Fouille de données - TP1

Polytech Montpellier - IG4

1 Présentation de WEKA

WEKA est un logiciel libre (java) qui permet d'appliquer de nombreaux algorithmes d'apprentissage et de fouille de données (règles, d'association, arbres de décision, K plus proches voisins, réseaux de neurones, etc).
Vous pouvez télécharger et installer weka sur vos ordinateurs en choisissant une version adaptée à votre système d'exploitation.

2 Format des fichiers

1. Lancez `weka`
2. Les bases de données exemples sont installées dans le répertoire `data`.

1. Ouvrez la base d'exemples `weather-nominal` (onglet `Preprocess - Open File`).
2. Combien y a-t-il d'exemples dans la base ?
3. Quels sont les attributs servant à décrire les exemples ? Pour chacun des attributs, quel est son type et les valeurs possibles ?
4. Visualisez la répartition des classes en fonction des valeurs d'attributs (onglet `Visualize`, attention à modifier la valeur de Jitter et de PointSize pour bien visualiser les exemples).
5. mêmes questions pour la base `iris.arff` En particulier, que constatez-vous sur l'interaction entre sepalwidth et sepallength ? et entre petalwidth et petallength ?
6. Ouvrez le fichier weather.arff grâce à un éditeur de textes. Décrivez l'en-tête et la structure du fichier.
7. On souhaite utiliser un fichier d'exemples qui n'est pas au format arff. Ouvrez le fichier bac.txt

   De manière générale, tout fichier au format CSV (Comma-Separated Values) est utilisable dans weka. Ouvrez la base `titanic.csv`

De nombreuses bases de données d'exemples pour l'apprentissage sont disponibles sur internet, par exemple sur le site de l'UCI ou sur https://www.kaggle.com/

3 Arbres de décision

3.1 Erreur

(a) Ouvrez la base titanic.
(b) Appliquez l'algorithme de classification C4.5 (J48 dans Weka).
(c) Dessinez l'arbre de décision obtenu.
(d) Combien d'exemples sont bien classés par cet arbre ?
(e) Combien sont mal classés ?
(f) Que représente la matrice de confusion ?
(g) Combien y a-t-il de Vrais positifs pour la classe SURVIVED = yes ? de Faux Positifs ?

3.2 Échantillons d'apprentissage et de test

(a) Avec quelle métode de test étaient obtenus les résultats précédents ?
(b) Plusieurs méthodes sont proposées sous Weka. À quoi servent-elles ? Que permettent-elles d'évaluer ? Rappelez les principes de chaque méthode de test proposée et ce à quoi correspondent les paramètres.
(c) Construisez un tableau de comparaison des résultats obtenus avec la base titanic pour chacune des méthodes de test disponibles dans weka en faisant varier les paramètres. Commentez ces résultats.
(d) Comment procéder pour un test de type *leave-one-out* ? Est-ce une bonne méthode de test pour cette base ? Et pour la base weather ?

3.3 Sélection d'attributs

Les filtres peuvent être utilisés pour modifier les fichiers de données et les données prises en compte, par exemple pour omettre des attributs dans la classification, ou des valeurs.
(a) Ouvrez la base iris.arff.
(c) Lancez la classification à l'aide de C4.5 (J48 sous weka).
(e) Modifiez la base des iris pour omettre les attributs petalwidth et petallength Quel nouvel arbre obtenez-vous ? Que se passe-t-il si vous supprimez plutôt les attributs sepalwidth et sepallength ? Pourquoi ?

3.4 Valeurs manquantes

La base labor.arff contient-elle des valeurs manquantes ?

Avec les filtres, remplacez les valeurs manquantes.

Quel taux de bonne classification obtient C4.5 ?

Quel est le taux de bonne classification sans remplacer les valeurs manquantes a priori ?

3.5 Règles de décision

Décrivez une méthode simple pour convertir un arbre de décision en base de régles de décision `Si ... Alors.`

Testez le module PART de weka basée sur J48 (onglet `Classify - Classifiers - rules`) sur plusieurs bases pour générer des règles de décision.

4 k-plus proches voisins

A quoi correspondent les paramètres proposés par WEKA pour les k plus proches voisins (méthode IBk dans `classifiers - lazy`) ?

4.1 Choix de k

(a) Ouvrez la base `soybean`, et testez les performances en classification selon la valeur attribuée à k.
(b) Avec une validation croisée à 10 échantillons, quelle valeur maximale de k est-il judicieux de choisir ?
(c) Quelle méthode est utilisée pour calculer la classe des exemples à classifier ?
(d) À partir de quelle valeur de k les résultats sont-ils mauvais ?
(e) Quelle méthode de calcul de la classe doit être privilégiée pour de grandes valeurs de k ?
(f) Pour $k = 100$, quels sont les résultats (taux de bonne classification) si on ne prend pas en compte la distance des voisins ? et si on la prend en compte ?

4.2 Normalisation

Exécutez la méthode des k plus proches voisins sur la base labor avec $k = 1$. Sur cette base, quelles sont les valeurs minimales et maximales prises par chacun des attributs ? La normalisation des attributs vous paraît-elle nécessaire ? Pourquoi ? Effectuez des tests comparatifs avec ou sans normalisation. La normalisation aurait-elle apporté quelque chose pour les arbres de décision ?

5 Clustering

Lancez la méthode kmeans en ignorant l'attribut classe sur les bases précédemment utilisées.
(a) Quels sont les centres des clusters créés ?
(b) Comment retrouver l'appartenance de exemples de la base aux clusters ?
(c) Que signifie l'information *Number of iterations* ?