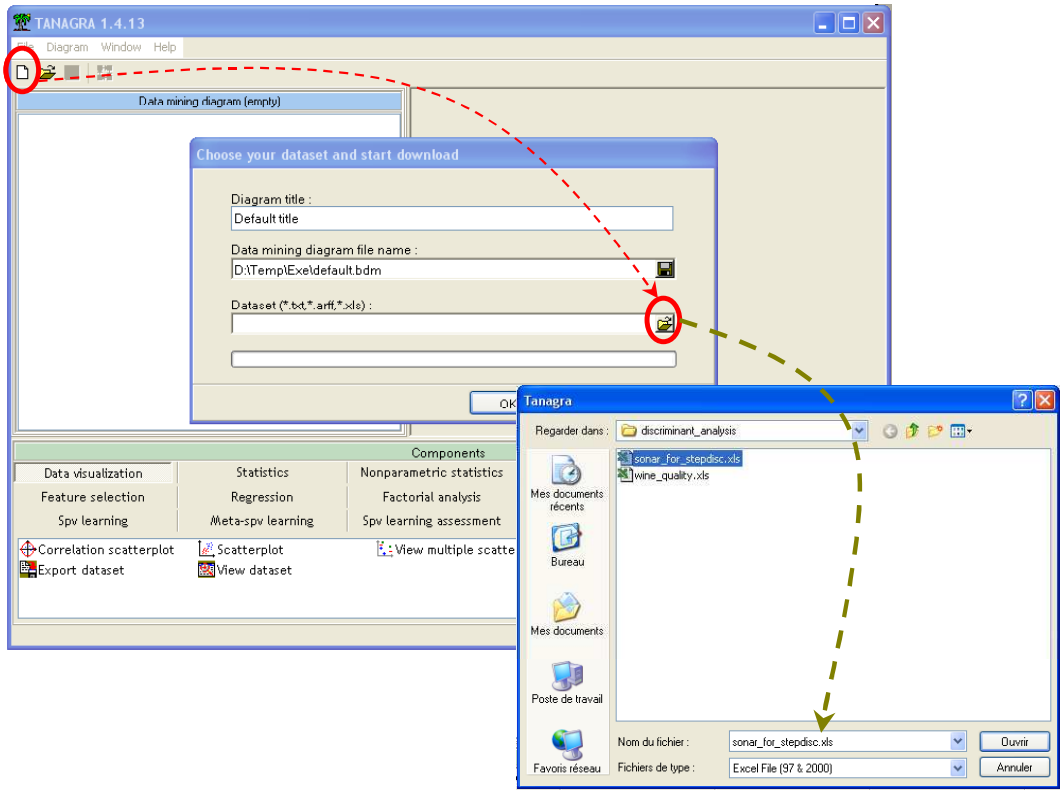
La sélection de variables est particulièrement recommandée car nous disposons de relativement peu d'observations (208 individus) pour 60 variables candidates. Il y a risque de sur-apprentissage.

Analyse discriminante linéaire dans TANAGRA

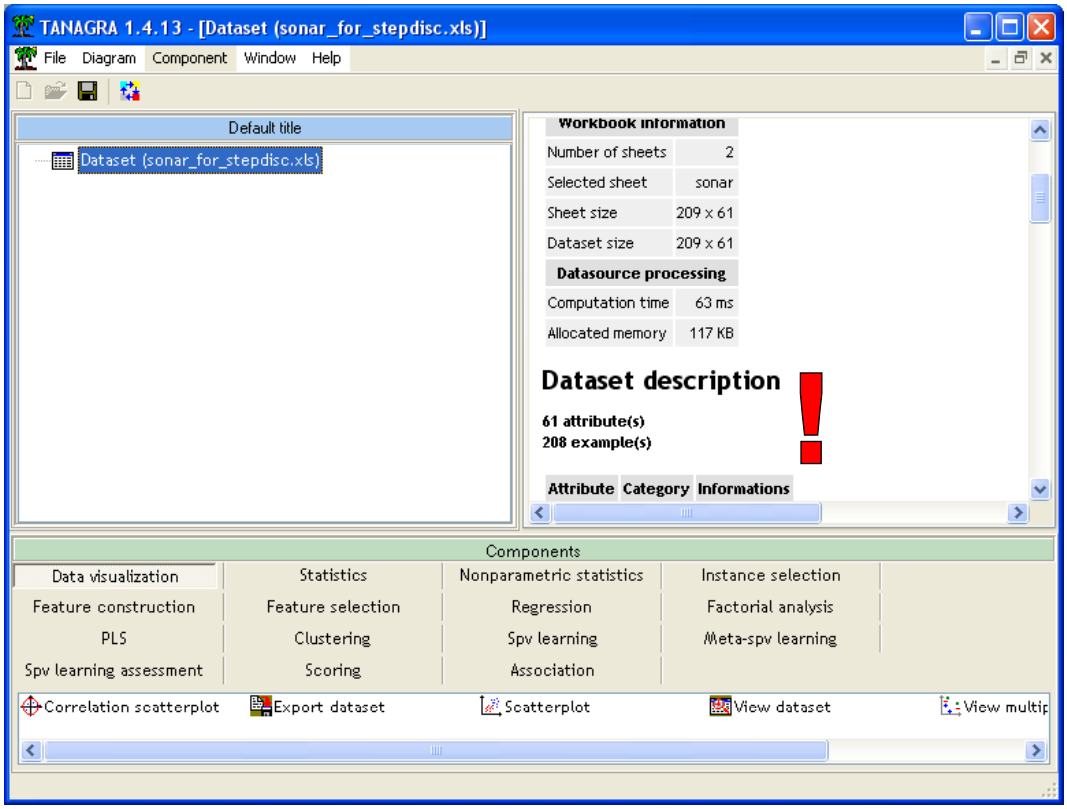
Dans un premier temps, nous mettons en œuvre une analyse discriminante linéaire, sans chercher à réduire le nombre de variables. Le résultat servira de référence, elle permettra d'évaluer l'efficacité de la méthode de sélection de variables STEPDISC.

