Subject

The main bottleneck of data mining software programming is the data management. In using a spreadsheet, which has many data manipulation functionalities, one can give priority to the core data mining methods development. XL-SIPINA is an attempt to embed the EXCEL© spreadsheet in a decision tree software.

Some commercial tools use this framework. They put forward some add-ons that are showed as new menus in EXCEL. XL-SIPINA relies on a different technology. It is stand-alone software; the spreadsheet is embedded in the main window with the OLE technology. The whole dataset is loaded into the main memory before the models computation. **Of course, XL-SIPINA cannot be executed if EXCEL is not installed on your computer.**

XL-SIPINA is an old test project. In the future, I think I plan to work on a free project such as GNUMERIC.

In this tutorial, we show the utilization of XL-SIPINA. **Last, if the embedding of the spreadsheet in the SIPINA is specific to XL-SIPINA version, the functionalities about interactive exploration of the tree, which are displayed in this tutorial, are also available in the stand-alone version (Research Version) of SIPINA.**

Using XL-SIPINA

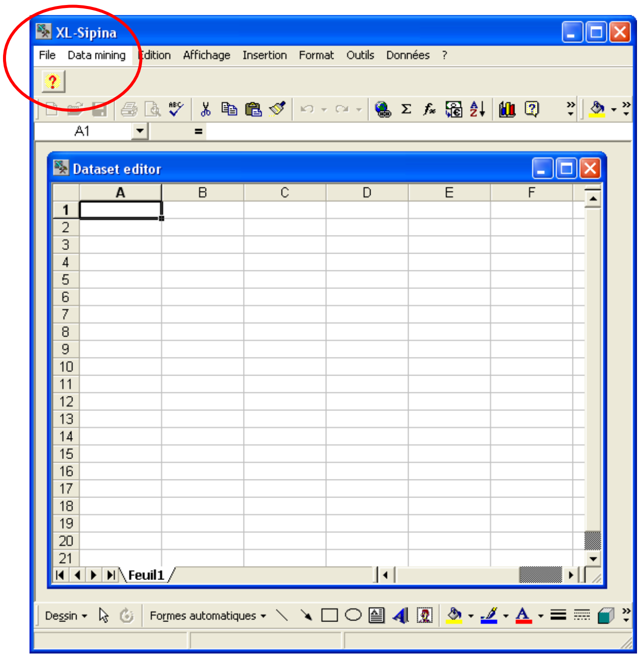
XL-SIPINA is mainly an exploratory tool. There are no functionalities about model assessment, prediction on a new dataset, etc. Because of the limitation of EXCEL (number of rows and columns), we cannot execute the software on a large dataset. In this situation, I recommend to use the SIPINA research version.

Dataset

We study the AUTOMOBILE DATABASE. We plan to explain the expert SYMBOLING annotation (a car is risky or not) from their characteristics (consumption, horsepower, etc.).

Starting XL-SIPINA

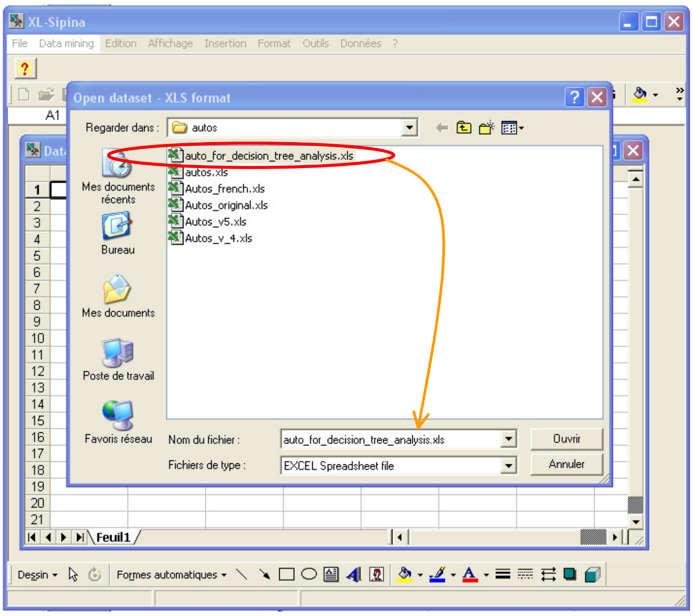
The main window of the software is the following. **If the EXCEL OLE server is not available, the software execution is interrupted**.



The two first menus belong to the software, the others belongs to EXCEL. A help button in the toolbar enables to show some help about the software handling.

Downloading the dataset

We click on the FILE/OPEN menu in order to download the dataset in the spreadsheet. We select the file in the "dataset" directory.



