Objectif

Montrer le fonctionnement des SVM (Support Vector Machine – Machines à vecteurs de support) dans TANAGRA.

Comparaison avec l'analyse discriminante linéaire, une méthode qui induit un séparateur linéaire dans l'espace de représentation.

L'implémentation est une traduction littérale en DELPHI du code source JAVA issu de la version 3-4 de WEKA (classe SMO.JAVA).

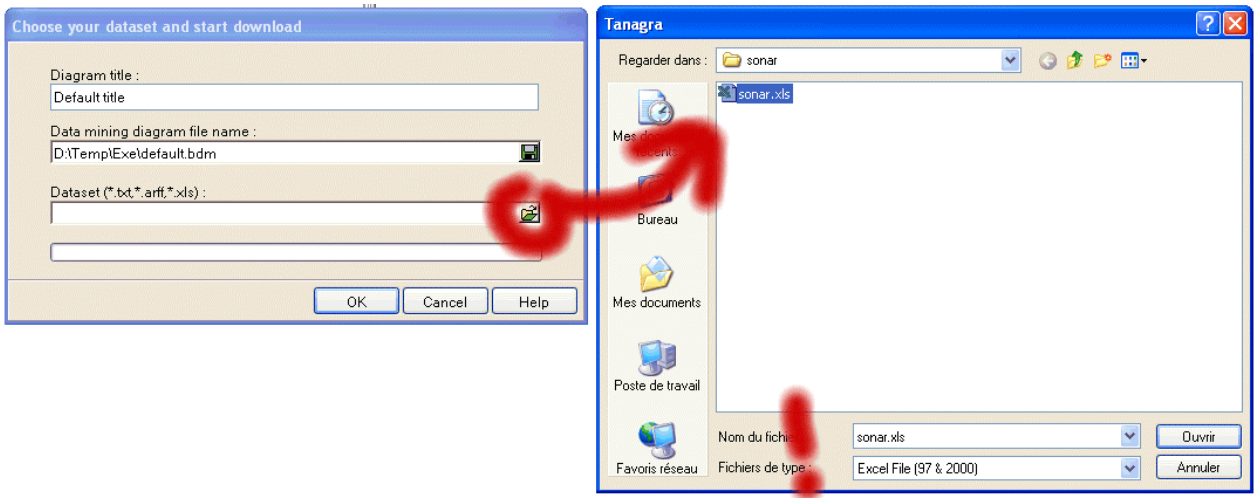
Fichier

Le fichier SONAR : prédire la nature d'un objet (roche ou mine) à partir des informations récupérées par un détecteur de mines. L'attribut classe possède deux modalités (le seul cas traité par notre implémentation des SVM pour l'instant), il y a 60 descripteurs, tous continus.

