Objectif

Un des principaux attraits des arbres de décision est la possibilité, pour l'utilisateur, d'intervenir lors d'élaboration du modèle. Cette propriété est précieuse dans les études réelles car elle donne à l'expert la faculté d'injecter dans le processus d'induction des connaissances externes qui permettent de mieux guider l'exploration.

TANAGRA ne dispose pas d'outils interactifs pour l'induction d'arbres de décision. Dans ce didacticiel nous utiliserons plutôt le logiciel SIPINA[1] qui est un de nos anciens projets, abandonné pour diverses raisons en 2000. Il propose de nombreux outils pour l'apprentissage supervisé, notamment une série de méthodes d'induction d'arbres. SIPINA présente essentiellement deux inconvénients qui seront corrigés dans TANAGRA : il propose uniquement des méthodes supervisées, il n'est pas possible de réaliser une classification ou une analyse factorielle par exemple ; l'utilisateur ne peut pas sauvegarder les paramètres de la séquence de traitements qu'il a définis, il est donc obligé de refaire toutes les opérations lors de la reprise d'un travail, ce qui n'est pas tenable dans un environnement professionnel.

Pour ma part, j'utilise parfois **SIPINA** en situation d'enseignement d'une part parce que la présentation graphique de l'arbre est plus avenante, et d'autre part parce qu'il est possible d'intervenir dans la construction de l'arbre, soit en élaguant manuellement des branches entières, soit en choisissant explicitement la variable de segmentation sur un sommet. A ma connaissance, **ORANGE** est le seul autre outil gratuit qui propose des fonctionnalités similaires via le composant INTERACTIVE TREE BUILDER. La présentation n'est pas graphique dans le cas d'ORANGE, l'arbre prend la forme d'un TREEVIEW, mais l'essentiel y est, il est possible d'intervenir dans le processus d'induction.

Dans ce didacticiel, à partir d'un exemple simple, nous verrons comment mettre en œuvre ces fonctionnalités avec nos deux logiciels. Nous verrons également que malgré leurs qualités, ils présentent des limites par rapport aux outils proposés dans le commerce.

Fichier

Nous utilisons le fichier IRIS_TREE.TXT (UCI IRVINE). Il comporte un attribut classe à trois modalités (type de fleur), 4 descripteurs, tous continus, et 150 exemples.

Nous avons sélectionné 75 observations pour l'apprentissage, 75 pour le test. Une colonne supplémentaire (STATUS) a été insérée pour indiquer le statut de chaque observation.

[1] http://eric.univ-lyon2.fr/~ricco/sipina.html, version recherche

Exploration avec SIPINA

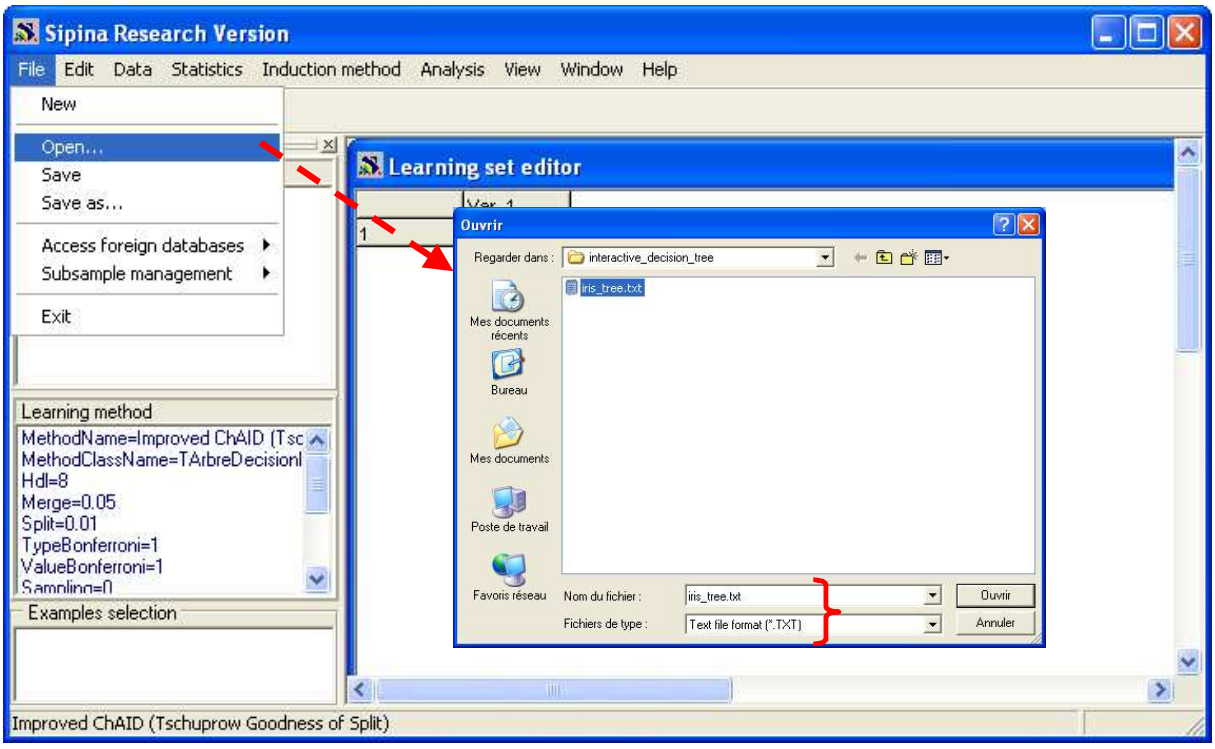
SIPINA est essentiellement piloté par menu, l'utilisateur doit réaliser une série de manipulation, dans le bon ordre, pour obtenir le résultat souhaité. Les traitements ne sont sauvegardés nulle part, l'utilisateur doit la redéfinir à chaque exécution du logiciel. Néanmoins ces tâches sont standardisées, nous les retrouvons quel que soit le mode de fonctionnement du logiciel.

Exécution de SIPINA

Au démarrage de SIPINA, la fenêtre principale est subdivisée en plusieurs parties : au centre la grille qui affiche les données chargées ; à gauche, une reprise des paramètres définis par l'utilisateur, à savoir la méthode utilisée, la subdivision en apprentissage et test de l'échantillon, et enfin, le problème de prédiction à résoudre.

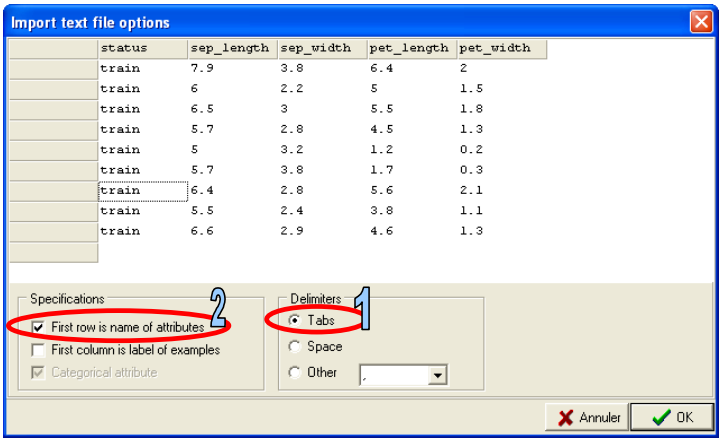
Importation des données

SIPINA accepte plusieurs formats. Nous utilisons le format texte avec séparateur tabulation dans ce didacticiel[2]. Pour importer les données, nous sélectionnons le menu FILE / OPEN. Dans la boîte de sélection de nom de fichier qui apparaît, nous précisons le bon format.



