Objectif

Confronter plusieurs algorithmes d'apprentissage sur des échantillons d'apprentissage et de test identiques. Utilisation comparée de TANAGRA, ORANGE et WEKA.

Très souvent, pour résoudre un problème d'apprentissage supervisé, nous sommes emmenés à choisir entre plusieurs algorithmes d'apprentissage. Parmi les critères d'évaluation figurent la précision des méthodes sur un échantillon test. Pour une expérimentation rigoureuse, il est fortement conseillé d'utiliser les mêmes échantillons d'apprentissage et de test, ainsi les méthodes seront directement comparables deux à deux, il est même possible de caractériser leur manière de classer, cela peut être intéressant lorsque les coûts de mauvais classement ne sont pas symétriques.

Dans ce didacticiel, nous montrons le détail des opérations sur les logiciels ORANGE, WEKA et TANAGRA. Nous verrons qu'ils procèdent avec une philosophie très différente, notamment dans la préparation des fichiers, mais au final nous obtenons des résultats similaires.

Nous avons choisi de mettre en compétition trois méthodes d'apprentissage pour illustrer notre propos : un SVM linéaire (Support Vector Machine), la régression logistique et un arbre de décision.

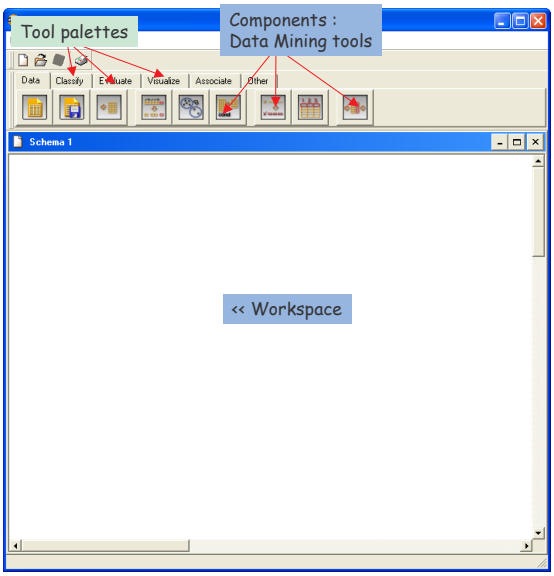
Fichier

Nous utilisons le fichier BREAST (UCI IRVINE). Il comporte un attribut classe binaire (tumeur bénigne ou maligne), 9 descripteurs, tous continus, et 699 exemples.

Nous avons sélectionné 499 observations pour l'apprentissage, 200 pour le test. **Nous utilisons la même subdivision pour nos trois logiciels.**

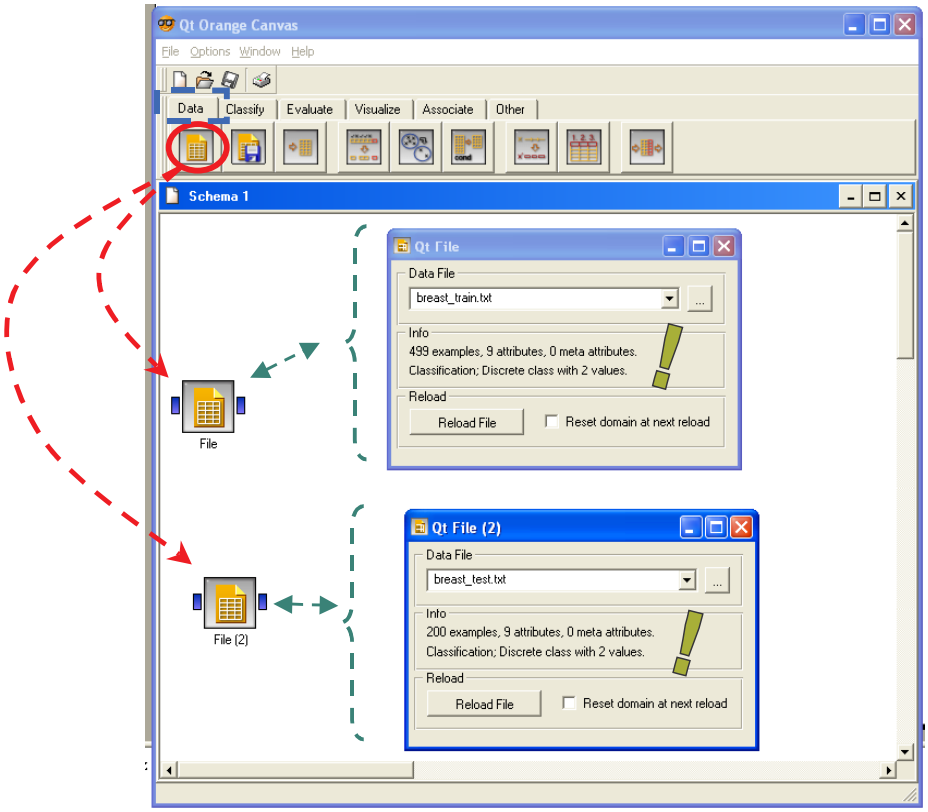
Comparer les méthodes avec ORANGE

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.



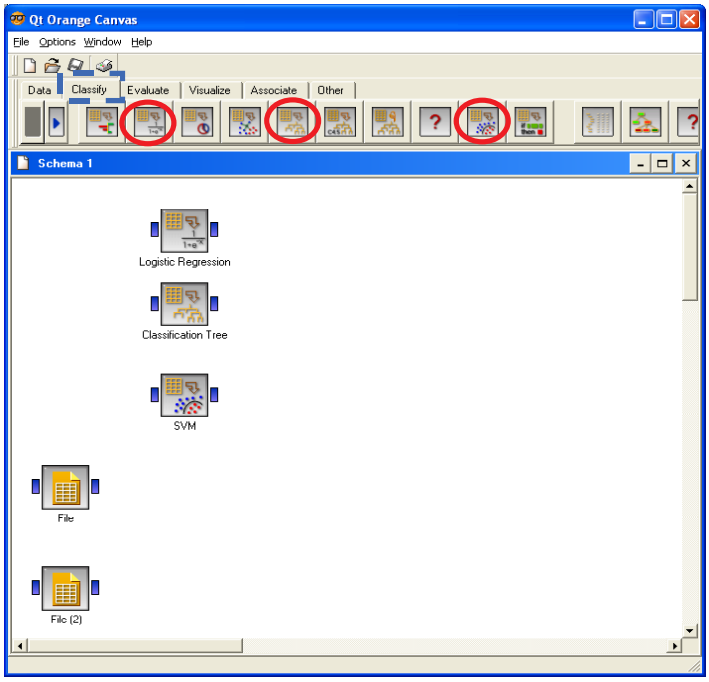
Préparer et charger les données

La meilleure manière de procéder avec ORANGE est de scinder les données en 2 distincts : BREAST_TRAIN.TXT pour l'apprentissage, BREAST_TEST.TXT pour le test. Nous plaçons donc deux fois le composant d'accès aux données dans l'espace de travail et nous les configurons en activant le menu OPEN.



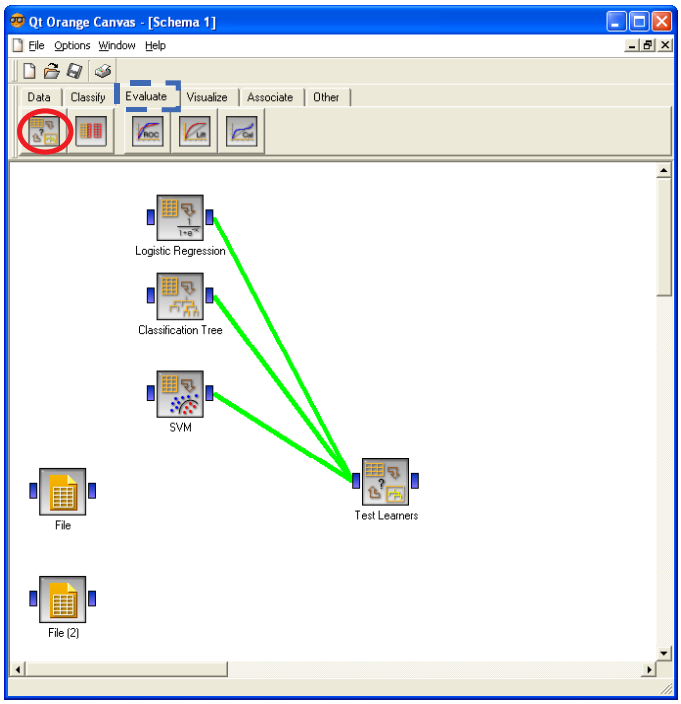
Placer les méthodes d'apprentissage

Nous voulons tester trois méthodes d'apprentissage, il nous faut donc les placer dans le diagramme de traitements. Ces composants sont situés dans l'onglet CLASSIFY.



