Clustering tree

Unsupervised learning or multi-objective predictive clustering tree

Ricco RAKOTOMALALA

Outline

1. Cluster analysis
2. Interpretation of groups (clusters)
3. Assigning a new instance to a cluster
4. Clustering tree
5. Conclusion

Cluster analysis

Clustering – Unsupervised learning

What is clustering?

X (all the descriptors are quantitative, for the moment)

No class-attribute

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 Golt 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 K 3.0 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	

The goal is to identify automatically similar objects and to gather them in groups.

Goal: grouping the objects according to their similarities (ex. Groups of customers with similar behavior, groups of cars with similar characteristic, etc.)

(1) The objects in the same group are similar

(2) The objects in distinct groups are different

Why is it interesting?

→ Discovering the underlying structure of the dataset

→ Summarizing the behaviors

→ Assigning a new instance to a group

→ Finding the “outliers” (atypical objects)

European « cars » dataset

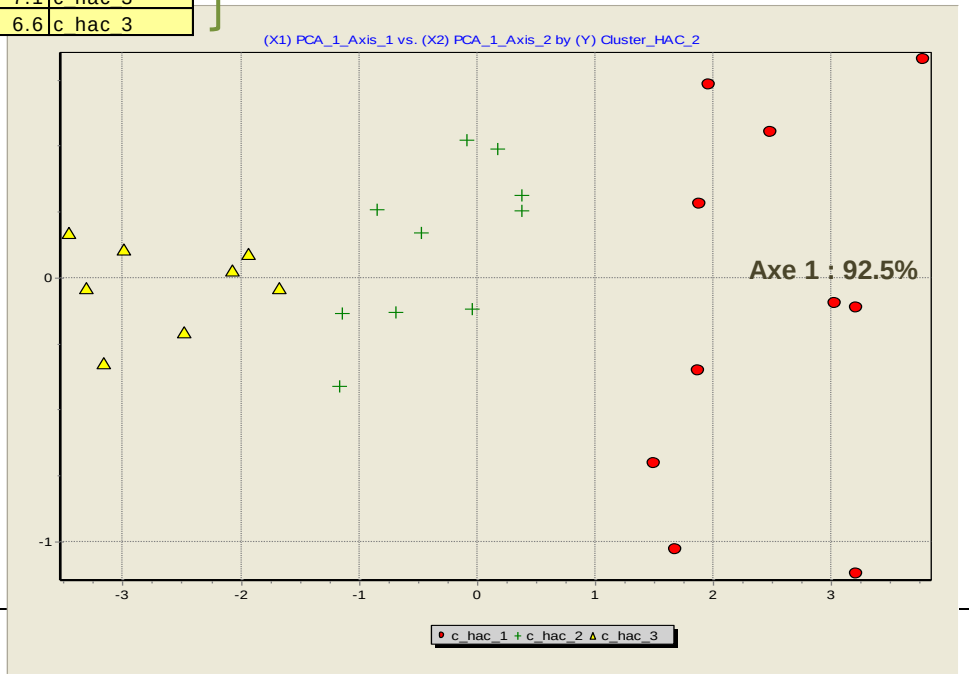
Modele	Prix	Cylindree	Puissance	Poids	Consommation	Groupe
Renault Safrane 2.2. V	36600	2165	101	1500	11.7	c_hac_1
Volvo 850 2.5	39800	2435	106	1370	10.8	c_hac_1
Hyundai Sonata 3000	38990	2972	107	1400	11.7	c_hac_1
Lancia K 3.0 LS	50800	2958	150	1550	11.9	c_hac_1
Mazda Hachtback V	36200	2497	122	1330	10.8	c_hac_1
Opel Omega 2.5i V6	47700	2496	125	1670	11.3	c_hac_1
Peugeot 806 2.0	36950	1998	89	1560	10.8	c_hac_1
Seat Alhambra 2.0	36400	1984	85	1635	11.6	c_hac_1
Toyota Previa salon	50900	2438	97	1800	12.8	c_hac_1
Volvo 960 Kombi aut	49300	2473	125	1570	12.7	c_hac_1
Opel Astra 1.6i 16V	25000	1597	74	1080	7.4	c_hac_2
Peugeot 306 XS 108	22350	1761	74	1100	9	c_hac_2
Seat Ibiza 2.0 GTI	22500	1983	85	1075	9.5	c_hac_2
VW Golt 2.0 GTI	31580	1984	85	1155	9.5	c_hac_2
Citroen ZX Volcane	28750	1998	89	1140	8.8	c_hac_2
Fiat Tempra 1.6 Liberty	22600	1580	65	1080	9.3	c_hac_2
Fort Escort 1.4i PT	20300	1390	54	1110	8.6	c_hac_2
Honda Civic Joker 1.4	19900	1396	66	1140	7.7	c_hac_2
Mitsubishi Galant	31990	1998	66	1300	7.6	c_hac_2
Nissan Primera 2.0	26950	1997	92	1240	9.2	c_hac_2
Daihatsu Cuore	11600	846	32	650	5.7	c_hac_3
Suzuki Swift 1.0 GLS	12490	993	39	790	5.8	c_hac_3
Fiat Panda Mambo L	10450	899	29	730	6.1	c_hac_3
VW Polo 1.4 60	17140	1390	44	955	6.5	c_hac_3
Opel Corsa 1.2i Eco	14825	1195	33	895	6.8	c_hac_3
Subaru Vivio 4WD	13730	658	32	740	6.8	c_hac_3
Toyota Corolla	19490	1331	55	1010	7.1	c_hac_3
Ford Fiesta 1.2 Zetec	19740	1242	55	940	6.6	c_hac_3

1

2

3

3 groups of cars: small, medium, large/van



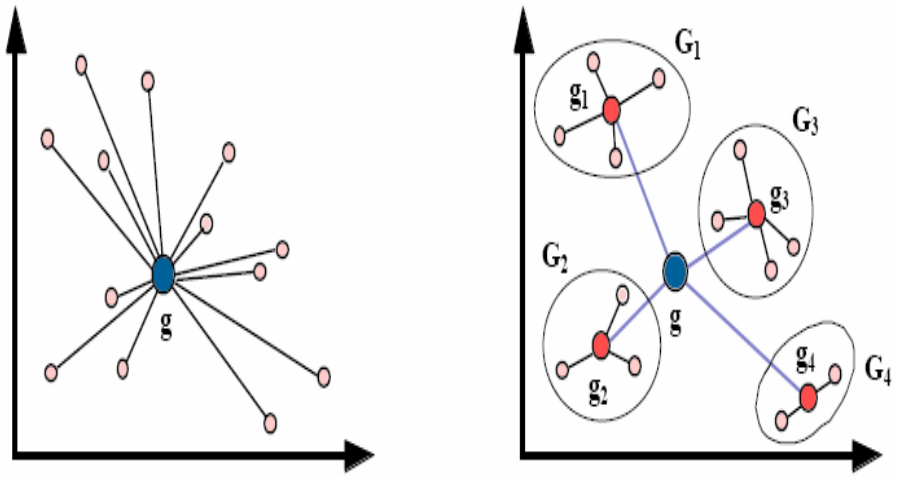
Objects in the two first principal components (PCA)

Small cars — Medium — Large/Van

Ricco Rakotomalala
Tanagra Tutorials - http://data-mining-tutorials.blogspot.fr/

Clustering instances – Main issues

Illustration in a scatter diagram (2-dimensional dataset)



How to calculate?

