TP1 Fouille de Données
Prise en main de Weka

23 janvier 2017
Université de Rouen - Lina Soualmia

1 Préambule

Le travail se fait seul. Les réponses aux questions sont à envoyer à fdd.m1.rouen@gmail.com avec comme objet : [FdD] TP1 GR(2, 3, ou 4) NOM sous la forme d'un fichier (pdf, txt). N'oubliez pas d'indiquer le numéro de la question et répondez-y de manière claire. Lisez **TOUTES** les indications. Le fichier est à envoyer en fin de séance au plus tard.

2 Analyse exploratoire avec Weka

Environnement de travail :
Système WEKA 3.6.x ;
Java

2.1 Introduction et prise en main

Le système Weka est développé à l'Université de Waikato, Nouvelle Zélande. Il permet de pré-traiter des données, de les analyser à l'aide d'une méthode de data mining et d'afficher le modèle résultant et ses performances. Weka est entièrement développé en Java. Il est diffusé sous licence publique GNU. Les ressources nécessaires à l'installation et à l'utilisation du système sont disponibles à l'adresse suivante : http://www.cs.waikato.ac.nz/ml/weka/.

Dans Weka, chaque méthode de transformation, de sélection d'attributs, d'apprentissage, de clustering ou de découverte d'associations est implémentée par une classe Java. La documentation des classes est accessible à partir du fichier `packages.html`.

Weka traite des données au format ARFF (Attribute Relation Format File) ou CSV (Coma Separated Values).

2.2 Le format ARFF

Les fichiers ARFF sont des fichiers texte d'extension `.arff`. L'entête du fichier définit les propriétés du jeu de données :

- @relation : nom du jeu de données. Apparaît en tête de fichier.
- @attribute : nom et format (type ou liste de valeurs) d'un attribut. Une ligne pour chaque attribut.
- @data : marqueur de début des données. Apparaît une seule fois avant les lignes de données. Chaque ligne de données représente un objet et contient la liste des valeurs des attributs séparées par des virgules. Par exemple, le fichier `weather.arff` est affiché ci-dessous :

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Les fichiers peuvent être créés et modifiés avec des éditeurs, des outils de bureautique ou des programmes simples.

3 Exercices

Weka est normalement installé sur les postes de la salle Info :

- ouvrez un terminal et tapez "weka"
- ou bien dans une fenêtre de commandes lancez la commande : `java -jar weka.jar`

Cela fait apparaître une fenêtre de sélection du mode d'exécution comportant 4 options :

- `Simple CLI` : mode d'exécution textuel. Les classes Java sont invoquées par ligne de commande.
- `Explorer` : client graphique. Les commandes sont définies par l'interface graphique.
- `Experimenter` : pour sauvegarder une suite d'exécutions.
- `KnowledgeFlow` : programmation par définition d'un flux de données.

Lancez l'interface graphique Explorer et la fenêtre de commande s'ouvre. Elle comporte 6 onglets :

- `Preprocess` : charger/filtrer un jeu de données.
- `Classify` : appliquer un algorithme de classification.
- `Cluster` : appliquer un algorithme de segmentation.
- `Associate` : appliquer l'extraction de règles d'association.
- `Select attributes` : appliquer la caractérisation analytique.
- `Vizualise` : visualisation en 2 dimensions de la répartition des valeurs des attributs.

Les onglets `Classify, Cluster, Associate`, et `Select attributes` correspondent chacun à une fonctionnalité de data mining. Le processus est identique pour toutes les fonctionnalités :

- on clique sur le bouton `Choose` et la liste des algorithmes s'affiche,
- on choisit un algorithme (une classe Java) dans la liste,
- on clique sur le nom de l'algorithme et la fenêtre des paramètres s'affiche,
- on définit les autres paramètres affichés sur l'onglet,
- on lance l'exécution et le résultat s'affiche dans le cadre de droite.

3.1 Question 1

Cliquez sur le bouton `Open file` et chargez le jeu de données `weather.arff` décrit ci-dessus. Ce jeu de données décrit en fonction des conditions climatiques – couverture du ciel, température, taux d'humidité et vent – s'il est possible ou non de jouer au tennis (attribut `play`). Le nom du jeu de données et le nombre d'instances et d'attributs apparaissent dans le cadre à gauche. Les informations affichées sont le nom de la relation (jeu de données chargé), le nombre d'instances (lignes) et le nombre d'attributs.

Déterminez le nombre d'instances et le nombre d'attributs du jeu de données `weather.arff`.

