Université de Bouira
Département d'Informatique
Module : Fouille de Données
Niveau M1-ISIL

Série TP 3

Classification. Les arbres de décision.

Travail 1 – Construction et Evaluation

Weka offre une large palette d'algorithmes et de paramétrages permettant la classification, la construction de modèles, et l'évaluation. Ces derniers sont accessibles via l'onglet *Classify* de l'explorateur *Weka Explorer*.

1. Depuis l'onglet *Preprocess*, chargez le jeu de données weather.numeric.arff dans Weka.
2. Cliquez sur l'onglet *Classify*. Identifiez ses différentes sections et options de test.

Dans cette partie, vous classifieriez les données en utilisant l'algorithme **J48**, qui est une méthode d'induction **d'arbres de décision** *C4.5*. Étant donné que l'algorithme C4.5 peut gérer les attributs numériques, contrairement à l'algorithme ID3, il n'est pas nécessaire de discrétiser aucun des attributs numériques du dataset chargé.

3. Depuis la section *Classifier*, cliquez sur *Choose* puis sélectionnez l'algorithme *J48 :* Choose => weka => classifiers => trees => J48.

Avant d'exécuter l'algorithme de classification, vous devez définir les options de test afin de sélectionner le type de validation/évaluation souhaité du modèle construit.

4. Définissez les options de test à partir de la section *Test Options*. Les options de test dont vous disposez sont :

   - **Use training set** : Évalue le classificateur sur la façon dont il prédit la classe des instances sur lesquelles il a été formé.
   - **Supplied test set** : Évalue le classificateur sur la façon dont il prédit la classe d'un ensemble d'instances chargées à partir d'un fichier. En cliquant sur le bouton 'Set ...', une boîte de dialogue s'ouvre vous permettant de choisir le fichier à tester.
   - **Cross-validation** : Évalue le classificateur par validation croisée, en utilisant le nombre de folds entrés dans le champ de texte "Folds".
   - **Percentage Split** : Évalue le classificateur sur la façon dont il prédit un certain pourcentage des données, qui est retenu pour les tests. La quantité de données contenue dépend de la valeur saisie dans le champ '%'.

Dans cette partie, vous évaluerez le classifier en fonction de la façon dont il prédit 66% des données testées.

5. Cochez le bouton radio *Percentage Split* et conservez le pourcentage par défaut 66% (on apprend le modèle sur 66% des exemples et on évalue le résultat sur les exemples restants). Cliquez sur le bouton *More Options*.
6. Identifiez ce qui est affiché dans les résultats.
7. Dans *Classifier Evaluation Options*, assurez-vous que les options suivantes sont cochées : *Output model*, *Output per-class stats*, *Output confusion matrix*, *Output entropy evaluation measures*, et *Store predictions for visualization*.
8. Cliquez ensuite sur le bouton *Start* pour générer le modèle prédictif (ici l'arbre de décision).

L'arbre construit (en mode texte + taille de l'arbre et nombre de feuilles) ainsi que les résultats obtenus (taux d'erreur, matrice de confusion) s'affichent dans la fenêtre de droite.

9. Extraire depuis les résultats obtenus, le nombre de feuilles ainsi que la taille de l'arbre (i.e. le nombre de ses nœuds).
10. Donner le nombre d'instances de chaque feuille.

11. Il est possible de faire afficher l'arbre obtenu en cliquant avec le bouton droit de la souris sur la ligne « *heure – trees.J48* » (en bas à gauche) puis en choisissant *Visualize tree*. Dans la fenêtre de visualisation de l'arbre, des options sont disponibles afin de faciliter cette visualisation (*feet to screen*, *auto scale*, etc.).

12. Il est aussi possible de visualiser les erreurs de classification en cliquant cette fois sur *Visualize classifier errors*.

13. Depuis la partie résumant l'évaluation – *Evaluation on test split*, extraire le taux d'erreur obtenu sur la base d'apprentissage ?
14. Combien d'instances ont-elles étaient utilisées comme base de test ?
15. Depuis la matrice de confusion, extraire les CP, CN, FP, FN. Expliquez ce qu'ils veulent dire.
16. Quelle est la sensitivité (*TP Rate*) par rapport à la classe Yes ?
17. Dans les options du classifier J48, passez la valeur du paramètre *unprunned* à *true* (on supprime le post-élagage de l'arbre). *Start*. Quel est le taux d'erreurs obtenu sur la base d'apprentissage ? Quelle est la taille de l'arbre ?
18. Réexécuter la classification en cochant cette fois *Use training set* puis Cross-validation dans les *Test Options*.

Travail II – Utilisation du classifier J48 sur le jeu de données bank-data

1. Depuis l'onglet *Preprocess*, chargez le jeu de données bank-data-transformed.arff (Vu en TP 2) dans Weka. Cliquez sur l'onglet *Classify*.

2. Appliquez l'algorithme *J48* sur ce dataset avec une évaluation qui utilise *Use training set* comme *Test Options*.
3. Quelle est l'erreur obtenue sur la base d'apprentissage ? Quelle est la taille de l'arbre ?
4. Dans les options du classifier J48, passez la valeur du paramètre *unprunned* à *true*. Quel est le taux d'erreurs obtenu sur la base d'apprentissage ? Quelle est la taille de l'arbre ?
5. Tester de nouveau ce classifier avec et sans élagage, en prenant cette fois comme *Test Options*, *Percentage Split* à 66%. Quels sont les taux d'erreurs obtenus avec et sans élagage ? Expliquez pourquoi on obtient ces résultats.
6. Expérimenter d'autres algorithmes des arbres de décision proposés par Weka. Lesquels offrent la meilleure précision ? Pour quelles valeurs de paramètres ?

Travail III – Utilisation de la sélection d'attributs

1. Ouvrez de nouveau le dataset bank-data-transformed.arff.
2. Les exemples de cette base sont décrits par quels et combien d'attributs ?
3. Appliquer sur cette base un filtrage des attributs (dans la partie *filter* de l'onglet *Preprocess*, appuyez sur le bouton *choose*, puis sélectionnez *filters => supervised => attribute => AttributeSelection => Apply*). Combien reste-t-il d'attributs ?
4. Tester de nouveau le classifier J48 (avec élagage) en prenant comme *Test Options*, *Percentage Split* à 66%. Quel est le taux d'erreurs obtenu ? Quelle est la taille de l'arbre ?
5. Comparez ces résultats aux résultats obtenus sans filtrage.

Il est aussi possible d'exécuter une sélection d'attributs automatiquement via l'onglet *Select Attributes*. Cet onglet permet de faire de la sélection d'attributs pertinents dans un objectif de sélection. C'est ici que vous pouvez trouver un moyen automatique de déterminer le meilleur couple d'attributs pour séparer les classes.

6. Depuis l'onglet *Select Attributes*, cliquez sur *Choose* de la section *Attribute Evaluator*, puis sélectionnez ***InfoGainAttributeEval***. La méthode de recherche *Ranker* est sélectionnée afin de lister les attributs en fonction des résultats d'évaluation.
7. Cochez *Cross-validation*. Cliquez sur *Start*.

Il est possible de sauvegarder le nouveau dataset réduit en cliquant droit sur la liste des résultats (en bas à gauche – Heure-Ranker+InfoGainAttributeEval) puis sélectionner *Save reduced data*.

Travail IV

1. Refaire les mêmes étapes du Travail I et du Travail II avec le jeu de données de l'exercice 1 de la série TD n° 2. Vérifiez l'arbre obtenu.