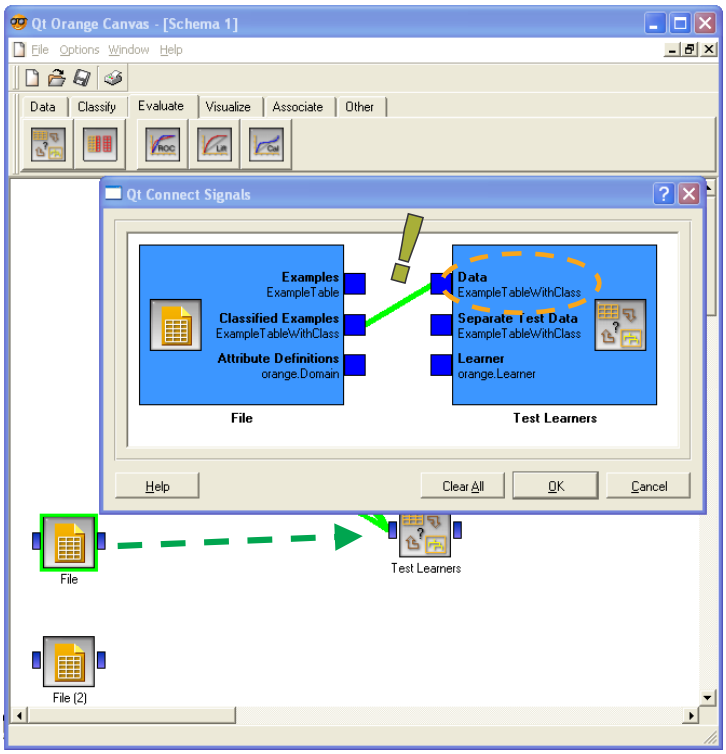
Placer le composant d'évaluation

Un seul composant d'évaluation permet de comparer directement les trois méthodes. Il s'agit de TEST LEARNERS situé dans l'onglet EVALUATE. Nous le plaçons dans le diagramme et nous le relions aux trois méthodes d'apprentissage.

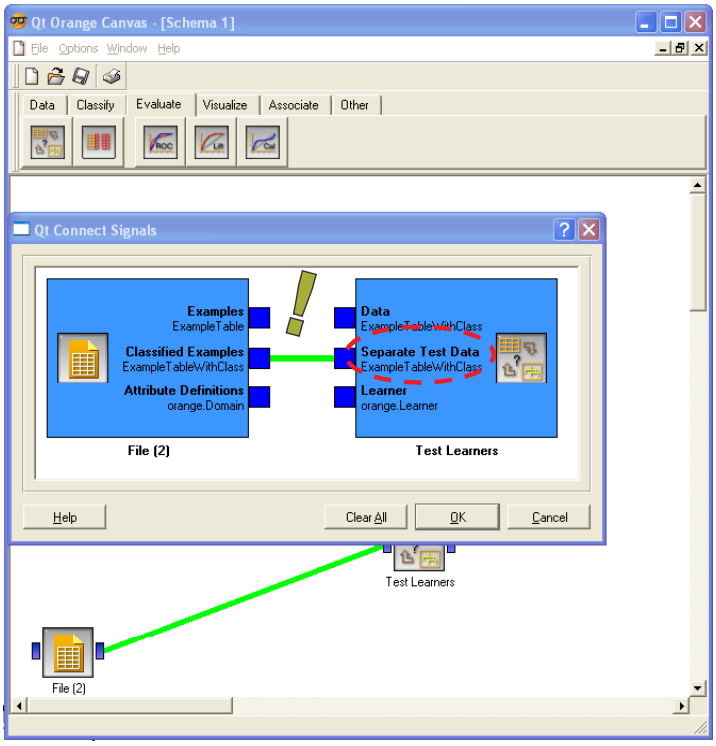


Nous devons maintenant lui spécifier quelles sont les données à utiliser pour l'apprentissage. Nous relions donc la première source de données [FILE] au composant TEST LEARNERS.

Une boîte de dialogue surgit au moment où nous établissons la connexion, elle est primordiale car elle nous permet de vérifier que nous transmettons bien les données d'apprentissage (DATA). L'apprentissage est automatiquement exécuté.

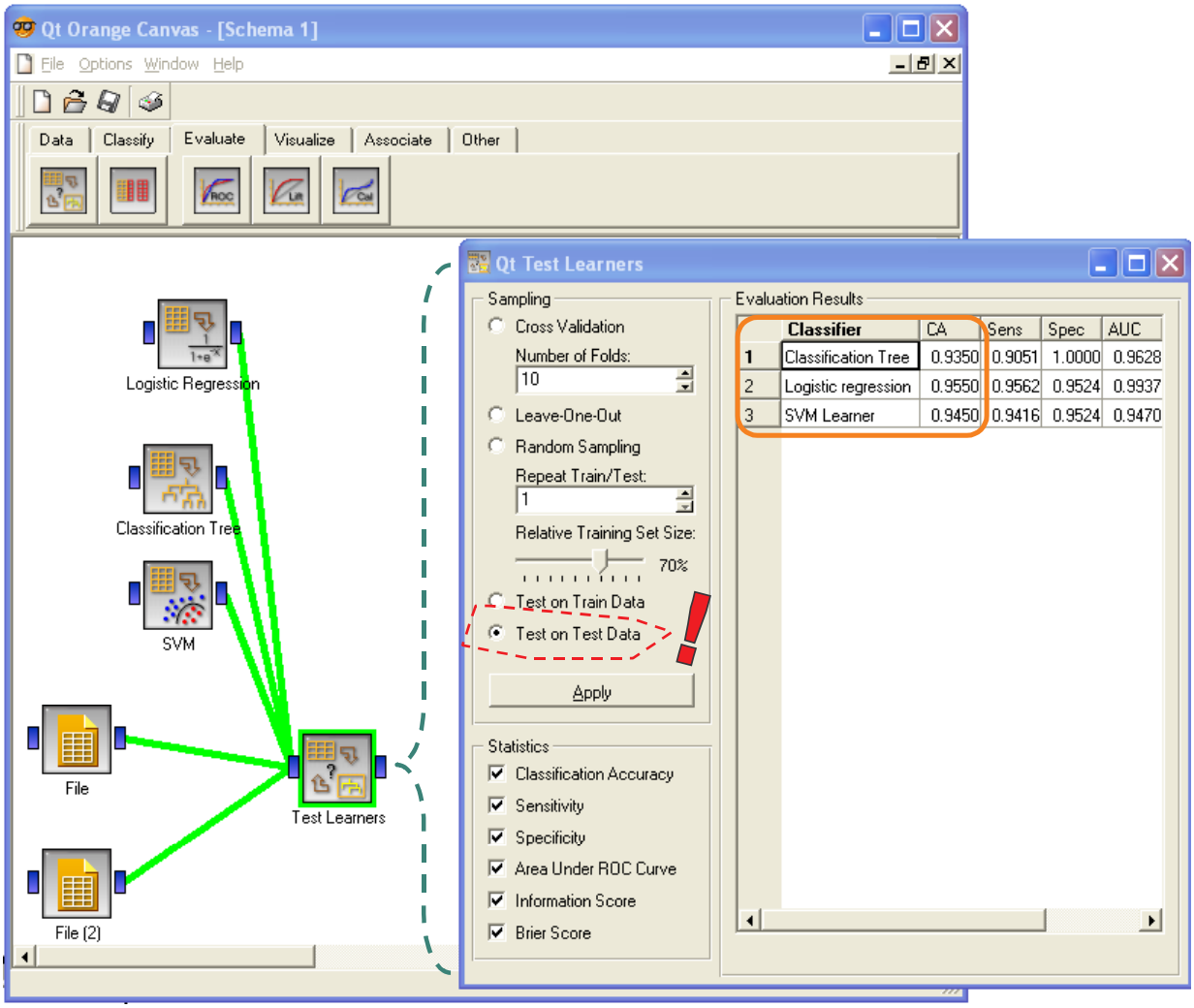


L'étape suivante consiste à relier la deuxième source de données [FILE (2)] à TEST LEARNERS. Comme il n'y a pas d'ambiguïté ici, la connexion est automatiquement établie, ORANGE considère que la deuxième source constitue l'ensemble test (SEPARATE TEST DATA). Nous pouvons modifier la nature de la connexion en double-cliquant sur le lien.



Voir les résultats

Pour voir les résultats, il faut activer le menu OPEN du composant de comparaison. Nous obtenons l'affichage suivant.