[2] Dans SIPINA, le point décimal est obligatoirement le caractère « . »

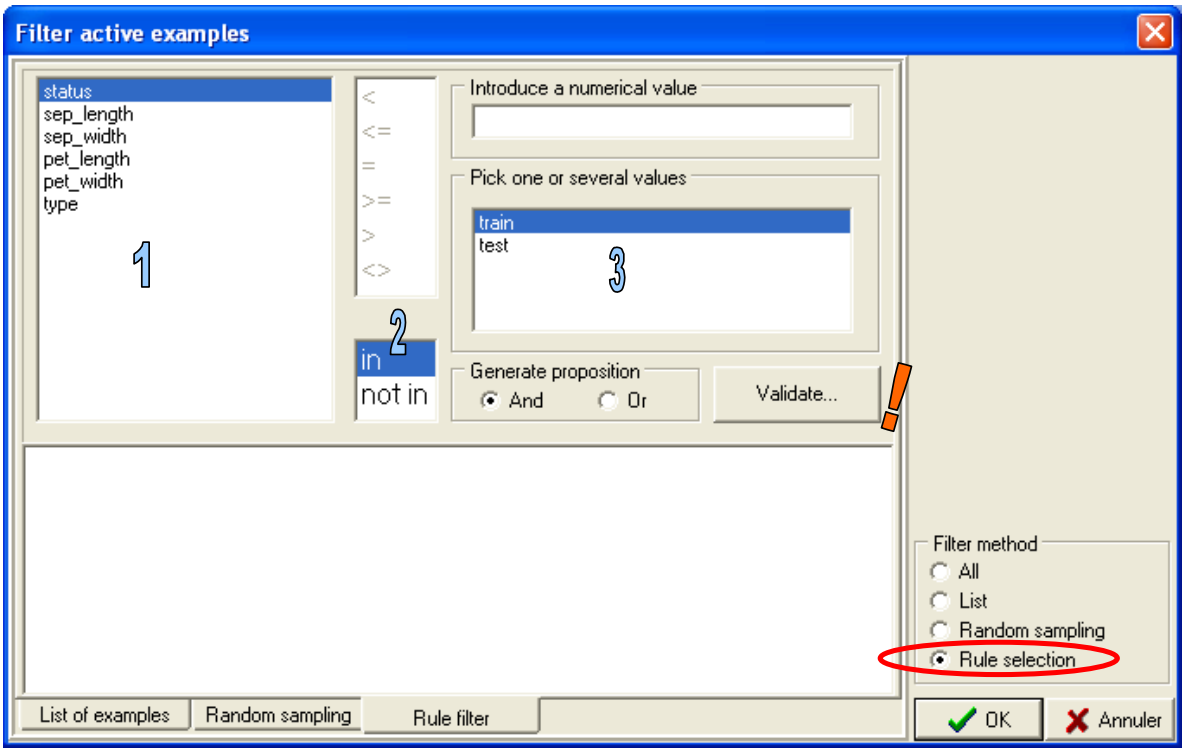
Une fenêtre apparaît, nous devons spécifier deux informations importantes : (1) le séparateur est le caractère « tabulation » ; (2) la première ligne des données indique le nom de chaque variable. Nous validons avec le bouton OK.



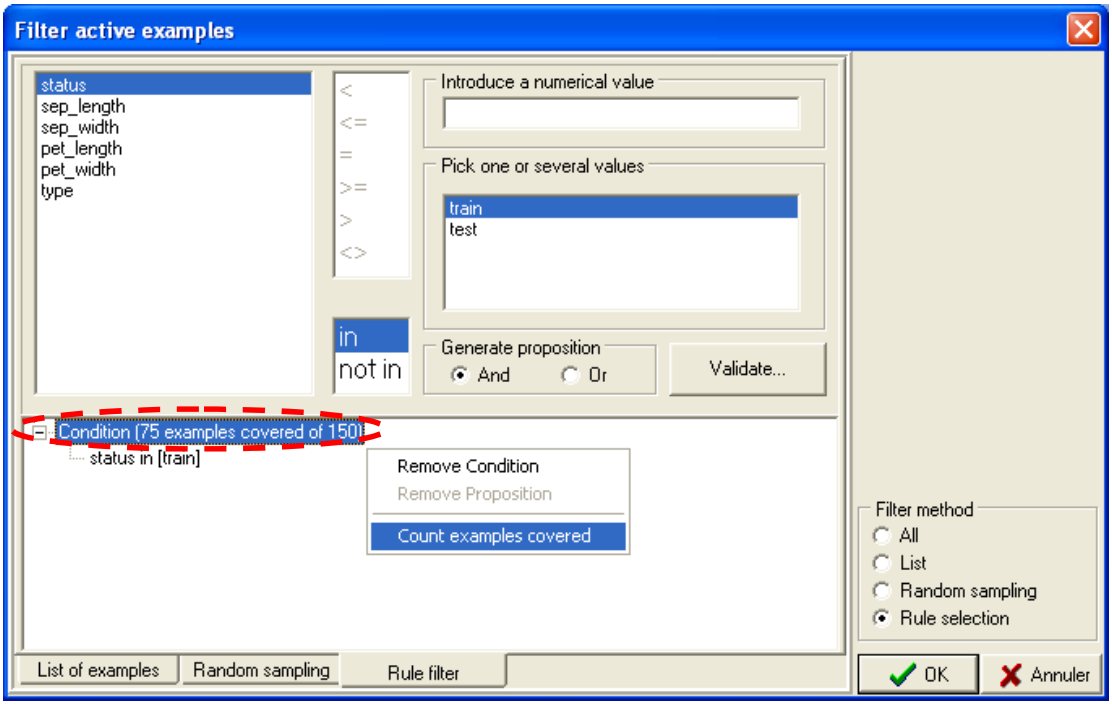
Les données sont alors automatiquement chargées dans la grille principale. Le type de chaque variable est déterminé par la première ligne de valeurs (la deuxième ligne du fichier s'il contient le nom des variables) : si elle est numérique, la variable est considérée continue ; si elle est alphanumérique, la variable est considérée discrète.

Sélectionner les observations en apprentissage et en test

Le rôle des observations est défini par la variable STATUS qui prend deux modalités TRAIN et TEST. Nous allons exploiter cette variable pour sélectionner les individus en apprentissage. Pour ce faire, nous sélectionnons le menu ANALYSIS / SELECT ACTIVE EXAMPLES, une boîte de dialogue apparaît, nous activons l'option RULE SELECTION.



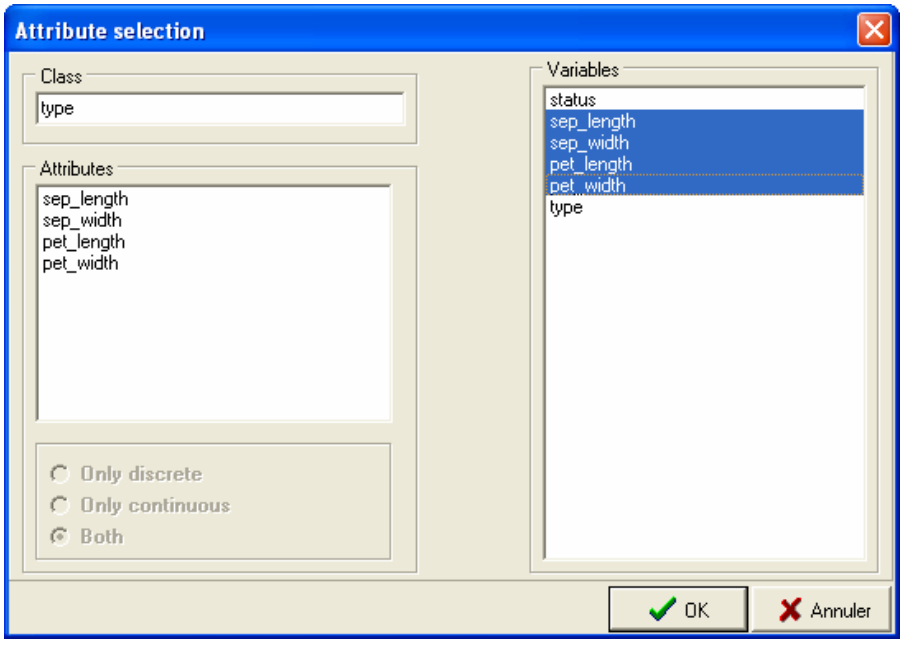
Pour définir la règle qui désigne les individus actifs, nous devons (1) sélectionner la variable, (2) choisir l'opérateur de la condition, (3) choisir la bonne modalité. Nous validons alors la condition (!). Le résultat apparaît dans la partie basse de la fenêtre, il est possible de compter les observations en activant le menu contextuel COUNT EXAMPLES COVERED.



Nous validons cette sélection en cliquant sur OK.