Nom du fichier : auto_for_decision_tree_analysis.xls

Fichiers de type : EXCEL Spreadsheet file

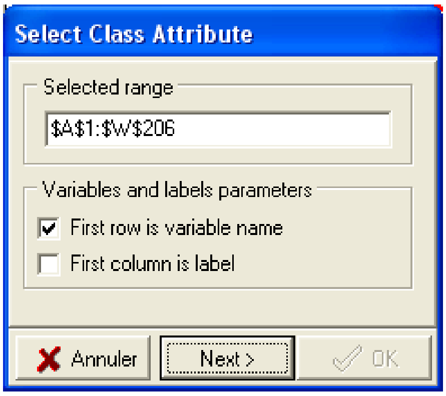
The dataset has 205 examples. RISKY is the class attribute. There are 22 descriptors, 8 of them are discrete. Our goal is to explain why the experts classify some cars as "positive".

Executing the decision trees algorithm

Before executing the learning algorithm, we must **select the dataset** we plan to analyze in the sheet. The columns must be adjacent; we cannot make a multiple selection.

Please, make sure that the dataset is well prepared. The dataset checking is not very powerful in the XL-SIPINA. For instance, it **cannot handle missing data**.

Then, we start the analysis with **DATA MINING / START LEARNING** menu. A dialog box enables to define/check the selected dataset (the selected cells in the sheet).



Select Class Attribute

Selected range

$A$1:$W$206

Variables and labels parameters

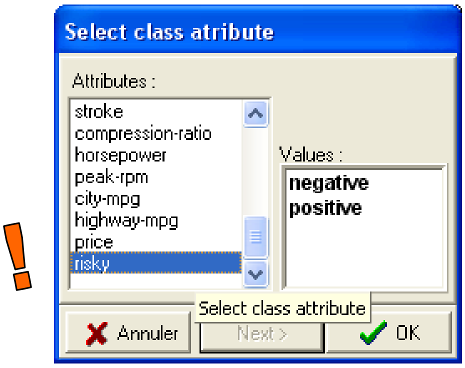
☑ First row is variable name

☐ First column is label

Two additional options are available. They enable to specify if the first row is the name of the attributes, and if the first column is the identification of the examples.

We select the **NEXT** button. **The dataset is parsed**. The software automatically detects if the attribute is continuous or discrete. The detection rule is based on the first row of the data. If the value can be transformed in a numeric value, the column is considered as a continuous attribute; it is turned into a discrete attribute in the other case. It is a difficult operation, some errors can occur during the processing. XL-SIPINA cannot handle missing data. We must treat them before the data analysis.

In the next step, we must **define the discrete class attribute**. In our tutorial, we select the RISKY attribute. Then we click on the OK button.



Select class atribute

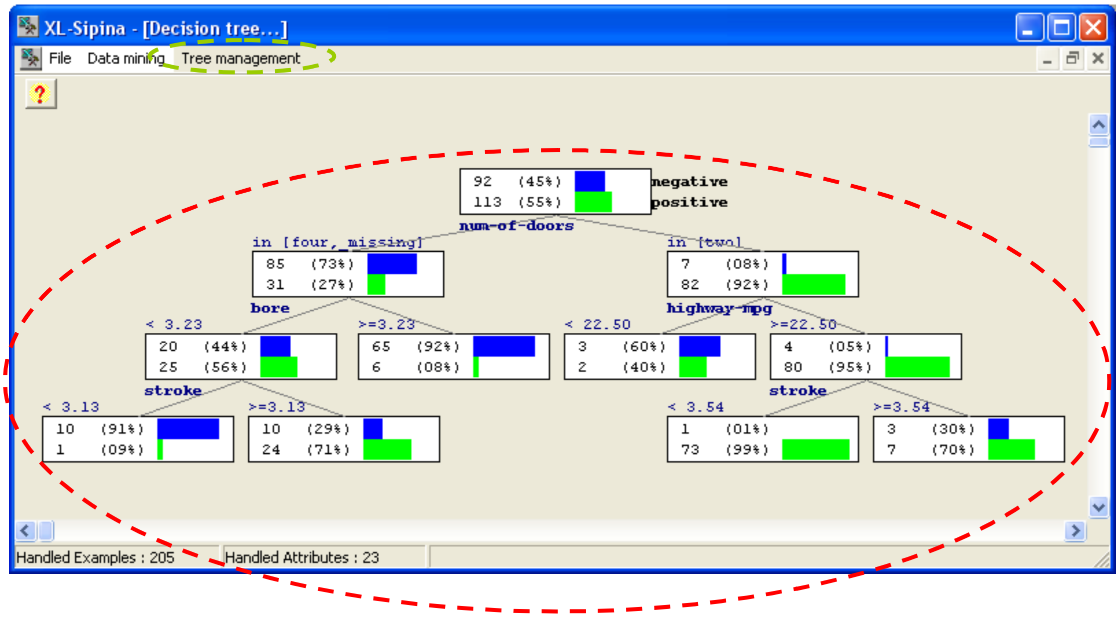
Attributes :

stroke
compression-ratio
horsepower
peak-rpm
city-mpg
highway-mpg
price
risky

Values :

negative
positive

The decision tree is built. It is displayed in a new window. EXCEL menus are hidden, and a new menu appears "TREE MANAGEMENT".



