Subject

Most of the time, the statistician must build groups of individuals and want to characterize them. The main interest of this very simple approach is that the results are easy to read and understand.

In this tutorial, we show how to build groups with some (target) attributes, and describe them with other (input) attributes.

Dataset

We use the « AUTOS.XLS » dataset with 205 examples.

We want to describe the cars starting from their consumption, price, horsepower and bodystyle according to their fuel-type (GAS or DIESEL) and aspiration (STD or TURBO).

Description and interpretation of groups with TANAGRA

Download the dataset

First of all, we import the dataset. We use the FILE / NEW menu.

Choose your dataset and	start download			
Diagram title : Default title			_	
Data mining diagram t D:\Temp\Exe\default.b	ile name : idm			
Dataset (*.txt*.arff.*xls):		(e)	
	ок	Tanagra	N	? 🗙
		Regarder dans .	didac data	
		Mes documents récents		
		Direau		
		Mes documents		
Data visualization	Statistics	Mes documents		
Feature selection Spv learning	Regression Meta-spv learning	Poste de travail		
Correlation scatterplot	Export dataset		Nom du fichier :	: autos.xls Ouvrir
		Favoris réseau	Fichiers de type	e: Excel File (97 & 2000) Annuler

Attribute selection

We want to build groups according to FUEL-TYPE and ASPIRATION; we set these attributes as TARGET. We want to describe groups according to BODY-STYLE, HORSEPOWER, CITY-MPG and PRICE; we set them as INPUT.

Define attribute statuses	Define attribute statuses
Parameters Attributes : C normalized-losses Input Input	Parameters Attributes : Inum-of-cylinders Inum-of-cylinders <td< th=""></td<>
OK Cancel Help	OK Cancel Help

Groups from one target attribute

We use the GROUP CARACTERISATION component for the construction of independent groups from the two target attributes.



This component gives two independent reports for each target attributes. They allow us to understand the specificities of each group.

			De	scription	of "fuel-type"				
	fuel-type:	=gas			ſ	uel-type=d	liesel		
Examples			[90).2 %] 185	Examples			[9.8 %] 20
Att - Desc	Test value	Group		Overral	Att - Desc	Test value	Group		Overral
Continuous attributes					Continuous attributes				
horsepower	2.4		106.40	104.26	city-mpg	3.6		30.30	25.22
price	-1.6	12	922.69	13207.13	price	1.6	15	838.15	13207.13
city-mpg	-3.6		24.67	25.22	horsepower	-2.4		84.45	104.26
Discrete attributes					Discrete attributes				
body-style=hatchback	2.9	[98.6 %]	37.3 %	34.1 %	body-style=sedan	2.7	[15.6 %]	75.0 %	46.8 %
body-style=convertible	0.8	[100.0 %]	3.2 %	2.9 %	body-style=wagon	0.4	[12.0 %]	15.0 %	12.2 %
body-style=hardtop	-0.3	[87.5%]	3.8 %	3.9 %	body-style=hardtop	0.3	[12.5 %]	5.0 %	3.9 %
body-style=wagon	-0.4	[88.0%]	11.9 %	12.2 %	body-style=convertible	-0.8	[0.0%]	0.0 %	2.9 %
body-style=sedan	-2.7	[84.4%]	43.8 %	46.8 %	body-style=hatchback	-2.9	[1.4%]	5.0 %	34.1 %
			De	scription	of "aspiration"				
	vointion			semption		unination_	hunha		
aspiration=std			0.02.470	¢	ispiration=	turbo		0.0.02.07	
Examples	Taskastas	0	[02	2.0%] 168	Examples	Taskastas	0	Ľ	18.U % J 37
Att - Desc	Test value	Group		Overral	Att - Desc	Test value	Group		Overral
Continuous attributes					Continuous attributes				
city-mpg	2.9		25.84	25.22	horsepower	3.4		124.43	104.26
price	-2.5	12	554.06	13207.13	price	2.5	16	172.44	13207.13
horsepower	-3.4		99.81	104.26	city-mpg	-2.9		22,41	25.22
Discrete attributes					Discrete attributes				
body-style=convertible	1.2	[100.0%]	3.6 %	2.9 %	body-style=wagon	0.3	[20.0%]	13.5 %	12.2 %
body-style=hardtop	0.4	[87.5%]	4.2 %	3.9 %	body-style=sedan	0.2	[18.8 %]	48.6 %	46.8 %
body-style=hatchback	-0.1	[81.4%]	33.9 %	34.1 %	body-style=hatchback	0.1	[18.6 %]	35.1%	34.1 %
body-style=sedan	-0.2	[81.3%]	46.4 %	46.8 %	body-style=hardtop	-0.4	[12.5 %]	2.7 %	3.9 %
body-style=wagon	-0.3	[80.0%]	11.9 %	12.2 %	body-style=convertible	-1.2	[0.0%]	0.0 %	2.9 %