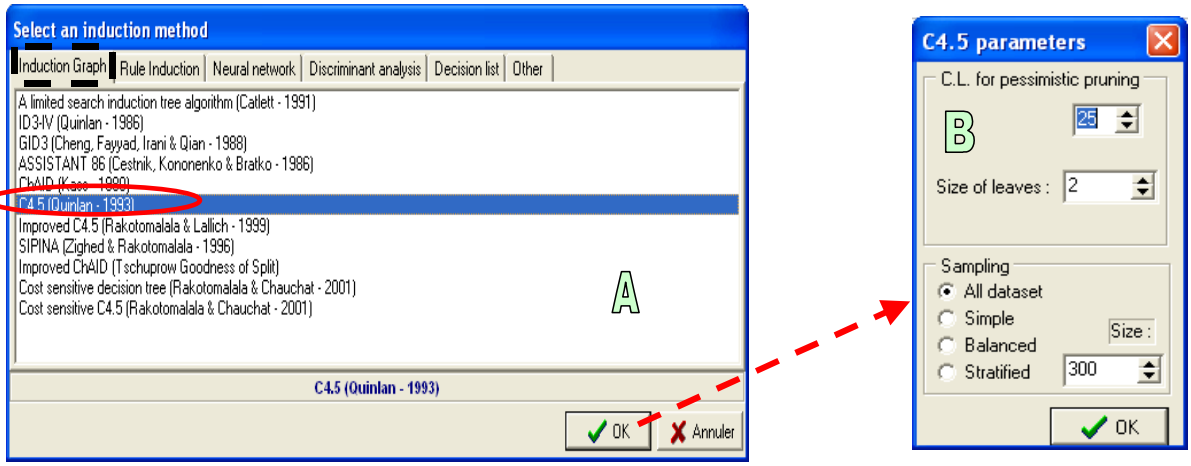
Sélectionner les variables de l'étude

Pour définir la variable cible et les variables prédictives, nous activons le menu ANALYSIS / DEFINE CLASS ATTRIBUTE. Une boite de dialogue permet de sélectionner les variables adéquates avec le principe du glisser / déposer. Bien entendu, il ne faut surtout pas utiliser la variable STATUS ici.

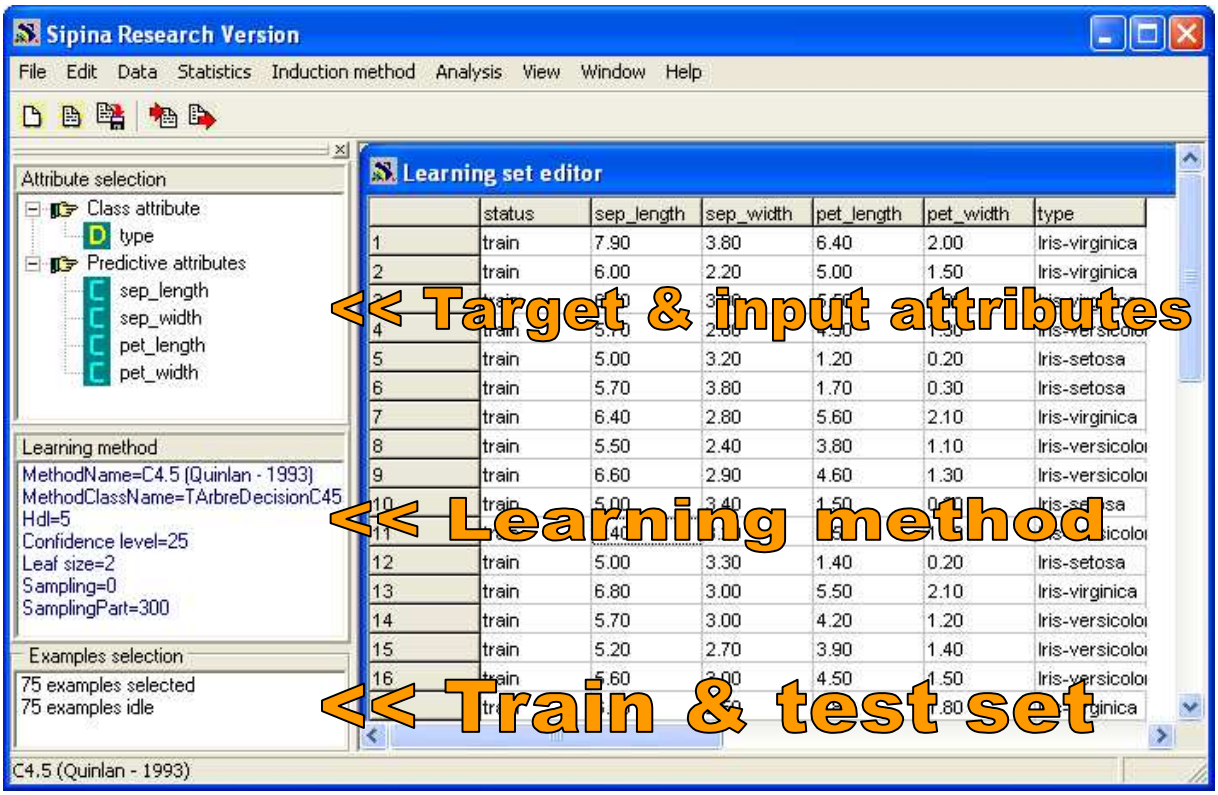


Définir la méthode d'apprentissage

Pour définir la méthode d'apprentissage, il faut activer le menu INDUCTION METHOD / STANDARD ALGORITHM. Une boîte de dialogue [A] apparaît, nous sélectionnons l'algorithme C4.5 dans l'onglet INDUCTION GRAPH. Après validation, une autre boîte [B] permet de paramétrer la méthode, nous nous contenterons des paramètres par défaut. Nous cliquons sur le bouton OK pour valider.

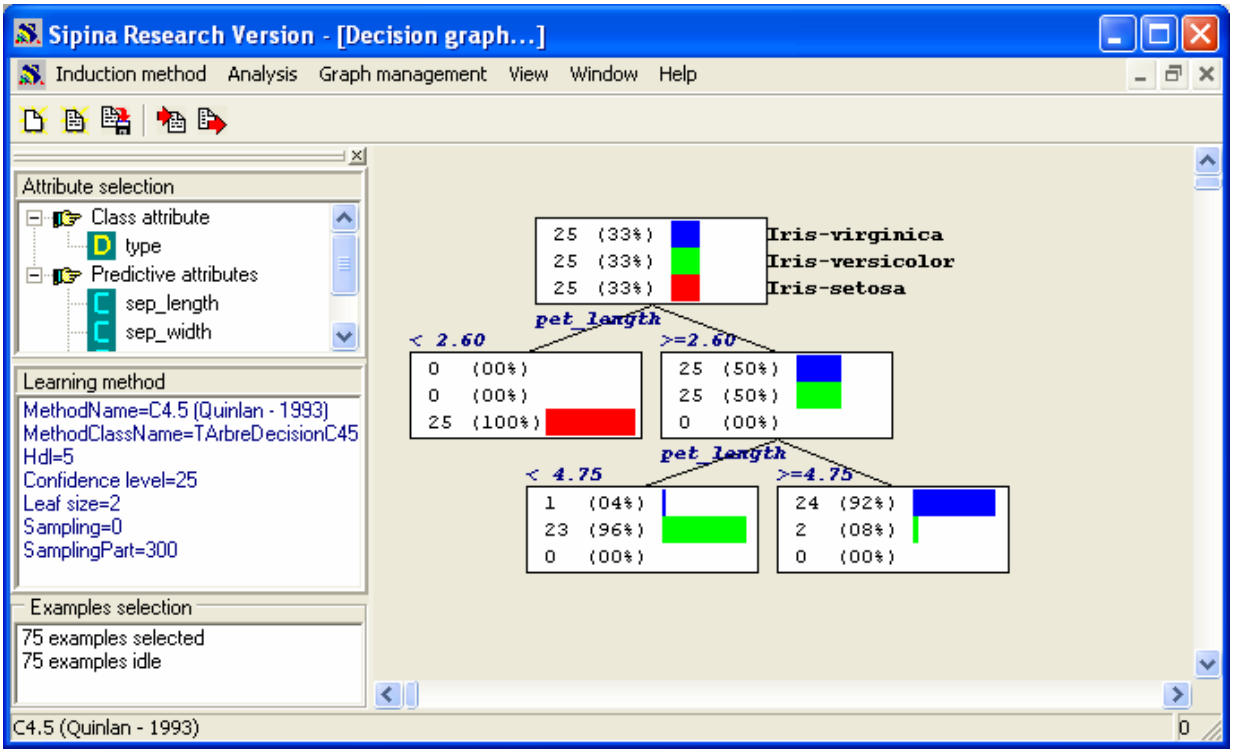


Après cette série d'opérations, nous avons sur la gauche de la fenêtre principale un résumé de toutes les manipulations entreprises. Un rapide coup d'œil permet de voir si tous les paramètres ont été introduits correctement.