The implemented algorithm is a variant of the CHAID method. The main difference is the utilization of the Tschuprow goodness of split criterion. The depth of the tree is limited to 4 in the default settings. We can modify that. We can also continue manually the construction of the tree.

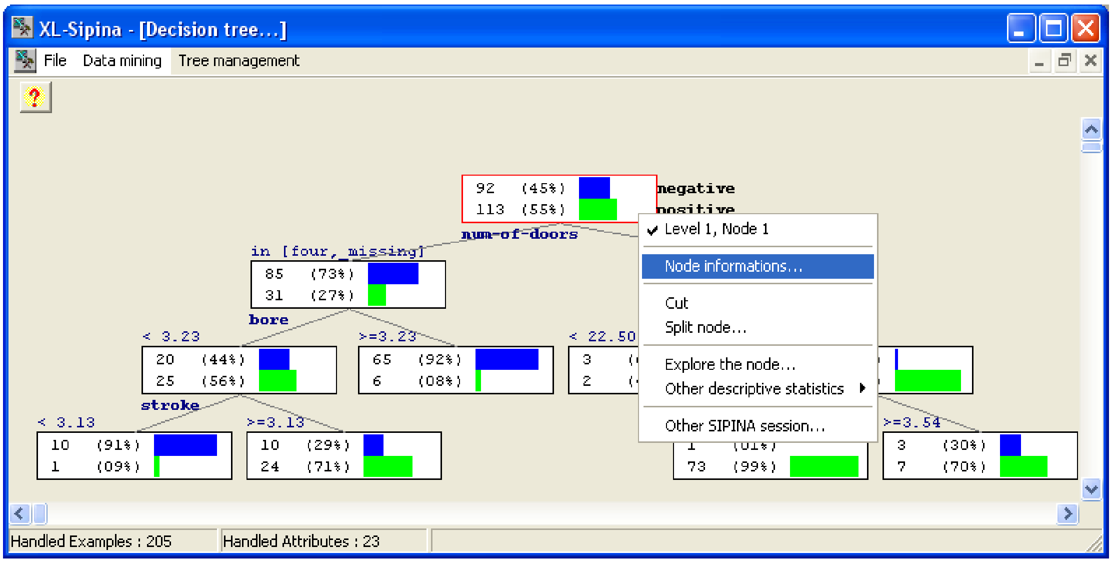
Exploring the tree

One of the main advantages of the decision trees is the possibility to interactively explore and modify the solutions suggested by the automatic induction. This is a convenient way to insert domain knowledge in the model.

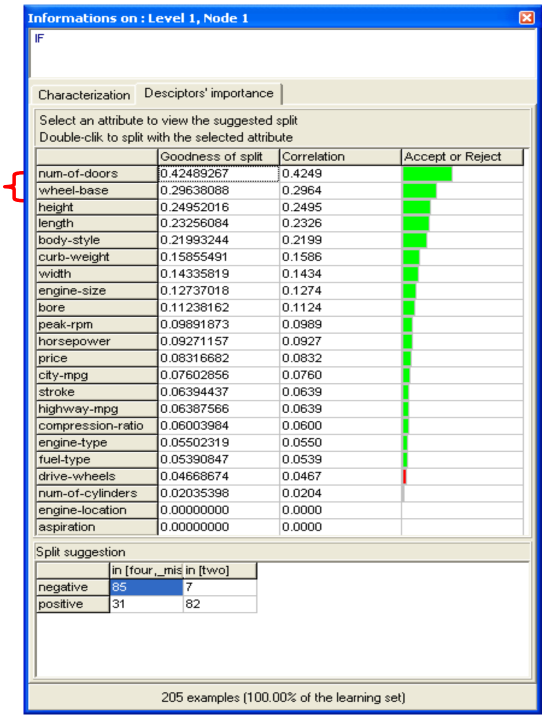
Seeing the other splits

The first split attribute in the root node is NUM-OF-DOORS. But we wonder: what is the goodness of the competitors, are they really bad in the discrimination?