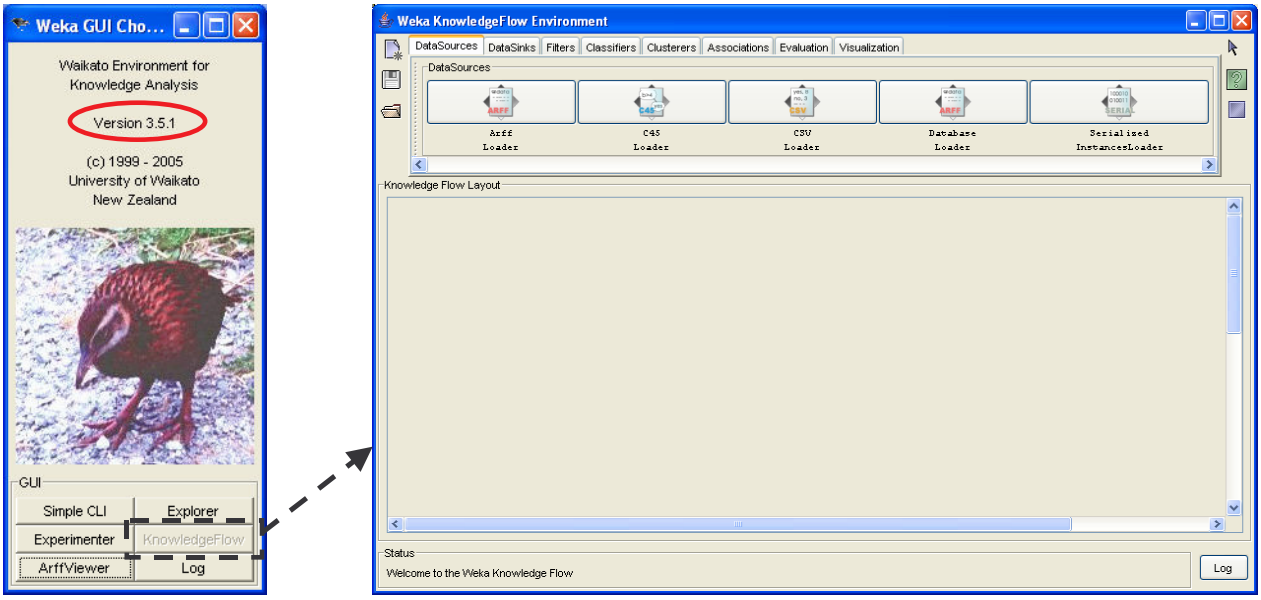
L'évaluation doit être réalisée sur l'ensemble test, nous vérifions que l'option est bien cochée. Plusieurs indicateurs sont disponibles dont le taux de bon classement qui nous intéresse :

- Arbre de décision : 93.5% (taux d'erreur 6.5%) ;
- Régression logistique : 95.5%
- SVM Linéaire[1] : 94.5%.

[1] Vérifiez que la méthode est correctement paramétrée, le noyau doit être linéaire (KERNEL – LINEAR).

Comparer les méthodes avec WEKA

Au lancement de WEKA, un panneau permet de choisir le mode d'exécution du logiciel. Nous choisissons le mode **KNOWLEDGE FLOW**. Nous avons utilisé la version **3.5.1** dans ce didacticiel. Nous parvenons alors dans l'espace de travail dans lequel nous allons définir nos traitements. Dans la partie haute, nous trouvons les icônes organisées dans des palettes, ils représentent les opérateurs de traitement.



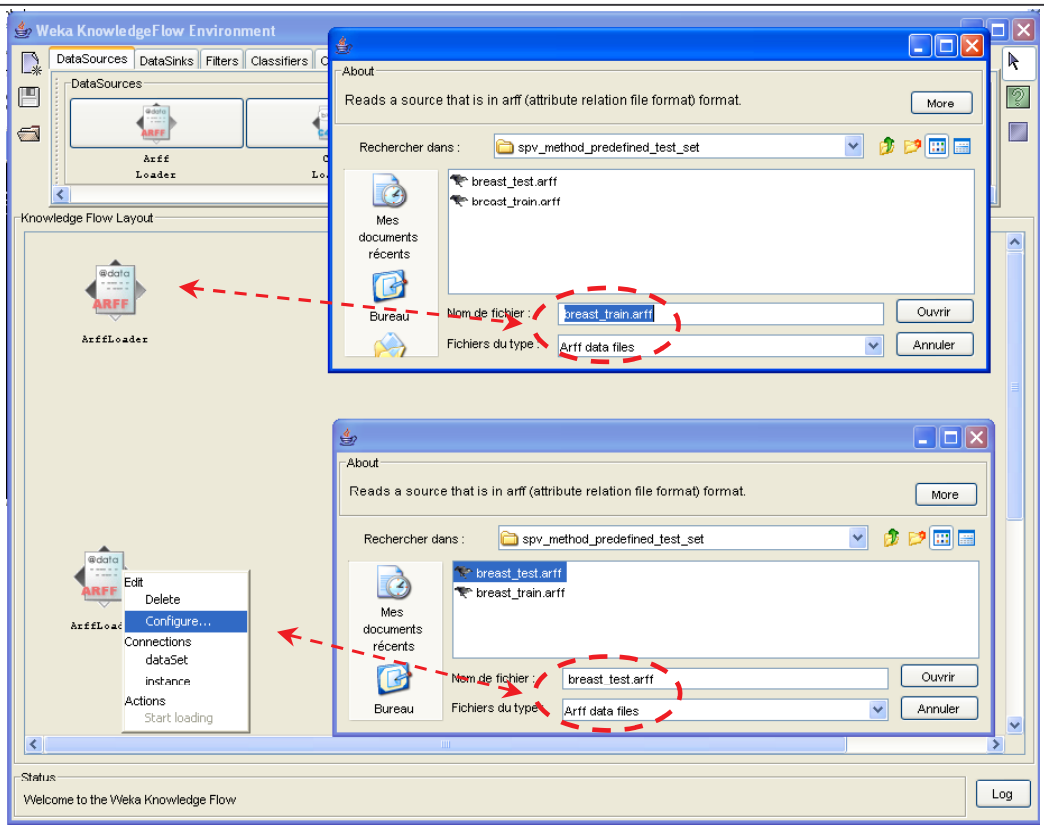
Préparer et charger les données

Dans WEKA, les données doivent être également dans deux fichiers séparés. Il est fondamental que les parties « description des attributs » des fichiers ARFF soient rigoureusement identiques, sinon la procédure échoue.

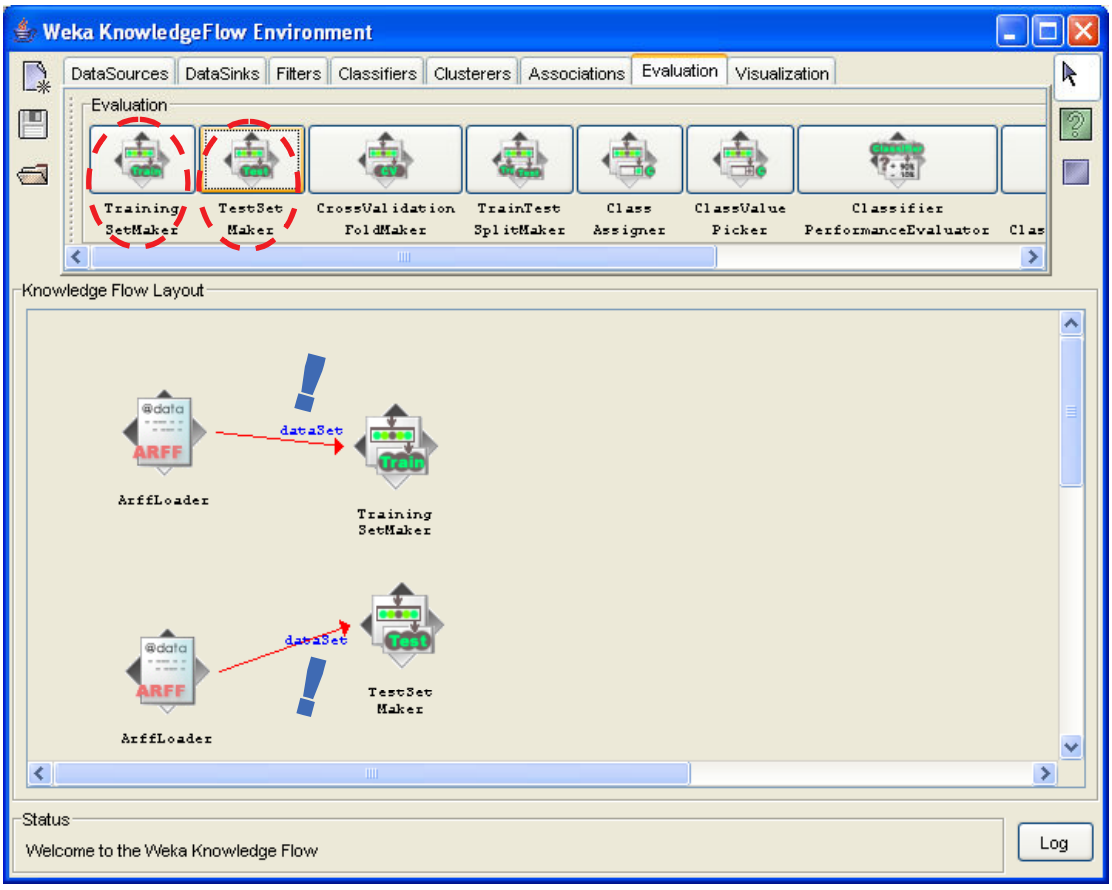
```
@relation breast_train.arff
@attribute clump REAL
@attribute ucellsize REAL
@attribute ucellshape REAL
@attribute mgadhesion REAL
@attribute sepics REAL
@attribute bnuclei REAL
@attribute bchromatin REAL
@attribute normnucl REAL
@attribute mitoses REAL
@attribute class {begnin,malignant}
```

```
@relation breast_test.arff
@attribute clump REAL
@attribute ucellsize REAL
@attribute ucellshape REAL
@attribute mgadhesion REAL
@attribute sepics REAL
@attribute bnuclei REAL
@attribute bchromatin REAL
@attribute normnucl REAL
@attribute mitoses REAL
@attribute class {begnin,malignant}
```

Nous plaçons donc deux fois les composants ARFF LOADER puis nous les configurons de manière à les brancher sur les fichiers adéquats.