FUEL-TYPE 90.2% of cars uses GAS, they are higher horsepower than the other cars (106.4 hp versus 104.2 hp for the whole dataset); they have higher consumption (24.67 mpg versus 25.22 mpg).

The TEST VALUE column shows the strength of the difference. The higher is the absolute value of this indicator, the higher is the difference between the mean computed in the subgroup and the mean computed on the whole dataset.

About the DISESEL cars, we see that they have lower consumption (30.30 mpg) and horsepower (84.45 hp) than the other cars. We see also that there is a significant presence of SEDAN (body-style) cars in this group: there are 46.8% in the whole dataset; there are 75% in

this subgroup [P(SEDAN / DIESEL) = 0.75, we can interpret this proportion as a **precision**, see the following cross-tabulation].

NB fuel-type	fuel-type		
body-style	diesel	gas	Total
convertible	0.00%	3.24%	2.93%
hardtop	5.00%	3.78%	3.90%
hatchback	5.00%	37.30%	34.15%
sedan	75.00%	43.78%	46.83%
wagon	15.00%	11.89%	12.20%
Total	100.00%	100.00%	100.00%

In another way, if the DIESEL represents 9.8% of the cars, we have 15.6% of SEDAN in this group: [P(DIESEL/SEDAN) = 0.156, we can interpret this value as a **recall**, see the following cross-tabulation].

NB fuel-type	fuel-type		
body-style	diesel	gas	Total
convertible	0.00%	100.00%	100.00%
hardtop	12.50%	87.50%	100.00%
hatchback	1.43%	98.57%	100.00%
sedan	15.63%	84.38%	100.00%
wagon	12.00%	88.00%	100.00%
Total	9.76%	90.24%	100.00%

ASPIRATION TURBO represents 18% of the dataset. They have higher horsepower (124.43 hp), price (16172 \$) and consumption (22.41 mpg).

The main weakness of the component is that we cannot create subgroups with two or more attributes; we cannot see the effect of the interaction of various values of attributes. **GROUP EPLORATION** allows us to **manually build groups with the combination of two or more attributes** and see the specificities of subgroups with comparative descriptive statistics.

Build groups with two or more attributes

We want to explore the specificities of TURBO-DIESEL cars. We add the GROUP EXPLORATION component in the diagram.



We see three areas in the visualization window.



The area [A – TARGET ATTRIBUTES] shows the target attributes that allow us to build groups; the area [B – GROUP DEFINITION] shows in a tree the subgroups; the area [C – SUBGROUP DESCRIPTION] shows the comparative descriptive statistics of the selected group in the tree.

We have 205 examples (the whole dataset) in the root of [B]. To build groups, we use "dragand-drop" from the target attributes to the node we want to explore in the tree.

Target attributes	Description of the subgroup on target attributes
(2) fuel-type	Group exploration 1
	Parameters
	Results
	Root node : 205 examples, 2 target attributes
Group definition	Computation time : 0 ms.
😑 Root (205 examples)	Created at 12/02/2006 18:52:12
fuel-type = gas (185 ex.)	
fuel-type = diesel (20 ex.)	