For checking that, we select the root of the tree. The box is colored in red. Then we use right-click in order to activate the popup menu. We select the NODE INFORMATION menu.



A window appears. In the descriptors' importance tab, we see the goodness of split for each predictive attribute. The index for NUM-OF-DOORS is 0.42489; it is 0.29638 for the next attribute (WHEEL-BASE). Goodness of split and correlation are identical indexes here. The ACCEPT/REJECT column is red if one of the stopping rules is activated.



Select an attribute to view the suggested split
Double-clik to split with the selected attribute

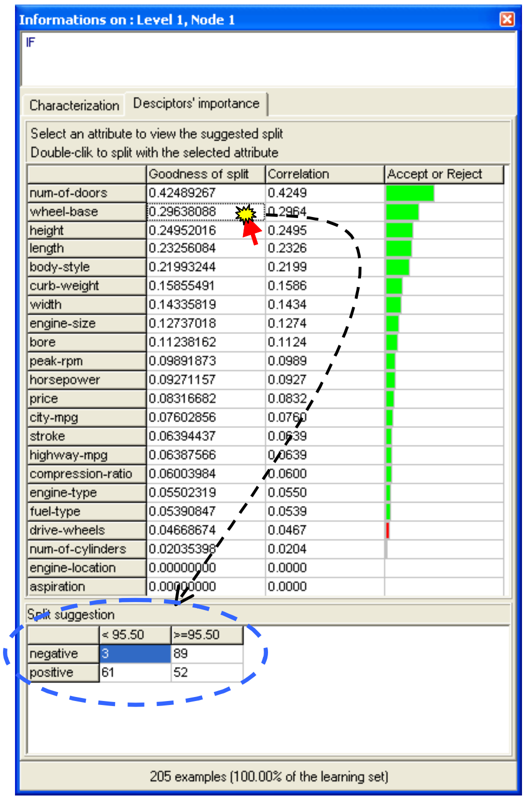
	Goodness of split	Correlation	Accept or Reject
num-of-doors	0.42489267	0.4249	🟩
wheel-base	0.29638088	0.2964	🟩
height	0.24952016	0.2495	🟩
length	0.23256084	0.2326	🟩
body-style	0.21993244	0.2199	🟩
curb-weight	0.15855491	0.1586	🟩
width	0.14335819	0.1434	🟩
engine-size	0.12737018	0.1274	🟩
bore	0.11238162	0.1124	🟩
peak-rpm	0.09891873	0.0989	🟩
horsepower	0.09271157	0.0927	🟩
price	0.08316682	0.0832	🟩
city-mpg	0.07602856	0.0760	🟩
stroke	0.06394437	0.0639	🟩
highway-mpg	0.06387566	0.0639	🟩
compression-ratio	0.06003984	0.0600	🟩
engine-type	0.05502319	0.0550	🟩
fuel-type	0.05390847	0.0539	🟩
drive-wheels	0.04668674	0.0467	🟥
num-of-cylinders	0.02035398	0.0204	
engine-location	0.00000000	0.0000	
aspiration	0.00000000	0.0000	

Split suggestion

	in [four,_mis	in [two]
negative	85	7
positive	31	82

205 examples (100.00% of the learning set)

When we select an attribute in the grid, the suggested split is showed in the bottom part of the window. For WHEELBASE, we see that the computed discretization cut point is 95.5.



Select an attribute to view the suggested split
Double-clik to split with the selected attribute

	Goodness of split	Correlation	Accept or Reject
num-of-doors	0.42489267	0.4249	🟩
wheel-base	0.29638088	0.2964	🟩
height	0.24952016	0.2495	🟩
length	0.23256084	0.2326	🟩
body-style	0.21993244	0.2199	🟩
curb-weight	0.15855491	0.1586	🟩
width	0.14335819	0.1434	🟩
engine-size	0.12737018	0.1274	🟩
bore	0.11238162	0.1124	🟩
peak-rpm	0.09891873	0.0989	🟩
horsepower	0.09271157	0.0927	🟩
price	0.08316682	0.0832	🟩
city-mpg	0.07602856	0.0760	🟩
stroke	0.06394437	0.0639	🟩
highway-mpg	0.06387566	0.0639	🟩
compression-ratio	0.06003984	0.0600	🟩
engine-type	0.05502319	0.0550	🟩
fuel-type	0.05390847	0.0539	🟩
drive-wheels	0.04668674	0.0467	🟥
num-of-cylinders	0.02035398	0.0204	
engine-location	0.00000000	0.0000	
aspiration	0.00000000	0.0000	