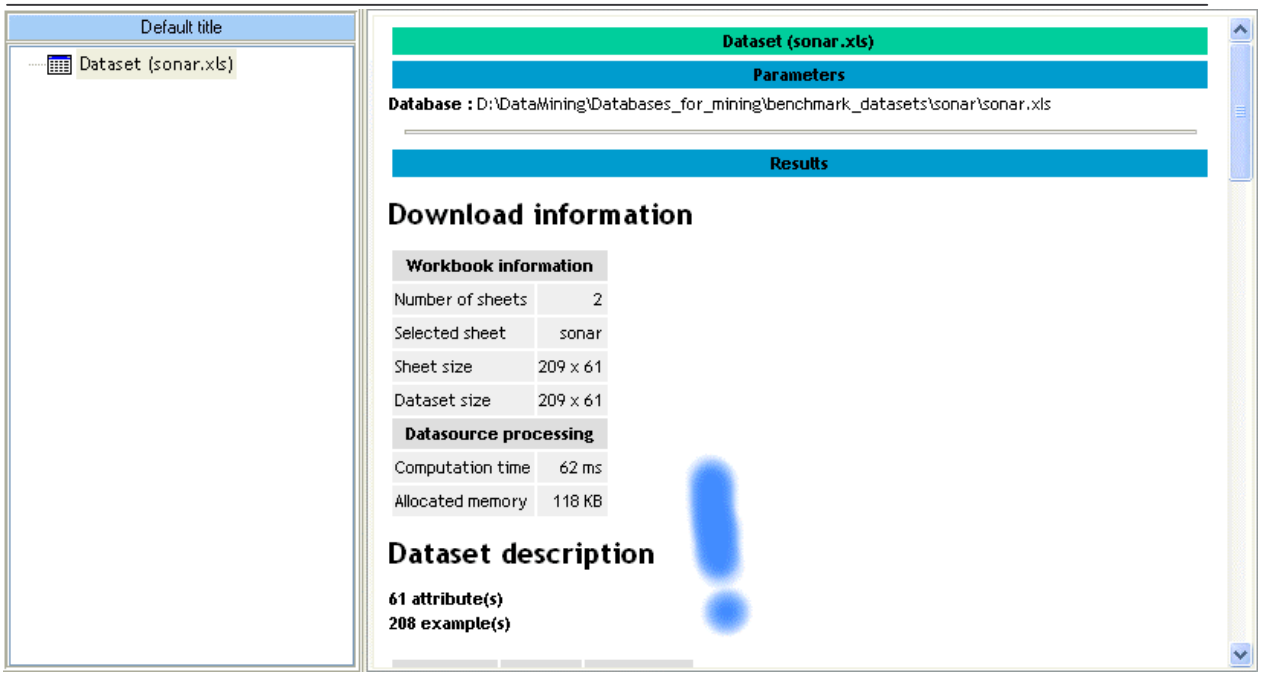
SVM

Charger le fichier de données

Importez le fichier SONAR.XLS dans TANAGRA en suivant la procédure habituelle « File / New ». Il faut sélectionner le type de fichier EXCEL dans la boîte de sélection.



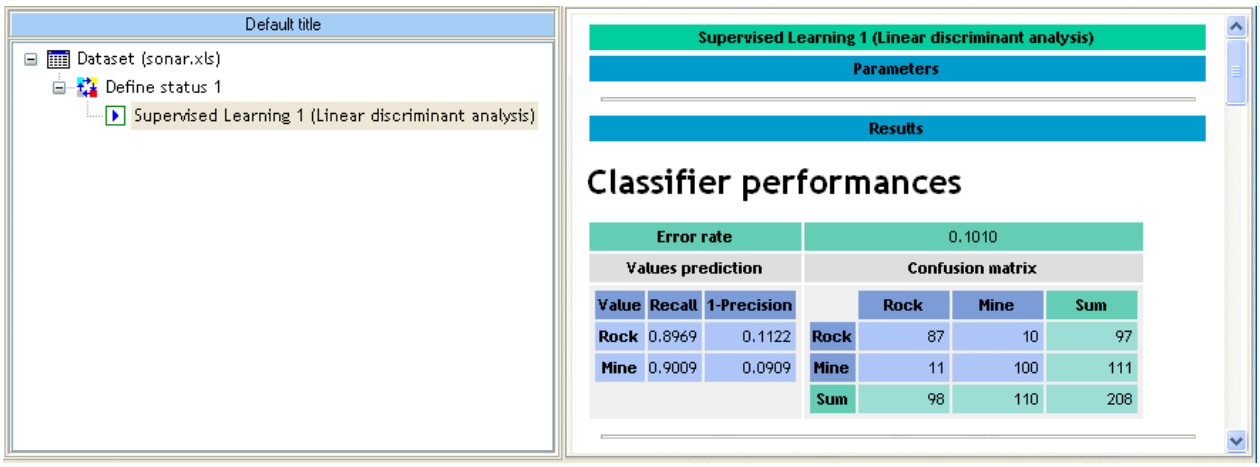
Les données sont importées, vérifiez que vous obtenez les mêmes spécifications que ci-dessous.



Analyse discriminante linaire

Dans un premier temps, nous réalisons une analyse discriminante linéaire. Comme son nom l'indique, cette méthode induit un séparateur linéaire dans l'espace de représentation, ses qualités et ses défauts sont bien connus maintenant, nous allons nous en servir comme méthode de référence.

