1 Objectif

Montrer le processus de post-élagage lors de l'induction d'un arbre de décision à l'aide de la méthode CART (Breiman et al., 1984 ; méthode C-RT dans TANAGRA).

Déterminer la bonne taille de l'arbre est une opération cruciale dans la construction d'un arbre de décision à partir de données. Elle détermine en grande partie ses performances lors de son déploiement dans la population[1]. Il y a deux extrêmes à éviter : l'arbre sous dimensionné, trop réduit, captant mal les informations utiles dans le fichier d'apprentissage ; l'arbre sur dimensionné, de taille exagérée, captant les spécificités du fichier d'apprentissage, spécificités qui ne sont pas transposables dans la population. Dans les deux cas, nous avons un modèle de prédiction peu performant.

Cette est souvent représentée par un graphique mettant en relation le nombre de feuille dans l'arbre et le taux d'erreur : sur l'échantillon d'apprentissage, servant à construire l'arbre, en décroissance permanente ; sur la population, inconnue, mais on suspecte qu'après un optimal, l'erreur va se dégrader à mesure que le nombre de feuilles devient exagéré ; sur un échantillon test enfin, échantillon à part n'ayant pas servi à construire le modèle, il sert à rendre compte (estimer) du phénomène de sur apprentissage (Figure 1).

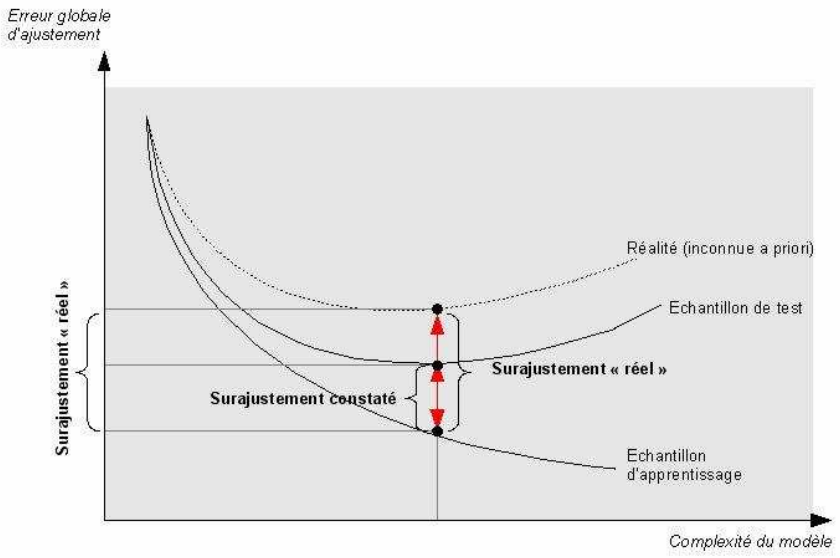


Figure 1 - Relation entre complexité du modèle et taux d'erreur (Source : **http://fr.wikipedia.org/wiki/Arbre_de_décision**)

Déterminer la bonne taille de l'arbre consiste donc à sélectionner, parmi les innombrables solutions que peuvent proposer l'induction, l'arbre le plus performant de la plus petite taille possible.

La performance est un critère clé puisqu'il s'agit avant tout d'un problème de prédiction, nous voulons produire un modèle efficace. La taille de l'arbre est un aspect tout aussi important car l'interprétation des résultats est bien souvent à l'origine de l'utilisation d'un arbre de décision dans un processus de fouille de données. Plus il sera réduit, plus facile sera la lecture du modèle et l'explicitation du phénomène de causalité qu'il traduit. De plus, un arbre de taille limitée utilise peu

[1] Rakotomalala R., « Arbres de décision », Revue MODULAD, n°33, pages 163 à 187, 2005 ; accessible en ligne http://www-rocq.inria.fr/axis/modulad/archives/numero-33/tutorial-rakotomalala-33/rakotomalala-33-tutorial.pdf

de variables, gage d'un déploiement facile lorsque nous voulons intégrer les règles de prédiction dans les systèmes d'information.

Dans leur ouvrage, Breiman et al. (1984)[2] sont certainement les premiers à avoir formuler clairement les enjeux de la détermination de la taille « optimale » d'un arbre de décision. Avec la méthode CART, ils ont popularisé une approche, la construction en deux temps d'un arbre, expansion – post-élagage, qui a été largement reprise par la suite par de nombreux auteurs, notamment par Quinlan avec la fameuse méthode C4.5 (1993).

Schématiquement, la technique consiste à construire l'arbre en deux temps. D'abord, un arbre comportant des feuilles pures au regard de la variable à prédire est construit, c'est la phase d'expansion (growing phase en anglais). L'arbre est généralement de très grande taille, avec très peu d'observations dans les feuilles. Puis, dans un deuxième temps, des séquences d'arbres de taille de plus en plus réduites sont évaluées, c'est la phase d'élagage (pruning phase en anglais). On sélectionne alors l'arbre réduit qui se comporte le mieux en termes d'erreur de prédiction.

Avec la méthode CART, dans sa version la plus simple, l'échantillon de données est simplement fractionné en deux portions pour assurer ces deux étapes. Un échantillon d'expansion (growing set) sert à construire l'arbre le plus grand possible. Cet arbre présente souvent une qualité de prédiction quasi-parfaite sur ces données. C'est tout à fait normal, les feuilles comportent très peu de contre-exemples. Puis, par le mécanisme du « coût complexité minimal » , de prime abord assez mystérieux, CART définit des séquences d'arbres de taille de plus en plus réduite. La principale idée à retenir dans ce dispositif est qu'il nous permet de lisser l'exploration des hypothèses, ce qui restreint d'autant les possibilités de trop coller aux données. Enfin, on injecte la seconde fraction des données (échantillon d'élagage, échantillon de validation, pruning set en anglais) pour sélectionner l'arbre le plus performant parmi les séquences d'arbres mis en avant à l'étape précédente.

Ce second échantillon n'ayant pas servi lors de la construction de l'arbre, il devrait mieux rendre compte des performances dans la population. Néanmoins, pour éviter de transposer le surajustement sur le fichier d'expansion en un surajustement sur l'échantillon d'élagage, CART intègre un mécanisme de préférence à la simplicité en choisissant, non pas l'arbre le plus précis sur l'échantillon d'élagage, mais le plus petit arbre dont l'erreur n'excède pas 1 écart-type de l'erreur optimale : c'est la fameuse règle de l'écart type (1-SE rule). De nombreuses études ont montré que cette stratégie permet de simplifier l'arbre tout en conservant ses performances en prédiction dans la population.

Dans ce didacticiel, nous mettons en œuvre la méthode CART. Nous essayons surtout de détailler les principaux repères à surveiller pour la guider, par un paramétrage adéquat, vers la solution souhaitable lors du processus de post-élagage.

2 Données

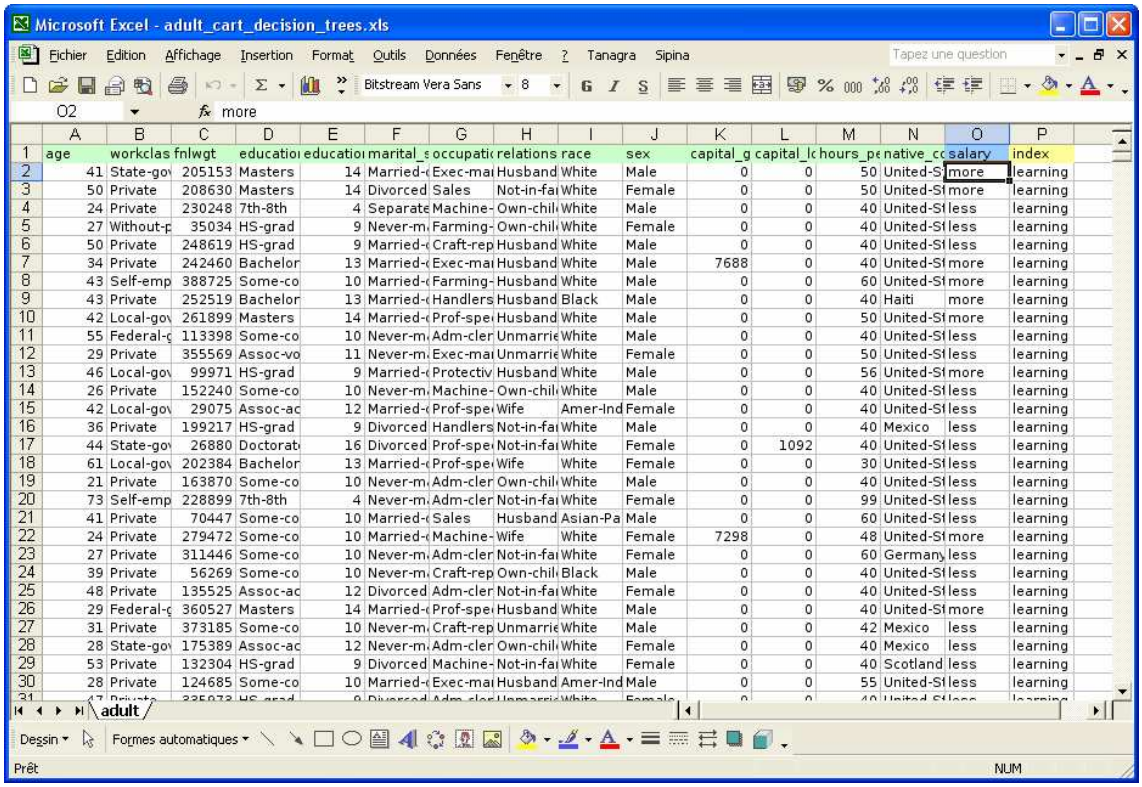
Nous utilisons le fichier ADULT_CART_DECISION_TREES.XLS[3] en provenance du serveur UCI[4]. Il comporte 48842 exemples et 14 variables. Nous voulons prédire la variable « CLASS », un individu a-t-il un revenu annuel supérieur ou inférieur à 50.000 $, à partir de ses caractéristiques signalétiques (niveau d'éducation, age, etc.). Nous avons subdivisé la base en 2 parties : un

