Master2 Informatique : Intelligence Artificielle (IA)
Module : *Fouille de données (FD)* Année universitaire : 2025/2026

TP N°4

(Environnement de travail : Weka)

La classification supervisée dans Weka :

L'onglet *Classify* permet d'appliquer des algorithmes de classification supervisée. Ces algorithmes construisent d'abord un modèle de classification appelé classifieur en utilisant un jeu d'apprentissage et testent ensuite ce classifieur sur un jeu de test. Les algorithmes disponibles sont classés par catégories : réseaux de neurones, arbre de décision, classifieurs bayésiens, fonction de régression, règles de classification et méta-heuristiques. Sur cet onglet:

- Le cadre Classifier permet de choisir l'algorithme utilisé. Cliquez sur le nom de l'algorithme sélectionné pour définir les paramètres de l'exécution.
- Le cadre Test options permet de choisir la méthode de formation du jeu d'apprentissage pour la construction du classificateur et du jeu de test pour son évaluation. Les options possibles sont :
  - Use training set : le jeu de données chargé sert de jeu d'apprentissage et de jeu de test. Cette méthode peut entrainer des problèmes de sur-évaluation et n'est que rarement utilisée.
  - Supplied test set : le jeu de données chargé sert de jeu d'apprentissage et le jeu de test est choisi en cliquant sur le bouton Set.
  - Cross-validation : pour une valeur K, le jeu est divisé en K partitions. L'une constitue le jeu de test, les autres forment le jeu d'apprentissage. Ce processus est répété K fois, chaque partition étant utilisée une fois comme jeu de test. Une valeur 10 pour K est en général conseillée.
  - Percentage split : définit le pourcentage du jeu de données chargé utilisé pour l'apprentissage. Le reste du jeu est utilisé comme jeu de test.
- La liste déroulante (Nom) permet de sélectionner l'attribut de classe dont on cherche à prédire les valeurs (ex : Play).
- Le cadre Classifier output affiche le classifieur généré ainsi que des statistiques sur l'exécution et la précision du classifieur. Selon la catégorie de l'algorithme utilisé, le classifieur lui-même pourra ou non être affiché. Certains algorithmes, tels que les réseaux de neurones par exemple, ne fournissent pas une description explicite du classifieur. D'autres, tels que les arbres de décision, fournissent une description explicite du classifieur.

La précision de l'application du classificateur sur le jeu de test est évaluée par une matrice de confusion qui indique pour chaque classe combien d'instances ont été correctement et incorrectement classées dans le jeu de test. Les 4 nombres calculés sont affichés dans la matrice de confusion :

- Vrais positifs : nombre d'instances yes classées yes
- Faux positifs : nombre d'instances no classées yes
- Vrais négatifs : nombre d'instances no classées no
- Faux négatifs : nombre d'instances yes classées no

Exercice 1 : Chargez le jeu de données *weather.arff* et sélectionnez l'algorithme de classification *J4.8* dans la catégorie *Classifiers / Trees*. Cet algorithme construit un arbre de décision par analyse du jeu d'apprentissage puis le teste sur le jeu de test. Dans le cadre Test *options* choisissez l'option *Cross validation* avec une valeur de 10 pour le paramètre *Folds*. Laissez les paramètres de l'algorithme à leur valeur par défaut.

Exercice 2 : Notez dans un tableau les caractéristiques de chacun des chemins allant de la racine de l'arbre à une feuille.

Exercice 3 : Indiquez dans un tableau pour l'algorithme *J4.8* les nombres de vrais positifs et négatifs, et les nombres de faux positifs et négatifs

Exercice 4 : Cliquez avec le bouton droit dans le cadre *Result list* et choisissez *Vizualise tree* afin d'afficher la représentation graphique l'arbre de décision généré par l'algorithme *J4.8.* Donnez le résultat obtenu dans ce cas.

Exercice 5 : Indiquez dans un tableau comment seront classées les instances suivantes dont la description vous est fournie à partir de l'arbre généré :

Outlook	Humidity	Windy	Temperature	Play
Sunny	83	True	78	?
Overcast	78	False	80	?
Rainy	84	False	76	?

Exercice 6 : Affichez la répartition des instances mal et bien classées pour l'exécution de l'algorithme J4.8 (bouton droit sur le nom de l'exécution dans la zone Result list et option Vizualise classifier errors) . Sélectionnez l'attribut Outlook pour l'axe X et Windy pour l'axe Y. Notez dans un cadre les combinaisons de valeurs de Outlook et Windy pour lesquelles des instances Play=no ont été mal classées (carrés rouges).
Notez dans un cadre les combinaisons de valeurs de outlook et windy pour lesquelles des instances Play=yes ont été mal classées (carrés bleus).

Comparaison des méthodes de construction d'arbres de décision

Exercice 7 : Appliquez l'algorithme *RandomTree* sur le jeu de données *weather.arff* avec l'option *Cross validation* avec une valeur de 10 pour le paramètre *Folds*. Laissez les paramètres de l'algorithme à leur valeur par défaut.
Refaites les même question des exercices 2,3,4 et 5.
Exercice 8 : Appliquez l'algorithme *ADTree* sur le jeu de données *weather.arff* avec l'option *Cross validation* avec une valeur de 10 pour le paramètre *Folds*. Laissez les paramètres de l'algorithme à leur valeur par défaut.
Refaites les même question des exercices 2,3,4 et 5.
Exercice 9 : Comparez les résultats obtenus par application des trois algorithmes *J4.8, RandomTree et ADTree*. Notez vos observations sur les différences observées.
Exercice 10 : A partir des matrices de confusion déterminez quel est l'algorithme qui a permis d'obtenir le classifieur ayant la plus grande précision (nombre minimum d'erreurs, c-à-d de faux positifs et négatifs).