Importation des données et création d'un diagramme

Nous créons un nouveau diagramme et nous importons les données. Nous activons pour cela le menu FILE/NEW. Notre fichier est au format EXCEL, les données doivent être situées sur la première feuille de calcul.

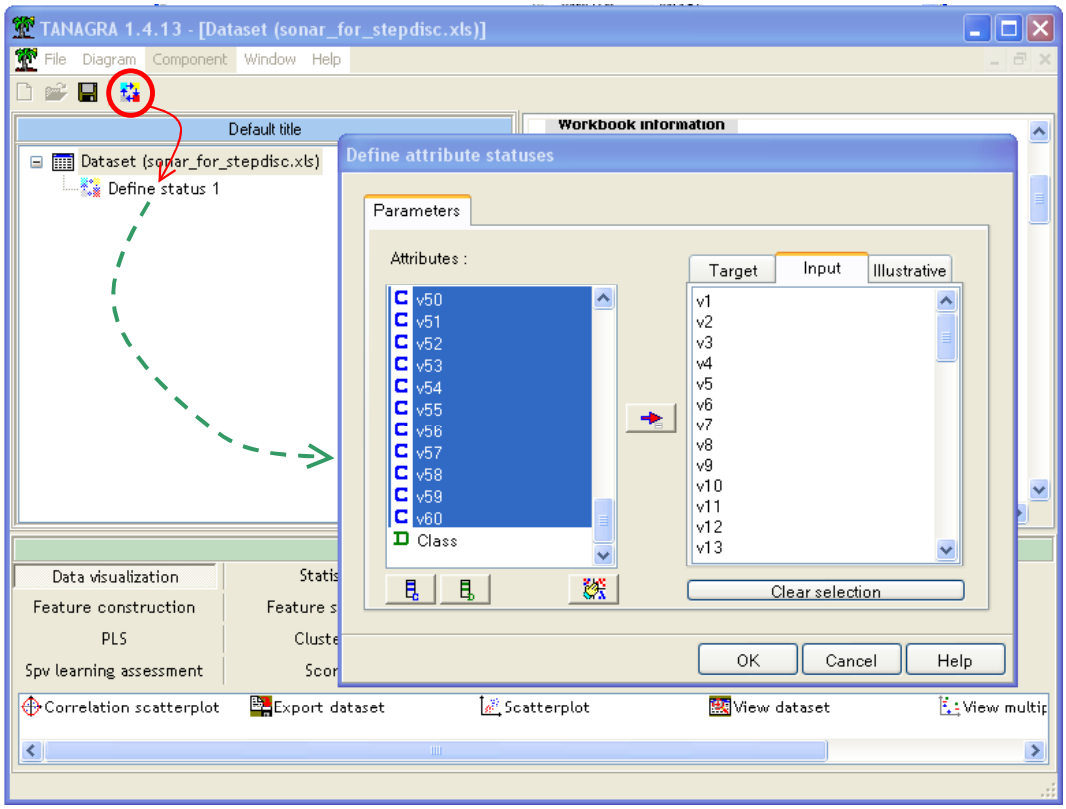


Vérifions que nous avons bien 61 variables et 208 observations dans notre fichier.

