Université Mohamed Boudiaf de M'sila **Resp.Module: Dr. Said GADRI**

Faculté des mathématiques et de l'informatique

Department: Informatique

Module: RNAA/M1IA

TRAVAUX PRATIQUES – ATELIER N° 01

Objectifs de l'atelier

Découvrir un environnement d'apprentissage automatique

Présentation de Weka 3.8

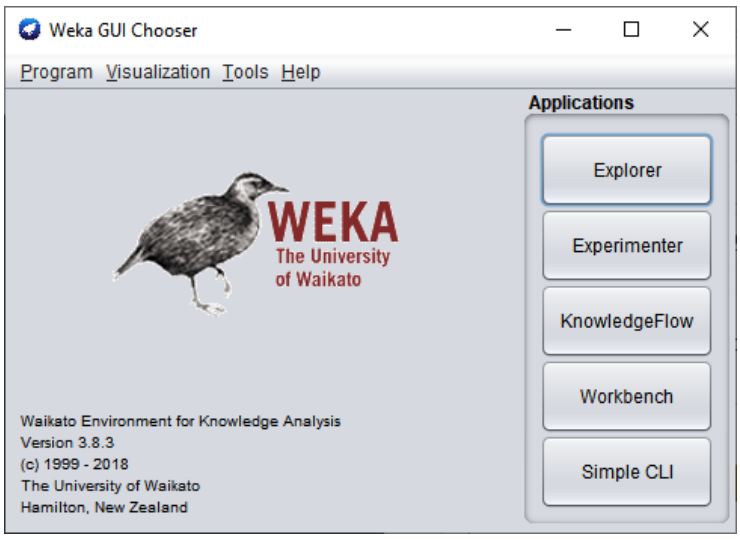
- Weka est un excellent logiciel pour faire l'apprentissage automatique (machine learning).
- Est un logiciel libre de data mining développé en java par l'université de Waikato et publié sous licence GNU (General Public License).
- Son développement a commencé en 1992 en C, puis en 1997 l'équipe de développement à décidé d'utiliser le langage java.
- Weka est utilisé par les data scientistes à des fins d'analyse de données.
- Il présente de nombreux avantages tels que : la gratuité, la portabilité, la facilité d'utilisation
- Offre une large collection de modèles de machine learning tels que : les Réseaux de neurones, les arbres décisionnels, ou encore les k-moyennes, etc.
- Dispose d'une interface graphique ce qui lui permet d'être manié par des débutants.

Installation de Weka

A partir de son site officiel, dans la section Download (version Windows 32 bits/64 bits)

Fonctions principales offertes par Weka

1. Le prétraitement de données en utilisant l'onglet "preprocessing"
2. La classification grâce à l'onglet "classification"
3. Le clustering, onglet "clustering "
4. La sélection de caractéristiques avec l'onglet "Feature Selection"
5. Visualiser des données avec l'onglet Visualise



Fonctions :

1. **Explorer :** Découvrir le jeu de données proposé par Weka.

Application : Explorer➔Open file ➔C:\programme➔weka 3.8.6➔Data➔Iris
(Interpréter le résultat).

L'onglet preprocess: permet de prétraiter les données que l'on insère dans Weka afin de les préparer à l'utilisation. Parmi les taches de prétraitement les plus importantes :

1. Sélection des attributs

Exemple 1 : utiliser un filtre supervisé (supervised filter)
Preprocess➔choose➔Filters➔choisir filter-type (supervised)➔AttributeSelection (pour sélectionner quelques attributs parmi plusieurs)

Avantages de sélection d'attributs :

- Sélectionner les attributs les plus pertinents pour résoudre le problème de classification.
- Alléger votre data set.
- Réduire le temps d'apprentissage
- Faciliter l'affichage des résultats

Exemple 2 : utiliser un filtre non supervisé (unsupervised filter)
Preprocess➔choose➔Filters➔choisir filter-type (unsupervised)➔AttributeSelection (pour sélectionner quelques attributs parmi plusieurs)

Types de filtres non supervisés

- Remove : Permet de supprimer des attributs.
- AddExpression : Permet de construire un nouvel attribut à partir d'une expression mathématique.
- AddID : Ajoute un id à chaque instance du jeu de données.
- Center : Centre les attributs du jeu de données de telle sorte que leur moyenne soit égale à 0.
- DateToNumeric : Convertis les dates en millisecondes.
- Standardise : Standardise les attributs du jeu de données de telle sorte que leur variance soit égale à 1.
- ReplaceMissingValues : Remplace les valeurs manquantes du jeu de données par la moyenne, ou le mode de chaque attribut.
- StringToWordVector : Convertis un texte en vecteur de valeur numérique représentant les occurrences de chacun des mots dans le texte.

2. Centrer et réduire les données (par exemple pour pouvoir les afficher)

Pour centrer réduire le jeu de données il suffit d'appliquer le filtre center puis le filtre standardise. Il est possible d'enregistrer le jeu de données centré et réduit au format qu'on souhaite en cliquant sur "Save ..."

3. ***Enregistrer un jeu de données avec le bouton SAVE:*** Une fenêtre "Enregistrer" apparaît.

On choisit un format favori parmi les suivants Arff, CSV, JSON, ...et on introduit le nom du fichier. Le format CSV est avantageux, car on peut importer le jeu de données dans Excel ou google sheet et faire des graphiques. Par contre le format ARFF est un format qui a été créé spécialement pour Weka.

Différentes zones sur la fenêtre principale :

1. La box Current Relation : indique le nom du jeu de données courant, son nombre d'exemple et d'attribut et la somme des poids de la totalité des exemples.
2. La box Attributes : affiche les attributs du jeu de données courant. Elle permet également de sélectionner ou de supprimer des attributs avec le bouton "Remove."
3. La boxSelected Attribute : affiche les propriété statistique (moyenne, min, max, histogramme, ...) de l'attribut couramment sélectionné.
4. Le bouton « VisualizeAll » : visualise l'histogramme de l'ensemble de tout les attributs simultanément.
5. Status box : nous donne des informations sur l'opération que Weka est en train ou d'effectuer. Le nombre d'opération en cours est indiqué par l'oiseau à droite.
6. Le bouton « log » : pour accéder à l'historique des actions que vous avez effectuées jusqu'à maintenant :

L'onglet Visualisation des données : Weka dispose de l'onglet visualisation qui propose des graphiques 2D montrant la répartition des instances du jeu de données par rapport aux caractéristiques sélectionnées par l'utilisateur. En cliquant sur un des graphiques, une fenêtre contenant ce dernier apparaît.

Bon Courage