1 Introduction

Gestion des données manquantes dans SIPINA.

Attention, pour reproduire pleinement les opérations décrites dans ce didacticiel, assurez vous de disposer de la version 3.1 de la version recherche de Sipina¹. Vous pouvez vérifier cela en observant le numéro de version dans la barre de titre du logiciel lorsqu'il est démarré.



L'appréhension des données manquantes est un problème difficile. La gestion informatique en ellemême ne pose pas de problème, il suffit de signaler la valeur manquante par un code spécifique. En revanche, son traitement avant ou durant l'analyse des données est très compliqué. Il faut prendre en considération deux aspects²:

 La nature de la valeur manquante. Est-elle complètement aléatoire c.-à-d. toutes les valeurs ont la même probabilité d'être manquante (ex. les personnes qui omettent d'indiquer leur revenu dans une fiche signalétique ne répondent pas à des caractéristiques particulières). Est-elle conditionnellement aléatoire c.-à-d. dans certaines conditions, l'occurrence d'une valeur manquante suit un processus aléatoire (par exemple, il n'y a pas de données manquantes, sauf parmi les cadres

¹ La page de téléchargement du logiciel est <u>http://eric.univ-lyon2.fr/~ricco/sipina_download.html</u>; chargez et installez la SIPINA RESEARCH VERSION.

² Ce petit fascicule est très intéressant pour comprendre les tenants et aboutissants du problème des données manquantes : P.D. Allison, « Missing Data », in Quantitative Applications in the Social Sciences Series n°136, Sage University Paper, 2002.

où l'absence de réponse est distribuée aléatoirement). Est-elle non aléatoire (par exemple, le nombre de cases pour inscrire les chiffres est limité à 4, toutes les personnes qui ont un salaire supérieur à 9999 euros mensuel ne peuvent pas inscrire leur salaire³). On considère généralement que nous nous inscrivons dans la première situation pour pouvoir travailler, mais rien n'est moins sûr dans les études réelles.

• La technique statistique que nous mettons en œuvre par la suite. En effet, certaines méthodes de traitement des données manquantes sont plus ou moins adaptées selon les techniques statistiques que nous utilisons.

Prenons l'exemple de la suppression des lignes du tableau des données. Nous supprimons du fichier toutes les observations comportant au moins une valeur manquante (*listwise deletion* ou *casewise deletion* en anglais). L'approche paraît primaire, voire brutale. Nous pouvons réduire considérablement la taille du fichier ainsi. Pourtant, on montre qu'elle est plus robuste que les méthodes sophistiquées (maximum de vraisemblance, imputation multiple), en termes de biais et variance des estimations, lorsque la formation des valeurs manquantes s'écarte du processus complètement aléatoire et que nous implémentons une régression linéaire ou logistique (Allison, 2001; pages 84-85).

Prenons un autre exemple, nous utilisons les informations fournies par les autres descripteurs pour « deviner » les valeurs manquantes des variables (ex. un arbre de décision, une régression linéaire ; on parle d'imputation déterministe). Ce faisant, nous renforçons artificiellement le lien entre les variables. Toutes les méthodes statistiques qui s'appuient sur la matrice des corrélations sont faussées, les écartstype des coefficients sont sous-estimées dans la régression (Allison, 2001 ; pages 11-12).

TANAGRA ayant une vocation essentiellement pédagogique, je ne voulais pas introduire des outils automatisés de gestion des données manquantes. Il ne me paraissait pas souhaitable que l'étudiant puisse cliquer sur un bouton et évacuer ce problème négligemment. Il doit préparer ses données en ayant pleinement conscience de ce qu'il fait avant de pouvoir lancer un traitement statistique dans de bonnes conditions.

Je n'avais pas ce type de scrupule du temps de SIPINA, qui a été pour moi un véritable laboratoire à idées. Plusieurs techniques ont été implémentées, je les redécouvre moi-même aujourd'hui. L'objectif de ce tutoriel est de montrer leur mise en œuvre et les conséquences des choix sur l'induction des arbres de décision avec la méthode C4.5 (Quinlan, 1993).

2 Données

Notre fichier provient du site de Gilles Hunault de l'Université d'Angers⁴. On veut prédire le ronflement chez des individus à partir de leurs caractéristiques (âge, poids, taille, etc.). Nous en avons extrait 30

³ Il paraît qu'il y en a. Ils ne sont pas enseignants-chercheurs en tous cas.

⁴ <u>http://www.info.univ-angers.fr/~gh/Datasets/datasets.htm</u>

observations, puis nous avons supprimé quelques valeurs totalement au hasard.

Nous manipulons plusieurs fichiers dans ce didacticiel :

- RONFLEMENT_ALL.FDM est le fichier complet au format SIPINA, sans données manquantes. Nous nous en servirons pour élaborer l'arbre de décision de référence.
- RONFLEMENT_WITH_MISSING.FDM est le fichier avec données manquantes. C'est la traduction au format SIPINA du fichier texte ci-dessous. Nous l'utiliserons pour montrer les différentes stratégies de traitement des données manquantes.

L'ensemble des fichiers sont réunis dans une archive accessible en ligne⁵.

Nous montrons ici le fichier au format texte, les valeurs manquantes sont symbolisées par le caractère «?»

in ronfle	ement_avec_i	missing.txt - I	Bloc-notes			. O <mark>X</mark>	
Fichier	Edition Fo	ormat Affic	hage ?				
AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE	*
65	105	196	8	non	oui	oui	
49	76	164	0	non	non	non	
35	108	194	0	non	oui	non	
51	100	190	3	non	non	oui	
66	93	182	?	?	oui	oui	
?	96	186	3	non	oui	non	
74	108	194	5	non	?	oui	
53	104	194	5	non	oui	oui	
40	112	193	?	non	oui	non	
46	110	196	0	non	?	non	
?	81	169	7	non	oui	oui	
68	108	194	0	oui	non	oui	
41	?	166	0	non	oui	non	
71	76	164	4	non	non	oui	
38	74	161	8	non	oui	oui	
48	91	180	?	oui	?	oui	
62	68	165	4	non	oui	non	
56	?	164	7	non	noņ	oui	
33	98	188	0	?	oui	non	
69	107	198	3	non	oui	non	
43	108	194	3	non	oui	non	
38	42	161	4	non	oui	non	
2	90	?	0	oui	?.	non	
64	54	159	4	?	oui	oui	
41	61	167	6	non	oui	oui	
61	98	188	0	non	noņ	ou1	
5/	60	166	4	2	oun	non	
39	2	196	3	non	noņ	non	
55	83	1/1	10	non	oui	non	
69	10/	198	2	non	oun	oun	
							Ŧ
•						•	зł

Nous avons préparé les fichiers pour faciliter les manipulations dans ce didacticiel. Mais notons que SIPINA sait importer des fichiers de données au format texte avec séparateur tabulation comportant des données manquantes. Il suffit de les symboliser par le caractère « ? » ou de laisser l'emplacement vide.

