INSTITUT
POLYTECHNIQUE
DE PARIS

Le Cnam
Institut Polytechnique de Paris

Projet RODM

Auteur :
Justine De Sousa
Natalia Jorquera

Superviseurs:
Zachari Ales

le cnam

13 avril 2022

Introduction

Il s'agit de générer des arbres de décisions optimaux. Ils sont générés par résolution d'un programme linéaire.

Les trois jeux de données *iris*, *seeds* et *wine* sont données. On en a sélectionné deux nouveaux : *tic-tac-toe* dans lequel il s'agit de déterminer si une configuration est gagnante pour le premier joueur et *cmc* (Choix de Méthode de Contraception) dans laquelle il s'agit de prédire la méthode de contraception utilisée par une femme en fonction de certaines caractéristiques comme son âge ou son niveau d'étude.

On a décidé d'implémenter un callback. Pour ce faire, nous supprimons les contraintes qui contrôlent le flux de données vers le nœud gauche et le nœud droit, et nous les ajoutons au cours de la résolution, coupant ainsi les points entiers.

1 Résultats obtenus

On exécute le programme sur les cinq jeux de données et pour chacunes des méthodes de séparation et de regroupement.

1.1 Sans regroupement

1.1.1 Séparations univariées

Le Tableau 1 montre les résultats obtenus pour le cas univarié sans regroupement, où l'on peut voir que lors de l'utilisation du callback la résolution est plus lente, et présente également un GAP légèrement plus élevé.

Instance	**D**	**Classique**				**Callback**			
		Temps	GAP	Erreurs		Temps	GAP	Erreurs	
				Train set	Test set			Train set	Test set
iris	2	11.2	0.0%	5	1	36.9	0.0%	5	1
	3	60.1	0.8%	1	1	60.0	13.2%	5	1
	4	60.2	2.6%	3	2	60.1	1.2e8%	76	24
seeds	2	60.1	1.3%	10	3	60.1	14.7%	18	4
	3	60.2	8.4%	13	8	60.1	25.4%	20	4
	4	60.3	7.7%	12	9	60.2	5499.8%	115	25
wine	2	60.1	5.2%	7	2	60.0	10.2%	6	1
	3	60.2	2.9%	4	2	60.0	19.3%	19	6
	4	60.4	17.4%	21	6	60.1	153.6%	80	21
tic-tac-toe	2	60.3	53.5%	267	65	60.1	107.0%	395	103
	3	60.6	45.6%	240	48	60.4	45.6%	240	48
	4	60.6	53.5%	267	65	60.5	7.66e8%	267	65
cmc	2	60.2	120.2%	643	162	61.2	1.178e9%	677	167
	3	60.5	378.9%	932	243	60.3	1.178e9%	677	167
	4	61.0	1.178e9%	677	167	60.8	1.178e9%	677	167

Table 1 – Résultats sans regroupement et séparation univariée

1.1.2 Séparations multivariées

Le Tableau 2 montre les résultats obtenus en utilisant la séparation multivariée, où l'on constate à nouveau que le callback est plus lent que la méthode classique, et l'on peut même constater que la solution obtenue lorsque le temps maximum autorisé est respecté est pire, étant donné que son GAP est plus grand.

Notons, par ailleurs, que la méthode par séparation multivariée est plus performante. En effet, pour les 3 plus petites instances, l'optimum est atteint la plupart du temps. Quant aux plus grandes instances, les deux méthodes (univariée et multivariée) sont limitées par le temps imposé. On ne peut pas conclure concernant le jeu de données *tic-tac-toe*. En revanche, pour un arbre contenant plus de 2 branchements, le multivarié est plus proche de l'optimum.

