Fiche TP N°1

Prise en main du logiciel WEKA

Ce TP a pour but de prendre en main l'outil *WEKA* (Waikato Environment for Knowledge Analysis), un outil open source proposant un ensemble de classes et d'algorithmes en Java implémentant les principaux algorithmes de Machine Learning.

1. Présentation et installation

L'outil WEKA implémente la plupart des algorithmes d'apprentissage supervisé et non supervisé. Il est disponible gratuitement à l'adresse http://www.cs.waikato.ac.nz/ml/weka/, dans des versions pour Unix et Windows.

Cet environnement peut être utilisé de 3 manières :

1. Via l'interface graphique, pour charger un fichier de données, lui appliquer un algorithme, vérifier son efficacité, qu'est suffisante pour les premières expériences. (C'est la méthode utilisée dans ce TP) ;
2. Via l'interface en CLI qu'est recommandée pour une utilisation plus avancée :
   - Elle offre des fonctionnalités supplémentaires ;
   - Elle utilise beaucoup moins de mémoire.
3. Utiliser les classes définies dans ses propres programmes pour créer d'autres méthodes, implémenter d'autres algorithmes, comparer ou combiner plusieurs méthodes.

Lors de son ouverture, une petite fenêtre intitulée *WEKA GUI Chooser* s'affiche avec quatre boutons. Lorsque le mode *Explorer* est choisi, une nouvelle fenêtre s'ouvre (Weka Knowledge Explorer) composée principalement :

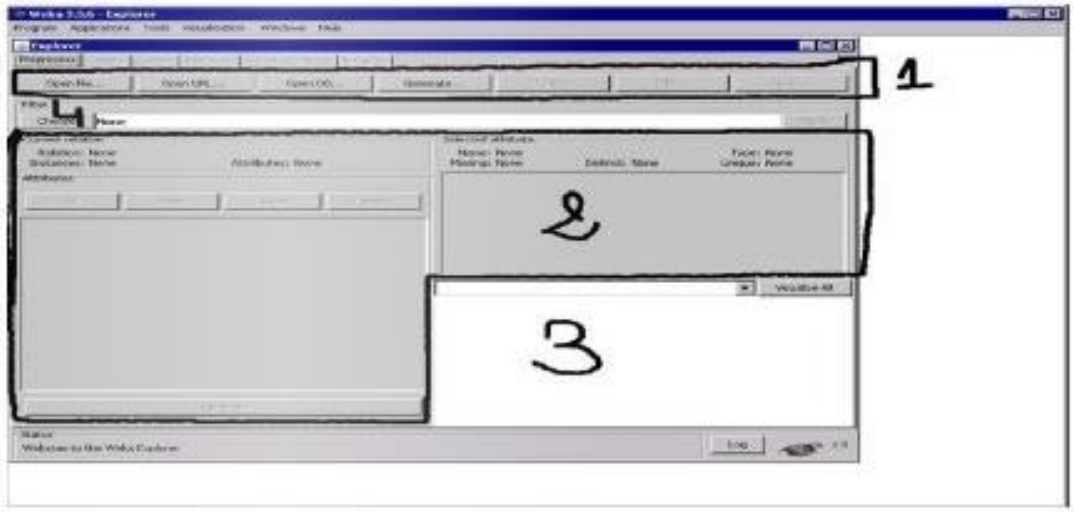
- De classes Java permettant de charger et de manipuler les données en utilisant le mode ***Preprocess ;***
- De classes pour les principaux algorithmes de classification supervisée et non supervisée, qui peuvent être utilisés via le mode ***Classify*** et ***Cluster*** ;
- D'outils de sélection d'attributs accessibles via le mode ***Select attributes*** ;
- D'outils d'application de l'algorithme de génération de règles d'association en utilisant l'onglet ***Associate*** ;
- De classes permettant de visualiser les résultats en utilisant le mode ***Visualize***.

2. Identification des composants du logiciel WEKA

1. Téléchargez WEKA et installez le.
2. Lancer le logiciel WEKA et identifier ses composants.

Après avoir lancé WEKA, la fenêtre qui vient de s'ouvrir permet de :

- Charger un fichier de données (partie 1).
- Explorer les données en affichant les infos par attributs ou par classe (partie 2 et 3).
- Modifier les données en utilisant des filtres (partie 4).