3.2 Question 2

La liste des attributs apparaît dans le cadre `Attributes`. Lorsqu'un attribut est sélectionné, en cliquant dessus, ses caractéristiques apparaissent dans le cadre `Selected attribute` à droite. Les informations affichées sont les suivantes :

- nom de l'attribut,
- type de l'attribut,
- nombre de valeurs manquantes,
- nombre de valeurs distinctes,
- nombre de valeurs uniques.

Sont également affichées des informations dépendant du type de l'attribut. S'il s'agit d'un attribut numérique :

- valeur minimale,
- valeur maximale,
- moyenne,
- déviation standard

et s'il s'agit d'un attribut symbolique (nominal) :

- liste des valeurs (appelées modalités)
- nombre d'instances possédant chaque valeur.

Sélectionnez successivement chaque attribut et indiquez le nombre de valeurs pour chacun des attributs nominaux.

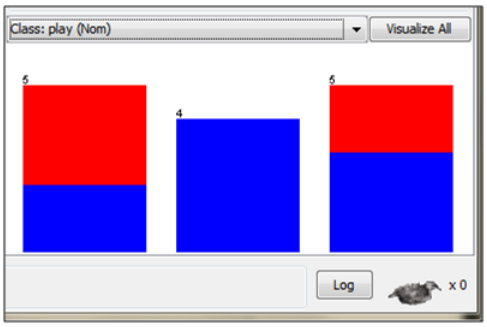
3.3 Question 3

Indiquez les valeurs minimales et maximales des attributs numériques.

La partie inférieure du cadre de droite permet d'observer la répartition des valeurs de l'attribut sélectionné en fonction d'un autre attribut (affiché dans la zone de choix `Colour`) dans un histogramme bleu et rouge. Chaque colonne de l'histogramme correspond à une valeur de l'attribut sélectionné et chaque couleur représente une valeur de l'attribut choisit dans la zone `Colour`.

Par exemple, dans la figure ci-dessous la colonne de gauche représente le nombre d'instances possédant la valeur `outlook=sunny`, celle du centre la valeur `outlook=overcast` et

celle de droite la valeur `outlook=rainy`. L'attribut de la zone `Colour` est `play` et la partie rouge représente les instances `play=no` et la partie bleue les instances `play=yes`.

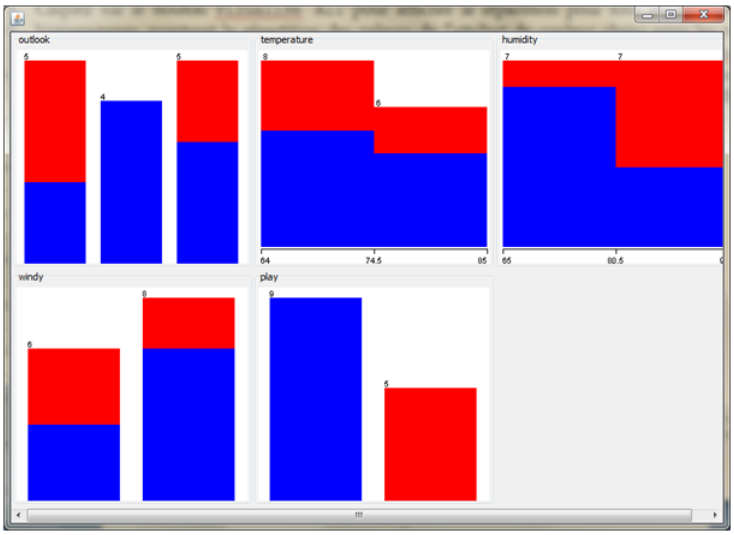


3.4 Question 4

Sélectionnez l'attribut `outlook` et affichez l'attribut `play` pour la couleur.

Quelle est la valeur de l'attribut `outlook` qui permet de prédire à coup sûr que l'on va jouer, c.à.d. qui est toujours associée à `play=yes` ?

Cliquez sur le bouton `Vizualise All` pour afficher la répartition pour tous les attributs. Les histogrammes montrant la répartition des valeurs de l'attribut de couleur choisi pour les valeurs de chaque attribut apparaissent. Pour les attributs numériques tels que `temperature` et `humidity` ci-dessous, l'intervalle de valeurs est divisé en 2 parties égales et la répartition des instances selon les valeurs de l'attribut de couleur ainsi que le nombre d'instances sont affichées pour les 2 parties.



3.5 Question 5

Quelle est la valeur de `play` la plus fréquente pour les instances `Humidity` ϵ[65, 80.5[et pour les instances `Humidity` ϵ [80.5, 96].

