

Arbres de classification

Apprentissage non-supervisé ou apprentissage multi-supervisé ?

Ricco RAKOTOMALALA

PLAN

1. Classification automatique, typologie, etc.
2. Interprétation des groupes
3. Classement d'un nouvel individu
4. Les arbres de classification
5. Bilan

La classification automatique

La typologie ou le Clustering ou l'Apprentissage non-supervisé

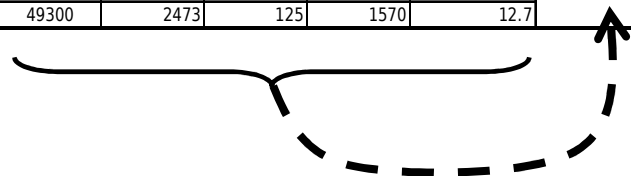
Classification automatique

Typologie, apprentissage non-supervisé

X (tous quantitatifs, pour l'instant)

Pas de Y à prédire

Modele	Prix	Cylindree	Puissance	Poids	Consommation	Groupe
Daihatsu Cuore	11600	846	32	650	5.7	
Suzuki Swift 1.0 GLS	12490	993	39	790	5.8	
Fiat Panda Mambo L	10450	899	29	730	6.1	
VW Polo 1.4 60	17140	1390	44	955	6.5	
Opel Corsa 1.2i Eco	14825	1195	33	895	6.8	
Subaru Vivio 4WD	13730	658	32	740	6.8	
Toyota Corolla	19490	1331	55	1010	7.1	
Opel Astra 1.6i 16V	25000	1597	74	1080	7.4	
Peugeot 306 XS 108	22350	1761	74	1100	9	
Renault Safrane 2.2. V	36600	2165	101	1500	11.7	
Seat Ibiza 2.0 GTI	22500	1983	85	1075	9.5	
VW Golf 2.0 GTI	31580	1984	85	1155	9.5	
Citroen ZX Volcane	28750	1998	89	1140	8.8	
Fiat Tempra 1.6 Liberty	22600	1580	65	1080	9.3	
Fort Escort 1.4i PT	20300	1390	54	1110	8.6	
Honda Civic Joker 1.4	19900	1396	66	1140	7.7	
Volvo 850 2.5	39800	2435	106	1370	10.8	
Ford Fiesta 1.2 Zetec	19740	1242	55	940	6.6	
Hyundai Sonata 3000	38990	2972	107	1400	11.7	
Lancia K3.0 LS	50800	2958	150	1550	11.9	
Mazda Hachtback V	36200	2497	122	1330	10.8	
Mitsubishi Galant	31990	1998	66	1300	7.6	
Opel Omega 2.5i V6	47700	2496	125	1670	11.3	
Peugeot 806 2.0	36950	1998	89	1560	10.8	
Nissan Primera 2.0	26950	1997	92	1240	9.2	
Seat Alhambra 2.0	36400	1984	85	1635	11.6	
Toyota Previa salon	50900	2438	97	1800	12.8	
Volvo 960 Kombi aut	49300	2473	125	1570	12.7	



Identifier les catégories (groupes) de voitures « similaires » (c.-à-d. qui se ressemblent)

Objectif : identifier des groupes d'observations ayant des caractéristiques similaires (ex. comportement d'achats de clients, caractère « polluant » de véhicules, etc.)

On veut que :

- (1) Les individus dans un même groupe se ressemblent le plus possible
- (2) Les individus dans des groupes différents se démarquent le plus possible

Pourquoi ?

- Identifier des structures sous-jacentes dans les données
- Résumer des comportements
- Affecter de nouveaux individus à des catégories
- Identifier les cas totalement atypiques

Classification automatique

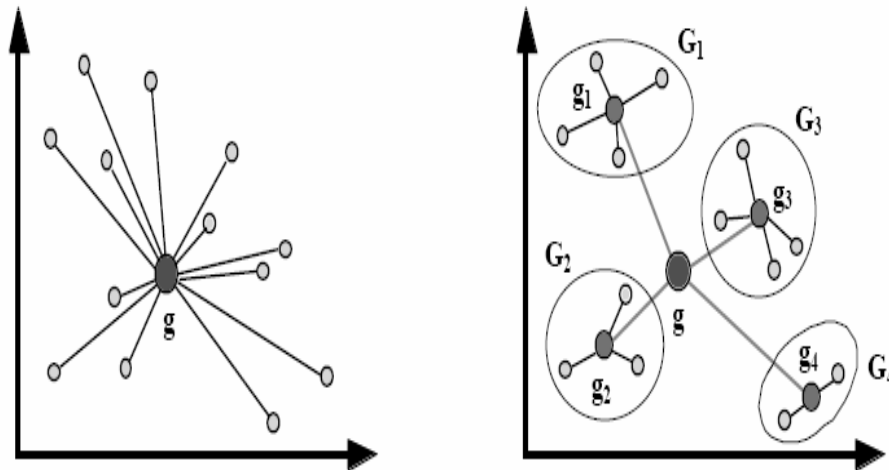
Objectifs

Principe : Constituer des groupes « naturels » de manière à ce que les individus dans un même groupe se ressemblent, et les individus dans des groupes différents soient dissemblables.

Autres visions :

- Identifier des groupes d'individus ayant un comportement (ou des caractéristiques) homogènes
- Proposer un résumé des données en explicitant ses principales dimensions (oppositions)
- Mettre en évidence les principales structures dans les données (définir des « concepts »)
- Construction d'une taxonomie (classification hiérarchique) d'objets (cf. taxonomie des espèces)

Illustration dans le plan



Points clés dans la constitution des groupes.

Quantifier :

- La proximité entre 2 individus
- La proximité entre 2 groupes
- La proximité entre 1 individu et un groupe (lors de la construction et l'affectation)

- Le degré de compacité d'un groupe
- L'éloignement global entre les groupes (séparabilité)

Classification ascendante hiérarchique

Une technique très populaire

Principe :

- Calculer la dissimilarité entre les individus
- Agglomérations successives en fusionnant en priorité les groupes les plus proches (cf. stratégies d'agrégation : saut minimum, méthode de WARD, etc.)
- Hauteur = Distance entre groupes

Avantages

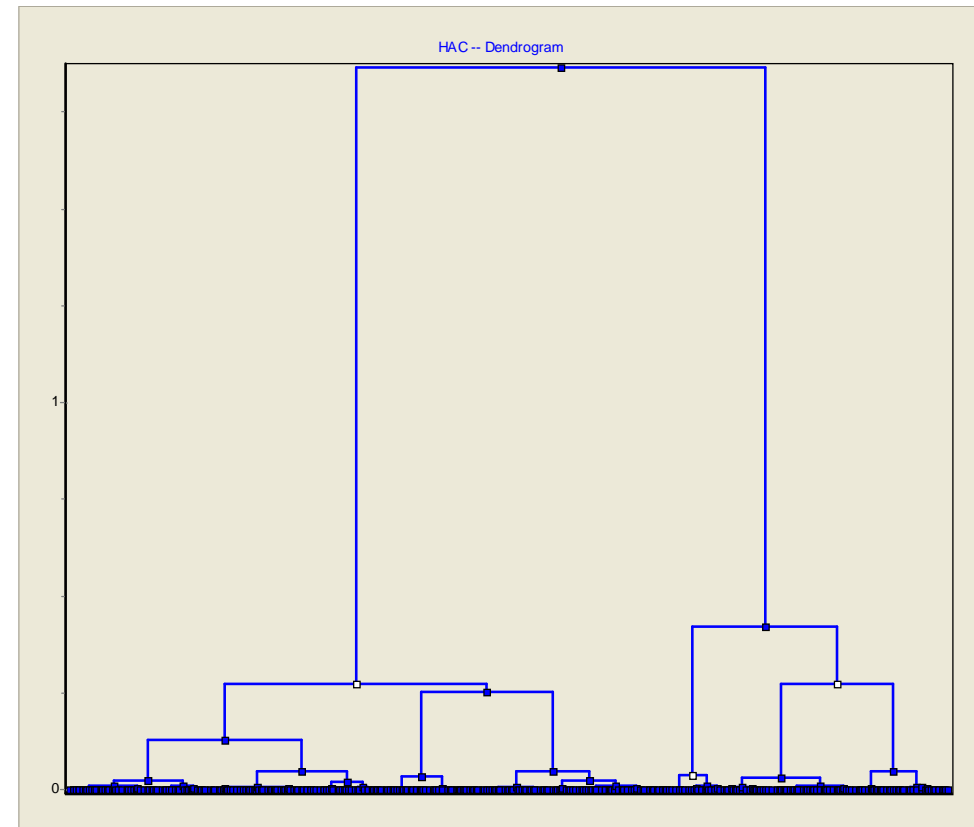
- Hiérarchie de partition (taxonomie)
- Indications sur la proximité entre groupes (choix du nombre de groupes → très difficile, il n'y a pas de solution « optimale »)
- Propose des solutions alternatives (que l'on peut interpréter ou approfondir)

Inconvénients

- Mise en œuvre sur des grandes bases (cf. stratégies mixtes)

Problèmes récurrents de la classification

- Détection du « bon » nombre de groupes
- Interprétation des groupes
- (Avec) L'utilisation des variables illustratives
- Classement d'un nouvel individu



→ 201 obs.

