Subject

Interactive induction of decision trees with SIPINA.

Various functionalities of SIPINA are not documented. In this tutorial, we show how to explore nodes of a decision tree, in order to obtain a better understanding of the characteristics of the subpopulation on a node. This is an important task, for instance when we want to validate the rules with an expert domain.

Dataset

We use the BLOOD_PRESSURE_LEVELS.XLS dataset, available on-line[1]. There are 399 examples. We want to predict the high blood pressure of a patient from their characteristics (smoke, sex, overweight, etc.). Here is the description of the attributes.

Attribute	Category	Informations
bpress_level	Discrete	2 values
gender	Discrete	2 values
smoke	Discrete	2 values
exercise	Continue	-
overweight	Continue	-
alcohol	Continue	-
stress	Continue	-
salt	Continue	-
income	Continue	-
education	Continue	-

We adopt a descriptive framework. We do not want to obtain the most accurate classifier, but rather to characterize the (groups of) individuals. In this point of view, the assessment relies mainly on the expertise. This is the physician who may confirm us if rules proposed by the tree are in adequacy with the domain knowledge or not. For this reason, we use all the dataset for the tree induction. The accuracy rate (or error rate) is not really a pertinent rating in this context.

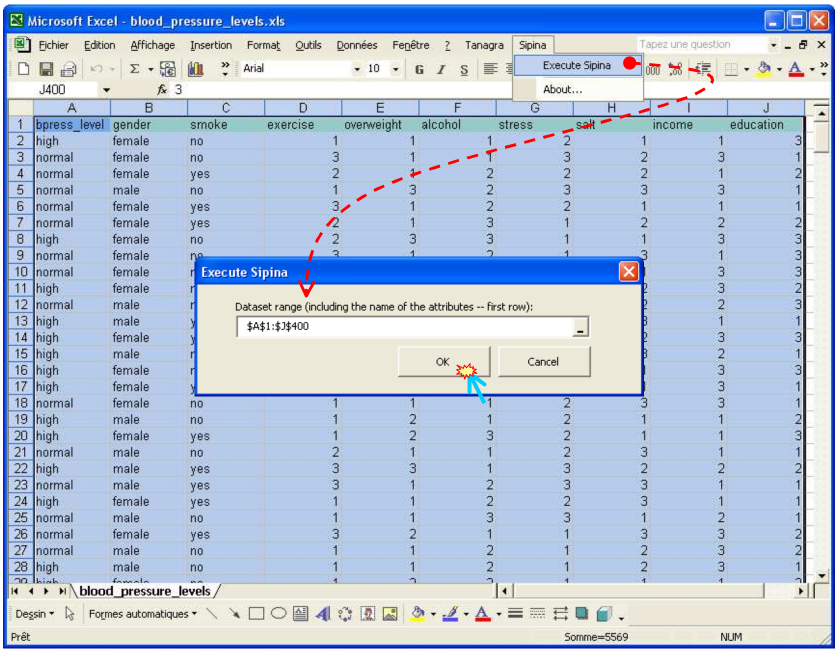
Induction of a decision tree using SIPINA

Data importation

The SIPINA add-in (SIPINA.XLA) is automatically installed on the hard disk. The simplest way to send the dataset from EXCEL to SIPINA is to use this add-in (see « SIPINA ADD-IN FOR EXCEL SPREADSHEET » on the website http://eric.univ-lyon2.fr/~ricco/sipina_download.html).

We select the range of cells, we activate the SIPINA / EXECUTE menu which is now available in EXCEL. A dialog box appears, we check if the selection is right, then we confirm by clicking on OK. Note, the first row of the cells range must correspond to the name of the attributes.

[1] http://eric.univ-lyon2.fr/~ricco/dataset/blood_pressure_levels.xls ; the original source of this dataset is http://www.math.yorku.ca/Who/Faculty/Ng/ssc2003/BPMain.htm


Microsoft Excel - blood_pressure_levels.xls
Fichier Edition Affichage Insertion Format Outils Données Fenêtre ? Tanagra Sipina
Execute Sipina
About...
Tapez une question
Arial
10
J400
3
blood_pressure_levels
Dessin
Formes automatiques
Prêt
Somme=5569
NUM

