Dr. Laura E. Raileanu
Fabien Dutoit

Analyse de données

Laboratoire Nº 2

A. Introduction

Ce laboratoire a pour but de mettre en pratique les concepts d'analyse de données vus en cours :

- Arbres de décisions;
- Clustering;
- Règles d'associations.

Ce laboratoire réutilise la base de données utilisée lors du premier laboratoire. Nous utiliserons le logiciel libre *Weka*[1] qui regroupe une collection d'outils pour l'analyse de données.

A la fin du laboratoire, vous remettrez votre rapport, au format *PDF*, qui contiendra les résultats obtenus aux différentes questions. Merci de bien vouloir rendre votre travail sur la page *CyberLearn* du cours. En cas de problème lors de la remise, vous pouvez l'envoyer par e-mail à fabien.dutoit@heig-vd.ch. Dernier délai pour le rendu :

Dimanche 27.11.2016 à 23:55

[1] Weka : http://www.cs.waikato.ac.nz/ml/weka/ et http://weka.wikispaces.com/

B. Installation et configuration

1. Installation de *Weka*

Comme pour le laboratoire précédent, celui-ci a été préparé et testé avec une machine sous *Windows 7*, tous les logiciels utilisés existent cependant sous *Mac OS* et *Linux*.

Vous trouverez les exécutables sur la page Internet du logiciel[2], son installation ne devrait pas vous poser de problème particulier.

2. Pilotes pour *MySQL*

Weka est un programme *Java*, il faudra donc lui fournir le pilote nécessaire pour qu'il puisse se connecter à la base de données. Ce pilote se trouve sur le site de *MySQL*[3].

Pour que *Weka* puisse le trouver, il faudra ajouter l'emplacement du fichier `mysql-connector-java-5.*.*-bin.jar` dans la variable d'environnement *CLASSPATH* (à créer ou compléter).

Si *Weka* était déjà en cours d'exécution, il faudra le redémarrer pour que le pilote soit pris en compte.

Sous *linux*, l'exécution d'un jar à partir de la commande `java -jar` ne tient pas compte de la variable *CLASSPATH* du système.

3. Connexion à *MySQL*

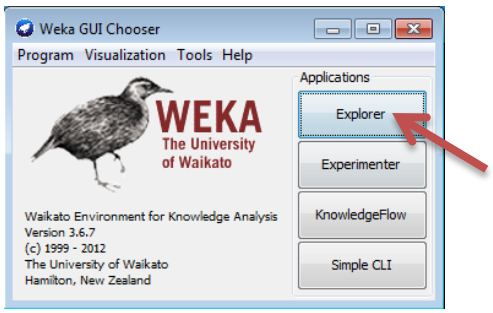
Pour que *Weka* puisse se connecter à la base de données, il faudra placer un fichier de configuration *DatabaseUtils.props* dans le dossier de l'utilisateur (`%USERPROFILE%` sous Windows). Si Weka ne le détecte pas, vous pouvez aussi le placer dans le dossier d'installation de Weka.
Habituellement un exemple de ce fichier se trouve dans l'archive *weka.jar*, pour des raisons de simplification, nous vous en fournissons une version adaptée. Il vous suffira juste de configurer l'URL de connexion à *MySQL* :

- Modifiez l'URL pour qu'elle corresponde à votre installation de *MySQL*.
-

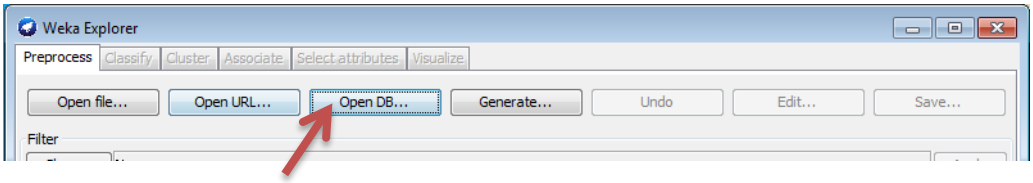
```
jdbcURL=jdbc:mysql://<host>:<port>/<dbname>?user=<username>&password=<password>
```

[2] Downloading Weka : http://www.cs.waikato.ac.nz/ml/weka/downloading.html
[3] Downloading Connector/J : http://www.mysql.com/downloads/connector/j/