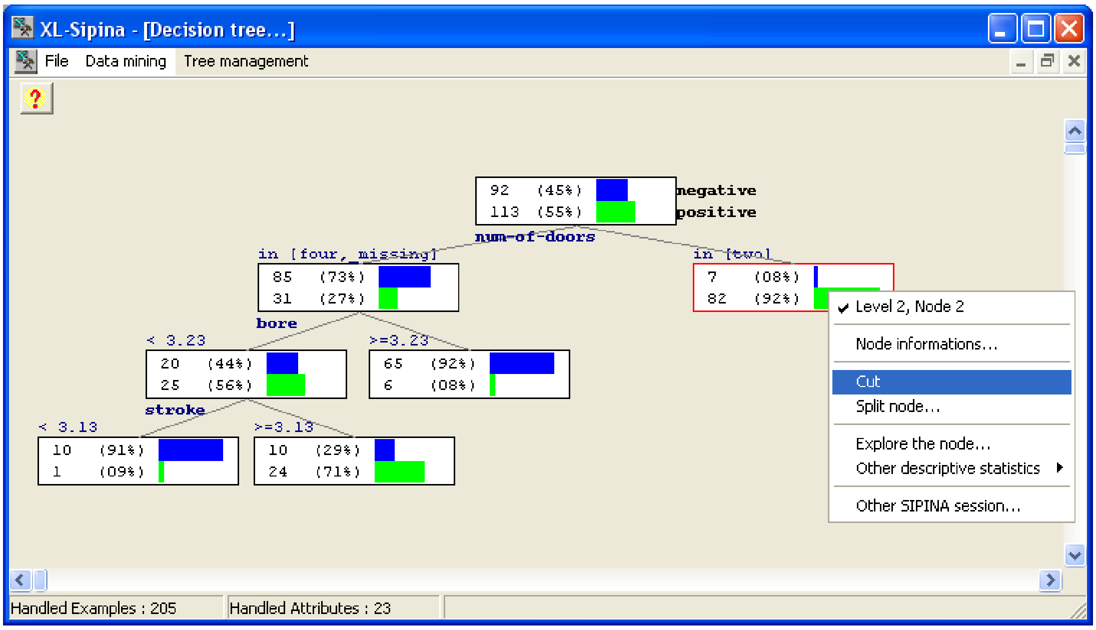
Split suggestion

	< 95.50	>=95.50
negative	3	89
positive	61	52

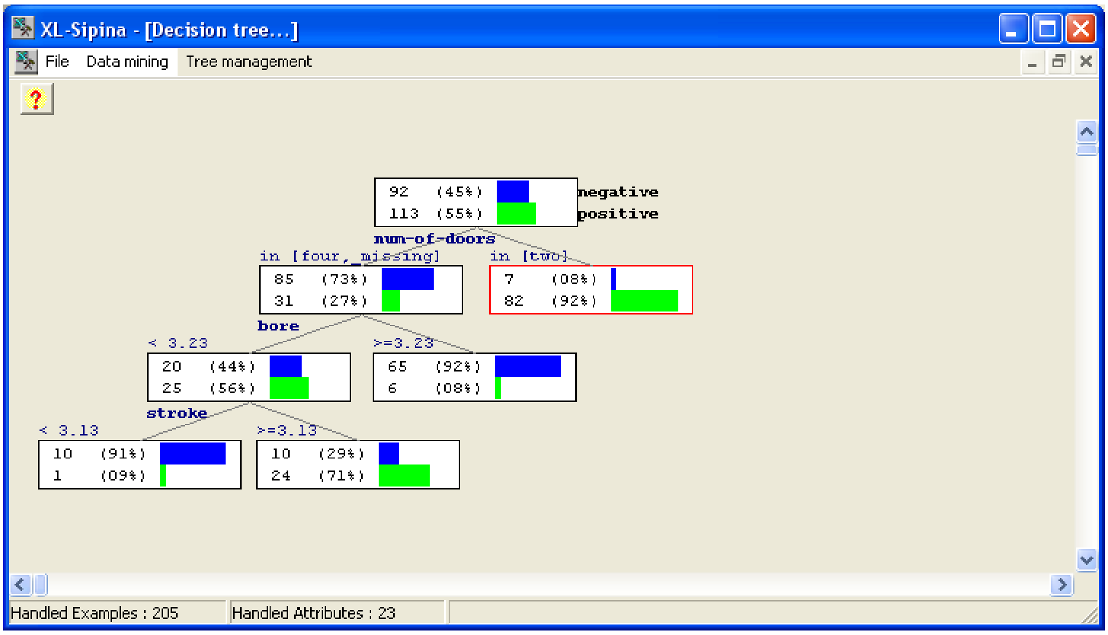
205 examples (100.00% of the learning set)

Pruning manually the tree

In order to prune the sub tree under the second node of the second level of the tree, we activate again the popup menu and we select the CUT option.