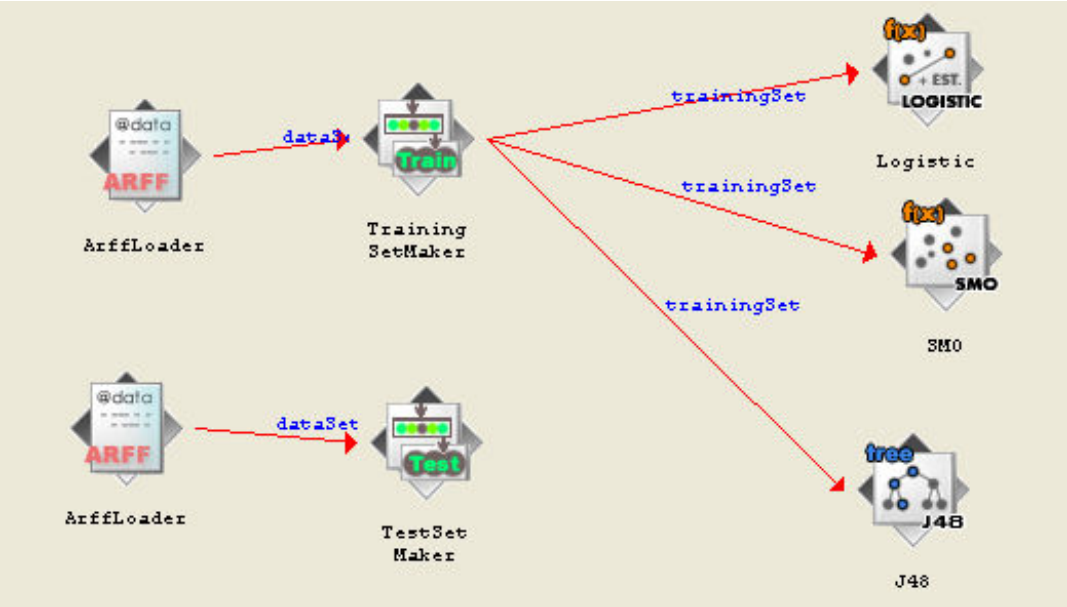


Nous devons maintenant préciser le statut de ces données pour la suite du diagramme. Nous insérons pour cela 2 composants : TRAINING SET MAKER et TEST SET MAKER (onglet EVALUATE). Nous effectuons les connexions adéquates.

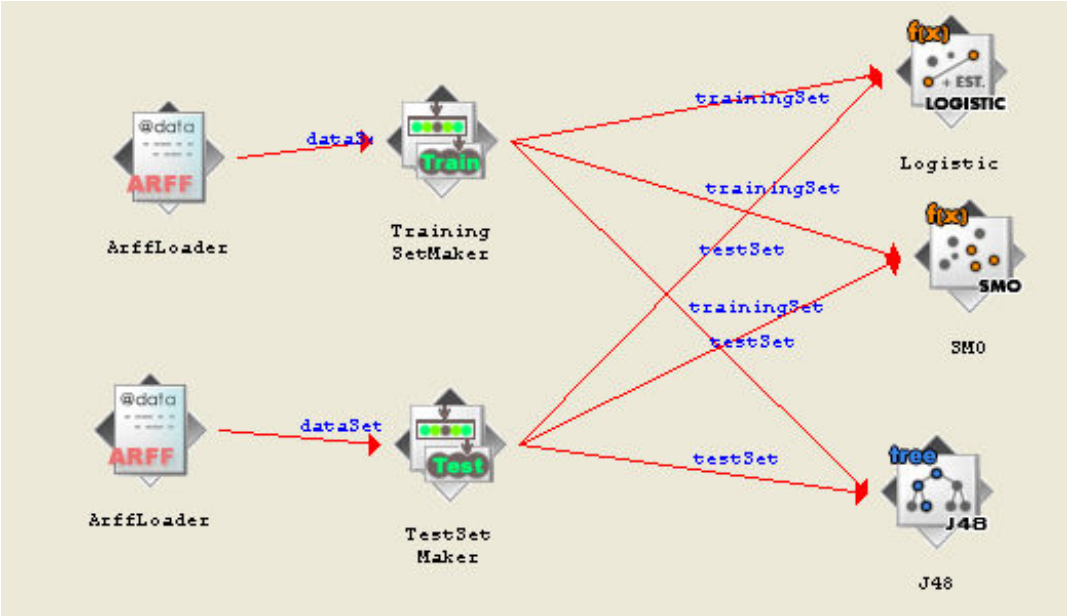


Méthodes d'apprentissage

Nous plaçons ensuite les trois méthodes d'apprentissage à évaluer situés dans l'onglet CLASSIFIERS. Attention, concernant le SVM (SMO), il faut bien vérifier que nous utilisons le noyau linéaire (exposant = 1, pas de noyau RBF). Nous devons alors connecter le TRAINING SET MAKER aux trois composants d'apprentissage,

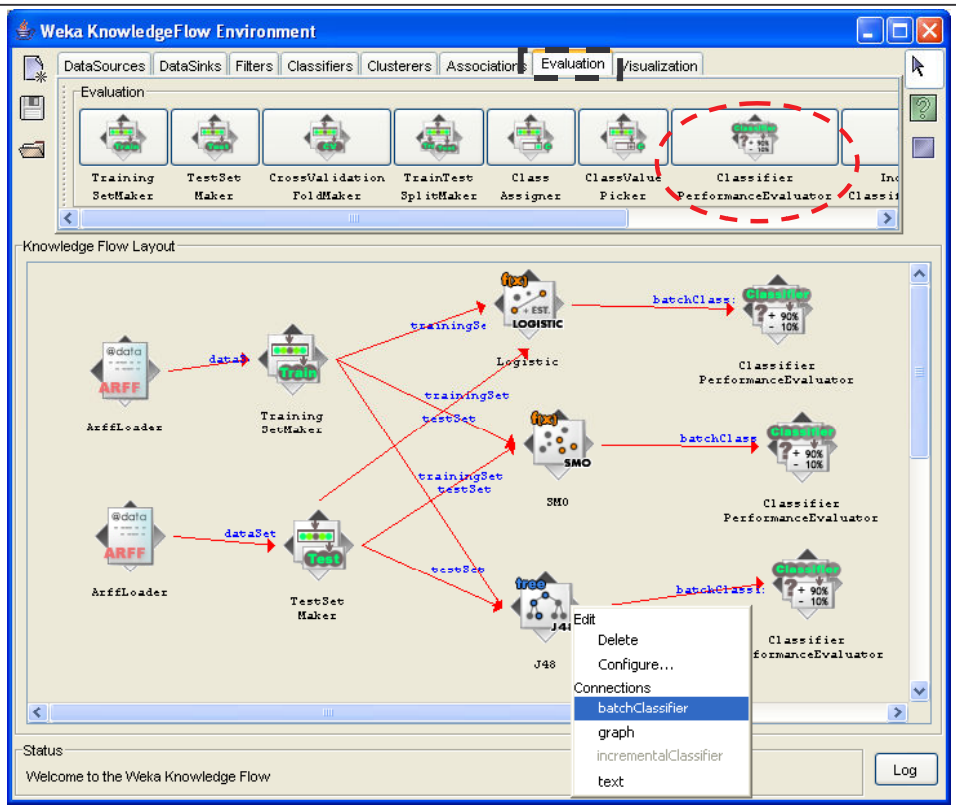


Nous procédons de la même manière en ce qui concerne le composant TEST SET MAKER.



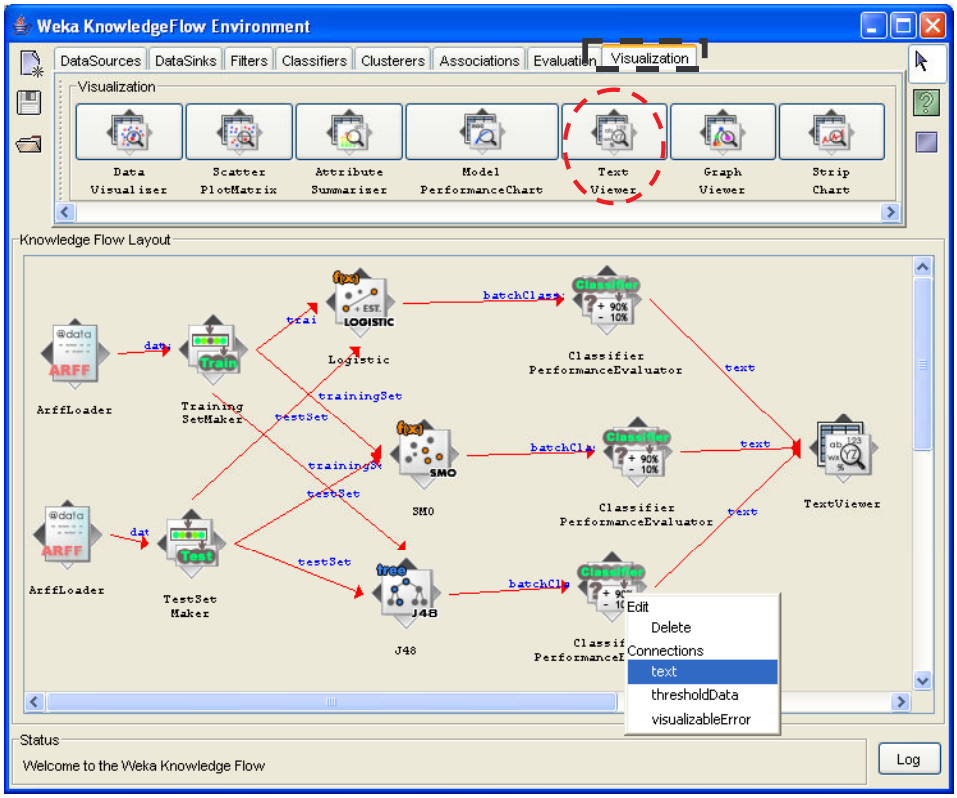
Composant d'évaluation

Nous associons un composant d'évaluation à chaque méthode d'apprentissage. C'est le rôle du composant CLASSIFIER PERFORMANCE EVALUATOR (onglet EVALUATION). Nous utilisons la connexion de type BATCH CLASSIFIER des méthodes d'apprentissage.