[2] Breiman L., Friedman J., Olshen R., Stone C., « Classification and Regression Trees », Chapman & Hall, 1984.

[3] Accessible en ligne : http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/adult_cart_decision_trees.zip

[4] http://archive.ics.uci.edu/ml/datasets/Adult

échantillon d'apprentissage de 10000 observations réservé pour la construction du modèle ; un échantillon test de 38842 individus pour en évaluer les performances. Une variable indicatrice INDEX permet de spécifier l'appartenance d'un individu à l'un ou l'autre des sous échantillons.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	age	workclas	fnlwgt	educatio	educatio	marital_s	occupatic	relations	race	sex	capital_g	capital_l	hours_pe	native_cc	salary	index
2	41	State-gov	205153	Masters	14	Married-(	Exec-mai	Husband	White	Male	0	0	50	United-St	more	learning
3	50	Private	208630	Masters	14	Divorced	Sales	Not-in-fa	White	Female	0	0	50	United-St	more	learning
4	24	Private	230248	7th-8th	4	Separate	Machine-	Own-chil	White	Male	0	0	40	United-St	less	learning
5	27	Without-p	35034	HS-grad	9	Never-m	Farming-	Own-chil	White	Female	0	0	40	United-St	less	learning
6	50	Private	248619	HS-grad	9	Married-(	Craft-rep	Husband	White	Male	0	0	40	United-St	less	learning
7	34	Private	242460	Bachelor	13	Married-(	Exec-mai	Husband	White	Male	7688	0	40	United-St	more	learning
8	43	Self-emp	388725	Some-co	10	Married-(	Farming-	Husband	White	Male	0	0	60	United-St	more	learning
9	43	Private	252519	Bachelor	13	Married-(	Handlers	Husband	Black	Male	0	0	40	Haiti	more	learning
10	42	Local-gov	261899	Masters	14	Married-(	Prof-spe	Husband	White	Male	0	0	50	United-St	more	learning
11	55	Federal-g	113398	Some-co	10	Never-m	Adm-cler	Unmarrie	White	Male	0	0	40	United-St	less	learning
12	29	Private	355569	Assoc-vo	11	Never-m	Exec-mai	Unmarrie	White	Female	0	0	50	United-St	less	learning
13	46	Local-gov	99971	HS-grad	9	Married-(	Protectiv	Husband	White	Male	0	0	56	United-St	more	learning
14	26	Private	152240	Some-co	10	Never-m	Machine-	Own-chil	White	Male	0	0	40	United-St	less	learning
15	42	Local-gov	29075	Assoc-ac	12	Married-(	Prof-spe	Wife	Amer-Ind	Female	0	0	40	United-St	less	learning
16	36	Private	199217	HS-grad	9	Divorced	Handlers	Not-in-fa	White	Male	0	0	40	Mexico	less	learning
17	44	State-gov	26880	Doctorat	16	Divorced	Prof-spe	Not-in-fa	White	Female	0	1092	40	United-St	less	learning
18	61	Local-gov	202384	Bachelor	13	Married-(	Prof-spe	Wife	White	Female	0	0	30	United-St	less	learning
19	21	Private	163870	Some-co	10	Never-m	Adm-cler	Own-chil	White	Male	0	0	40	United-St	less	learning
20	73	Self-emp	228899	7th-8th	4	Never-m	Adm-cler	Not-in-fa	White	Female	0	0	99	United-St	less	learning
21	41	Private	70447	Some-co	10	Married-(	Sales	Husband	Asian-Pa	Male	0	0	60	United-St	less	learning
22	24	Private	279472	Some-co	10	Married-(	Machine-	Wife	White	Female	7298	0	48	United-St	more	learning
23	27	Private	311446	Some-co	10	Never-m	Adm-cler	Not-in-fa	White	Female	0	0	60	Germany	less	learning
24	39	Private	56269	Some-co	10	Never-m	Craft-rep	Own-chil	Black	Male	0	0	40	United-St	less	learning
25	48	Private	135525	Assoc-ac	12	Divorced	Adm-cler	Not-in-fa	White	Female	0	0	40	United-St	less	learning
26	29	Federal-g	360527	Masters	14	Married-(	Prof-spe	Husband	White	Male	0	0	40	United-St	more	learning
27	31	Private	373185	Some-co	10	Never-m	Craft-rep	Unmarrie	White	Male	0	0	42	Mexico	less	learning
28	28	State-gov	175389	Assoc-ac	12	Never-m	Adm-cler	Own-chil	White	Female	0	0	40	Mexico	less	learning
29	53	Private	132304	HS-grad	9	Divorced	Machine-	Not-in-fa	White	Female	0	0	40	Scotland	less	learning
30	28	Private	124685	Some-co	10	Married-(	Exec-mai	Husband	Amer-Ind	Male	0	0	55	United-St	less	learning

Notre objectif est de produire, en nous appuyant sur la méthodologie CART, un arbre de décision à la fois performant et peu complexe c.-à-d. comportant le moins de feuilles (règles) possible.

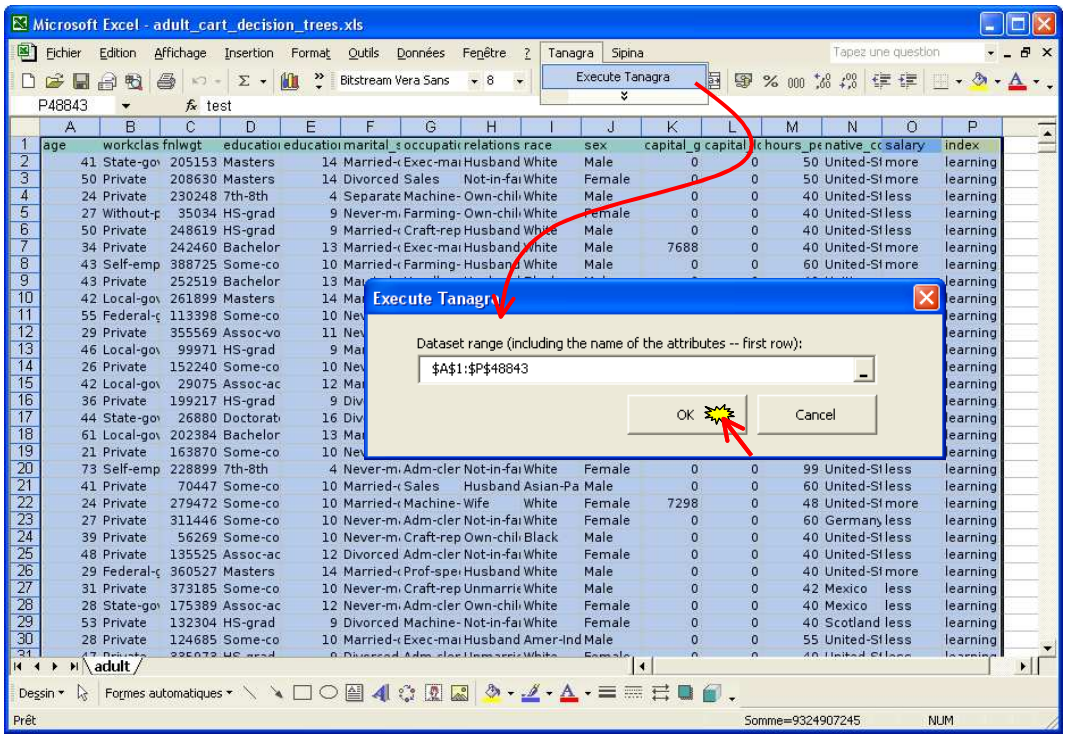
Bien entendu, pour que l'évaluation soit crédible, l'échantillon test ne doit être utilisé qu'en dernier ressort, pour comparer les performances des modèles alternatifs proposés. En aucune manière, il ne doit être mis à contribution pour guider l'exploration des solutions.