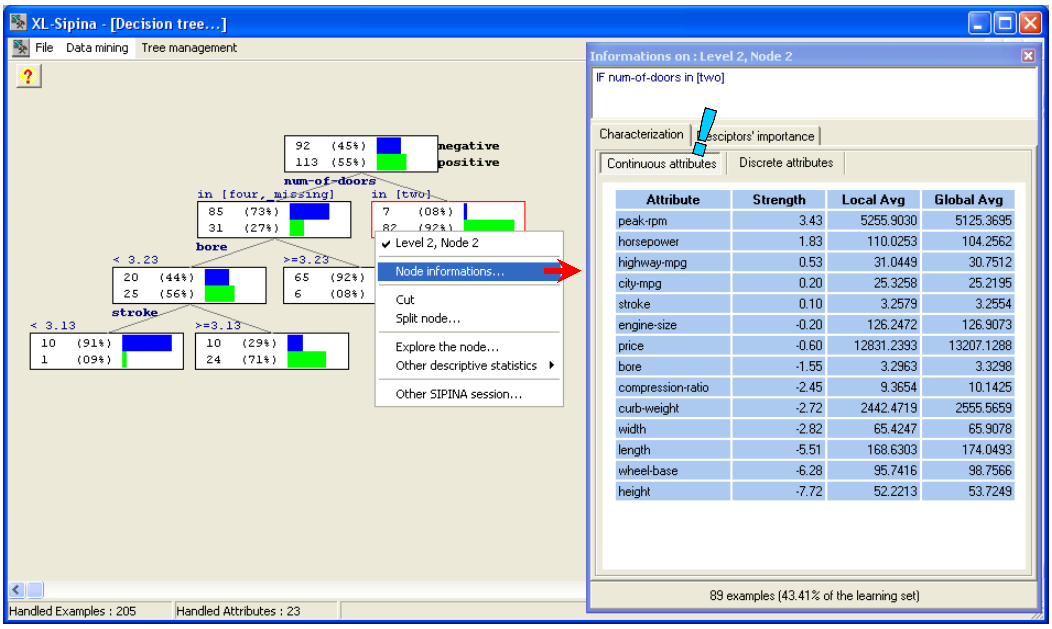
The visual organization of the tree can be refreshed with the F5 shortcut.



Exploration of a node (exploration of a subsample of the dataset)

The selected node is described with the rule NUM-OF-DOORS = TWO. But perhaps other attributes enable also to characterize this subpopulation. Their importance is hidden in the tree description.

We have three tools in the popup menu. First, **EXPLORE THE NODE** shows the local dataset. Second, **OTHER DESCRIPTIVE STATISTICS** enables to compute various descriptive statistics on the local examples. Third, the CHARACTERIZATION tab of the **NODE EXPLORATION** window enables to quickly compare the statistics on the local node (local examples) and the root node (all examples).



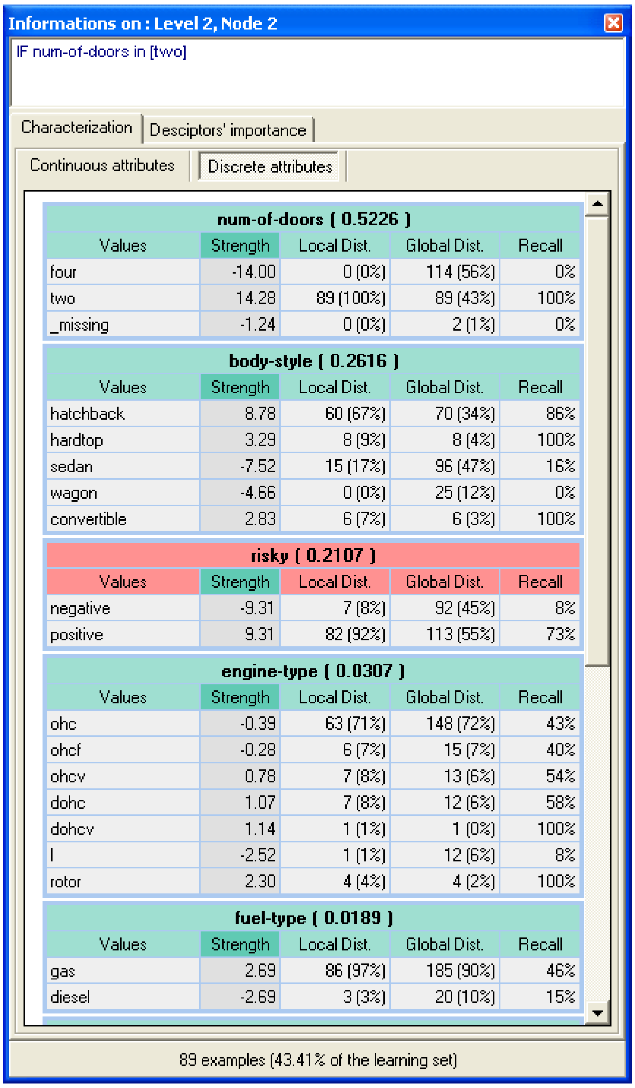
IF num-of-doors in [two]