3.6 Question 6

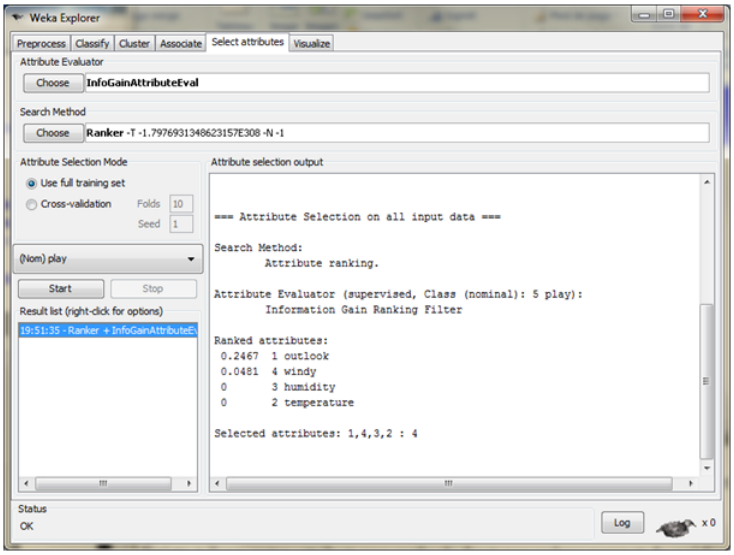
Existe-t-il un autre attribut que `outlook` dont une valeur permet de prédire à coup sûr que l'on va jouer (toujours associée à `play=yes`) ou que l'on ne va pas jours (toujours associée à `play=no`) ?

Les méthodes de l'onglet `Select attributes` permettent de déterminer l'utilité de chaque attribut pour la prédiction de la valeur d'un autre attribut. Cette étape permet d'avoir un premier aperçu général des données. La fenêtre se divise en plusieurs zones :

- `attribute evaluator` : algorithme utilisé. Cliquer sur le nom permet de définir ses paramètres.
- `Search method` : méthode de recherche utilisée.
- Liste déroulante (`Nom`) : sélection de l'attribut dont on veut prédire la valeur (ex :play)
- `Attribute selection mode` : choix du jeu d'apprentissage.
- `Attribute selection output` : résultat de l'algorithme.
- `Result list` : historique des résultats des exécutions. Cliquer sur une ligne avec le bouton droit affiche un menu d'options
- `Status` : état actuel (occupé/inactif). Le nombre d'exécutions en cours est affiché à droite.
- `Log` : historique des messages d'information.

3.7 Question 7

Dans la zone `Attribute selection output` ci-dessous, le résultat est une liste ordonnée des attributs dans l'ordre décroissant de leur utilité pour distinguer les valeurs de l'attribut `play` : le premier est celui qui permet le mieux de distinguer les valeurs `play=yes` et `play=no`, puis sont affichés le second le plus utile, le troisième ... etc. Pour chaque attribut une mesure numérique de son utilité pour la prédiction des valeurs de l'attribut choisi est affichée.



Exécutez une analyse pour chaque algorithme ci-dessous (InfoGain ...etc.) en sélectionnant pour le paramètre `Search method` la méthode `ranker`. Notez l'ordre des attributs (par leur numéro) dans un tableau à 2 colonnes (col1 : Algorithme ; col2 : liste des attributs par ordre décroissant d'utilité)

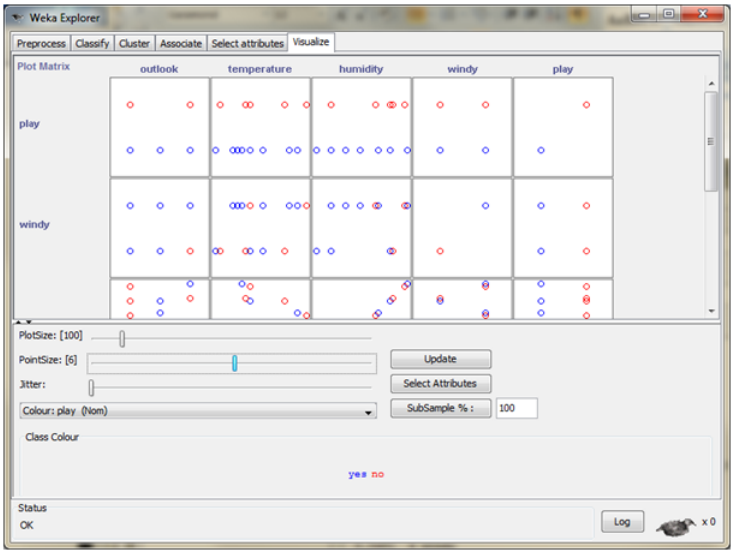
`InfoGainAttributeEval` : ?
`ReliefAttributeEval` : ?
`GainRatioAttributeEval` : ?
`SymetricalUnsertAttributeEval` : ?
`OneRAttributeEval` : ?
`ChiSquaredAttributeEval` : ?