We find again the previous results: 90.2% (185) cars use GAS, 9.8% (20) use DIESEL. If we select the DIESEL node, we obtain the same descriptive statistics.

Target attributes	Description of	of the subgi	roup on target atti	ributes	
(2) fuel-type (2) aspiration	Rule : fuel-type = diesel	Rule : fuel-type = diesel			
	Subgroup = Local				
	Examples			[9.8 %] 20	
	Att - Desc	Test value	Group	Overral	
	Continuous attributes				
Group definition	city-mpg	3.6	30.30	25.22	
🖃 Root (205 examples)	price	1.6	15838.15	13207.13	
fuel-type = gas (185 ex.)	horsepower	-2.4	84.45	104.26	
idel-type = diesel (20 ex.)	Discrete attributes				
	body-style=sedan	2.7	[15.6 %] 75.0 %	46.8 %	
	body-style=wagon	0.4	[12.0 %] 15.0 %	12.2 %	
	body-style=hardtop	0.3	[12.5 %] 5.0 %	3.9 %	
	body-style=convertible	-0.8	[0.0%] 0.0%	2.9 %	
	body-style=hatchback	-2.9	[1.4%] 5.0%	34.1 %	
					-

If we want to explore a new subgroup e.g. TURBO-DIESEL subgroup, we must add the ASPIRATION attribute on the FUEL-TYPE = DIESEL node of the tree.

Target attributes	Description	of the subgi	roup on tar;	get atti	ributes	
(2) fuel-type (2) aspiration	Rule : fuel-type = diesel					^
	S	ubgroup =	Local			
	Examples			I	[9.8%]20	
	Att - Desc	Test value	Group		Overral	
	Continuous attributes					
Group definition	city - npg	3.6		30.30	25.22	
Root (205 examples)	Drice	1.6	15	838.15	13207.13	
fuel-type = gas (185 ex.)	horsepower	-2.4		84.45	104.26	
aspiration = std (7 ex.)	Discrete attributes					
aspiration = turbo (13 ex.)	body-style=sedan	2.7	[15.6 %]	75.0 %	46.8 %	
	body-style=wagon	0.4	[12.0 %]	15.0 %	12.2 %	
	body-style=hardtop	0.3	[12.5 %]	5.0 %	3.9 %	
	body-style=convertible	-0.8	[0.0%]	0.0 %	2.9 %	
	body-style=hatchback	-2.9	[1.4 %]	5.0 %	34.1 %	

We click on the new node to see the characteristics of the group.

Target attributes	Description	of the subgr	roup on target attr	ributes		
(2) fuel-type (2) aspiration	Rule : fuel-type = diesel && aspiration = turbo					
	s	ubgroup =	Local			
	Examples			[6.3%] 13		
	Att - Desc	Test value	Group	Overral		
	Continuous attributes					
Group definition	price	2.8	19159.15	13207.13		
Root (205 examples)	city-mpg	0.9	26.77	25.22		
fuel-type = gas (185 ex.)	horsepower	-0.5	98.62	104.26		
aspiration = std (7 ex.)	Discrete attributes	screte attributes				
aspiration = turbo (13 ex.)	body-style=sedan	1.7	[9.4%] 69.2%	46.8 %		
	body-style=wagon	1.2	[12.0 %] 23.1 %	12.2 %		
	body-style=hardtop	0.7	[12.5 %] 7.7 %	3.9 %		
	body-style=convertible	-0.6	[0.0%] 0.0%	2.9 %		
	body-style=hatchback	-2.7	[0.0%] 0.0%	34.1 %		

There are 13 (6.3%) TURBO-DIESEL cars.

Of course, we can add more than two attributes in the tree; we can also remove uninteresting nodes. The only limitation of this component is that the target attributes must be discrete.