Attribute	Strength	Local Avg	Global Avg
peak-rpm	3.43	5255.9030	5125.3695
horsepower	1.83	110.0253	104.2562
highway-mpg	0.53	31.0449	30.7512
city-mpg	0.20	25.3258	25.2195
stroke	0.10	3.2579	3.2554
engine-size	-0.20	126.2472	126.9073
price	-0.60	12831.2393	13207.1288
bore	-1.55	3.2963	3.3298
compression-ratio	-2.45	9.3654	10.1425
curb-weight	-2.72	2442.4719	2555.5659
width	-2.82	65.4247	65.9078
length	-5.51	168.6303	174.0493
wheel-base	-6.28	95.7416	98.7566
height	-7.72	52.2213	53.7249

89 examples (43.41% of the learning set)

According to the data type, two tabs are available. We can compare the values of continuous or discrete attributes. The "strength" index states the importance of the difference between the value of the statistic on all the examples and the examples corresponding to the selected node. The used statistic is the average for continuous attribute, and the proportion for the discrete one.

In our tutorial, we see that the cars with NUM-OF-DOORS=TWO have also, in average, high PEAK-RPM and HORSEPOWER. They have weak HEIGHT and WHEELBASE.



Informations on : Level 2, Node 2

IF num-of-doors in [two]

num-of-doors (0.5226)

Values	Strength	Local Dist.	Global Dist.	Recall
four	-14.00	0 (0%)	114 (56%)	0%
two	14.28	89 (100%)	89 (43%)	100%
_missing	-1.24	0 (0%)	2 (1%)	0%

body-style (0.2616)

Values	Strength	Local Dist.	Global Dist.	Recall
hatchback	8.78	60 (67%)	70 (34%)	86%
hardtop	3.29	8 (9%)	8 (4%)	100%
sedan	-7.52	15 (17%)	96 (47%)	16%
wagon	-4.66	0 (0%)	25 (12%)	0%
convertible	2.83	6 (7%)	6 (3%)	100%

risky (0.2107)

Values	Strength	Local Dist.	Global Dist.	Recall
negative	-9.31	7 (8%)	92 (45%)	8%
positive	9.31	82 (92%)	113 (55%)	73%

engine-type (0.0307)

Values	Strength	Local Dist.	Global Dist.	Recall
ohc	-0.39	63 (71%)	148 (72%)	43%
ohcf	-0.28	6 (7%)	15 (7%)	40%
ohcv	0.78	7 (8%)	13 (6%)	54%
dohc	1.07	7 (8%)	12 (6%)	58%
dohcv	1.14	1 (1%)	1 (0%)	100%
l	-2.52	1 (1%)	12 (6%)	8%
rotor	2.30	4 (4%)	4 (2%)	100%

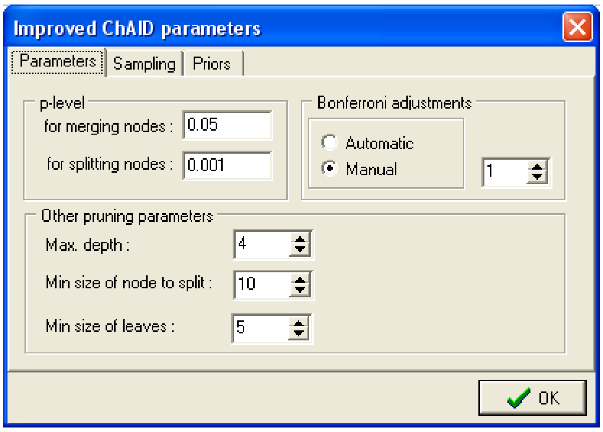
fuel-type (0.0189)

Values	Strength	Local Dist.	Global Dist.	Recall
gas	2.69	86 (97%)	185 (90%)	46%
diesel	-2.69	3 (3%)	20 (10%)	15%

89 examples (43.41% of the learning set)

The cars with HATCHBACK and HARDTOP-CONVERTIBLE style are more frequent in this group. At this step of the analysis, the expert domain can give more explanation about these phenomena, test the difference on other attributes, try another splits, etc. The cooperation between the statistician and the expert domain becomes very important starting from this step.

