TD 4IF MID-FD 2015/2016 – TD 2 (C. Rigotti – J-F. Boulicaut)

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

Travail en binôme

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

http://www.knime.org/documentation

Classification supervisée

Premiers pas vers la classification supervisée avec illustrations sur le jeu de données IRIS.

Parmi les quatre attributs numériques du jeu IRIS, quels sont les couples semblant porter assez d'information pour séparer les espèces ? Tester en vérifiant si ce ou ces couples permettent d'obtenir de bonnes classifications et regarder également ce qui se passe lorsque l'on utilise le ou les attributs calculés au moyen d'une « PCA ».

Comprendre la notion de jeu d'apprentissage (training) et de jeu de test. Certains composants partitionnent eux-mêmes les données en jeu d'apprentissage et jeu de test, d'autres n'utilisent en entrée que le jeu d'apprentissage et fournissent en sortie un modèle (port de sortie bleu, e.g., composants « Learner ») que l'on peut alors appliquer sur le jeu test (composant « Predictor »). Certains composants existent en deux versions (par exemple « J48 » dans Mining->Decision-Trees et « J48 » dans Weka->Trees.

Essayer d'abord la prédiction de l'espèce de fleur au moyen d'un arbre de décision à apprendre (« Decision Tree Learner ») puis appliquer ((« Decision Tree Predictor »). On peut utiliser le composant de partitionnement Data-Manipulation -> Row -> Partitioning pour construire les ensembles d'apprentissage et de test.
Evaluer la qualité de la classification par rapport à la classe de référence : mesure de « Entropy Scorer », table de contingence et taux d'erreur (composant « Scorer »).
Découvrir la méthode classique d'estimation de la qualité d'une méthode via la validation croisée (« cross validation » dans le menu « Meta »).

Comparer les résultats obtenus avec plusieurs méthodes, par exemple les arbres de décision (« J48 »), la classification à base de règles (« JRip »), ou encore les K plus proches voisins (« K Nearest Neighbor – KNN »).

Données BEARS (Benchmark UCI, datasets/bears)

Précaution : ce jeu comporte plus d'attributs, certains composants peuvent demander plus d'attention lors de la configuration/utilisation pour éviter de trop grandes consommations de ressources ou des affichages surchargés et/ou illisibles.

Identifier des « outliers », des individus ayant des caractéristiques hors normes (par classification et/ou visualisation). Sont-ils intéressants ? Les garder ? Les supprimer ?

Pour avoir une notion de distance qui ait du sens lors d'un « clustering » : choix des attributs, normalisation (standardisation) éventuelle.

Clustering avec « K-Means » et « Fuzzy C-Means » : utiliser dans un premier temps le « clustering hiérarchique » pour avoir une idée du nombre raisonnable de clusters à fixer ensuite comme paramètre.

Ne pas hésiter à se concentrer sur des sous-ensembles d'attributs (raisonnablement choisis) afin de pouvoir interpréter les résultats.

Utiliser des attributs dérivés (composant « Math Formula » ou « Java Snippet » ou encore les composant pour « scripts Python »), pour calculer, par exemple, l'Indice de Masse Corporel = (poids/(taille*taille)) qui peut être un descripteur plus intéressant pour la classification, qu'elle soit supervisée ou non.

Construire un modèle de classification lisible par l'être humain (e.g., arbre de décision ou classifieur à base de règles) pour « expliquer » a posteriori les clusters formés. Essayer d'expliquer les clusters avec les attributs utilisés pour un « clustering », et essayer aussi de le faire seulement avec des attributs n'ayant pas servi au « clustering ».

Exercice : discrétiser un attribut en utilisant « K-Means » (par exemple, le poids ou la taille en trois groupes : petit, moyen, grand). Construire un classifieur pour prévoir cette valeur discrète (avec l'attribut discrétisé et aussi sans). Remarque : si la méthode de discrétisation ne semble pas appropriée à l'attribut choisi, il existe aussi d'autres composants permettant de réaliser des discrétisations : « CAIM » et « Numeric Binner ». Evaluer la qualité du modèle de classification et estimer l'erreur par validation croisée.

Exercice : Peut-on prévoir le genre ?

Essayer des méthodes de régression (menu Statistics : Linear Correlation, Linear/Polynomial Regression Learner/Predictor). Peut-on arriver à prévoir le poids ?

Découverte de motifs

La méthode des règles d'association s'applique sur des données transactionnelles ou assimilées. On propose de traiter les données sur les champignons (Jeu de données « Mushroom ») et 'en extraire certaines règles d'association significatives. Ceci va demander de transformer les données avant de pouvoir utiliser des composants comme « Association Rules » ou « Apriori » de Weka.