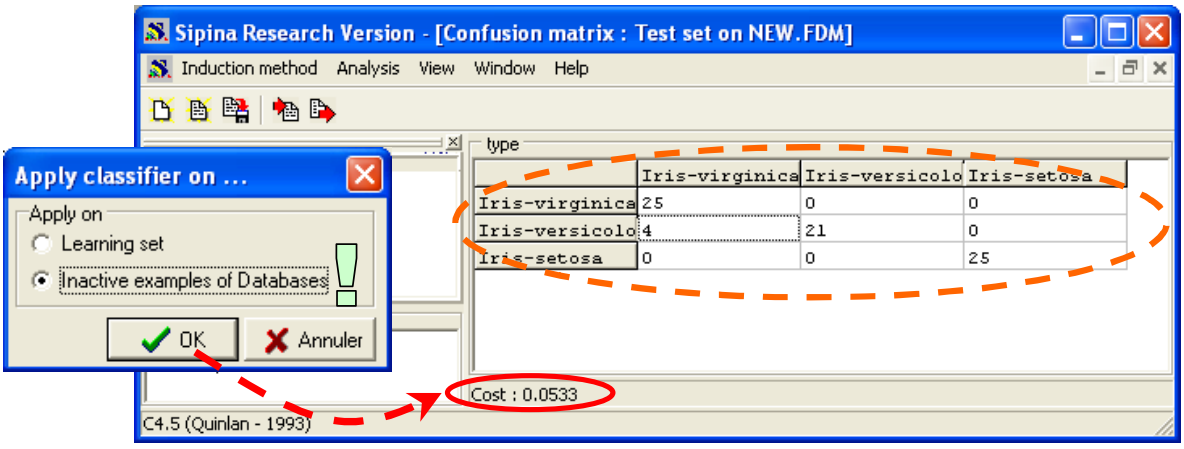
Analyse automatique et évaluation

Nous activons le menu ANALYSIS / LEARNING pour lancer la construction de l'arbre avec la méthode C4.5. L'arbre s'affiche automatiquement.



L'arbre est très simple, 3 règles permettent de déterminer le type d'IRIS à partir de ses caractéristiques physiques, il semble que seule la variable PET_LENGTH soit réellement pertinente.

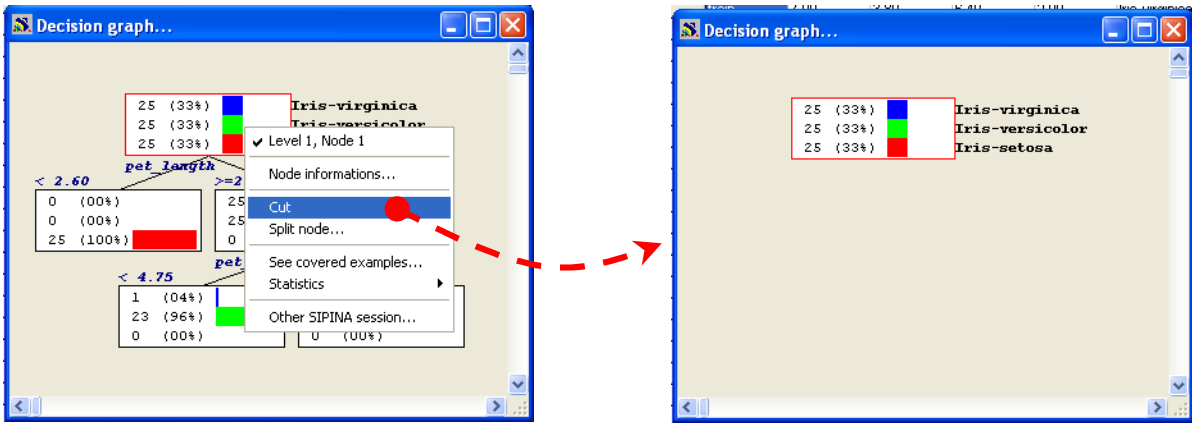
Pour évaluer les performances de ce modèle sur l'échantillon test, il nous faut activer le menu ANALYSIS /TEST et sélectionner l'option INACTIVE EXAMPLES OF DATABASE, nous obtenons la matrice de confusion en test et le taux d'erreur associé (5.33%).



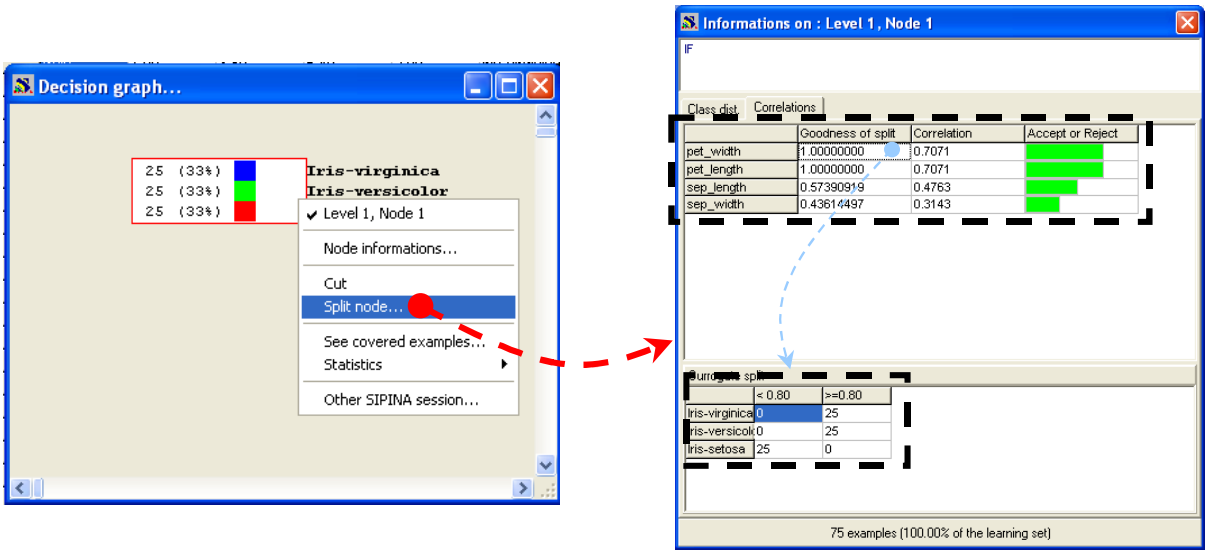
Exploration interactive

Seule la variable PET_LENGTH a été utilisée dans notre arbre. On peut se demander dans quelle mesure les autres variables pourraient participer à la construction du modèle de prédiction. Dans cette deuxième partie, nous voulons développer nous même l'arbre en analysant les segmentations candidates sur chaque sommet.

Nous faisons afficher l'arbre en activant le menu WINDOW / DECISION GRAPH. Dans un premier temps, nous devons élaguer l'arbre à partir de sa racine. La manœuvre est relativement simple, nous sélectionnons ce premier sommet, puis avec le menu contextuel, nous cliquons sur le menu CUT.



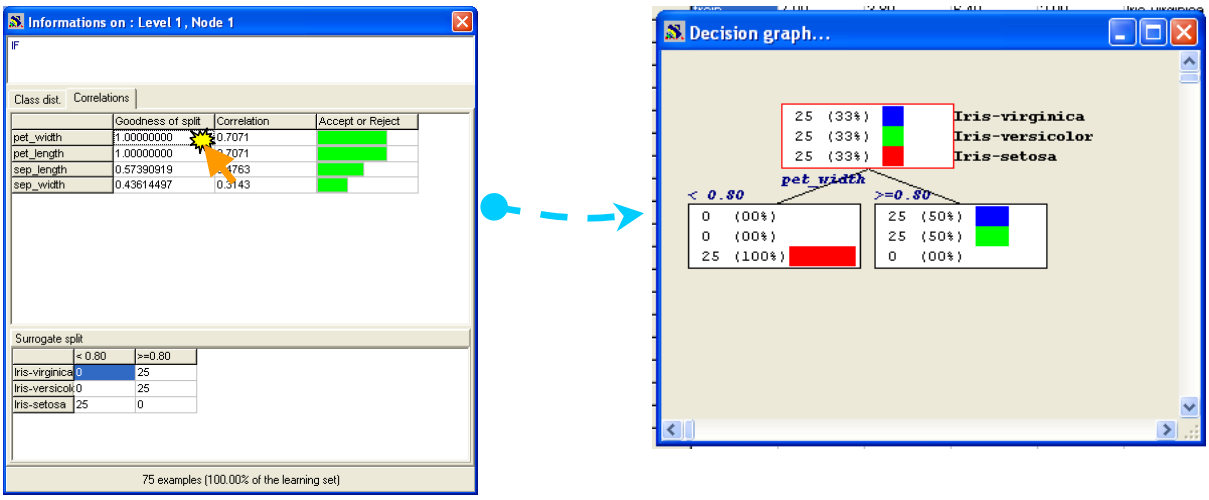
Dans un second temps, nous affichons les segmentations candidates en cliquant sur l'option SPLIT NODE du menu contextuel. Une nouvelle fenêtre apparaît avec l'ensemble des segmentations candidates sur le sommet sélectionné.