De même nous pouvons utiliser la macro complémentaire pour envoyer les données d'Excel vers SIPINA⁶. Il faut simplement laisser la cellule vide lorsque l'on a une valeur manquante. Un classeur au format XLS accompagne les données pour que l'utilisateur puisse réaliser lui même les tests.

⁵ <u>http://eric.univ-lyon2.fr/~ricco/dataset/ronflement_missing_data.zip</u>

⁶ http://tutoriels-data-mining.blogspot.com/2008/03/connexion-excel-sipina.html

3 Traitement de la base complète

3.1 Chargement des données

Après avoir démarré SIPINA, nous chargeons le fichier complet RONFLEMENT_ALL.FDM en actionnant le menu FILE / OPEN.

File	Edit Data Statistics Induction	method Analy	sis View Wind	low Help			
	New						
	Open	Learning set ed	litor			^	
	Save	- lvor 1					
	Save as	St Ouvrir					
	Export	Regarder dans :	🍌 missing_data		• • •		
	Subsample management	9	Nom Da	te de m Type th missing.fdm	Taille	Nots-clés	
	Exit	Emplacements récents	ronflement_all	.fdm			
Lear Meth Meth Hdl=	ning method odName=Improved ChAID (Tsc A odClassName=TArbreDecisionI	Bureau					
Merg Split= Type	e=0.05 :0.001 Bonferroni=1	Maison				/	
Value Samp Exar	Bonterron=1 bling=0 mples selection						
1 exa 0 exa	mples selected mples idle	Ordinateur					
 Impro	ved ChAID (Tschuprow Goodness of	Réseau	Nom du fichier :	ronflement_all.fdm	6	•	Ouvrir
			Types de fichiers :	Data Manager File(*	.FDM)	-	Annuler

Les 30 observations sont affichées dans la grille de données.

