TP4 – **Sélection de modèle**

{Yann.Esposito}@lif.univ-mrs.fr

13 octobre 2005

☞ Nous allons étudier à travers ce TP les différentes façon de faire de la sélection d'attributs.

1 Les données

Allez sur le site : http://www.lif.univ-mrs.fr/~esposito/ Dans la section enseignement, téléchargez tous les fichier nécessaire au TP n°4.

Lancez Weka (téléchargez-le si nécessaire)

Les données sont des photos de 40 personnes. Chaque personne a été prise 10 fois en photo. Il faudra apprendre sur les 5 premières photos et reconnaitre chaque personne sur les photos restantes.

Les fichiers ont été traité. Les photos n'ont plus qu'une taille de 46×56 pixels. Les bitmaps ont été transformés en format csv chaque attribut `xi_j` correspondant à la couleur sur 8 bits du pixel de situé sur la $i^{\text{ème}}$ colonne et $j^{\text{ème}}$ ligne. La classe est de type nominal, de v1 à v40 correspondant à chaque visage.

Il faut noter que ni SAS ni Weka n'étaient capable d'ouvrir les fichiers correspondant aux images non réduites (taille $92 \times 112 = 10304$ attributs).

Remarquez que le nombre d'instances est bien inférieur au nombre d'attributs.

2 Sélection manuelle

2.1. Quel est l'influence de l'attribut `x1_1` ?

2.2. Utilisez le filtre non supervisé d'instances `RemoveWithValues` de façon à ne conserver que les instances correspondant à la première photo.

(a) Quel est le type de distribution correspond le mieux à chaque pixel ; constante, linéaire ou exponentiel ?

(b) Celà vous parraît-il normal ?

2.3. Recommencez à utiliser un filtre de façon à conserver uniquement les instances correspondant aux deux premier visages.

(a) Quel est le type de distribution maintenant ?

2.4. Apprenez avec l'algorithm j48 en utilisant un pourcentage split de 66%. Il faut attendre un peu (80 secondes sur un G4 à 800MHz). Est-ce que les résultats sont aussi bon que ce à quoi vous vous attendiez ?

☞ Vous avez pu remarquer que le temps d'utilisation est fastidieux. À partir de maintenant, nous allons diminuer le nombre d'attribut de façon aveugle mais en essayant de contrôller la perte d'informations.

2.5. utilisez le filtre *RandomProjection*. Choisissez de conserver seulement 5% du nombre d'attribut. Avez-vous beaucoup perdu de qualité d'apprentissage avec J48 ?

2.6. Supprimez les attributs de manière aléatoire. J'ai fabriquer des listes d'attributs de façon totallement aléatoire (voir sur mon site). Faite un copier-coller de la liste d'attributs en utilisant le filtre d'attributs `Remove` en n'oubliant pas d'inverser le résultat.

(a) Vérifiez les différences de résultats obtenus.

(b) Sur le meilleur des échantillons, supprimer aussi les données dans le fichier de test et réévaluez la qualité sur l'échantillon test. Puis recommencez un apprentissage en utilisant tout l'échantillon d'entraînement et en testant la qualité sur l'échantillon test.

3 Sélection automatique

☞ Nous allons maintenant utiliser des méthodes automatiques qui permettent de faire un choix non aveugle des attributs à sélectionner.

3.1. Cliquez sur l'onglet `Select Attributes`. Le nombre d'attribut est trop grand pour pouvoir utiliser les valeurs par défaut.

(a) Choisissez l'évaluateur d'attributs (Attribute Evaluator) `InfoGainAttributeEval`. cet évaluateur va classer les attributs en fonction de leur gain d'information par rapport à la classe.
Choisissez comme méthode de recherche Ranker. Avec cette méthode, tous les attributs sont listés avec un valeur de qualité associé. La liste est alors données par ordre décroissant d'intérêt.
Lancez la sélection et sélectionnez eviron le dixième des premiers attributs choisis. Évaluez la qualité de l'apprentissage en utilisant ces nouveaux attributs.

(b) recommencez en sélectionnant environ la moitié des attributs

(c) faite une autre sélection après avoir enlevé deux tiers des attributs (un peu moins de 800) en utilisant la méthode `CfSubsetEval` et en sélectionnant la méthode de recherche `BestFirst`. Évaluez une nouvelle fois le nouveau choix de sélection en prenant bien soin de vérifier les gains et les pertes de qualités avant et après suppression des attributs.

✳