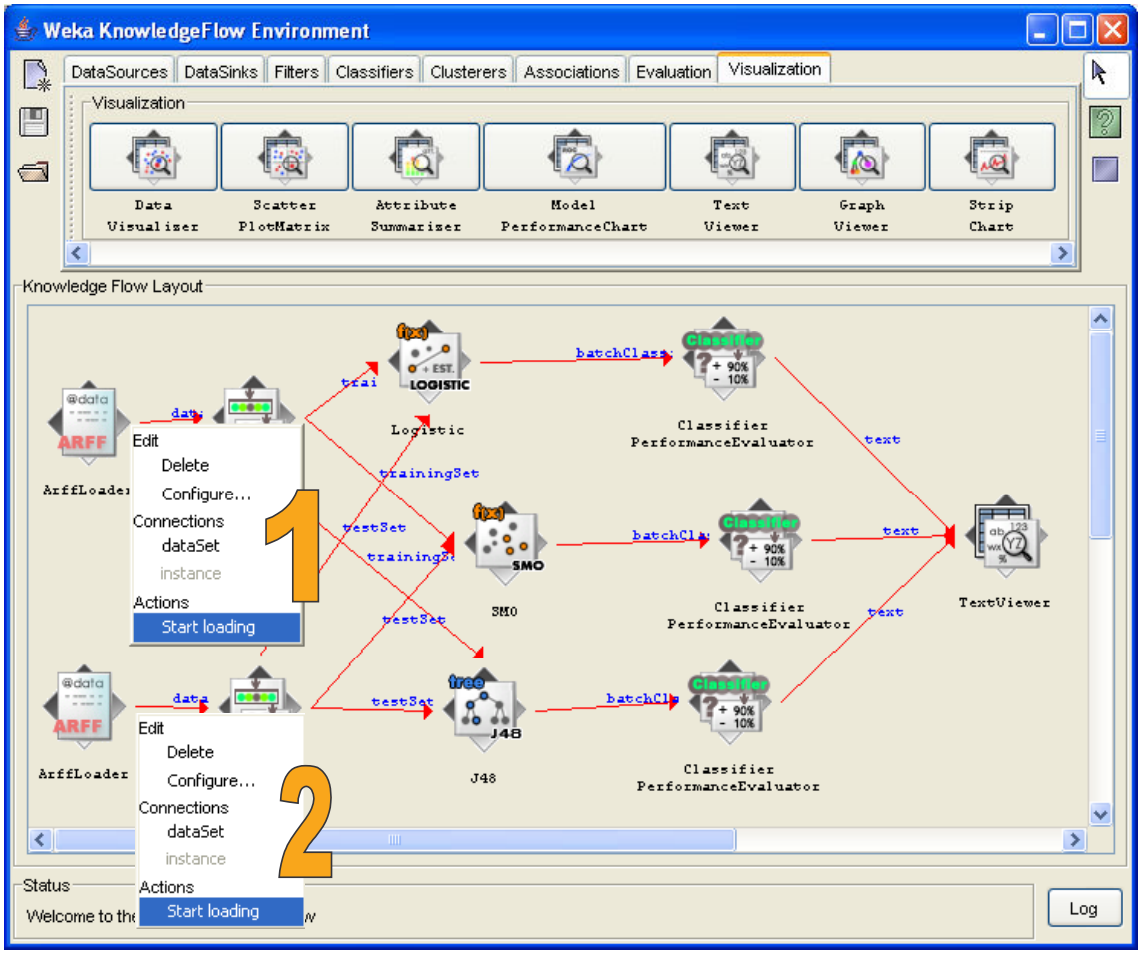
Composant de visualisation des résultats

Il nous reste alors à visualiser les résultats en utilisant le composant TEXT VIEWER (VISUALIZATION). La principale astuce ici est qu'il est possible d'afficher les résultats dans un seul outil, ce qui facilite les comparaisons. Nous utilisons la connexion de type TEXT.

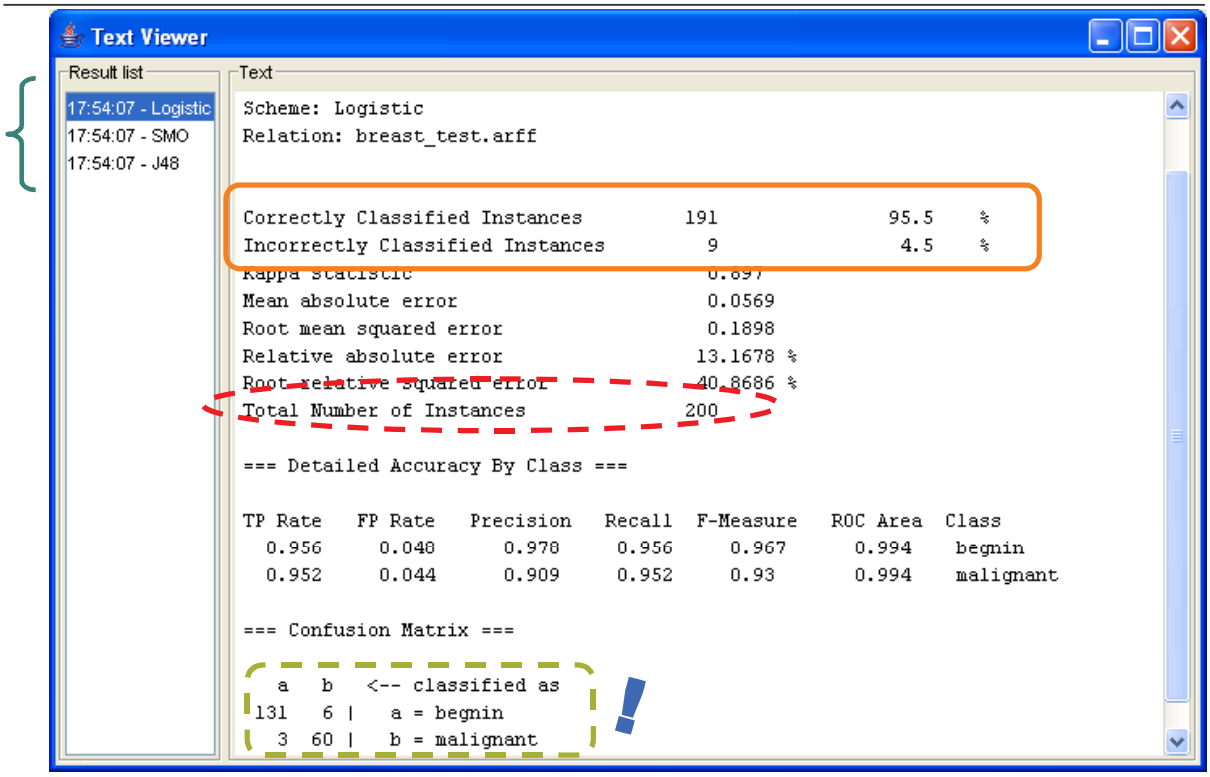


Exécution du diagramme

L'exécution du diagramme se fait en deux temps, [1] d'abord en activant le menu START LOADING du composant de données branché sur les données d'apprentissage, ce qui déclenche la construction des modèles de prédiction ; [2] puis en activant le menu START LOADING du composant de données branché sur les données de test, ce qui déclenche l'évaluation des performances des modèles.



En sélectionnant le menu SHOW RESULTS du composant TEXT VIEWER, nous pouvons recenser les performances de chaque méthode sur l'ensemble test. Nous avons le détail des calculs avec la matrice de confusion.



Comme précédemment, nous nous intéressons essentiellement au taux de bon classement dans ce didacticiel, nous retiendrons donc pour chaque méthode :

- Arbre de décision : 93.5% (taux d'erreur 6.5%) ;
- Régression logistique : 95.5%
- SVM Linéaire : 95.5%.