AGE POIDS TAILLE ALCOUL FEMME TABAC RONFLE Nthrbute selection 65.00 105.00 196.00 8.00 non oui oui 2 49.00 76.00 164.00 0.00 non oui oui oui 3 35.00 108.00 194.00 0.00 non oui non 4 51.00 100.00 190.00 3.00 non oui non 5 66.00 93.00 182.00 5.00 non oui	File Edit Data Statistics I	nduction r	nethod An	alysis View	Window	Help			- 8
AGE POIDS TAILLE ALCOOL FEMME TABAC RONFLE Nithbute selection 1 65.00 105.00 196.00 8.00 non oui oui oui oui oui oui non	5 🗄 🗳 🕞								
Attribute selection 1 65.00 105.00 196.00 8.00 non oui oui 2 49.00 76.00 164.00 0.00 non non non 3 35.00 108.00 194.00 0.00 non oui non 4 51.00 100.00 190.00 3.00 non oui non 5 66.00 93.00 182.00 3.00 non oui oui 6 70.00 96.00 186.00 3.00 non oui oui 7 74.00 108.00 194.00 5.00 non oui oui 9 40.00 112.00 198.00 5.00 non oui non 11 40.00 81.00 199.00 0.00 non oui non 12 68.00 108.00 194.00 0.00 non oui non 14 71.00 76.00		1	AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
earning method 1 40.00 76.00 164.00 0.00 non non non 3 35.00 108.00 194.00 0.00 non oui non 4 51.00 100.00 190.00 3.00 non oui oui 5 66.00 93.00 182.00 5.00 non oui oui 6 70.00 96.00 186.00 3.00 non oui oui oui 8 53.00 104.00 194.00 5.00 non oui oui 9 40.00 112.00 193.00 5.00 non oui oui 11 40.00 81.00 169.00 7.00 non oui oui 12 68.00 108.00 194.00 0.00 non oui oui 12 68.00 108.00 196.00 0.00 non oui oui oui 13	attribute selection	1	65.00	105.00	196.00	8.00	non	oui	oui
searing method 3 35.00 108.00 194.00 0.00 non oui non 4 51.00 100.00 190.00 3.00 non non oui oui 6 70.00 96.00 186.00 3.00 non oui oui 6 70.00 96.00 186.00 3.00 non oui oui 7 74.00 108.00 194.00 5.00 non oui oui oui 9 40.00 112.00 193.00 5.00 non oui non 10 46.00 110.00 196.00 7.00 non oui non 12 68.00 108.00 194.00 0.00 non non oui 13 41.00 69.00 166.00 0.00 non non oui 15 38.00 74.00 168.00 0.00 oui oui non 16		2	49.00	76.00	164.00	0.00	non	non	non
earning method 4 51.00 100.00 190.00 3.00 non non oui 5 66.00 93.00 182.00 5.00 non oui oui 6 70.00 96.00 188.00 3.00 non oui oui 7 74.00 108.00 194.00 5.00 non oui oui 9 40.00 112.00 193.00 5.00 non oui non 10 46.00 110.00 196.00 0.00 non oui non 11 40.00 81.00 169.00 7.00 non oui non 12 68.00 108.00 194.00 0.00 non oui non 13 41.00 68.00 166.00 0.00 non oui non 14 71.00 76.00 164.00 4.00 non non oui non 15 38.00		3	35.00	108.00	194.00	0.00	non	oui	non
samples selection 5 66.00 93.00 182.00 5.00 non oui oui 6 70.00 96.00 186.00 3.00 non oui non 7 74.00 108.00 194.00 5.00 non oui oui 9 40.00 112.00 193.00 5.00 non oui oui 9 40.00 110.00 196.00 0.00 non oui non 10 46.00 110.00 196.00 0.00 non oui non 11 40.00 81.00 169.00 7.00 non oui non 12 68.00 108.00 194.00 0.00 non oui non 13 41.00 69.00 166.00 0.00 non non oui oui 14 71.00 76.00 166.00 non non oui oui non 15		4	51.00	100.00	190.00	3.00	non	non	oui
earning method 6 70.00 96.00 186.00 3.00 non oui non 7 74.00 108.00 194.00 5.00 non oui oui oui 8 53.00 104.00 194.00 5.00 non oui non oui oui non oui oui oui oui oui oui oui oui oui non oui o		5	66.00	93.00	182.00	5.00	non	oui	oui
P 74.00 108.00 194.00 5.00 non oui oui 8 53.00 104.00 194.00 5.00 non oui oui 9 40.00 112.00 193.00 5.00 non oui non 10 46.00 110.00 196.00 0.00 non oui non 11 40.00 81.00 166.00 0.00 non oui non 12 68.00 108.00 194.00 0.00 non oui non 13 41.00 69.00 166.00 0.00 non non oui non 14 71.00 76.00 164.00 8.00 non noi oui noi 15 38.00 74.00 168.00 0.00 noi oui noi 16 48.00 91.00 180.00 0.00 noi noi non 17 62.00		6	70.00	96.00	186.00	3.00	non	oui	non
8 53.00 104.00 194.00 5.00 non oui oui 9 40.00 112.00 193.00 5.00 non oui non 10 46.00 110.00 196.00 0.00 non oui non 11 40.00 81.00 169.00 7.00 non oui oui 12 68.00 108.00 194.00 0.00 oui non oui oui 13 41.00 69.00 166.00 0.00 non oui oui oui 14 71.00 76.00 164.00 4.00 non oui		7	74.00	108.00	194.00	5.00	non	oui	oui
9 40.00 112.00 193.00 5.00 non oui non 10 46.00 110.00 196.00 0.00 non oui non 11 40.00 81.00 169.00 7.00 non oui oui oui 12 68.00 108.00 194.00 0.00 oui non oui oui non oui oui non oui oui non oui oui oui oui oui oui oui oui oui non oui oui oui oui oui oui oui non oui oui<		8	53.00	104.00	194.00	5.00	non	oui	oui
10 46.00 110.00 196.00 0.00 non oui non 11 40.00 81.00 169.00 7.00 non oui oui 12 68.00 108.00 194.00 0.00 oui non oui 13 41.00 69.00 166.00 0.00 non oui non 14 71.00 76.00 164.00 4.00 non oui oui 15 38.00 74.00 161.00 8.00 non oui oui oui 16 48.00 91.00 180.00 0.00 oui oui oui 17 62.00 68.00 164.00 7.00 non oui oui 19 33.00 98.00 188.00 0.00 non oui non 20 69.00 107.00 198.00 3.00 non oui non 21 43.00 90.00		9	40.00	112.00	193.00	5.00	non	oui	non
11 40.00 81.00 169.00 7.00 non oui oui 12 68.00 108.00 194.00 0.00 oui non oui 13 41.00 69.00 166.00 0.00 non oui non oui 14 71.00 76.00 164.00 4.00 non non oui oui 15 38.00 74.00 161.00 8.00 non oui oui oui 16 48.00 91.00 180.00 0.00 oui oui oui non 17 62.00 68.00 165.00 4.00 non non oui non 18 56.00 58.00 168.00 0.00 oui oui non 20 69.00 107.00 198.00 3.00 non oui non 21 43.00 108.00 194.00 3.00 non oui non </td <td></td> <td>10</td> <td>46.00</td> <td>110.00</td> <td>196.00</td> <td>0.00</td> <td>non</td> <td>oui</td> <td>non</td>		10	46.00	110.00	196.00	0.00	non	oui	non
12 68.00 108.00 194.00 0.00 oui non oui 13 41.00 69.00 166.00 0.00 non oui non 14 71.00 76.00 164.00 4.00 non oui oui 15 38.00 74.00 161.00 8.00 non oui oui 16 48.00 91.00 180.00 0.00 oui oui oui 17 62.00 68.00 166.00 4.00 non oui oui oui 18 56.00 58.00 164.00 7.00 non non oui non 20 69.00 107.00 198.00 3.00 non oui non 21 43.00 108.00 194.00 3.00 non oui non 22 49.00 90.00 179.00 0.00 non oui non 23 49.00		11	40.00	81.00	169.00	7.00	non	oui	oui
