Classification associative

Alhousseynou BALL[*] and Bakary SIDIBE[*]

[*]Etudiants en M2 Datascience - Ecole Polytechnique X

11th March 2020

Contents

Introduction

Un des enjeux majeurs des scientifiques de données est de mettre en place de puissants algorithmes afin de synthétiser les informations pour optimiser la prise de décisions. C'est dans ce cadre que la classification associative a été mise en place. Elle est une sorte de modèle d'apprentissage supervisé qui utilise des règles d'association pour attribuer une valeur cible. L'article de Bertsimas et al. (2012) propose une nouvelle approche de Mixed Integer Optimization(MIO) appelée Ordered Rules for Classification pour cette tâche de classification. Cette méthode comporte deux parties:

- La partie explore une frontière particulière des solutions dans l'espace des règles
- La deuxième partie apprend à établir un classement optimal pour les règles permettant ainsi de construire une liste de décision qui est simple et perspicace.

L'objectif de notre document est de mettre en pratique cet algorithme sur des données réelles et puis évaluer ses performances.
Ce document est structuré en cinq principales parties. La première partie décrit brièvement la base de données et propose une représentation binaire des variables à retenir pour la prédiction de la variable cible. La deuxième partie porte sur la généralisation des règles de classification et l'ordonnancement de ces dernières. La troisième partie évalue les performances obtenues. La quatrième partie traitera des questions d'ouverture. Enfin, la cinquième partie donnera les étapes à faire afin de reproduire ce travail.

1 Données et représentation binaire

1.1 Présentation des données

Les données sont composées de 189 patients diagnostiqués ou non de la maladie rénale chronique(ckd or notckd). On cherche à prédire si une personne est atteinte de la maladie rénale chronique ou pas en se basant sur certaines variables.

variables	description
age	age
bp	blood pressure
sg	specific gravity
al	albumin
su	sugar
pc	pus cell
pcc	pus cell clumps
ba	bacteria

(a)

variables	description
bgr	blood glucose random
bu	blood urea
sc	serum creatinine
sod	sodium
pot	potassium
hemo	hemoglobin
pcv	packed cell volume
wbcc	white blood cell count

(b)

variables	description
rbcc	red blood cell count
htn	hypertension
dm	diabetes mellitus
cad	coronary artery disease
appet	appetite
pe	pedal edema
ane	anemia
class	class

(c)

Figure 1: Description Kidney data

	age	bp	sg	al	su	pc	...	appet	pe	ane	class
0	48	70	1.005	4	0	abnormal	...	poor	yes	yes	ckd
1	53	90	1.020	2	0	abnormal	...	poor	no	yes	ckd
2	63	70	1.010	3	0	abnormal	...	poor	yes	no	ckd
3	68	70	1.015	3	1	normal	...	poor	yes	no	ckd
4	68	80	1.010	3	2	abnormal	...	poor	yes	no	ckd

Figure 2: Les 5 premières observation

1.2 Représentation binaire

Avant d'appliquer un modèle de classification, il faut dans un premier temps choisir les variables pertinentes puis dans un second temps rendre les variables sélectionnées binaires. Pour le choix des variables pertinentes, une étude de corrélation entre la variable cible et les prédicateurs a été menée. Étant donné que les variables sont mixtes, deux mesures de corrélation sont utilisées: rapport de corrélation et la corrélation de Pearson. Une fois cette première sélection faite, nous allons étudier la multicollinearité entre les variables avec l'algorithme VIF pour sélectionner finalement les variables qui garantissent une bonne performance du modèle.

- rapport de corrélation: Le rapport de corrélation mesure le niveau de liaison entre une variable qualitative et une variable quantitative. Il correspond au rapport entre la variance entre les groupes (groupe de positifs et groupe de négatifs) et la variance totale. Sa valeur varie entre 0 et 1. Un rapport de corrélation proche de 1 révèle une forte corrélation entre les variables en jeu, et un rapport de corrélation proche de 0 révèle une faible liaison entre les variables. Le choix des variables à retenir se base sur le test de nullité du rapport de corrélation. Il suppose que la quantité $F = \frac{r(n-p)}{(p-1)(1-r)}$ suit une loi de Fischer de paramètres p-1 et n-p avec r le rapport de corrélation, n le nombre total d'individus et p le nombre de groupes. Si F est supérieur au quantile d'ordre 0,99 de la loi de Fisher à p-1 et n-p degrés de liberté, alors l'hypothèse nulle de nullité du rapport est rejetée au seuil de 1% et donc on conclut en la corrélation entre la variable à expliquer et la variable explicative. On a n=189 et p=2, alors le quantile d'ordre 0,99 vaut $F_{1,187}^{0.99} = 6.78$. En se référant sur la figure 3, la variable **pot** ne se sera pas considérée dans la modélisation.

