TD1 – **WEKA**

{Francois.Denis, Yann.Esposito}@lif.univ-mrs.fr

24 novembre 2004

☞ Ce TP sera dédié à la prise en main de WEKA qui permettra l'introduction de diverses notions essentielles à l'apprentissage.

0 Ouverture

Dans le cas où le logiciel WEKA n'est pas déjà installé, veuillez l'installer en allant sur le site http://www.cs.waikato.ac.nz/~ml/weka/. Téléchargez alors le fichier d'installation avec java 1.4 (fichier `weka-3-4-3jre.exe`. Exécutez-le. Une fois l'installation terminée, lancez WEKA.

☞ Lors de son ouverture Weka ouvre une petite fenêtre avec quatre boutons. Choisissez `Explorer`. Lorsque le mode `Explorer` s'ouvre, vous pouvez voir 6 onglets : `Preprocess, Classify, Cluster, Associate, Select attributes, Visualize`.
Le mode `Preprocess` permet de charger un fichier de données d'en avoir un aperçut rapide et d'appliquer des filtres à ces données.
Les modes `Classify`, `Cluster` et `Associate` permettent de faire de l'apprentissage proprement dit. `Classify` permet de faire de l'apprentissage supervisé (les données sont étiquetées), `Cluster` permet d'utiliser des algorithmes d'apprentissage non supervisés et `Associate` permet de générer des règles d'associations des données. (Dans ce TP, nous ne nous intéresserons pas aux modes `Cluster` et `Associate`).
Le mode `Select attributes` permet de sélectionner les attributs les plus significatifs en utilisant des algorithmes supervisés.
Le mode `Visualize` permet de visualiser des graphiques en 2D des données. En particulier cela permet de se faire une meilleure idée de l'organisation de celles-ci.

1 Les données

☞ WEKA utilise (entre autres) le format de fichier ARFF pour enregistrer les données. Il s'agit d'une liste d'exemple auxquels sont associées des valeurs d'attributs.

Ouvrez avec un éditeur quelconque (notepad par exemple) le fichier `contactlenses.arff` qui se trouve dans le répertoire `data` du répertoire d'installation de WEKA. Tout d'abord remarquons que l'on peut insérer des commentaires dans ces fichiers en commançant la ligne avec le caractère %. La première ligne qui n'est pas un commentaire est `@relation contact-lenses`. Il s'agit de donner le nom contact-lenses au fichier de données.

Il y a ensuite une liste de lignes de la forme commançant par `@attribute`. Cette ligne définit un *attribut*. La ligne `@attribute age {young, pre-presbyopic, presbyopic}` définit l'attribut `age` qui peut avoir les valeurs `youg` ou `pre-presbyopic` ou `presbyopic`.

Une fois que les attributs ont tous été définis, la ligne `@data` indique le début des données. Il s'agit d'une liste de valeurs d'attributs ordonnés. Chaque ligne correspond à un exemple. Ainsi la ligne `young,myope,no,reduced,none` correspond à l'instance

âge	spectacle-prescrip	astigmatism	tear-prod-rate	contact-lenses
young	myope	no	reduced	none

1.1. Lancez WEKA. Cliquez sur `Explorer`, une autre fenêtre s'ouvre. Sur cette dernière cliquez sur `Open File`, allez dans le répertoire `data` puis ouvrez le fichier iris.arff.

(a) Combien y a-t-il d'instances ? d'attributs ?

(b) Quels sont les noms des attributs et leur numéro ?

(c) À quoi correspond la zone `Selected attribute` à droite juste sous le bouton `Apply` ? Vérifiez ce qui se produit lorsque vous cliquez sur différents attributs. À quoi correspondent les valeur ; `Name, Type, Missing, Distinct` et `Unique` ?

(d) A quoi correspond le petit graphique en bas à droite ? A quoi sert l'onglet qui se trouve juste en haut à gauche du graphique ? Que fait le bouton `Vizualize all` ?

2 Filtres sur les données

☞ Il est souvent nécessaire de prétaiter les données avant d'utiliser un algorithme d'apprentissage. Cela permet notamment de supprimer les instances correspondant à des erreurs de mesures ou d'éliminer des attributs superflus. Cela peut aussi permettre d'uniformiser les données.

Il existe un grand nombre de filtres que WEKA peut appliquer aux données. Pour les choisir, il suffit de cliquer sur le bouton `Choose` dans la zone `filter` en haut de la fenêtre.

Vous avez le choix entre des filtres supervisés (qui utilisent la classe des données) et non supervisés. Une fois ce choix accomplis, vous avez le choix entre des filtres d'attributs ou d'instances.