3 Arbre de décision – La méthode CART

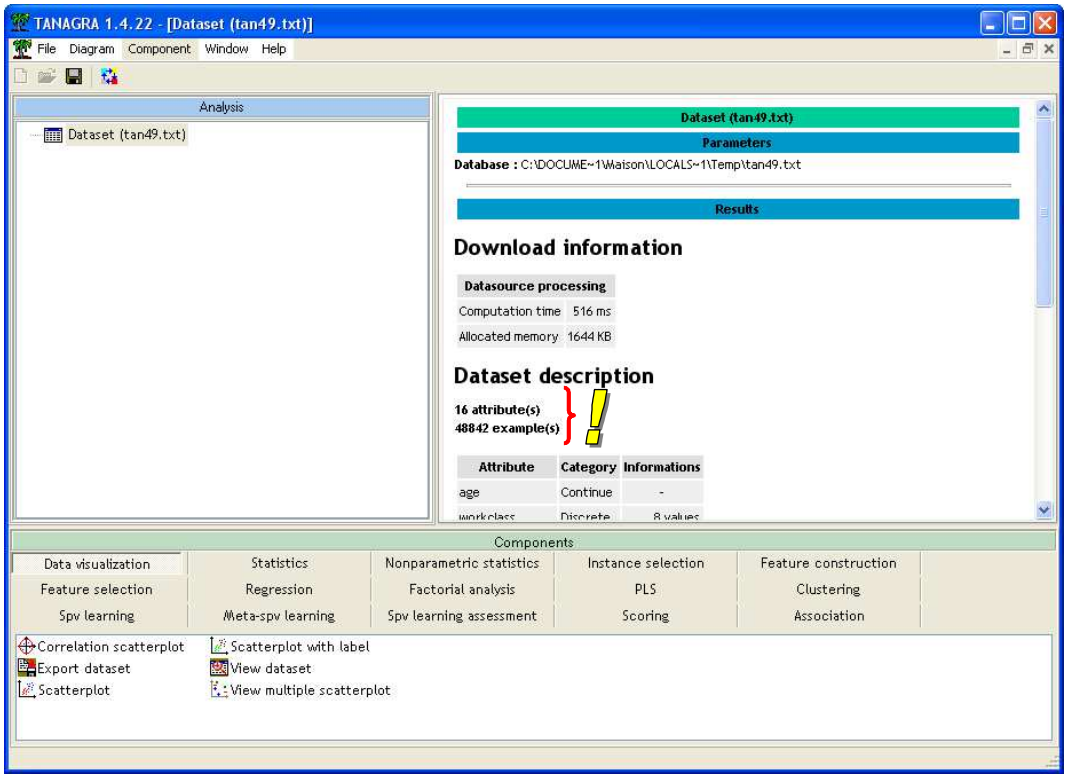
3.1 Créer un diagramme et importer les données dans TANAGRA

Le plus simple pour créer un diagramme et importer les données dans TANAGRA est d'ouvrir le fichier dans le tableur EXCEL. A l'aide de la macro complémentaire fournie avec le logiciel[5], il est possible d'envoyer les données directement du tableur vers TANAGRA. Pour ce faire, nous sélectionnons la plage de données et activons le menu TANAGRA/EXECUTE TANAGRA. Une boîte de dialogue apparaît, indiquant les coordonnées de la plage de cellules. Si la sélection est correcte, nous pouvons valider en cliquant sur OK.

[5] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Excel_AddIn.pdf

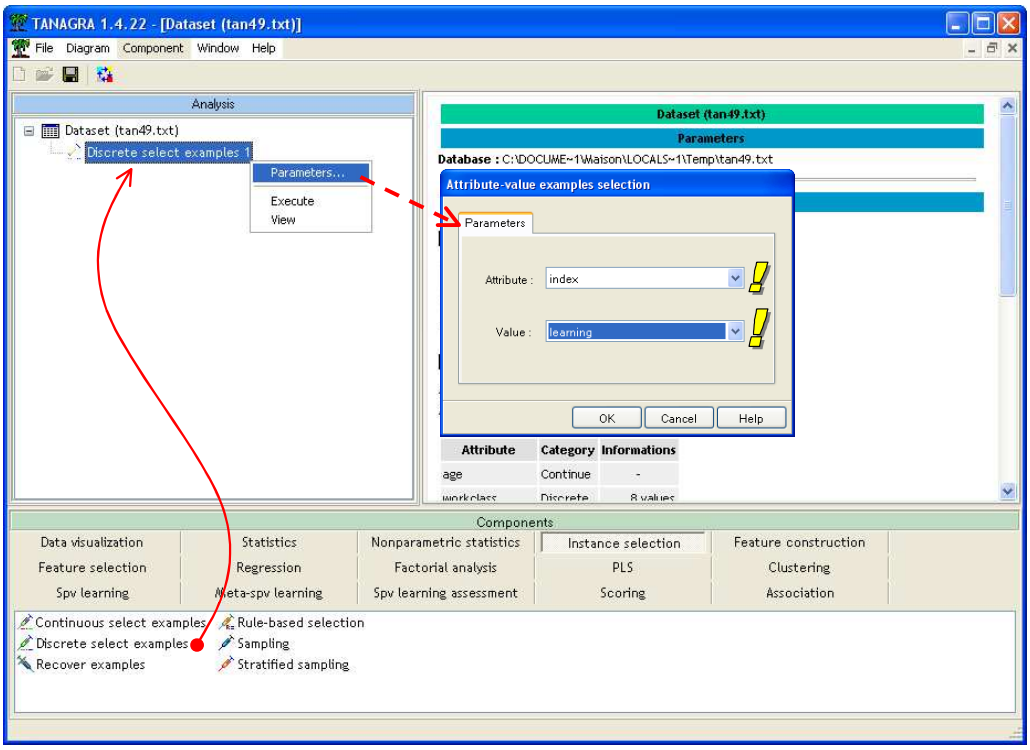


TANAGRA est automatiquement démarré, nous disposons bien de 48842 observations avec 15 variables (incluant l'index définissant les échantillons apprentissage et test).

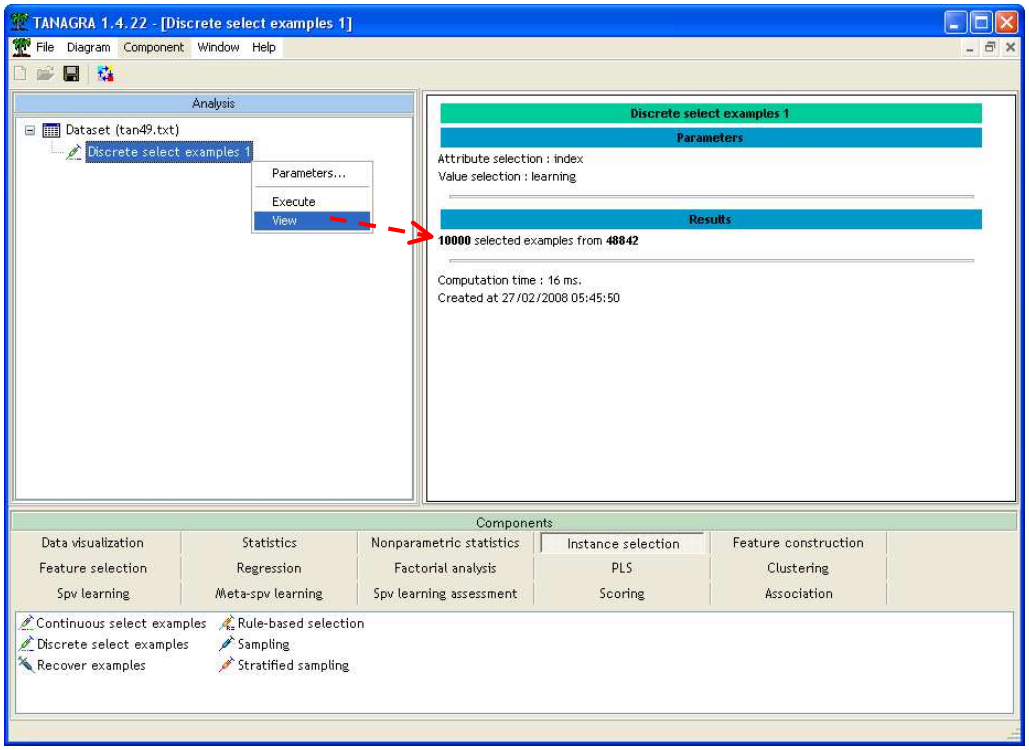


3.2 Subdiviser les données

Nous introduisons le composant DISCRETE SELECT EXAMPLES (onglet INSTANCE SELECTION) dans le diagramme pour désigner les observations dédiées à l'apprentissage. Nous le paramétrons (menu contextuel PARAMETERS) : INDEX joue le rôle de variable de contrôle, les individus actifs correspondent à la valeur LEARNING.

