Université de Carthage	Module : Data Mining
INSAT	Sections : GL
Département Mathématiques et Informatique	Niveau : 4ème année
Année Universitaire : 2019 - 2020	Enseignante : Sana Hamdi

TP N°1 : Découverte du logiciel Weka

Objectif :

Ce TP a pour but de vous familiariser avec Weka (*Waikato environment for knowledge analysis*), une plateforme d'algorithmes écrite en java pour la résolution de problèmes du data mining. En effet, après avoir installé Weka, ce TP va vous permettre de :

- Explorer l'interface « Explorer » de Weka
- Découvrir quelques datasets
- Construire un classificateur
- Interpréter les résultats
- Utiliser les filtres
- Visualiser les données

I- Présentation et Installation

Weka est disponible gratuitement à l'adresse www.cs.waikato.ac.nz/ml/weka, dans des versions pour Unix et Windows. Weka peut s'utiliser de plusieurs façons :

- Par l'intermédiaire d'une interface utilisateur: c'est la méthode utilisée dans ce TP.
- Sur la ligne de commande.
- Par l'utilisation des classes fournies à l'intérieur de programmes Java

- Téléchargez Weka et installez-le.
- Vous pouvez Télécharger la présentation d'Eibe Frank à http://liris.cnrs.fr/marc.plantevit/ENS/TP/weka.ppt

II- Premiers pas

Weka est maintenant installé sur votre ordinateur. Pour aller plus vite :

1. Créez un raccourci de l'application sur votre bureau

2. Copiez le répertoire data (C:\Program Files\Weka-3-8\data) sous votre répertoire Documents.
3. Après avoir lancé Weka, vous obtenez la fenêtre intitulée Weka GUI Chooser: choisissez l'Explorer. La nouvelle fenêtre qui s'ouvre alors (Weka Knowledge Explorer) présente six onglets :

- ***Preprocess :*** pour choisir un fichier, inspecter et préparer les données.
- ***Classify :*** pour choisir, appliquer et tester différents algorithmes de classification : là, il s'agit d'algorithmes de classification supervisée.
- ***Cluster :*** pour choisir, appliquer et tester les algorithmes de segmentation.
- ***Associate :*** pour appliquer l'algorithme de génération de règles d'association.
- **Select Attributes** : pour choisir les attributs les plus prometteurs.
- ***Visualize :*** pour afficher (en deux dimensions) certains attributs en fonctions d'autres.

III- Les Données

Les données sont sous un format ARFF -pour Attribute-Relation File Format-. Des exemples de données sont disponibles dans le répertoire data. Ouvrez dans un éditeur un de ces fichiers d'exemples et regardez son format.

Dans l'onglet **Preprocess**, cliquez sur **Open File** et ouvrez par exemple le fichier weather.nominal.arff : il a 14 instances et cinq attributs; les attributs sont appelés *Outlook, temperature, humidity, windy* et *play*. Le premier attribut, *outlook*, est sélectionné par défaut (vous pouvez en choisir d'autres en cliquant dessus) et n'a pas de valeurs manquantes, trois valeurs distinctes, et aucune valeur unique; les valeurs réelles sont *sunny, overcast* et *rainy*, et se produisent respectivement cinq, quatre, et cinq fois. Un histogramme en bas à droite indique la fréquence à laquelle chacune des deux valeurs de la classe, *play*, apparaît pour chaque valeur de l'attribut *outlook*. L'attribut *outlook* est utilisé car il apparaît dans la zone située au-dessus de l'histogramme, mais vous pouvez en faire un autre pour chacun des autres attributs.

Vous pouvez supprimer un attribut en cochant sa case et en utilisant le bouton Supprimer. ***All*** sélectionne tous les attributs, ***None*** ne sélectionne aucun et ***Invert*** inverse la sélection en cours. Vous pouvez annuler une modification en cliquant sur le bouton ***Undo***. Le bouton ***Edit*** fait apparaître un éditeur qui vous permet d'examiner les données, de rechercher des valeurs particulières et de les modifier, ainsi que de supprimer des occurrences et des attributs. Un

clic droit sur les valeurs et les en-têtes de colonnes ouvre les menus contextuels correspondants.

1. Ouvrez le dataset **iris.arff**
2. Décrivez-le en terminant le paragraphe suivant :

   iris.arff contient la description de …….. de spécimens d'iris de … sortes différentes. Chaque description est composée de … attributs (dimensions des …… et des ……), et d'un cinquième attribut qui est la ……..…. de cet exemple (i.e. la sorte d'iris à laquelle il appartient).

IV- Visualisation des données

Pour une première visualisation des données du dataset **iris.arff**, ouvrez ce dataset sur Weka et passez dans la fenêtre **Visualize**. Vous y voyez un ensemble de 25 graphiques (que vous pouvez ouvrir en cliquant dessus), qui représentent chacun une vue sur l'ensemble d'exemples selon deux dimensions possibles, la couleur des points étant leur classe. Sur le graphique, chaque point représente un exemple : on peut obtenir le descriptif de cet exemple en cliquant dessus. La couleur d'un point correspond à sa classe (détaillé dans la sous-fenêtre Class colour). Au départ, le graphique n'est pas très utile, car les axes représentent le numéro de l'exemple.