Commencez par regarder les filtres non supervisés. Essayez de comprendre ce qu'ils font avec leur description (il suffit de cliquer sur le nom du filtre à droite du bouton `Choose`). Puis appliquez les (cliquez sur `Apply`) pour vérifier qu'il font bien ce à quoi vous vous attendez. En particulier :

2.1. Que font les filtres non supervisés d'attributs `remove`, `normalize`, `standardize`.

2.2. Que font les filtres non supervisés d'instances `normalize`, `RemoveMisclassified`, `resample`

2.3. Que font les filtres supervisés d'attributs `AttributeSelection` et `StratifiedRemoveFolds`.

2.4. Expliquez l'intérêt de tous ces filtres.

3 Visualisation des données

☞ La visualisation de données peut permettre de se faire une idée de l'organisation de celles-ci.

Cliquez sur l'onglet `Visualize`. Vous voyez un tableau de graphiques étiquetés par les attributs des données. Chaque graphique correspond à un graphique où chaque point représente une instance colorée en fonction de sa classe. La position du point est données en abscisse par l'attribut situé en haut du tableau et en ordonnée par l'attribut situé à gauche du tableau.

3.1. Vérifiez que la diagonale du tableau ne contient des graphiques ou toutes les instances sont sur la droite x=y.

3.2. Toujours pour le jeu de données `iris.arff` pouvez vous dire s'il existe un attribut unique avec lequel on va pouvoir bien classer ? Si oui lequel ?

3.3. Existe-t-il deux attributs avec lesquels le classement peut être très bon ? Si oui lesquels ?

☞ En cliquant sur une croix dans une fenêtre de visualisation, l'ensemble des instances correspondant à cette croix sont décrit de façon plus précise dans une autre fenêtre. En particulier cette fenêtre permet d'accéder au numéro des instances cliquées.

3.4. Qu'elle est l'utilité du `Jitter` ?

3.5. Expliquez comment on peut simuler un graphique tridimensionnel en changeant la coloration des instances.

4 Classification

☞ La classification est l'apprentissage supervisé. Les algorithmes de classification prennent en entrée un ensemble de données étiquetées et renvoient des modèles qui permettent de classer de nouvelles données non étiquetées.

Par défaut le classifieur choisi est `ZeroR`. Il s'agit du classifieur qui choisit la classe majoritaire. Lancez ce classifieur sur le jeu de données `iris.arff`.

Dans la zone `Classifier output` vous pouvez voir les informations que le classifieur renvoie.

Après la ligne

```
=== Run information ===
```

les information sont données :

- `Scheme` donne le classifieur utilisé
- `Relation` donne le nom de la relation utilisée (les données)
- `Instances` donne le nombre d'instances
- `Attributes` liste les attributs
- `Test mode` Donne la façon d'apprendre validation croisée, ensemble de test...

Après la ligne

```
=== Classifier model (full training set) ===
```

Le modèle rendu par l'algorithme est donné sous format texte.

Après la ligne

```
=== Summary ===
```

l'algorithme renvoie un résumé des résultats obtenus. puis le détail de l'efficacité de l'algorithme par classe et enfin la matrice de confusion.

4.1. L'algorithme OneR choisit un seul attribut et choisit de classer en fonction de celui-ci. Lancez l'algorithme OneR sur le jeu de données.

(a) Quel est le taux de réussite de l'algorithme ? L'attribut choisi est-il le même que celui que vous aviez choisi avec la visualisation ?

4.2. Les classifieurs sont classés en 7 classes :

- bayes – comporte notamment naïve Bayes et les réseaux Bayesiens ;
- functions – comporte les réseaux de neurones, les régressions linéaires... ;
- lazy – comporte IB1 (le plus proche voisin) et IBk (les k plus proches voisins) ;
- meta – comporte des algorithmes comme le Boosting notamment AdaBoost ;
- misc – Divers algorithmes exotiques ;
- trees – contient C4.5 sous le nom J48 ;
- rules – des règles d'apprentissages comme le choix de la classe majoritaire ou apprendre en utilisant un unique attribut.

4.3. Essayez chacun des algorithmes décrits en explorant les paramètres de chacun des algorithmes. En particulier testez le comportement de J48 et de naïve Bayes.

4.4. Expliquez et constatez les différentes options de tests, c'est-à-dire :
- utiliser l'ensemble d'apprentissage comme ensemble test ;
- utiliser un ensemble donné ;
- utiliser de la validation croisée ;
- utiliser une découpe de l'échantillon d'apprentissage.

(a) Expliquez les problèmes liés à l'utilisation de l'ensemble d'apprentissage pour l'ensemble qui permet d'effectuer les tests de qualité de l'apprentissage.

(b) Expliquez pourquoi il vaut parfois mieux utiliser la découpe de l'ensemble d'apprentissage que de la validation croisée.

(c) Qu'est-ce que faire de la validation croisée où le nombre de "Folds" est égal au nombre d'instances ?