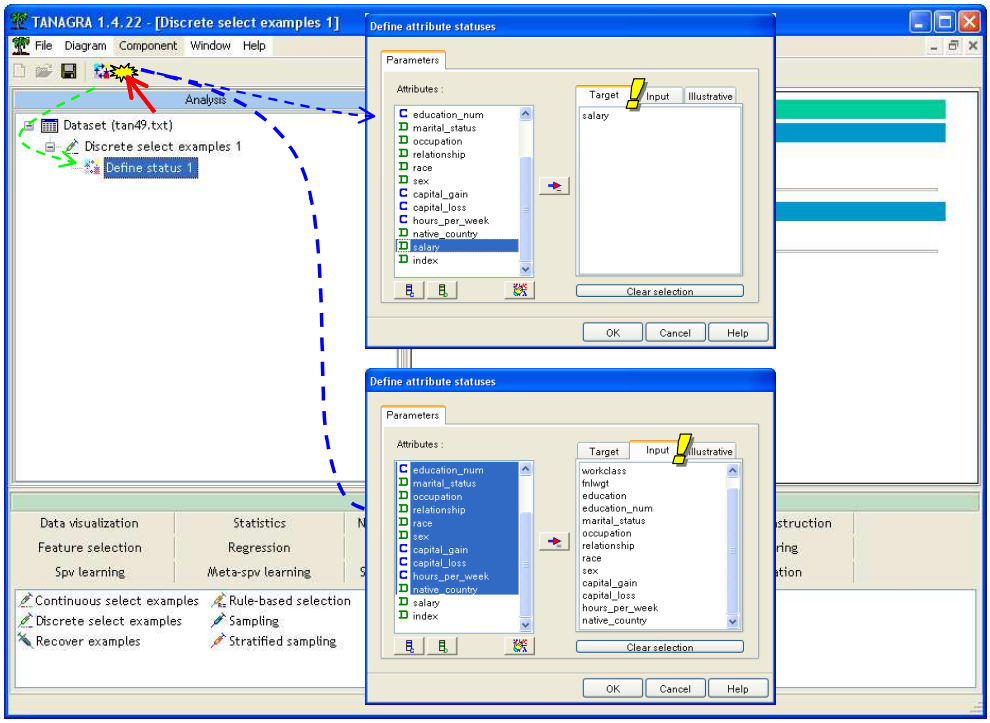


Après validation (bouton OK), le menu contextuel VIEW permet d'afficher le résultat de la sélection, 10000 individus sont sélectionnés pour l'induction de l'arbre.



3.3 Variable dépendante et variables prédictives

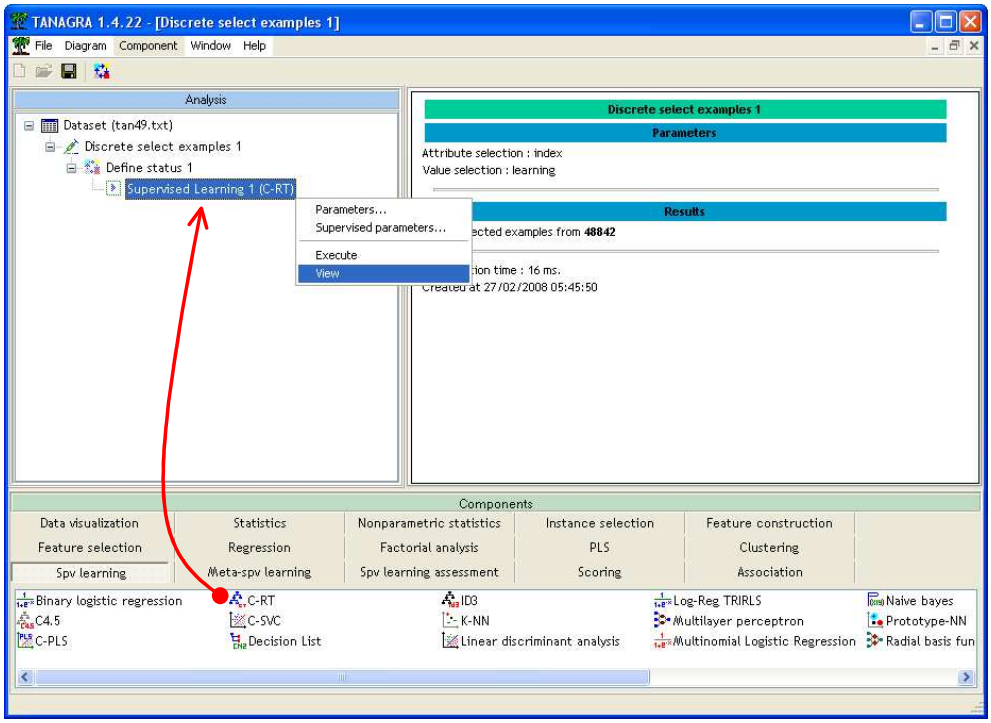
Pour décrire le problème de prédiction à résoudre, nous introduisons le composant DEFINE STATUS dans le diagramme. Le plus simple est d'actionner le raccourci dans la barre d'outils. Nous plaçons en TARGET le variable CLASS, en INPUT les autres variables, allant de AGE à NATIVE COUNTRY. Bien évidemment, il n'y a pas lieu de sélectionner INDEX ici.



3.4 Apprentissage avec la méthode C-RT

Le composant C-RT correspond à la méthode CART telle qu'elle est décrite dans l'ouvrage de référence (Breiman et al., 1984). Dans la phase d'expansion, l'indice de Gini est utilisé pour choisir les variables de segmentation. Dans la phase d'élagage, l'implémentation repose sur l'algorithme décrit dans le chapitre 10 de l'ouvrage.

Nous introduisons le composant C-RT (onglet SPV LEARNING) dans le diagramme. Nous activons le menu VIEW pour obtenir les résultats. Selon la puissance de la machine, le résultat sera plus ou moins long à venir (0.5 secondes sur ma machine).



Détaillons les différentes sections du rapport fourni par TANAGRA.

3.4.1 Matrice de confusion

Classifier performances

Error rate	0.1490

Values prediction

Value	Recall	1-Precision
more	0.5474	0.2431
less	0.9453	0.1295

Confusion matrix

	more	less	Sum
more	1298	1073	2371
less	417	7212	7629
Sum	1715	8285	10000

La matrice de confusion confronte les vraies valeurs et les valeurs prédites de SALARY sur les 10000 observations ayant participé à l'apprentissage (growing + pruning). Elle est accompagnée du taux d'erreur qui est de 0.1490 dans notre exemple. Etant calculé sur les données ayant servi à construire l'arbre, cet indicateur est souvent optimiste. Néanmoins, l'importance de l'écart dépend en partie de l'aptitude de la technique à coller exagérément aux données.

3.4.2 Partition des données

Data partition

Growing set	6700
Pruning set	3300

TANAGRA nous indique que parmi les 10000 observations dédiées à l'apprentissage, il a réservé 6700 observations pour l'expansion de l'arbre (growing set), 3300 pour le post élagage (pruning set). La partition a été effectuée de manière aléatoire.

3.4.3 Séquence d'arbres

Trees sequence (# 32)

N°	# Leaves	Err (growing set)	Err (pruning set)
32	1	0.2363	0.2388
28	6	0.1466	0.1539
20	39	0.1193	0.1479
1	205	0.0904	0.1700

Nous détaillerons ce tableau plus loin. A ce stade, on se contentera de constater que l'arbre le plus grand lors de la phase d'expansion comporte 205 feuilles, le taux d'erreur sur le growing set est de 0.0904, sur le pruning set 0.1700. L'arbre optimal sur le pruning set comporte 39 feuilles, avec un taux d'erreur 0.1479 (pruning set).

En appliquant la règle de l'écart type (1-SE RULE), **l'arbre retenu par CART comporte 6 feuilles avec un taux d'erreur de 0.1539 sur le pruning set** (à comparer avec les 205 feuilles obtenues initialement !). Nous détaillerons plus loin la procédure de calcul utilisée par CART.

Enfin l'arbre trivial composé de la seule racine présente un taux d'erreur de 0.2388.