Instance	D	Classique				Callback			
		Temps	GAP	Erreurs		Temps	GAP	Erreurs	
				Train set	Test set			Train set	Test set
iris	2	2.8	0.0%	1	3	12.2	0.0%	1	0
	3	3.8	0.0%	0	1	60.0	0.8%	0	1
	4	60.3	0.8%	1	3	60.1	0.8%	1	2
seeds	2	5.8	0.0%	0	1	6.6	0.0%	0	1
	3	19.3	0.0%	0	1	60.1	5.7%	8	3
	4	60.4	17.5%	3	2	60.1	1.68e8%	115	25
wine	2	1.5	0.0%	0	1	2.8	0.0%	0	1
	3	0.9	0.0%	0	1	15.8	0.0%	0	2
	4	26.6	0.0%	0	0	60.1	1.42e8%	96	23
tic-tac-toe	2	60.3	43.4%	223	51	60.1	7.66e8%	267	65
	3	60.6	53.5%	267	65	60.1	7.66e8%	267	65
	4	60.5	53.5%	267	65	60.3	7.66e8%	267	65
cmc	2	60.3	135.1%	677	167	60.1	187.3%	768	194
	3	61.3	135.1%	677	167	60.3	1.178e9%	677	167
	4	60.7	135.1%	677	167	60.7	1.178e9%	677	167

Table 2 – Résultats sans regroupement et séparation multivariée

1.2 Avec regroupement

1.2.1 Séparations univariées

Les résultats sont présentés dans le Tableau 3, où l'on peut voir que moins on utilise de clusters, plus la résolution est rapide. Cependant, elle présente plus d'erreurs que lorsqu'on utilise un grand nombre de clusters. Remarquons que ce dernier cas se rapproche de la méthode sans regroupement. On peut également voir que le callback est plus lent.

Instance	D	nb clusters	Classique				Callback			
			Temps	GAP	Erreurs		Temps	GAP	Erreurs	
					Train set	Test set			Train set	Test set
iris	2	3	0.1	0.0%	38	12	2.5	0.0%	38	12
		24	0.6	0.0%	5	1	0.5	0.0%	5	1
		48	1.5	0.0%	5	1	4.7	0.0%	5	1
		72	3.8	0.0%	5	1	8.5	0.0%	5	1
		96	13.5	0.0%	5	1	23.4	0.0%	5	1
		120	13.0	0.0%	5	1	60.0	2.6%	5	1
	3	3	0.1	0.0%	38	12	0.4	0.0%	38	12
		24	5.6	0.0%	4	1	17.6	0.0%	4	1
		48	4.7	0.0%	0	1	17.6	0.0%	0	1
		72	4.9	0.0%	0	4	24.2	0.0%	0	1
		96	44.9	0.0%	0	1	60.0	4.3%	2	1
		120	15.2	0.0%	0	1	60.0	6.2%	7	3
	4	3	0.1	0.0%	38	12	0.1	0.0%	38	12
		24	23.4	0.0%	3	1	60.0	2.6%	3	2
		48	5.8	0.0%	0	2	60.0	1.7%	2	1
		72	16.4	0.0%	0	1	60.0	44.6%	6	2
		96	25.5	0.0%	0	3	60.0	179.1%	40	12
		120	27.4	0.0%	0	4	60.1	7.1%	7	3
seeds	2	3	0.0	0.0%	53	17	0.0	0.0%	53	17
		33	1.1	0.0%	17	4	1.6	0.0%	17	4
		67	2.9	0.0%	14	3	7.9	0.0%	14	3
		100	9.1	0.0%	10	3	60.0	9.9%	17	5
		134	48.9	0.0%	10	3	60.0	17.1%	25	4
		168	60.1	3.8%	10	3	60.0	19.1%	20	4
	3	3	0.1	0.0%	53	17	0.0	0.0%	51	15
		33	17.0	0.0%	34	13	60.0	29.3%	26	6
		67	58.0	0.0%	6	3	60.0	12.8%	14	3
		100	60.1	6.3%	9	3	60.0	57.0%	23	10
		134	60.1	9.1%	14	4	60.0	24.4%	15	5
		168	60.1	9.1%	14	7	60.0	28.2%	26	9
	4	3	0.1	0.0%	53	17	0.2	0.0%	53	17
		33	46.8	0.0%	21	6	60.0	40.0%	25	7