☞ Lorsque des classifications sont effectuées, dans la zone `Result list`, en cliquant avec le bouton droit, une liste d'options apparaissent. En particulier il y a l'option `visualize classifier errors` qui permet d'enter en mode visualisation et qui encadre les instances erronées.
Une autre façon de constater ce qui s'est produit est de cliquer sur le bouton `More options...`.
Une fenêtre d'option s'ouvre dans laquelle on peut choisir l'option `Output predictions`

4.5. Est-ce que les instances qui posent problèmes sont souvent les mêmes ? Par exemple entre j48 et NaiveBayes. Préférez le mode `percentage split` pour faire cette vérification.

5 Étude approfondie des filtres de prétraitement

5.1. Établissez un protocole qui permet de montrer qu'un filtre est utile pour un classifieur donné.

5.2. Vous allez évaluez différents filtres pour le classifieur J48 en utilisant votre protocole.

(a) Quels sont les filtres qui semblent les plus intéressant pour J48 et pourquoi ?

(b) Quels sont les filtres pour lesquels vous ne ferez pas de tests et pourquoi ?

(c) Quels sont les filtres les plus efficaces ?

(d) Les résultats sont-ils étonnant ?

6 choix d'attributs

☞ Lorsque le nombre de données est trop grand ou que l'algorithme d'apprentissage à utiliser demande trop de ressources, il est souvent nécessaire de sacrifier de l'information pour gagner du temps de calcul.
Se pose alors la question du choix des informations à éliminer.

Allez dans `Select attributes` et choisissez comme évaluateur d'attribut `CfsSubsetEval` et comme méthode de recherche `BestFirst`. Cliquez sur `start`.

6.1. Interprétez ce qu'il y a dans la zone `Attribute selection output`. Ce résultat correspond-t-il au choix que vous aviez fait lors de la visualisation des données ?

6.2. Apprenez en utilisant seulement ces deux attributs. La qualité de l'apprentissage est-elle meilleure ou moins bonne ? Expliquez pourquoi, en particulier pour J48.

6.3. Expliquez comment la sélection d'attributs peut améliorer la qualité de l'apprentissage.

7 Recherche du meilleur modèle

☞ Les multiples algorithmes d'apprentissage peuvent avoir des taux de réussite parfois très grand. Pourtant face à de nouvelles données, les modèles générés par ces algorithmes peuvent s'avérer très médiocres. Il va être question de faire le bon choix parmi les modèles.

7.1. Téléchargez le fichier `learn.arff` à l'adresse
http://lif.univ-mrs.fr/~esposito/pub/learn.arff.

Il s'agit d'un fichier de données. Vous devez apprendre le meilleur modèle possible sachant que le modèle sera testé sur un ensemble de test qui ne vous est pas fourni.

(a) Dans la fenêtre `preprocess` vous pouvez avoir une idée de l'organisation des données. D'après ces visualisations, les données ont elles l'air facile à apprendre ? Justifiez votre réponse.

(b) La façon naturelle de procéder pour accomplir un choix de modèles est de fabriquer quelques modèles en modifiant légèrement quelques aspects (utilisation de filtres peu agressifs, modification légère des paramètres). Puis comparer entre eux ces premiers modèles. Vérifier ensuite quelles sont les modification les plus prometteuses et continuer dans cette voie.

Remarquez que cette façon de procéder est aussi un méta algorithme d'apprentissage.

☞ Sauvegarder tous les modèles intermédiaires que vous utilisez. À la fin du TP, vous devrez avoir choisi 2 modèles qui sont d'après vous les meilleurs. Il faudra alors tester leur qualité avec l'échantillon test. Puis vous vérifierez que ces modèles sont bien meilleurs que les modèles écartés.
Pensez qu'un taux de classification à peine meilleur n'est pas significatif. Le choix que vous accomplirez dans vos modèles doit être justifié. Certaines mesures semblent d'ailleurs être de meilleure qualité que d'autres.
Essayez de visualiser les données, l'intuition humaine est souvent meilleure que des algorithmes lorsque le nombre d'informations à traiter n'est pas trop important.

7.2. Faite un résumer des tests que vous avez accomplis. À quels moment avez vous accomplis des choix et pourquoi ? Donnez les justifications de votre parcours. Le travail ne doit pas être

une recherche aléatoire du meilleur modèle mais bien le fruit d'un protocole de recherche du meilleur modèle que vous énoncerez clairement.

Bien que le mode `Experiment` soit dédié à ce genre de problèmes, je vous demande d'accomplir ce travail uniquement en mode `Explorer`.

☞ N'oubliez pas que le but de cet exercice est de vous familiariser avec le comportement des algorithmes d'apprentissages, des filtres de prétraitement et de la visualisation. Ne perdez pas de temps en optimisation fastidieuses si elles ne vous permettent pas d'utiliser pleinement partie des fonctionnalités de WEKA.

Bonne chance.