Modifying the parameter settings



Improved ChAID parameters

p-level

for merging nodes : 0.05

for splitting nodes : 0.001

Bonferroni adjustments

○ Automatic

● Manual 1

Other pruning parameters

Max. depth : 4

Min size of node to split : 10

Min size of leaves : 5

In order to modify the parameters of the decision tree induction algorithm, we must first stop the current analysis (DATA MINING / STOP LEARNING). Then we click on the DATA MINING / PARAMETERS menu. The following dialog box appears.

The major parameters are in the PARAMETERS tab. They mainly enable to control the size of the tree.

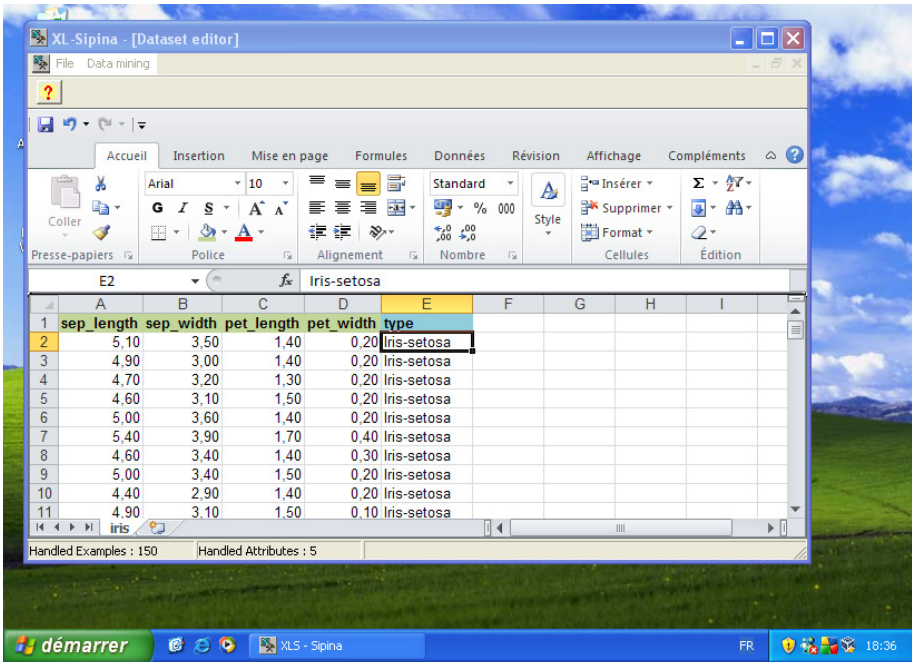
Conclusion

XL-SIPINA was essentially an attempt in order to evaluate the difficulty of the embedding a spreadsheet in data mining software. The solution works fine as we can see in this tutorial. But, this forces us to compile a specific version of SIPINA.

Epilogue -- 2010/08/29

Because I have hindsight now (2010/08/29), **I realize that the "add-in" solution which enables to incorporate SIPINA into Excel is ultimately the most viable**. It is simple, reliable (the two go together often), and very powerful (a few seconds are enough to transfer a database with 100,000 observations and 22 variables). Therefore, the solution I use myself when I treat the data files.

Another aspect has excited my curiosity recently. I wanted to know how XL-SIPINA reacted faced to the latest versions of Excel (Office 2007 and Office 2010). I have noted that the tool works fine. All induction trees features are active. We were able to reproduce all the operations described in this tutorial. But the interface is a little modified. The Office Ribbon is incorporated in the software when it is started. Here is a screenshot for Excel 2010.



E2 Iris-setosa

	A	B	C	D	E	F	G	H	I
1	sep_length	sep_width	pet_length	pet_width	type				
2	5,10	3,50	1,40	0,20	Iris-setosa				
3	4,90	3,00	1,40	0,20	Iris-setosa				
4	4,70	3,20	1,30	0,20	Iris-setosa				
5	4,60	3,10	1,50	0,20	Iris-setosa				
6	5,00	3,60	1,40	0,20	Iris-setosa				
7	5,40	3,90	1,70	0,40	Iris-setosa				
8	4,60	3,40	1,40	0,30	Iris-setosa				
9	5,00	3,40	1,50	0,20	Iris-setosa				
10	4,40	2,90	1,40	0,20	Iris-setosa				
11	4,90	3,10	1,50	0,10	Iris-setosa				

Handled Examples : 150 Handled Attributes : 5