3. Identifier l'onglet *Preprocess* de préparation des données dans la rubrique *Explorer*.
4. Les bases de données exemples sont installées dans le répertoire data. WEKA utilise le format de fichier **ARFF** (pour *Attribute-Relation File Format)* par défaut pour enregistrer les données. Ouvrez le jeu de données d'exemples *weather.arff* (onglet Preprocess / Open File).
5. Combien y a-t-il d'exemples dans le dataset ?
6. Quels sont les attributs servant à d´écrire les exemples ? Pour chacun des attributs, quel est son type et les valeurs possibles ?
7. Visualisez la répartition des classes en fonction des valeurs d'attributs (onglet *Visualize*, attention à modifier la valeur de Jitter et de PointSize pour bien visualiser les exemples).
8. Ouvrez le fichier *weather.arff* grâce à un éditeur de textes. Décrivez l'en-tête et la structure du jeu de données.
9. Mêmes questions pour le dataset *iris.arff* .

3. Exploration des données avec WEKA

Il est souvent nécessaire de prétraiter les données avant d'utiliser un algorithme d'apprentissage. Cela permet notamment de supprimer les instances correspondant à des erreurs de mesures ou d'éliminer des attributs superflus. Cela peut aussi permettre d'uniformiser les données. Il existe un grand nombre de filtres que WEKA peut appliquer aux données. Pour les choisir, il suffit de cliquer sur le bouton *Choose* dans la zone *Filter* (parties 4).

Vous avez le choix entre des filtres supervisés (qui utilisent la classe des données) et non supervisés. Une fois ce choix accomplis, vous avez le choix entre des filtres d'attributs ou d'instances. Avec un clic droit sur le nom du filtre, il est possible d'obtenir plus de détails sur son fonctionnement. Il est également possible de modifier certains paramètres. Pour appliquer

un filtre, il suffit de cliquer sur *Apply*. En particulier, nous nous intéressons à ceux permettant de :

- Normaliser les valeurs numériques.
- Discrétiser des données.
- Remplacer des valeurs manquantes.
- Eliminer des attributs non pertinents.

Le jeu de données *diabetes.arff* contient différents attributs qui peuvent être utiles pour prédire si l'individu est infecté par le diabète ou non. Ce jeu de données contient des données extraites de la population américaine.

A. Normalisation

Le filtre normalise toutes les valeurs numériques de l'ensemble de données donné (à l'exception de l'attribut de classe, s'il est défini). Par défaut, les valeurs résultantes sont dans [0,1] pour les données utilisées pour calculer les intervalles de normalisation. Mais avec les paramètres d'échelle et de translation, il est possible de changer cela, par exemple, avec échelle = 2.0 et translation = -1.0, vous obtenez des valeurs dans l'intervalle [-1, +1].

- Sélectionnez le filtre *Normalize* dans le dossier unsupervised / attribute. Puis faites un clic droit sur le filtre afin d'obtenir un descriptif du filtre. Appliquez le sur les données *diabetes.arff*.

B. 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 (c'est-à-dire transformer son ensemble de valeurs en un nombre fini d'éléments). Le filtre *Discretize* permet de rendre discret un attribut continu.

- Ici, il y a plusieurs attributs numériques. Discrétiser ces attributs en utilisant le filtre de WEKA.

C. Remplacement de valeurs manquantes

Sélectionnez le filtre *replacemissingValue* dans le dossier unsupervised / attribute. Puis faites un clic droit sur le filtre afin d'obtenir un descriptif du filtre. Appliquez le sur les données *diabetes.arff*.

D. Sélection d'attributs en utilisant *Ranking*

Les données comportent souvent des attributs inutiles. Pour cela, on garder les attributs qui sont utiles : 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. L'algorithme *Ranker* de WEKA peut être appliqué pour fournit un classement égal et raffiné à tous les attributs. Après le classement de tous les attributs, on peut omettre l'attribut de rang inférieur pour obtenir des résultats précis.

– Sélectionnez le mode *Select attributes*, dans le dossier *attributeSelection*, selectionnez *Ranker*, puis faites un clic droit dessus afin d'obtenir un descriptif de la méthode. Appliquez la sur le jeu de données *diabetes.arff*.

E. Sauvegarde

Sauvegardez les modifications que vous avez effectuées sur le jeu de données. N'oubliez pas de le renommer pour ne pas écraser le fichier déjà existant.

F. Classification

Après avoir effectué le prétraitement des données (Preprocess), il est possible de faire du Machine Learning avec les modes *Classify* et *Cluster*, ce qui fera l'objet du prochain TP.