13 41.00 69.00 166.00 0.00 non oui non 14 71.00 76.00 164.00 4.00 non oui oui oui 15 38.00 74.00 161.00 8.00 non oui non oui oui oui oui oui non oui oui non oui oui non oui oui oui oui non oui non oui oui non oui non oui non oui non oui non oui oui non oui		12	68.00	108.00	194.00	0.00	oui	non	oui
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii		13	41.00	69.00	166.00	0.00	non	oui	non
15 38.00 74.00 161.00 8.00 non oui oui 16 48.00 91.00 180.00 0.00 oui oui oui oui 17 62.00 68.00 166.00 4.00 non oui oui oui 18 56.00 186.00 164.00 7.00 non non oui non 19 33.00 98.00 188.00 0.00 oui oui non 20 69.00 107.00 198.00 3.00 non oui non 21 43.00 108.00 194.00 3.00 non oui non 22 49.00 90.00 179.00 0.00 oui non oui 23 49.00 90.00 179.00 0.00 non oui oui 24 64.00 54.00 159.00 4.00 non oui oui oui		14	71.00	76.00	164.00	4.00	non	non	oui
16 48.00 91.00 180.00 0.00 oui non 17 62.00 68.00 166.00 7.00 non oui oui non oui oui non oui oui oui oui		15	38.00	74.00	161.00	8.00	non	oui	oui
17 62.00 68.00 165.00 4.00 non oui non 18 56.00 58.00 164.00 7.00 non non oui non 19 33.00 98.00 188.00 0.00 oui oui non 20 69.00 107.00 198.00 3.00 non oui non 21 43.00 108.00 194.00 3.00 non oui non 22 38.00 42.00 161.00 4.00 non oui non 24 64.00 54.00 159.00 4.00 non oui oui non 25 41.00 61.00 167.00 6.00 non oui oui oui 30ueBonferroni=1 40.00 57.00 60.00 166.00 4.00 oui oui non 27 57.00 60.00 166.00 4.00 oui oui non		16	48.00	91.00	180.00	0.00	oui	oui	oui
18 56.00 58.00 164.00 7.00 non non oui 19 33.00 98.00 188.00 0.00 oui oui non 20 69.00 107.00 198.00 3.00 non oui non 20 69.00 107.00 198.00 3.00 non oui non 20 69.00 107.00 198.00 3.00 non oui non ethodClassName=TArbreDecision 21 43.00 108.00 194.00 3.00 non oui non 23 49.00 90.00 179.00 0.00 oui non non 24 64.00 54.00 159.00 4.00 non oui oui 25 41.00 61.00 167.00 60.00 non oui oui 26 61.00 98.00 188.00 0.00 non non non non 27		17	62.00	68.00	165.00	4.00	non	oui	non
19 33.00 98.00 188.00 0.00 oui oui non 20 69.00 107.00 198.00 3.00 non oui non 20 69.00 107.00 198.00 3.00 non oui non ethod/Name=Improved Ch4ID [Tsc ethod[JassName=TArbreDecision] dH8 22 38.00 42.00 161.00 4.00 non oui non 22 49.00 90.00 179.00 0.00 oui non oui non 24 64.00 54.00 159.00 4.00 non oui oui oui 25 41.00 61.00 167.00 6.00 non oui		18	56.00	58.00	164.00	7.00	non	non	oui
20 69.00 107.00 198.00 3.00 non oui non aaming method 21 43.00 106.00 194.00 3.00 non oui non ethodName=Improved Ch4ID [Tsc ethodClassName=TArbreDecision] dl=8 22 38.00 42.00 161.00 4.00 non oui non 22 38.00 42.00 161.00 169.00 non oui non non 23 49.00 90.00 179.00 0.00 oui non non 24 64.00 54.00 159.00 4.00 non oui oui 25 41.00 61.00 167.00 6.00 non oui oui oui 26 61.00 98.00 188.00 0.00 non non non 27 57.00 60.00 166.00 4.00 oui non non xampling=0 29 55.00 83.00 171.00 <td< td=""><td></td><td>19</td><td>33.00</td><td>98.00</td><td>188.00</td><td>0.00</td><td>oui</td><td>oui</td><td>non</td></td<>		19	33.00	98.00	188.00	0.00	oui	oui	non
Parining method 21 43.00 108.00 194.00 3.00 non oui non ethodMame=Improved ChAID [Tsc ethodClassName=TArbreDecision] dH8 erge=0.05 pie=0.001 ppeB.001 23 49.00 90.00 179.00 0.00 oui non oui non 24 64.00 54.00 159.00 4.00 non oui		20	69.00	107.00	198.00	3.00	non	oui	non
Bit Dock Bit Dock	earning method	21	43.00	108.00	194.00	3.00	non	oui	non
ethodClassName=TArbreDecision1 dH-8 erge=0.05 pit=0.001 pit=0.001 pit=0.001 pit=0.001 pit=0.001 pit=0.001 pit=0.001 pit=0.001 anpling=0 23 49.00 90.00 179.00 0.00 oui non non 24 64.00 54.00 159.00 4.00 non oui	ethodName=Improved ChAID (Tsc 🔺	22	38.00	42.00	161.00	4.00	non	oui	non
24 64.00 54.00 159.00 4.00 non oui oui pile 0.001 25 41.00 61.00 167.00 6.00 non oui oui pypeBonterroni=1 ampling=0 7 57.00 60.00 166.00 4.00 oui oui oui 28 39.00 119.00 196.00 3.00 non oui non 30 69.00 107.00 198.00 2.00 non oui oui	ethodClassName=TArbreDecisionI	23	49.00	90.00	179.00	0.00	oui	non	non
Erge=1005 bit=0.001 25 41.00 61.00 167.00 6.00 non oui oui ypeBonterroni=1 alueBonterroni=1 alueBonterroni=1 alueBonterroni=1 alueBonterroni=1 alueBonterroni=1 alueBonterroni=1 alueBonterroni=1 alueBonterroni=1 alueBonterroni=1 alueBonterroni=1 alueBonterroni=1 alueBonterroni=1 27 57.00 60.00 166.00 non oui oui non 27 57.00 60.00 166.00 3.00 non non non 28 39.00 119.00 196.00 3.00 non non non 30 69.00 107.00 198.00 2.00 non oui oui	di=8 oor	24	64.00	54.00	159.00	4.00	non	oui	oui
PpeBonferroni=1 alueBonferroni=1 ampling=0 28 61.00 98.00 188.00 0.00 non non oui 27 57.00 60.00 166.00 4.00 oui oui non 28 39.00 119.00 196.00 3.00 non non non 29 55.00 83.00 177.00 10.00 non oui non 30 69.00 107.00 198.00 2.00 non oui oui	olit=0.001	25	41.00	61.00	167.00	6.00	non	oui	oui
27 57.00 60.00 166.00 4.00 oui oui non ampling=0 28 39.00 119.00 196.00 3.00 non non non samples selection 29 55.00 83.00 171.00 10.00 non oui non 30 69.00 107.00 198.00 2.00 non oui oui	ypeBonferroni=1	26	61.00	98.00	188.00	0.00	non	non	oui
ampling=U 28 39.00 119.00 196.00 3.00 non non non xamples selection 29 55.00 83.00 171.00 10.00 non oui non 30 69.00 107.00 198.00 2.00 non oui oui	alueBonferroni=1	27	57.00	60.00	166.00	4.00	oui	oui	non
Image: selection 29 55.00 83.00 171.00 10.00 non oui non 30 69.00 107.00 198.00 2.00 non oui oui	ampling=U	28	39.00	119.00	196.00	3.00	non	non	non
30 69.00 107.00 198.00 2.00 non oui oui	Examples selection	29	55.00	83.00	171.00	10.00	non	oui	non
		30	69.00	107.00	198.00	2.00	non	oui	oui