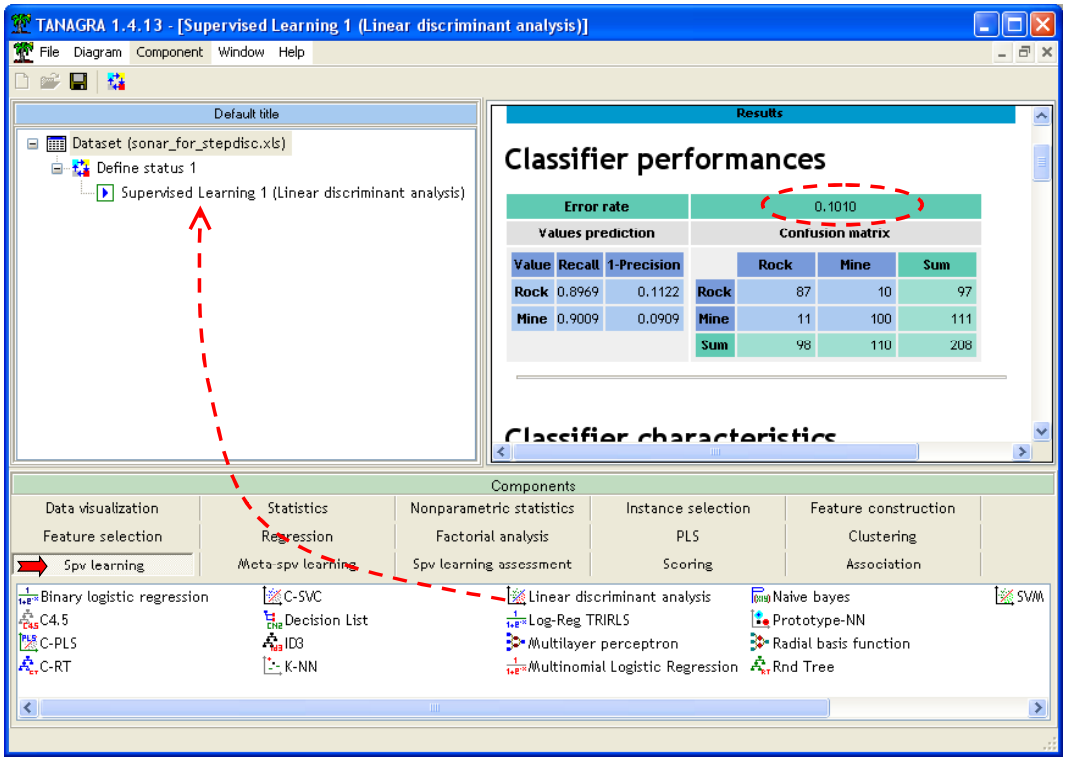


Analyse discriminante

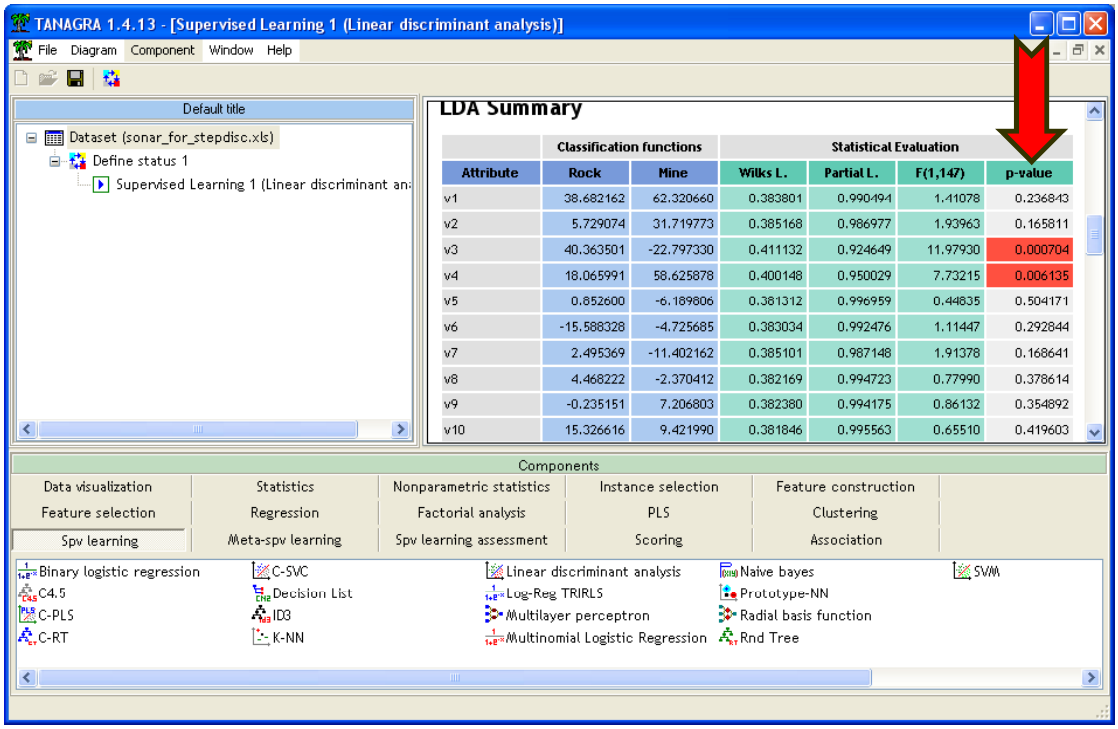
Nous plaçons le composant DEFINE STATUS dans le diagramme à partir du raccourci dans la barre d'outil. Nous plaçons en INPUT les variables V1…V60 ; CLASS est la variable à prédire (TARGET).



Puis nous insérons le composant Analyse Discriminante Linéaire.