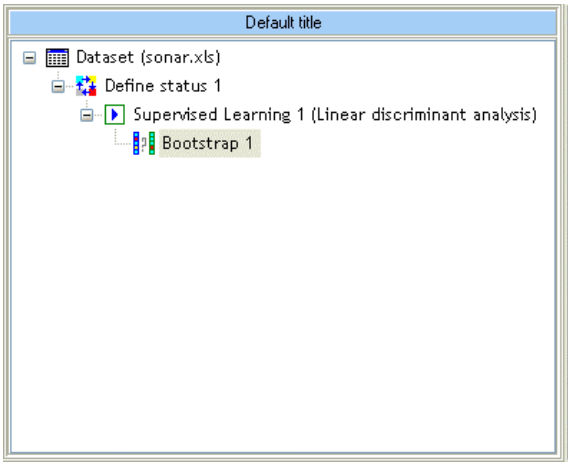
Le diagramme de traitements correspondant est le suivant, nous avons mis en TARGET l'attribut « CLASS » (discret et binaire), en INPUT les autres attributs (continus). Le taux d'erreur en resubstitution est de 10%, ce qui semble satisfaisant par rapport au taux d'erreur du classifieur par défaut qui est de 46% (97/208).

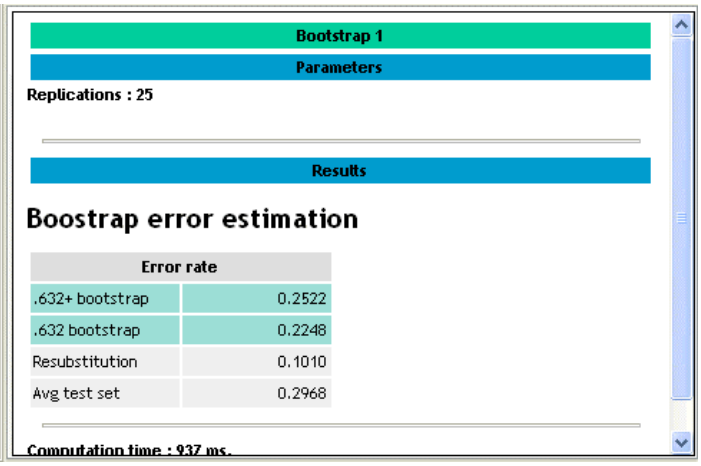


Il faut être très prudent par rapport à ce résultat, nous savons que le taux d'erreur en resubstitution est un estimateur biaisé du taux d'erreur, le biais est d'autant plus fort dans

notre cas que la dimensionnalité (60 descripteurs) est élevée par rapport à la taille de l'échantillon (208 observations).

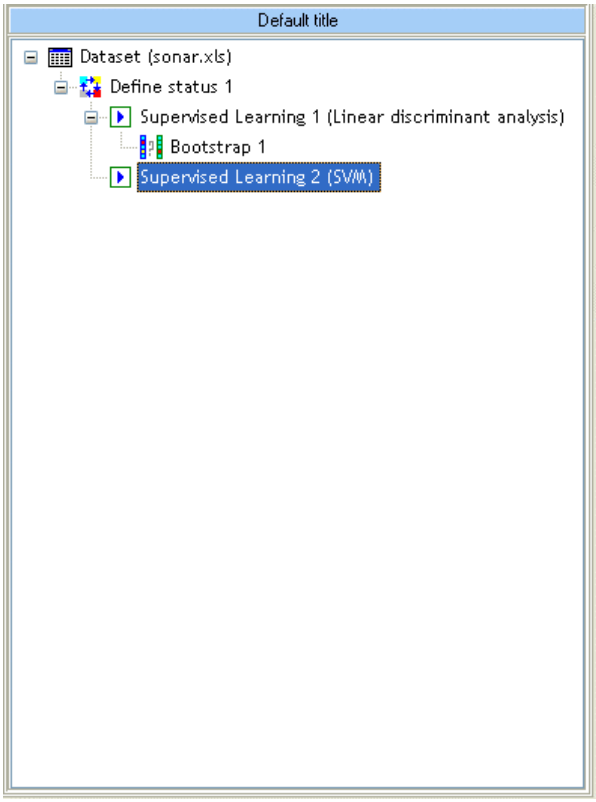
Pour calculer de manière adéquate le taux d'erreur, nous mettons en œuvre la procédure d'évaluation BOOTSTRAP. Les résultats sont édifiants, le « vrai » taux d'erreur est en réalité proche de 25%.