- The similarity between two instances
- The similarity between two groups
- The similarity between 1 instance and 1 group (during the learning phase and the deployment phase)

- The compactness of a group
- The global separability of groups

Agglomerative Hierarchical Clustering

Outline of the algorithm

- Calculating the distance between pairs of instances
- Successive agglomerations by merging the most similar groups (depending on the linkage criterion e.g. single linkage, complete linkage, Ward, etc.)
- Height = Distance between groups

Advantage

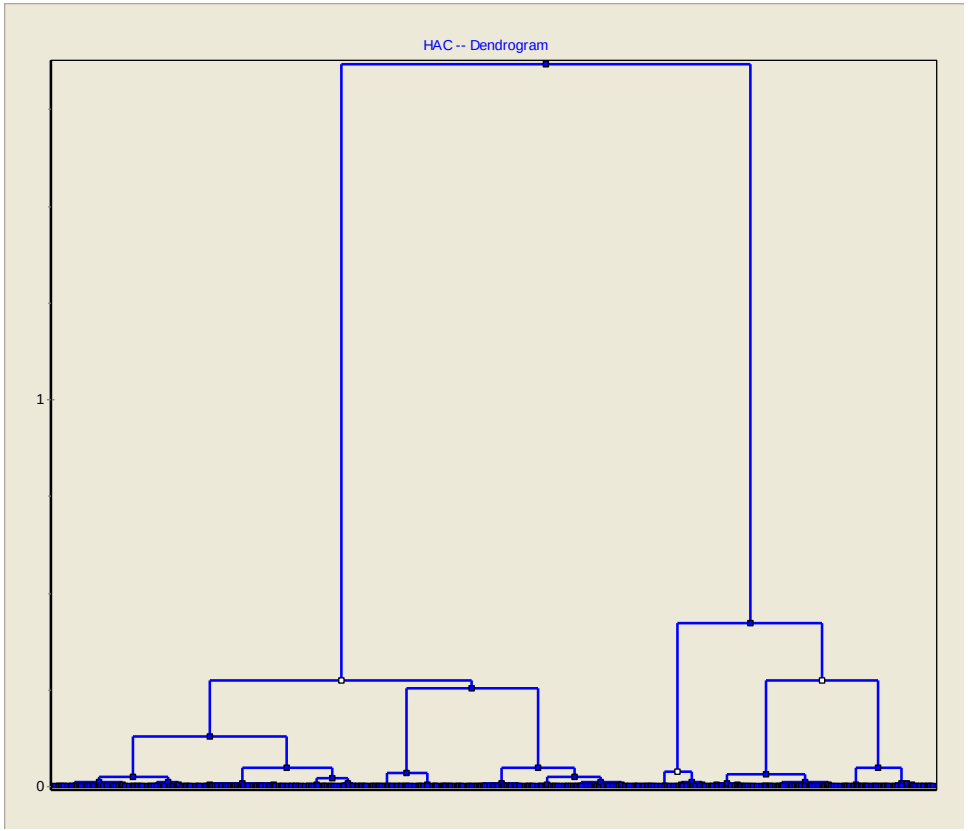
- Partition hierarchy
- Give indications about the similarity between groups
- Propose alternative solutions (nested solutions)

Inconvenient

- Not practical for large dataset

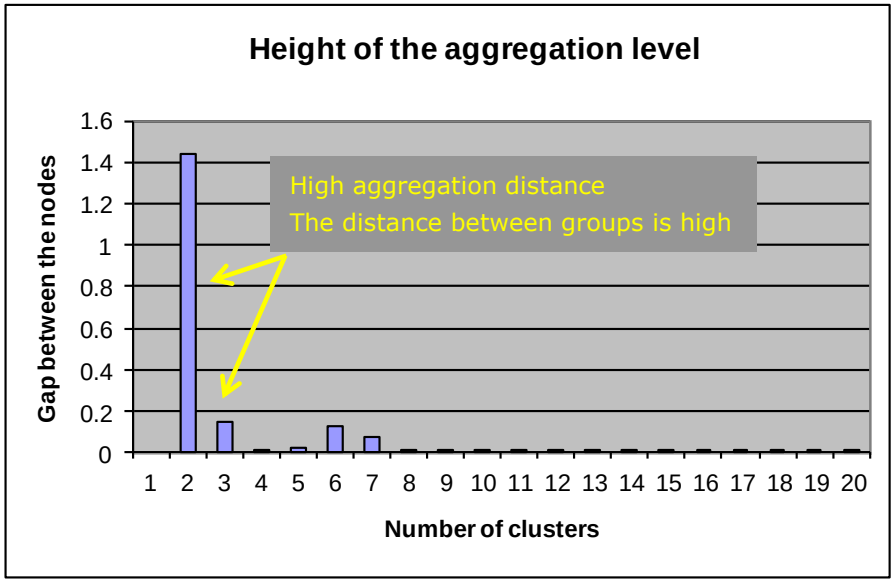
Recurring issues in the clustering process

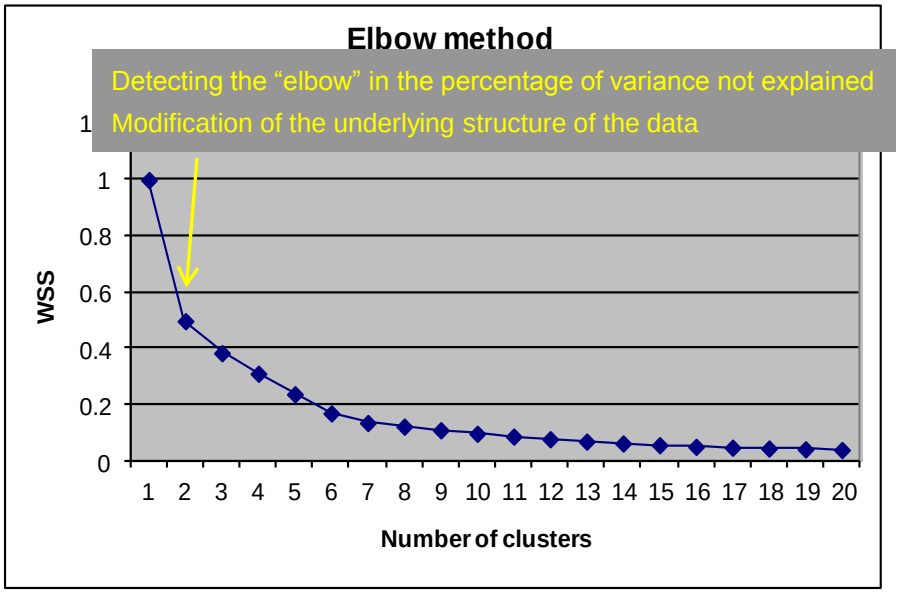
- Determining the number of clusters
- Interpreting the groups using the active variables
- (And) using the illustrative variables
- Assigning a new observation to a group



→ 201 instances
→ 4 active variables (“costs”: price, highway mpg and city mpg, insurance)
→ 2 first principal components → 92% of available information
→ Ward’s method
→ 3 groups are highlighted (the “2 groups” solution is often trivial)

Determining the number of clusters





Choosing the right number of groups remains an open issue

We must consider the interpretation of the groups and the specification of the analysis