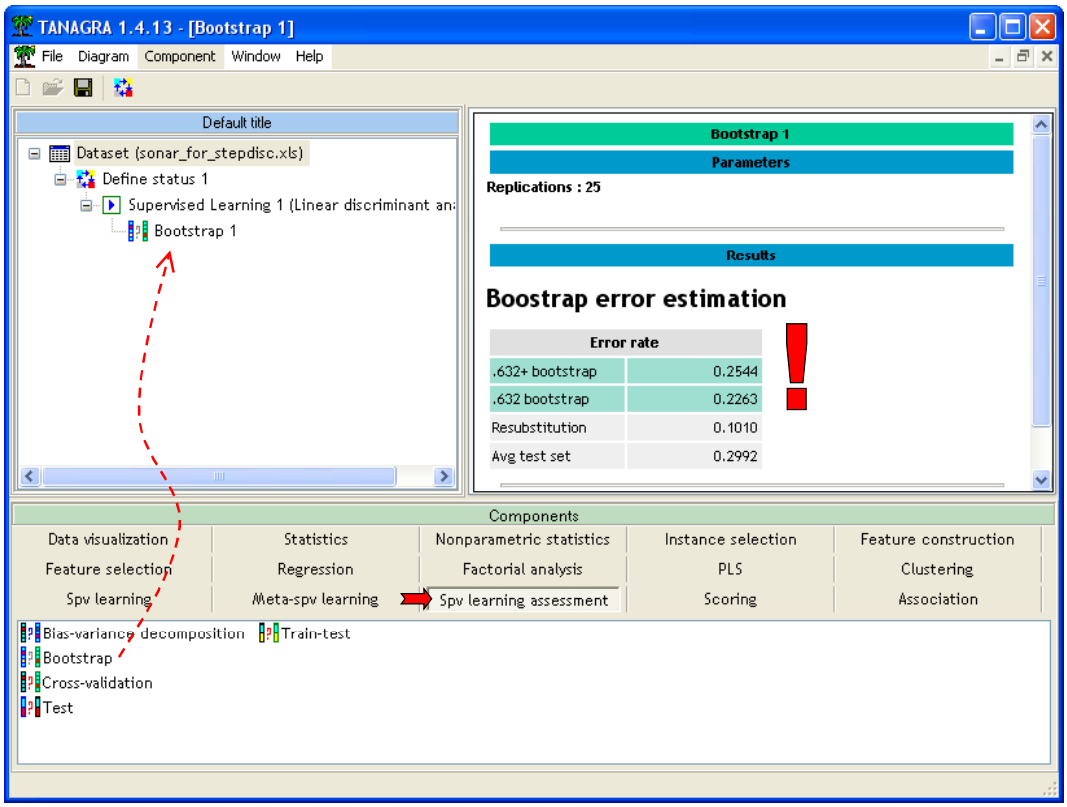


La matrice de confusion indique un taux de mauvais classement de **0.1010**. En consultant le détail des résultats, nous constatons qu'à un niveau de signification de 5%, plusieurs variables ne sont pas déterminantes dans la modélisation c.-à-d. leur retrait n'entraînerait pas une dégradation significative du Lambda de Wilks.



Le taux d'erreur calculé sur les données ayant servi à l'apprentissage – dit erreur en resubstitution – est très souvent trop optimiste. Optimisme d'autant plus marqué qu'il y a sur-apprentissage : le modèle « colle » trop aux données. Afin d'obtenir une évaluation plus réaliste des performances,

nous utilisons des méthodes de ré-échantillonnage. Nous plaçons à cet effet le composant BOOTSTRAP dans le diagramme.



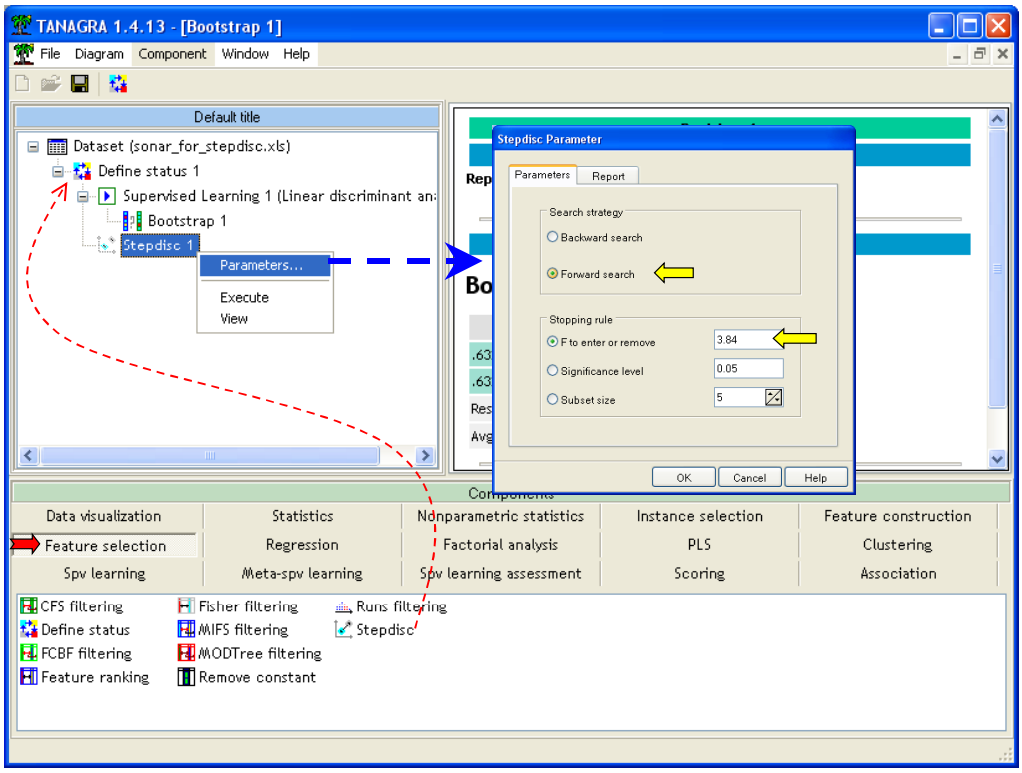
Les résultats montrent que lorsque nous déploierons le modèle de prédiction dans population, la probabilité de faire une mauvaise prédiction est de **0.2544**, et non de 0.1010 comme semblait l'indiquer l'évaluation en resubstitution.

Sélection de variables STEPDISC

Dans un deuxième temps, nous procédons à une sélection de variables à l'aide de la méthode STEPDISC. Nous évaluerons alors les conséquences de cette sélection sur les performances de l'analyse discriminante.