4

		67	60.2	5.0%	5	3	60.4	71.4%	30	8
		100	60.2	9.8%	14	3	60.0	1.68e8%	115	25
		134	60.3	13.5%	17	3	60.1	1.68e8%	115	25
		168	60.3	20.0%	27	9	60.5	1.68e8%	115	25
wine	2	3	0.1	0.0%	56	15	0.0	0.0%	56	15
		28	1.2	0.0%	10	5	1.0	0.0%	8	4
		56	2.2	0.0%	7	3	8.5	0.0%	7	3
		85	24.4	0.0%	5	2	60.0	10.0%	5	2
		113	27.8	0.0%	6	2	60.0	16.4%	11	3
		142	60.1	2.9%	5	2	60.0	5.2%	5	2
	3	3	0.1	0.0%	56	15	0.1	0.0%	56	15
		28	6.6	0.0%	0	3	42.6	0.0%	0	2
		56	34.7	0.0%	1	4	60.0	10.9%	6	3
		85	60.1	8.4%	11	5	60.0	20.3%	17	2
		113	60.1	6.8%	6	1	60.0	25.7%	22	9
		142	60.1	3.6%	5	3	60.0	14.5%	10	5
	4	3	0.2	0.0%	56	16	0.4	0.0%	56	15
		28	18.7	0.0%	1	3	60.0	36.5%	16	10
		56	60.2	5.2%	2	3	60.0	36.5%	15	4
		85	60.2	6.0%	1	1	60.1	26.8%	13	4
		113	60.3	9.2%	7	3	60.1	153.6%	86	21
		142	60.3	13.6%	17	5	60.1	246.3%	74	15
tic-tac-toe	2	2	0.0	0.0%	267	65	0.0	0.0%	267	65
		153	2.2	0.0%	267	65	5.4	0.0%	267	65
		306	10.2	0.0%	267	65	60.2	8.6%	267	65
		459	60.2	15.1%	267	65	60.1	47.4%	267	65
		612	60.3	59.3%	267	65	60.1	200.4%	499	127
		766	60.2	45.6%	240	48	60.1	65.4%	296	77
	3	2	0.0	0.0%	267	65	0.1	0.0%	267	65
		153	8.8	0.0%	267	65	28.9	0.0%	267	65
		306	60.2	0.0%	267	65	60.1	493.8%	395	103
		459	60.4	128.7%	240	48	60.1	50.0%	267	65
		612	60.3	186.9%	499	127	60.1	7.66e8%	267	65
		766	60.4	72.5%	322	72	60.2	7.66e8%	267	65
	4	2	0.0	0.0%	267	65	0.0	0.0%	267	65
		153	20.5	0.0%	267	65	60.1	7.66e8%	267	65
		306	60.2	48.3%	267	65	60.2	7.66e8%	267	65
		459	60.3	53.5%	267	65	60.2	207.6%	499	127
		612	60.5	53.5%	267	65	60.2	7.66e8%	267	65
		766	60.6	7.66e8%	267	65	60.3	7.66e8%	267	65
cmc	2	3	0.0	0.0%	677	167	0.0	0.0%	677	167
		235	60.0	86.3%	676	166	60.0	135.1%	677	167
		471	60.2	134.7%	676	166	60.0	225.4%	715	172
		706	60.3	126.1%	655	157	60.1	344.5%	911	229
		942	60.4	135.1%	677	167	60.1	1.178e9%	677	167
		1178	60.4	124.8%	654	156	60.1	1.178e9%	677	167
	3	3	0.0	0.0%	677	167	0.1	0.0%	677	167
		235	60.2	134.7%	676	166	60.1	139.4%	677	167
		471	60.4	334.7%	907	228	60.2	1.178e9%	677	167
		706	60.4	172.7%	738	188	60.4	1.178e9%	677	167
		942	60.6	127.0%	659	175	60.3	1.178e9%	677	167
		1178	61.0	135.1%	677	167	60.2	1.178e9%	677	167
	4	3	0.1	0.0%	677	167	0.1	0.0%	677	167
		235	60.3	177.2%	676	166	60.2	1.178e9%	677	167
		471	60.7	135.1%	677	167	60.2	166.5%	677	167
		706	61.3	187.3%	768	194	60.3	1.178e9%	677	167
		942	61.7	1.178e9%	677	167	60.5	1.178e9%	677	167
		1178	62.3	1.178e9%	677	167	60.6	1.178e9%	677	167

Table 3 – Résultats avec regroupements et séparation univariée

1.2.2 Séparations multivariées

Le Tableau 4 montre que, comme dans le cas univarié, l'utilisation d'un plus grand nombre de clusters entraîne une résolution plus rapide mais plus d'erreurs de classification.