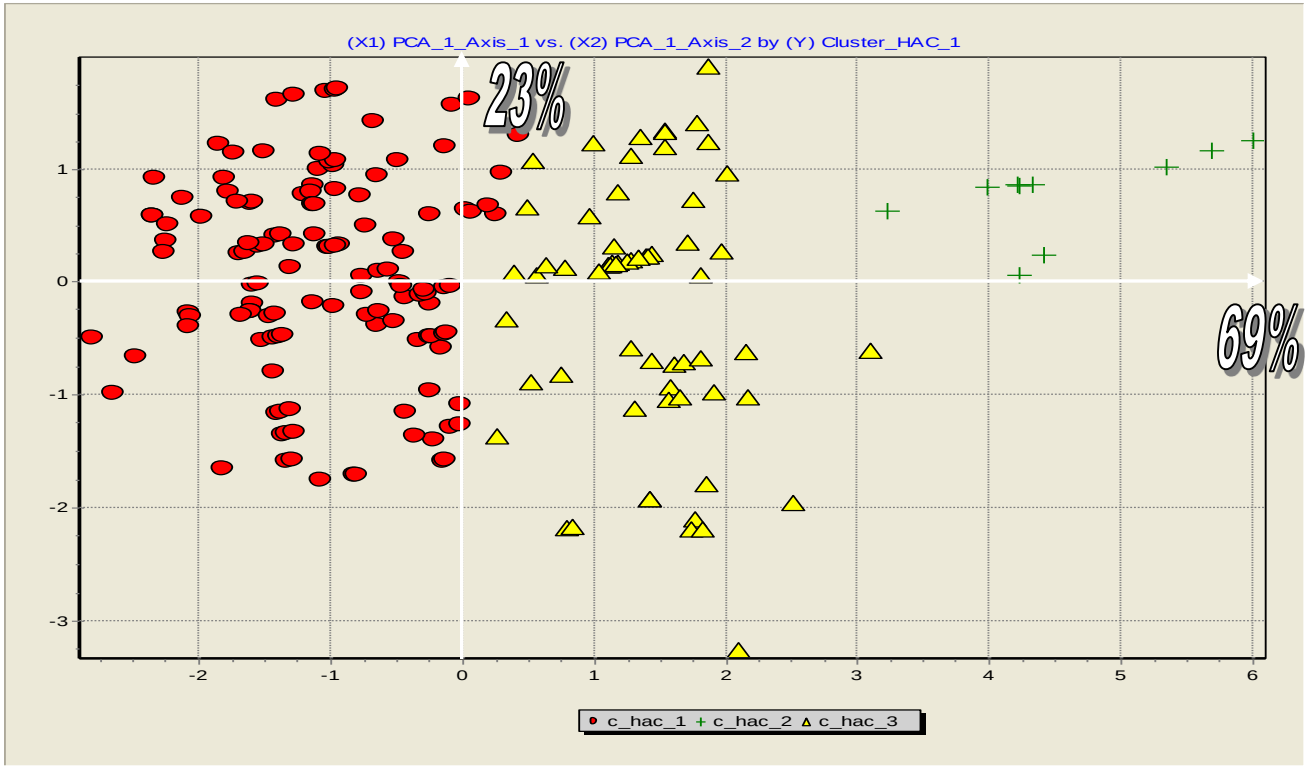
Interpretation of the groups

Understanding the nature of groups

Using active and illustrative variables

Interpretation using a factor analysis

Principle: Explain the relative position of groups by interpreting the factors.



Advantage

Multivariate approach.

Visualizing the objects can help to choose the right number of clusters.

Drawback

- Giving a interpretation of factors is not always obvious. Especially when more than two factors are relevant.
- The groups were constructed without taking into account the orthogonality constraint of the factor analysis.

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 %)*

Interpretation using conditional descriptive statistics

Results

Description of "Cluster_HAC_1"

Cluster_HAC_1=c_hac_1			
Examples			[63.7 %] 128
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
normalized-losses	-4.7	113.61 (26.82)	121.33 (30.55)
price	-10.0	8899.70 (2714.81)	12904.96 (7492.72)
conso-autoroute	-11.0	6.93 (0.99)	8.02 (1.86)
conso-ville	-11.5	8.33 (1.30)	9.90 (2.56)
Discrete attributes : [Recall] Accuracy			

Cluster_HAC_1=c_hac_2			
Examples			[5.0 %] 10
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
price	10.3	36835.50 (4483.50)	12904.96 (7492.72)
conso-autoroute	8.2	12.74 (1.43)	8.02 (1.86)
conso-ville	7.4	15.76 (1.16)	9.90 (2.56)
normalized-losses	0.5	126.30 (9.09)	121.33 (30.55)
Discrete attributes : [Recall] Accuracy			

Cluster_HAC_1=c_hac_3			
Examples			[31.3 %] 63
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
conso-ville	8.4	12.16 (1.29)	9.90 (2.56)
conso-autoroute	7.5	9.48 (0.80)	8.02 (1.86)
price	5.5	17244.13 (4297.33)	12904.96 (7492.72)
normalized-losses	4.7	136.24 (34.26)	121.33 (30.55)
Discrete attributes : [Recall] Accuracy			

Aim. Comparing the features (mean, proportion) computed on the whole dataset and on the concerned group.

Advantage. Easy to calculate and to understand. The test value (vt) allows to identify the significance of the difference..

Inconvenient. Univariate approach. Do not take into account the relation between the variables. Hard to read when we have a large number of variables.

Test value (VT): Statistical test. Comparison to an expected value (reference).

1 : The reference is computed on the whole dataset

2 : The samples are nested

3 : The VT is biased on the active variables, but not on the illustrative variables

4 : Critical value +/- 2 (roughly normal distribution)

5 : Instead of the comparison with a critical value, it is better to identify the extreme values, the oppositions, etc.

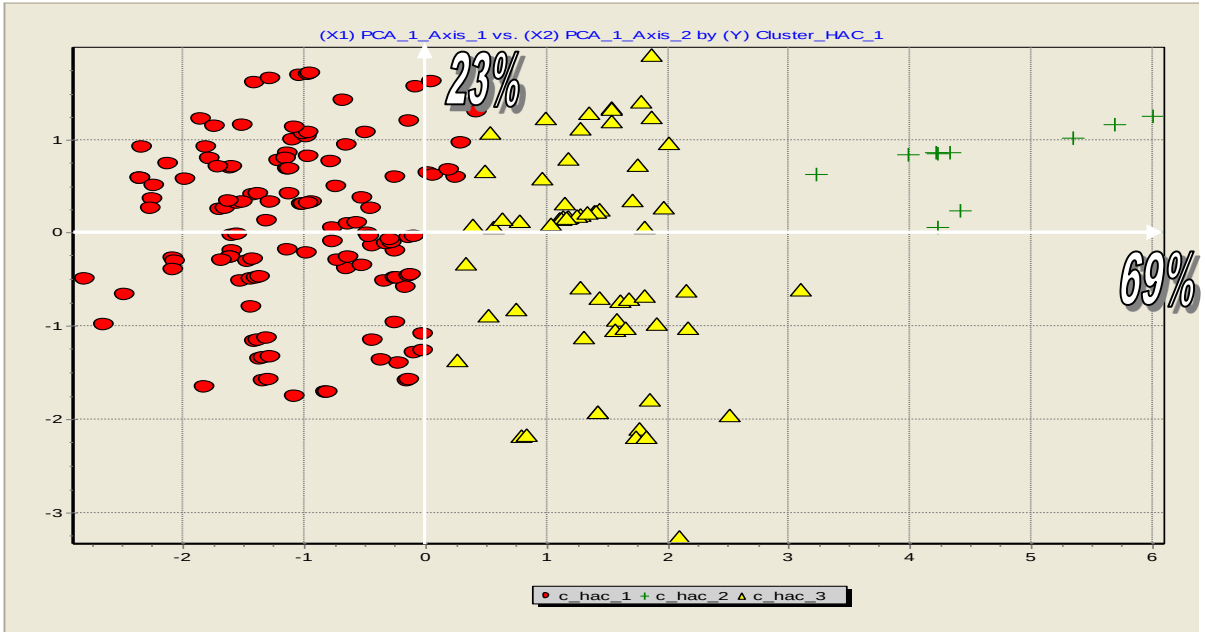
Quantitative (Mean) [group k]

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

Categorical (Proportion) [group k, level j]

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

Factor analysis or conditional descriptive statistics ?