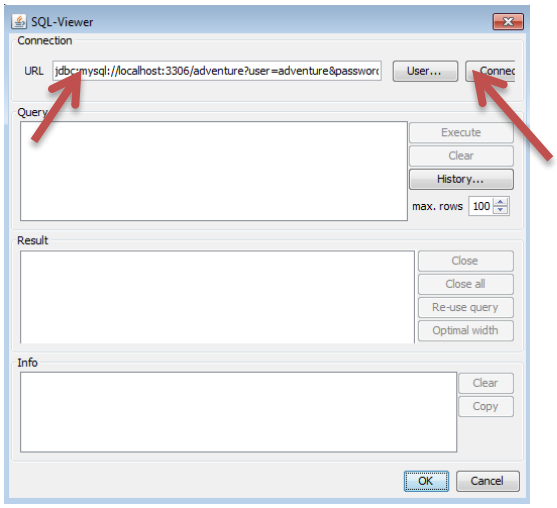
Vous pouvez maintenant ouvrir *Weka*, il est aussi possible d'ouvrir à la place *Weka (with console)* qui vous permettra de débugger plus facilement :



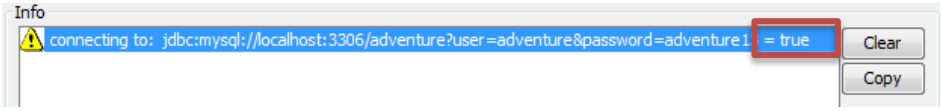
Nous allons maintenant tester la connexion sur la base de données :



Vous devriez retrouver l'URL du fichier de configuration dans la fenêtre suivante :



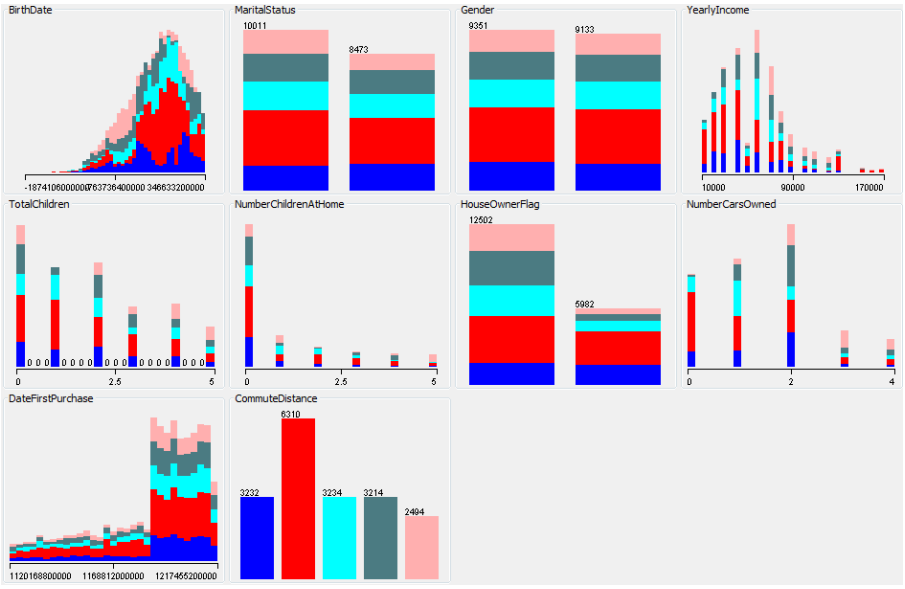
Si c'est le cas vous pouvez essayer de vous connecter, en cas de succès vous verrez s'afficher cette ligne dans les infos :



Vous pourrez ensuite tester une requête, par exemple :

```
SELECT * FROM dimcustomer
```

Si les premières lignes de table s'affichent, vous pouvez cliquer sur OK et passer à la suite. Cela vous ramènera à l'écran précédent où cette fois-ci vous aurez une liste avec tous les attributs de la table et une analyse statistique. Vous vous rendrez vite compte que certains champs, comme les clés ou les numéros de téléphones, n'apportent pas grand-chose. Nous avons néanmoins quelques courbes intéressantes :



En cas de difficulté de connexion à la base de données MySQL, vous pouvez utiliser à la place le fichier de données à importer, `offline_data.arff` à dé-zipper et ouvrir avec *Weka*.

C. Analyse de données

Pour les points suivants nous allons garder uniquement les attributs suivants :

- MaritalStatus;
- Gender;
- YearlyIncome;
- TotalChildren;
- EnglishEducation;
- EnglishOccupation;
- HouseOwnerFlag;
- NumberCarsOwned.