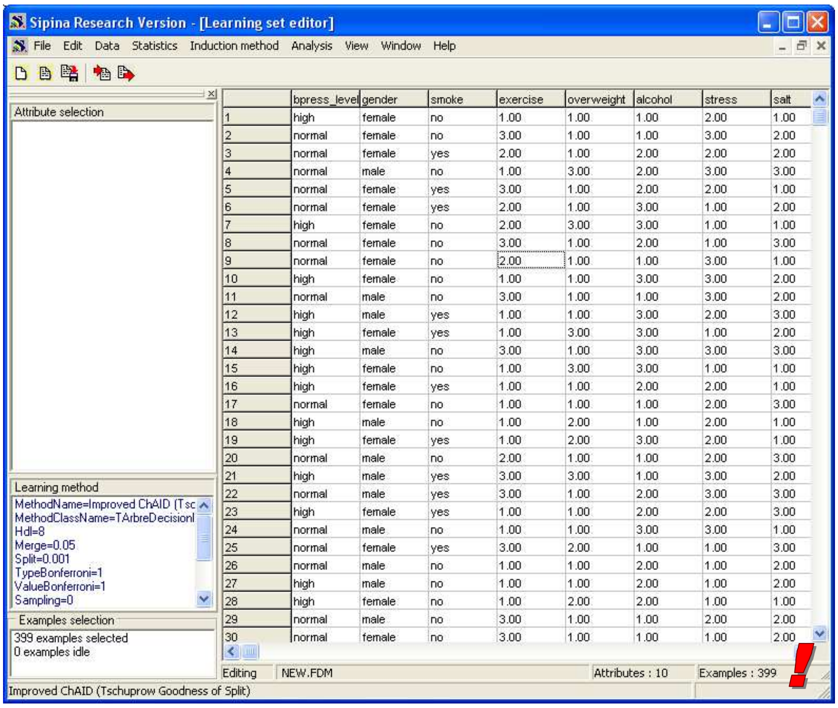
	A	B	C	D	E	F	G	H	I	J
1	bpress_level	gender	smoke	exercise	overweight	alcohol	stress	salt	income	education
2	high	female	no	1	1	1	2	1	1	3
3	normal	female	no	3	1	1	3	2	3	1
4	normal	female	yes	2	1	2	2	2	1	2
5	normal	male	no	1	3	2	3	3	3	1
6	normal	female	yes	3	1	2	2	1	1	1
7	normal	female	yes	2	1	3	1	2	2	2
8	high	female	no	2	3	3	1	1	3	3
9	normal	female	no	3	1	2	1	3	1	3
10	normal	female							3	3
11	high	female							3	2
12	normal	male							2	3
13	high	male							1	1
14	high	female							3	3
15	high	male							2	1
16	high	female							3	3
17	high	female							3	1
18	normal	female	no	1	1	1	2	3	3	1
19	high	male	no	1	2	1	2	1	1	2
20	high	female	yes	1	2	3	2	1	1	3
21	normal	male	no	2	1	1	2	3	1	1
22	high	male	yes	3	3	1	3	2	2	2
23	normal	male	yes	3	1	2	3	3	1	1
24	high	female	yes	1	1	2	2	3	1	1
25	normal	male	no	1	1	3	3	1	2	1
26	normal	female	yes	3	2	1	1	3	3	2
27	normal	male	no	1	1	2	1	2	3	2
28	high	male	no	1	1	2	1	2	3	1

Execute Sipina

Dataset range (including the name of the attributes -- first row): $A$1:$J$400

OK Cancel

SIPINA is started automatically. The data are transferred through the clipboard. The number of variables and the number of observations are displayed in the status bar (10 variables and 399 observations). Note: SIPINA does not handle missing data, we must treat them before.


Sipina Research Version - [Learning set editor]
File Edit Data Statistics Induction method Analysis View Window Help
Attribute selection
Learning method
Examples selection
Editing
NEW.FDM

	bpress_level	gender	smoke	exercise	overweight	alcohol	stress	salt
1	high	female	no	1.00	1.00	1.00	2.00	1.00
2	normal	female	no	3.00	1.00	1.00	3.00	2.00
3	normal	female	yes	2.00	1.00	2.00	2.00	2.00
4	normal	male	no	1.00	3.00	2.00	3.00	3.00
5	normal	female	yes	3.00	1.00	2.00	2.00	1.00
6	normal	female	yes	2.00	1.00	3.00	1.00	2.00
7	high	female	no	2.00	3.00	3.00	1.00	1.00
8	normal	female	no	3.00	1.00	2.00	1.00	3.00
9	normal	female	no	2.00	1.00	1.00	3.00	1.00
10	high	female	no	1.00	1.00	3.00	3.00	2.00
11	normal	male	no	3.00	1.00	1.00	3.00	2.00
12	high	male	yes	1.00	1.00	3.00	2.00	3.00
13	high	female	yes	1.00	3.00	3.00	1.00	2.00
14	high	male	no	3.00	1.00	3.00	3.00	3.00
15	high	female	no	1.00	3.00	3.00	1.00	1.00
16	high	female	yes	1.00	1.00	2.00	2.00	1.00
17	normal	female	no	1.00	1.00	1.00	2.00	3.00
18	high	male	no	1.00	2.00	1.00	2.00	1.00
19	high	female	yes	1.00	2.00	3.00	2.00	1.00
20	normal	male	no	2.00	1.00	1.00	2.00	3.00
21	high	male	yes	3.00	3.00	1.00	3.00	2.00
22	normal	male	yes	3.00	1.00	2.00	3.00	3.00
23	high	female	yes	1.00	1.00	2.00	2.00	3.00
24	normal	male	no	1.00	1.00	3.00	3.00	1.00
25	normal	female	yes	3.00	2.00	1.00	1.00	3.00
26	normal	male	no	1.00	1.00	2.00	1.00	2.00
27	high	male	no	1.00	1.00	2.00	1.00	2.00
28	high	female	no	1.00	2.00	2.00	1.00	1.00
29	normal	male	no	3.00	1.00	1.00	2.00	2.00
30	normal	female	no	3.00	1.00	1.00	1.00	2.00

MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

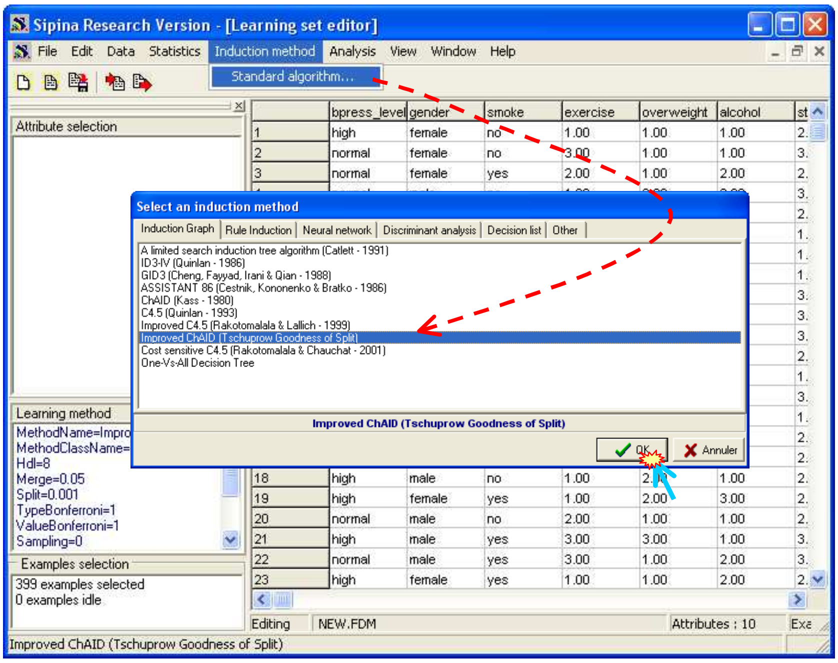
399 examples selected
0 examples idle

Attributes : 10
Examples : 399

Improved ChAID (Tschuprow Goodness of Split)

Choosing the learning method

The first step is to select the induction algorithm. We activate the INDUCTION METHOD / STANDARD ALGORITHM menu. A dialog box appears with the list of available methods.



A limited search induction tree algorithm (Catlett - 1991)
ID3-IV (Quinlan - 1986)
GID3 (Cheng, Fayyad, Irani & Qian - 1988)
ASSISTANT 86 (Cestnik, Kononenko & Bratko - 1986)
ChAID (Kass - 1980)
C4.5 (Quinlan - 1993)
Improved C4.5 (Rakotomalala & Lallich - 1999)
Improved ChAID (Tschuprow Goodness of Split)
Cost sensitive C4.5 (Rakotomalala & Chauchat - 2001)
One-Vs-All Decision Tree

Improved ChAID (Tschuprow Goodness of Split)

MethodName=Impro
MethodClassName=
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

399 examples selected
0 examples idle

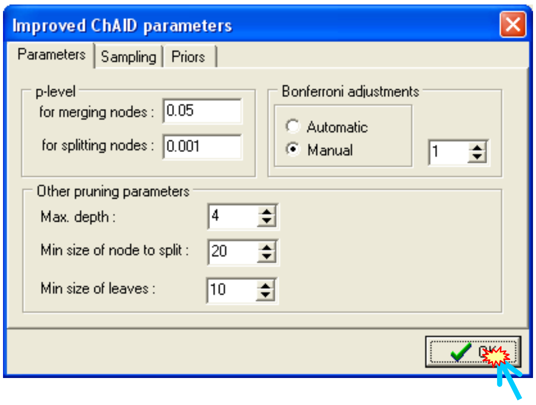
	bpress_level	gender	smoke	exercise	overweight	alcohol	st
1	high	female	no	1.00	1.00	1.00	2.
2	normal	female	no	3.00	1.00	1.00	3.
3	normal	female	yes	2.00	1.00	2.00	2.
18	high	male	no	1.00	2.00	1.00	2.
19	high	female	yes	1.00	2.00	3.00	2.
20	normal	male	no	2.00	1.00	1.00	2.
21	high	male	yes	3.00	3.00	1.00	3.
22	normal	male	yes	3.00	1.00	2.00	3.
23	high	female	yes	1.00	1.00	2.00	2.

Editing NEW.FDM Attributes : 10 Exa

Improved ChAID (Tschuprow Goodness of Split)

Various supervised learning algorithms are available. The most interesting ones are the decision tree methods. SIPINA can learn automatically the tree from the dataset. But it is also possible to interactively build the tree, guided by the domain knowledge. For most complete software, with other learning algorithms than trees, try **TANAGRA** which is also freely available (http://eric.univ-lyon2.fr/~ricco/tanagra/).

We select the IMROVED CHAID method. It is rather simple; it builds a short tree, useful in a first data exploration. We click on OK. A new dialog box appears, it enables us to set the parameters of the algorithm. We validate the default parameters.