3.4.4 Description de l'arbre

Tree description

Number of nodes	11
Number of leaves	6

Decision tree

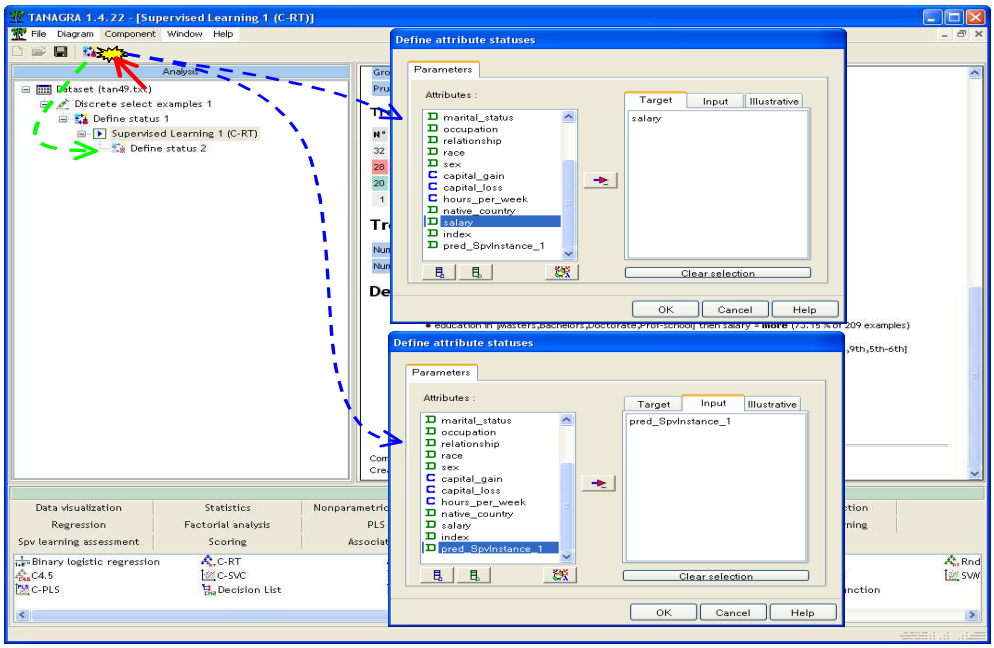
- relationship in [Husband,Wife]
  - education in [Masters,Bachelors,Doctorate,Prof-school] then salary = **more** (73.15 % of 209 examples)
  - education in [7th-8th,HS-grad,Some-college,Assoc-voc,Assoc-acdm,11th,10th,Preschool,12th,1st-4th,9th,5th-6th]
    - capital_gain < 5095.5000
      - capital_loss < 1794.0000 then salary = **less** (72.96 % of 1953 examples)
      - capital_loss >= 1794.0000 then salary = **more** (71.79 % of 78 examples)
    - capital_gain >= 5095.5000 then salary = **more** (96.43 % of 112 examples)
- relationship in [Not-in-family,Own-child,Unmarried,Other-relative]
  - capital_gain < 7073.5000 then salary = **less** (94.87 % of 3608 examples)
  - capital_gain >= 7073.5000 then salary = **more** (93.22 % of 59 examples)

La dernière section décrit l'arbre de décision produit lors de l'induction. On se bornera à noter que les attributs RELATIONSHIP, EDUCATION, CAPITAL_GAIN et CAPITAL_LOSS semblent les plus déterminants.

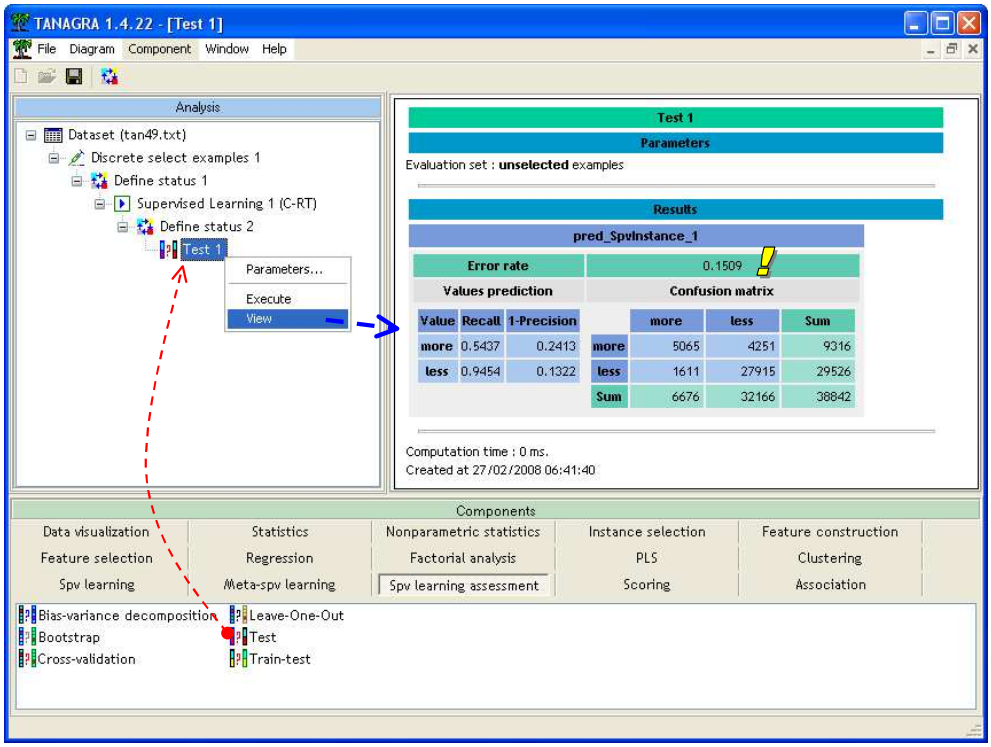
3.5 Evaluation sur l'échantillon test

Les ensembles growing et pruning participent, chacun à leur manière, à l'élaboration du modèle de prédiction. A ce titre, ils fournissent une estimation optimiste des performances puisque l'arbre est optimisé pour ces données. Pour obtenir une évaluation réellement non biaisée, il faut utiliser un ensemble test qui n'a jamais participé, de près ou de loin, à l'apprentissage. C'est à ce stade que nous mettons à contribution les individus « Index = test ».

Dans un premier temps, nous introduisons de nouveau le composant DEFINE STATUS dans le diagramme. Nous devons indiquer à TANAGRA la variable à prédire observée SALARY (TARGET) et la variable PRED_SPVINSTANCE_1 produite automatiquement par le composant C-RT (INPUT). Cette nouvelle colonne dans nos données comporte les valeurs prédites par l'arbre de décision, tant sur les données sélectionnées que sur les données non sélectionnées.



Dans un deuxième temps, nous introduisons le composant TEST (onglet SPV LEARNING ASSESSMENT) dans le diagramme. Elle est paramétrée par défaut pour construire la matrice de confusion, et calculer le taux d'erreur, sur les données non sélectionnées.



Nous activons le menu VIEW. Nous obtenons un taux d'erreur de 0.1509 calculé sur les 38842 individus que nous avons mis de côté initialement.

Ce n'est qu'une estimation bien sûr. Mais étant calculé sur un effectif aussi élevé, nous pouvons penser qu'il est relativement fiable : l'intervalle de confiance à 95% est [0.1473 ; 0.1545].

4 Quelques variantes autour du post-élagage

4.1 La 0-SE RULE

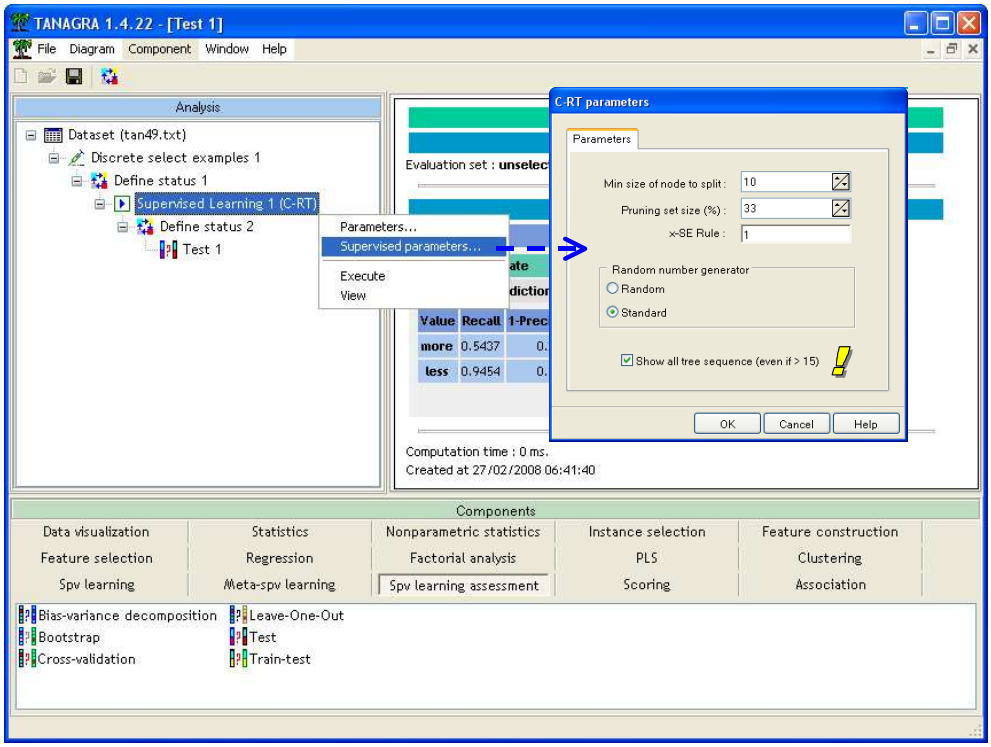
Une question revient très souvent chez les utilisateurs : pourquoi ne pas choisir directement l'arbre qui minimise l'erreur sur le pruning set ? Cette partie des données n'a pas servi à l'expansion de l'arbre, nous serons ainsi assurés de choisir un arbre « optimal ».