Instance	D	nb clusters	Classique				Callback			
			Temps	GAP	Erreurs		Temps	GAP	Erreurs	
					Train set	Test set			Train set	Test set
iris	2	3	0.4	0.0%	38	12	3.4	0.0%	38	12
		24	1.0	0.0%	1	0	2.3	0.0%	1	0
		48	1.3	0.0%	1	0	5.9	0.0%	1	3
		72	2.4	0.0%	1	3	10.5	0.0%	1	4
		96	2.4	0.0%	1	0	17.9	0.0%	1	0
		120	1.9	0.0%	1	3	31.1	0.0%	1	0
	3	3	0.5	0.0%	38	12	6.8	0.0%	38	12
		24	1.7	0.0%	0	4	3.9	0.0%	0	4
		48	1.6	0.0%	0	4	14.9	0.0%	0	3
		72	2.4	0.0%	0	0	48.3	0.0%	0	3
		96	6.5	0.0%	0	4	45.0	0.0%	0	4
		120	2.7	0.0%	0	4	60.1	0.8%	0	1
	4	3	1.4	0.0%	38	12	3.5	0.0%	38	12
		24	1.6	0.0%	0	0	1.4	0.0%	0	2
		48	4.0	0.0%	0	4	20.1	0.0%	0	0
		72	3.9	0.0%	0	0	60.0	10.1%	10	12
		96	9.5	0.0%	0	0	60.1	1.2e8%	76	24
		120	55.2	0.0%	0	1	60.1	0.8%	1	2
seeds	2	3	0.4	0.0%	53	17	0.3	0.0%	53	17
		33	1.5	0.0%	1	1	1.8	0.0%	1	3
		67	1.9	0.0%	0	0	2.1	0.0%	0	1
		100	1.6	0.0%	0	0	2.7	0.0%	0	0
		134	2.3	0.0%	0	1	8.2	0.0%	0	0
		168	5.2	0.0%	0	1	7.9	0.0%	0	1
	3	3	0.7	0.0%	53	17	6.1	0.0%	53	17
		33	1.6	0.0%	0	0	2.4	0.0%	0	0
		67	4.1	0.0%	0	2	22.5	0.0%	0	1
		100	4.3	0.0%	0	2	54.9	0.0%	0	1
		134	15.2	0.0%	0	3	60.0	44.8%	3	5
		168	20.7	0.0%	0	1	60.1	5.7%	8	3
	4	3	2.9	0.0%	53	17	9.8	0.0%	53	17
		33	7.2	0.0%	0	3	26.1	0.0%	0	2
		67	15.9	0.0%	0	3	60.0	6.3%	3	5
		100	36.3	0.0%	0	2	60.1	1.2%	1	3
		134	24.2	0.0%	0	3	60.1	1.68e8%	115	25
		168	60.1	7.7%	9	8	60.1	1.68e8%	115	25
wine	2	3	0.2	0.0%	0	2	0.1	0.0%	0	1
		28	0.7	0.0%	0	0	0.7	0.0%	0	1
		56	1.3	0.0%	0	1	0.4	0.0%	0	2
		85	1.0	0.0%	0	2	0.4	0.0%	0	2
		113	1.0	0.0%	0	1	5.0	0.0%	0	2
		142	1.2	0.0%	0	2	2.5	0.0%	0	1
	3	3	0.2	0.0%	0	1	0.5	0.0%	0	1
		28	2.7	0.0%	0	2	0.8	0.0%	0	3
		56	2.2	0.0%	0	1	1.0	0.0%	0	1
		85	4.8	0.0%	0	2	5.6	0.0%	0	2
		113	7.2	0.0%	0	1	36.2	0.0%	0	3
		142	4.2	0.0%	0	1	15.8	0.0%	0	2
	4	3	0.5	0.0%	0	1	0.3	0.0%	0	1
		28	7.5	0.0%	0	2	1.6	0.0%	0	1
		56	7.5	0.0%	0	2	28.0	0.0%	0	2
		85	11.7	0.0%	0	3	60.0	1.4%	1	9
		113	15.8	0.0%	0	0	60.0	1.42e8%	96	23
		142	39.0	0.0%	0	1	60.1	1.42e8%	96	23
tic-tac-toe	2	2	0.7	0.0%	267	65	0.5	0.0%	267	65
		153	60.2	18.4%	116	32	60.0	27.0%	143	31
		306	60.2	27.7%	154	36	60.1	53.5%	267	65
		459	60.2	28.7%	139	38	60.0	186.9%	499	127
		612	60.3	53.5%	267	65	60.1	186.9%	499	127