Results

Description of "Cluster_HAC_1"

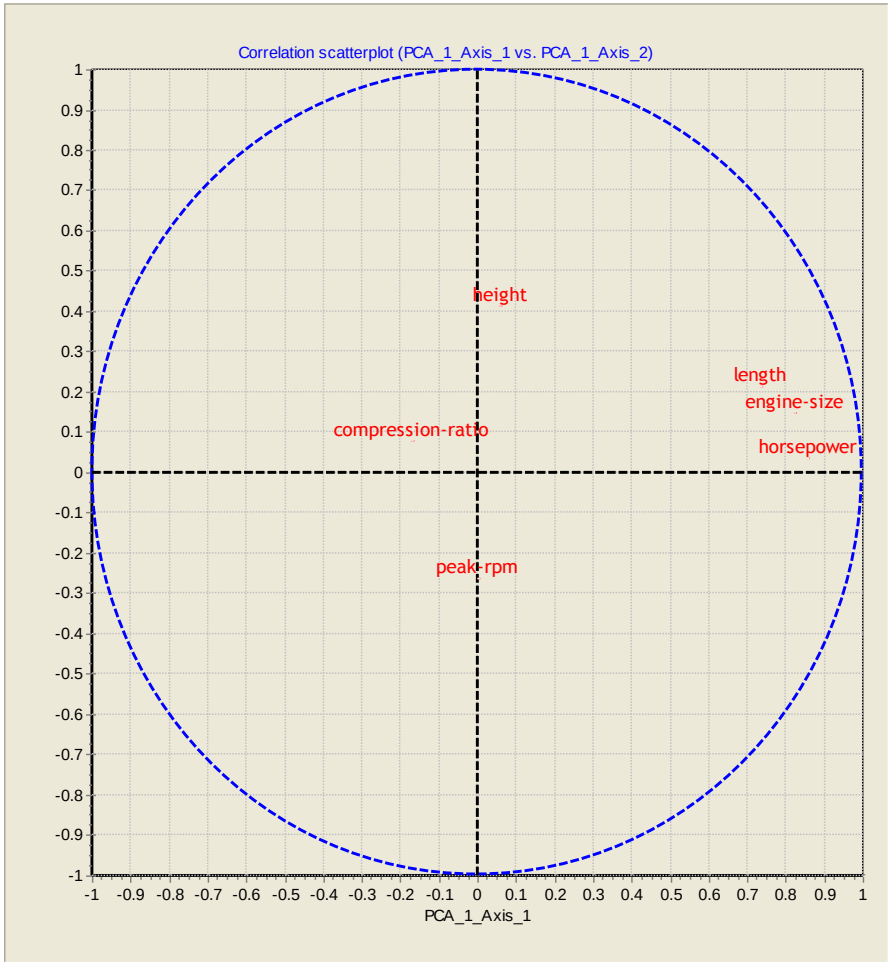
Cluster_HAC_1=c_hac_1				Cluster_HAC_1=c_hac_2				Cluster_HAC_1=c_hac_3			
Examples		[63.7 %] 128		Examples		[5.0 %] 10		Examples		[31.3 %] 63	
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)				Continuous 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			

The approaches are complementary.

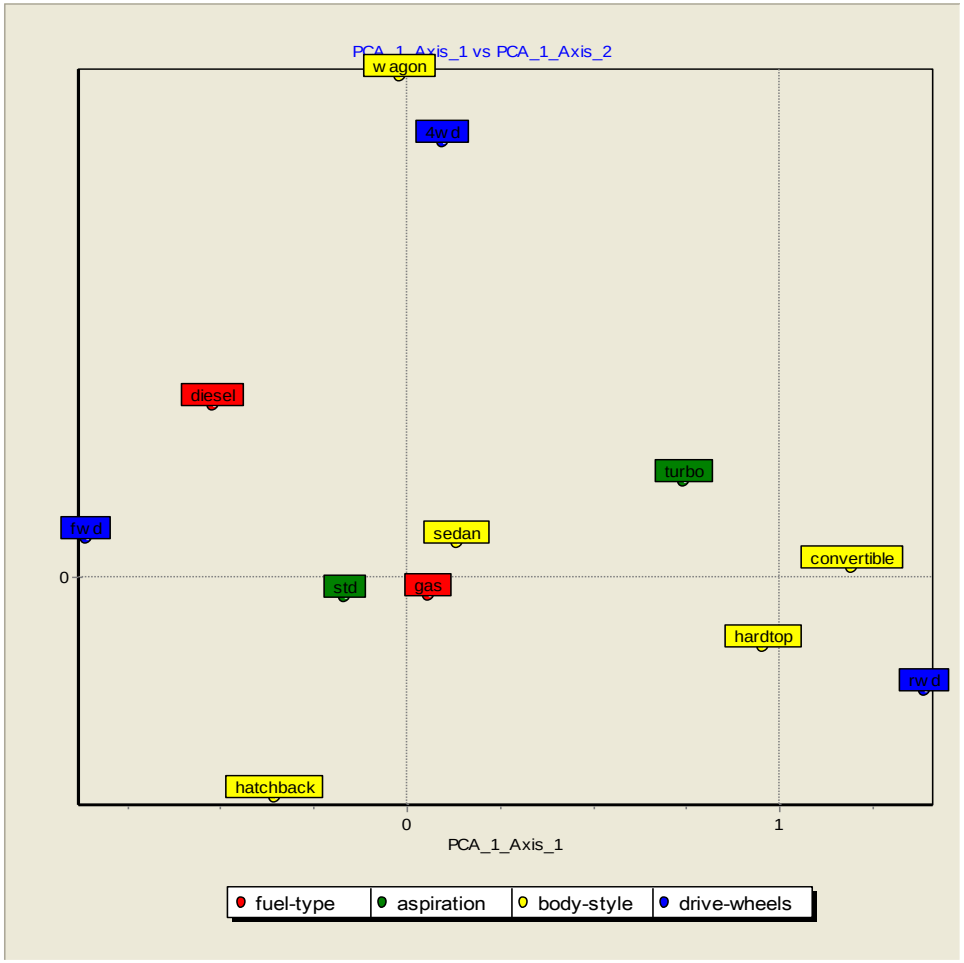
Illustrative variables

Variables not used for the construction of clusters. But used for a better interpretation and/or to highlight other types of characteristics of the objects.

Quantitative variables



Categorical variables



The same difficulties inherent to the reading of factor analysis results.

Illustrative variables

Using the conditional descriptive statistics

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	Overral	Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
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%**

The importance of the variables can be detected one by one.