1. Changez les axes pour mettre la largeur des pétales en abscisse, et la longueur des sépales en ordonnées.
2. Proposez un ensemble de deux règles simples permettant de classer les exemples selon leur genre: quelle erreur commettrez-vous ? Les petits rectangles sur la droite de la fenêtre représentent la distribution des exemples, pour l'attribut correspondant, par rapport à l'attribut (ou la classe) codé par la couleur. En cliquant du bouton gauche sur un de ces rectangles, vous le choisissez comme axe des X, le bouton droit le met sur l'axe des Y.
3. En mettant la classe sur l'axe des X, quels sont à votre avis les attributs qui, pris seuls, permettent le mieux de discriminer les exemples? Si les points sont trop serrés, le potentiomètre Jitter, qui affiche les points "à peu près" à leur place, vous permet de les visualiser un peu plus séparément: cela peut être utile si beaucoup de points se retrouvent au même endroit du plan.

V- Construction d'un arbre de décision

Pour voir ce que l'algorithme C4.5 (qui produit un modèle de type arbre de décision) fait avec un dataset, nous allons utiliser l'algorithme J48, l'implémentation open source de C4.5 par Weka.

1- Ouvrez le fichier **weather.nominal.arff**

2- Dans l'onglet **Classify**, choisissez le classificateur **J48** sous la section **trees**

3- Cliquez sur **Start**.

La figure 1 ci-dessous, montre l'illustration des résultats après avoir analysé les données météorologiques.

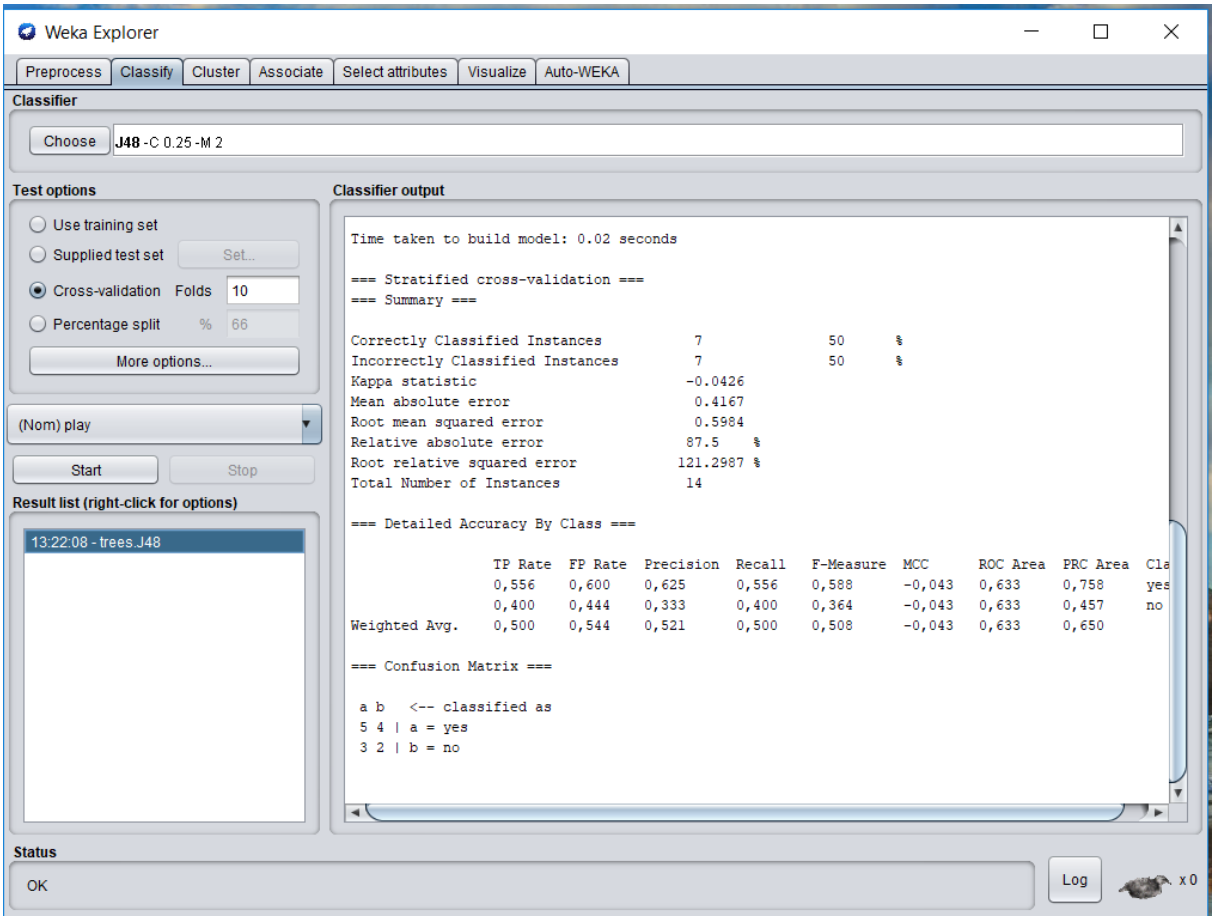

Figure 1 : Utilisation de J48

4- Faites la même chose pour le dataset glass.arff

VI- Observation des résultats

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     weather.symbolic
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------
outlook = sunny
|   humidity = high: no (3.0)
|   humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)

