Objectif

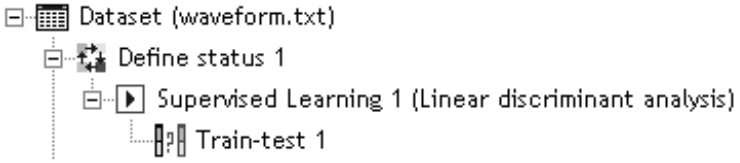
Montrer que dans certains cas, il peut être très avantageux d'utiliser les axes factoriels pour les calculs. Dans ce didacticiel, nous prendrons les deux premiers axes factoriels dans un problème de prédiction à l'aide de l'analyse discriminante linéaire.

Fichier

Les fameuses « Ondes de Breiman » (fichier WAVEFORM), très intéressant car étant généré, on connaît exactement le taux théorique d'erreur du classifieur bayesien, càd le meilleur taux que l'on peut atteindre sur ces données, il est de 0.14 (page 55 du livre de Breiman et al.). On ne peut pas faire mieux. Autre intérêt pour ce fichier, il a été maintes fois utilisé dans des groupes de travail, notamment pour comparer les performances des algorithmes d'apprentissage.

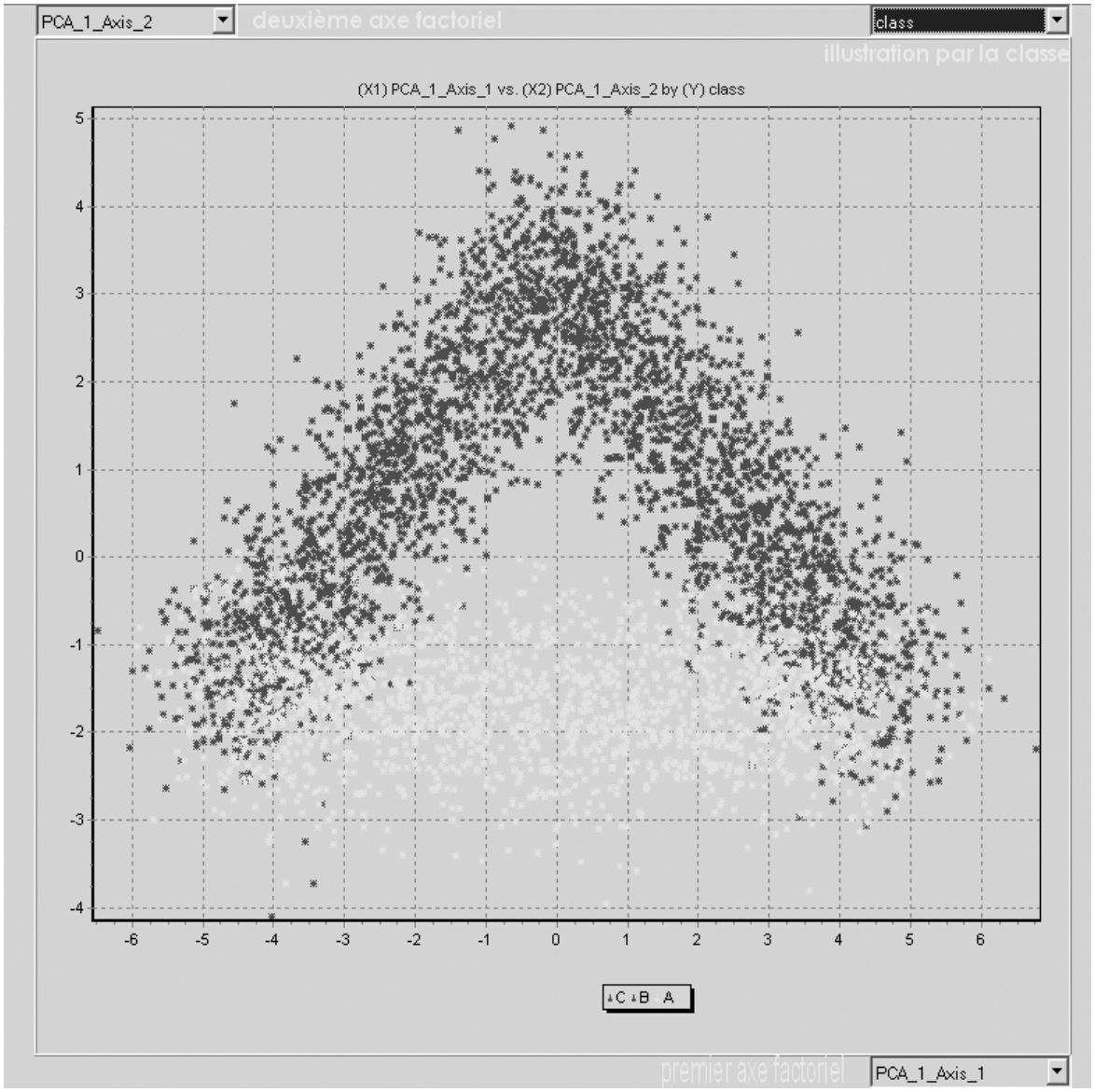
Etapes de l'expérimentation

1. Charger les données, il comporte 5000 individus et 22 variables.
2. Définir le statut des variables, « class » étant celle à prédire, les autres, toutes continues, sont les descripteurs.
3. Brancher une analyse discriminante (LDA) avec un méta-apprentissage simple, exécuter la filière, le taux d'erreur apparent (en re-substitution) est de 0.1350.
4. Pour se rapprocher des conditions d'expérimentations d'un projet Inter-PRC « Méthodes Symboliques Numériques de Discrimination », il y a quelques années (cf. Gallinari & Gascuel), nous allons évaluer le classifieur par une subdivision « apprentissage – test » en plaçant 300 individus pour l'apprentissage, et le reste, 4700 individus, en test. Cette procédure est répétée 10 fois. Pour ce faire, nous plaçons un composant « Train-Test » à la suite de l'apprentissage par LDA, avec les paramètres suivants : Train set Proportion = 0.06 (0.06 x 5000 = 300), et Répétition = 10. Le taux d'erreur mesuré en test est alors de #0.20. Ce qui est cohérent avec les résultats présentés dans la littérature, on est loin en tous les cas du 0.135 ci-dessus.