3.2 Choix de l'algorithme de traitement

L'étape suivante consiste à choisir l'algorithme d'apprentissage. Nous actionnons le menu INDUCTION METHOD / STANDARD ALGORITHM, nous choisissons la méthode C4.5 (Quinlan, 1993) et nous validons les paramètres par défaut. Nous noterons principalement que la méthode ne segmente pas un nœud si les feuilles subséquentes contiennent moins de 2 observations. Nous nous en souviendrons plus loin.

	💦 Sip	oina Rese	arch Version 3.	1		-				
	File	Edit D	ata Statistics	Induction method	Analysis	View W	indow	Help		
	B	B 🙀		Standard algor	ithm					
	1	Calasta	n induction me	× r		_		The second		
	Attribu	Select a	n induction me	hod			1			
	/	Inducti	on Graph Rule	Induction Neural netwo	ork Discrir	minant analys	is Dec	cision list Oth	ner	
		A limite	d search inductio (Quinlan - 1986)	n tree algorithm (Catlett -	1991)					
		GID3 (Cheng, Fayyad, Ir	ani & Qian - 1988)						
	\sim	ChAID	FANT 86 (Cestnik (Kass - 1980)	k, Kononenko & Bratko -	1986)					
-		C4.5 (0	luinlan - 1993) od C4 5 (Pokotor	aalala % Lalliob						
	Learn	Improv	ed ChAID (Tschu	prow Goodness of Split)		C4 E		X		
	Metho	Cost se	nsitive C4.5 (Rak All Decision Tre	:otomalala & Chauchat - e	2001)	.4.5 param	eters		2	
	Hdl=8			-		C.L. for pe	essimistic	c pruning		
	Merge Split=I							25 🚖		
	TypeE	<u> </u>			C4.					
	Samp					Size of lea	ves: 2	2 🔹		• • • •
ŕ	Exam									Annuler
				15	48.00	∟ ⊏ Sampling :				
1				< <u> </u>	140.00	 All data 	aset		4	
Ī	Impro	ved ChA	D (Tschuprow	Goodness of Split)		⊖ Simple		Size :	Exec.Time:0 m	
						C Balanc	ed F	5000	CONCERNMENT OF THE OWNER	
						io stratine	su la			
								🗸 ОК		

3.3 Définition du problème à traiter

Nous devons maintenant définir la variable à prédire et les variables explicatives. Nous actionnons le menu ANALYSIS / DEFINE CLASS ATTRIBUTES. Par glisser déposer, nous plaçons RONFLE en CLASS, les autres variables en ATTRIBUTES. Nous validons la sélection.

File Edit Data Statistics Induction m	nethod Analysis View Window Help	
<u>Ľ 🖹 👺 📭 – – – – – – – – – – – – – – – – – – </u>	Define class attribute	
Attribute selection	Learnin Set weight field Set priore	ME
Learning method	Set positive Learning Stop analys	Variables AGE POIDS TAILLE ALCOOL FEMME TABAC RONFLE
MethodName=C4.5 (Quinlan - 1993) MethodClassName=TArbreDecisionC45 Hdl=5 Confidence level=25 Leaf size=2 11	Classificatic Test	
Sampling=0 SamplingPart=5000	Error measu	
Examples selection 14 15 16	Feature sele	
	III Personnal t	🖌 ΟΚ 🛛 🖌 όρο

3.4 Arbre de décision

Il ne reste plus qu'à lancer les traitements. Nous actionnons le menu ANALYSIS / LEARNING. Nous obtenons l'arbre de décision.



Figure 1 - Arbre sur le fichier sans valeurs manquantes

Nous avons un arbre à 3 niveaux. Les variables déterminantes sont le TABAC, l'ALCOOL et l'AGE.

4 Traitement des valeurs manquantes

L'idée maintenant est de travailler à partir du fichier comportant des valeurs manquantes. Nous souhaitons étudier dans quelle mesure les options proposées par SIPINA pour les traiter nous éloignent de l'arbre ci-dessus (Figure 1) lorsque nous lançons la méthode C4.5.

Nous devons stopper l'analyse courante en actionnant le menu ANALYSIS / STOP ANALYSIS. Puis vider la grille de données en cliquant sur le menu FILE / NEW.

Nous pouvons charger le fichier comportant les valeurs manquantes en cliquant sur le menu FILE / OPEN. Nous sélectionnons cette fois-ci le fichier RONFLEMENT_WITH_MISSING.FDM.

🔉 Sipina Research Version 3.1 - [Lea	arning set	editor]	10.00	-	1.00.00			
🔉 File Edit Data Statistics I	nduction i	method An	alysis View	Window	Help			_ & ×
🖰 🖹 👺 📑								
		AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
Attribute selection	1	65.00	105.00	196.00	8.00	non	oui	oui
	2	49.00	76.00	164.00	0.00	non	non	non
	3	35.00	108.00	194.00	0.00	non	oui	non
	4	51.00	100.00	190.00	3.00	non	non	oui
	5	66.00	93.00	182.00			oui	oui
	6		96.00	186.00	3.00	non	oui	non
	7	74.00	108.00	194.00	5.00	non		oui
	8	53.00	104.00	194.00	5.00	non	oui	oui
	9	40.00	112.00	193.00		non	oui	non
	10	46.00	110.00	196.00	0.00	non		non
	11		81.00	169.00	7.00	non	oui	oui
	12	68.00	108.00	194.00	0.00	oui	non	oui
	13	41.00		166.00	0.00	non	oui	non
	14	71.00	76.00	164.00	4.00	non	non	oui
	15	38.00	74.00	161.00	8.00	non	oui	oui
	16	48.00	91.00	180.00		oui		oui
	17	62.00	68.00	165.00	4.00	non	oui	non
	18	56.00		164.00	7.00	non	non	oui
	19	33.00	98.00	188.00	0.00		oui	non
	20	69.00	107.00	198.00	3.00	non	oui	non
Learning method	21	43.00	108.00	194.00	3.00	non	oui	non
MethodName=C4.5 (Quinlan - 1993)	22	38.00	42.00	161.00	4.00	non	oui	non
MethodClassName=TArbreDecisionC45	23		90.00		0.00	oui		non
Hdl=5 Confidence louiel=25	24	64.00	54.00	159.00	4.00		oui	oui
Leaf size=2	25	41.00	61.00	167.00	6.00	non	oui	oui
Sampling=0	26	61.00	98.00	188.00	0.00	non	non	oui
SamplingPart=5000	27	57.00	60.00	166.00	4.00		oui	non
	28	39.00		196.00	3.00	non	non	non
Examples selection	29	55.00	83.00	171.00	10.00	non	oui	non
	30	69.00	107.00	198.00	2.00	non	oui	oui
	Editing	D:\DataMir	ing\Databas	es for minir	ng\dataset fo	r soft Attri	butes : 7	Examples : 30
C4.5 (Quinlan - 1993)			_			E	xec.Time : 10	5 ms. //

Figure 2 - Grille de données avec valeurs manquantes

Nous observons les cases vides dans la grille de données.

