Acquisition de connaissances
4INFO
2014-2015

Antoine CHENON
Guillaume ROY
Mathilde LEPARQUIER

TP3 : Génération de règles d'association

1 Fouille de données sous Weka

Question 1.1.

L'analyse du fichier **weather.nominal.arff** nous donne le résultat suivant :

```
Apriori
=======

Minimum support: 0.15 (2 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 47

Size of set of large itemsets L(3): 39

Size of set of large itemsets L(4): 6

Best rules found:

 1. outlook=overcast 4 ==> play=yes 4    conf:(1)
 2. temperature=cool 4 ==> humidity=normal 4    conf:(1)
 3. humidity=normal windy=FALSE 4 ==> play=yes 4    conf:(1)
 4. outlook=sunny play=no 3 ==> humidity=high 3    conf:(1)
 5. outlook=sunny humidity=high 3 ==> play=no 3    conf:(1)
 6. outlook=rainy play=yes 3 ==> windy=FALSE 3    conf:(1)
 7. outlook=rainy windy=FALSE 3 ==> play=yes 3    conf:(1)
 8. temperature=cool play=yes 3 ==> humidity=normal 3    conf:(1)
 9. outlook=sunny temperature=hot 2 ==> humidity=high 2    conf:(1)
10. temperature=hot play=no 2 ==> outlook=sunny 2    conf:(1)
```

Weka retrouve les règles d'association les plus pertinentes.

Question 1.2.

Apriori - **confidence** ======= Minimum support: 0.15 (2 instances) Minimum metric <confidence>: 0.9 Number of cycles performed: 17 Generated sets of large itemsets: Size of set of large itemsets L(1): 12 Size of set of large itemsets L(2): 47 Size of set of large itemsets L(3): 39 Size of set of large itemsets L(4): 6	Apriori - **lift** ======= Minimum support: 0.3 (4 instances) Minimum metric <lift>: 1.1 Number of cycles performed: 14 Generated sets of large itemsets: Size of set of large itemsets L(1): 12 Size of set of large itemsets L(2): 9 Size of set of large itemsets L(3): 1
Apriori - **leverage** ======= Minimum support: 0.3 (4 instances) Minimum metric <leverage>: 0.1 Number of cycles performed: 14 Generated sets of large itemsets: Size of set of large itemsets L(1): 12 Size of set of large itemsets L(2): 9 Size of set of large itemsets L(3): 1	Apriori - **conviction** ======= Minimum support: 0.25 (3 instances) Minimum metric <conviction>: 1.1 Number of cycles performed: 15 Generated sets of large itemsets: Size of set of large itemsets L(1): 12 Size of set of large itemsets L(2): 26 Size of set of large itemsets L(3): 4

Ces 4 mesures permettent d'affiner les règles d'association selon le critère choisi.

Question 1.3.

Pour la règle suivante : 1. outlook=overcast 4 ==> play=yes 4 nous obtenons les valeurs

- Confidence = $\frac{4}{4}=1$
- Lift = $\frac{4/4}{9/14}=\frac{14}{9}$
- Leverage = $\frac{4}{14}-(\frac{4}{9})\times(\frac{9}{14})=0.10$

Question 1.4.

Il s'agit du ratio entre la fréquence de non-apparition de la partie droite de la règle et la fréquence de non-apparition de la partie droite de la règle avec la partie gauche de la règle. Autrement dit, on mesure la dépendance pour les contre-exemples.

Questions 1.5. à 1.9.

Cf. document joint.

Lorsque l'on applique l'algorithme Apriori, on obtient les données suivantes :

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     adult1-weka.filters.unsupervised.attribute.Discretize-B3-M-420.0-Rfirst-last
Instances:    250
Attributes:   15
              age
              workclass
              fnlwgt
              education
              education-num
              marital-status
              occupation
              relationship
              race
              sex
              capital-gain
              capital-loss
              hours-per-week
              native-country
              gain
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.75 (187 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 5

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 7

Size of set of large itemsets L(3): 1

Best rules found:

 1.  gain= <=50K 191 ==>  capital-gain='(-inf-11365]' 190    conf:(0.99)
 2.  hours-per-week='(27.333333-53.666667]' 199 ==>  capital-gain='(-inf-11365]' 195    conf:(0.98)
 3.  capital-loss='(-inf-735.333333]' 232 ==>  capital-gain='(-inf-11365]' 226    conf:(0.97)
 4.  native-country= United-States 217 ==>  capital-gain='(-inf-11365]' 211    conf:(0.97)
 5.  race= White 202 ==>  capital-gain='(-inf-11365]' 196    conf:(0.97)
 6.  capital-loss='(-inf-735.333333]'  native-country= United-States 202 ==>  capital-gain='(-inf-11365]' 196    conf:
(0.97)
 7.  native-country= United-States 217 ==>  capital-loss='(-inf-735.333333]' 202    conf:(0.93)
 8.  capital-gain='(-inf-11365]'  native-country= United-States 211 ==>  capital-loss='(-inf-735.333333]' 196    conf:
(0.93)
 9.  capital-gain='(-inf-11365]' 244 ==>  capital-loss='(-inf-735.333333]' 226    conf:(0.93)
10.  race= White 202 ==>  capital-loss='(-inf-735.333333]' 187    conf:(0.93)
```

En activant l'option « car », on obtient les règles pour les contre-exemples.