→ 4 var. actives de « coûts » (prix, conso. ville et autoroute, prime assurance)

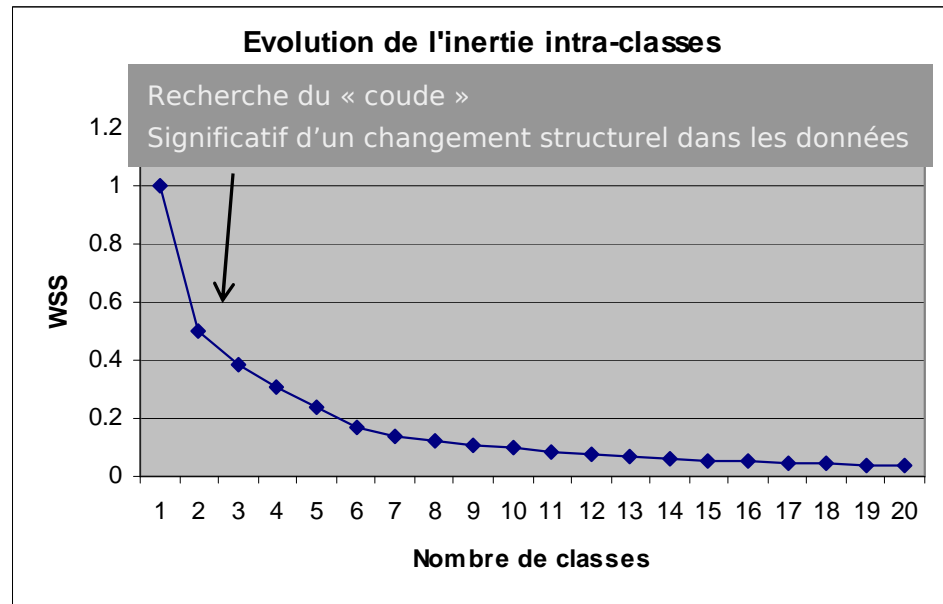
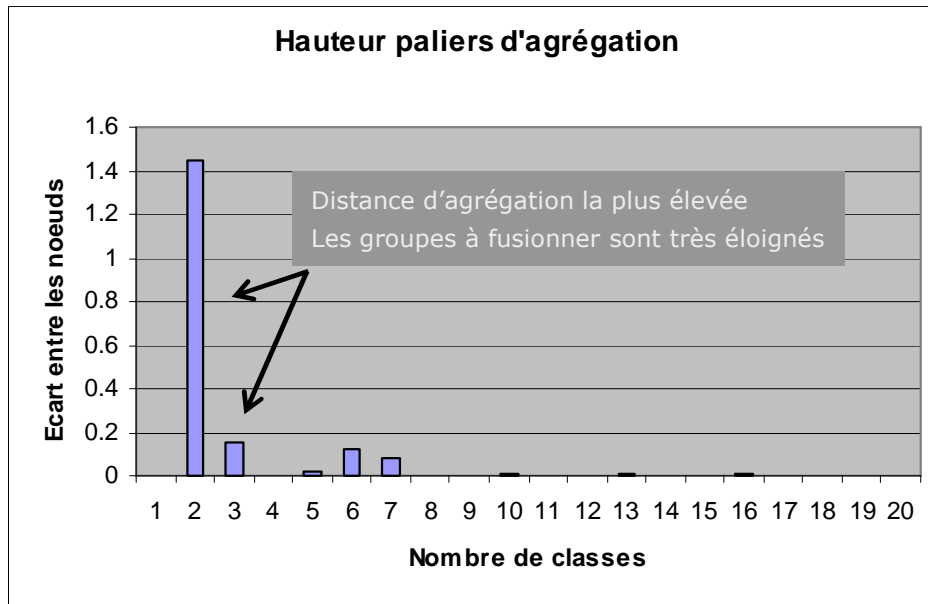
→ 2 premiers axes factoriels → 92% inertie

→ Méthode de Ward (les axes ne doivent pas être réduites !!!)

→ Solution à 3 classes demandée (solution à 2 classes souvent triviale)

Détection du « bon » nombre de groupe

Quelques repères



Le choix du « bon » nombre de classes reste ouvert.
Il est indissociable de l'interprétation des classes et du cahier
des charges de l'étude.



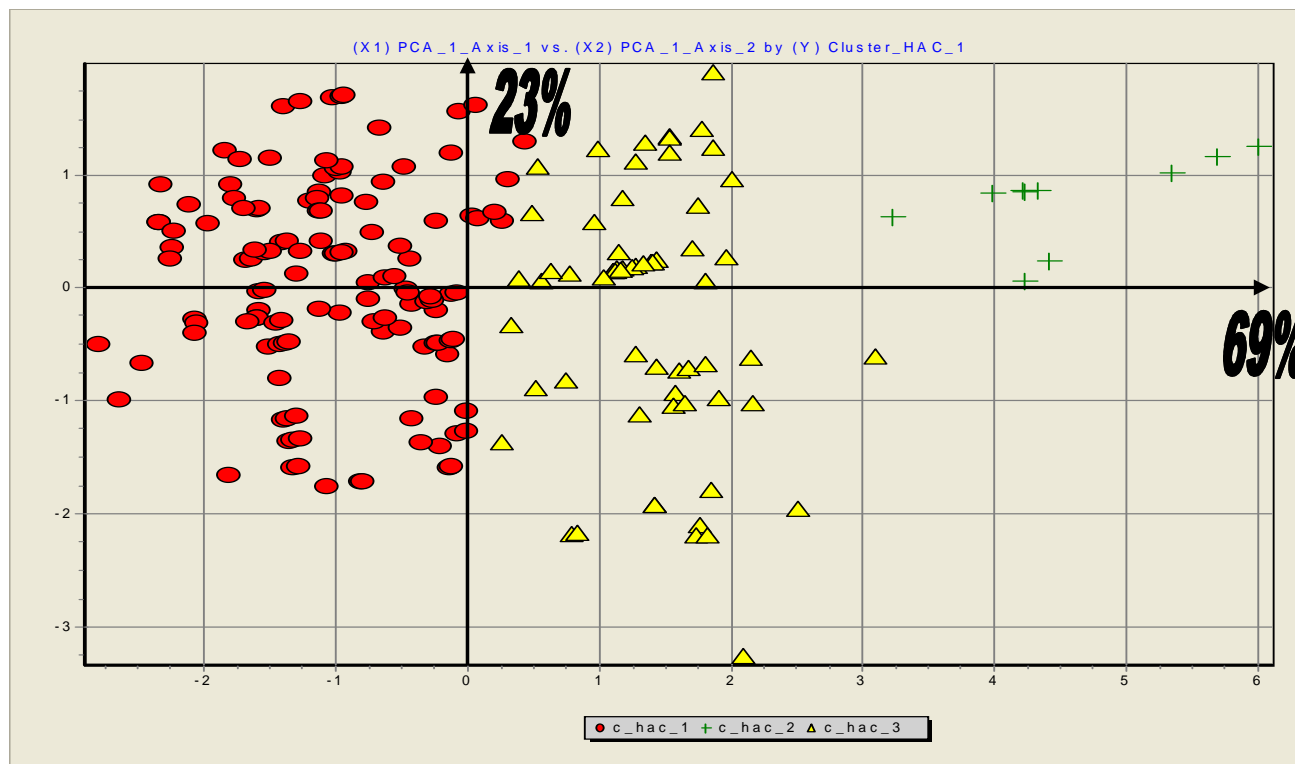
Interprétation et description des groupes

Comprendre l'appartenance d'un individu à un groupe
Utilisation des variables actives et des variables illustratives

Interprétation des groupes sur les variables actives

Analyse factorielle

Principe : Comprendre pourquoi les observations dans un même groupe sont ensemble. Comprendre ce qui différencie les groupes.



Avantage

Vision multivariée des données

Inconvénients

A la difficulté d'interprétation des groupes vient s'ajouter la difficulté d'interprétation des axes factoriels.

D'autant plus compliqué lorsque + des 2 premiers axes sont significatifs.

Attribute	Axis_1		Axis_2		Axis_3		Axis_4	
	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
-								
price	0.8927	80 % (80 %)	0.1479	2 % (82 %)	-0.4254	18 % (100 %)	-0.0177	0 % (100 %)
normalized-losses	0.3183	10 % (10 %)	-0.9474	90 % (100 %)	-0.0319	0 % (100 %)	0.0092	0 % (100 %)
conso-ville	0.9613	92 % (92 %)	0.0498	0 % (93 %)	0.2383	6 % (98 %)	-0.1287	2 % (100 %)
conso-autoroute	0.9678	94 % (94 %)	0.1257	2 % (95 %)	0.1661	3 % (98 %)	0.1412	2 % (100 %)
Var. Expl.	2.759	69 % (69 %)	0.9377	23 % (92 %)	0.2663	7 % (99 %)	0.0369	1 % (100 %)

Remarque

Aide au choix du nombre de groupes. Dans notre exemple, on peut se demander si « 4 groupes » n'est pas plus approprié.