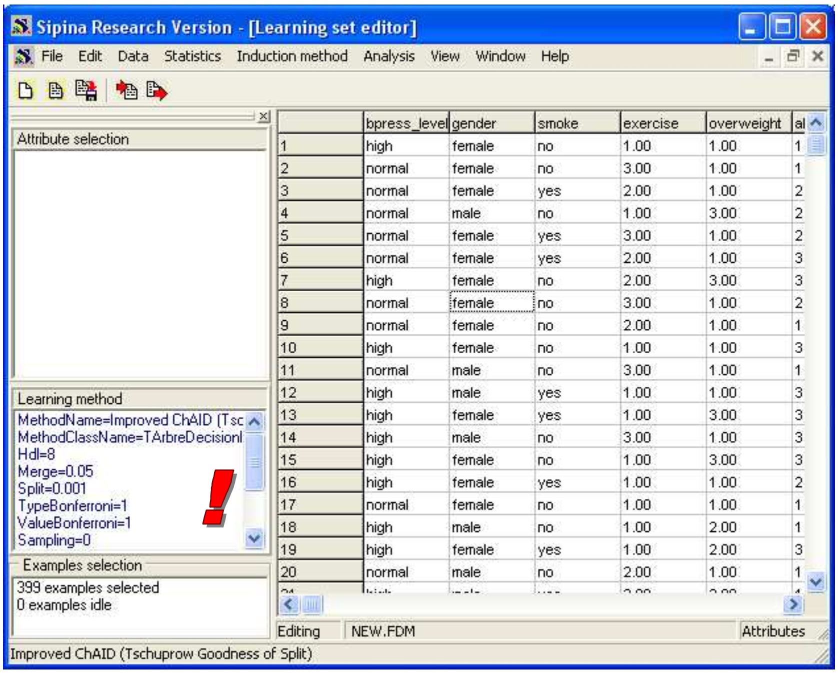


p-level
for merging nodes : 0.05
for splitting nodes : 0.001

Bonferroni adjustments
○ Automatic
● Manual
1

Other pruning parameters
Max. depth : 4
Min size of node to split : 20
Min size of leaves : 10

The user's choice (method and parameters) are displayed in the middle part of the project explorer.


Sipina Research Version - [Learning set editor]
File Edit Data Statistics Induction method Analysis View Window Help
Attribute selection
Learning method
Examples selection
!
Editing
NEW.FDM
Attributes

	bpress_level	gender	smoke	exercise	overweight	al
1	high	female	no	1.00	1.00	1
2	normal	female	no	3.00	1.00	1
3	normal	female	yes	2.00	1.00	2
4	normal	male	no	1.00	3.00	2
5	normal	female	yes	3.00	1.00	2
6	normal	female	yes	2.00	1.00	3
7	high	female	no	2.00	3.00	3
8	normal	female	no	3.00	1.00	2
9	normal	female	no	2.00	1.00	1
10	high	female	no	1.00	1.00	3
11	normal	male	no	3.00	1.00	1
12	high	male	yes	1.00	1.00	3
13	high	female	yes	1.00	3.00	3
14	high	male	no	3.00	1.00	3
15	high	female	no	1.00	3.00	3
16	high	female	yes	1.00	1.00	2
17	normal	female	no	1.00	1.00	1
18	high	male	no	1.00	2.00	1
19	high	female	yes	1.00	2.00	3
20	normal	male	no	2.00	1.00	1

MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

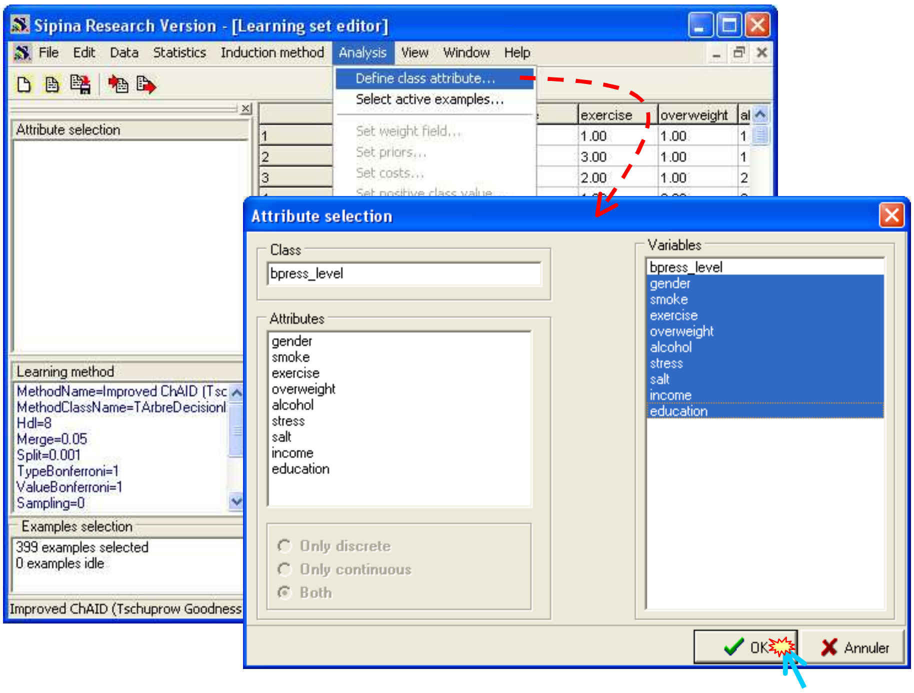
399 examples selected
0 examples idle

Improved ChAID (Tschuprow Goodness of Split)

Class attribute and predictive attributes

In the next step, we must specify the class attribute (the attribute that we want to predict) and the predictive ones (the descriptors).

We activate the ANALYSIS/DEFINE CLASS ATTRIBUTE menu. A dialog box appears. Using drag and drop principle, we set BPRESS LEVEL as CLASS (TARGET), and the other attributes as ATTRIBUTES (INPUT).