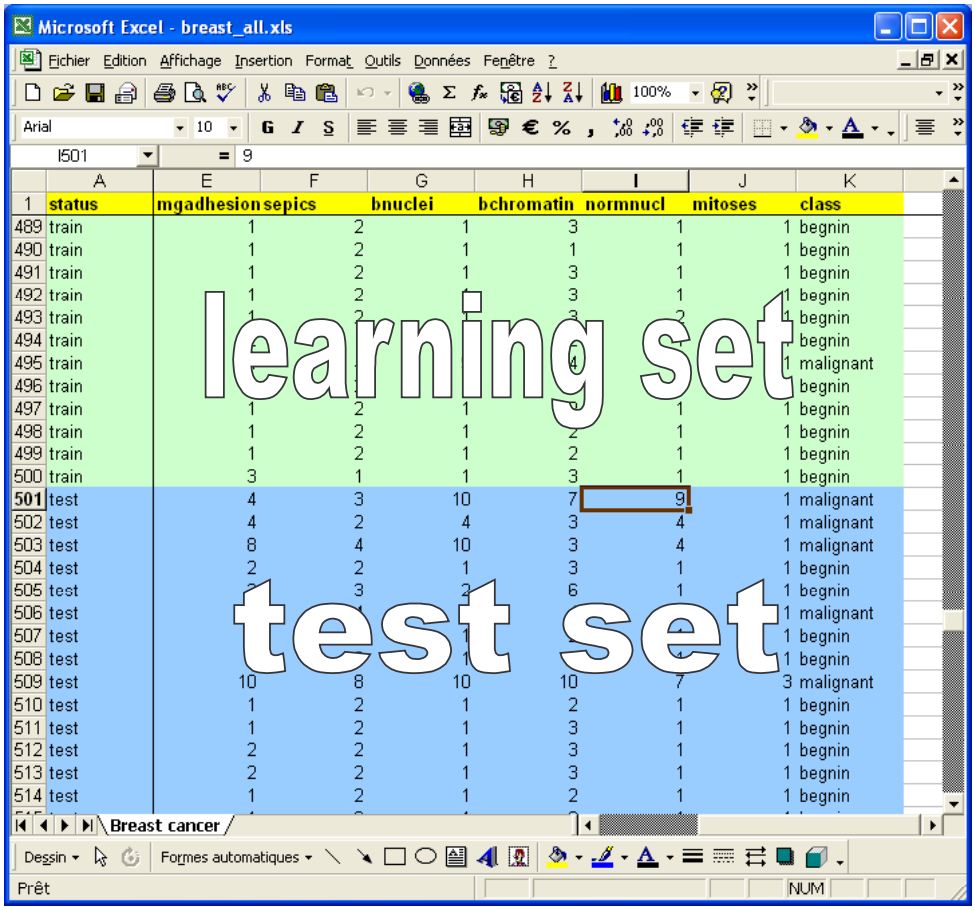
Notons que les SVM et la régression logistique présentent globalement les mêmes performances en termes de taux de bons classements ; la matrice de confusion, et donc la structure de l'erreur, est différente en revanche.

Comparer les méthodes avec TANAGRA

Par rapport aux deux autres logiciels de ce didacticiel, TANAGRA utilise un arbre pour représenter les traitements. Cela simplifie sa structure, mais induit une contrainte forte, il n'est pas possible de spécifier deux sources de données. Il est dès lors nécessaire de préparer différemment les données.

Préparer les données

Nous travaillons sur le fichier BREAST_ALL.XLS[2]. Toutes les observations ont été réunies dans un seul et même fichier, nous avons ajouté une nouvelle colonne (STATUS) qui sert à distinguer les observations dévolues à l'apprentissage de ceux destinées au test. Le fichier XLS se présente donc de la manière suivante.



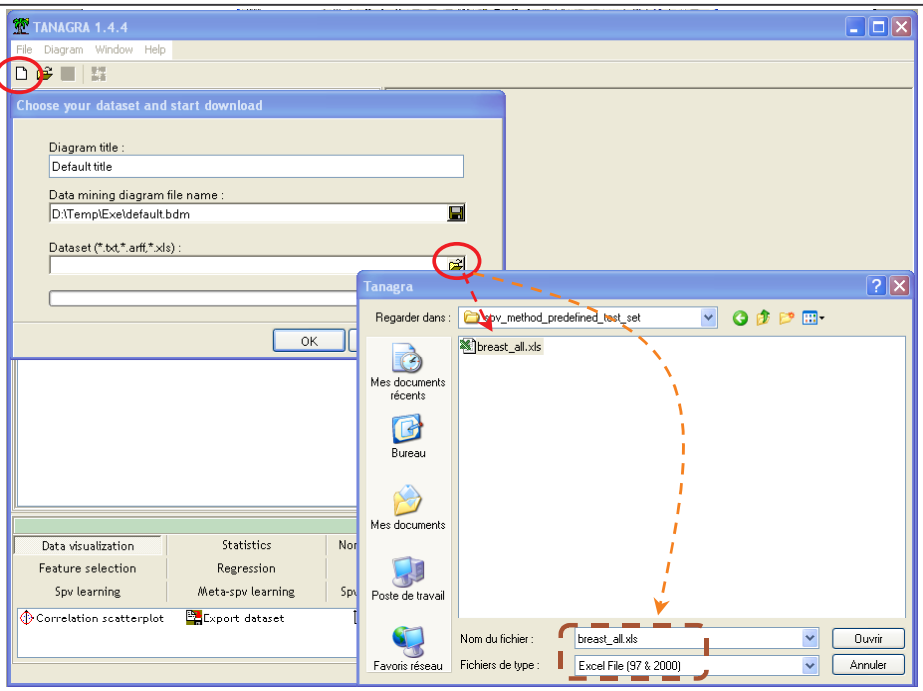
	status	mgadhesion	sepics	bnuclei	bchromatin	normnucl	mitoses	class
489	train	1	2	1	3	1	1	begnin
490	train	1	2	1	1	1	1	begnin
491	train	1	2	1	3	1	1	begnin
492	train	1	2	1	3	1	1	begnin
493	train	1	2	1	3	2	1	begnin
494	train	1	2	1	1	1	1	begnin
495	train	1	2	1	4	1	1	malignant
496	train	1	2	1	2	1	1	begnin
497	train	1	2	1	2	1	1	begnin
498	train	1	2	1	2	1	1	begnin
499	train	1	2	1	2	1	1	begnin
500	train	3	1	1	3	1	1	begnin
501	test	4	3	10	7	9	1	malignant
502	test	4	2	4	3	4	1	malignant
503	test	8	4	10	3	4	1	malignant
504	test	2	2	1	3	1	1	begnin
505	test	1	3	2	6	1	1	begnin
506	test	4	1	10	4	1	1	malignant
507	test	1	1	1	2	1	1	begnin
508	test	1	1	1	2	1	1	begnin
509	test	10	8	10	10	7	3	malignant
510	test	1	2	1	2	1	1	begnin
511	test	1	2	1	3	1	1	begnin
512	test	2	2	1	3	1	1	begnin
513	test	2	2	1	3	1	1	begnin
514	test	1	2	1	2	1	1	begnin

Importer les données

La préparation étant réalisée, il faut fermer EXCEL[3] et lancer TANAGRA. La première étape consiste à créer un nouveau diagramme et importer le fichier BREAST_ALL.XLS.

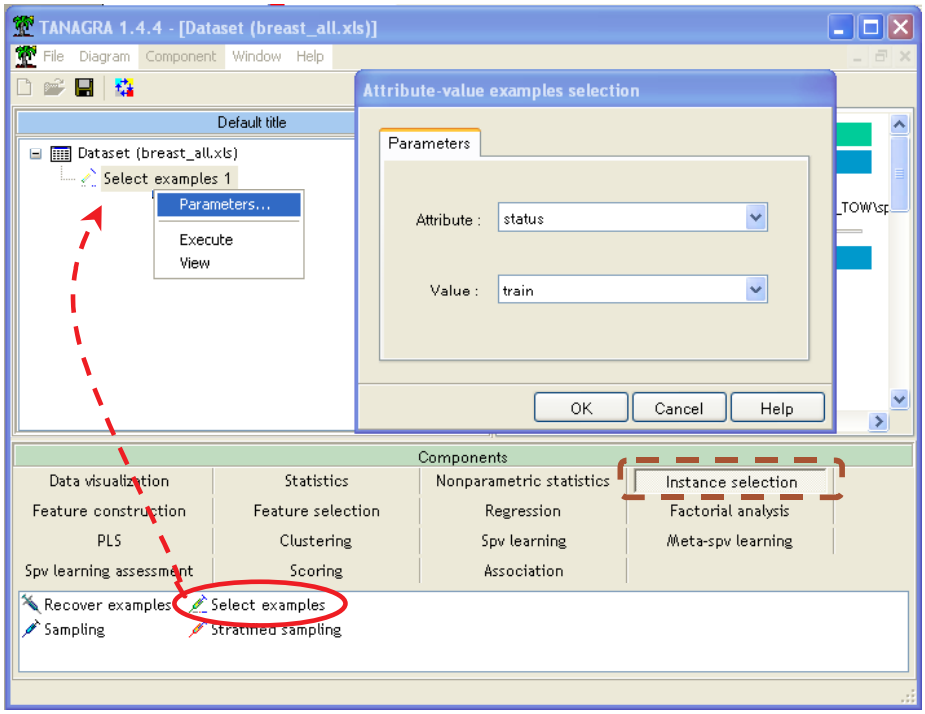
[2] Tanagra peut lire directement le format XLS, les données doivent être situées dans la première feuille du classeur, elles doivent être alignées en haut et à gauche (en A1).

[3] EXCEL verrouille les fichiers qu'il est en train d'éditer, il est donc très important que le classeur XLS soit fermé avant que l'on tente de l'ouvrir dans TANAGRA.