Interprétation des groupes sur les variables actives

Statistiques descriptives comparatives – Principe de la « valeur test »

Results											
Description of "Cluster_HAC_1"											
Cluster_HAC_1=c_hac_1				Cluster_HAC_1=c_hac_2				Cluster_HAC_1=c_hac_3			
Examples [63.7 %] 128				Examples [5.0 %] 10				Examples [31.3 %] 63			
Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
normalized-losses	-4.7	113.61 (26.82)	121.33 (30.55)	price	10.3	36835.50 (4483.50)	12904.96 (7492.72)	conso-ville	8.4	12.16 (1.29)	9.90 (2.56)
price	-10.0	8899.70 (2714.81)	12904.96 (7492.72)	conso-autoroute	8.2	12.74 (1.43)	8.02 (1.86)	conso-autoroute	7.5	9.48 (0.80)	8.02 (1.86)
conso-autoroute	-11.0	6.93 (0.99)	8.02 (1.86)	conso-ville	7.4	15.76 (1.16)	9.90 (2.56)	price	5.5	17244.13 (4297.33)	12904.96 (7492.72)
conso-ville	-11.5	8.33 (1.30)	9.90 (2.56)	normalized-losses	0.5	126.30 (9.09)	121.33 (30.55)	normalized-losses	4.7	136.24 (34.26)	121.33 (30.55)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			

Principe : Comparer les paramètres [moyenne, proportion] calculés pour la totalité de l'échantillon et pour le groupe concerné.

Avantage : Simplicité. La valeur test donne une idée de l'importance de l'écart

Inconvénient : Analyse univariée. Ne tient pas compte explicitement des liaisons entre les variables [on doit le faire nous même].

Valeur test : Statistique du test de comparaison à un standard

Remarques

- 1 : Le standard est le paramètre calculé sur la totalité de l'échantillon
- 2 : Ce n'est pas un vrai test (échantillons non indépendants)
- 3 : La statistique a tendance à être exagérée sur les variables actives, ce n'est pas le cas pour les variables illustratives
- 4 : Seuil critique +/- 2 (très lointaine référence à la loi normale)
- 5 : Mieux vaut se concentrer sur les valeurs extrêmes, les oppositions et les décrochements

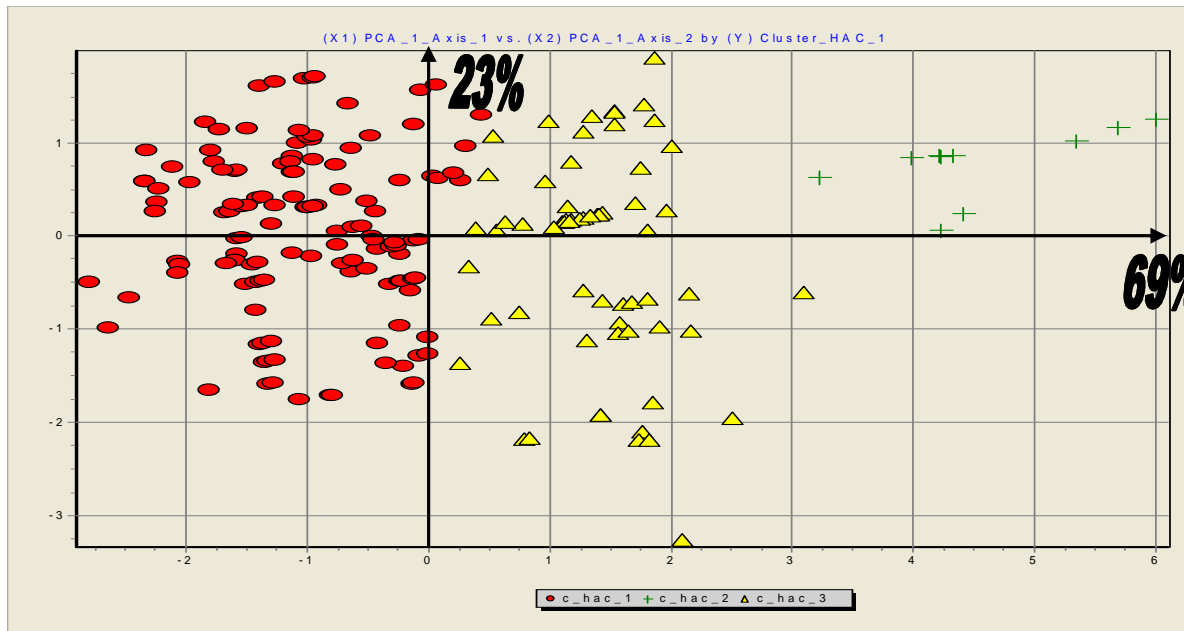
Var.Quantitative (Moyenne) [Classe k]

$$vt = \frac{\bar{x}_k - \bar{x}}{\sqrt{\frac{n - n_k}{n - 1} \times \frac{\sigma_x^2}{n_k}}}$$

Var.Qualitative (Proportion) [Classe k, Caractère j]

$$vt = \frac{S - E(S)}{\sigma_s} = \frac{n_{kj} - \frac{n_k \times n_j}{n}}{\sqrt{n_k \frac{n - n_k}{n - 1} \frac{n_j}{n} \left(1 - \frac{n_j}{n}\right)}}$$

Description factorielle ou Valeur test ?



Description condensée mais (parfois) réductrice

Ex. la différence entre C2 et C3 sur l'axe 1 tient essentiellement au prix en réalité.

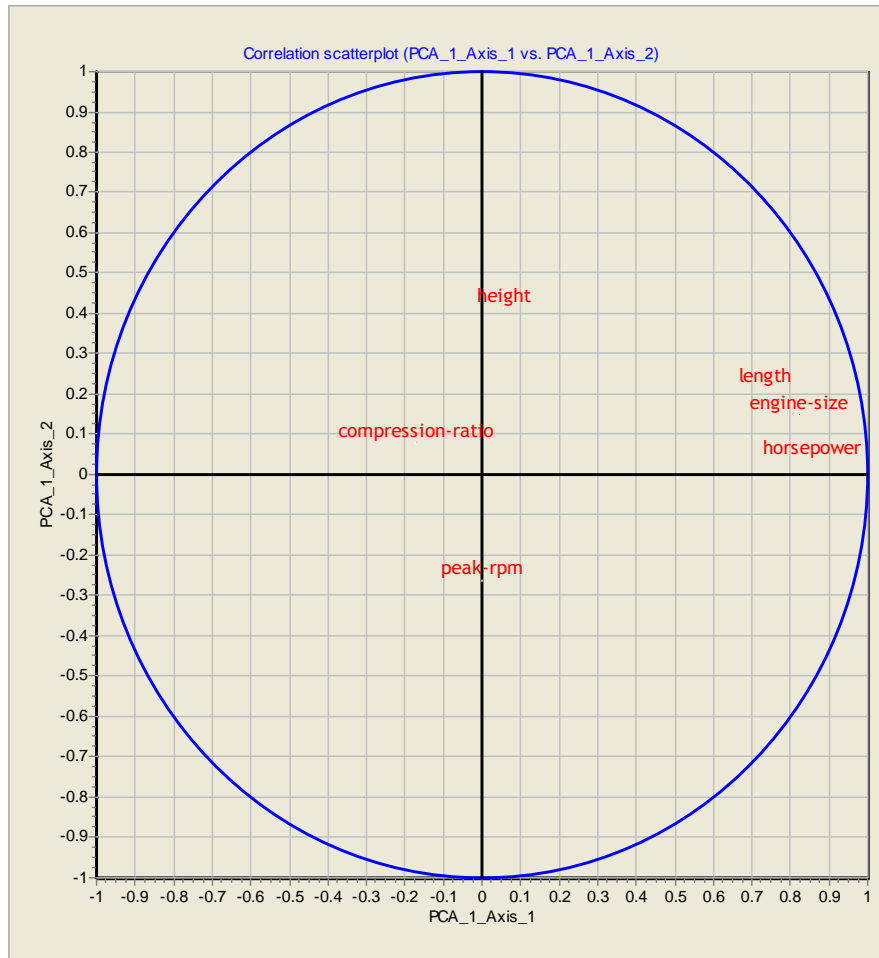
Results											
Description of "Cluster_HAC_1"											
Cluster_HAC_1=c_hac_1				Cluster_HAC_1=c_hac_2				Cluster_HAC_1=c_hac_3			
Examples		[63.7 %] 128		Examples		[5.0 %] 10		Examples		[31.3 %] 63	
Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
normalized-losses	-4.7	113.61 (26.82)	121.33 (30.55)	price	10.3	36835.50 (4483.50)	12904.96 (7492.72)	conso-ville	8.4	12.16 (1.29)	9.90 (2.56)
price	-10.0	8899.70 (2714.81)	12904.96 (7492.72)	conso-autoroute	8.2	12.74 (1.43)	8.02 (1.86)	conso-autoroute	7.5	9.48 (0.80)	8.02 (1.86)
conso-autoroute	-11.0	6.93 (0.99)	8.02 (1.86)	conso-ville	7.4	15.76 (1.16)	9.90 (2.56)	price	5.5	17244.13 (4297.33)	12904.96 (7492.72)
conso-ville	-11.5	8.33 (1.30)	9.90 (2.56)	normalized-losses	0.5	126.30 (9.09)	121.33 (30.55)	normalized-losses	4.7	136.24 (34.26)	121.33 (30.55)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			

Description détaillée mais vite (trop) touffue, à mesure que le nombre de variables et de groupes augmente...

