TP4 - Génération de règles d'association

Paul Chaignon - Ulysse Goarant

25 février 2014

1 Fouille de données sous Weka

Question 1.1

Weka détermine les itemsets fréquents (en indiquant leur nombre par taille) et les règles associées de confiance suffisante.

Question 1.2

En fonction de la mesure, les règles générées varient. Certaines sont cependant communes à des mesures différentes.

Question 1.3

Règle : $outlook = overcast \Rightarrow play = yes$
Calcul de la confiance :

$$\frac{N(outlook = overcast \cap play = yes)}{N(outlook = overcast)} = \frac{4}{4} = 1$$

Calcul du lift :

$$\frac{N(omega) \times N(outlook = overcast \cap play = yes)}{N(outlook = overcast) \times N(play = yes)} = \frac{70 \times 4}{4 \times 4} = 17.5$$

Calcul du levier :

$$P(outlook = overcast \cap play = yes) - P(outlook = overcast) \times P(play = yes) =$$

Question 1.8

Les règles obtenues selon les paramètres par défaut sont les suivantes :

1. marital-status=_Never-married capital-gain=_faible-gain-place
capital-loss=_faible-perte-placement 67 ==> gain=_<=50K 64 conf:(0.96)
2. age=_jeune workclass=_Private capital-gain=_faible-gain-place
gain=_<=50K 63 conf:(0.95)
3. marital-status=_Never-married capital-gain=_faible-gain-place
gain=_<=50K 70 conf:(0.95)

```
          4. workclass=_Private education-num=_peu-eduque capital-gain=_
gain=_<=50K 69     conf:(0.95)
          5. age=_jeune workclass=_Private 70 ==> gain=_<=50K 66
conf:(0.94)
          6. workclass=_Private education-num=_peu-eduque capital-gain=_
capital-loss=_faible-perte-placement 67 ==> gain=_<=50K 63     conf:(0.94)
          7. workclass=_Private education-num=_peu-eduque 76 ==>
gain=_<=50K 71     conf:(0.93)
          8. age=_jeune capital-gain=_faible-gain-placement native-country
gain=_<=50K 67     conf:(0.93)
          9. marital-status=_Never-married capital-loss=_faible-perte-plac
gain=_<=50K 66     conf:(0.93)
         10. workclass=_Private education-num=_peu-eduque capital-loss=_
gain=_<=50K 65     conf:(0.93)
```

Elles sont toutes associées à l'attribut *gain=_ <=50K*. Ces personnes sont caractérisées entre autres comme étant jeune, d'un niveau de formation modeste, et n'ayant peu ou pas de gains à l'aide de placements.

Question 1.9

L'option *car* impose que l'attribut à droite dans les règles soit le gain ce qui facilite grande l'interprétation des règles puisque notre objectif était d'obtenir des informations relatives au gain.

2 Étude de cas : articles de presse

Question 2.1

Le nombre de mots considérés joue un rôle prépondérant dans les calculs au moment de la formation des itemsets fréquents.

Question 2.2

Dans notre modélisation, une transaction correspond à un article, les items correspondent aux mots. Nous avons écrit un script pour former un fichier *ARFF* à partir d'un ensemble d'articles (un article par ligne). Les attributs correspondent à la présence ou l'abscence de chacun des mots précédents. En appliquant la méthode Apriori, on espère former des règles qui nous indiquent le sujet d'un article en fonction des mots présents dans celui-ci.