Variables	Fisher
hemo	363.49
pcv	362.36
sg	272.41
al	253.46
rbcc	189.74
sc	98.07
sod	85.78

(a)

Variables	Fisher
bu	81.68
bgr	67.14
su	47.80
age	31.66
wbcc	26.91
bp	26.65
pot	2.40

(b)

Figure 3: Fisher corrélation

- corrélation de Pearson (Khi2): La corrélation de Pearson permet d'étudier la liaison entre deux variables qualitatives. Étant donné un tableau représentant la distribution jointe de deux variables, la statistique de test est donnée par : $\mathcal{X}^2 = n(\sum_{i,j} \frac{n_{i,j}^2}{n_{i.}n_{.j}} - 1)$ avec nij est l'effectif contenu dans case repérée par la ligne i et la colonne j, ni. est l'effectif marginal de la ligne i , n.j est celui de la colonne j, et n est l'effectif total. On remarque que, d'après la figure 4, l'ensemble des variables qualitatives sont corrélées avec la variable cible.

variables	stats	pvalue	criticals
htn	120.07	0.0	6.63
dm	96.59	0.0	6.63
pc	63.66	0.0	6.63
appet	46.02	0.0	6.63
pe	43.94	0.0	6.63
ane	33.90	0.0	6.63
pcc	28.16	0.0	6.63
cad	24.45	0.0	6.63
ba	19.04	0.0	6.63

Figure 4: Corrélation Khi2

- Variance inflation factor: Le VIF quantifie le niveau de multicolinéarité entre les variables à partir de la valeur du R^2 simple obtenue après régression multiple des variables explicatives sur la variable à expliquer. Il s'exprime par la relation : $VIF = \frac{1}{1-R^2}$ et la multicolinéarité est forte lorsque VIF > 10. En se référant à la figure 5, on n'a pas de problème de multicolinéarité forte. Toutes les variables restantes peuvent être utilisées pour la classification. Cependant, en nous basant sur la précision et le recall, nous avons choisi les variables qui minimisent l'erreur de classement. Ces variables sont: hemo, pcv, sg, htn, dm, pc.

variables	VIF
hemo	5.7880
pcv	5.6128
sc	4.7766
htn	4.6965
bu	4.4951
al	4.0704
bgr	3.6284
su	3.4614

(a)

variables	VIF
dm	3.4578
rbcc	3.1112
pc	2.8822
sg	2.6139
ane	2.5461
pe	2.5270
appet	2.0966

(b)

variables	VIF
pcc	1.9609
sod	1.8676
ba	1.7982
cad	1.5307
bp	1.3363
wbcc	1.3101
age	1.3092

(c)

Figure 5: Variance inflation factor

2 Généralisation des règles et ordonnancement

Dans un premier temps, il était question d'implémenter l'algorithme du RuleGen qui fournit les règles de classification. Le nombre total de règles obtenu est de 79.

Algorithme RuleGen

Entrées mincov, iter_lim

pour toute classe y faire

- $(\mathcal{R}_Y, \bar{s},\ \text{iter},\ c_{\max}) \leftarrow (\varnothing, 0, 1, n)$
- Répéter
  - si iter $= 1$ alors
    - $\bar{s}, b \leftarrow$ Résoudre $P\,(y, c_{\max})$
    - iter $=$ iter $+ 1$
  - $\mathcal{R}_Y \leftarrow \mathcal{R}_Y \cup b$
  - Ajouter la contrainte(*)
  - si iter $<$ iter_lim alors
    - stemp, $b \leftarrow$ Résoudre $P\,(y, c_{\max})$
    - si sTemp $< \bar{s}$ alors
      - $c_{\max} \leftarrow \min\,(c_{\max} - 1, \sum_{i=1}^{n} x_i)$
      - iter $\leftarrow 1$
    - sinon
      - iter $\leftarrow$ iter $+ 1$
  - sinon
    - $c_{\max} \leftarrow c_{\max} - 1$
    - iter $\leftarrow 1$
  - jusqu'à Cmax $< n$ mincov
- retourner $\mathcal{R}_Y$

Le nombre de règles du classifieur est donné par la figure 6.