Subdivision Apprentissage – Test

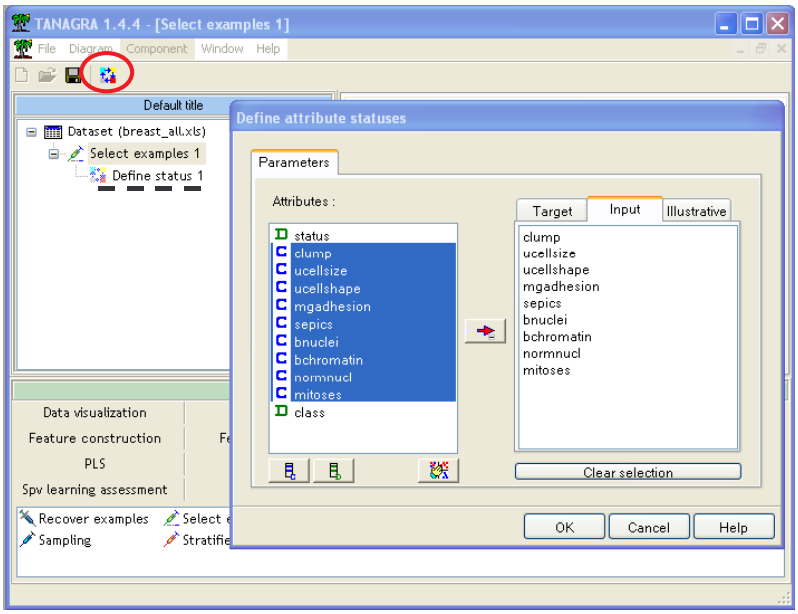
Nous utilisons le composant SELECT EXAMPLES (INSTANCE SELECTION) pour subdiviser les données dans TANAGRA. Les individus actifs correspondent à ceux qui ont la modalité TRAIN pour la variable STATUS, nous le précisons en activant le menu PARAMETERS du composant.



En exécutant le composant (menu VIEW), nous constatons que 499 individus maintenant sont sélectionnés pour les calculs.

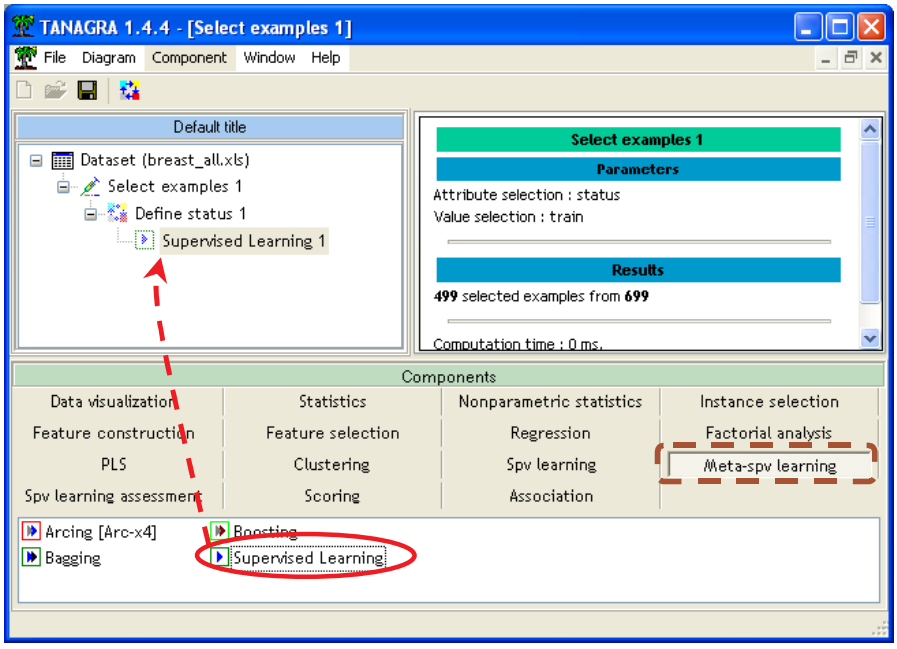
Définition des variables

Nous ajoutons le composant DEFINE STATUS pour sélectionner la variable TARGET (CLASS) et les variables input (toutes les variables continues, il ne faut surtout pas sélectionner la variable STATUS).

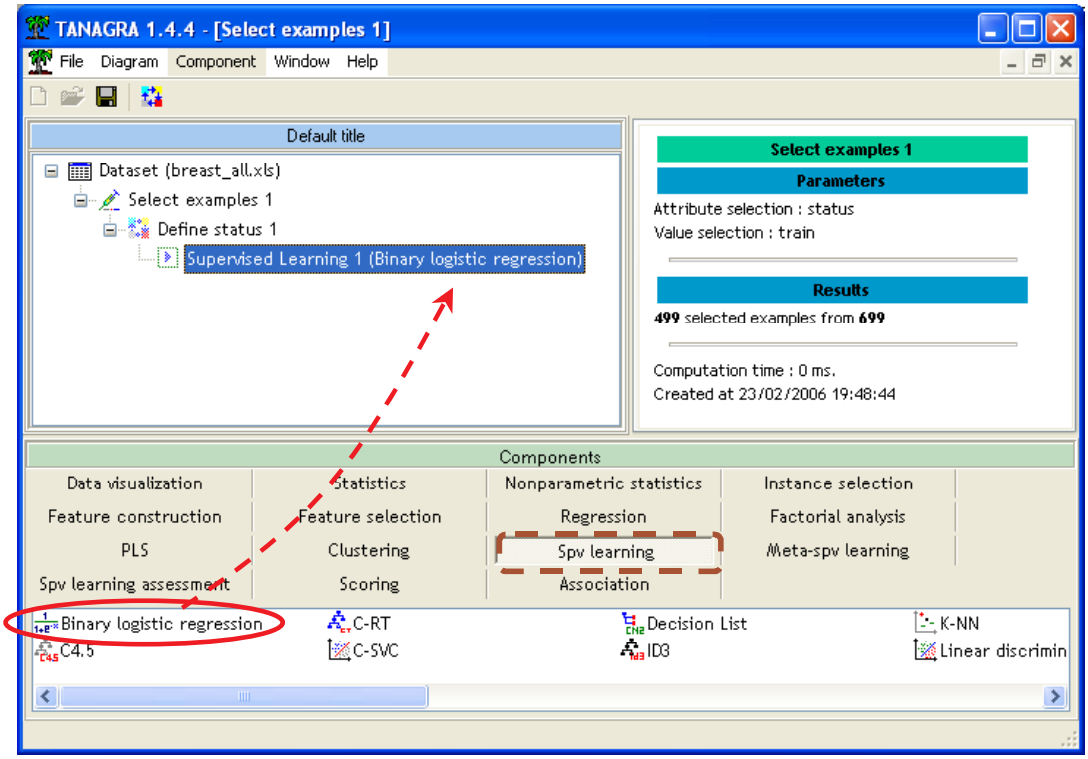


Méthodes d'apprentissage

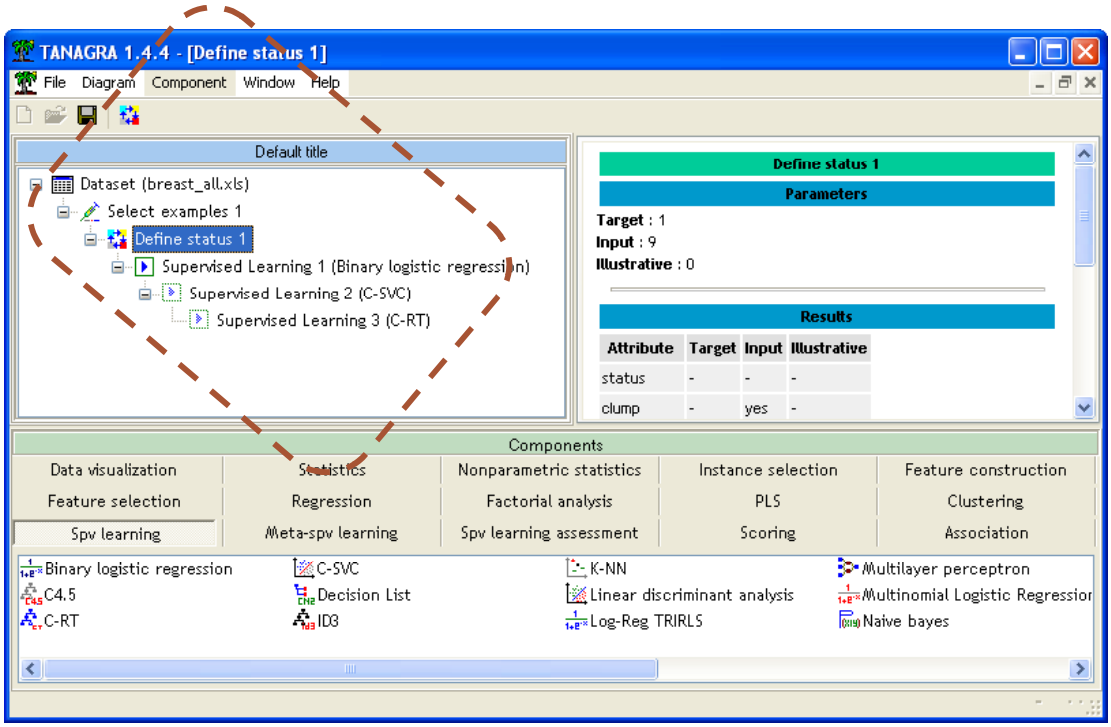
Nous devons placer les trois méthodes d'apprentissage que nous voulons évaluer. Nous détaillons cette opération pour la régression logistique. Insérer une méthode d'apprentissage se fait toujours en deux temps dans TANAGRA, tout d'abord placer l'opérateur meta-apprentissage qui permet de définir la stratégie d'utilisation des méthodes. Nous utiliserons un apprentissage simple dans notre cas. Nous insérons le composant SUPERVISED LEARNING (onglet META SPV-LEARNING) dans le diagramme.