		766	60.5	53.5%	267	65	60.1	7.66e8%	267	65
	3	2	2.5	0.0%	267	65	6.7	0.0%	267	65
		153	60.3	1.3%	10	3	60.0	428.3%	419	105
		306	60.4	39.0%	215	61	60.1	7.66e8%	267	65
		459	60.2	53.5%	267	65	60.1	7.66e8%	267	65
		612	60.6	53.5%	267	65	60.3	7.66e8%	267	65
		766	60.3	53.5%	267	65	60.2	7.66e8%	267	65
	4	2	17.3	0.0%	267	65	26.5	0.0%	267	65
		153	60.4	53.5%	267	65	60.1	7.66e8%	267	65
		306	60.5	53.5%	267	65	60.1	7.66e8%	267	65
		459	60.7	53.5%	267	65	60.3	7.66e8%	267	65
		612	60.9	53.5%	267	65	60.3	7.66e8%	267	65
		766	61.2	53.5%	267	65	60.3	7.66e8%	267	65
cmc	2	3	1.8	0.0%	677	167	1.4	0.0%	677	167
		235	60.1	108.1%	662	160	60.0	199.0%	763	195
		471	60.4	130.5%	661	156	60.0	440.4%	765	201
		706	60.4	135.1%	677	167	60.1	1.178e9%	677	167
		942	60.3	135.1%	677	167	60.1	1.178e9%	677	167
		1178	60.5	135.1%	677	167	60.1	187.3%	768	194
	3	3	7.5	0.0%	677	167	9.0	0.0%	677	167
		235	60.3	135.1%	677	167	60.1	1.178e9%	677	167
		471	60.5	135.1%	677	167	60.1	1.178e9%	677	167
		706	61.9	135.1%	677	167	60.3	1.178e9%	677	167
		942	61.1	135.1%	677	167	60.2	1.178e9%	677	167
		1178	61.4	135.1%	677	167	60.2	1.178e9%	677	167
	4	3	61.0	35.1%	677	167	45.8	0.0%	677	167
		235	61.0	135.1%	677	167	60.1	1.178e9%	677	167
		471	62.8	135.1%	677	167	60.1	1.178e9%	677	167
		706	63.6	135.1%	677	167	60.3	1.178e9%	677	167
		942	62.1	135.1%	677	167	60.3	1.178e9%	677	167
		1178	65.3	135.1%	677	167	60.5	1.178e9%	677	167

Table 4 – Résultats avec regroupements et séparation multivariée

Conclusion

En général, les résultats montrent que l'utilisation de clusters pour regrouper des données similaires permet de trouver un arbre de classification beaucoup plus rapidement que lorsqu'il n'est pas utilisé. De plus, il y a un compromis entre la vitesse et la fidélité, car en utilisant un plus grand nombre de clusters, c'est-à-dire des groupes plus petits, la résolution devient plus lente, mais un arbre de classification plus précis est obtenu. Ceci est congruent, puisque l'utilisation des clusters remplace les données par un identifiant pour chaque cluster, réduisant ainsi le nombre de contraintes et la fidélité du programme. D'autre part, on peut constater que la méthode de callback est toujours plus lente. Cela peut être dû à de multiples facteurs, comme la complexité du modèle sans les restrictions de flux, ce qui signifie que lors de la résolution d'un premier modèle relaxé, cela nous oblige à passer par plus de nœuds dans le processus branch & cut que lors de l'ajout de ces restrictions dès le début. Une autre possibilité est le nombre de cœurs utilisés, puisque lors de l'utilisation du callback, le nombre de cœurs est limité à un, alors que la résolution du modèle complet permet l'utilisation de plusieurs cœurs.