Il y a plusieurs réponses possibles. La première repose sur le bon sens : pourquoi reporter sur l'échantillon pruning ce que nous voulions justement éviter sur l'échantillon growing ? A savoir éviter de trop optimiser sur un échantillon au risque d'ingérer indûment les spécificités des données ?

La seconde réponse repose sur l'étude de la courbe opposant le nombre de feuilles de l'arbre (sa complexité) et le taux d'erreur sur le pruning set. Voyons comment obtenir cette courbe avec TANAGRA.

4.1.1 Courbe d'erreur en fonction de la complexité de l'arbre

Pour obtenir le détail de la courbe d'erreur, nous activons le menu contextuel SUPERVISED PARAMETERS du composant SUPERVISED LEARNING 1 (C-RT) dans le diagramme. Nous sélectionnons l'option SHOW ALL TREE SEQUENCE.



Nous actionnons le menu VIEW, le tableau retraçant les erreurs est détaillé maintenant. Nous remarquons plusieurs choses : le nombre de feuilles des arbres qui ont été testés n'est pas régulier, le mécanisme de coût complexité permet de réduire considérablement les solutions à évaluer ; l'erreur sur l'échantillon growing diminue constamment à mesure que le nombre de feuilles augmente ; l'erreur sur le pruning set diminue rapidement d'abord, semble stagner sur un palier, puis se dégrade lorsque le nombre de feuilles devient exagéré. Nous reproduisons le tableau (Tableau 1) et le graphique correspondant (Figure 2).

Nous constatons que les solutions allant d'un arbre avec 7 feuilles à un arbre avec 43 feuilles sont similaires en termes de taux d'erreur sur l'échantillon pruning. L'arbre « optimal » comporte 39 feuilles, il propose un taux d'erreur de 0.1479. Mais nous comprenons aisément que nous avons tout intérêt à choisir un arbre proche du « coude » dans la courbe d'erreur (Figure 2). Nous conservons un bon niveau de performances avec un arbre réduit.

N°	# Leaves	Err (growing set)	Err (pruning set)
32	1	0.2363	0.2388
31	3	0.1748	0.1806
30	4	0.1593	0.1664
29	5	0.1516	0.1567
28	6	0.1466	0.1539
27	7	0.1446	0.1497
26	11	0.1391	0.1488
25	16	0.1340	0.1488
24	21	0.1293	0.1485
23	22	0.1284	0.1488
22	28	0.1248	0.1485
21	33	0.1221	0.1494
20	39	0.1193	0.1479
19	43	0.1176	0.1479
18	46	0.1164	0.1506
17	51	0.1148	0.1521
16	73	0.1082	0.1552
15	80	0.1064	0.1558
14	83	0.1057	0.1555
13	91	0.1039	0.1552
12	94	0.1033	0.1552
11	107	0.1009	0.1567
10	116	0.0994	0.1570
9	139	0.0960	0.1606
8	154	0.0942	0.1642
7	158	0.0937	0.1642
6	168	0.0927	0.1648
5	174	0.0921	0.1670
4	188	0.0910	0.1679
3	194	0.0907	0.1691
2	199	0.0906	0.1700
1	205	0.0904	0.1700

Tableau 1 - Complexité de l'arbre et taux d'erreur growing / pruning

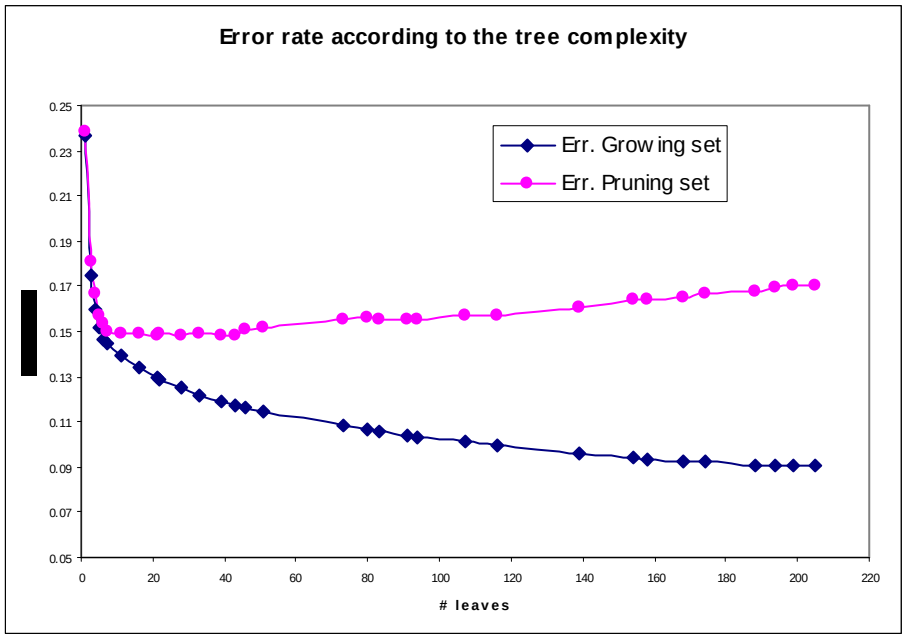


Figure 2 - Evolution du taux d'erreur en fonction de la complexité de l'arbre

4.1.2 Fonctionnement de la règle de l'écart type : 1-SE RULE

Comment a procédé CART pour choisir l'arbre à 6 feuilles ?

L'idée des auteurs de la méthode est tout simplement de sélectionner, dans les séquences de solutions proposées, le plus petit arbre dont les performances ne s'écartent pas significativement de l'arbre optimal sur le pruning set. Ils calculent pour cela une erreur seuil qui s'apparente à la borne haute de l'intervalle de confiance de l'erreur.

L'erreur « optimale » de l'arbre comportant 39 feuilles est $\varepsilon = 0.1479$, son écart-type estimé

$$\sigma = \sqrt{\frac{\varepsilon \times (1-\varepsilon)}{n}} = \sqrt{\frac{0.1479 \times (1-0.1479)}{3300}} = 0.00617977$$

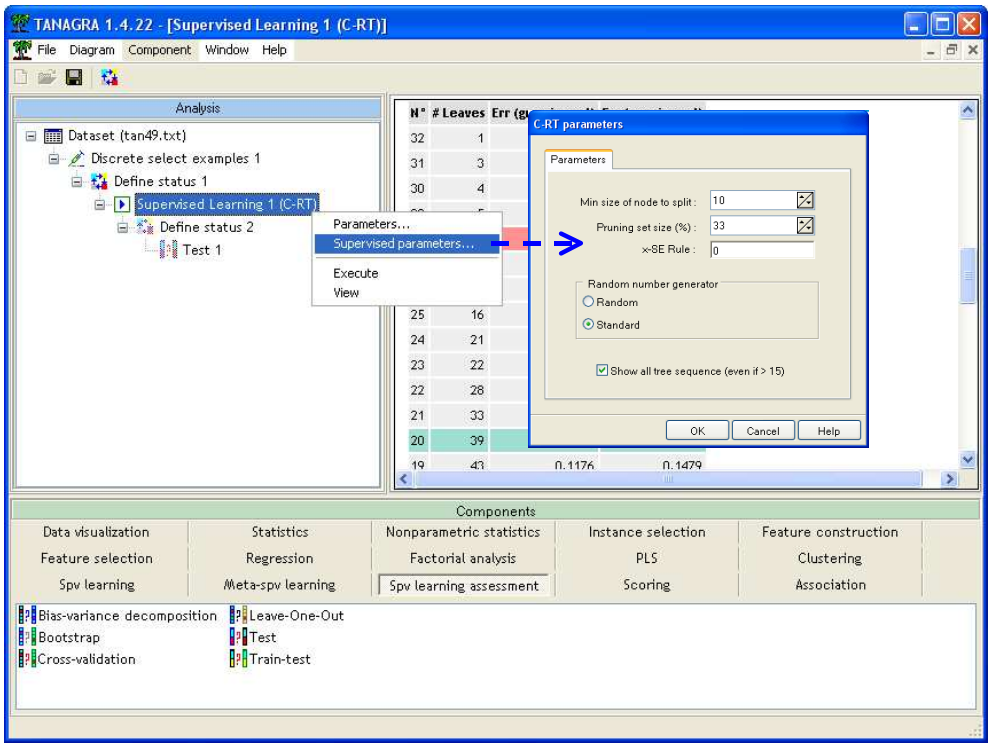
La borne haute définie par la règle de l'écart type (1-SE RULE ; θ = 1) est

$$\varepsilon_{seuil} = \varepsilon + \theta \times \sigma = \varepsilon + 1 \times \sigma = 0.1541$$

Nous cherchons dans notre tableau **le plus petit arbre dont l'erreur sur le pruning set est en dessous de ce seuil**. Il s'agit de l'arbre **comportant 6 feuilles**, avec un taux d'erreur de 0.1539. C'est le mécanisme que CART met en place pour sélectionner l'arbre final, il illustre à merveille le principe de parcimonie.