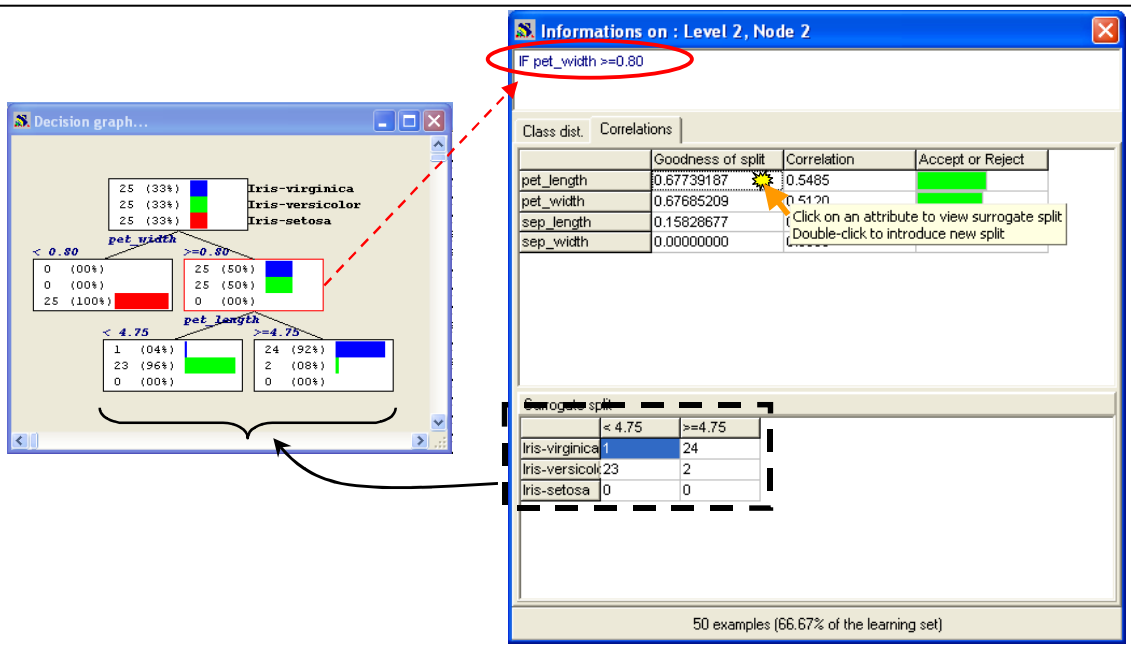
Nous observons que finalement PET_WIDTH et PET_LENGTH ont exactement le même pouvoir prédictif. Ils proposent un GAIN RATIO (GOODNESS OF SPLIT) de 1.0 et un T de TSCHUPROW (CORRELATION) de 0.7071 : en réalité, la variable PET_LENGTH a été sélectionnée dans l'analyse automatique à cause de l'ordre interne utilisé lors des calculs. La dernière colonne indique si la segmentation proposée répond positivement (en vert) ou non (en rouge) aux conditions d'admissibilité.

En sélectionnant la valeur de GOODNESS OF SPLIT en face de la variable PET_WIDTH, nous voyons s'afficher dans la partie basse de la fenêtre la segmentation proposée. Nous observons dans ce cas que le seuil de discrétisation est de 0.8. La distribution dans les feuilles générées est affichée.

Si nous voulons valider une segmentation avec la variable PET_WIDTH, il faut tout simplement double-cliquer sur la valeur du GOODNESS OF SPLIT.

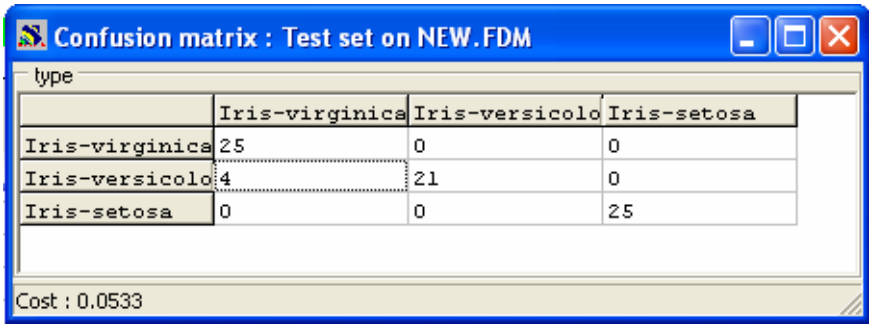


La feuille à gauche est uniquement constitué d'IRIS-SETOSA. Nous segmentons la feuille à droite, toujours avec la même démarche, avec la variable PET_LENGTH. Nous obtenons l'arbre suivant.



Bien entendu, c'est un exemple d'école, choisir entre PET_LENGTH ou PET_WIDTH ne nous paraît pas capital. Il en est tout autrement lorsque nous travaillons sur une étude réelle. L'importance des variables n'est pas le même. Certaines peuvent s'avérer plus fiables que d'autres, plus facilement mesurables, plus pertinentes pour l'interprétation, etc. Il est évident dans ce cas que le travail d'exploration dépasse les compétences du statisticien ou de l'informaticien, l'expert du domaine (le banquier, le spécialiser CRM, le médecin, etc.) a un rôle déterminant dans ce processus d'exploration.

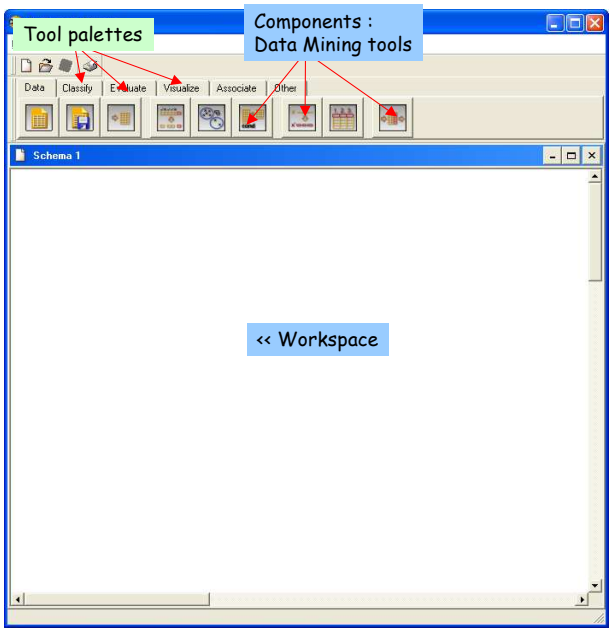
Par la suite, nous utilisons le même canevas pour évaluer les performances de ce modèle construit manuellement, à savoir calculer la matrice de confusion sur l'échantillon test (ANALYSIS / TEST). Dans cet exemple, voici ce qu'il en ressort.



Les performances sont les mêmes, ce qui n'est guère étonnant compte tenu des données que nous avons. Nous aurions dégradé significativement les performances en revanche si nous avions introduit une des deux variables SEP_LENGTH ou SEP_WIDTH dans l'arbre.

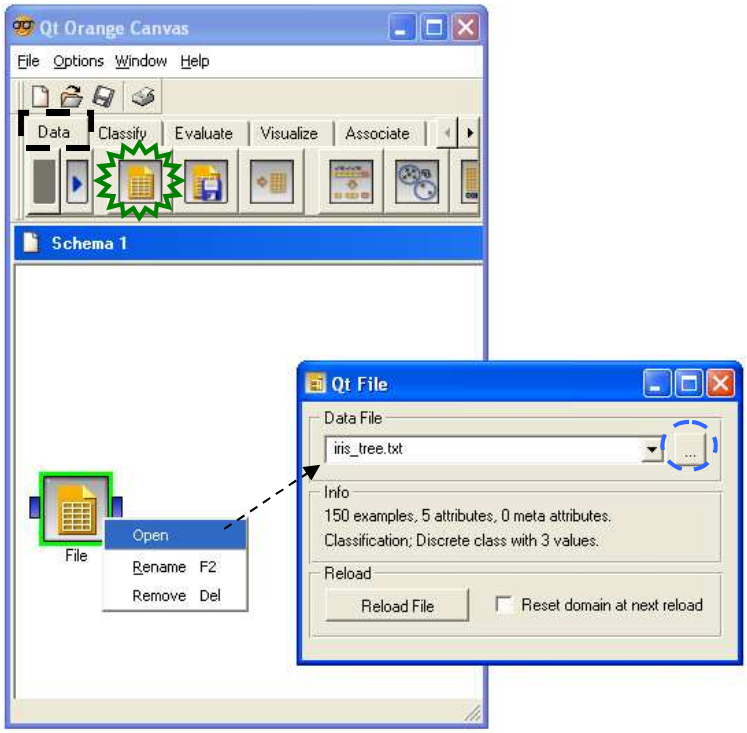
Exploration interactive avec ORANGE

ORANGE propose une interface composée de deux parties distinctes : un espace pour définir les traitements ; une palette d'outils située dans la partie haute de la fenêtre principale.