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16
1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 6: Les règles du classifieur: kidney data

Interprétation :

- si dm = 'yes' et hemo ∈ [11.3-13.9[, on prédit que l'individu est malade('ckd')
- sinon, si htn = 'yes' et hemo ∈ [11.3-13.9[, on prédit que l'individu est malade('ckd')
- sinon, si sg ∈ [0-1.015[, on prédit que l'individu est malade('ckd')
- sinon, si hemo ∈ [0-11.3[, on prédit que l'individu est malade('ckd')
- sinon, si sg ∈ [1.015-1.02[, on prédit que l'individu est malade('ckd')
- sinon, on prédit que l'individu n'est pas malade('notckd')

Les caractéristiques de la machine sont les suivants et les temps de calculs sont les suivants:

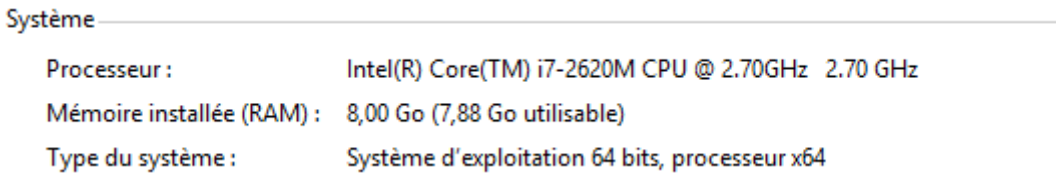
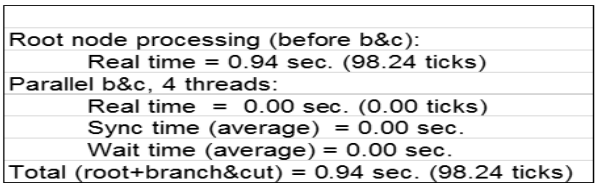
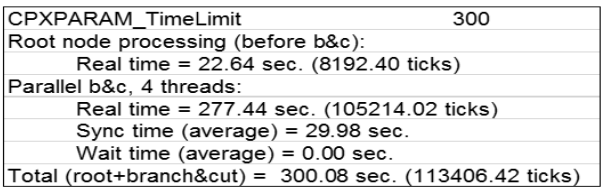


Figure 7: informations sur la machine



(a) Generating the rules



(b) Sorting the rules

Figure 8: Performance sur Kidney data

Le solveur utilisé est le CPLEX version 12.9.

3 Performances

Les performances du classifieur seront évaluées à partir de deux métriques: précision et recall.

Le classifieur prédit bien la classe des individus, car tous les individus étiquetés comme étant malades sont réellement malades.

Class	Prec.	Recall	Size
0	1.0	1.0	77
1	1.0	1.0	49
avg	1.0	1.0	
w. avg	1.0	1.0	

(a) Train results

Class	Prec.	Recall	Size
0	1.0	1.0	38
1	1.0	1.0	25
avg	1.0	1.0	
w. avg	1.0	1.0	

(b) Test results

Figure 9: Performance sur Kidney data

4 Questions d'ouvertures

4.1 performances du classifieur sur d'autres variables

Le classifieur donne de bons résultats avec les nouvelles variables. Cependant, comparé avec les résultats obtenus ci-dessus, il semble être moins performant lorsqu'on utilise d'autres vecteurs de caractéristiques. La précision et le recall ont diminué globalement de 2%.

Class	Prec.	Recall	Size
0	0.98	1.0	79
1	1.0	0.96	47
avg	0.99	0.98	
w. avg	0.98	0.98	

(a) Train results

Class	Prec.	Recall	Size
0	0.97	1.0	36
1	1.0	0.96	27
avg	0.99	0.98	
w. avg	0.98	0.98	

(b) Test results

Figure 10: Performance sur Kidney data

Le nombre de règles du classifieur sont devenus plus importants(12 règles contre 6).

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
1	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	1	0	0	0
1	0	0	0	0	1	0	0	0	0	0	0	1
0	1	0	0	0	1	0	0	0	0	0	0	0
0	1	0	0	1	0	0	0	0	1	0	0	0
1	0	0	0	0	0	1	0	0	0	0	1	0
1	0	1	0	0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	1	1
1	0	0	0	0	0	0	0	0	0	0	1	0
0	1	0	1	0	0	0	0	0	1	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 11: Les règles du classifieur: kidney data

4.2 Application sur d'autres données

La méthode décrite ci-dessus a été appliquée sur les données de vote du Congrès américain de 1984. La description des variables est fournie en annexe à la figure 13. La base de données disponible sur UCI Machine Learning comporte des valeurs manquantes. Ainsi, une imputation par le mode a été effectuée.
Le classifieur utilise les 6 règles de la figure 12a.

