Cycle ingénieur : Deuxième année - Parcours MI

Etude de cas

19 octobre 2018

Applications de certains algorithmes de fouille données sur un problème de classification

Etude de cas avec R

Cette dernière partie nécessite la rédaction d'un programme R ainsi que la rédaction des réponses aux questions associées.

Soit l'ensemble de données joint nommé "flagsData.csv" décrivant un ensemble de drapeaux de certains pays en utilisant 30 attributs qui sont les suivants :

1. **name** : Name of the country concerned
2. **landmass** : 1=N.America, 2=S.America, 3=Europe, 4=Africa, 4=Asia, 6=Oceania
3. **zone** : Geographic quadrant, based on Greenwich and the Equator ; 1=NE, 2=SE, 3=SW, 4=NW
4. **area** : in thousands of square km
5. **population** : in round millions
6. **language** : 1=English, 2=Spanish, 3=French, 4=German, 5=Slavic, 6=Other Indo-European, 7=Chinese, 8=Arabic, 9=Japanese/Turkish/Finnish/Magyar, 10=Others
7. **religion** : 0=Catholic, 1=Other Christian, 2=Muslim, 3=Buddhist, 4=Hindu, 5=Ethnic, 6=Marxist, 7=Others
8. **bars** : Number of vertical bars in the flag
9. **stripes** : Number of horizontal stripes in the flag
10. **colours** : Number of different colours in the flag
11. **red** : 0 if red absent, 1 if red present in the flag

12. **green** : same for green
13. **blue** : same for blue
14. **gold** : same for gold (also yellow)
15. **white** : same for white
16. **black** : same for black
17. **orange** : same for orange (also brown)
18. **mainhue** : predominant colour in the flag (tie-breaks decided by taking the topmost hue, if that fails then the most central hue, and if that fails the leftmost hue)
19. **circles** : Number of circles in the flag
20. **crosses** : Number of (upright) crosses
21. **saltires** : Number of diagonal crosses
22. **quarters** : Number of quartered sections
23. **sunstars** : Number of sun or star symbols
24. **crescent** : 1 if a crescent moon symbol present, else 0
25. **triangle** : 1 if any triangles present, 0 otherwise
26. **icon** : 1 if an inanimate image present (e.g., a boat), otherwise 0
27. **animate** : 1 if an animate image (e.g., an eagle, a tree, a human hand) present, 0 otherwise
28. **text** : 1 if any letters or writing on the flag (e.g., a motto or slogan), 0 otherwise
29. **topleft** : colour in the top-left corner (moving right to decide tie-breaks)
30. **botright** : Colour in the bottom-left corner (moving left to decide tie-breaks)

Question 4.1 Lire ce fichier avec la commande

```
flags <-read.csv(file="flagsData.csv",sep=";")
```

Pour que cette commande marche, il faut que le fichier contenant les données soit dans le même répertoire de travail de R. Pour connaître le répertoire de travail de R, exécuter la commande

```
getwd()
```

Vous pouvez ensuite copier le fichier des données dans le répertoire de travail. une deuxième solution consiste à spécifier le répertoire de travail PATH [1] dans la commande ainsi :

```
flags <-read.csv(file="PATH/flagsData.csv",sep=";")
```

Effectuer ensuite les transformations de types suivantes :

```
flags$landmass<-as.factor(flags$landmass)
flags$zone<-as.factor(flags$zone)
flags$language<-as.factor(flags$language)
flags$religion<-as.factor(flags$religion)
flags$red<-as.factor(flags$red)
flags$orange<-as.factor(flags$orange)
flags$green<-as.factor(flags$green)
flags$blue<-as.factor(flags$blue)
flags$gold<-as.factor(flags$gold)
flags$white<-as.factor(flags$white)
flags$black<-as.factor(flags$black)
flags$crescent<-as.factor(flags$crescent)
flags$triangle<-as.factor(flags$triangle)
flags$icon<-as.factor(flags$icon)
flags$animate<-as.factor(flags$animate)
flags$text<-as.factor(flags$text)
```

Quels sont les types des attributs présents dans ce jeu de données ? Choisir une variable de chaque type et représenter graphiquement la distribution de ses valeurs. Si vous repérez un attribut identificateur, supprimez le.

Question 4.2 Notre but est d'apprendre les valeurs de la variable **landmass** en fonction des autres attributs (ou d'un sous ensemble). Diviser l'ensemble de données en un jeu d'apprentissage et un jeu de test en prenant 70% pour l'apprentissage et 30% pour le test. Vérifier que toutes les valeurs de la classe sont représentées dans les deux jeux.

1. Ici, PATH est une variable contenant le répertoire de travail trouvé par la commande getwd()

Question 4.3 A l'aide du package rpart, trouver l'arbre de décision appris à partir de l'ensemble d'apprentissage avec le paramètre de contrôle *minsplit* = 10. Représenter graphiquement cet arbre.

quels sont les attributs les plus discriminants pour la détermination de la classe ? Écrire l'ensemble des règles qui prédisent la sortie : *landmass=4*.

Question 4.4 Trouver la matrice de confusion sur les données de l'ensemble de test. Quelle classe a la meilleure précision ? Quelle est celle qui a le meilleur rappel ?

Question 4.5 Appliquer l'algorithme de Random Forest sur le jeux d'apprentissage, donner la précision du classifieur pour chaque valeur de classe en se basant sur l'ensemble d'apprentissage. A L'aide des deux fonctions

```
importance(rf)  varImpPlot(rf)
```

que vous appliquez au modèle résultant, trouver les dix meilleurs attributs selon le gain de Gini.

Question 4.6 Nous souhaitons appliquer l'algorithme k-means sur la totalité des données ; choisir parmi les dix attributs trouvés dans la question précédente, les attributs opportuns. Appliquer l'algorithme k-means sans prendre en compte la variable classe *landmass*, avec K=8, K=10,k=12. Relevez l'inertie totale de l'ensemble de données. Relevez l'inertie inter et l'inertie intra pour chacune des trois valeurs du paramètre k. Pour construire un dataFrame en prenant en compte certains attributs d'un autre data frame, vous pouvez utiliser :

```
newDF<-data.frame(oldDF$att1, .., oldDF$attk, ..)
```

Question 4.7 Trouver la matrice de correspondance classes-clusters pour k=10. Est-ce qu'il y a des clusters purs ? Est-ce qu'il a des drapeaux de classes différentes qui se ressemblent ? Commentez cette matrice.

Question 4.8 Nous souhaitons appliquer l'algorithme apriori sur la totalité des données ; choisir parmi les dix attributs trouvés dans la question 4.5, les attributs opportuns. Prenez en compte également la variable classe *landmass*. Triez les règles par rapport à leur confiance et relevez l'ensemble des règles obtenues. Trouvez ensuite les règles qui prédisent la sortie : *landmass=4* et comparez le résultat avec celui trouvé dans la question 3.