4.1 Suppression des observations

Une première stratégie, très simple, mais non dénuée d'intérêt malgré l'a priori négatif qui l'accompagne, est la suppression des observations comportant au moins une valeur manquante. Si les valeurs manquantes sont aléatoirement réparties avec une proportion assez faible, la stratégie est tout à fait viable. En revanche, si elles sont très nombreuses ou concentrées sur une des variables, l'approche a des conséquences catastrophiques, on peut littéralement « vider » le fichier de ses observations.

Nous actionnons le menu STATISTICS / MISSING DATA / DELETE EXAMPLES. Nous voyons immédiatement le tableau rétrécir, il ne reste plus que 16 observations, près de 50% de l'effectif initial.

🔉 Sipina Research Version 3.1 - [Lea	arning set eo	litor]	-		1.00.00			
🔉 File Edit Data Statistics I	nduction me	thod Ana	lysis View	Window	Help			_ & ×
🕂 🕂 🎆 🕞								
×		AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
Attribute selection	1	65.00	105.00	196.00	8.00	non	oui	oui
	2	49.00	76.00	164.00	0.00	non	non	non
	3	35.00	108.00	194.00	0.00	non	oui	non
	4	51.00	100.00	190.00	3.00	non	non	oui
	8	53.00	104.00	194.00	5.00	non	oui	oui
	12	68.00	108.00	194.00	0.00	oui	non	oui
I	14	71.00	76.00	164.00	4.00	non	non	oui
Learning method	15	38.00	74.00	161.00	8.00	non	oui	oui
MethodName=C4.5 (Quinlan - 1993)	17	62.00	68.00	165.00	4.00	non	oui	non
MethodClassName=TArbreDecisionC45	20	69.00	107.00	198.00	3.00	non	oui	non
Hdl=5 Confidence level=25	21	43.00	108.00	194.00	3.00	non	oui	non
Leaf size=2	22	38.00	42.00	161.00	4.00	non	oui	non
Sampling=0	25	41.00	61.00	167.00	6.00	non	oui	oui
SamplingPart=5000	26	61.00	98.00	188.00	0.00	non	non	oui
	29	55.00	83.00	171.00	10.00	non	oui	non
Examples selection	30	69.00	107.00	198.00	2.00	non	oui	oui
					1			
<u> </u>	Editing	D:\DataMini	ng\Database	es for minin	q∖dataset fo	r soft Attri	butes : 7	Examples : 16
C4.5 (Quinlan - 1993)							Exec.Time : 1	6 ms.

Qu'importe, nous réitérons l'analyse précédente. Nous définissons la variable à prédire et les variables prédictives (ANALYSIS / DEFINE CLASS ATTRIBUTES) puis nous lançons les traitements (ANALYSIS / LEARNING). Nous obtenons l'arbre suivant.



Finalement, malgré la réduction drastique du fichier de données, nous obtenons un arbre quasiidentique à l'arbre sur les données complètes (Figure 1). C4.5 n'a pas pu développer le nœud à droite parce qu'il dispose de trop peu d'observations, il ne peut pas produire des feuilles avec moins de 2

observations conformément au paramétrage par défaut de la méthode. Si nous forçons quand même la segmentation, nous obtiendrions l'arbre suivant.



C'est la copie conforme de l'arbre sur les données complètes. Seuls les effectifs sont différents, ce qui est tout à fait normal. Que les bornes de discrétisation soient les mêmes est un sacré coup de chance en revanche.

Lorsque nous utilisons la stratégie de suppression d'individus, nous constatons que les modifications sont minimes sur l'arbre produit dans la mesure où les valeurs manquantes sont aléatoirement répartis, et qu'ils sont proportionnellement faibles. En revanche, il faut modifier le paramétrage de la méthode pour l'adapter à la réduction des effectifs.

4.2 Remplacement des valeurs – A

Voyons ce qu'il se passe si nous utilisons des stratégies de remplacement des valeurs manquantes : la moyenne pour les variables quantitatives, la valeur la plus fréquente (le mode) pour les qualitatives.

Nous stoppons l'analyse courante (ANALYSIS / STOP ANALYSIS), puis vidons la grille (FILE / NEW). Nous rechargeons le fichier comportant des valeurs manquantes (FILE / OPEN). Nous retrouvons notre grille de départ (Figure 2).

Nous actionnons maintenant le menu STATISTICS / MISSING DATA / REPLACE BY VALUES. Sipina affiche une boîte de dialogue listant les variables comportant au moins une valeur manquante. On notera ainsi que la variable à prédire RONFLE est complètement renseignée, elle est absente de la liste.

🔉 File Edit Data	Statistics 1	Induction me	thod Ana	alysis View	Windov	v Help	- 8	×
🖰 🛅 🖺 🕞	Descri	ptive statistic	s 🕨					
Attribute selection	Transf	orm	۱.	POIDS	TAILLE	ALCOOL	FEMME	
Attribute selection	Missin	g data	+	Delete e	examples	-	non	
		3	35.00	Replace	by value	s		-
		4	51.00	100.00	190.00	3.00	non	
		5	66.00	93.00	182.00	Replace missing	dita	
		6		96.00	186.00	Penlace MD fo	r the followin	a variabl
		7	74.00	108.00	194.00	Replace MD To	i ule lollowi	iy vanabi
		8	53.00	104.00	194.00	AGE		
Learning method	1000	9	40.00	112.00	193.00	TAILLE		
MethodName=U4.5 (Quini MethodClassName=TArbr	ian - 1993) eDecisionC45	10	46.00	110.00	196.00	ALCOOL		
Hdl=5	00/00/30/1040	11		81.00	169.00	FEMME		
Confidence level=25		12	68.00	108.00	194.00	TADAL		
Leat size=2 Sempling=0		13	41.00		166.00			
Sampling=0 SamplingPart=5000		14	71.00	76.00	164.00			
		15	38.00	74.00	161.00			
Examples selection		16	48.00	91.00	180.00			
		17	62 00	68 00	165 00		🗸 ОК	
		Editing	D:\DataMin	ing\Database	s for mir	ning\dataset for	soft Attribu	u //
C4.5 (Quinlan - 1993)						Ex	ec.Time:0 n	n //

Nous sélectionnons toutes les variables puis nous cliquons sur OK. Pour chaque variable, Sipina propose une valeur de substitution. Pour les variables quantitatives, il propose la moyenne. Nous cliquons sur OK pour l'AGE.