4.1.3 Performances de l'arbre 0-SE RULE (θ = 0)

Néanmoins, nous voulons évaluer les performances de l'arbre optimal comportant 39 feuilles. Nous paramétrons de nouveau le composant SUPERVISED LEARNING 1 (C-RT) en imposant la valeur 0-SE RULE.



Nous relançons l'arbre en clinquant sur le menu VIEW, l'arbre choisit cette fois-ci comporte 39 feuilles. Le taux d'erreur sur le growing set est de 0.1193, sur le pruning set de 0.147…

Trees sequence (# 32)

N°	# Leaves	Err (growing set)	Err (pruning set)
32	1	0.2363	0.2388
31	3	0.1748	0.1806
30	4	0.1593	0.1664
29	5	0.1516	0.1567
28	6	0.1466	0.1539
27	7	0.1446	0.1497
26	11	0.1391	0.1488
25	16	0.1340	0.1488
24	21	0.1293	0.1485
23	22	0.1284	0.1488
22	28	0.1248	0.1485
21	33	0.1221	0.1494
20	39	0.1193	0.1479

...et sur l'ensemble de l'échantillon d'apprentissage (growing + pruning) de 0.1287.

Classifier performances

Error rate	0.1287

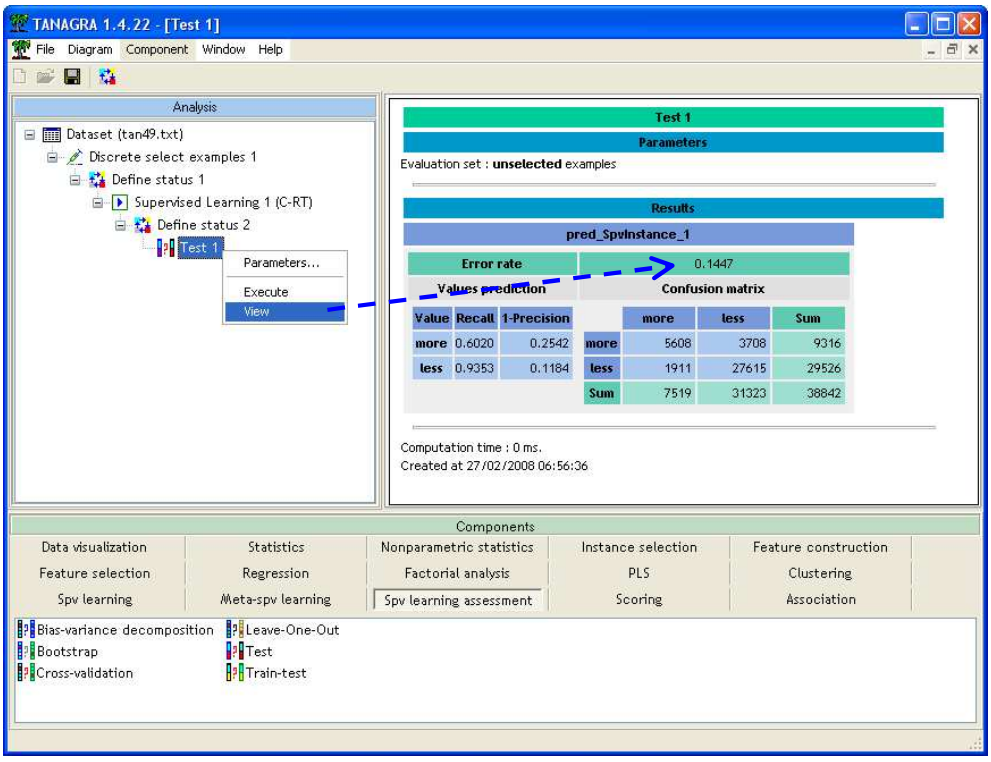
Values prediction

Value	Recall	1-Precision
more	0.6326	0.2171
less	0.9455	0.1077

Confusion matrix

	more	less	Sum
more	1500	871	2371
less	416	7213	7629
Sum	1916	8084	10000

Voyons maintenant ce qu'il en est sur l'échantillon test, nous activons pour cela le menu VIEW du composant TEST 1 au bout du diagramme de traitement.



Le taux d'erreur en test est 0.1447, avec un intervalle de confiance à 95% égal à [0.1412 ; 0.1482]. Il n'est pas significativement différent de l'erreur de l'arbre à 6 feuilles produit par la 1-SE RULE (Paragraphe 3.5, page 8).

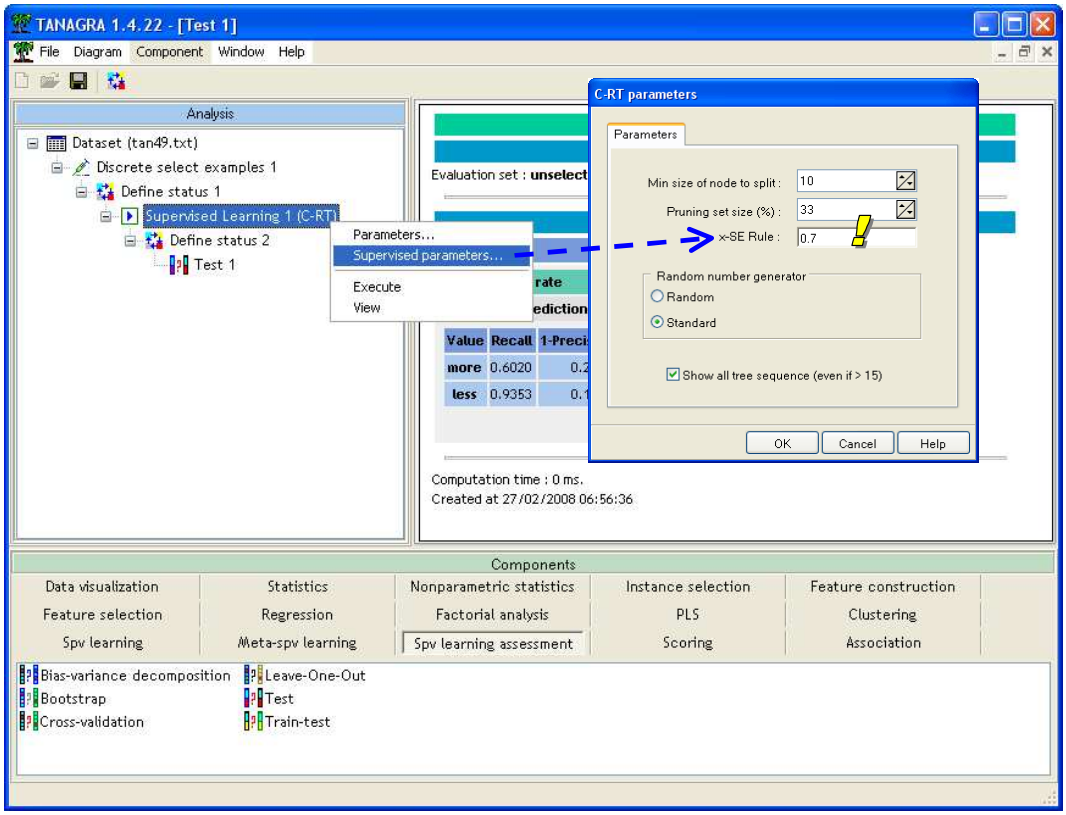
4.2 Exploiter la courbe d'erreur du post élagage

La règle de l'écart type (1-SE RULE) vise à produire un arbre plus simple tout en conservant un bon niveau de performances. Intuitivement, nous comprenons qu'elle cherche à se rapprocher du « coude » dans le graphique de l'erreur (Figure 2), l'endroit où nous avons épuisé l'information utile et commençons à ingérer les spécificités du fichier de données dans l'arbre. Nous constatons également qu'elle est assez finalement approximative, dépendante du paramètre de pénalisation θ. Dans cette section, nous allons essayer d'utiliser les outils à notre disposition (le Tableau 1 et la Figure 2), pour définir l'arbre correspondant à la solution souhaitée.