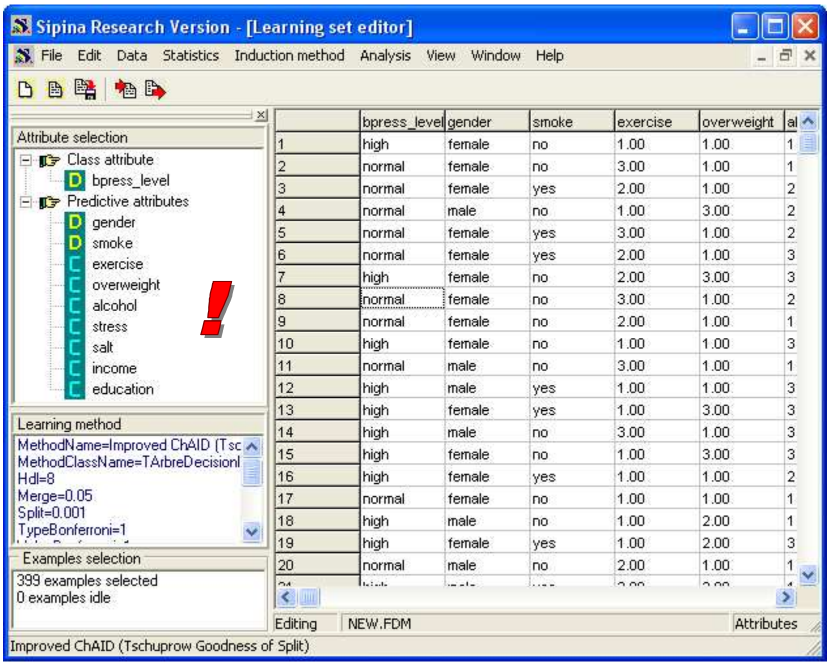

Sipina Research Version - [Learning set editor]
File Edit Data Statistics Induction method Analysis View Window Help
Define class attribute...
Select active examples...
Set weight field...
Set priors...
Set costs...
Set positive class value...
exercise
overweight
al
Attribute selection
Learning method
Examples selection
Attribute selection
○ Only discrete
○ Only continuous
● Both
✓ OK
✗ Annuler

Class: bpress_level

Attributes:
gender
smoke
exercise
overweight
alcohol
stress
salt
income
education

Variables:
bpress_level
gender
smoke
exercise
overweight
alcohol
stress
salt
income
education

We validate the selection with OK. The user's choices are displayed in the top part of the project explorer.


Sipina Research Version - [Learning set editor]
File Edit Data Statistics Induction method Analysis View Window Help
!

Attribute selection
- Class attribute
  - D bpress_level
- Predictive attributes
  - D gender
  - D smoke
  - C exercise
  - C overweight
  - C alcohol
  - C stress
  - C salt
  - C income
  - C education

Learning method
MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1

Examples selection
399 examples selected
0 examples idle

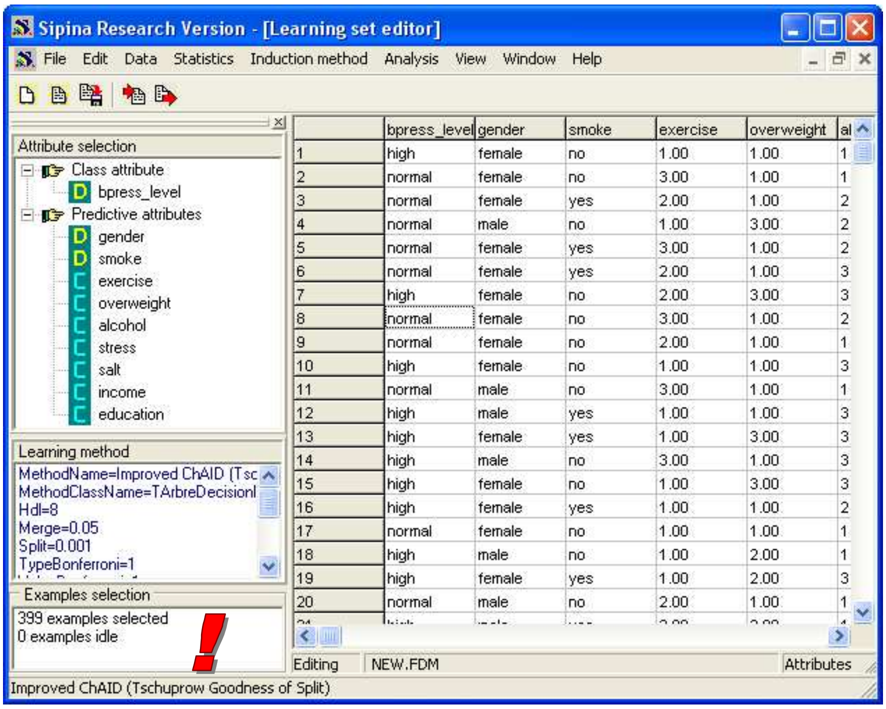
	bpress_level	gender	smoke	exercise	overweight	al
1	high	female	no	1.00	1.00	1
2	normal	female	no	3.00	1.00	1
3	normal	female	yes	2.00	1.00	2
4	normal	male	no	1.00	3.00	2
5	normal	female	yes	3.00	1.00	2
6	normal	female	yes	2.00	1.00	3
7	high	female	no	2.00	3.00	3
8	normal	female	no	3.00	1.00	2
9	normal	female	no	2.00	1.00	1
10	high	female	no	1.00	1.00	3
11	normal	male	no	3.00	1.00	1
12	high	male	yes	1.00	1.00	3
13	high	female	yes	1.00	3.00	3
14	high	male	no	3.00	1.00	3
15	high	female	no	1.00	3.00	3
16	high	female	yes	1.00	1.00	2
17	normal	female	no	1.00	1.00	1
18	high	male	no	1.00	2.00	1
19	high	female	yes	1.00	2.00	3
20	normal	male	no	2.00	1.00	1

Editing NEW.FDM Attributes
Improved ChAID (Tschuprow Goodness of Split)

Some of the input attributes are maybe irrelevant. Induction tree algorithm can highlight automatically the most relevant predictive attributes. It is one of its key points.