Replace missing o	lata with
AGE	
Mean	53.037037
C Other	0
	X Annuler

Idem pour le POIDS, la TAILLE et l'ALCOOL.

Replace missing data with	Replace missing data with	Replace missing data with
POIDS	TAILLE	ALCOOL
Mean 89,555556	Mean 180.62069	© Mean 3.444444
C Other 0	C Other 0	C Other 0
🗶 <u>A</u> nnuler	🗶 Annuler 🚺 🗹 OK	Annuler

Lorsque nous arrivons au stade de la variable qualitative FEMME. Sipina propose par défaut de

26/10/2009

Sipina Research Version 3.1 - [Lea	arning set ed	itor]						
💸 File Edit Data Statistics I	nduction me	thod Ana	lysis View	Window	Help			
Y 🚯 🛤 🛤								
X		AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
Attribute selection	1	65.00	105.00	196.00	8.00	non	oui	oui
	2	49.00	76.00	164.00	0.00	non	non	non
	3	35.00	108.00	194.00	0.00	non	oui	non
	4	51.00	100.00	190.00	3.00	non	non	oui
	5	66.00	93.00	182.00	3.44	non	oui	oui
	6	53.04	96.00	186.00	3.00	non	oui	non
	7	74.00	108.00	194.00	5.00	non	oui	oui
	8	53.00	104.00	194.00	5.00	non	oui	oui
	9	40.00	112.00	193.00	3.44	non	oui	non
	10	46.00	110.00	196.00	0.00	non	oui	non
	11	53.04	81.00	169.00	7.00	non	oui	oui
	12	68.00	108.00	194.00	0.00	oui	non	oui
	13	41.00	89.56	166.00	0.00	non	oui	non
	14	71.00	76.00	164.00	4.00	non	non	oui
	15	38.00	74.00	161.00	8.00	non	oui	oui
	16	48.00	91.00	180.00	3.44	oui	oui	oui
	17	62.00	68.00	165.00	4.00	non	oui	non
	18	56.00	89.56	164.00	7.00	non	non	oui
	19	33.00	98.00	188.00	0.00	non	oui	non
	20	69.00	107.00	198.00	3.00	non	oui	non
	21	43.00	108.00	194.00	3.00	non	oui	non
earning method AethodNama-C4 5 (Quinlan, 1992)	22	38.00	42.00	161.00	4.00	non	oui	non
iethodClassName=TArbreDecisionC45	23	53.04	90.00	180.62	0.00	oui	oui	non
Idl=5	24	64.00	54.00	159.00	4.00	non	oui	oui
Confidence level=25	25	41.00	61.00	167.00	6.00	non	oui	oui
ear size=2 ampling=0	26	61.00	98.00	188.00	0.00	non	non	oui
amplingPart=5000	27	57.00	60.00	166.00	4.00	non	oui	non
	28	39.00	89.56	196.00	3.00	non	non	non
Examples selection	29	55.00	83.00	171.00	10.00	non	oui	non
	30	69.00	107.00	198.00	2.00	non	oui	oui

Il ne nous reste plus qu'à lancer les traitements (choisir la variable à prédire et les prédictives, lancer

La gi

FEMME		TABAC
Mode	1	Mode
C Choose value	1	C Choose va
C "Missing" value		C "Missing" v

Mode	1
C Choose value	1
C "Missing" value	
× Anr	nuler

Sipina

remplacer la valeur manquante par le mode, en l'occurrence FEMME = 1. Nous validons. Idem pour TABAC.

l'analyse). Nous obtenons l'arbre suivant.

Sipina Sipina Research Version 3.1 - [Decision tree...] - 8 × 🔉 Induction method Analysis Tree management View Window Help 及 🗕 🙀 🖻 = × * Attribute selection 🖃 🗊 Class attribute . 15 (50%) D RONFLE ui Ξ 15 (50%) 🖻 🐨 Predictive attributes TABAC AGE POIDS in [oui] in [nen] 10 (43%) (71%) 5 13 (57% 2 (29% Learning method MethodName=C4.5 (Quinlan - 1993) MethodClassName=TArbreDecisionC45 ALCOOL AGE < 4.50 =4.50 50.00 =50.00 (100%) (25%) (86%) 0 (00%) Hdl=5 4 6 5 Confidence level=25 12 (75% (14% 2 (100% 0 (00%) Leaf size=2 Sampling=0 AGE 63.00 >=63.00 SamplingPart=5000 3 (08%) (75%) 1 11 (92% (25%) Examples selection 30 examples selected 0 examples idle • C4.5 (Quinlan - 1993) Exec.Time: 31 ms

Les trois premiers niveaux sont identiques à notre arbre de référence (Figure 1). Les effectifs sont par contre différents. Les modifications introduites dans les données incitent C4.5 à introduire une segmentation supplémentaire dans la partie gauche de l'arbre. Après coup, ça paraît évident. Nous avons artificiellement favorisé TABAC = OUI (augmenté son poids, il s'y trouve un individu supplémentaire) en remplaçant les valeurs manquantes avec le mode. L'algorithme pense pouvoir trouver des informations intéressantes de ce côté.

4.3 Remplacement des valeurs – B

Voyons maintenant une autre stratégie de remplacement. Nous utilisons la moyenne toujours pour les variables quantitatives, nous créons une nouvelle modalité « valeur manquante » (_MISSING_) pour les variables qualitatives.

De nouveau, nous stoppons les traitements (ANALYSIS / STOP ANALYSIS) et vidons la grille (FILE / NEW). Nous rechargeons le fichier RONFLEMENT_WITH_MISSING.FDM pour revenir à notre point de départ (Figure 2). Nous cliquons sur STATISTICS / MISSING DATA / REPLACE BY VALUES. La boîte de dialogue indiquant les variables comportant des valeurs manquantes est affichée. Nous les sélectionnons toutes puis nous validons.

