Weka : Présentation

1 Présentation

Weka (Waikato Environment for Knowledge Analysis) est un ensemble d'outils permettant de manipuler et d'analyser des fichiers de données, implémentant la plupart des algorithmes d'intelligence artificielle, dont les arbres de décision et les réseaux de neurones.

Il est écrit en java, disponible sur le web [1] , et s'appuie sur le livre

Data Mining, practical machine learning tools and techniques with Java implementations
Witten & Frank
Editeur : Morgan Kauffman

Il se compose principalement :

- De classes Java permettant de charger et de manipuler les données.
- De classes pour les principaux algorithmes de classification supervisée ou non supervisée.
- D'outils de sélection d'attributs, de statistiques sur ces attributs.
- De classes permettant de visualiser les résultats.

On peut l'utiliser à trois niveaux :

- Via l'interface graphique, pour charger un fichier de données, lui appliquer un algorithme, vérifier son efficacité.
- Invoquer un algorithme sur la ligne de commande.
- Utiliser les classes définies dans ses propres programmes pour créer d'autres méthodes, implémenter d'autres algorithmes, comparer ou combiner plusieurs méthodes.

2 Liens et ressources

Le site de Weka contient :

- La version courante (évolution rapide).
- Un lien vers la javadoc (pas tout à fait à jour).
- Quelques tutoriaux.
- La liste de diffusion répond aux questions de tout type concernant Weka (y compris les questions de débutants ...)

La javadoc en local : `/usr/local/weka-3-4/doc`

La FAQ locale : `www.lifl.fr/~decomite/weka` (... si vous avez des questions ...)

1. `www.cs.waikato.ac.nz/ml/weka`

3 Installation, initialisation

`Weka` est installé dans les salles TP. Si vous voulez l'employer chez vous :

- Charger l'archive zip à partir du site de `Weka`
- Décompressez-la.
- C'est tout....
- Les classes sont dans `weka.jar`
- Les sources dans `weka-src.jar`
- `Tutorial.pdf` est une présentation assez poussée des fonctionnalités de `Weka` .

4 Interface graphique

Elle se lance par `java -jar $WEKAHOME/weka.jar`
(à condition d'avoir lancé `setenv WEKAHOME /usr/local/weka-3-4`).
On a alors le choix entre :

Simple CLI Un interpréteur de ligne de commande.

Explorer une interface graphique.

Experimenter Un outil, pas très bien décrit dans la doc, pour tester des schémas de fouille.

Choisissez l'Explorer, on obtient un écran ressemblant à la figure 1.

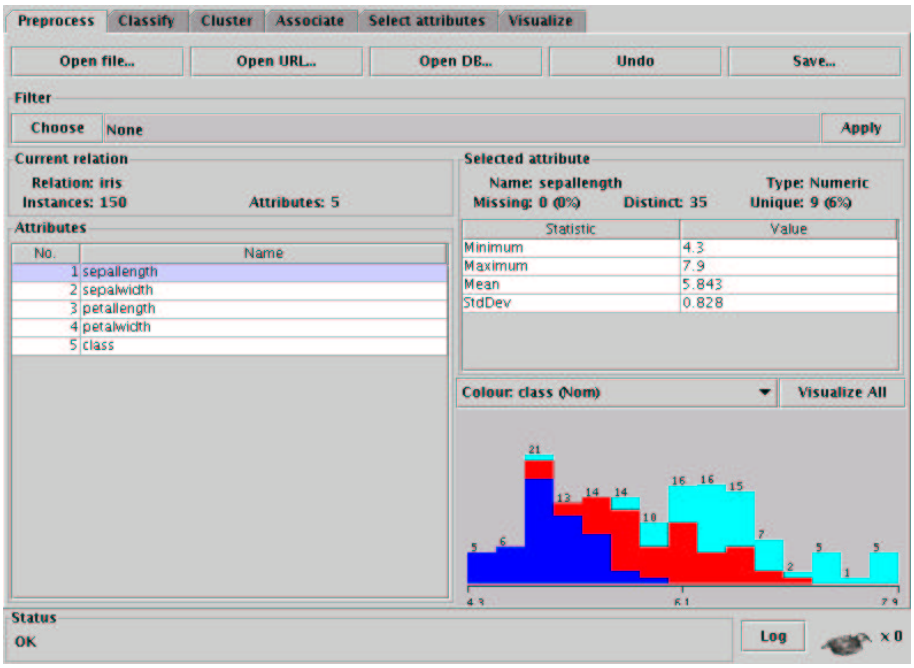

FIG. 1 – *La fenêtre principale de l'Explorer*

Les six onglets correspondent soit à des étapes du processus d'apprentissage, soit à des classes d'algorithmes de classification (supervisée ou non):

Preprocess : La saisie des données, l'examen et la sélection des attributs, les transformations d'attributs.

Classify : Les méthodes de classification.

Cluster : Les méthodes de segmentation (clustering).

Associate : Les règles d'association.

Select attributes : L'étude et la recherche de corrélations entre attributs.

Visualize : représentations graphiques des données.

Les onglets `Cluster`,`Associate` *et* `Select attributes` *ne sont pas décrits dans ce TP.*

4.1 Preprocess

Un exemple de début de fichier est donné sur la figure 2.