Number of Leaves  :         5

Size of the tree :

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           7               50      %
Incorrectly Classified Instances         7               50      %
Kappa statistic                          -0.0426
Mean absolute error                       0.4167
Root mean squared error                   0.5984
Relative absolute error                  87.5    %
Root relative squared error             121.2987 %
Total Number of Instances                14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,556     0,600     0,625      0,556     0,588      -0,043   0,633     0,758     yes
               0,400     0,444     0,333      0,400     0,364      -0,043   0,633     0,457     no
Weighted Avg.  0,500     0,544     0,521      0,500     0,508      -0,043   0,633     0,650

=== Confusion Matrix ===

 a b   ← classified as
 5 4 | a = yes
 3 2 | b = no
```

Figure 2 : Output de l'algorithme J48

1- Dessiner manuellement l'arbre de décision proposée par l'algorithme J48 appliqué sur **weather.nominal.arff**
2- Vérifier avec le **Tree Visualizer** (clic droit sur le modèle concerné dans la liste des résultats)

VII- Utilisation des filtres :

Le fichier bank-data.csv contient des données extraites d'un recensement de la population américaine. Le but de ces données est initialement de prédire si quelqu'un gagne plus de 50.000 dollars par an. On va d'abord transformer un peu les données :

1. Transformation des données

Récupérer le fichier bank-data.csv, revenez à la fenêtre **Preprocess**.

- Tout d'abord ouvrez le fichier bank-data.csv
- La sous-fenêtre **Filters** vous permet de manipuler les filtres. Le fonctionnement général est toujours le même :

– Vous choisissez un ensemble de filtres, chaque filtre, avec ces options, étant choisi dans le menu déroulant du haut de la sous-fenêtre, puis ajouté à la liste des filtres par la commande **Add**.
– On applique les filtres avec la commande **Apply** Filters.
– On peut alors remplacer le fichier précédemment chargé par les données transformées.
– Ce fichier devient alors le fichier de travail.
– Le bouton **Save** sauvegarde ces données transformées dans un fichier.

2. Sélection des attributs

Les données comportent souvent des attributs inutiles : numéro de dossier, nom, date de saisie… Il est possible d'en supprimer « à la main », à condition de connaître le domaine. On peut aussi lancer un algorithme de data mining, et regarder les attributs qui ont été utilisés: soient ceux-ci sont pertinents, et il est important de les garder, soient ils sont tellement liés à la classe qu'à eux seuls ils emportent la décision (pensez à un attribut qui serait la copie de la classe). Weka a automatisé cette recherche des attributs pertinents dans le filtre **AttributeSelection**, qui permet de définir les attributs les plus pertinents selon plusieurs méthodes de recherche (*search*), en utilisant plusieurs mesures possibles de la pertinence d'un attribut (*eval*).

Ici l'attribut id est une quantité qu'on peut ignorer pour la fouille : supprimez-le !

3. Discrétisation

Certains algorithmes ont besoin d'attributs discrets pour fonctionner, d'autres n'acceptent que des attributs continus (réseaux de neurones, plus proches voisins). D'autres encore acceptent indifféremment des attributs des deux types. Weka dispose de filtres pour discrétiser des valeurs continues. Le filtre **Discretize** permet de rendre discret un attribut continu et ceci de plusieurs façons :

– En partageant l'intervalle des valeurs possibles de l'attribut en intervalles de taille égale.

– En le partageant en intervalles contenant le même nombre d'éléments.

– En fixant manuellement le nombre d'intervalles (bins).

– En laissant le programme trouver le nombre idéal de sous intervalles.

Ici il y a plusieurs attributs numériques : "*children*", "*income*", "*age*".

1. Discrétiser *age* et *income* en utilisant le filtre Weka et en forçant le nombre d'intervalles à 3. Sauver le fichier transformé par exemple dans bank1.arff.

2. L'attribut *children* est numérique mais ne prend que 4 valeurs : 0,1, 2,3 ; pour le discrétiser, on peut soit utiliser le filtre, soit le faire à la main dans le fichier arff.

Remarque: si vous éditez directement le fichier, vous pouvez en profiter pour rendre les données plus lisibles, par exemple en traduisant le nom des attributs, en donnant des noms aux intervalles obtenus par la discrétisation...