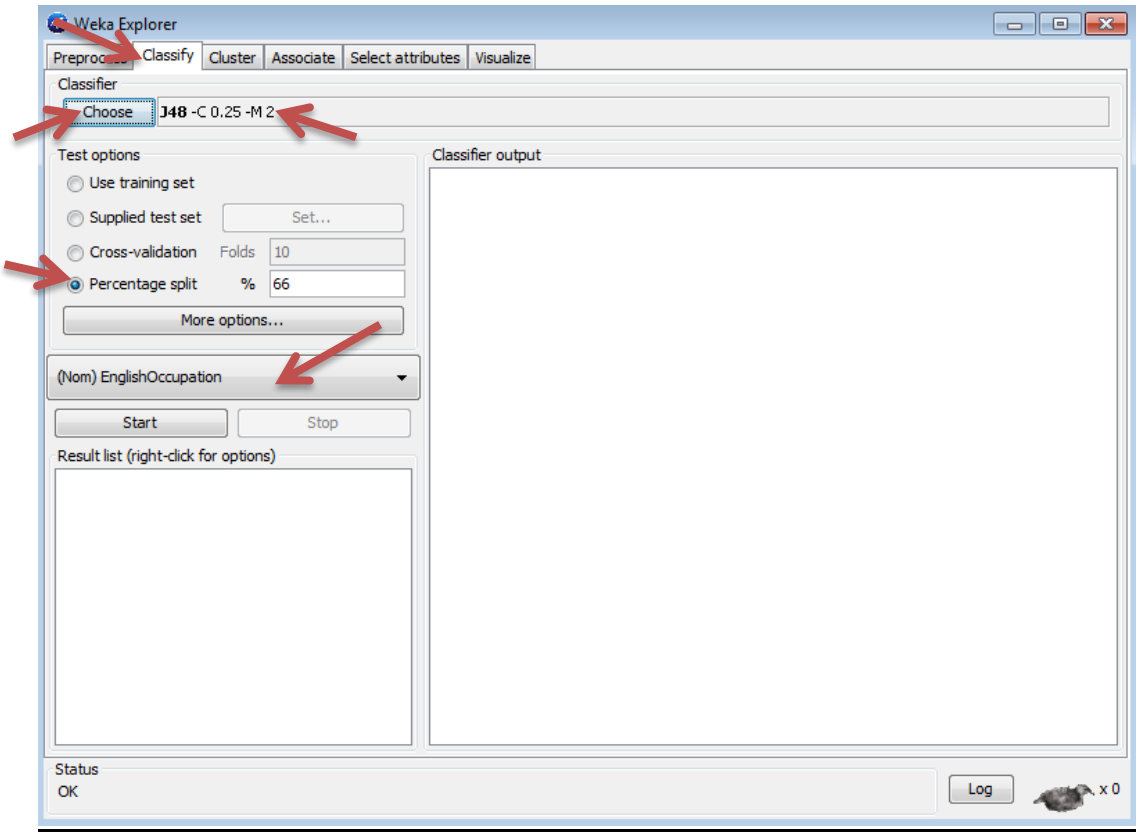
N'hésitez pas à faire un clic droit sur un des résultats pour obtenir des représentations graphiques de votre analyse.

Attention, certains algorithmes peuvent prendre plusieurs minutes avant de converger.

1. Arbres de décisions

- Une fois les attributs superflus retirés, allez dans l'onglet *Classify*. Choisissez un algorithme de type arbre, ses paramètres, le paramètre de test (split de préférence) ainsi que l'attribut à prédire: *EnglishOccupation*.

Vous pourrez ensuite lancer la classification.



Essayez plusieurs algorithme et paramètres. Dans votre rapport, reproduisez l'arbre obtenu lors d'une exécution, indiquez les paramètres choisis, et commentez une partie intéressante de l'arbre ainsi que le taux d'erreur obtenu.

2. Clustering

L'onglet *Cluster* est très semblable au précédent, la principale différence est que cette fois-ci tous les attributs deviennent des variables d'entrées, il n'y aura donc plus de prédiction.

Comme pour le point précédent, essayez plusieurs algorithmes avec plusieurs paramètres. Choisissez une exécution qui vous semble intéressante et reportez-la dans votre rapport (algorithme, paramètres et les clusters) et discutez des caractéristiques d'un des clusters.

3. Règles d'associations

Weka ne supporte pas les entrées numériques pour les tables associatives, nous allons donc devoir changer les types des attributs *NumberCarsOwned*, *YearlyIncome* et *TotalChildren*.

Pour cela, retournez dans l'onglet *Preprocess* où vous sélectionnerez les 3 variables à modifier. Dans la partie *Filter*, choisissez le filtre :

```
unsupervised/attribute/NumericToNominal
```

que vous appliquerez.

Une fois ceci réalisé, allez dans l'onglet *Associate* et comme pour les points précédents, testez plusieurs algorithmes (principalement Apriori, FilteredAssociator, PredictiveApriori et Tertius) avec plusieurs choix de paramètres. Dans le rapport, indiquez l'exécution qui vous semble la plus intéressante (algorithme, paramètres et max. env. 10 règles) et discutez de la pertinence de 3 de ces règles.