We can filter the results to visualize only the results with a high |VT|

The results can be hard to read when we have a large number of variables

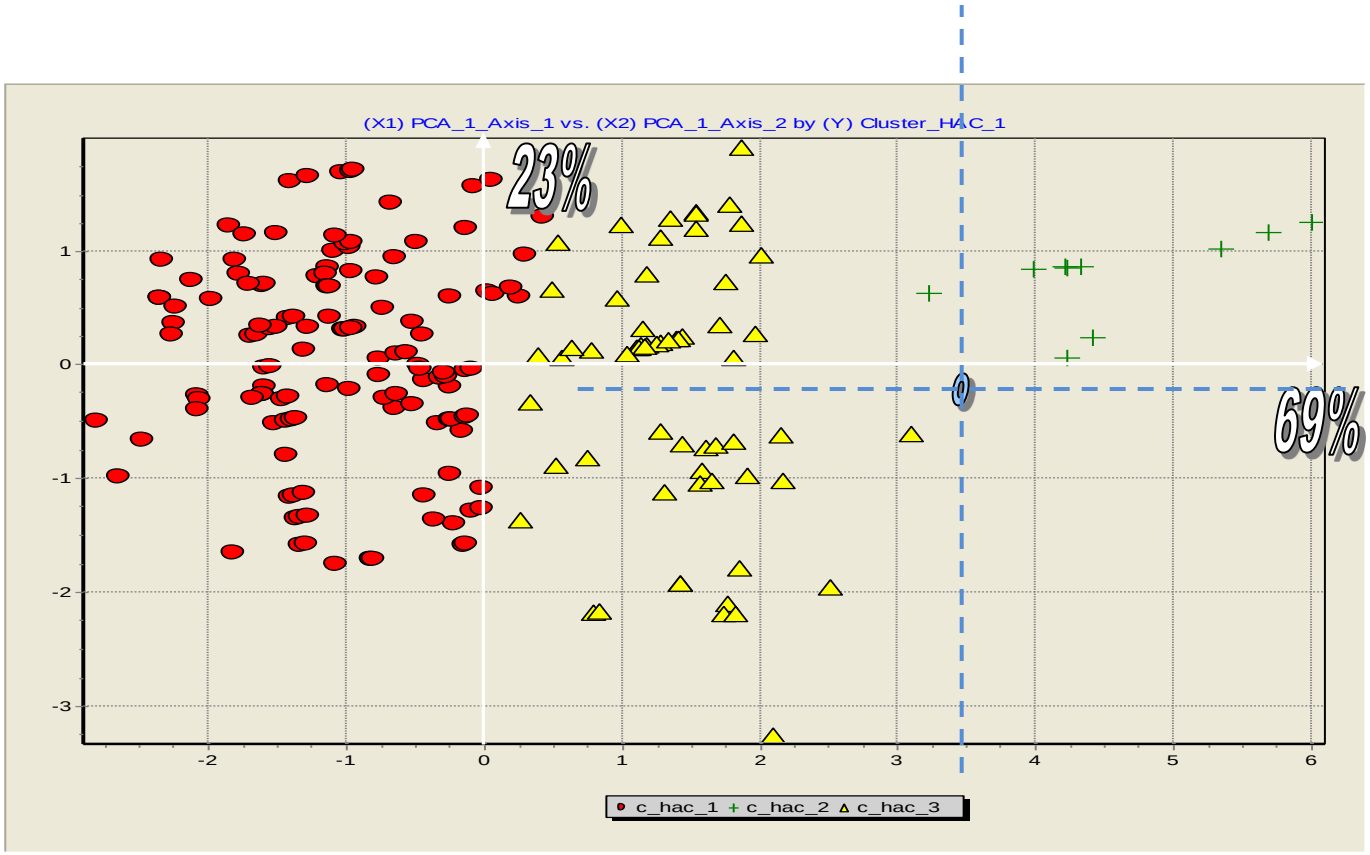
Assigning an instance to a cluster

Classification based on the active and/or illustrative variables

Classifying a new instance

Idea: Finding the cluster which is the closest

Caution: This must rely on the same distance and the same aggregation strategy used during the groups' construction (e.g. single linkage, ward's criterion, etc.)



Which group for o ?

Single linkage : Δ

Complete linkage : +

Not easy in practice, we must dispose to all the instances.

The interpretation is difficult. Why an instance is assigned to a group?

Classifying a new instance as from the illustrative variables

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	-

Two types of variables

To identify the group membership

E.g. Intrinsic characteristics of cars

E.g. Identification of the customers (age, sex, etc.)

Multi-objective supervised learning

To characterize the homogeneity of the groups

E.g. To establish groups according to the costs of the cars

E.g. To establish groups according to the customer's behavior

Issue: The distance-based classification is no longer really relevant

Goal: We want to obtain an easy-to-use classification rule (usable in the information systems)

Observation: The framework is similar to the supervised learning process, but we have several target attributes. The definition of the groups is based on multivariate characteristics.

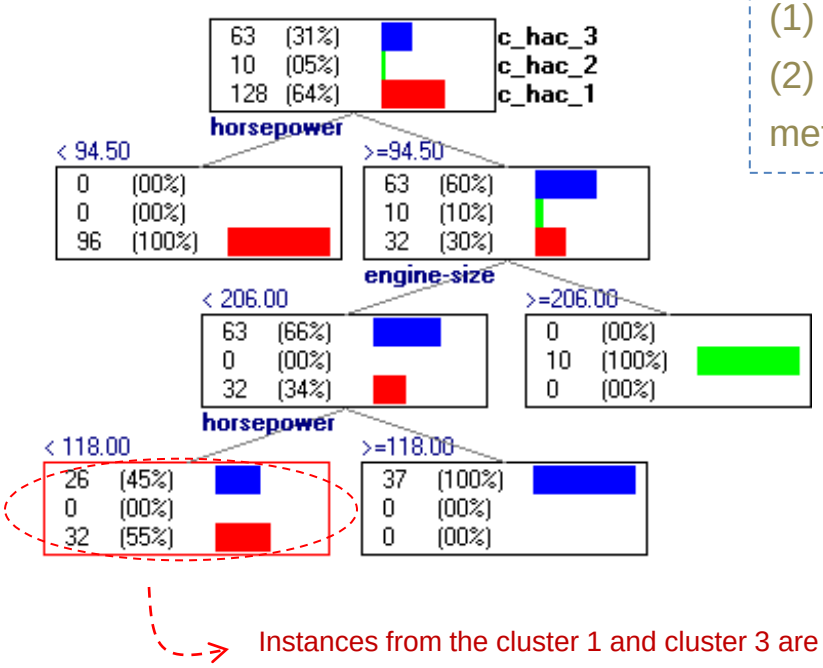
Classification of a new instance

Using a classification tree

Process in two-steps:

(1) Create clusters using a standard cluster analysis approach

(2) Use the clusters as target attribute in a supervised learning method (e.g. classification tree approach)

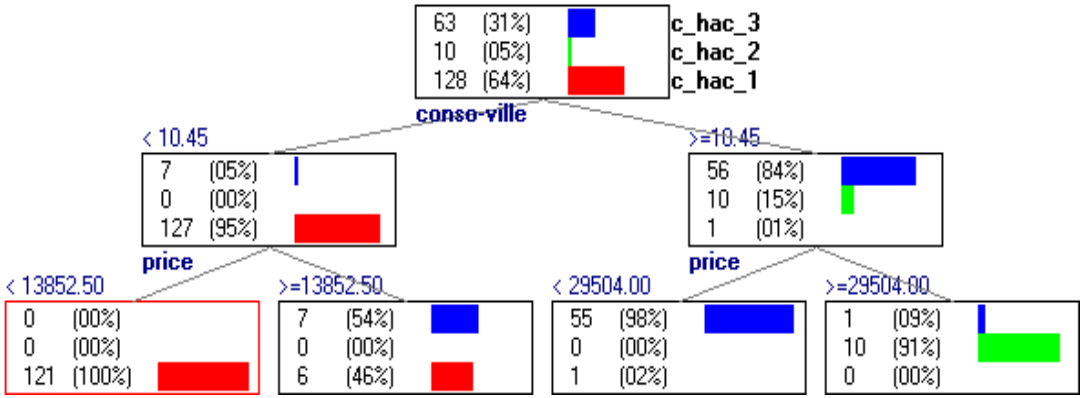


Instances from the cluster 1 and cluster 3 are gathered in this leaf.