Replace missing data
Replace MD for the following variables
AGE POIDS TAILLE ALCOOL FEMME TABAC
✓ ок

Encore une fois, nous utilisons la moyenne pour les variables continues AGE, POIDS, TAILLE et ALCOOL.

| Replace missing data with |
|---------------------------|---------------------------|---------------------------|---------------------------|
| AGE | POIDS | TAILLE | ALCOOL |
| ☞ Mean | @ Mean 89.555556 | | © Mean 3.4444444 |
| C Other 0 | C Other 0 | C Other 0 | C Other 0 |
| X Annuler | 🗶 Annuler | K Annuler K | 🗶 Annuler |

Pour les variables qualitatives, nous sélectionnons maintenant l'option « MISSING » VALUE.

Replace missing data with		Replace missing data with
C Mode		C Mode 1
C Choose value 1	-	C Choose value 1
		• "Missing" value
🗙 Annuler 🛛 🗸 OK		🗶 Annuler 🛛 🚺 🗸 OK

Nous observons une nouvelle grille de données. La modalité «_MISSING_ » saute aux yeux dans les colonnes des variables qualitatives. Le nombre total d'observations n'est pas modifié.

🕅 File Edit Data Statistics I	nduction	method An	alysis View	Window	Help			-
í 🖹 🖺 🛸								
+ <u>×</u>		AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
Attribute selection	1	65.00	105.00	196.00	8.00	non	oui	oui
	2	49.00	76.00	164.00	0.00	non	non	non
	3	35.00	108.00	194.00	0.00	non	oui	non
	4	51.00	100.00	190.00	3.00	non	non	oui
	5	66.00	93.00	182.00	3.44	missing	oui	oui
	6	53.04	96.00	186.00	3.00	non	oui	non
	7	74.00	108.00	194.00	5.00	non	missing	oui
	8	53.00	104.00	194.00	5.00	non	oui	oui
	9	40.00	112.00	193.00	3.44	non	oui	non
	10	46.00	110.00	196.00	0.00	non	missing	non
	11	53.04	81.00	169.00	7.00	non	oui	oui
	12	68.00	108.00	194.00	0.00	oui	non	oui
	13	41.00	89.56	166.00	0.00	non	oui	non
	14	71.00	76.00	164.00	4.00	non	non	oui
	15	38.00	74.00	161.00	8.00	non	oui	oui
	16	48.00	91.00	180.00	3.44	oui	missing	oui
	17	62.00	68.00	165.00	4.00	non	oui	non
	18	56.00	89.56	164.00	7.00	non	non	oui
	19	33.00	98.00	188.00	0.00	missing	oui	non
	20	69.00	107.00	198.00	3.00	non	oui	non
in	21	43.00	108.00	194.00	3.00	non	oui	non
earning method tathadMassa-C4 E (Quislan 1993)	22	38.00	42.00	161.00	4.00	non	oui	non
1ethodClassName=TArbreDecisionC45	23	53.04	90.00	180.62	0.00	oui	missing	non
ldl=5	24	64.00	54.00	159.00	4.00	missing	oui	oui
Confidence level=25	25	41.00	61.00	167.00	6.00	non	oui	oui
ear size=∠ ampling=0	26	61.00	98.00	188.00	0.00	non	non	oui
amplingPart=5000	27	57.00	60.00	166.00	4.00	missina	oui	non
	28	39.00	89.56	196.00	3.00	non	non	non
Examples selection	29	55.00	83.00	171.00	10.00	non	oui	non
	30	69.00	107.00	198.00	2.00	non	oui	oui

Il ne reste plus qu'à relancer les traitements (Définir la variable à prédire et les explicatives, puis lancer l'apprentissage). Nous obtenons un nouvel arbre.



L'arbre ressemble peu ou prou à celui construit sur la totalité des données. Nous notons qu'une nouvelle branche s'est formée sur la droite, composée à partir de la valeur _MISSING_ de TABAC. C4.5 en extrait une segmentation parfaite, avec des feuilles pures. C'est typiquement un artefact. Les valeurs manquantes ayant été introduites totalement au hasard, cette nouvelle règle n'est certainement pas reproductible sur un autre fichier.

5 Conclusion

N'allons surtout pas tirer des conclusions définitives à partir d'un petit exemple didactique. Bien sûr d'autres techniques existent. Certaines sont de bon sens : une colonne est quasi-vide, il faut virer la variable, une ligne est quasi-vide, il faut retirer les observations. D'autres sont plus techniques : la méthode du maximum de vraisemblance, l'imputation multiple (Allison, 2001 : chapitres 4 et 5).

Autre aspect important, nous avons privilégié l'analyse qualitative des résultats dans ce didacticiel, en comparant les arbres produits subséquemment (j'adore) au prétraitement des valeurs manquantes. Une approche plus mécanique est possible. On la retrouve souvent dans les publications. L'idée consiste à analyser les conséquences du traitement sur les performances du modèle de prédiction en généralisation. La démarche est schématiquement la suivante :

- Nous scindons une base sans valeurs manquantes en échantillon apprentissage (APP) et test (TEST).
- Nous créons un modèle à partir de APP, mesurons le taux d'erreur sur TEST. Ce sera la référence.
- Nous bruitons APP en retirant au hasard une certaine proportion de valeurs. Nous avons un fichier d'apprentissage avec données manquantes (APP-MD). On peut faire varier cette proportion.
- Nous appliquons sur APP-MD les différentes stratégies de traitement des données manquantes pour construire un modèle de prédiction dont nous mesurons les performances sur TEST. Nous avons ainsi une série de taux d'erreur que l'on peut comparer avec la référence. Nous écrivons un article pour montrer quelle est la meilleure méthode en traitant plusieurs bases UCI IRVINE.

C'est très bien. Il y a deux remarques à faire par rapport à ce schéma : on ne traite toujours que les valeurs manquantes totalement aléatoires dans ce cas ; l'autre piste à creuser est la gestion de données manquantes lorsque nous appliquons le modèle c.-à-d. lorsque les individus de l'échantillon fichier test eux-mêmes ne sont pas décrits complètement. Il existe des travaux très intéressants qui en parlent. L'idée mérite d'être creusée à mon avis. La gestion des données manquantes est au moins aussi importante lors de la phase de classement que lors de la phase d'apprentissage.