Ce sous-menu permet de charger un fichier de données. On peut charger un fichier au format spécifique de Weka, le format `arff` , qui se compose :

- du nom de la relation : `@relation 'labor-neg-data'`
- de la liste des noms d'attributs (`@attribute 'duration' real` ...) et du type de valeurs :
  - nominale : `@attribute 'cost-of-living-adjustment' {'none','tcf','tc'}`
  - réelle : `@attribute 'working-hours' real` et
- de la liste des instances : chaque ligne représente une description, par la liste des valeurs de chacun de ses attributs. Une valeur manquante est remplacée par un point d'interrogation.

```
@relation 'labor-neg-data'
@attribute 'duration' real
@attribute 'wage-increase-first-year' real
@attribute 'wage-increase-second-year' real
@attribute 'wage-increase-third-year' real
@attribute 'cost-of-living-adjustment' {'none','tcf','tc'}
@attribute 'working-hours' real
@attribute 'pension' {'none','ret_allw','empl_contr'}
@attribute 'standby-pay' real
@attribute 'shift-differential' real
@attribute 'education-allowance' {'yes','no'}
@attribute 'statutory-holidays' real
@attribute 'vacation' {'below_average','average','generous'}
@attribute 'longterm-disability-assistance' {'yes','no'}
@attribute 'contribution-to-dental-plan' {'none','half','full'}
@attribute 'bereavement-assistance' {'yes','no'}
@attribute 'contribution-to-health-plan' {'none','half','full'}
@attribute 'class' {'bad','good'}
@data
1,5,?,?,?,40,?,?,2,?,11,'average',?,?,'yes',?,'good'
2,4.5,5.8,?,?,35,'ret_allw',?,?,'yes',11,'below_average',?,'full',?,'full','good'
?,?,?,?,?,38,'empl_contr',?,5,?,11,'generous','yes','half','yes','half','good'
3,3.7,4,5,'tc',?,?,?,?,'yes',?,?,?,?,'yes',?,'good'
```

FIG. 2 – *Le début d'un fichier ARFF*

On peut aussi charger un fichier au format CSV (`Open File`), ou encore des données à partir d'une requête SQL (`Open DB`), ou sur le web (`Open URL`).

Chargez un fichier arff (vous en trouverez dans `/usr/local/weka-3-4/data/`). Si tout se passe bien, vous voyez apparaître la liste des attributs à gauche de la fenêtre. Cliquer sur l'un d'eux affiche un certain nombre de statistiques à droite :

- Les valeurs minimales, maximales et moyennes pour les attributs numériques.
- Les répartitions par valeur pour les attributs nominaux.

Dans les deux cas, sont calculés le nombre de valeurs manquantes, de valeurs distinctes et de valeurs uniques.

Sous ces statistiques élémentaires, un petit graphique indique la répartition des exemples pour l'attribut courant, sous forme d'un histogramme où la couleur indique la proportion d'éléments de chaque classe dans chaque colonne.

Le bouton `Visualize All` permet de voir tous les histogrammes en même temps, ce qui permet, avec un peu d'expérience, de se faire une idée de la répartition des données par classe et par attribut.

On peut définir des *filtres* qui modifieront le fichier de données :

- calculer de nouveaux attributs (AddExpression).
- modifier le type d'un attribut (NumericToBinary,NumericTransform,StringToNominal ...)
- remplacer les valeurs manquantes (ReplaceMissingValues).
- supprimer des instances selon un certain critère (enlever si mal classé : RemoveMisclassified, enlever un certain pourcentage d'exemples : RemovePercentage...).
- supprimer des attributs.

Certains des noms de filtres sont explicites, certains sont documentés ... mais pas tous. La javadoc de `Weka` peut vous apporter les informations qui vous manquent.

Pour utiliser un ou des filtres :

- Choisissez-le avec `Choose`.
- Cliquez ensuite sur le bouton central pour fixer ces paramètres.
- Appliquer (bouton `Apply`)

4.2 Classify

Sous l'onglet `Classify`, choisissez comme classifieur `weka.trees.J48` : c'est la version `Weka` de c4.5 (arbres de décision).

Choisissez 'Use training set' pour créer un arbre standard, puis 'start' pour lancer la construction de l'arbre : la fenêtre 'Classifier output' contient les résultats du calcul : description de l'arbre, taux d'erreurs, matrice de confusion ...

On peut visualiser l'arbre de décision en cliquant du bouton droit, dans la fenêtre 'Result list', sur la dernière commande effectuée.

Les réseaux de neurones sont dans `weka.classifiers.functions.MultilayerPerceptron`.

5 Ligne de commande

Les variables d'environnements WEKAHOME et CLASSPATH doivent être initialisées. La séquence de lancement à utiliser est la suivante :

```
setenv WEKAHOME /usr/local/weka-3-4
setenv CLASSPATH $CLASSPATH:$WEKAHOME/weka.jar
```

Tous les algorithmes implémentés dans `Weka` possèdent une méthode `main`, et peuvent donc être appelés depuis la ligne de commande. Par exemple, on peut invoquer la construction d'arbres de décision par :

```
java -cp $WEKAHOME/weka.jar weka.classifiers.trees.J48 -t iris.arff
```

où `iris.arff` est un fichier d'apprentissage [2].

Toutes les options utilisables dans l'interface graphique sont accessibles sur la ligne de commande, le fichier `/usr/local/weka-3-4/Tutorial.pdf` en fournit la syntaxe ...

2. Des exemples de fichiers se trouvent dans le répertoire `/usr/local/weka-3-4/data`