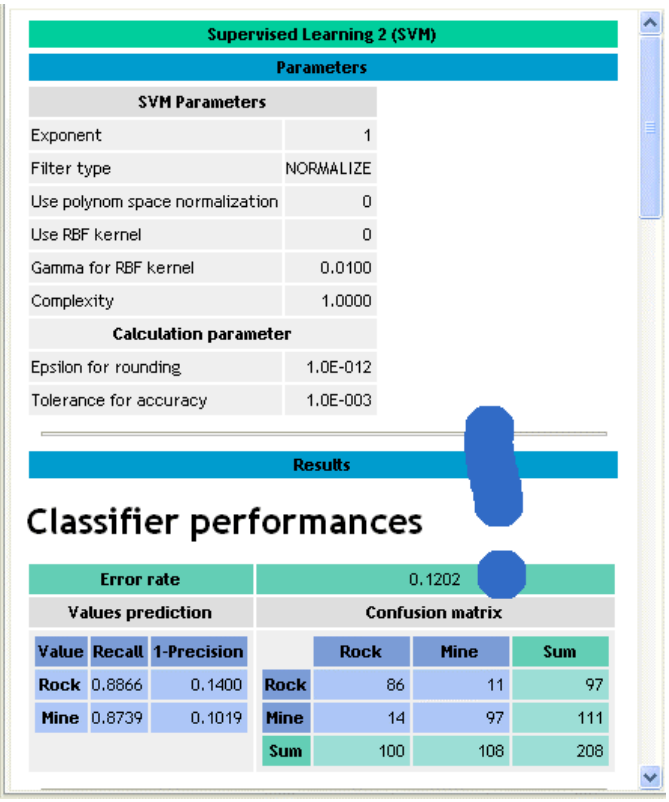




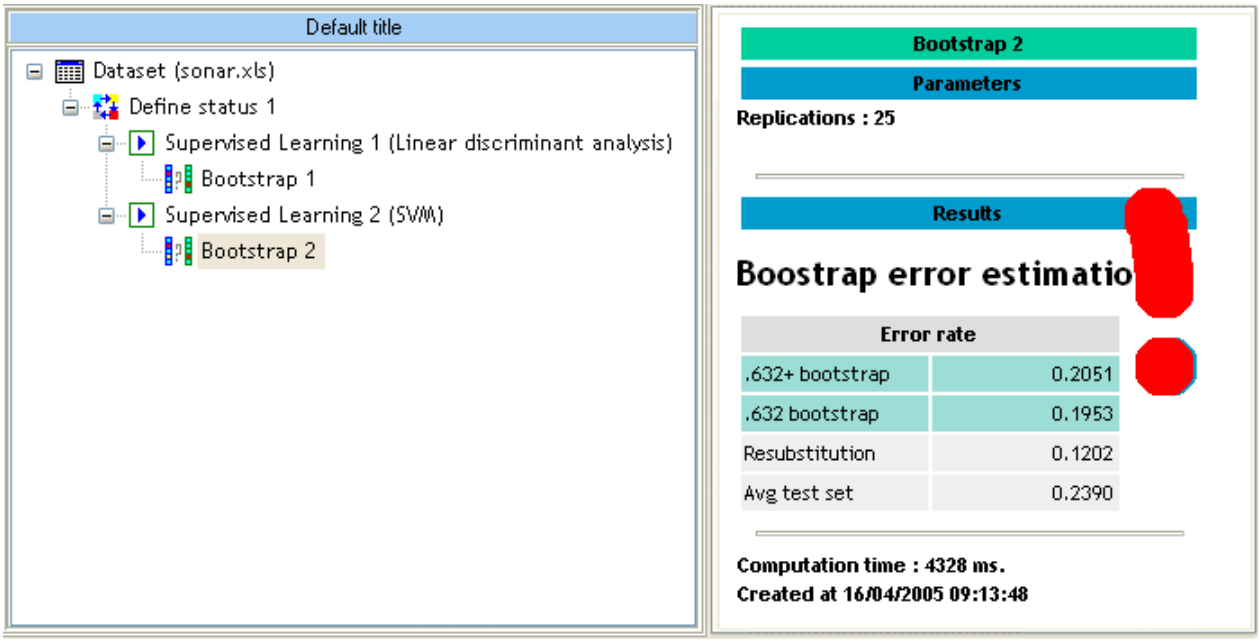
SVM linéaire

Plaçons maintenant le composant SVM, le paramétrage par défaut permet d'obtenir un séparateur linéaire, nous pouvons comparer les résultats.





Les résultats semblent similaires (taux d'erreur en resubstitution de 12%). Voyons ce qu'il en est en ce qui concerne l'estimation BOOTSTRAP de l'erreur.



Le SVM linéaire se démarque, le « vrai » taux d'erreur est autour de 20%, nettement amélioré par rapport à l'analyse discriminante.