Puis dans un deuxième temps, nous intégrons la méthode d'apprentissage proprement dite dans l'opérateur précédent. Nous avons choisi le composant BINARY LOGISTIC REGRESSION (onglet SPV LEARNING), il est un peu lent par rapport aux autres approches disponibles dans TANAGRA, il a l'avantage de fournir une série de statistiques supplémentaires.



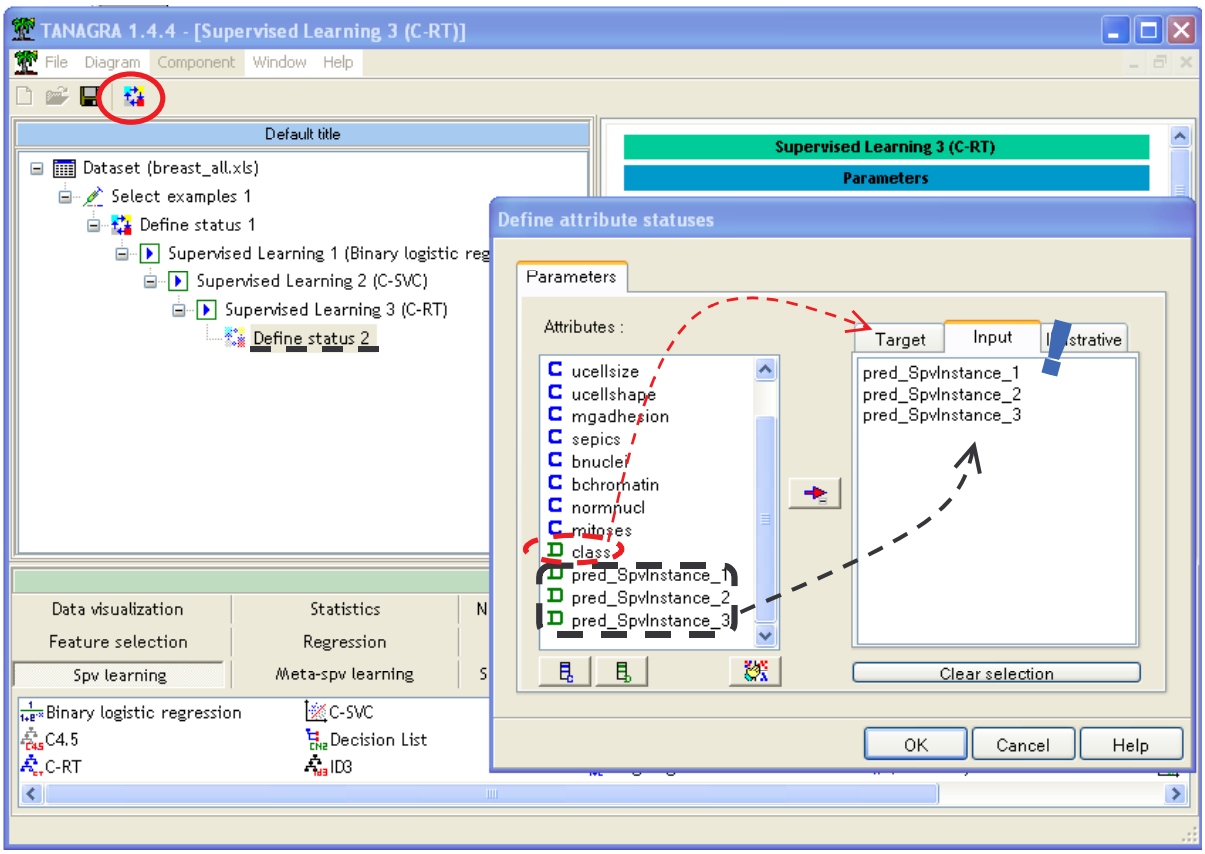
Nous procédons de la même manière pour insérer les SVM (C-SVC) et l'arbre de décision (C-RT). Nous obtenons le diagramme suivant.



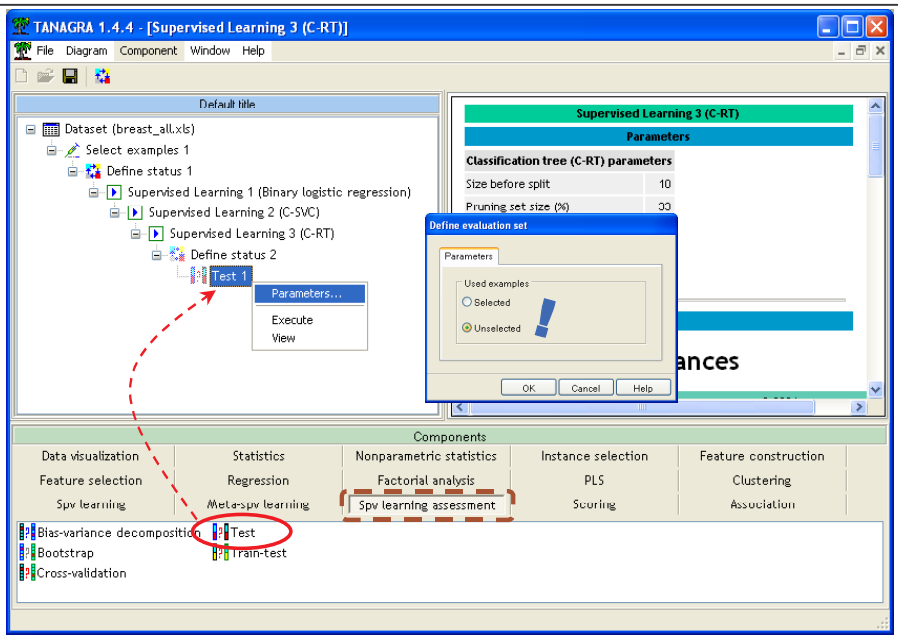
Pour exécuter toute la chaîne de traitements, il suffit d'activer le menu VIEW du dernier composant du diagramme.

Comparer les performances sur l'ensemble test

Pour comparer les performances, nous devons tout d'abord insérer de nouveau un composant DEFINE STATUS dans le diagramme en cliquant sur le raccourci dans la barre d'outil. Nous plaçons en TARGET la variable à prédire CLASS, puis nous définissons en INPUT les variables générées par les méthodes d'apprentissage. Ces variables correspondent aux projections effectuées par les modèles de prédiction sur la totalité de l'échantillon.



Il ne nous reste plus qu'à insérer et paramétrer le composant d'évaluation, il s'agit de TEST situé dans l'onglet SPV LEARNING ASSESMENT. Nous devons bien spécifier que l'évaluation est réalisée sur l'échantillon n'ayant pas servi à l'apprentissage, c.-à-d. les individus non-sélectionnés.



Nous cliquons sur le menu VIEW, nous obtenons les résultats suivants.

Test 1

Parameters

Evaluation set : **unselected** examples

Results

pred_SpvInstance_1

Error rate	0.0450

Values prediction

Value	Recall	1-Precision
begnin	0.9562	0.0224
malignant	0.9524	0.0909

Confusion matrix

	begnin	malignant	Sum
begnin	131	6	137
malignant	3	60	63
Sum	134	66	200

pred_SpvInstance_2

Error rate	0.0550

Values prediction

Value	Recall	1-Precision
begnin	0.9416	0.0227
malignant	0.9524	0.1176

Confusion matrix

	begnin	malignant	Sum
begnin	129	8	137
malignant	3	60	63
Sum	132	68	200

pred_SpvInstance_3

Error rate	0.0750

Values prediction

Value	Recall	1-Precision
begnin	0.9051	0.0159
malignant	0.9683	0.1757

Confusion matrix

	begnin	malignant	Sum
begnin	124	13	137
malignant	2	61	63
Sum	126	74	200

Dans le cas de TANAGRA, les taux de bons classements sont :

- Arbre de décision : 92.5% (taux d'erreur 7.5%) ;
- Régression logistique : 95.5%
- SVM Linéaire : 94.5%.

Conclusion

Dans ce didacticiel, nous avons montré qu'il était assez aisé, avec les logiciels ORANGE, WEKA et TANAGRA, de comparer les performances de méthodes supervisées en procédant à l'induction sur le même ensemble d'apprentissage, puis en les évaluant sur le même ensemble de test.

Un autre aspect que nous n'avons pas mis en œuvre dans ce didacticiel, mais qui pourrait s'avérer très instructif, serait de croiser les étiquetages des méthodes pour vérifier si elles classent de la même manière ou pas. Nous l'avons fait par ailleurs avec TANAGRA, nous avons constaté que le SVM linéaire et la régression logistique classent quasiment de la même manière (2 cas désaccords) ; elles se démarquent assez sensiblement de l'arbre de décision (12 cas de désaccords à chaque fois).

Enfin, il est normal que les résultats soient légèrement dissemblables d'un logiciel à l'autre. Les méthodes s'appuient sur des heuristiques, les résultats reposent en partie sur la stratégie mise en œuvre et… l'implémentation. Ce dernier point n'est pas négligeable, loin de là. Cela aurait été néanmoins inquiétant si l'on avait obtenu des taux de bons classements très différents.