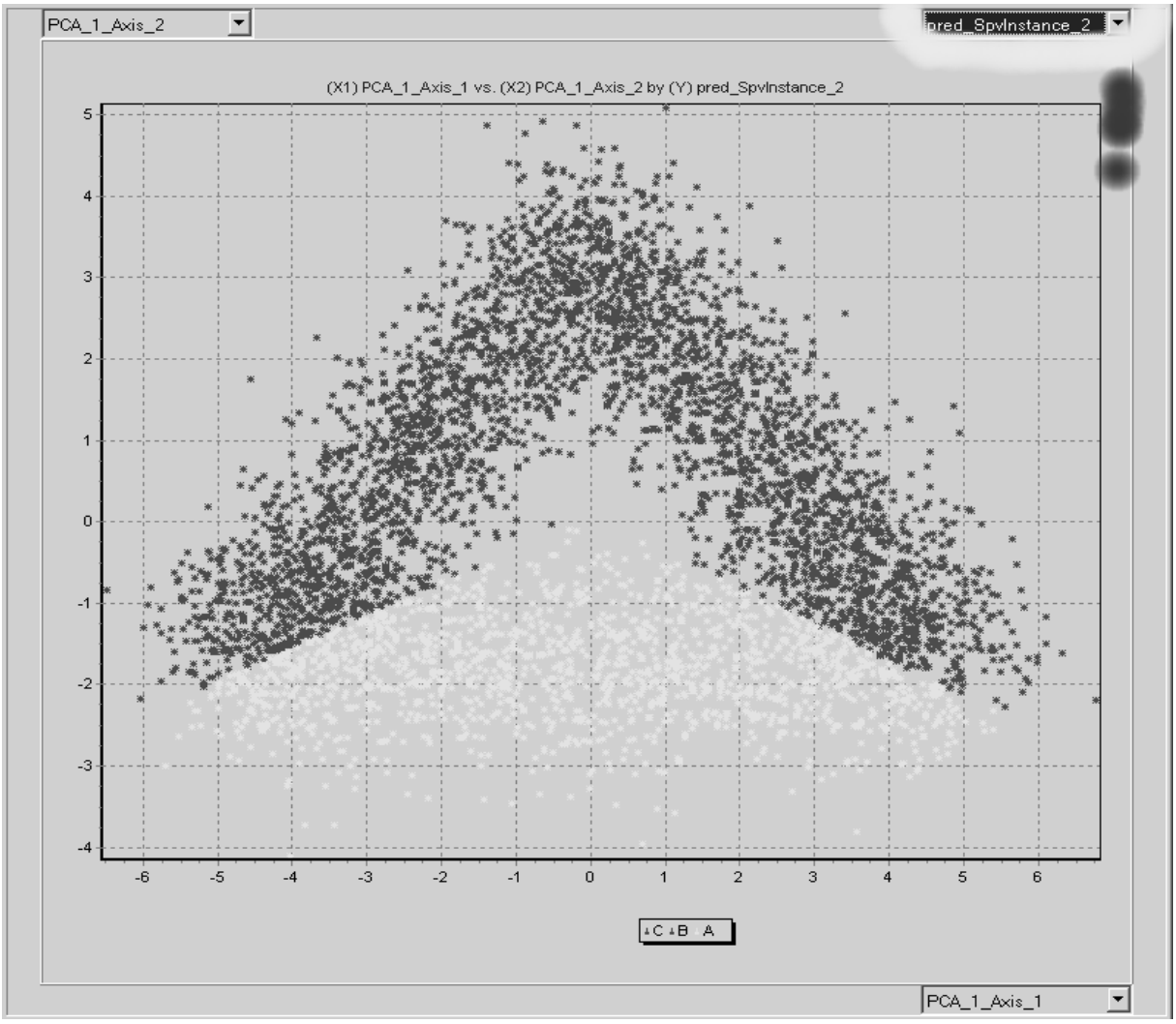
5. Essayons maintenant de voir ce qui se passe si l'on veut construire un classifieur à partir des axes factoriels de l'ACP. Placer un « Define Status » sous les données, sélectionnez toutes les variables continues comme « Input ».
6. Placez alors à la suite le composant « Principal Component Analysis » (ACP), laissez les paramètres par défaut. L'ACP va alors générer plusieurs variables correspondant aux axes factoriels, elles sont maintenant disponibles en aval du composant.
7. On se rend compte alors que les deux premiers axes résument déjà plus de la moitié de l'information disponible, en projetant les individus dans le premier plan et en les illustrant par leur classe d'appartenance, utilisez le composant Scatterplot que vous

placerez après l'ACP, on se rend compte que le meilleur séparateur, dans le plan tout du moins, paraît évident.



8. Reste alors à redéfinir le statut des attributs en mettant en « Target », la classe à prédire toujours, et en « Input », les deux premiers axes factoriels. Placez à la suite de ce composant une LDA comme précédemment et lancer l'apprentissage. Le taux d'erreur apparent est de 0.1350.
9. Essayons de nouveau d'évaluer le modèle d'apprentissage en construisant une évaluation « Train-Test » (300 contre 4700 individus), répétés 10 fois**. Attention ! Sur ce chemin partant de la racine jusqu'au composant « Train-Test », pour chaque session d'évaluation, les 300 individus sont bien utilisés pour l'apprentissage sur chaque composant, l'ACP est donc réalisée sur 300 individus, les 4700 autres étant projetés sur les axes comme individus supplémentaires, c'est la condition sine qua non pour que cette expérimentation soit valide (!!!)**. Et là, on constate que l'erreur mesurée en test est de #0.15, surclassant largement la première approche ci-dessus.
10. Pour s'en convaincre, effectuons la projection des individus sur les axes factoriels, en les illustrant à l'aide de la prédiction effectuée par l'analyse discriminante à partir des

axes factoriels. On constate qu'il a bien construit plusieurs séparation linéaire, une par modalité de la variable à prédire, et que les erreurs surviennent surtout au niveau des chevauchements entre les groupes d'individus.



11. A ce stade, le diagramme de traitements est le suivant.