Selection of the learning sample

We use mainly an interactive approach in this tutorial. We use all the dataset for the tree induction. It is the default selection. We see the dataset selection in the bottom part of the project explorer.


Sipina Research Version - [Learning set editor]
File Edit Data Statistics Induction method Analysis View Window Help
!

Attribute selection
- Class attribute
  - D bpress_level
- Predictive attributes
  - D gender
  - D smoke
  - C exercise
  - C overweight
  - C alcohol
  - C stress
  - C salt
  - C income
  - C education

Learning method
MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1

Examples selection
399 examples selected
0 examples idle

	bpress_level	gender	smoke	exercise	overweight	al
1	high	female	no	1.00	1.00	1
2	normal	female	no	3.00	1.00	1
3	normal	female	yes	2.00	1.00	2
4	normal	male	no	1.00	3.00	2
5	normal	female	yes	3.00	1.00	2
6	normal	female	yes	2.00	1.00	3
7	high	female	no	2.00	3.00	3
8	normal	female	no	3.00	1.00	2
9	normal	female	no	2.00	1.00	1
10	high	female	no	1.00	1.00	3
11	normal	male	no	3.00	1.00	1
12	high	male	yes	1.00	1.00	3
13	high	female	yes	1.00	3.00	3
14	high	male	no	3.00	1.00	3
15	high	female	no	1.00	3.00	3
16	high	female	yes	1.00	1.00	2
17	normal	female	no	1.00	1.00	1
18	high	male	no	1.00	2.00	1
19	high	female	yes	1.00	2.00	3
20	normal	male	no	2.00	1.00	1

Editing NEW.FDM Attributes
Improved ChAID (Tschuprow Goodness of Split)

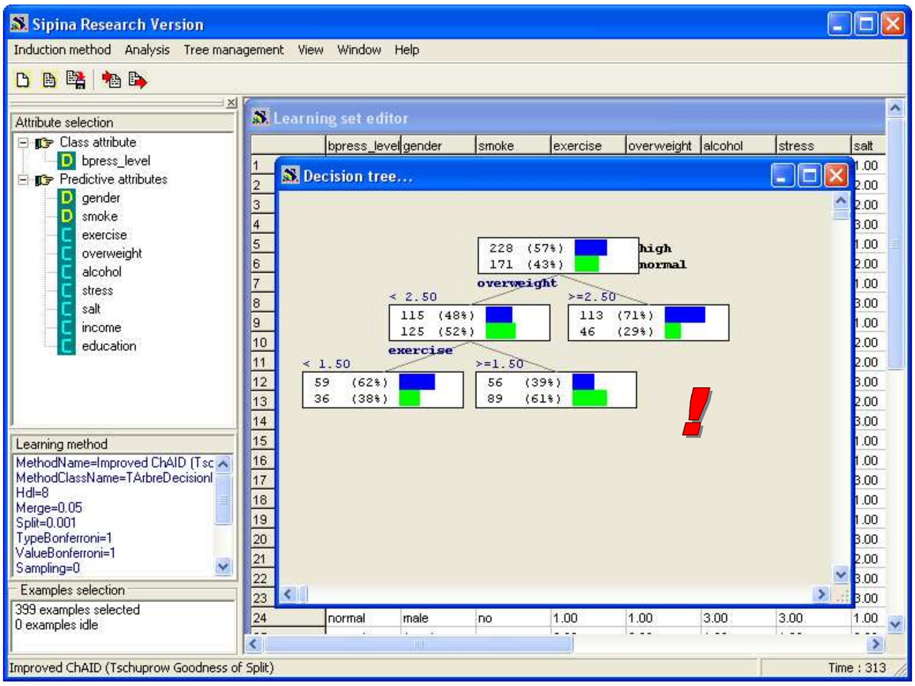
Note on subsample definition: If we want to subdivide the dataset into learning and test sample. We use the ANALYSIS / SELECT ACTIVE EXAMPLES menu. We can subdivide randomly the dataset, we can also use a rule based selection, etc.

Start problem analysis

We want to build the tree using the selected method. We activate the ANALYSIS / LEARNING menu. From the decision tree (Figure 1), we extract the following rules.

Rule *(**If** Premise **Then** Conclusion)*	Confidence	Lift	Support
If overweight >= 2.5 Then Blood pressure = high	71%	1.25	40%[2]
If overweight < 2.5 ET Exercise < 1.5 Then Blood Pressure = high	62%	1.08	24%
If overweight < 2.5 ET Exercise >= 1.5 Then Blood pressure = low	61%	1.40	36%

A rule is relevant if it has a high confidence and a high support.



