Module Business Intelligence
5A STI - ASL
TP Fouille de Données avec WEKA

Information

Ce TP est noté. Vous devrez m'envoyer un document en format .pdf à l'adresse suivante : benjamin.nguyen@insa-cvl.fr avec comme sujet [ASL-TP-WEKA] et votre nom, dès la fin du TP.
Même si WEKA est une bibliothèque Java, et peut donc être appelé au sein de programmes Java, nous l'utiliserons exclusivement via l'interface graphique.
Les questions auxquelles il faut donner une réponse sont en *italique*.

Préparation

Téléchargez la version de WEKA pour l'OS que vous utilisez à l'adresse suivante : http://www.cs.waikato.ac.nz/ml/weka/downloading.html

Vous pouvez exécuter WEKA en vous plaçant dans le répertoire correspondant et en tapant la commande : `java -Xmx1000M -jar weka.jar`

Une fenêtre se lance, choisissez le bouton "Explorer".
Pour choisir un jeu de données, prenez l'onglet "preprocess", cliquez sur le bouton "open file", naviguez vers le répertoire d'installation de WEKA, entrez dans le répertoire "data", vous trouverez les fichiers de données (.arff). Dès que vous avez choisi un fichier, les autres onglets (classification, clustering, règles d'association) deviennent actifs. Il vous est également possible, en restant sur l'onglet "preprocess" de visualiser le jeu de données, et d'appliquer des prétraitements.

Le Golfeur : Fichier : weather.nominal.arff et weather.numeric.arff
Le fichier contient des informations qui expliquent si on doit jouer au golf ou pas, selon les conditions météo.

1) Visualisez le contenu de ce fichier en format .arff, et en utilisant le visualisateur de WEKA.

Chaque ligne du fichier weather.nominal.arff contient les informations suivantes :

outlook = sunny, overcast, rainy (ciel = soleil, couvert, pluie)
temperature = hot, mild, cold (température = chaud, moyen, froid)
humidity = high, normal (humidité = haute, normale)
windy = TRUE, FALSE (vent = vrai, faux)
play = yes, no (on joue au golf = oui, non)

Combien y-a-t-il d'instances dans ce fichier ?

Dans le cas du fichier weather.numeric.arff, les données associées sont numériques, et non plus discrètes.

2) On va chercher à créer un classifieur pour savoir si c'est une bonne idée de jouer au golf, compte tenu des conditions climatiques. On utilisera les fichiers weather.nominal.arff et weather.numeric.arff et on comparera les résultats.

Utilisez un classifieur de type kNN (lazy/IBk). En cliquant sur le bouton "more options" cochez "output predictions".

Sans changer les paramètres de base, lancez le classifieur en cross-validation Folds = 10, et donnez la qualité des résultats, exprimés par la matrice de confusion. Comparez les résultats sur weather.nominal et weather.numeric en expliquant d'où peuvent provenir les différences de qualité sur les résultats.

En utilisant le fichier weather.nominal.arff, créez un fichier golf.arff avec une seule ligne, avec les valeurs outlook = sunny, temperature = hot, humidity = normal, windy = FALSE, et play = yes. Après avoir lancé le classifieur sur weather.nominal.arff en cross validation folds 10, avec les paramètres de base, cliquez sur le bouton start pour générer le modèle. Cliquez ensuite sur "supplied test set" et choisissez votre fichier golf.arff. Cliquez avec le bouton droit sur la ligne lazy.IBk et choisissez re-evaluate model on current test set. Vous êtes en train de tester le classifieur, construit en utilisant le fichier weather.nominal.arff sur le fichier golf.arff.

Comment faire pour voir la prédiction du classifieur ? Quelle est cette prédiction ?

Créez un nouveau fichier golf2.arff avec les valeurs valeurs outlook = sunny, temperature = hot, humidity = normal, windy = FALSE, et play = no et effectuez la même manipulation.

Qu'obtenez-vous, est-ce que le classifieur a fait la même prédiction que précédemment ?

3) On utilise maintenant un classifieur par arbre : trees/J48

Donnez l'arbre résultant de la classification de weather.nominal.arff avec les paramètres de base. Appliquez ***à la main*** *ce classifieur à un tuple dont la valeur est* outlook = sunny, temperature = hot, humidity = normal, windy = FALSE. *Appliquez automatiquement le classifieur pour voir si le résultat est le même.*
Est-ce que le classifieur conseille de jouer au golf si : outlook = sunny, temperature = cold, humidity = normal, windy = TRUE ? *Justifiez.*

Quel arbre obtenez-vous en utilisant le classifieur trees/FT ? Donnez la matrice que confusion. Comment faut-il l'interpréter ?

4) On va utiliser le fichier weather.nominal.arff, mais avec un traitement sous forme de règles d'associations, en utilisant l'algorithme **a priori**. Choisissez donc l'onglet Associate et l'algorithme a priori.

Lancez l'algorithme ***a priori*** *avec les paramètres de base. Parmi les 10 meilleures règles d'association, donnez celles qui permettent de conclure au fait qu'on joue ou pas au golf. Combien en avez-vous ? Comparez ces règles par rapport aux résultats des algorithmes de classification par arbre utilisés précédemment. Commentez.*

5) On va utiliser le fichier weather.nominal.arff, mais avec un traitement sous forme de clustering.
On commence par utiliser l'algorithme **SimpleKMeans**. Choisissez donc l'onglet Cluster et l'algorithme SimpleKMeans.

Appuyez sur le bouton classes to clusters evaluation et choisissez "play". Cela permet de comparer les résultats du clustering aux valeurs attendues.

Lancez l'algorithme, en choisissant numClusters = 2. Combien d'instances sont dans le cluster 0 et combien sont dans le cluster 1 ? Combien d'instances play=yes sont dans le cluster 0 et dans le cluster 1 ? Idem pour le nombre d'instances play=no ? Commentez la qualité du clustering.

Faites de même avec le fichier weather.numeric.arff

On utilise maintenant l'algorithme **DBSCAN**, sur le fichier weather.numeric.arff.

Trouvez un paramétrage permettant de classifier correctement au moins 75% des points qui font partie d'un cluster. Combien avez-vous de clusters ? Combien avez-vous de points dans chaque cluster ?

6) D'après un énoncé de Marc Plantevit (Université de Lyon I)

Un supermarché décide de traiter ses tickets de caisse avec du datamining. Les données sont les suivantes pour une semaine donnée.

NoClient	Produit1	Produit2	Produit3	Produit4	Produit5
1	X			X	X
2	X	X			X
3					X
4			X	X	X
5	X	X	X	X	X
6	X				X
7	X			X	X
8		X	X		

Construisez le fichier .arff correspondant à ces données.

Faites une extraction des règles d'association avec un support de 0.5 puis un support de 0.1

Supposons que tous les produits ont initialement le même prix de 10 EUR, que sur chaque produit le supermarché fait 5 EUR de bénéfices et que le patron du supermarché vous demande de suggérer 1 produit pour faire une promotion la semaine suivante. On suppose aussi que les clients vont tous acheter au moins les mêmes produits que la semaine précédente.

Quel produit lui suggérez vous ? Quel rabais suggérez vous ? A combien estimez vous les bénéfices ou pertes suite à cette opération de promotion ? Justifiez.