Acquisition de connaissances 2
4INFO
2013-2014

TP4 : Génération de règles d'association

Cette séance de travaux pratiques vise à concrétiser les différents principes abordés dans le cours sur l'apprentissage non-supervisé de règles d'associations. La séance comprend 2 parties : l'une essentiellement « presse-bouton » s'appuie sur le logiciel d'apprentissage Weka déjà utilisé lors des précédentes séances ; l'autre, plus libre, reproduisant un problème réel.

Un compte-rendu vous est demandé pour cette séance. Celui-ci doit présenter des réponses concises, mais argumentées, à toutes les questions de l'énoncé. Ce compte-rendu est à déposer sous moodle au ***format PDF*** *dans un* ***délai d'une semaines*** *après la séance (faites en sorte que votre CR porte* ***les noms des deux*** *étudiants du binôme).*

1 Fouille de données sous Weka

1.1 Présentation de Weka

Weka est une plateforme de fouille de données qui implémente une collection d'algorithmes d'apprentissage. Les algorithmes peuvent être directement appliqués sur une base de données ou appelés à partir d'un code Java. Weka contient des outils pour le pré-traitement des données, le classement, la régression, les règles d'association et la visualisation.

Les données sont sous un format ARFF -pour *Attribute-Relation File Format*-. Ce format est simple et il est facile de convertir des données issues par exemple d'un tableur en ARFF.

1.2 Test de Weka avec l'exemple du golf

Weka est installé sur vos machines : /soft/local/stow/weka-3-5-6. Suivez les étapes suivantes :

1. Lancez Weka, puis l'Explorer (dans Application). Choisissez le fichier weather.nominal.arff parmi les données fournies avec Weka [1] : il s'agit d'un petit exemple où un golfeur décrit chacune de ses journées par des indices météorologiques auxquels s'ajoutent l'information selon laquelle l'observateur est allé jouer au golf ou non ce jour là.
2. Dans l'onglet *Associate*, choisissez *APriori*.
3. Vérifiez que tout fonctionne en lançant l'algorithme sans modifier les paramètres du programme.

Question 1.1
Identifiez à quelles étapes et à quels résultats d'Apriori, vu en cours, correspondent les informations retournées par Weka ?

En cliquant dans la fenêtre en face du bouton *Choose*, on a accès aux paramètres de l'algorithme. Le bouton *More* détaille chacune de ces options. Voici la liste des paramètres les plus utiles :

car : indique si les conclusions des règles générées doivent porter uniquement sur l'attribut classe (*true*) ou si celles-ci sont libres (*false*).

1. Dans le sous-répertoire data/ de Weka.

classIndex : numéro de l'attribut classe si *car* est mis à vrai.
lowerBoundMinSupport : valeur minimale du seuil pour le support. Le seuil part d'une valeur initiale et décroît régulièrement jusqu'à cette borne.
upperBoundMinSupport : valeur initiale du seuil pour le support.
delta : le support minimal décroit de cette quantité, jusqu'à ce qu'ait été trouvé le nombre de règles demandé ou qu'ait été atteinte la valeur minimale du support.

metricType : mesure qui permet de classer les règles. Quatre mesures sont proposées :
- **Confidence** : la confiance (vu en cours).
- **Lift** : $\frac{conf(A,B)}{conf(\emptyset,B)}$.
- **leverage** : $sup(A \cup B) - sup(A)sup(B)$.
- **Conviction** : $\frac{conf(\emptyset,\neg B)}{conf(A,\neg B)}$.

minMetric : valeur minimale de la mesure choisie, en dessous de laquelle l'algorithme ne cherchera plus de règles.

numRules : nombre de règles que l'algorithme doit produire au maximum.

Question 1.2
Sur le fichier weather.nominal.arff, comparez les règles produites selon la mesure choisie, que constatez vous ?

Question 1.3
Pour une règle que vous choisirez, vérifiez les calculs de confiance, lift et leverage en donnant les détails du calcul.

Question 1.4
Décrivez en quelques mots la notion de conviction (de façon informelle).

1.3 Weka pour l'étude de la population américaine

Téléchargez le fichier adult1.csv depuis le moodle et chargez le dans Weka. Ce fichier contient un extrait du recensement de la population américaine. À partir des attributs répertoriés dans la table 1, le but d'origine de ces données est de prédire si quelqu'un gagne plus de 50 000 dollars par an.

1.3.1 Netoyage et normalisation des données

Note : Pour cette partie vous n'avez pas besoin de faire apparaitre de réponse dans votre CR. Vous devrez seulement joindre en annexe le fichier adult_discretized_VotreNOM.arff final.

Les données comportent parfois des attributs inutiles, par exemple le numéro de dossier ou encore la date de saisie. Il est possible d'en supprimer « à la main », à condition de connaître le domaine du problème. On peut aussi lancer un algorithme de fouille de données et regarder les attributs qui apparaissent dans les motifs : soit ceux-ci sont pertinents et il est important de les garder, soit ils sont tellement liés à un autre attribut qu'ils génèrent des règles évidentes ou inintéressantes.

