

TP n°3

Fouille de données

Modélisation prédictive

Environnement de travail :

Système WEKA 3.4

Java 2 version JDK 1.4

Méthodes pour la modélisation prédictive dans WEKA

L'implémentation d'une méthode d'apprentissage particulière est réalisée par une classe Java. Les classes qui implémentent une méthode de classification supervisée ou une méthode de régression sont regroupées dans le package **weka.classifiers**. La classe **weka.classifiers.Classifier** définit la structure générale commune à toutes les méthodes de classification.

Exercice 1 – Règles de classification – Arbres de décision

Pour cet exercice, on utilisera les données du fichier **weather.nominal.arff** étudié au TP précédent

Effectuer une première analyse du jeu de données en utilisant les informations données par le fichier texte et les mesures et histogrammes fournis par Weka.

Pouvez-vous en tirer des informations sur la valeur prédictive des attributs ?

1.1 OneR

- Rechercher à partir du jeu **weather.nominal**, un modèle prédictif en utilisant la méthode implémentée par **OneR** avec les paramètres par défaut
- Analyser le résultat
- Quel est l'attribut choisi par la méthode?
- Quel est le taux d'exemples mal classés au total?
- Quel est le taux d'exemples mal classés dans chaque classe?
- Quel est la probabilité pour qu'un terrain impraticable soit classé "praticable"?

1.2 J48

- Rechercher à partir du jeu **labor2**, un modèle prédictif en utilisant la méthode implémentée par **J48** avec les paramètres par défaut

Les valeurs affichées pour chaque règle, indiquent le nombre moyen d'exemples qui atteignent la feuille de l'arbre (c.a.d vérifient la partie gauche de la règle).

- L'arbre de décision extrait peut être visualisé en cliquant bouton droit sur la ligne correspondant à l'expérience dans le panneau **Result list**
- Observer le modèle extrait ; a-t-il de meilleures performances :
- Quel est le taux d'exemples mal classés au total?
- Quel est le taux d'exemples mal classés dans chaque classe?
- Quel est la probabilité pour qu'un terrain impraticable soit classé "praticable"?

Exercice 2 – Règles de classification – Arbres de décision

Pour cet exercice, on utilisera les données du fichier **vote.arff**.

Ouvrir le fichier avec un éditeur de texte et observer les différents attributs.

Effectuer une première analyse de jeu de données en utilisant les informations données par le fichier texte et les mesures et histogrammes fournis par Weka.

Pouvez-vous en tirer des informations sur la valeur prédictive des attributs ?

2.1 J48

- Rechercher à partir du jeu **vote**, un modèle prédictif en utilisant la méthode implémentée par **J48** avec les paramètres par défaut
- Analyser le résultat
- Quel est le taux d'exemples mal classés au total?
- Quel est le taux d'exemples mal classés dans chaque classe?
- Quel est la probabilité pour qu'un vote "républicain" soit classé "democrate"? et réciproquement?
- Utiliser les outils de WEKA pour supprimer des attributs et observez les conséquences sur les résultats obtenus.
- Refaire les tests en utilisant d'autres classifieurs et comparez les résultats aux résultats précédents.