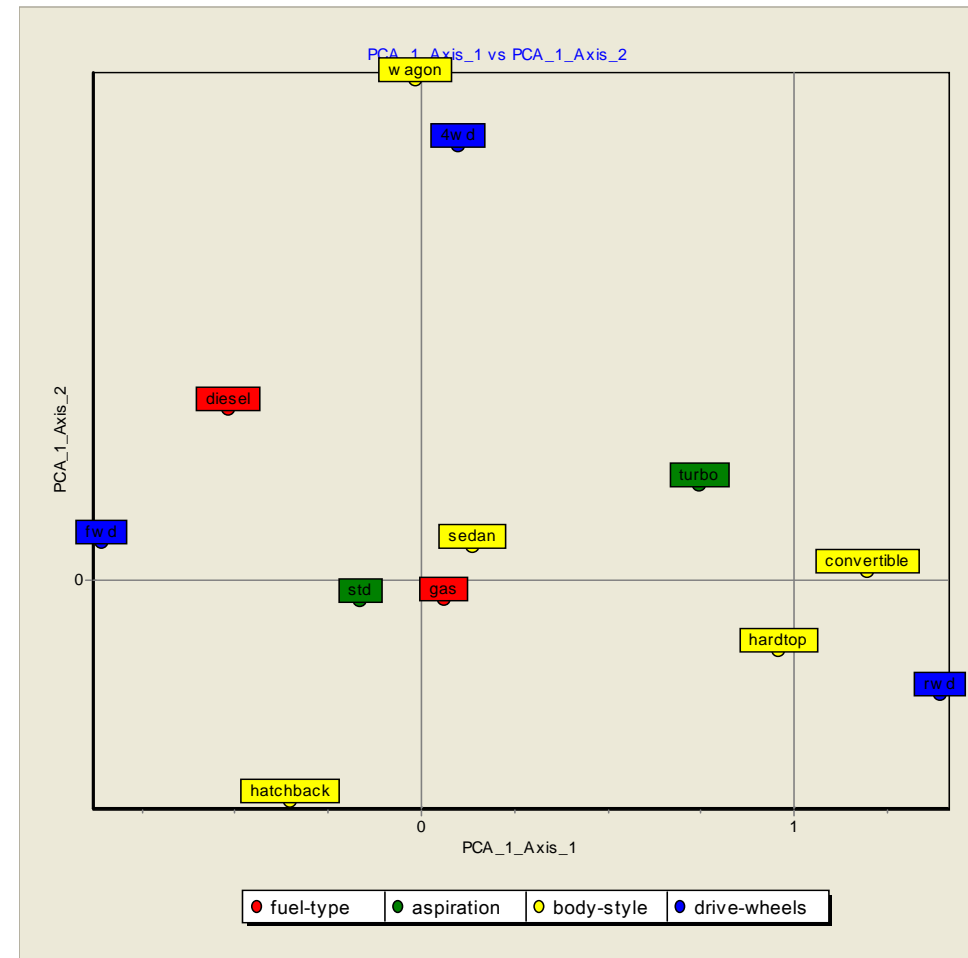
Variables illustratives

Comprendre, illustrer et désigner les groupes – Projection factorielle

Variables quantitatives



Variables qualitatives



Toujours les mêmes difficultés inhérentes à la lecture d'une analyse factorielle

Variables illustratives

Comprendre, illustrer et désigner les groupes – S'appuyer sur les valeurs tests

Cluster_HAC_1=c_hac_1				Cluster_HAC_1=c_hac_2				Cluster_HAC_1=c_hac_3						
[63.7 %] 128				[5.0 %] 10				[31.3 %] 63						
Examples	Att - Desc	Test value	Group	Overall	Examples	Att - Desc	Test value	Group	Overall	Examples	Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)						
compression-ratio		0.9	10.34 (4.05)	10.16 (4.01)	engine-size		10.1	254.90 (44.09)	126.00 (41.22)	horsepower		7.3	131.84 (32.25)	102.79 (37.84)
peak-rpm		-0.2	5106.65 (480.37)	5111.94 (471.36)	horsepower		6.9	183.80 (29.64)	102.79 (37.84)	length		6.2	182.25 (8.99)	174.17 (12.43)
height		-0.8	53.66 (2.20)	53.77 (2.45)	length		5.7	196.09 (7.79)	174.17 (12.43)	engine-size		4	143.11 (29.69)	126.00 (41.22)
engine-size		-8.4	107.50 (17.57)	126.00 (41.22)	height		0.1	53.85 (2.89)	53.77 (2.45)	peak-rpm		0.8	5153.17 (466.36)	(471.36)
length		-8.6	168.48 (9.83)	174.17 (12.43)	peak-rpm		-1.3	4920.00 (359.94)	5111.94 (471.36)	height		0.8	53.97 (2.86)	53.77 (2.45)
horsepower		-10.2	82.17 (17.90)	102.79 (37.84)	compression-ratio		-1.4	8.43 (1.09)	10.16 (4.01)	compression-ratio		-0.2	10.06 (4.18)	10.16 (4.01)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy						
drive-wheels=fwd		7.5	[84.9 %] 78.9 %	59.20%	drive-wheels=rwd		4.3	[13.7 %] 100.0 %	36.30%	drive-wheels=rwd		6.3	[58.9 %] 68.3 %	36.30%
aspiration=std		3.6	[69.5 %] 89.1 %	81.60%	body-style=sedan		2.1	[8.3 %] 80.0 %	47.80%	aspiration=turbo		4.5	[62.2 %] 36.5 %	18.40%
num-of-doors=_missing		1.1	[100.0 %] 1.6 %	1.00%	style=convertible		1.6	[20.0 %] 10.0 %	2.50%	style=convertible		1.4	[60.0 %] 4.8 %	2.50%
body-style=hatchback		0.9	[68.1 %] 36.7 %	34.30%	aspiration=std		1.5	[6.1 %] 100.0 %	81.60%	num-of-doors=two		0.7	[34.1 %] 46.0 %	42.30%
drive-wheels=4wd		0.9	[77.8 %] 5.5 %	4.50%	body-style=hardtop		1.3	[16.7 %] 10.0 %	3.00%	body-style=wagon		0.5	[36.0 %] 14.3 %	12.40%
num-of-doors=four		0.4	[64.9 %] 57.8 %	56.70%	fuel-type=gas		1.1	[5.5 %] 100.0 %	90.00%	fuel-type=diesel		0.4	[35.0 %] 11.1 %	10.00%
body-style=hardtop		0.2	[66.7 %] 3.1 %	3.00%	num-of-doors=four		0.2	[5.3 %] 60.0 %	56.70%	style=hatchback		0.1	[31.9 %] 34.9 %	34.30%
fuel-type=diesel		0.1	[65.0 %] 10.2 %	10.00%	num-of-doors=two		-0.1	[4.7 %] 40.0 %	42.30%	fuel-type=gas		-0.4	[30.9 %] 88.9 %	90.00%
body-style=wagon		0	[64.0 %] 12.5 %	12.40%	doors=_missing		-0.3	[0.0 %] 0.0 %	1.00%	num-of-doors=four		-0.5	[29.8 %] 54.0 %	56.70%
fuel-type=gas		-0.1	[63.5 %] 89.8 %	90.00%	drive-wheels=4wd		-0.7	[0.0 %] 0.0 %	4.50%	drive-wheels=4wd		-0.6	[22.2 %] 3.2 %	4.50%
body-style=sedan		-0.3	[62.5 %] 46.9 %	47.80%	fuel-type=diesel		-1.1	[0.0 %] 0.0 %	10.00%	body-style=sedan		-0.6	[29.2 %] 44.4 %	47.80%
num-of-doors=two		-0.6	[61.2 %] 40.6 %	42.30%	body-style=wagon		-1.2	[0.0 %] 0.0 %	12.40%	body-style=hardtop		-0.8	[16.7 %] 1.6 %	3.00%
body-style=convertible		-2.1	[20.0 %] 0.8 %	2.50%	aspiration=turbo		-1.5	[0.0 %] 0.0 %	18.40%	doors=_missing		-1	[0.0 %] 0.0 %	1.00%
aspiration=turbo		-3.6	[37.8 %] 10.9 %	18.40%	style=hatchback		-2.3	[0.0 %] 0.0 %	34.30%	aspiration=std		-4.5	[24.4 %] 63.5 %	81.60%
drive-wheels=rwd		-8.1	[27.4 %] 15.6 %	36.30%	drive-wheels=fwd		-3.9	[0.0 %] 0.0 %	59.20%	drive-wheels=fwd		-6	[15.1 %] 28.6 %	59.20%