C'est une constante remarquable de cette méthode, elle résiste très bien au sur-apprentissage même dans des dimensions très élevées (nous avons mené des tests sur les fichiers des protéines -- cf. didacticiel NIPALS – avec approximativement 7000 descripteurs pour 100 observations, la méthode tient la route alors qu'une très grande majorité des descripteurs ne sont pas pertinents).

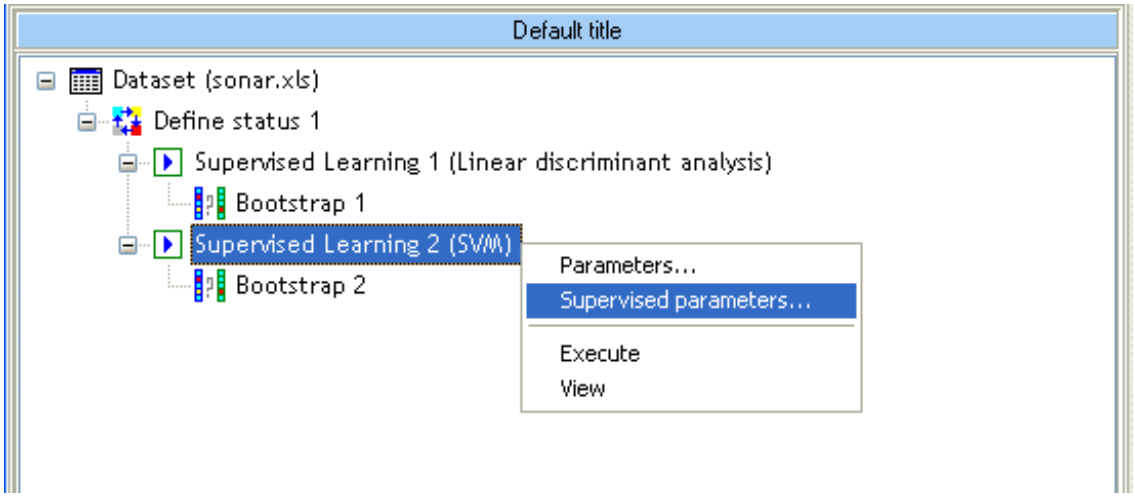
Pourquoi ? Par rapport à l'analyse discriminante, le biais de représentation est le même, nous cherchons un séparateur linéaire ; l'explication vient du biais de préférence qui est très contraignant dans les SVM, produisant ainsi un classifieur remarquablement stable, la fameuse « malédiction de la dimensionnalité » a du mal à frapper.

SVM avec noyau polynomial de degré 2

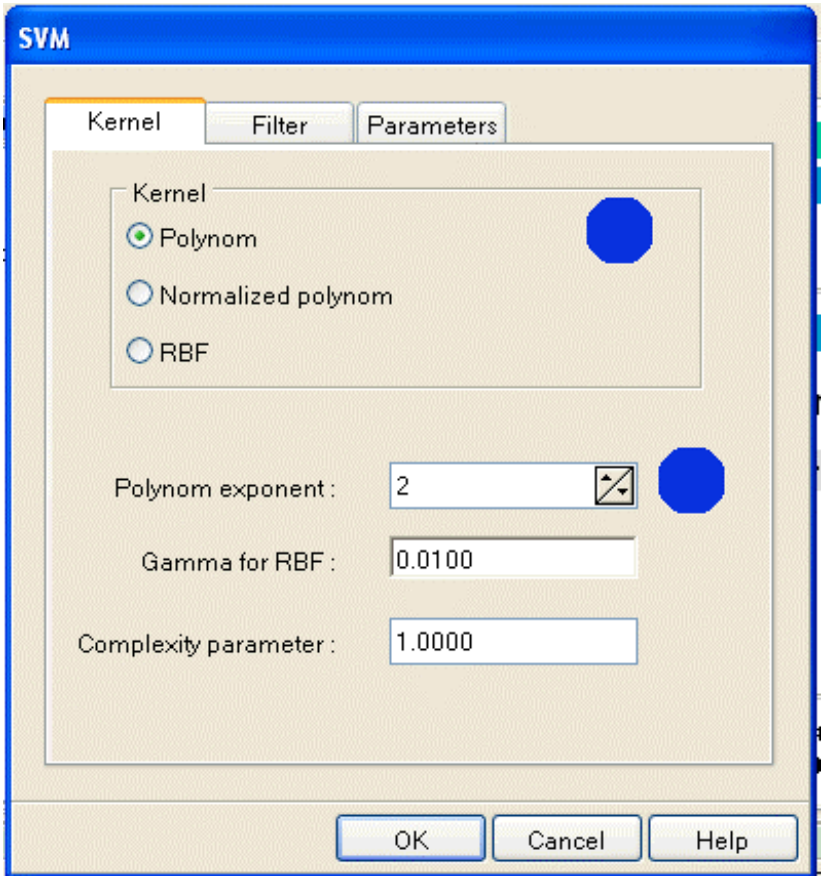
L'autre avantage des SVM est qu'il est possible de se projeter dans un espace de plus grande dimension : nous savons qu'un séparateur linéaire dans un espace de combinaisons des variables bien choisies permet d'induire un classifieur non-linéaire dans l'espace initial : **le secret des SVM réside dans l'utilisation des fameuses fonctions noyau qui permettent cette projection sans avoir à générer explicitement ces combinaisons de variables.**