3.8 Question 8

Comparez les résultats obtenus et identifiez l'ordre des attributs le plus fréquent (par leur nom) ?

L'onglet `Visualize` permet de représenter les instances comme des points dans les espaces bi-dimensionnels des données. Chaque dimension (axes X et Y) est un attribut et il y a donc un graphique pour chaque couple d'attributs. Chaque instance est un point de couleur rouge ou bleu en fonction de la valeur d'un troisième attribut qui est en général un attribut de classe (par exemple `play`). Ce mode de visualisation permet de se faire une première idée de la répartition des données. Les options sont :

- `PlotSize` : taille des graphiques
- `PointSize` : taille des points bleus et rouges qui représentent les instances
- `Jitter` : déplacer les points d'une distance aléatoire afin de distinguer les points confondus
- `Colour` : choix de l'attribut représenté dans l'espace bi-dimensionnel (`play` dans la figure ci-dessous)
- `Update` : cliquer pour mettre à jour l'affichage après modification des paramètres ci-dessous
- `Select attributes` : permet de choisir les attributs utilisés comme dimensions
- `SubSample` % : taille de l'échantillon affiché (pour les très grands jeux de données)
- `Class Colour` : permet de changer la couleur des points en cliquant sur `yes` et `no`.

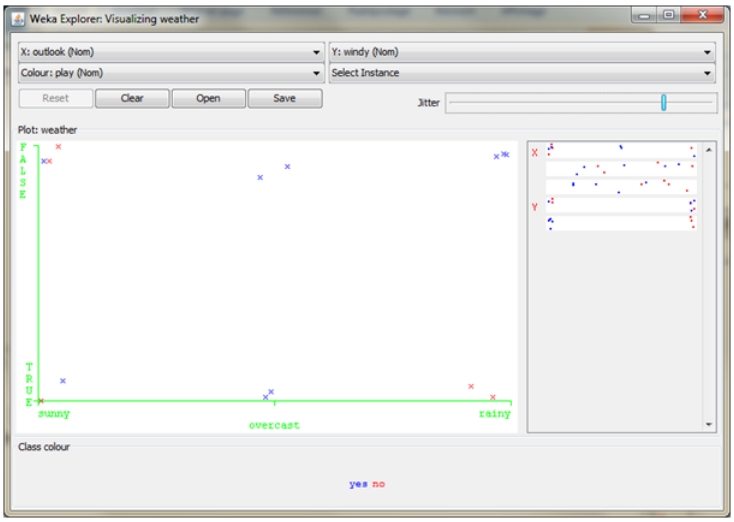


Cliquez sur un des graphiques afin d'ouvrir la fenêtre d'affichage bi-dimensionnel détaillé telle que représentée dans la figure ci-dessous.

Les options sont :

- X : choix de l'attribut représenté sur l'axe horizontal (p. ex. : outlook)
- Y : choix de l'attribut représenté dans l'espace bi-dimensionnel (p. ex. : Windy)
- Colour : choix de l'attribut représenté dans l'espace bi-dimensionnel (p. ex. : Play)

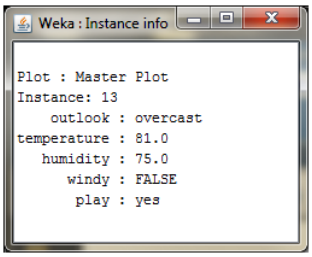
Cliquez sur un point et les caractéristiques de l'instance qu'il représente s'affichent.



3.9 Question 9

Choisissez l'attribut `outlook` pour l'axe X, `windy` pour l'axe Y et `play` pour la couleur. Indiquez le nombre d'instances `play = yes` et `play=no` pour chaque combinaison de valeurs de `outlook` et `windy`.

Afin d'afficher les informations sur une instance particulière, vous pouvez cliquer dessus. Une fenêtre apparaît indiquant pour chaque attribut la valeur de cette instance. (Instance : 13 : numéro de la ligne).



3.10 Question 10

Cliquez sur l'instance correspondant à la valeur `play=no` pour `outlook=sunny` et `windy=true`. Quelles sont ses autres caractéristiques ?