Visibilité individuelle des variables

Possibilité de filtrer de manière à ne faire apparaître que les |v.t.| élevés

Mêmes difficultés de lecture dès que les variables et les groupes augmentent

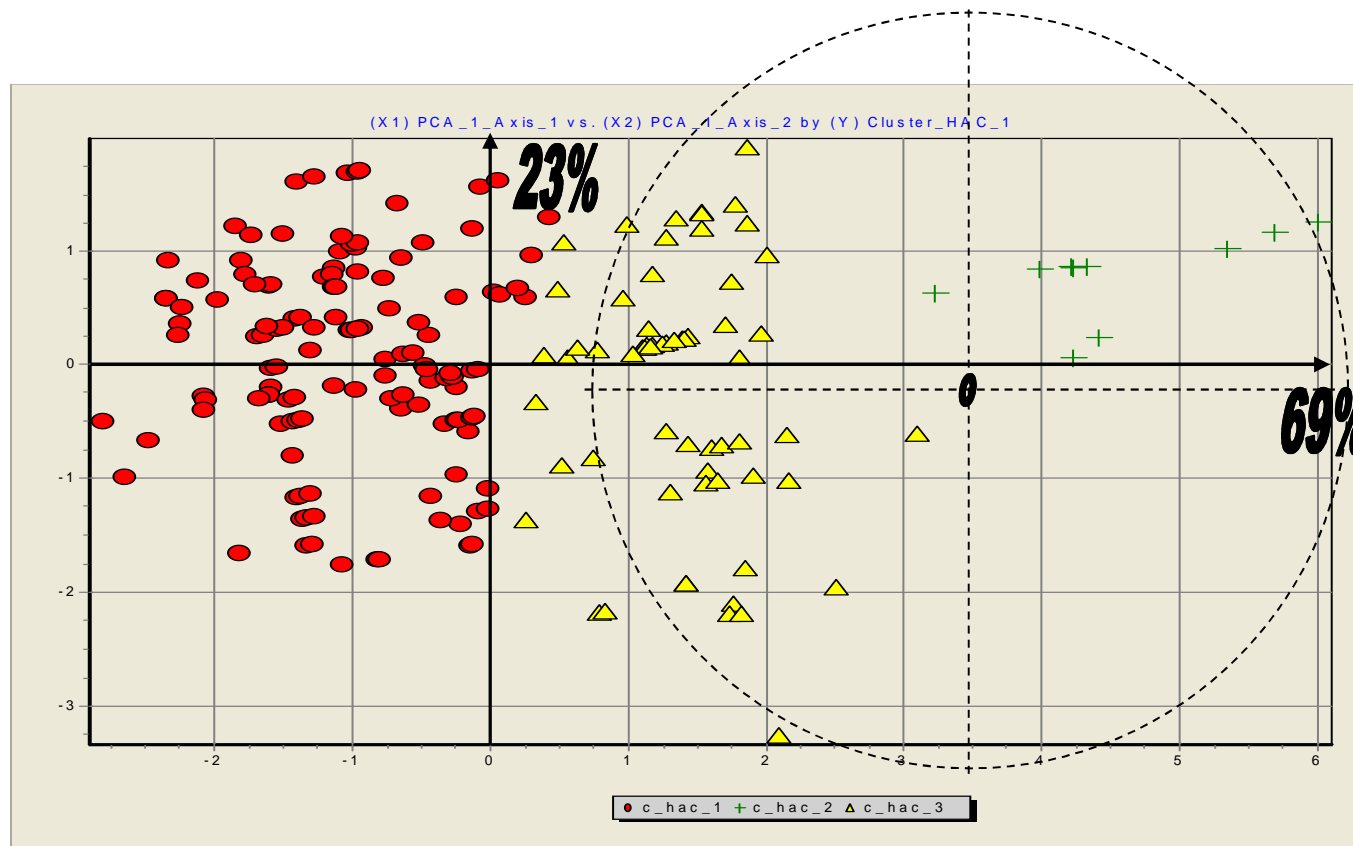
Classement d'un nouvel individu

Affectation d'un nouvel individu à un groupe sur la base
des variables actives ou illustratives

Classement d'un individu

Sur la base des variables actives

Principe : Se baser sur la proximité pour attribuer un individu à un groupe
Attention : « Proximité » doit être en rapport avec la stratégie d'agrégation utilisée (ex. saut minimum, saut maximum, Ward, etc.)



Quel groupe pour o ?
Saut minimum : Δ
Saut maximum : +

Difficile à industrialiser : il faudrait disposer de tous les individus, tout le temps ?

Difficile à interpréter : quels sont les caractères qui associent (le plus) cet individu au groupe ?

Classement d'un individu

Sur la base des variables illustratives

Attribute	Category	Informations
fuel-type	Discrete	2 values
aspiration	Discrete	2 values
num-of-doors	Discrete	3 values
body-style	Discrete	5 values
drive-wheels	Discrete	3 values
wheel-base	Continue	-
length	Continue	-
width	Continue	-
height	Continue	-
curb-weight	Continue	-
num-of-cylinders	Discrete	7 values
engine-size	Continue	-
compression-ratio	Continue	-
horsepower	Continue	-
peak-rpm	Continue	-
price	Continue	-
normalized-losses	Continue	-
conso-ville	Continue	-
conso-autoroute	Continue	-

Pour désigner l'appartenance à un groupe

Ex. Caractéristiques intrinsèques des véhicules

Ex. Descriptifs signalétiques de la clientèle (âge, sexe, etc.)

Pour élaborer les groupes homogènes

Ex. Constituer des groupes en termes de coûts d'utilisation

Ex. Comportement d'achats / à différents produits

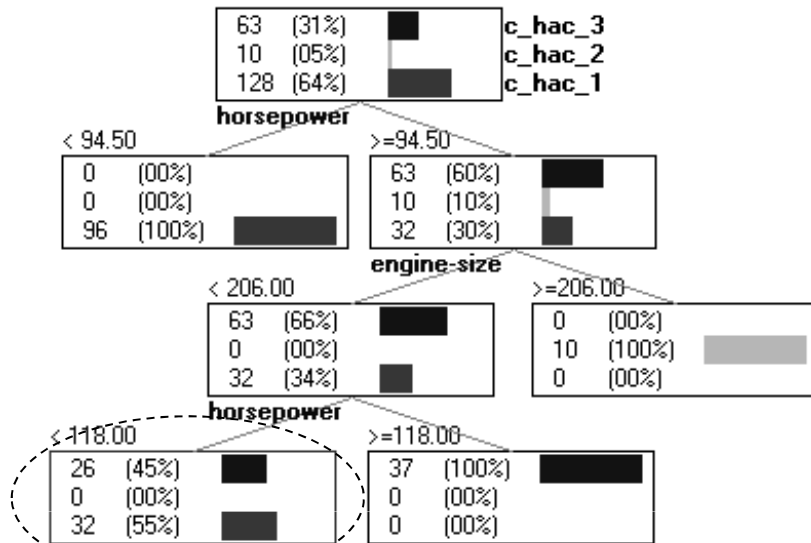
Problème : le principe de la proximité devient très difficile à gérer

Objectif : Produire une règle d'affectation facilement industrialisable (intégrable facilement dans les systèmes d'information)

Remarque : On se retrouve dans un cadre proche du supervisé, sauf que l'on essaie de prédire des groupes qui résument des caractéristiques multivariées (par le clustering)

Classement des individus

Basé sur des règles – Utilisation des arbres de décision



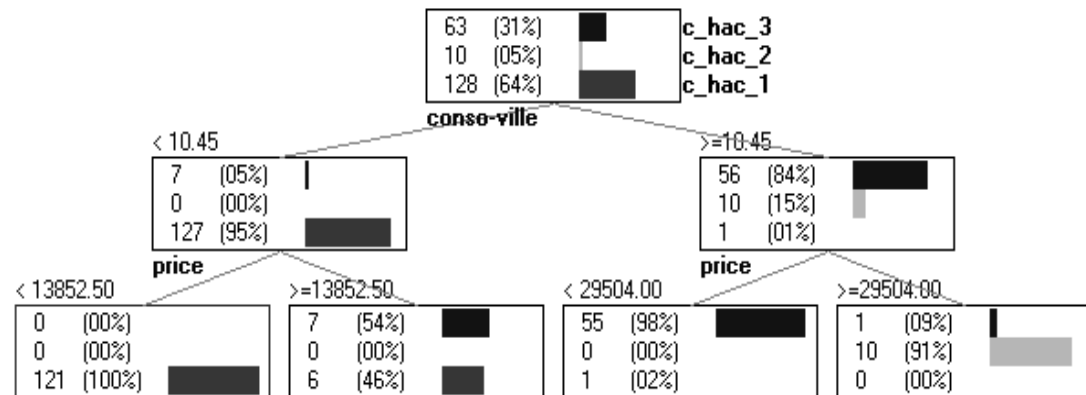
--> Souci : Sont rassemblés des véhicules situés dans des groupes différents

→ La simple lecture de la fiche technique permet de déterminer le « coût global » d'usage d'un véhicule

Avantage : Très simple à mettre en œuvre

Inconvénient : Des groupes peuvent ne pas être « purs » (mélange de classes). Normal les processus d'apprentissage ne sont pas les mêmes, et ne sont pas (forcément) liés

Remarque : Cette démarche peut être étendue à une caractérisation à l'aide des variables actives (avec les mêmes inconvénients...)