Dans ce qui suit, nous utilisons un noyau polynomial de degré 2, ce qui revient à « construire » un espace composé du carré des variables et des produits croisés entre les variables.

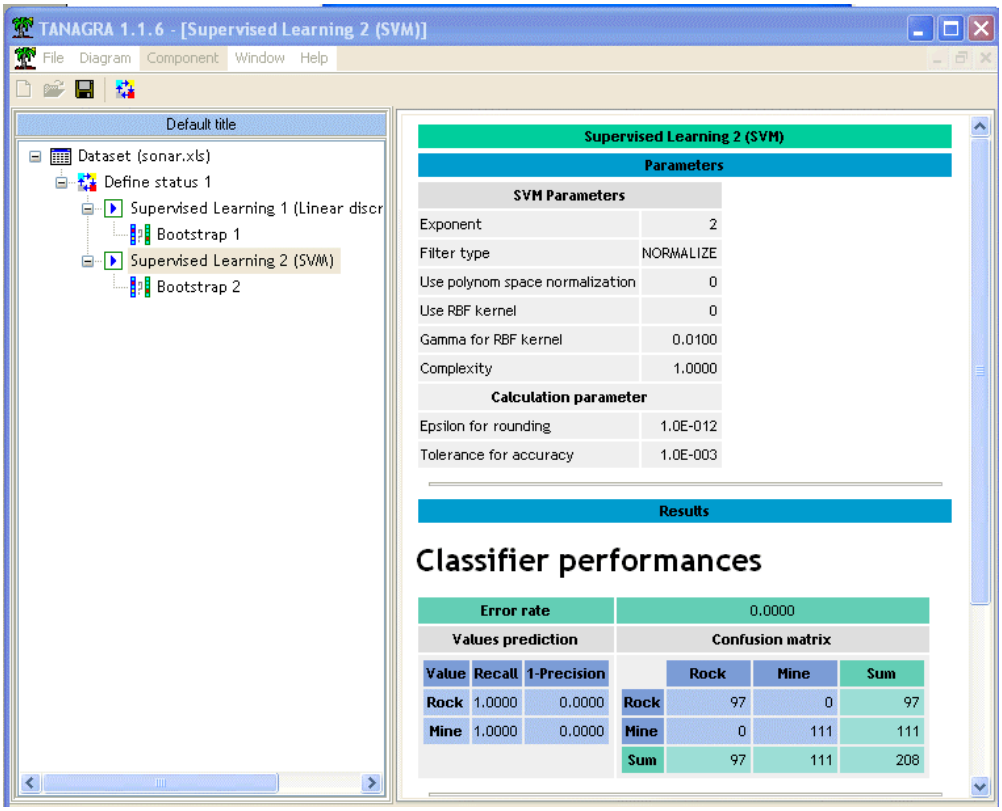
Le paramétrage de la méthode est réalisé via un clic droit sur le composant SVM



Puis la modification du degré du noyau polynomial.



Le taux d'erreur en resubstitution (0%) est très trompeur, dans un tel espace il est usuel qu'en apprentissage nous trouvons une courbe de séparation parfaite.



Le taux d'erreur du SVM à noyau polynomial de degré 2 estimé à l'aide du BOOTSTRAP est d'à peu près 15% sur le fichier SONAR (par comparaison, l'analyse discriminante propose un taux d'erreur de 25% et un SVM linéaire 20%).