Advantage: Very easy to perform

Inconvenient : Some leaves of the tree can be not pure according to the target attribute i.e. we have a mixture of instances from different clusters in one leaf.

Remark: this process can be extended to the characterization of the clusters with the active variables (with the same drawbacks).



Clustering tree

An extension of the regression tree

Clustering tree

Goal : Using the segmentation process in a multivariate problem i.e. linking the creation of "homogeneous" groups with the construction of the classification rule.

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

Used for establishing groups' homogeneity

Used for describing the groups

Remark: Y = X is a specific case → usual clustering problem

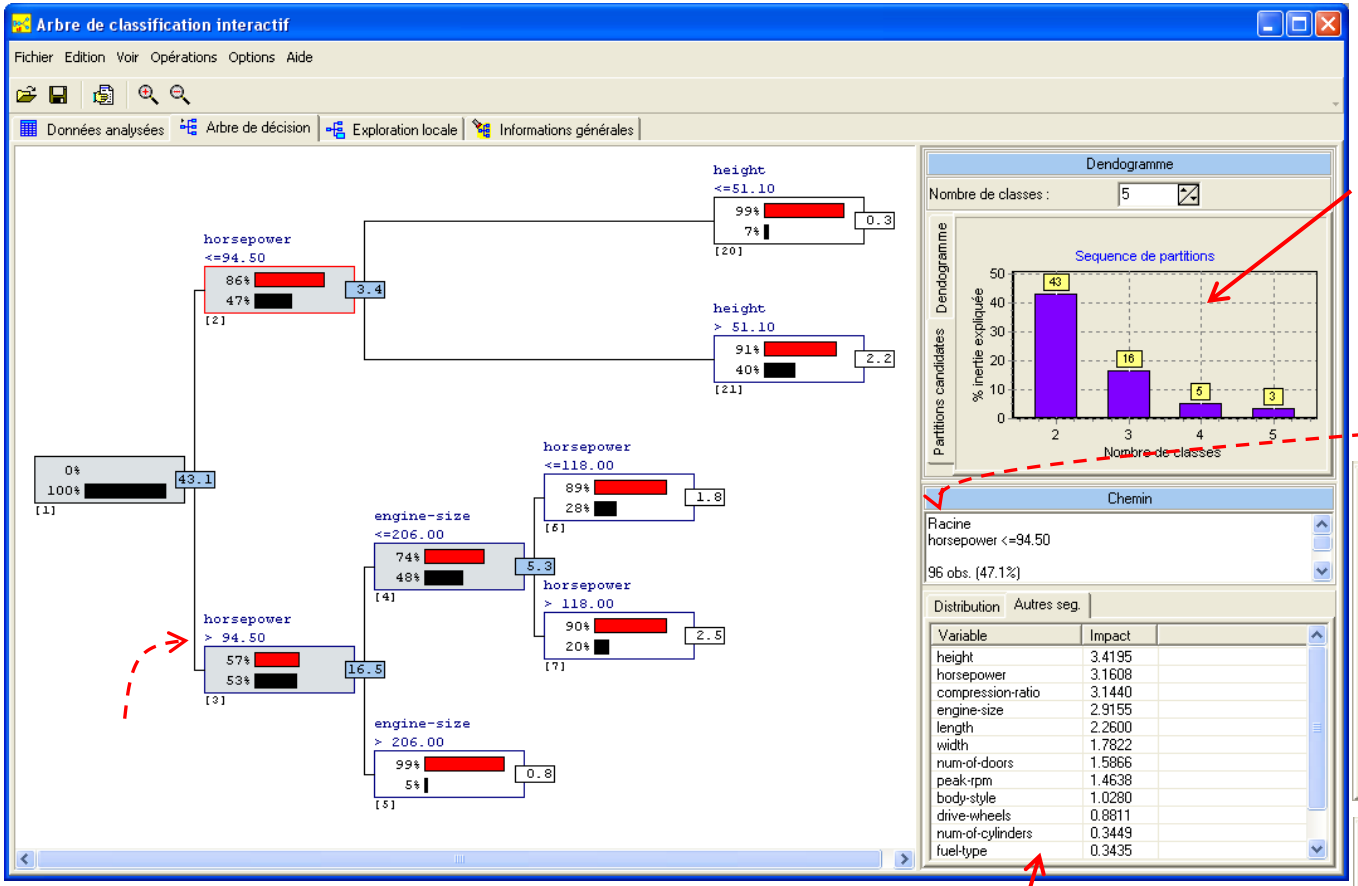
Attribute	Category	Informations	
fuel-type	Discrete	2 values	X
aspiration	Discrete	2 values	
num-of-doors	Discrete	3 values	
body-style	Discrete	5 values	
drive-wheels	Discrete	3 values	
wheel-base	Continue	-	
length	Continue	-	
width	Continue	-	
height	Continue	-	
curb-weight	Continue	-	
num-of-cylinders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

Same issues as for the classification/regression tree algorithm

- How to choose the splitting variable for a node
- How to calculate the cutting value for a continuous attribute
- How to merge the levels of a categorical variable for a binary split
- How to determine the right size of the tree and thus the right number of groups

→ The assignment of a new individual to a group (leaf) is easy by reading the tree

Clustering tree – "CARS" dataset



Percentage of explained variance (inertia) for each merging

[Detecting the right number of groups]

Issue n°3

Description of the current group

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

Goodness of split of variables for the current node

Issue n°1

Cutting value for a continuous attribute

[Merging strategy for levels of categorical attributes]

Issue n°2

Clustering tree

Selection of the splitting attribute

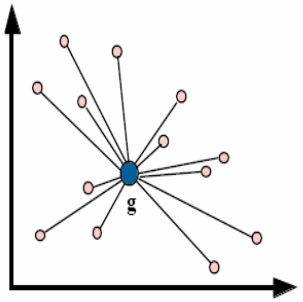
Measuring the homogeneity: Generalizing the variance notion to the inertia (multidimensional variance).

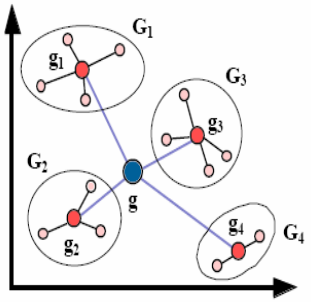
Measure: Generalization of the variance decomposition, Huygens theorem.