Attribute selection
- Class attribute
  - D bpress_level
- Predictive attributes
  - D gender
  - D smoke
  - C exercise
  - C overweight
  - C alcohol
  - C stress
  - C salt
  - C income
  - C education

Learning method

```
MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0
```

Examples selection

399 examples selected
0 examples idle

	bpress_level	gender	smoke	exercise	overweight	alcohol	stress	salt
24	normal	male	no	1.00	1.00	3.00	3.00	1.00

Improved ChAID (Tschuprow Goodness of Split) Time : 313

Figure 1 – Decision tree on the "blood pressure" dataset

We can evaluate the information provided by a rule by computing the ratio between the proportion of class in the whole dataset and the proportion of the class in the rule covered examples: it is the LIFT value. If the LIFT is upper than 1, the rule is interesting. In the first rule, the LIFT is 1.40 = 71%/57%.

Interactive exploration of the tree

Now begins really the work of domain experts. The goal is to obtain a better characterization of rules associated to the leaves of the tree. For that purpose, we have to determine the role of the variables which do not apparently appears in the tree. Are they really irrelevant or are they masked by the selected variables?

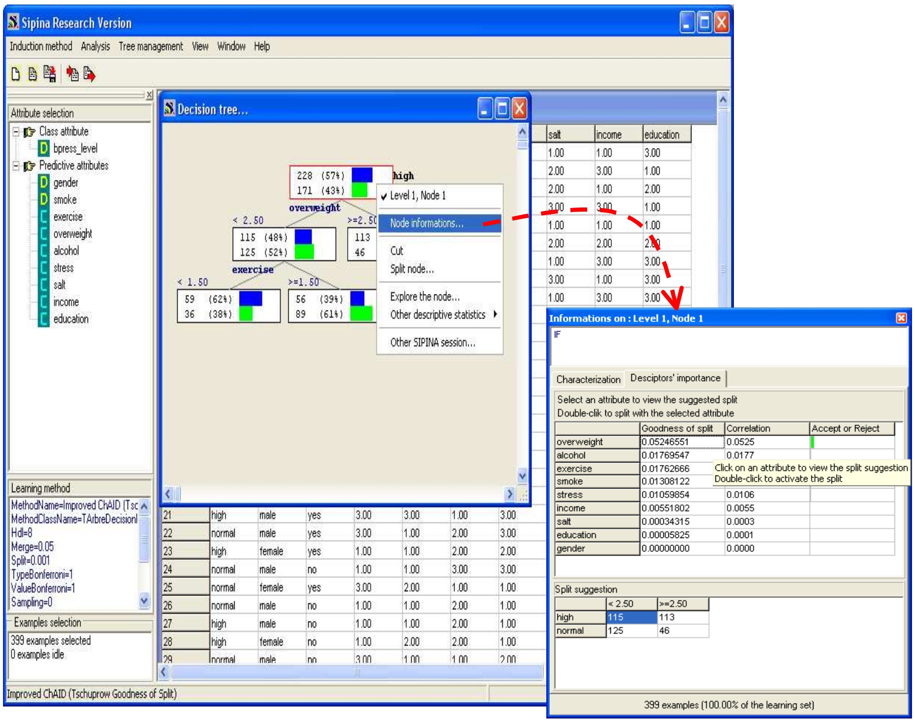
[2] The definition of the support for the predictive rules (supervised learning) is different from that used in the association rules induction. Here, it is the number of examples covered by the premise divided by the whole dataset size i.e. 40% = (113+46)/399.

Alternative splitting variables

On the root of the tree, the first split uses "overweight". Is it the only relevant variable? What about the other predictive variables?

Indeed, the technique chooses simply the best variable in the sense of a given criterion. The other variables may be masked, even if they have almost an equivalent quality. This is not absolutely harmless. If we choose to split with another variable on the place of that automatically detected by the method, it is possible that the other variables occurring in the low parts of the tree are completely different. We obtain very different rules[3].

We are going to study the alternative solutions of the variable "overweight" during the segmentation of the root of the tree. For that purpose, we make a click with the right button of the mouse on this node. In the contextual menu which appears, we activate the option NODE INFORMATION …



MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

399 examples selected
0 examples idle

Improved ChAID (Tschuprow Goodness of Split)

salt	income	education
1.00	1.00	3.00
2.00	3.00	1.00
2.00	1.00	2.00
3.00	3.00	1.00
1.00	1.00	1.00
2.00	2.00	2.00
1.00	3.00	3.00
3.00	1.00	3.00
1.00	3.00	3.00

21	high	male	yes	3.00	3.00	1.00	3.00
22	normal	male	yes	3.00	1.00	2.00	3.00
23	high	female	yes	1.00	1.00	2.00	2.00
24	normal	male	no	1.00	1.00	3.00	3.00
25	normal	female	yes	3.00	2.00	1.00	1.00
26	normal	male	no	1.00	1.00	2.00	1.00
27	high	male	no	1.00	1.00	2.00	1.00
28	high	female	no	1.00	2.00	2.00	1.00
29	normal	male	no	3.00	1.00	1.00	2.00

Informations on : Level 1, Node 1

IF

Characterization | Desciptors' importance

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

	Goodness of split	Correlation	Accept or Reject
overweight	0.05246551	0.0525	
alcohol	0.01769547	0.0177	
exercise	0.01762666		
smoke	0.01308122		
stress	0.01059854	0.0106	
income	0.00551802	0.0055	
salt	0.00034315	0.0003	
education	0.00005825	0.0001	
gender	0.00000000	0.0000	

Click on an attribute to view the split suggestion
Double-click to activate the split

