Introduction à Weka (algorithme Random Forest)

ING2 GI EISTI 2017-2018

PRÉSENTATION ET OBJECTIF

Dans ce TP nous allons continuer la découverte de la plateforme WEKA et de son API pour l'étude des algorithmes basés sur les arbres de décisions et notamment les "Random Forest". **Ce TP est à faire sous Linux.**

ALGORITHME ID3

Si tous les attributs sont de type qualitatif/énuméré/nominal, il est possible d'utiliser l'algorithme ID3 (Iterative Dichotomiser), construisant un arbre de décision par la méthode de sélection via le calcul d'entropie vue en cours au premier semestre

1. Lancez le GUI de WEKA puis sélectionner l'explorer
2. Sélectionnez l'onglet Preprocess et chargez le fichier `weather.nominal.arff` (version de weather avec uniquement des attributs qualitatifs)
3. Sélectionnez ensuite l'onglet Classify et choisissez l'algorithme ID3
4. Lancez l'algorithme sur le jeux de données avec le bouton start, sans modifier les paramètres.
5. Utilisez le bouton Log pour trouver quelle classe de l'API le GUI a utiliser
6. Reproduisez le résultat en ligne de commande via la commande `java -cp weka.jar …` (vous avez accès à la complétion automatique)
7. Faites varier les différents paramètres disponibles via l'API

ALGORITHME C4.5 (J4.8)

L'algorithme C4.5 améliore l'algorithme ID3 à deux niveaux : il permet la gestion des données non qualitatives (quantitatives) et la gestion des données manquantes. L'implémentation disponible de cet algorithme dans WEKA est la version J4.8.

Reproduisez les questions de l'exercice ci-dessus en utilisant maintenant le fichier `weather.numeric.arff` (Ne pas oublier que les températures sont en Fahrenheit.)

ALGORITHME RANDOM FOREST

Les deux algorithmes d'arbres de décision que nous venons de tester ont un défaut majeur, leur mauvaise généralisation (risque d'overfitting). Pour y remédier, le méta-algorithme des Random Forest génère des arbres à partir de vecteurs aléatoires et choisi la valeur retournée à la majorité.

Reproduisez les questions de l'exercice précédent.

UTILISATION EN PRODUCTION

Nous allons tester les algorithmes J4.8 et Random Forest sur un exemple célèbre de Machine Learning, le fameux MNIST (**http://yann.lecun.com/exdb/mnist/**)

1. Utilisez le programme JAVA disponible dans l'archive pour transformer les images de la base MNIST en fichiers ARFF
2. Utilisez J4.8 et RandomForest via l'API en faisant varier les paramètres et tracez les graphes suivant pour chaque algorithme :
    a. temps d'apprentissage / nombre d'exemples
    b. qualité de l'apprentissage / nombre d'exemples