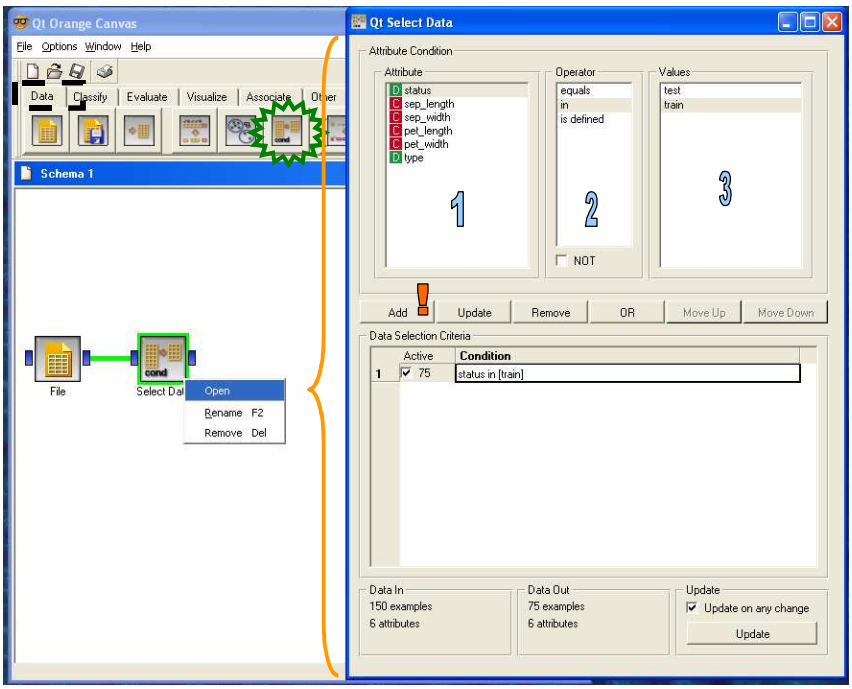
Préparer et charger les données

Nous utilisons le même fichier de données (IRIS_TREE.TXT). Nous plaçons donc le composant DATA dans le diagramme.



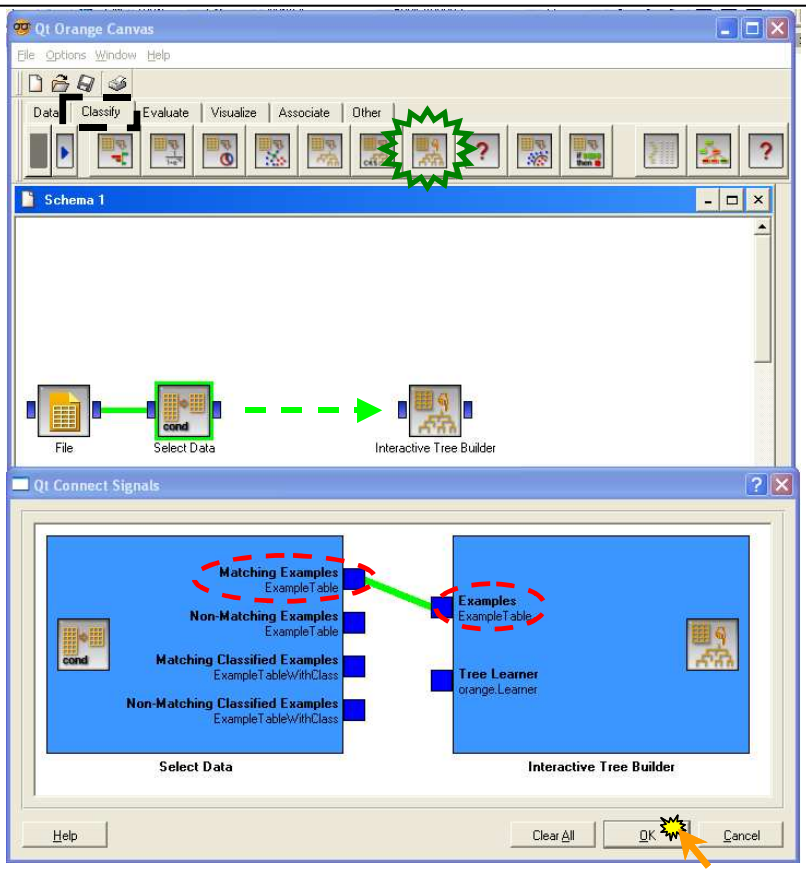
Sélection des observations pour l'apprentissage

Nous devons maintenant subdiviser l'échantillon en nous appuyant sur la variable STATUS. Pour ce faire, nous insérons le composant SELECT DATA (onglet DATA) dans le diagramme, nous le connectons au composant précédent, puis nous le paramétrons en cliquant sur le menu contextuel OPEN. Tout comme dans SIPINA, avec une démarche assez similaire, il est possible de définir les conditions permettant de réaliser une restriction sur notre ensemble de données.

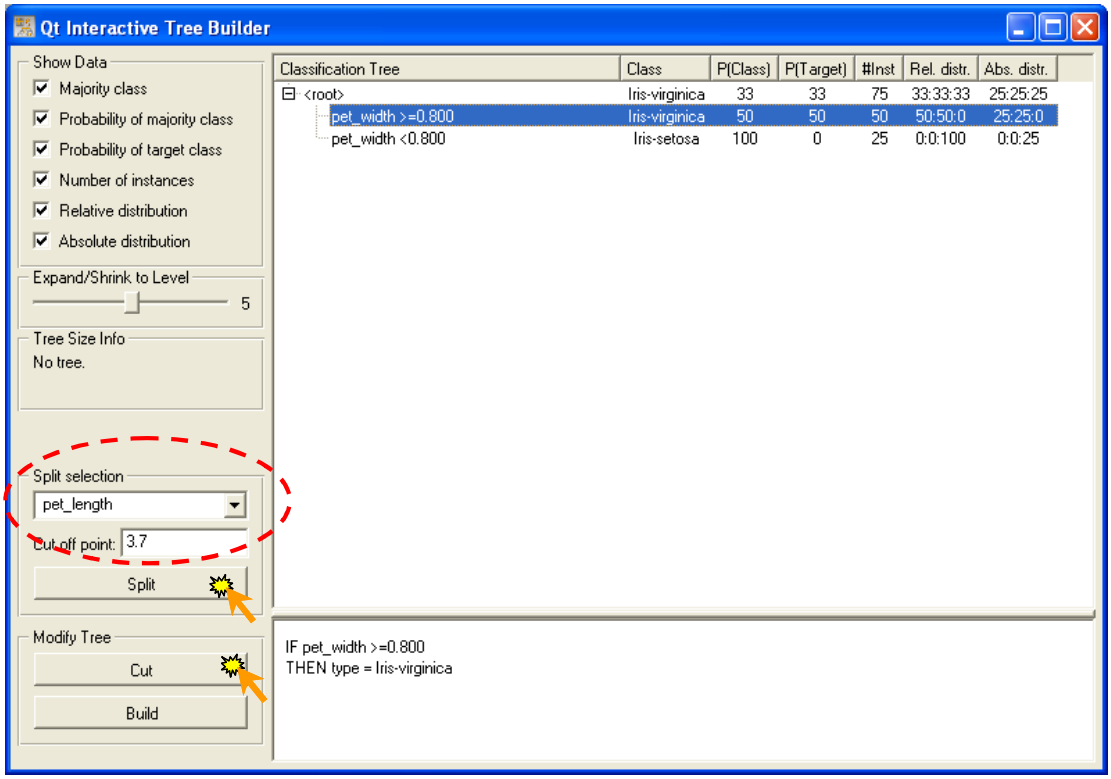


Analyse interactive

Les 75 observations, correspondant aux données d'apprentissage, sont maintenant sélectionnées. Nous plaçons à la suite le composant de construction interactive d'arbre de décision, il s'agit d'INTERACTIVE TREE BUILDER (onglet CLASSIFY). Lors de la connexion avec le composant SELECT DATA, une boîte de dialogue permet de préciser la nature de la connexion, nous validons la proposition par défaut.