Arbres de classification

S'appuyer sur le principe de la segmentation pour constituer des groupes homogènes

Arbres de classification

Principe

Objectif : S'appuyer sur le processus de segmentation dans un problème multivarié c.-à-d. lier le processus de constitution de groupes homogènes avec la construction de la règle d'affectation

$$(Y_1, \dots, Y_I) = f(X_1, \dots, X_J; \alpha)$$

↑
Sert à caractériser
l'homogénéité des groupes

↑
Sert à construire
les groupes

Remarque : $Y = X$ est un cas particulier tout à fait licite

Attribute	Category	Informations
fuel-type	Discrete	2 values
aspiration	Discrete	2 values
num-of-doors	Discrete	3 values
body-style	Discrete	5 values
drive-wheels	Discrete	3 values
wheel-base	Continue	-
length	Continue	-
width	Continue	-
height	Continue	-
curb-weight	Continue	-
num-of-cylinders	Discrete	7 values
engine-size	Continue	-
compression-ratio	Continue	-
horsepower	Continue	-
peak-rpm	Continue	-
price	Continue	-
normalized-losses	Continue	-
conso-ville	Continue	-
conso-autoroute	Continue	-

X

Y

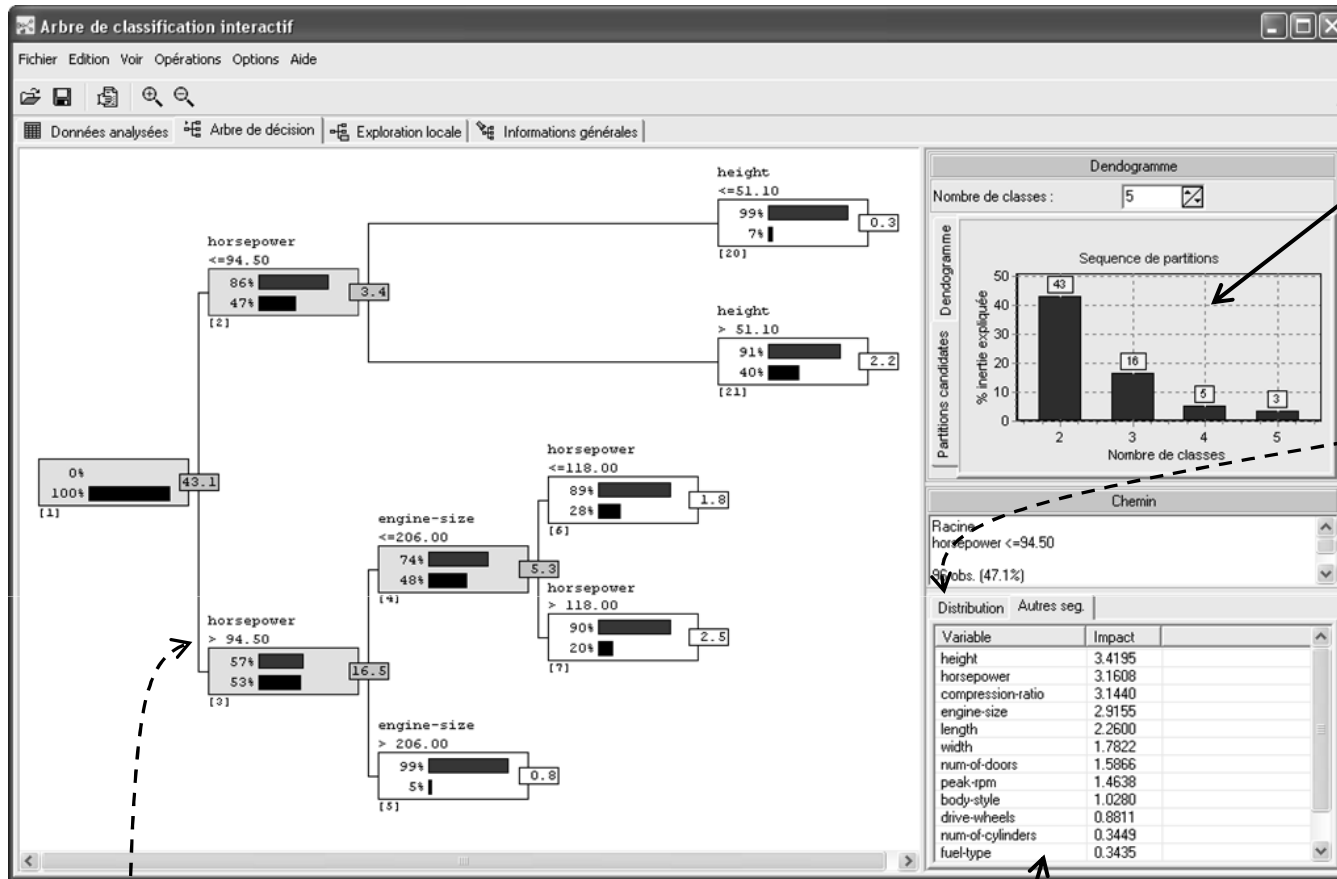
Mêmes éléments à traiter que pour un arbre de décision

- Critère pour le choix de la variable de segmentation
- Borne de découpage d'un descripteur continu
- Regroupement des modalités d'une variable catégorielle
- Détection de la « bonne » taille de l'arbre (et donc du bon nombre de groupes)

→ L'affectation d'un individu à un groupe est immédiat

Arbres de classification

Exemple des voitures



Hauteur des paliers pour identifier le nombre « pertinent » de classes
 [Choix de la bonne taille de l'arbre]
Problème n°3

---Description des classes

Distribution | Autres seg.

Var continues | Var catégorielles

Attribut	Moy. segment	Moy. racine	Valeur test
normalized-losses	110.52	121.34	-4.81
wheel-base	95.87	98.78	-6.50
engine-size	101.89	127.00	-8.10
width	64.55	65.92	-8.57
price	8114.34	13222.95	-8.72
length	166.06	174.09	-8.75
curb-weight	2164.60	2557.21	-10.14
conso-autoroute	6.61	8.04	-10.36
horsepower	73.61	104.32	-10.44
conso-ville	7.88	9.96	-10.81

Distribution | Autres seg.

Var continues | Var catégorielles

Attribut (Modalité)	% segment	% racine	Valeur test
drive-wheels (fwd)	89.58	58.33	8.51
num-of-cylinders (four)	98.96	77.45	6.91
aspiration (std)	96.88	81.86	5.23
num-of-cylinders (eight)	0.00	2.45	-2.13
num-of-cylinders (five)	0.00	5.39	-3.21
num-of-cylinders (six)	0.00	11.76	-4.91
aspiration (turbo)	3.13	18.14	-5.23
drive-wheels (rwd)	5.21	37.25	-8.90

Calcul de la borne de discrétisation
 [Et critère de regroupement pour les variables catégorielles]
Problème n°2

Choix des variables de segmentation
Problème n°1

Arbres de Classification

Critère pour le choix de la variable de segmentation

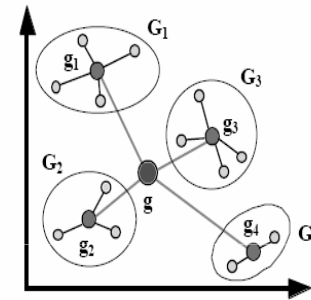
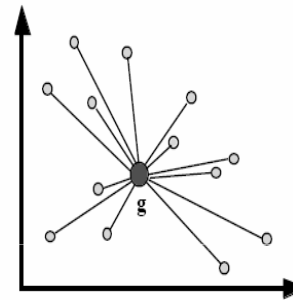
Mesurer l'homogénéité des groupes : Généraliser la notion de variance à la notion d'inertie

Critère : Généralisation multivariée de la décomposition de la variance, le théorème d'Huygens.