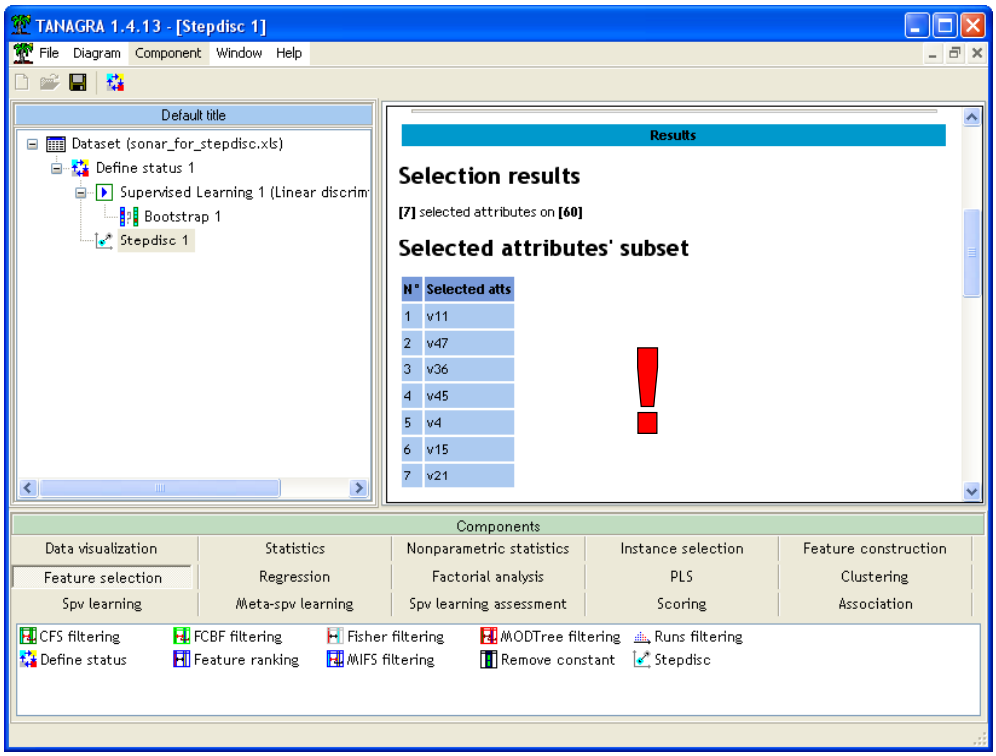
STEPDISC

Nous plaçons le composant STEPDISC à la suite du DEFINE STATUS 1. Puis nous le paramétrons en activant le menu contextuel PARAMETERS.



Nous voulons bien une recherche FORWARD, nous choisissons comme règle d'arrêt de comparer la valeur de F calculée avec le seuil 3.84[2].

L'exécution du composant (Menu contextuel VIEW) nous indique que **7** variables ont été sélectionnées par la procédure.



[2] Proposée par défaut dans SPSS.

Le détail des calculs est affiché dans un tableau. Pour ne pas alourdir la présentation, seules les 5 (paramétrable) meilleures solutions sont affichées à chaque étape.