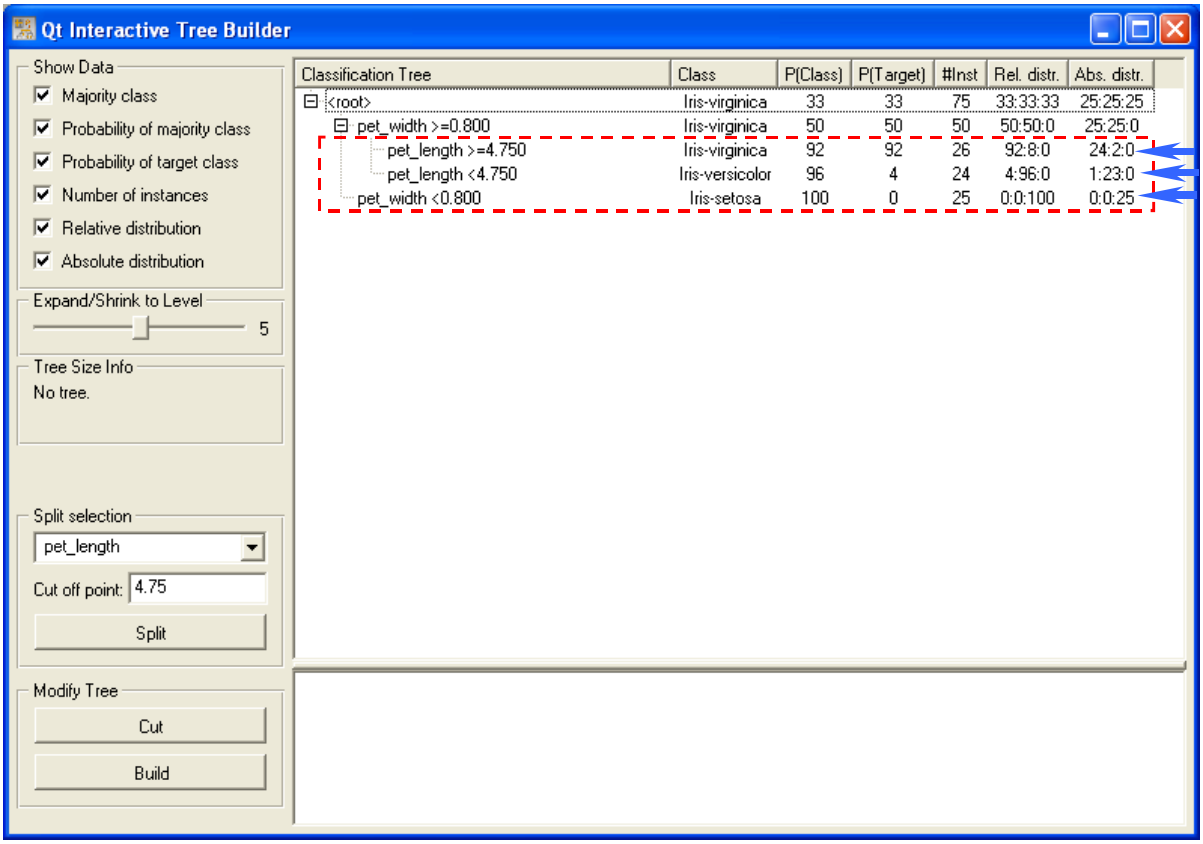
Nous cliquons alors sur l'option OPEN du menu contextuel pour faire apparaître l'outil d'exploration interactive.



Il est possible de construire l'arbre manuellement en choisissant la variable de segmentation et en cliquant sur SPLIT. Le seuil de discrétisation proposé est par défaut la moyenne de la variable sur le nœud (ex. ici la moyenne de PET_LENGTH sur le sommet considéré est de 3.7, nous avons vu avec SIPINA que le bon seuil de coupure est 4.75), l'avantage avec ORANGE est que l'on peut ajuster manuellement ce seuil.

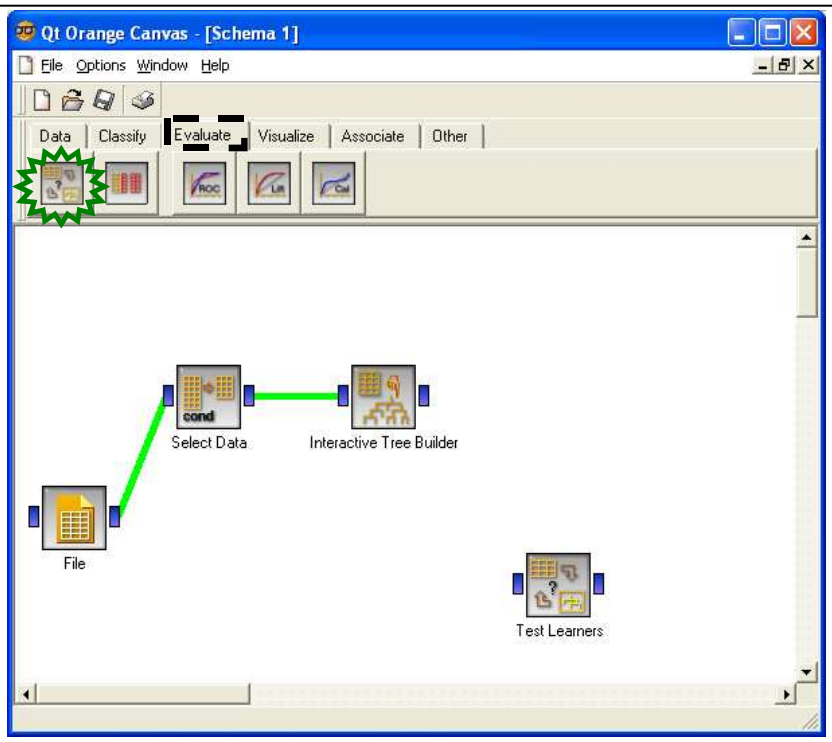
Après avoir développé à la main les premiers sommets, ORANGE propose une autre option très intéressante, nous pouvons demander au logiciel de finaliser de manière autonome le reste de l'arbre avec l'option BUILD.

Dans notre cas, nous nous contenterons de reproduire l'arbre qui a été construit avec SIPINA, nous obtenons donc le résultat suivant. Nous distinguons les trois feuilles de l'arbre.