- Choisir alors la variable qui maximise le gain d'inertie c.-à-d. l'inertie inter-classes $B = T - W$
- On veut produire des sous-groupes homogènes
- On veut que les barycentres soient éloignés les uns des autres

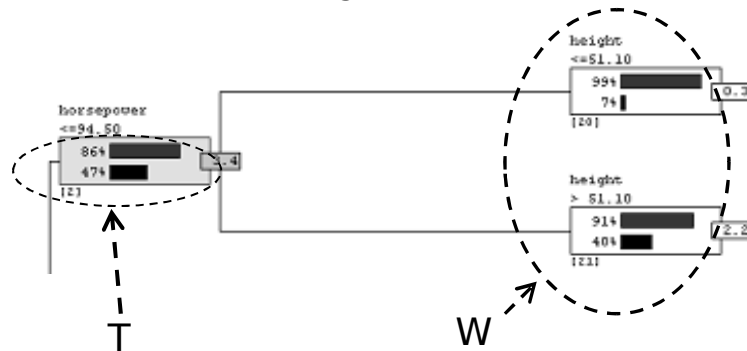
$$T = B + W$$

$$\sum_{i=1}^n p_i d^2(i, g) = \sum_{k=1}^K p_k d^2(g_k, g) + \sum_{k=1}^K \sum_{i \in G_k} p_i d^2(i, g_k)$$



Concrètement

Pour un sommet à segmenter



Variable	Impact
height	3.4195
horsepower	3.1608
compression-ratio	3.1440
engine-size	2.9155
length	2.2600
width	1.7822
num-of-doors	1.5866
peak-rpm	1.4638
body-style	1.0280
drive-wheels	0.8811
num-of-cylinders	0.3449
fuel-type	0.3435
aspiration	0.1460
engine-location	0.0000

Trier les variables selon le gain d'inertie et choisir celle qui le maximise (Faire le parallèle avec CART)

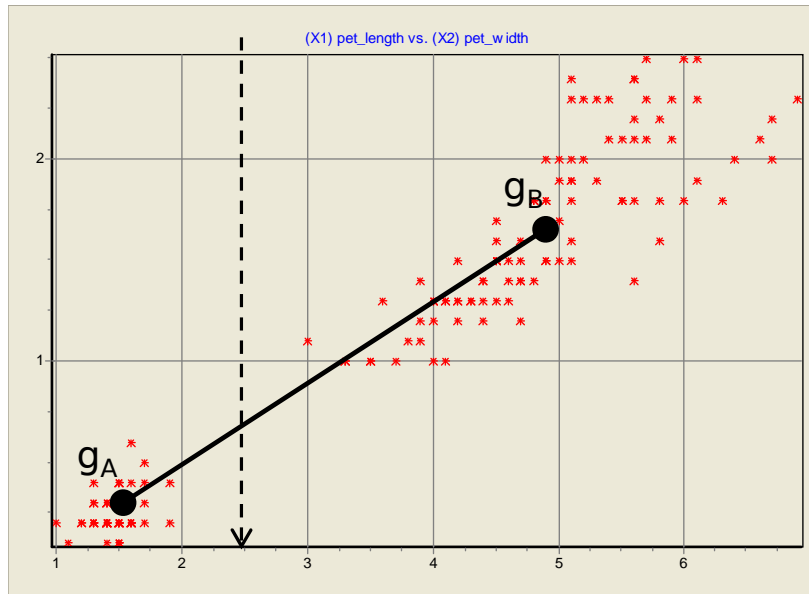
[Le praticien peut intervenir lors de la construction de l'arbre !!!]

Arbres de classification

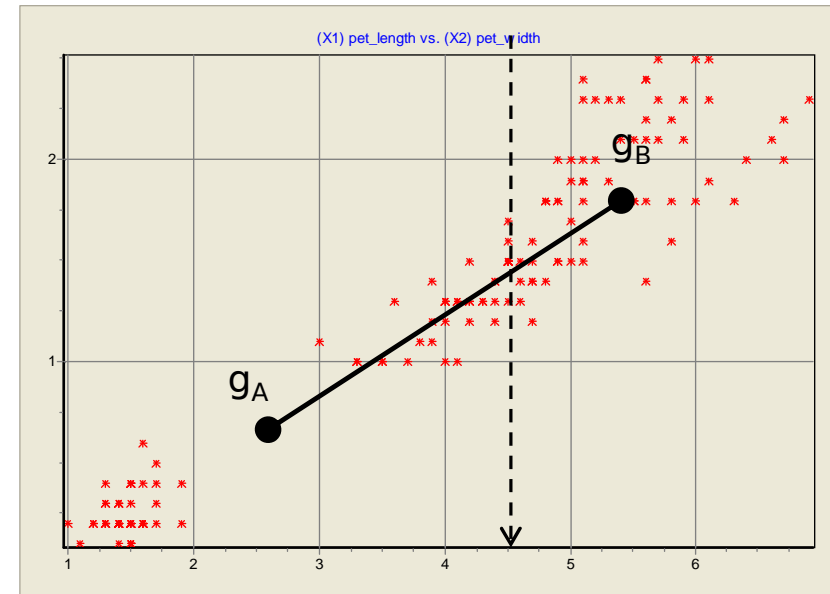
Traitement des variables continues

Principe : Idem les arbres de décision. Pour découper une variable X, la technique consiste à tenter toutes les segmentations possibles (entre 2 points successifs) et choisir la borne qui maximise le gain d'inertie

Exemple de 2 situations (parmi d'autres) à comparer



$$\Delta = B/T = 82\% \quad !$$



$$\Delta = B/T = 63\%$$

Formule simplifiée du gain d'inertie
 Critère de WARD

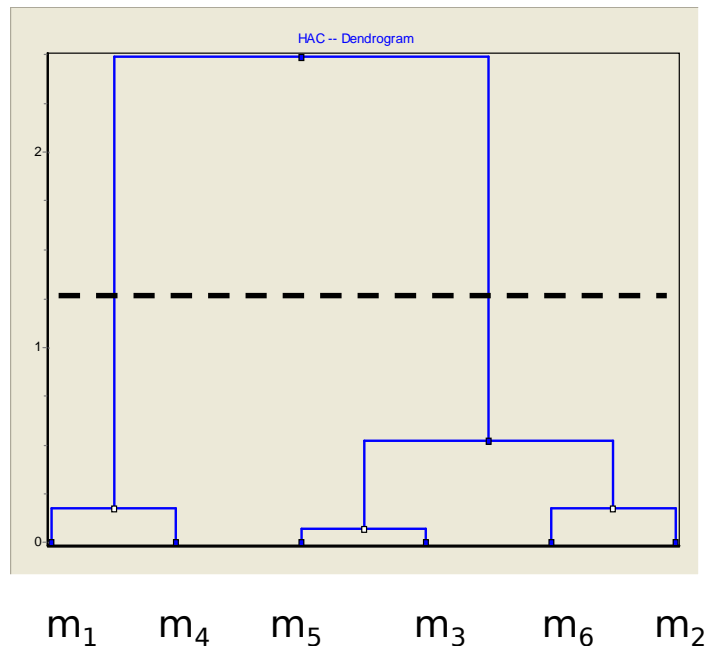
$$B = \frac{n_A \times n_B}{n_A + n_B} d^2 (g_A, g_B)$$

Tri : $O(n \log n)$... on peut pré-trier la base
 Recherche : $O(n)$... très rapide

Arbres de classification

Regroupement des modalités (var. catégorielles)

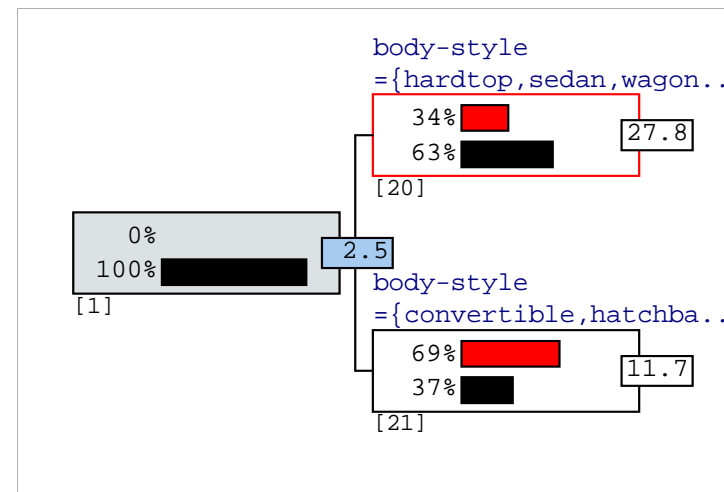
Principe : Par convention l'arbre est binaire pour éviter la fragmentation des données (cf. CART). Comment procéder au regroupement des modalités des variables catégorielles à L (L > 2) modalités ?



Démarche : Regrouper en priorité les modalités correspondant à des groupes d'individus proches (Minimiser la perte d'inertie consécutive à un regroupement – Critère de WARD).