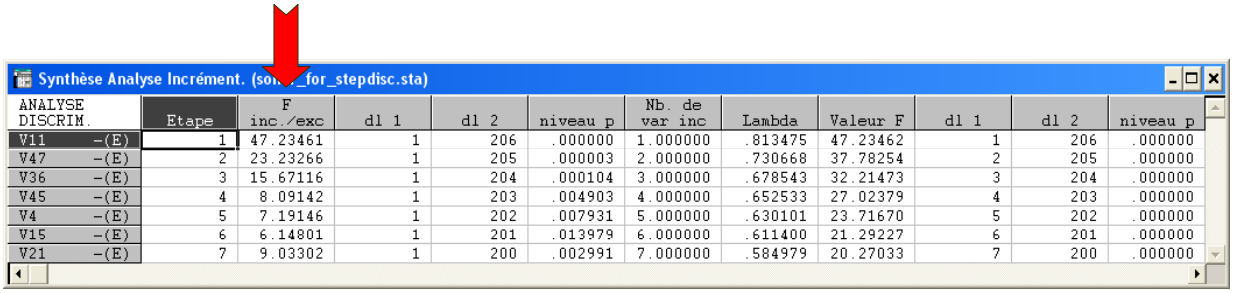
Detailed results

N°	d.f	Best	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	(1, 206)	**v11** L : 0.813 F : 47.23 p : 0.0000	**v11** L : 0.813 F : 47.23 p : 0.0000	**v12** L : 0.846 F : 37.36 p : 0.0000	**v49** L : 0.882 F : 27.57 p : 0.0000	**v45** L : 0.884 F : 27.11 p : 0.0000	**v10** L : 0.884 F : 26.94 p : 0.0000
2	(1, 205)	**v47** L : 0.731 F : 23.23 p : 0.0000	**v47** L : 0.731 F : 23.23 p : 0.0000	**v46** L : 0.738 F : 21.11 p : 0.0000	**v49** L : 0.743 F : 19.37 p : 0.0000	**v48** L : 0.749 F : 17.69 p : 0.0000	**v45** L : 0.750 F : 17.46 p : 0.0000
3	(1, 204)	**v36** L : 0.679 F : 15.67 p : 0.0001	**v36** L : 0.679 F : 15.67 p : 0.0001	**v37** L : 0.689 F : 12.45 p : 0.0005	**v21** L : 0.693 F : 11.01 p : 0.0011	**v35** L : 0.696 F : 10.09 p : 0.0017	**v22** L : 0.697 F : 9.78 p : 0.0020
4	(1, 203)	**v45** L : 0.653 F : 8.09 p : 0.0049	**v45** L : 0.653 F : 8.09 p : 0.0049	**v44** L : 0.655 F : 7.31 p : 0.0074	**v4** L : 0.657 F : 6.53 p : 0.0113	**v21** L : 0.658 F : 6.35 p : 0.0125	**v43** L : 0.660 F : 5.81 p : 0.0168
5	(1, 202)	**v4** L : 0.630 F : 7.19 p : 0.0079	**v4** L : 0.630 F : 7.19 p : 0.0079	**v21** L : 0.635 F : 5.58 p : 0.0191	**v31** L : 0.637 F : 4.91 p : 0.0279	**v54** L : 0.638 F : 4.55 p : 0.0342	**v23** L : 0.638 F : 4.46 p : 0.0359
6	(1, 201)	**v15** L : 0.611 F : 6.15 p : 0.0140	**v15** L : 0.611 F : 6.15 p : 0.0140	**v16** L : 0.612 F : 6.07 p : 0.0146	**v23** L : 0.616 F : 4.60 p : 0.0331	**v21** L : 0.616 F : 4.53 p : 0.0346	**v22** L : 0.617 F : 4.27 p : 0.0401
7	(1, 200)	**v21** L : 0.585 F : 9.03 p : 0.0030	**v21** L : 0.585 F : 9.03 p : 0.0030	**v20** L : 0.589 F : 7.57 p : 0.0065	**v22** L : 0.592 F : 6.58 p : 0.0111	**v31** L : 0.594 F : 5.86 p : 0.0164	**v23** L : 0.597 F : 4.85 p : 0.0288
8	(1, 199)	-	**v52** L : 0.577 F : 2.89 p : 0.0908	**v54** L : 0.578 F : 2.30 p : 0.1312	**v49** L : 0.579 F : 2.15 p : 0.1440	**v43** L : 0.579 F : 2.03 p : 0.1556	**v31** L : 0.580 F : 1.65 p : 0.2008

Nous observons que la variable V11 a été introduite à la première étape, le F calculé est de 47.23. A la seconde étape, la variable V47 a été sélectionnée. Nous constatons le choix entre V47 et V46 s'est jouée à peu de choses (F = 23.23 contre F = 21.11). En travaillant sur un autre échantillon, il se peut très bien que V46 passe devant V47. La sélection incrémentielle continue ainsi jusqu'à la huitième étape où la meilleure variable V52 propose un F calculé (2.89) inférieur au seuil que l'on s'est choisi (3.84). Le processus est donc stoppé.

N'ayant aucune connaissance dans le domaine, nous n'allons pas nous risquer à interpréter les résultats. Tout juste ferons-nous remarquer que la variable V12 qui arrive en seconde position dans la première étape, n'apparaît plus dans les meilleures solutions par la suite. Ca laisse à penser que cette variable est fortement corrélée avec V11, le fait d'avoir introduit cette dernière dans la sélection exclut V12. Si les variables V11 et V12 revêtent des significations particulières pour l'expert, il pourra s'appuyer sur ces informations pour guider la recherche dans la bonne direction.

N.B. A titre de comparaison, nous constatons que STATISTICA fournit exactement la même séquence de résultats.

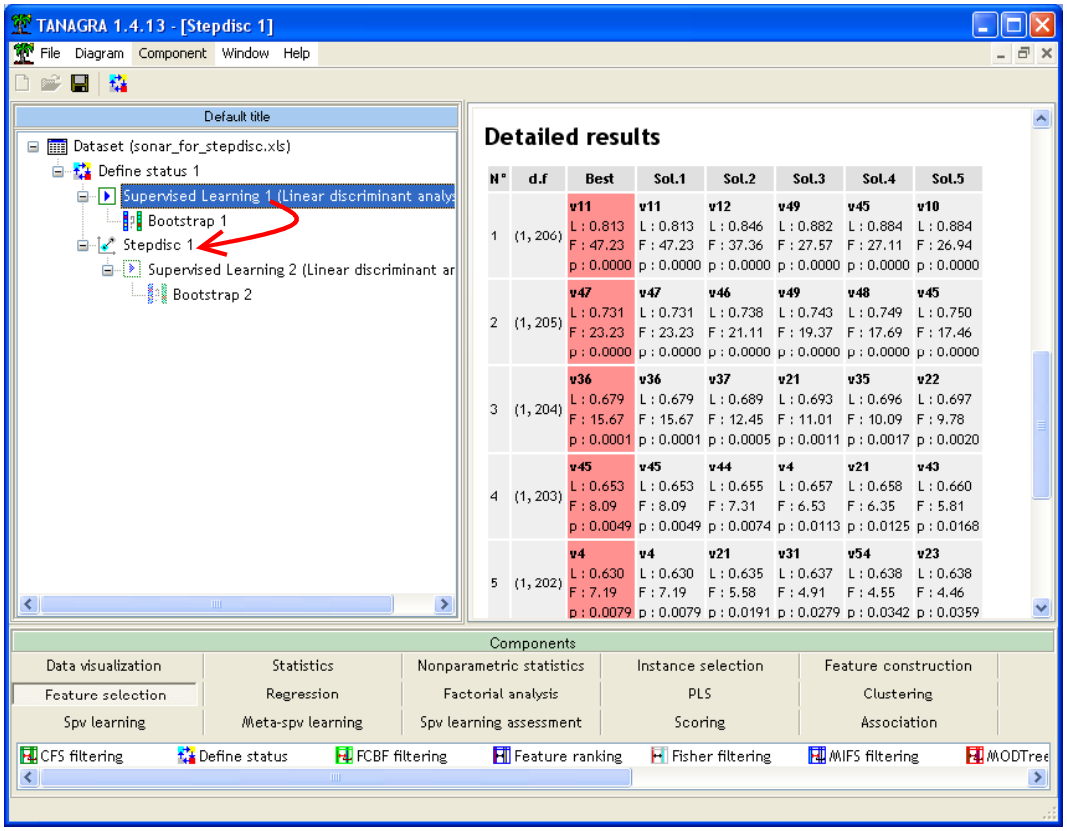


Synthèse Analyse Incrément. (so…_for_stepdisc.sta)

