AIRIAU Pierre-Marie, LERAY Maud

TP3 Acquisition de connaissances

Génération de règles d'association

1) Fouille de données sous Weka

1.2- Test de Weka avec l'exemple du golf

=== 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: weather.symbolic

Instances: 14

Attributes: 5

outlook

temperature

humidity

windy

play

=== Associator model (full training set) ===

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)

Qu.1.1) "Minimal support" correspond au "minsup" du cours. "Minimum metric <confidence>" représente le seuil de confiance. Ici par exemple, on cherche à déterminer les règles qui s'appliquent à au moins 15% des transactions, avec une confiance de plus de 90%. Les 4 lignes commençant par "Size of set..." représentent les "frequent itemsets" évoqués dans le cours, dont le support (nombre en bout de ligne) est supérieur au "minsup". Ensuite, les lignes 1 à 10 qui suivent "Best rules found" sont les règles d'association construites par l'algorithme (on remarque que leur confiance est toujours égale à 1).

Qu.1.2) Test supplémentaire : en baissant le "minMetric" à 0.8 au lieu de 0.9, en reprenant la mesure "Confidence", on obtient deux règles différentes (les 9 et 10), qui n'ont pas une confiance égale à 1 (une a 0.8 et une autre a 0.86). Les huit premières règles sont en revanche identiques au résultat de la qu.1.1. On comprend donc que les règles sont classées par "conf" décroissantes. Aussi, il y a maintenant un "large itemset" en moins. On constate aussi que le nombre de cycles effectués a diminué : il est passé de 17 à 15. Enfin, l'attribut "minimal support" est passé de 0.15 à 0.25, ce qui signifie que 25% des transactions respectent ces règles, à 80% de confiance.

En mettant le "minLift" à 1.1 et en choisissant la mesure "Lift", on observe 3 "large itemsets" au lieu de 4 à la qu.1.1, et le nombre de cycles est passé de 17 à 14. Le "minimal support" est maintenant de 0.3. Les règles trouvées sont très différentes de celles de la qu.1.1, et leurs confiances respectives sont variables : entre 0.44 et 1. Aussi, de nouveaux attributs sont affichés : lev ("leverage"), lift et conv ("conviction"). Les règles sont maintenant classées par "lift" décroissants, ceux-ci varient entre 1.33 et 2. Elles sont affichées ci-dessous.

Best rules found:

1. temperature=cool 4 ==> humidity=normal 4 conf:(1) < lift:(2)> lev:(0.14) [2] conv:(2)

2. humidity=normal 7 ==> temperature=cool 4 conf:(0.57) < lift:(2)> lev:(0.14) [2] conv:(1.25)

3. humidity=high 7 ==> play=no 4 conf:(0.57) < lift:(1.6)> lev:(0.11) [1] conv:(1.13)

4. play=no 5 ==> humidity=high 4 conf:(0.8) < lift:(1.6)> lev:(0.11) [1] conv:(1.25)

5. outlook=overcast 4 ==> play=yes 4 conf:(1) < lift:(1.56)> lev:(0.1) [1] conv:(1.43)

6. play=yes 9 ==> outlook=overcast 4 conf:(0.44) < lift:(1.56)> lev:(0.1) [1] conv:(1.07)

7. humidity=normal windy=FALSE 4 ==> play=yes 4 conf:(1) < lift:(1.56)> lev:(0.1) [1] conv:(1.43)

8. play=yes 9 ==> humidity=normal windy=FALSE 4 conf:(0.44) < lift:(1.56)> lev:(0.1) [1] conv:(1.07)

9. humidity=normal 7 ==> play=yes 6 conf:(0.86) < lift:(1.33)> lev:(0.11) [1] conv:(1.25)

10. play=yes 9 ==> humidity=normal 6 conf:(0.67) < lift:(1.33)> lev:(0.11) [1] conv:(1.13)

Qu.1.3) Prenons la règle n°3 : conf = 0.57, lift = 1.6 et lev = 0.11. Par le calcul, on trouve :

conf = 4/7 = 0.5714

lift = (4/14)/((7/14) * (5/14)) = (4*14)/(5*7) = 1.6

lev = (4/14) - ((7/14) * (5/14)) = 0.1071

où :

- 14 est le nombre d'instances ;

- 4 est le nombre d'instances ayant humidity=high ET play=no ;

- 7 est le nombre d'instances ayant humidity=high ;

- 5 est le nombre d'instances ayant play=no.

Qu.1.4) La conviction représente le ratio de la fréquence observée de prévisions fausses d'une règle si les attributs concernés sont indépendants, divisée par la fréquence observée de prévisions fausses.

1.3- Weka pour l'étude de la population américaine

(voir fichier adult_discretized_AIRIAU_LERAY.arff)

1.3.2- Fouille des données

Qu.1.9) On observe qu'en appliquant l'option car, toutes les règles trouvées sont fonctions du gain :

Best rules found:

1. marital-status= Never-married capital-gain-bin='(-inf-704.5]' capital-loss-bin='(-inf-670]' 67 ==> gain= <=50K 64 conf:(0.96)

2. age-bin='(-inf-31.5]' workclass= Private capital-gain-bin='(-inf-704.5]' 66 ==> gain= <=50K 63 conf:(0.95)

3. marital-status= Never-married capital-gain-bin='(-inf-704.5]' 74 ==> gain= <=50K 70 conf:(0.95)

4. age-bin='(-inf-31.5]' workclass= Private 70 ==> gain= <=50K 66 conf:(0.94)

5. age-bin='(-inf-31.5]' capital-gain-bin='(-inf-704.5]' native-country= United-States 72 ==> gain= <=50K 67 conf:(0.93)

6. marital-status= Never-married capital-loss-bin='(-inf-670]' 71 ==> gain= <=50K 66 conf:(0.93)

7. age-bin='(-inf-31.5]' capital-gain-bin='(-inf-704.5]' capital-loss-bin='(-inf-670]' native-country= United-States 68 ==> gain= <=50K 63 conf:(0.93)

8. marital-status= Never-married 78 ==> gain= <=50K 72 conf:(0.92)

9. age-bin='(-inf-31.5]' native-country= United-States 77 ==> gain= <=50K 71 conf:(0.92)

10. age-bin='(-inf-31.5]' capital-loss-bin='(-inf-670]' native-country= United-States 73 ==> gain= <=50K 67 conf:(0.92)