TP1 – **Prise en main de WEKA**

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

22 septembre 2005

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

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-3-4-5jre.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 la fenêtre `Weka Explorer` s'ouvre, vous pouvez y voir 6 onglets : `Preprocess, Classify, Cluster, Associate, Select attributes, Visualize`. Le mode `Preprocess` permet de charger un fichier de données d'en avoir un aperçu 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 de manière fine avec des algorithmes supervisés. Le mode `Visualize` permet de visualiser les données ce qui permet d'améliorer l'idée que l'on se fait de leur organisation.

1 Les données

☞ WEKA utilise 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 commançant par `@attribute`. Ces lignes définissent un *attribut*. La ligne

```
@attribute age {young, pre-presbyopic, presbyopic}
```

définit l'attribut `age` qui peut avoir les valeurs `young` 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

age	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 valeurs ; `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 prétraiter les données avant d'utiliser un algorithme d'apprentissage. On dit souvent que l'on nettoie les données. 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 accompli, 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`, `RandomProjection`.

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

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

2.4. Expliquez quels peuvent-être les intérêts 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 que des graphiques où 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.

☞ La zone `Test options` permet de choisir de quelle façon l'évaluation des performances du modèle appris se fera.
- L'option `Use training set` utilise l'ensemble d'entraînement pour cette évaluation.
- L'option `Supplied test set` va utiliser un autre fichier.
- Lorsque l'option `Cross-validation` est sélectionnée, l'ensemble d'apprentissage est coupé en 10 (si `Folds` vaut 10). L'algorithme va apprendre 10 fois sur 9 parties et le modèle sera évalué sur le dixième restant. Les 10 évaluations sont alors combinées.
- Avec l'option `Percentage split`, c'est un pourcentage de l'ensemble d'apprentissage qui servira à l'apprentissage et l'autre à l'évaluation.

Par défaut le classifieur choisit est `ZeroR`. Il s'agit du classifieur qui choisit la classe majoritaire. Choisissez le mode d'évaluation `Use training set`. Lancez le 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 le mode d'évaluation du modèle : 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 ===
```

s'affiche un résumé de l'évaluation du modèle :

```
Correctly Classified Instances         100               66.6667 %
Incorrectly Classified Instances        50               33.3333 %
Kappa statistic                          0.5
Mean absolute error                      0.2222
Root mean squared error                  0.3333
Relative absolute error                 50      %
Root relative squared error             70.7107 %
Total Number of Instances              150
```

Tout d'abord le nombre d'instances bien classées ainsi que le pourcentage que cela représente. Puis le nombre d'instances mal classées avec la proportion associée. Puis vient la `kappa statistic` ; il s'agit de la valeur suivante :

$$\frac{P(A) - P(E)}{1 - P(E)}$$

où $P(A)$ correspond à la proportion d'instances bien classées et $P(E)$ correspond à l'espérance de bien classer par chance. Plus précisemment $P(E)$ est calculée à l'aide de la matrice de confusion (voir plus loin) avec la formule suivante :

$$P(E) = \frac{\sum_i (\sum_j M_{i,j}) \times (\sum_j M_{j,i})}{N^2}$$

où N est le nombre d'instances. Un coefficient de 1 signifie un modèle parfait alors que 0 correspond à un modèle qui se trompe toujours. On peut considérer qu'une valeur inférieure à $0,7$ est faible et qu'une valeur supérieure à $0,8$ est élevée.

Les 4 indicateurs suivants ne sont pertinents que pour la régression. `Mean absolute error` : il s'agit de l'erreur de prédiction moyenne. `Root squared error` : si cette valeur est significativement plus élevée que la `Mean absolute error` cela signifie qu'il y a des instances pour lesquelles l'erreur de prédiction est significativement plus grande que l'erreur de prédiction moyenne.

`Relative absolute error` et `Root relative squared error` correspondent aux deux même valeur que précédemment mais comparativement au classifieur de classe majoritaire.

Vient ensuite la matrice de confusion. La case à la ligne i et à la colonne j indique le nombre d'élément de classe i prédite comme faisant partie de la classe j.

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 choisit 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. Nous allons nous intéresser aux algorithmes `ZeroR`, `OneR`, `Naïve Bayes`, `IB1`, `IBk` et `J48`. Essayez chacun de ces algorithmes en explorant les paramètres modifiables.

4.4. Évaluez l'impact des différentes méthodes d'évaluations sur chaque algorithme.
- utilisez l'ensemble d'apprentissage comme ensemble test ;
- utilisez un ensemble donné ;
- utilisez la validation croisée ;
- utilisez une découpe de l'échantillon d'apprentissage.

(a) Expliquez les problèmes liés à l'utilisation de l'ensemble d'apprentissage pour l'évaluation du modèle.

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

(c) Expliquez l'intérêt de 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

5.6. Dans un premier temps nous allons étudier l'influence de certains filtres sur `Naive Bayes`. Il faut appliquer les filtres sur tous les attributs à l'éxception de la classe. Pour cela sélectionnez tous les attribut en cliquant sur `All` dans la zone `Attributes`. Désélectionnez ensuite class. De plus toutes les évaluation se feront avec `Percentage split` à 66%.

(a) Quelle est l'influence du filtre `Normalize` ? Expliquez.

(b) Quelle est l'influence du filtre `Standardize` ? Expliquez.

(c) Quelle est l'influence de `RandomProjection` en utilisant 3 dimensions ? En utilisant 1 dimension (valeur de `seed` 42) ? Que vous inspirent ces résultats ?

6 choix d'attributs

☞ Il faut souvent limiter le nombre d'attributs pour améliorer les résultats d'apprentissage ainsi que du temps de calcul. La question va être de savoir quelles informations doivent-être écartées.

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

6.7. 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 ?

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

7 Recherche du meilleur modèle

☞ Il va être question de trouver le meilleur modèle d'apprentissage.

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

Le but du jeu va être de trouver le meilleur modèle possible sachant qu'il existe un autre jeu de donné test qui servira une fois vos choix de modèles accomplis.

Pensez à utiliser le choix d'attributs, des filtres, faites des visualisations. Utilisez plusieurs classifieurs. Certains d'entre eux peuvent être inutilisables. Pensez à trouvez des moyens pour les utiliser malgré tout.

En fait, tous les moyens sont bons pour trouver la solution. Lorsque vous estimerez avoir trouvé le meilleur modèle, envoyez-le moi par email (en précisant le titre WEKA). Je testerai ce modèle avec un fichier test. Bonne chance.