Introduction au data mining
L3 MIS - STA 1616 - 2010
V. Monbet
Règles d'association

L'objectif de ce TD est de comprendre l'algorithme Apriori ainsi que les mesures de qualité utilisées pour valider la génération de régles d'association.

1 A la main...

Les données que nous considérons ici portent sur des propriétés de champignons. Les variables sont les suivantes :

1. cap-surface `{fibrous,grooves,scaly,smooth}`
2. bruises `{bruises,no}`
3. gill-size `{broad,narrow}`
4. habitat `{grasses,leaves,meadows,paths,urban,waste,woods}`
5. poisonousness `{edible,poisonous}`

Et les données :

```
scaly,bruises,broad,waste,edible
smooth,no,narrow,woods,poisonous
fibrous,no,broad,grasses,edible
scaly,bruises,broad,woods,edible
scaly,no,narrow,leaves,poisonous
scaly,bruises,broad,paths,edible
smooth,no,broad,leaves,edible
scaly,no,broad,woods,poisonous
scaly,no,narrow,woods,poisonous
smooth,no,broad,leaves,edible
fibrous,no,broad,paths,poisonous
fibrous,bruises,broad,woods,edible
smooth,bruises,narrow,grasses,poisonous
fibrous,no,broad,paths,poisonous
smooth,bruises,narrow,grasses,poisonous
scaly,no,narrow,leaves,poisonous
```

```
scaly,no,narrow,woods,poisonous
fibrous,no,broad,grasses,edible
scaly,bruises,broad,woods,edible
fibrous,no,broad,grasses,edible
```

1. Trouvez, à la main, les règles d'association de cet ensemble de données en suivant l'algorithme Apriori. On choisira un support minimum de 25% et une confiance minimale de 90%. Commencez par générer les ensembles candidats fréquents niveau par niveau et une fois que vous avez généré tous les ensembles d'items fréquents produisez à partir de ceux ci les règles ayant une confiance suffisante. Montrez en détails toutes les étapes du procédé.
2. Dans l'algorithme `apriori`, la qualité des règles est mesurée par leur confiance et leur support. Dans un second temps, il est usuel de calculer aussi le **lift**. Utilisez l'article `evaluationEtValidationDeLInteretDesReglesDAssociation.pdf` pour trouver la définition du **lift**. Puis calculez le **lift** pour les règles que vous avez extraites. Discutez les résultats.

Notez que cet ensemble de données contient des mesures répétées et tenez en compte.

2 Avec R

Nous allons maintenant étudier le jeu de données `titanic.dat` en utilisant le package `R` intitulé `arules`. Le fichier `titanic.txt` décrit les données. Installer le package à partir du fichier `.zip`.

1. Lire les données

   ```
   T <- read.table(file.choose(),header=TRUE)
   dim(T)
   for (k in 1:4) T[,k] <- as.factor(T[,k])
   ```

2. Transformer les données de façon à ce qu'elles soient reconnues comme des transactions. En pratique on construit un tableau de données binaires.

   ```
   Titanic <- as(T, "transactions")
   Titanic
   summary(Titanic)
   ```

3. Tracer un diagramme en batons des items les plus fréquents, en choisissant par exemple un support minimum de 10% pour commencer.

   ```
   itemFrequencyPlot(Titanic, support = 0.1, cex.names = 0.8)
   ```

   Quel est le nombre de règles extraites ? Que se passe t'il si on fait varier le seuil du support ?

4. Utiliser l'algorithme Apriori pour extraire les meilleures règles. Vous choisirez de façon pertinente les seuils de support et de confiance.

```
rules <- apriori(Titanic, parameter = list(support = 0.01,confidence = 0.6))
rules
summary(rules)
```

Discuter les résultats

5. Extraire les règles contenant l'item `Survived=1`. Et expliquer les règles qui ont la plus grande confiance.

```
rulesSurvived <- subset(rules, subset = rhs %in% "Survived=1" & lift>1.6)
inspect(head(SORT(rulesSurvived, by = "confidence"), n = 3))
```

Traduire les résultats en français et les interpréter.

6. Répondre aux mêmes questions pour les règles contenant l'item `Survived=0`, en choisissant un seuil de lift à 1.