ANALYSE DISCRIM.	Etape	F inc./exc	dl 1	dl 2	niveau p	Nb. de var inc	Lambda	Valeur F	dl 1	dl 2	niveau p
V11 -(E)	1	47.23461	1	206	.000000	1.000000	.813475	47.23462	1	206	.000000
V47 -(E)	2	23.23266	1	205	.000003	2.000000	.730668	37.78254	2	205	.000000
V36 -(E)	3	15.67116	1	204	.000104	3.000000	.678543	32.21473	3	204	.000000
V45 -(E)	4	8.09142	1	203	.004903	4.000000	.652533	27.02379	4	203	.000000
V4 -(E)	5	7.19146	1	202	.007931	5.000000	.630101	23.71670	5	202	.000000
V15 -(E)	6	6.14801	1	201	.013979	6.000000	.611400	21.29227	6	201	.000000
V21 -(E)	7	9.03302	1	200	.002991	7.000000	.584979	20.27033	7	200	.000000

STEPDISC + Analyse discriminante + Evaluation bootstrap

Au-delà de l'analyse fine des résultats, un des objectifs de la sélection de variables est de produire un espace de représentation plus performant. Pour évaluer cette idée, nous plaçons à la suite du composant STEPDISC la séquence Analyse Discriminante – Bootstrap. Le plus simple est de procéder par glisser-déposer dans le diagramme lui-même.



Nous activons le menu VIEW de l'analyse discriminante. Nous constatons que l'erreur en resubstitution est maintenant de 0.1875, et que les 7 variables sont toutes significatives au niveau de signification de 5%.