x1	x2	x3	x4	x5	x6
1	1	0	1	1	0
1	1	0	1	0	1
1	1	0	0	0	1
0	0	1	0	0	0
1	1	0	0	0	0
0	0	0	0	0	0

(a) règles du classifieur

Class	Prec.	Recall	Size
0	0.98	0.96	180
1	0.93	0.96	110
avg	0.95	0.96	
w. avg	0.96	0.96	

(b) Train results

Class	Prec.	Recall	Size
0	0.97	0.97	87
1	0.95	0.95	58
avg	0.96	0.96	
w. avg	0.96	0.96	

(c) Test results

Figure 12: Performance sur house votes

Interprétation

- si physician = 'yes', salvador = 'yes' et education = 'yes', on prédit que l'individu votera pour les républicains
- sinon, si physician = 'yes', salvador = 'yes' et nicaraguan = 'yes', on prédit que l'individu votera pour les républicains
- sinon, si physician = 'yes' et nicaraguan = 'yes', on prédit que l'individu votera pour les républicains
- sinon, si adoption = 'yes', on prédit que l'individu votera pour les démocrates
- sinon, si physician = 'yes', on prédit que l'individu votera pour les républicains
- sinon, on prédit que l'individu votera pour les démocrates

5 Étapes pour reproduire ce travail

Pour reproduire le travail, voici les étapes à suivre:

Installation

1. installer:
   (a) Julia
   (b) CPLEX (v12.8 ou v12.9): attention, la 12.10 n'est pas compatible avec Julia pour l'instant.
2. installer les packages: tapez "julia" dans une console, puis:
   (a) using Pkg
   (b) Pkg.add("JuMP")
   (c) Pkg.add("DataFrames")
   (d) Pkg.add("CSV")
   (e) ENV["CPLEX_STUDIO_BINARIES"] = "votre chemin d'installation de cplex (plus d'infos en dessous)"
   (f) Pkg.add("CPLEX")

Le chemin à indiquer pour CPLEX :

Linux
ENV["CPLEX_STUDIO_BINARIES"] = "/path/to/cplex/bin/x86-64_linux"
OSX
ENV["CPLEX_STUDIO_BINARIES"] = "/path/to/cplex/bin/x86-64_osx"
Windows
ENV["CPLEX_STUDIO_BINARIES"] = "C:/IBM/CPLEX_Studio128/cplex/bin/x64_win64"
‘‘‘

Kidney data

1. Uitliser les variables **class, htn, dm, hemo, pcv, sg, pc** puis les rendre binaire en respectant ces intervalles(quartiles):
   (a) hemo: [0.0-11.3[,[11.3-13.9[, []13.9-15.5[, [15.5-Inf[
   (b) pcv: [0.0-34.0[, [34.0-42.0[, [42.0-48.0[, [48.0-Inf[
   (c) sg: [0.0-1.015[,[1.015-1.02[,[1.02-1.025[, [1.025-Inf[
2. ouvrir **julia** puis se placer dans le dossier **src** et exécuter la commande **include('main.jl')**

Pour reproduire les questions d'ouvertures, appliquer la procédure décrite ci-dessus en utilisant ces vecteurs caractéristiques : **al, sc, rbcc, htn, dm** et la binarisation suivante:

- al: [0.0-1.0[, [1.0-Inf[
- sc: [0.0-0.8[, [0.8-1.1[, [1.1-2.2[, [2.2-Inf[
- rbcc: [0.0-4.0[, [4.0-4.8[, [4.8-5.5[, [5.5-Inf[

housevotes data

Les variables utilisées sont **physician,adoption,salvador,education,nicaraguan**

Conclusion

Dans ce projet, nous avons mis en pratique la méthode Ordered Rules for Classification (ORC) proposée par l'article de Bertsimas et al. (2012). L'avantage majeur de cette méthode par rapport à d'autres méthodes de classification en machine learning est son interpretabilité. De plus, les performances sont globalement très bonnes.

Annexes

variables	description
class	Class
hand	handicapped-infants
water	water-project-cost-sharing
adoption	adoption-of-the-budget-resolution
physician	physician-fee-freeze
salvador	el-salvador-aid
religious	religious-groups-in-schools
satellite	anti-satellite-test-ban
nicaraguan	aid-to-nicaraguan-contras

(a)

variables	description
missile	mx-missile
immigration	immigration
synfuels	synfuels-corporation-cutback
education	education-spending
superfund	superfund-right-to-sue
crime	crime
exports	duty-free-exports
administration	export-administration-act-south-africa

(b)

Figure 13: Description housevotes data

References

Bertsimas, D., A. Chang, and C. Rudin (2012). An Integer Optimization Approach to Associative Classification.