Weka a automatisé cette recherche des attributs pertinents dans l'onglet *Select attributes*, qui permet de définir les attributs les plus pertinents selon plusieurs méthodes de recherche (*search*), en utilisant plusieurs méthodes (*eval*) (Nous n'allons pas l'utiliser aujourd'hui).

Attribut	Domaine de définition
age	$\mathbb{R}$
workclass	{Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked}
fnlwgt	$\mathbb{R}$
education	{Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool}
education-num	$\mathbb{R}$
marital-status	{Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse}
occupation	{Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces}
relationship	{Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried}
race	{White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black}
sex	{Female, Male}
capital-gain	$\mathbb{R}$
capital-loss	$\mathbb{R}$
hours-per-week	$\mathbb{R}$
native-country	{United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands}
gain	{50K, <=50K}

TABLE 1 – Description des attributs utilisés dans le fichier adult1.csv

Question 1.5
Supprimez les attributs qui ne vous paraissent pas intéressant pour l'étude. Notons que cette étape se fait généralement avec l'aide d'un expert du domaine ou par des approches automatiques. Dans ce cas, vous serez votre propre expert.

Les algorithmes de génération de règles d'association ont besoin d'attributs discrets pour fonctionner. Le filtre non-supervisé *Discretize* permet de rendre discret un attribut continu et ceci de plusieurs façons :

- en partageant l'intervalle des valeurs possibles de l'attribut en intervalles de tailles égales ;
- en le partageant en intervalles contenant le même nombre d'éléments ;
- en fixant manuellement le nombre d'intervalles (*bins*) ;
- en laissant le programme trouver le nombre idéal de sous-intervalles.

Question 1.6
Discrétiser les attributs numériques de notre problème (cf. table 1) en utilisant le filtre non-supervisé *Discretize* et en forçant le nombre d'intervalles à 3, avec une répartition à peu près équivalente entre les 3 intervalles. Sauver le fichier transformé sous le nom adult_discretized_VotreNOM.arff.

Question 1.7
Les noms des attributs ne sont pas significatifs. Éditez le fichier adult_discretized.arff et rempla-

cez les noms d'attributs dus à la discrétisation par des noms d'attributs explicites pour vous. Attention de bien remplacer les noms des attributs dans l'entête et dans les descriptions.

1.3.2 Fouille des données

Question 1.8
Appliquer l'algorithme APriori et tentez d'interpréter les règles produites pour différents jeux de paramètres (Il n'est pas nécessaire d'être exhaustif, jouez avec les paramètres et ne commentez que ce qui pous parait pertinent).

Question 1.9
Activez l'option *car*, fixez la classe attribut à celle correspondant au champ *gain*, puis lancez Apriori (en prenant soin de fixer *metricType* à « *confidence* »). Qu'obtenez vous ?

2 Étude de cas : Articles de presse

Nous proposons d'analyser des articles de presse tirés du journal Le Monde et datés d'avril 2002. Nous cherchons à en tirer des règles d'associations entre mots ou groupes de mots afin de mettre au jour les différents acteurs et événements marquants de ce mois d'avril. Pour des raisons algorithmiques, nous ne voudrions considérer, pour l'ensemble des articles, que les 200 mots jugés les plus discriminants tout au long du mois.

Question 2.1
À quelle étape particulière d'Apriori le nombre de mots considérés pour représenter un article joue-t-il un rôle prépondérant dans la complexité des calculs ?

Le fichier articles.10p.txt présent sur le moodle répertorie des articles du mois d'avril 2002. Chaque ligne y représente un article sous forme lemmatisée, c'est-à-dire où toutes les formes féminines et plurielles ont été ramenées au masculin singulier et où tous les verbes conjugués ont été ramenés vers leur forme infinitive. L'ensemble des mots discriminants pour ce mois est disponible au même endroit sous le nom de mots.lst.

Il vous est demandé de proposer une méthode permettant de découvrir des liens entre mots ou groupes de mots à partir du texte des articles. Vous êtes libres d'effectuer tous les traitements que vous voulez sur les données à condition que ceux-ci soient précisés et justifiés dans votre compte-rendu. Utilisez ensuite Weka pour générer les règles. Vous pouvez utiliser weka en ligne de commande pour extraire les règles d'association (exemple : java -classpath weka.jar weka.associations.Apriori -t data/weather.nominal.arff).

Question 2.2
Décrivez votre méthode, en particulier votre modélisation (les transactions, les items, la relation qui les lie). De plus, si vous utilisez un script/programme pour modifier les données, joignez le source en annexe de votre CR. Donnez des exemples de règles avec leurs mesures d'intérêt.