Réitérer en fusionnant au fur et à mesure les groupes : bref, CAH sur les modalités

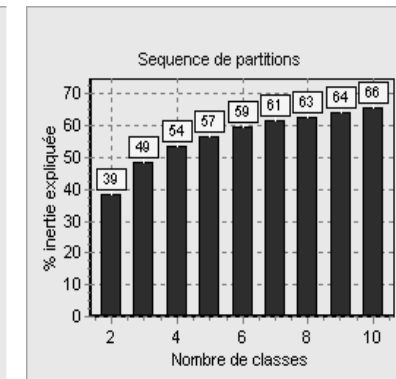
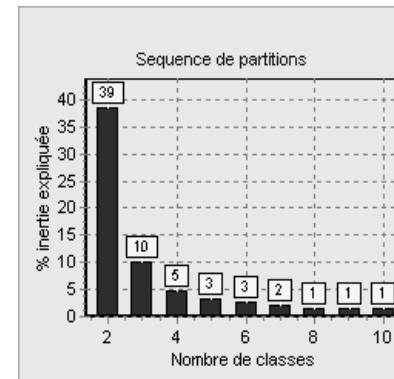
Ex. Segmentation sur la variable « body-style »



Arbres de classification

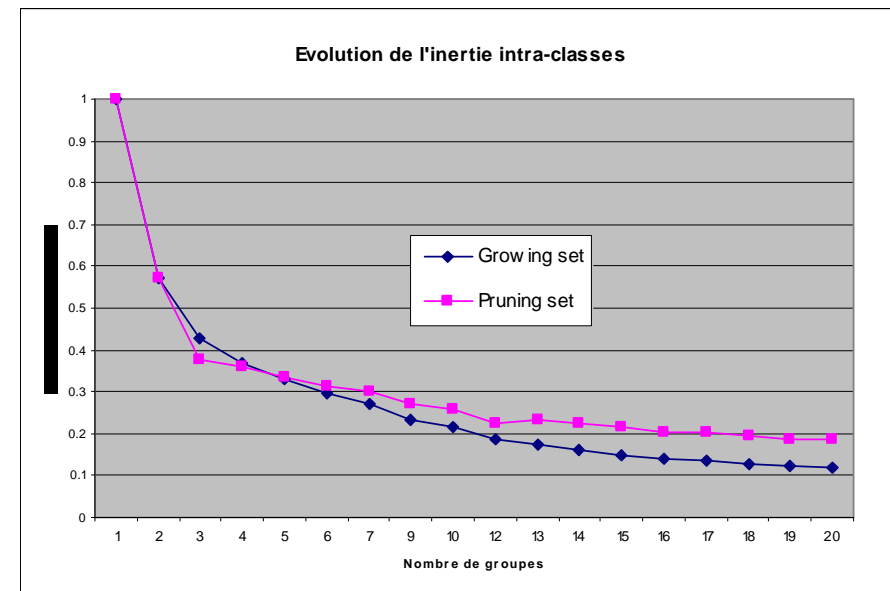
Choix du nombre adéquat de groupes

Méthode classique : Comme pour la CAH. S'inspirer des paliers (gain) et du coude (cumulée) de l'inertie expliquée, en association avec l'interprétation des groupes, pour choisir le « bon » nombre de classes.



Méthode inspirée de CART : Subdiviser le fichier en « growing » et « pruning ». Choisir le plus petit arbre avec une inertie expliquée « convenable » [Le « coude » toujours et encore]

Remarque : Attention à l'instabilité sur les petits fichiers. Comme CART, ne pas utiliser cette technique lorsque les effectifs sont faibles.



AUTOS : 3 groupes semblent le plus adéquat

Rappel : La CAH en proposait 3

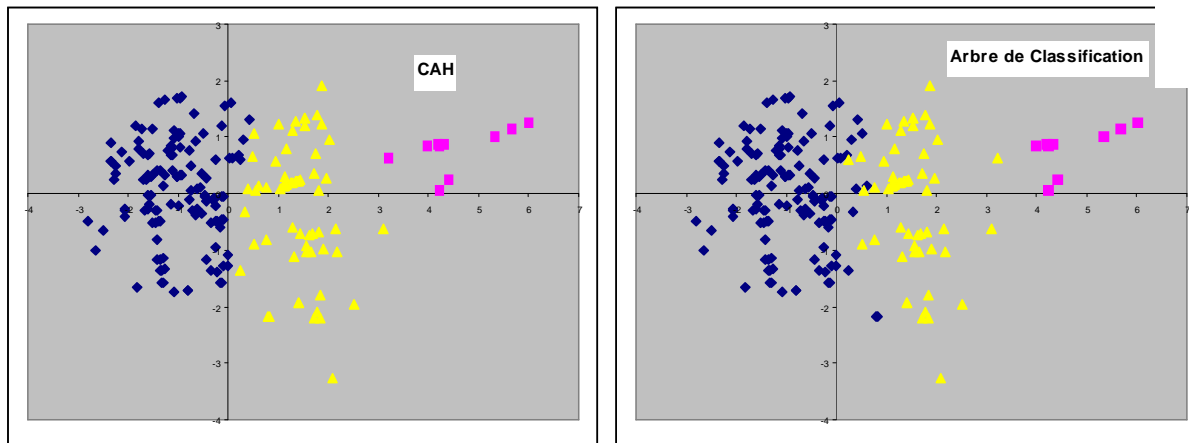
Arbres de classification

Comparaison des résultats avec la CAH

Question : Les arbres de classification introduisent une contrainte supplémentaire (exploration monothétique) dans la recherche de la solution. Est-ce que ça les pénalise ?

[On se place dans le cadre où les variables actives servent également de variables de segmentation, sinon les résultats ne seraient pas comparables]

Comparaison des solutions dans le 1er plan factoriel



Tree

- conso-ville < 10.4500 then **cluster n°1**, with 134 examples (66.67%)
- conso-ville ≥ 10.4500
 - price < 31925.0000 then **cluster n°2**, with 58 examples (28.86%)
 - price ≥ 31925.0000 then **cluster n°3**, with 9 exemples (4.48%)

Croisement des groupes construits

	c_ct_1	c_ct_2	c_ct_3	Sum
c_hac_1	127	1	0	128
c_hac_2	0	1	9	10
c_hac_3	7	56	0	63
Sum	134	58	9	201

Les performances des deux méthodes (en termes d'inertie expliquée) sont similaires dans la grande majorité des cas (cf. références)

Ex. AUTOS

- B (CAH) = 61%
- B (Arbres) = 61%

Arbres de classification

Régularisation et traitement des variables actives discrètes – Quelques pistes

Idée générique de la régularisation

Projection factorielle des données

Prendre les X premiers axes (ex. la moitié)

Distance euclidienne simple sur les axes pour les calculs d'inerties

Idée : Lisser les données pour n'utiliser que l'information « utile » et évacuer les fluctuations aléatoires (N.B. Prendre tous les axes = Travailler dans l'espace original)

Variables actives qualitatives : ACM (Analyse factorielle des correspondances multiples)

Variables actives continues (quantitatives) : ACP (Analyse en composantes principales)

Mélange variables actives qualitatives et quantitatives : Découper les variables quantitatives en intervalles (ex. méthode « fréquences égales » symétrise les distributions) + ACM

Bilan -- Les arbres de classification

Une technique de classification automatique : vise à créer des groupes homogènes au regard d'un certain nombre de variables actives

- + Permet de produire directement une règle d'affectation « industrialisable »
- + Interprétation directe des groupes à l'aide des mêmes règles
- + Rapidité/capacité à traiter de grandes bases (similaire aux arbres de décision)
- + Possibilité de guider la recherche des classes (construction interactive – s'appuyer sur la connaissance du domaine pour produire des groupes pertinents)

- + Possibilité de dissocier variables actives [expliquées] (pour apprécier l'homogénéité des groupes ex. comportement d'achat) et variables de segmentation [explicatives] (pour élaborer et expliquer les groupes ex. caractéristiques signalétiques des personnes)

- Même outils d'interprétations que les autres méthodes de typologie (projection factorielle, comparaisons univariées, etc.)
- Possibilité de description selon les variables actives et illustratives

- Problème toujours ouvert : choisir le « bon » nombre de groupes

Bibliographie

En ligne

M. Chavent, C. Guinot, Y. Lechevallier, M. Tenenhaus - « Méthodes divisives de classification et segmentation non supervisée : recherche de la typologie de la peau humaine saine », RSA, Vol. 47, N°4, pp. 87-99, 1999.

http://www.numdam.org/numdam-bin/fitem?id=RSA_1999_47_4_87_0

M. Gettler-Summa, C. Pardoux - « La classification automatique »

<http://www.ceremade.dauphine.fr/~touati/EDOGEST-seminaires/Classification.pdf>

G. Bisson - « Évaluation et catégorisation »

<http://www-lipn.univ-paris13.fr/seminaires/A3CTE/Presentations/A3CTE-Evaluation.Bisson.pdf>

Ouvrages

R. Rakotomalala, T. Le Nouvel - « Interactive Clustering Tree : Une méthode de classification descendante adaptée aux grands ensembles de données », RNTI-A-1, Numéro Spécial : « Data Mining et Apprentissage statistique : Application en assurance, banque et marketing », pp. 75-94, 2007.

J.P. Nakache, J. Confais - « Approche pragmatique de la classification », TECHNIP, 2004.

L. Lebart, A. Morineau, M. Piron - « Statistique exploratoire multidimensionnelle », DUNOD, 2004.