I. Classification sous l'environnement Explorer

1. Classification à partir de la visualisation des données

Pour une première approche sur la classification de la base Iris.arff déjà chargée, passez dans la fenêtre **Visualize**. Vous y voyez un ensemble de 25 graphiques (que vous pouvez ouvrir en cliquant dessus), qui représentent chacun une vue sur l'ensemble d'exemples selon deux dimensions possibles, la couleur des points étant leur classe. Sur le graphique, chaque point représente un exemple : on peut obtenir le descriptif de cet exemple en cliquant dessus. La couleur d'un point correspond à sa classe (détaillé dans la sous-fenêtre Class colour). Au départ, le graphique n'est pas très utile, car les axes représentent le numéro de l'exemple.

- Changez les axes pour mettre la largeur des pétales en abscisse, et la longueur des sépales en ordonnées.
- Proposez un ensemble de règles simples permettant de classer les exemples selon leur genre. Les petits rectangles sur la droite de la fenêtre représentent la distribution des exemples, pour l'attribut correspondant, par rapport à l'attribut (ou la classe) codé par la couleur. En cliquant du bouton gauche sur un de ces rectangles, vous le choisissez comme axe des X, le bouton droit le met sur l'axe des Y.
- En mettant la classe sur l'axe des X, quels sont à votre avis les attributs qui, pris seuls, permettent le mieux de discriminer les exemples ? Si les points sont trop serrés, le potentiomètre Jitter, qui affiche les points "à peu près" à leur place, vous permet de les visualiser un peu plus séparément : cela peut être utile si beaucoup de points se retrouvent au même endroit du plan.

2. Classification supervisée : modèle arbre de décision

- Pour classifier des données, cliquez sur l'onglet **Classify**. Ensuite, choisissez une option de test.

La zone Test options permet de choisir de quelle façon l'évaluation des performances du modèle appris se fera. Les options suivantes sont fournies :

- L'option **Use training set** utilise l'ensemble d'entraînement pour cette évaluation.
- L'option **Supplied test set** va utiliser un autre fichier.
- Lorsque l'option **Cross-validation** est sélectionnée, l'ensemble d'apprentissage est coupé en 10 (si Folds vaut 10). L'algorithme va apprendre 10 fois sur 9 parties et le modèle sera évalué sur le dixième restant. Les 10 évaluations sont alors combinées. Cette option sera utilisée pour ce TP.
- Avec l'option **Percentage split**, c'est un pourcentage de l'ensemble d'apprentissage qui servira à l'apprentissage et l'autre à l'évaluation.

- Ensuite, cliquez sur le bouton **Choose** de **Classifier** pour choisir un algorithme parmi ceux proposés par WEKA. Dans la fenêtre qui s'ouvre,

développez le dossier **trees** et choisir l'algorithme **J48** (l'implémentation de C4.5). Cliquez sur **Start** pour effectuer l'analyse.

Dans la partie **Classifier output** vous aurez des statistiques sur le fichier exploité, à savoir le nombre d'instances **Total Number of Instances** de votre fichier, le nombre d'instances correctement classifiées **Correctly Classified Instances** et incorrectement classifiées **Incorrectly Classified Instances** ainsi que d'autres statistiques. Sur le même écran vous avez aussi la **matrice de confusion** : ***Confusion Matrix*** de cette analyse.

- Une fois une méthode choisie, il est possible de modifier les valeurs des paramètres par défaut de WEKA en cliquant directement sur le nom de la méthode (e.g. en cliquant sur "J48 -C 0.25 -M 2"). Vous pouvez modifier la valeur du "confidence factor" qui contrôle le degré d'élagage. Vous pourrez également modifier le nombre minimal d'exemples qui doivent figurer dans une feuille de l'arbre.
- Modifier l'option de test à **Percentage split : 2/3 entrainement et 1/3 test.** Quelle est la méthode la plus performante ?
- Pour afficher l'arbre de décision, cliquez droit dans la partie **Result list**. Ensuite, Choisissez l'option **Visualize tree**.
- Analysez cet arbre

II. Classification sous l'environnement KnowledgeFlow

- Fermez l'environnement **'Explorer'** et ouvrez celui **' KnowledgeFlow'**.

  Sous cet environnement, les modules sont classés en 8 catégories :

  - **Datasources** qui permet de récupérer des données sous la forme de fichier .arff, csv ou de base de données ;
  - **Datasinks** qui permet de sauvegarder des sorties de données sous la forme de fichier arff, csv ou de base de données ;
  - **Filters** qui permet de filtrer les données ;
  - **Classifiers** présentant les différentes méthodes de classification ;
  - **Clusterers** présentant les différentes méthodes de clustering ;
  - **Association** présentant les différentes méthodes d'association ;
  - **Evaluation** qui permet l'évaluation des modèles crées par les modules classifiers, clusterer et association
  - **Visualization** qui permet de visualiser les données, les modèles et les évaluations sous la forme de texte, de courbe et de graphe.

<u>Sélection de données</u>

Sélectionnez le module ' **Arff Loader**' dans la catégorie **datasources** permettant d'obtenir un fichier.arff. Configurez le pour exploiter le fichier iris.arff.

Visualisation des données

- Afin de visualiser la répartition des attributs, installez le module de visualisation '**Attribute Summarizer**'.
- Installez également le module '**DataVisualizer**', pour visualiser les exemples
- liez les deux modules de visualisation via un **'dataset'** au module de sélection des données. Chose faite, lancez les données à travers le graphe en appuyant sur **start loading** (bouton droit sur le module de sélection des données). Visualisez les informations fournies par les deux modules de visualisations grâce aux options **'Show Summaries'** et **'Show Plot'**

Sélection de la classe

- Afin de spécifier quel attribut nous souhaitons utiliser comme classe, vous devez sélectionner le module '**ClassAssigner'** qui se trouve dans la catégorie '**Evaluation'.**
- liez ce module au celui de sélection des données via un **'dataset'** et configurez le pour que la classe soit l'attribut Class.

Séparation des données de test et d'apprentissage

- Trouvez le module permettant d'effectuer une séparation des données en un groupe de test et d'apprentissage. Configurez-le pour que le groupe soit scindé en deux moitiés égales.
- Liez ce module au module '**ClassAssigner'**

Classification

- Installer le module **J48** et reliez-le au module de Séparation des données de test et d'apprentissage.

Visualisation du modèle appris

Pour visualiser le modèle appris par J48, il est possible d'utiliser les graphes ou les fichiers textes.

- Installez ces deux modules de visualisation et relancez les données pour que les nouveaux modules installés les reçoivent. Visualisez le modèle crée en cliquant droit sur les deux modules de visualisation.