**TD 4IF MID-FD 2013/2014 – TD 1 (C. Rigotti – J-F. Boulicaut)**

# Prise en main de la plate-forme de fouille de données KNIME

## Exploitation des données IRIS (benchmark UCI) – Travail en binôme

Site Knime: http://www.knime.org http://www.knime.org/documentation

Vous pouvez facilement installer sur votre poste/environnement de travail la plate-forme KNIME. Sur le département, elle a été installée directement sous « C: ». Il faut veiller à déterminer un espace de travail sur lequel vous avez des droits en écriture (typiquement C:\temp).

Des fichiers ont été déposés à votre attention dans \\servif-home\fic-eleves\Espace Pedagogique\4IF\Modeles et Outils Mathematiques\Analyse et fouille de donnees

Pour installer de nouvelles extensions, un retour vers http://www.knime.org sera nécessaire.

Même si les plantages sont rares, il faut penser à faire des sauvegardes régulières et veiller à avoir les droits d'écriture sur le répertoire choisi pour les sauvegardes de l'espace de travail.

## Visualisation et exploration de données

Lecture : Possibilité d'utiliser différents formats, e. g., CSV et autres (Fonction 'configure' du composant « File Reader »).

Essai avec sur IrisDataset/data.all

Visualisation de la table : Utilisation du composant « Interactive Table » (e.g., choix du mode de représentation des attributs), comprendre la sémantique de ces données.

Domaine et dispersion des attributs : Utilisation des composants « Box Plot », « Scatter Plot », « Scatter Matrix », et « Interactive Histogram ».

Mise en évidence d'objets : Utilisation des composants « Color Manager », « Shape Manager » à essayer sur un attribut de classe mais aussi sur des attributs numériques.

Marquage de points : Composants « HiLite » et « HiLite Filter » avec mise à jour des informations sur les domaines des attributs au moyen de « Domain Calculator » (e.g., après sélection dans une table).

Sélection de colonnes, de lignes.

Calcul de statistiques simples au moyen des composants du groupe « Statistics » et outils utilisables pour la normalisation des valeurs numériques (« Normalizer »).

Découverte du composant « Parallel Coordinates »

Etude des corrélations linéaires « Linear correlation »

Découverte de la réduction du nombre de dimensions au moyen de l'analyse en composantes principales (composant « PCA »). Voir (au sens de visualiser) comment la projection des données sur une ou deux composantes principales impacte la discrimination entre espèces.

## « Clustering »

Voir certaines méthodes classiques via les composants « K-Means », «Fuzzy c-Means », et « Hierarchical clustering »

Bien étudier les paramétrages de ces méthodes et composants : choix des ensemble d'attributs retenus (descripteurs) avec ou sans processus de normalisation, choix de la mesure de similarité, nombre de clusters, pondération selon la distance entre groupes dans les méthodes hiérarchiques (« single », « average », « complete »).

Observer et comprendre les formes des clusters assez différentes avec « single » et « complete».

Qualité du clustering : visualisation de la répartition des objets (usage de « shape manager » ou de « color manager » puis des composants de visualisation), sommes des carrés des distances intra et inter-clusters (fournies par Fuzzy C-means), mesure par rapport à des groupes déjà connus et donc selon un critère externe (e.g., lorsque l'on a déjà des labels qui identifient des groupes comme les espèces d'IRIS déjà identifiées) avec Mining -> Scoring -> Entropy Scorer pour une mesure d'entropie ou avec Mining -> Scoring -> Scorer pour avoir une table de contingence.

Tester l'impact du mélange aléatoire (composant « shuffle ») du fichier de données sur les résultats (utiliser « Looper » – méta composant - du menu « Meta » pour répéter un traitement, et agréger les résultats au moyen d'un « Scorer »). Comment utiliser les composants disponibles pour évaluer la stabilité d'un algorithme comme « K-Means » ?

Essayer la construction d'un arbre de décision (« Decision Tree Learner ») pour expliquer les structures de classification obtenues (groupements calculés) .