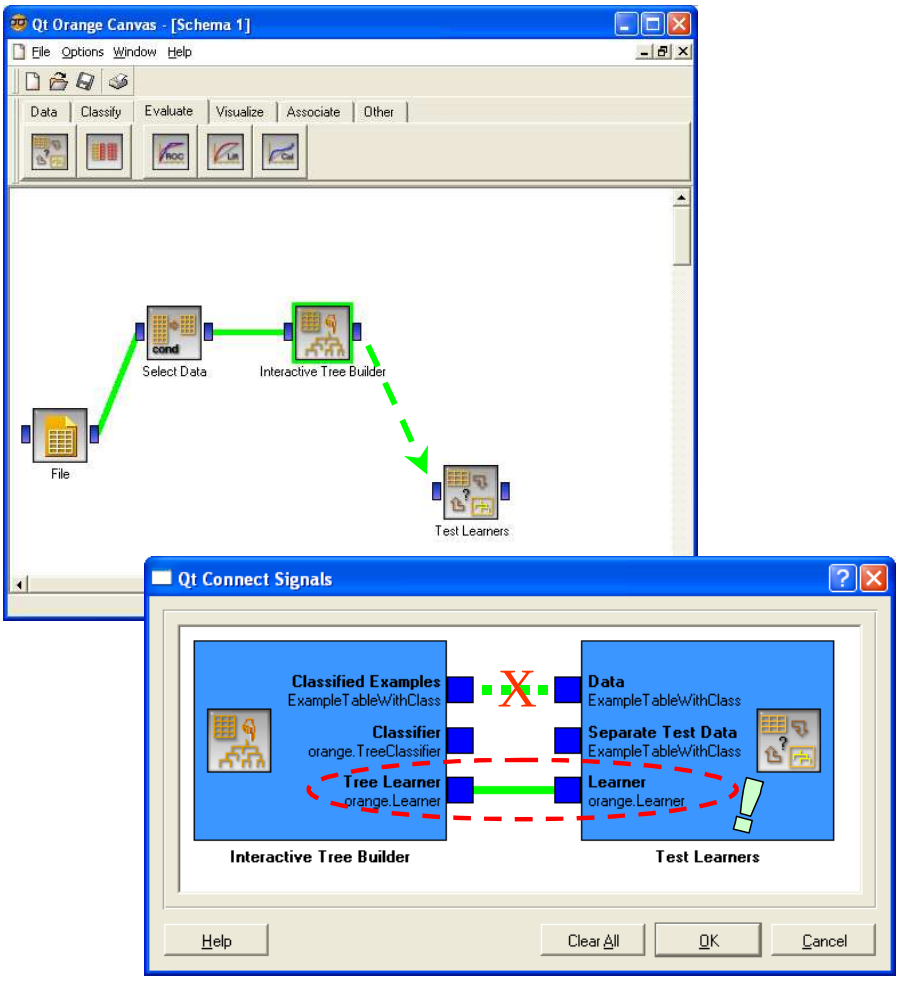


Evaluation sur l'échantillon test

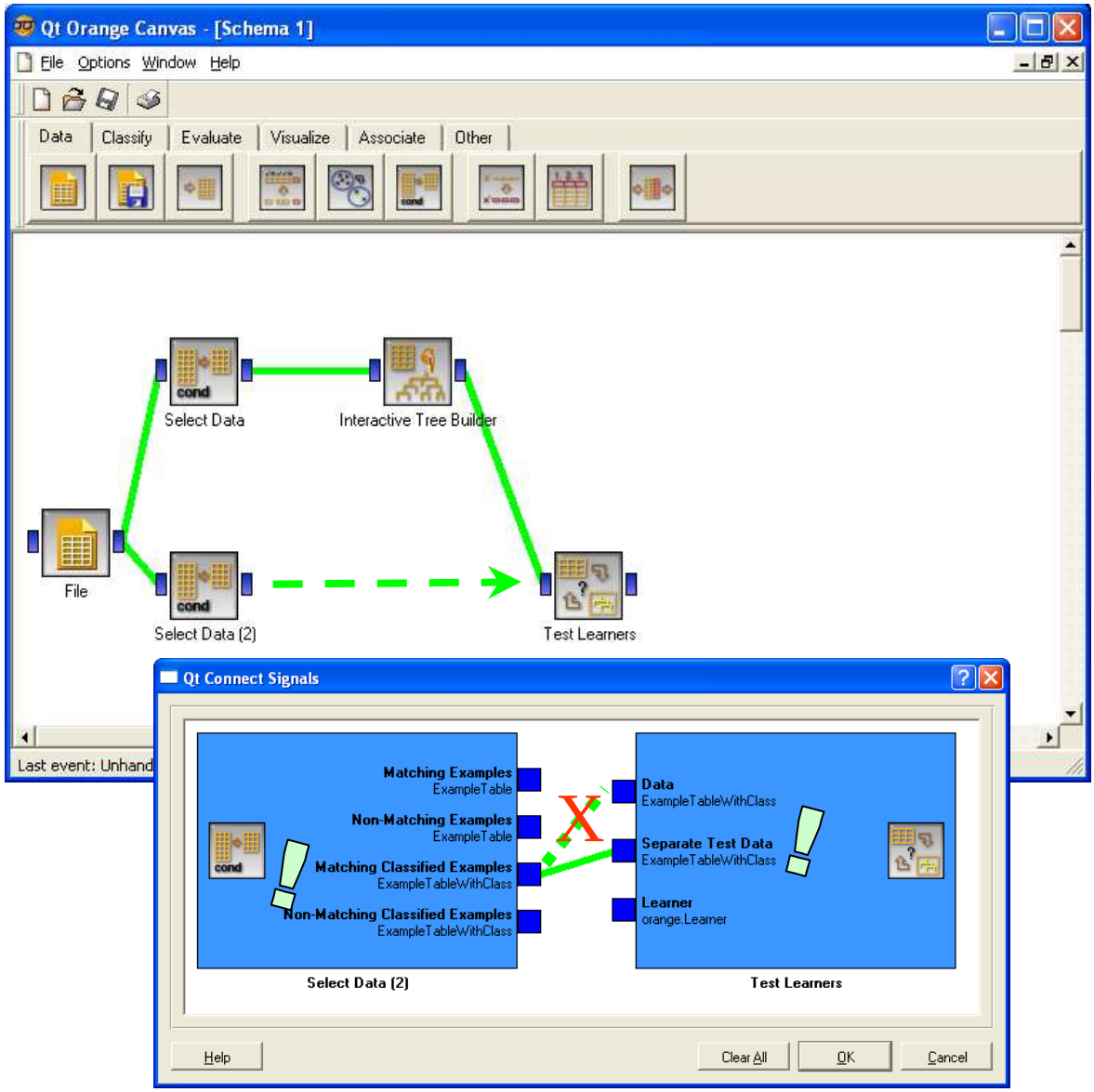
Il nous reste à évaluer les performances de cet arbre. Nous plaçons à cet effet le composant TEST LEARNERS de l'onglet EVALUATE. Nous devons lui connecter le modèle de prédiction et les données en test.



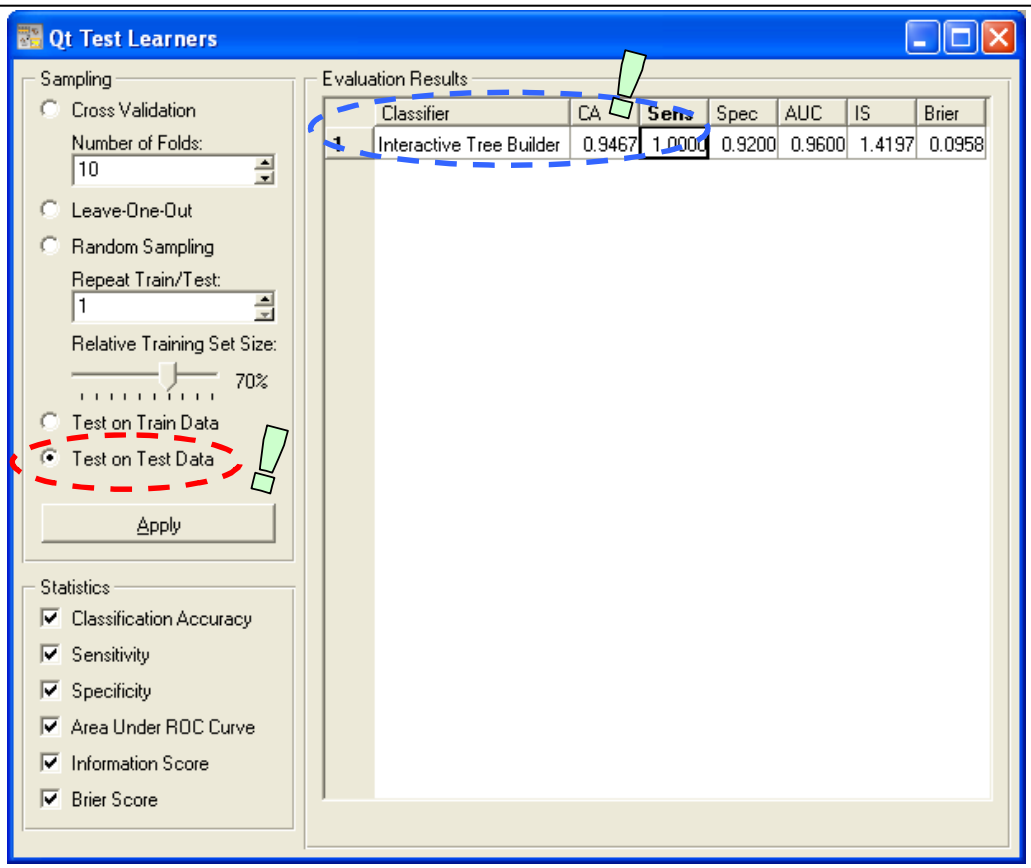
Connectons dans un premier temps le classifieur. Pour ce faire, nous relions l'arbre à TEST LEARNERS. De nouveau la boîte permettant de spécifier la nature de la liaison apparaît. Cette fois-ci la connexion est de type LEARNER.



Dans un deuxième temps, nous devons spécifier les données en test. Le plus simple est disposer un nouveau composant SELECT DATA dans le diagramme, de sélectionner les données STATUS = TEST dans la condition de restriction, puis de le relier à TEST LEARNERS en veillant à ce que nous transmettions bien les données en test.



La dernière étape consiste alors à activer le menu OPEN du composant TEST LEARNERS, puis veiller à ce que l'option TEST ON TEST DATA soit bien cochée. Les résultats sont directement affichés. Nous retrouvons la précision déjà annoncée dans SIPINA, ce qui est normal dans la mesure où nous avons construit le même arbre. Le contraire eut été inquiétant.



Conclusion

Ces deux logiciels donnent aux praticiens le pouvoir d'intervenir dans un processus d'exploration des connaissances lors de la construction d'un arbre de décision. Ils leur manquent néanmoins certaines fonctionnalités qui les mettraient au même niveau que les outils commerciaux.

Il n'est pas possible avec SIPINA d'ajuster les propriétés d'une segmentation : modifier la valeur seuil lorsqu'on utilise un descripteur continu, regrouper les modalités lorsque l'on segmente avec une variable discrète.

ORANGE ne classe pas les variables candidates selon leur pertinence sur un nœud. Si nous pouvons modifier le seuil de découpage des variables continues, il aurait été souhaitable de se voir proposer la valeur optimale et non pas la moyenne locale, l'utilisateur pourra ensuite moduler la valeur selon ses connaissances. Tout comme avec SIPINA, rien n'a été prévu pour que l'on puisse moduler les regroupements des modalités des variables candidates discrètes.

Il reste que ce sont des outils puissants, ce sont les seuls de ce type (à ma connaissance) que l'on peut trouver gratuitement sur le web. Il est important de le souligner.