Classifier performances

Error rate	0.1875

Values prediction

Value	Recall	1-Precision
Rock	0.8041	0.2041
Mine	0.8198	0.1727

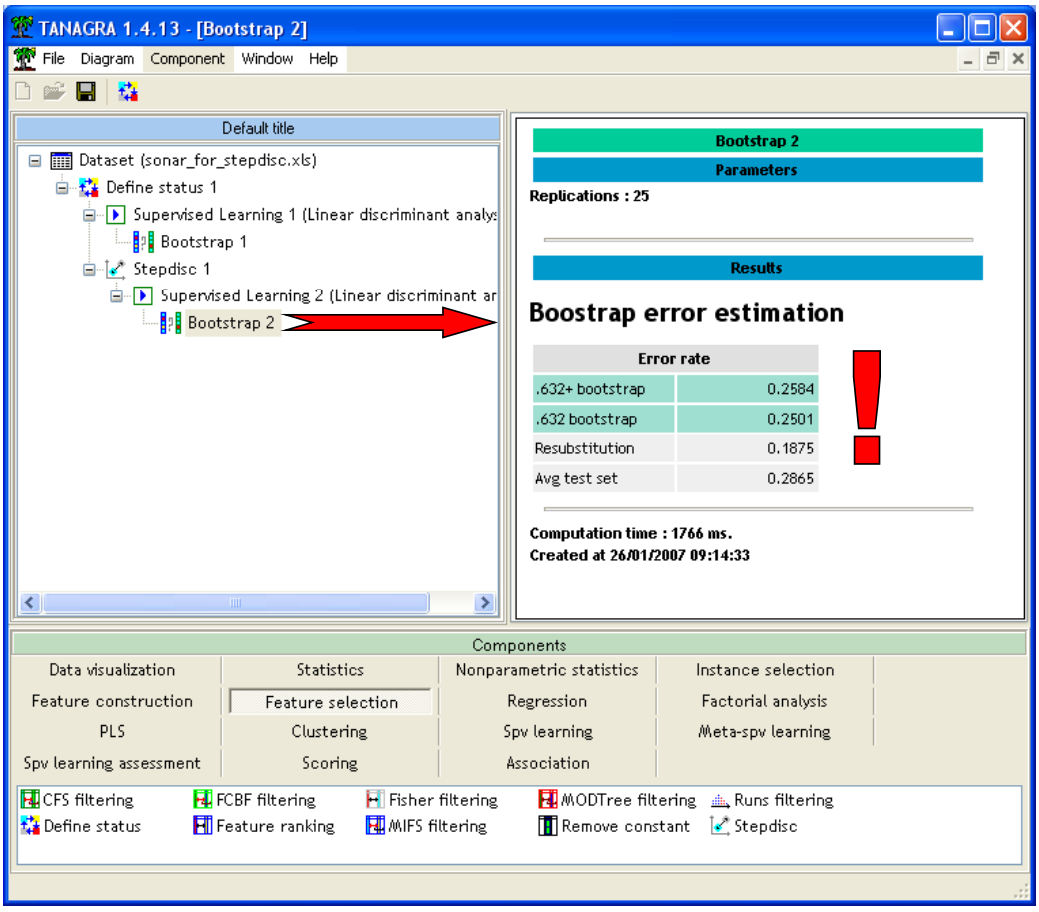
Confusion matrix

	Rock	Mine	Sum
Rock	78	19	97
Mine	20	91	111
Sum	98	110	208

LDA Summary

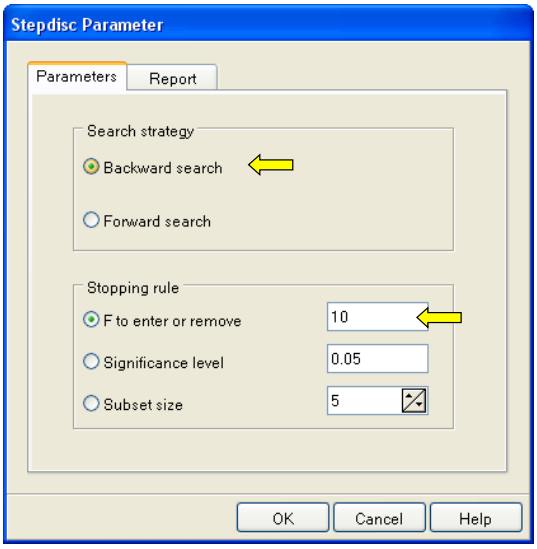
	Classification functions		Statistical Evaluation			
Attribute	Rock	Mine	Wilks L.	Partial L.	F(1,200)	p-value
v11	8.637494	16.695728	0.656707	0.890776	24.52325	0.000002
v47	29.753123	36.964342	0.600735	0.973772	5.38679	0.021298
v36	10.181178	6.494436	0.650366	0.899462	22.35513	0.000004
v45	-7.377507	-2.882663	0.601739	0.972147	5.73010	0.017601
v4	2.268385	15.468078	0.613508	0.953499	9.75378	0.002055
v15	3.120895	-0.313122	0.616228	0.949291	10.68357	0.001272
v21	11.106784	13.585777	0.611400	0.956787	9.03302	0.002991
constant	-8.278780	-11.421379	-			

Reste alors à évaluer les performances de tout ce processus à l'aide du composant bootstrap. Il nous indique qu'en déploiement, notre modèle a une probabilité d'erreur de **0.2584**, très proche finalement du premier modèle, sauf qu'ici nous n'utilisons plus que 7 variables.



Sélection BACKWARD

Voyons maintenant ce qu'il en est si l'on met en œuvre la sélection BACKWARD. Nous revenons sur le composant STEPDISC et nous activons le menu PARAMETERS. Nous paramétrons la méthode de la manière suivante[3].



Après calcul (menu VIEW), nous constatons que **7** variables ont été retenues pour la discrimination. Le détail des opérations nous indique que la variable V26 a été la première variable exclue, puis la variable V45, etc.

Selection results

[7] selected attributes on **[60]**

Selected attributes' subset

N°	Selected atts
1	v4
2	v12
3	v30
4	v31
5	v32
6	v36
7	v49

Detailed results

N°	d.f	Best	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	(1, 147)	**v26** L : 0.380 F : 0.00 p : 0.9929	**v26** L : 0.380 F : 0.00 p : 0.9929	**v45** L : 0.380 F : 0.00 p : 0.9846	**v33** L : 0.380 F : 0.01 p : 0.9210	**v46** L : 0.380 F : 0.01 p : 0.9048	**v38** L : 0.380 F : 0.03 p : 0.8683
2	(1, 148)	**v45** L : 0.380 F : 0.00 p : 0.9854	**v45** L : 0.380 F : 0.00 p : 0.9854	**v33** L : 0.380 F : 0.01 p : 0.9208	**v46** L : 0.380 F : 0.02 p : 0.9021	**v38** L : 0.380 F : 0.03 p : 0.8680	**v58** L : 0.380 F : 0.04 p : 0.8407
3	(1, 149)	**v33** L : 0.380 F : 0.01 p : 0.9212	**v33** L : 0.380 F : 0.01 p : 0.9212	**v46** L : 0.380 F : 0.02 p : 0.8784	**v38** L : 0.380 F : 0.03 p : 0.8682	**v37** L : 0.380 F : 0.04 p : 0.8403	**v58** L : 0.380 F : 0.04 p : 0.8394
4	(1, 150)	**v46** L : 0.380 F : 0.03 p : 0.8713	**v46** L : 0.380 F : 0.03 p : 0.8713	**v38** L : 0.380 F : 0.03 p : 0.8687	**v58** L : 0.380 F : 0.04 p : 0.8438	**v37** L : 0.380 F : 0.04 p : 0.8431	**v43** L : 0.380 F : 0.06 p : 0.8118

[3] La valeur seuil 10 est fixée en référence à la valeur par défaut proposée par STATISTICA.

Comparaison FORWARD - BACKWARD

En comparant les résultats fournis par les deux approches, nous observons que le sous-ensemble final diffère assez sensiblement. Nous recensons les variables dans le tableau suivant.

N°	Forward	Backward
1	v4	v4
2	v11	v12
3	v15	v30
4	v21	v31
5	v36	v32
6	v45	v36
7	v47	v49

Seules les variables V4 et V36 ont été sélectionnées dans les deux cas.

Bien souvent ce type de résultats laisse l'utilisateur perplexe. Mais quel est le bon sous-ensemble de variables alors ?

Il n'y a pas de réponse définitive en réalité. Il faut garder à l'esprit qu'il s'agit là de techniques numériques, avec ses avantages (rapidité entre autres) et ses inconvénients (calcul + tri : une différence de quelques millièmes est une différence entérinée). Il faut s'en servir avant tout pour mieux défricher les données, cerner les phénomènes sous-jacents. Ici commence le vrai rôle du data miner.