- Selecting the variable which maximizes the inertia gain c.-à-d. between inertia B = T - W
- Inside a group, the instances are similar, close to the conditional centroid
- The groups' centroids are distant each other (or distant to the global centroid)

$$T = B + W$$

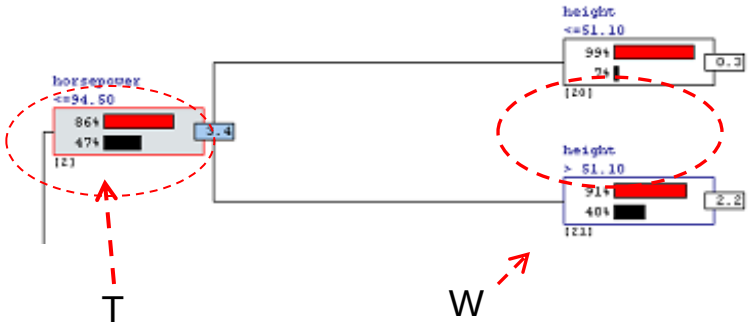
$$\sum_{i=1}^{n} p_i d^2(i,g) = \sum_{k=1}^{K} p_k d^2(g_k,g) + \sum_{k=1}^{K} \sum_{i \in G_k} p_i d^2(i,g_k)$$

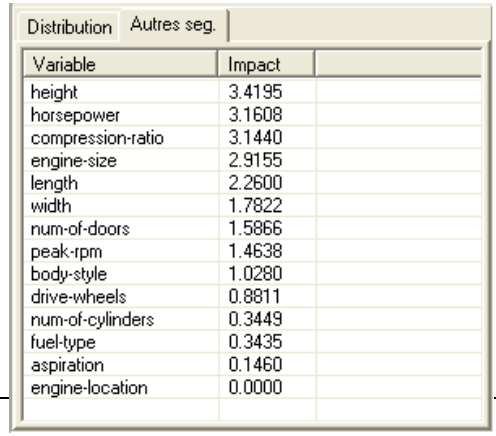




In concrete terms

For a node to split





Distribution | Autres seg.

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

Sorting the variables according to the inertia gain and selecting the one with the highest value.

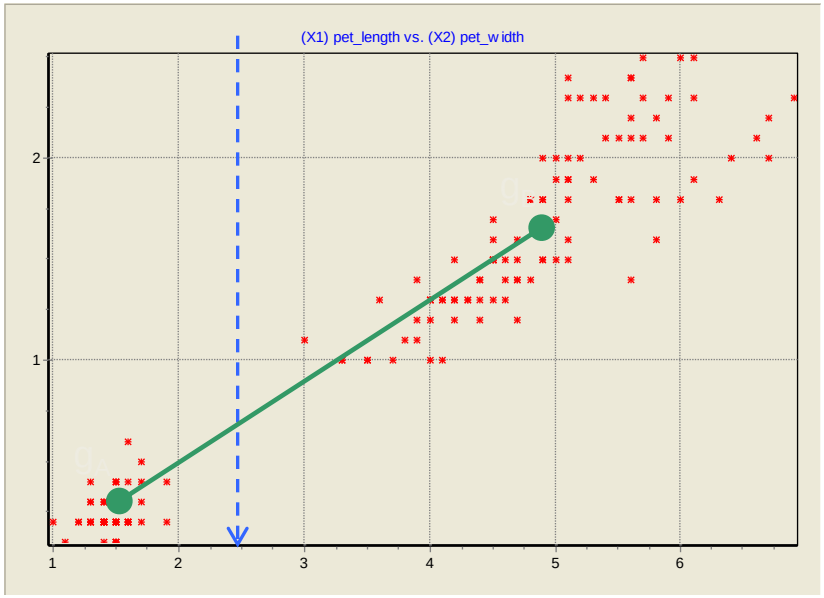
[Like for some decision tree tools, the user can choose interactively the splitting variable on a node]

Clustering tree

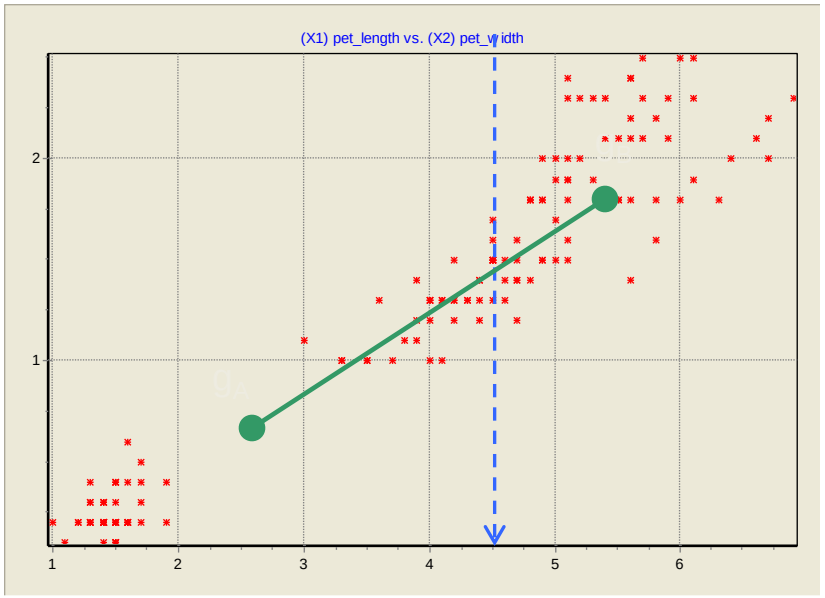
Handling continuous attribute

Principle: Like for classification tree learning, we try all of the cut points after sorting the variable, we select the one which maximizes the goodness of split measure

E.g. Two competing cutting points, which is the better?



Δ = B/T = 82%



Δ = B/T = 63%

Another formula for the inertia gain

WARD's criterion

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

Sorting: O(n log n)…

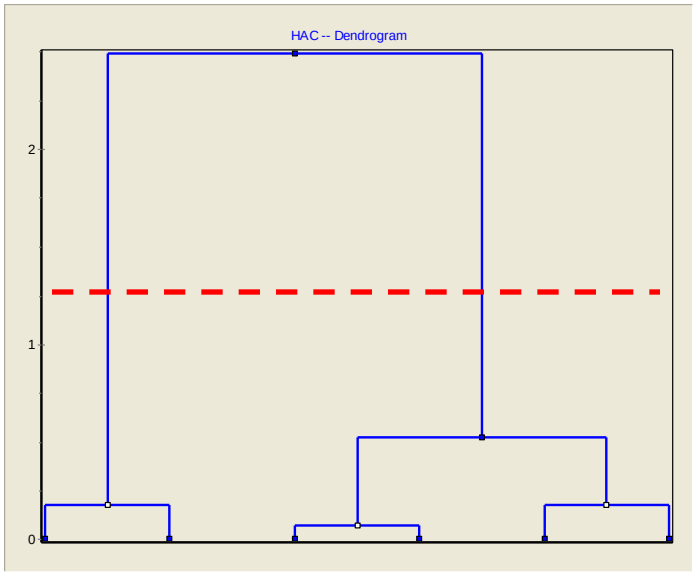
Detecting the cut point: O(n)…

Clustering tree

Binarization for categorical attribute

Principle: The “binarization” enables to avoid the data fragmentation (see CART algorithm)

How to merge the L levels (L > 2) of a categorical attribute?