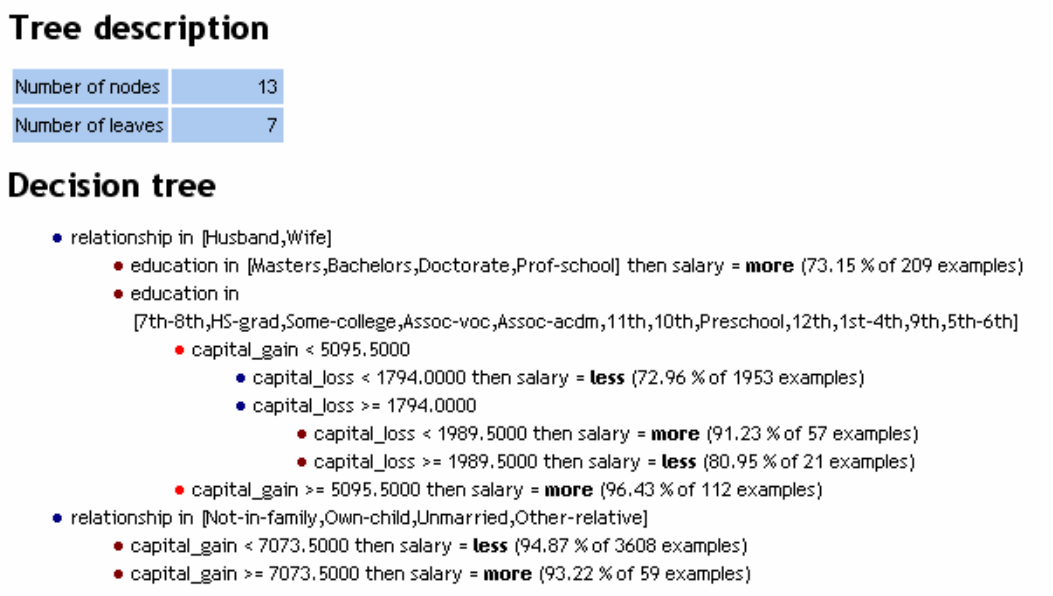
4.2.1 Déterminer le paramètre θ

A la lumière de la Figure 2, nous souhaitons produire l'arbre à 7 feuilles. En effet, c'est le modèle le plus simple situé sur le palier des meilleures solutions (arbre à 7 feuilles jusqu'à l'arbre à 43 feuilles, Tableau 1). Le taux d'erreur sur le pruning set est de 0.1497. Pour déterminer la valeur du paramètre θ permettant de produire cet arbre, il faut définir l'erreur seuil de manière à ce qu'elle soit située entre 0.1497 (arbre à 7 feuilles) et 0.1539 (arbre à 6 feuilles). En tâtonnant un peu, nous obtenons, entre autres, θ = 0.7, avec un erreur seuil $\varepsilon_{seuil} = 0.1479 + 0.7 \times 0.006 = 0.1522$.

Nous modifions le paramètre du composant SUPERVISED LEARNING 1 (C-RT), nous introduisons la valeur θ = 0.7 (Attention au point décimal selon la version de votre système d'exploitation).



L'arbre retenu comporte bien 7 feuilles.



Tree description

Number of nodes	13
Number of leaves	7

Decision tree

- relationship in [Husband,Wife]
  - education in [Masters,Bachelors,Doctorate,Prof-school] then salary = **more** (73.15 % of 209 examples)
  - education in [7th-8th,HS-grad,Some-college,Assoc-voc,Assoc-acdm,11th,10th,Preschool,12th,1st-4th,9th,5th-6th]
    - capital_gain < 5095.5000
      - capital_loss < 1794.0000 then salary = **less** (72.96 % of 1953 examples)
      - capital_loss >= 1794.0000
        - capital_loss < 1989.5000 then salary = **more** (91.23 % of 57 examples)
        - capital_loss >= 1989.5000 then salary = **less** (80.95 % of 21 examples)
    - capital_gain >= 5095.5000 then salary = **more** (96.43 % of 112 examples)
- relationship in [Not-in-family,Own-child,Unmarried,Other-relative]
  - capital_gain < 7073.5000 then salary = **less** (94.87 % of 3608 examples)
  - capital_gain >= 7073.5000 then salary = **more** (93.22 % of 59 examples)

Remarque 1 : Dans les logiciels où nous pouvons intervenir manuellement dans la construction de l'arbre, il n'est pas nécessaire de re-définir le paramétrage de la méthode, il suffit d'élaguer ou segmenter interactivement les sommets qui nous intéressent. Cette fonctionnalité est assez rare dans les logiciels libres[6], elle est en revanche incontournable dans les outils commerciaux (SPAD, SAS-EM, SPSS-ANSWERTREE, etc.).

Remarque 2 : Dans certains logiciels, il est possible de jouer sur d'autres paramètres pour étudier les solutions alternatives (voir le « paramètre de complexité » par exemple). Mais au final, il s'agit toujours de constituer des séquences d'arbres, puis d'essayer de détecter le modèle pour lequel une segmentation supplémentaire apporte une amélioration marginale.

4.2.2 Performances de la solution θ = 0.7 (Arbre à 7 feuilles)

Il nous reste maintenant à évaluer les performances de l'arbre à 7 feuilles. Nous activons le menu VIEW du composant TEST 1, nous obtenons un taux d'erreur de 0.1489, avec un intervalle de confiance à 95% égal à [0.1454 ; 0.1524]. Le tableau récapitulatif suivant retrace l'ensemble des opérations.

Theta-SE RULE	#Leaves	Err.Test	95% Conf.Interval
1	6	0.1509	0.1473 ; 0.1545
0.7	7	0.1489	0.1454 ; 0.1524
0	39	0.1447	0.1412 ; 0.1482

Manifestement, l'arbre à 6 feuilles suffit largement pour assurer un niveau de performances satisfaisant.

5 Conclusion

Parmi les innombrables variantes des techniques d'apprentissage des arbres de décision, CART est probablement celle qui détecte le mieux la bonne profondeur de l'arbre. Elle produit de ce fait, bien souvent, des modèles performants.

[6] SIPINA et ORANGE sont parmi les rares logiciels libres à proposer des procédures simples. Voir à ce sujet le didacticiel http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Interactive_Tree_Builder.pdf

En analysant la procédure de post-élagage, nous constatons qu'il est possible de simplifier encore l'arbre « optimal » détecté sur l'échantillon d'élagage (pruning set). La règle de l'écart type 1-SE RULE est la procédure standard de CART. Nous pouvons la paramétrer efficacement en exploitant au mieux les informations fournies par la courbe de l'erreur (Figure 2). L'objectif est de produire un arbre efficace avec une complexité réduite, mettant en jeu peu de variables, plus facile à manipuler et à interpréter.

CART est très populaire en France. La grande majorité des ouvrages en langue française la mettent systématiquement en avant dès qu'il s'agit de décrire une technique d'induction d'arbres de décision, au détriment des deux autres grandes références que sont CHAID (Kass, 1980) et C4.5 (Quinlan, 1993). Au delà des clivages culturels, il semble que CART s'impose naturellement dans les études plus ou moins automatisées où l'on cherche avant tout à produire un arbre performant. Il faut par ailleurs que l'on dispose de suffisamment d'observations, la partition growing/pruning participant à la fragmentation des données[7].

En revanche, lorsque nous travaillons sur de très grands ensembles de données, CART, comme toute les méthodes s'appuyant sur le post-élagage, s'avère très gourmande en temps de calcul. En effet, l'arbre maximal élaboré lors de la phase d'expansion peut comporter un nombre invraisemblable de feuilles. Inutilement d'ailleurs puisque la très grande majorité des branches seront élaguées par la suite. Ce travers est exacerbé lorsque la base est composée en grande partie de descripteurs continus. Dans ce cas, surtout lors de la phase exploratoire d'appréhension des données où nous essayons avant tout de déceler assez rapidement les relations entre les variables, nous préfèrerons la méthode CHAID basée sur un mécanisme de pré-élagage : la technique essaie de définir une règle d'arrêt judicieuse durant l'expansion de l'arbre.

[7] CART intègre certes une procédure de validation croisée adaptée aux petits effectifs pour la sélection de la taille adéquate de l'arbre. Mais elle est compliquée, difficile à faire passer auprès des non initiés.