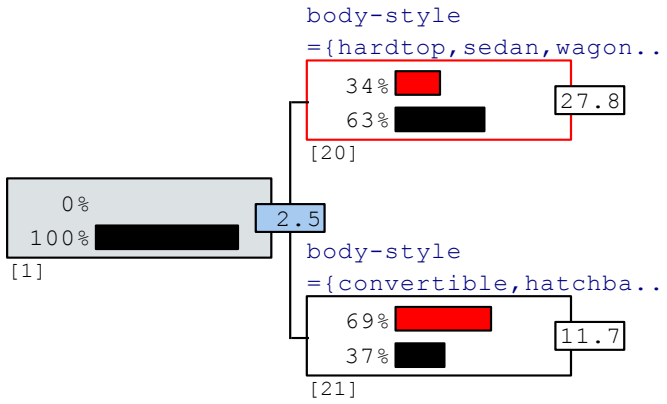
m_1 m_4 m_5 m_3 m_6 m_2

Approach:

- Gather together the levels which are similar according the Ward's criterion
- Continue the process (agglomerative approach) until we obtain 2 groups

→ Thus, we perform a hierarchical agglomerative clustering on levels of the categorical variable

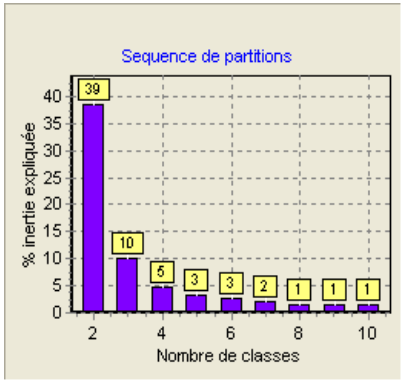
E.g. Splitting a node with “body-style” (Cars dataset)

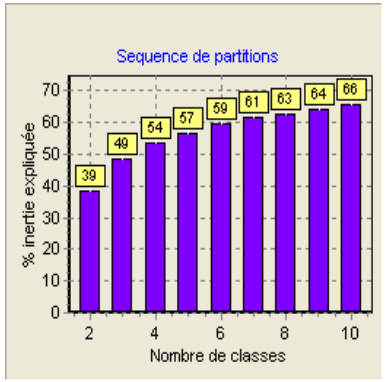


Clustering tree

Determining the right number of groups

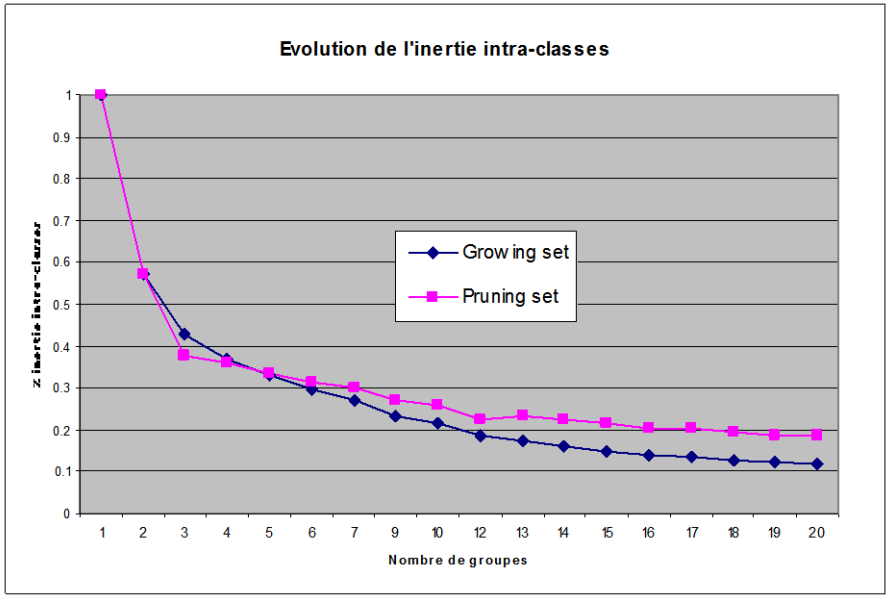
Standard approach: As for HAC, considering the aggregation levels (gain) and the elbow in the cumulative explained inertia help to determine the right number of groups. The interpretation of the results is also an essential component.





Suggested by CART approach: We subdivide the dataset into growing and pruning sets. We detect the elbow in the within-class inertia curve according to the number of groups (the number of leaves in the tree)

Note: This approach is valid only if we have a large dataset.



AUTOS : “3 groups” seems the right solution

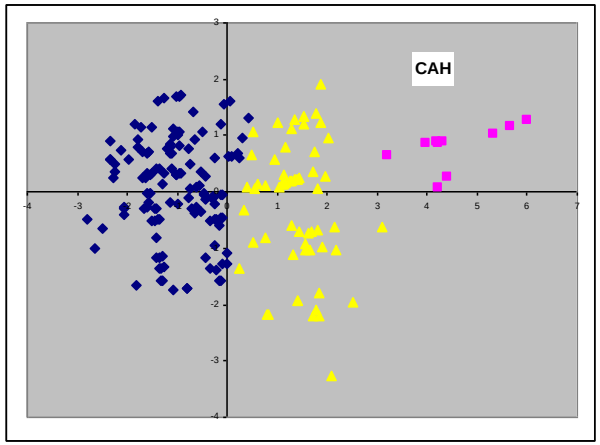
The cumulative explained inertia above suggests the same solution

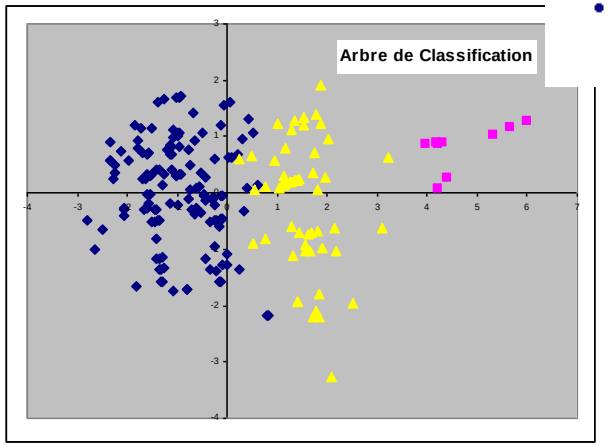
Clustering tree

Comparing the obtained groups with the standard HAC algorithm

Issue:: The clustering tree algorithm has an additional constraint compared to the well-known clustering algorithms, this is a “monothetic” approach i.e. only one variable is used to create the subgroups at each step. What is the influence of this constraint on the quality of the partition?

Comparison of the groups in the first factorial map (PCA)





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 examples (4.48%)

Contingency table between the groups

	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

The performances of the two approaches - in terms of explained inertia - are very similar in most cases (see references)

Ex. CARS: B (HAC) = 61% - B (Tree) = 61%

Clustering tree

Regularization and the handling the discrete variables

Idea underlying the regularization for clustering tree

Perform a factor analysis

Pick the most relevant factors

Use the Euclidean distance for the calculation of the inertia

Why? By removing the irrelevant factors, we reduce the influence of sampling fluctuations and treat only the relevant information from the data (if we use all the factors, we obtain the same results than dealing with the original variables).

Categorical active variables: we use Multiple Correspondence Analysis (MCA)

Quantitative active variables: we use Principal Component Analysis (PCA)

Mixture of categorical and quantitative active variables: (1) we discretize the continuous variables – e.g. equal frequency – and we perform a MCA ; or (2) we use the factorial analysis for mixed data (FAMD)

Clustering tree - Conclusion

A clustering approach which detects homogeneous groups on the basis of similarities between objects :

+ We obtain an easy to use classification rule
+ The rules give a first interpretation of the constitution of the groups
+ The approach can deal with very large databases (such as decision tree learning algorithm)
+ We can guide the process in interactive way (e.g. by selecting the splitting variable on a node)
+ We have the same tools than the usual clustering approach for the interpretation of the groups (factor analysis, conditional descriptive statistics)
+ We can interpret the groups with active and illustrative variables

+ The ability to distinguish the explained variables (which enable to define the homogeneity of the groups) and the explicative variables (which enable to define the groups) is a very significant feature ➔ predictive clustering tree

- The detection of the “right” number of clusters remains an open issue

References

Article

M. Chavent, « A monothetic clustering method », Pattern Recognition Letters, 19, pp. 989—996, 1998.

H. Blockeel, L. de Raedt, J. Ramon, "Top-Down Induction of Clustering Trees", ICML'98, pp. 55-63, 1998.

Chapter of book

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.

Tutorial

Tanagra, "Clustering trees", November 2008.