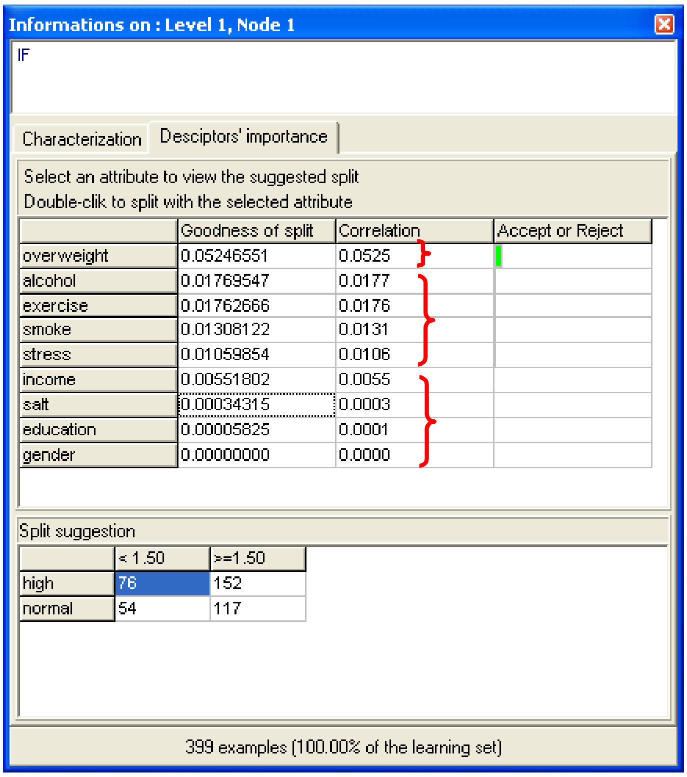
Split suggestion

	< 2.50	>=2.50
high	115	113
normal	125	46

399 examples (100.00% of the learning set)

A new window appears. We observe the list of predictive variables and their respective contributions if we use them. This method uses the TSCHUPROW's T in order to characterize a split.

[3] But, if the rules seem different, they classify the examples in the same way. In this point of view, the trees are similar.



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

	Goodness of split	Correlation	Accept or Reject
overweight	0.05246551	0.0525	🟩
alcohol	0.01769547	0.0177	
exercise	0.01762666	0.0176	
smoke	0.01308122	0.0131	
stress	0.01059854	0.0106	
income	0.00551802	0.0055	
salt	0.00034315	0.0003	
education	0.00005825	0.0001	
gender	0.00000000	0.0000	

Split suggestion

	< 1.50	>=1.50
high	76	152
normal	54	117

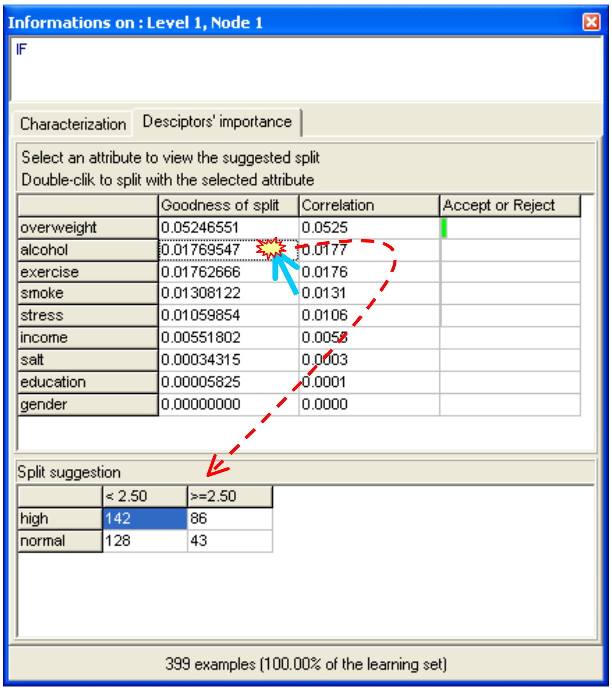
399 examples (100.00% of the learning set)

We observe mainly three groups of variables:

1. OVERWEIGHT is really the most interesting variable for splitting the root node. In the low part of the window appears the associated partitioning. It is the same that we observe in the graphical representation of the tree.
2. ALCOHOL, EXERCISE, SMOKE and STRESS are the next. They bring less information than OVERWEIGHT, but they still provide information about BLOOD PRESSURE discrimination problem.
3. Then, INCOME, SALT, EDUCATION and GENDER seem not relevant, at this step, for the detection of individuals with high BLOOD PRESSURE.

We wonder what happens if we choose another splitting variable, ALCOHOL for instance. Because the expert points out that it is important. In that case, criteria which are not numeric come into play in the study.

In a first step, we do not want to modify the tree, we want to see only the resulting partition if we use ALCOHOL. For that, we click on GOODNESS OF SPLIT value for each variable. In the case of ALCOHOL, we activate the corresponding box; the bottom part of the window reflects the associated segmentation.



Informations on : Level 1, Node 1

IF

Characterization | Desciptors' importance

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

	Goodness of split	Correlation	Accept or Reject
overweight	0.05246551	0.0525	🟩
alcohol	0.01769547	0.0177	
exercise	0.01762666	0.0176	
smoke	0.01308122	0.0131	
stress	0.01059854	0.0106	
income	0.00551802	0.0055	
salt	0.00034315	0.0003	
education	0.00005825	0.0001	
gender	0.00000000	0.0000	

Split suggestion

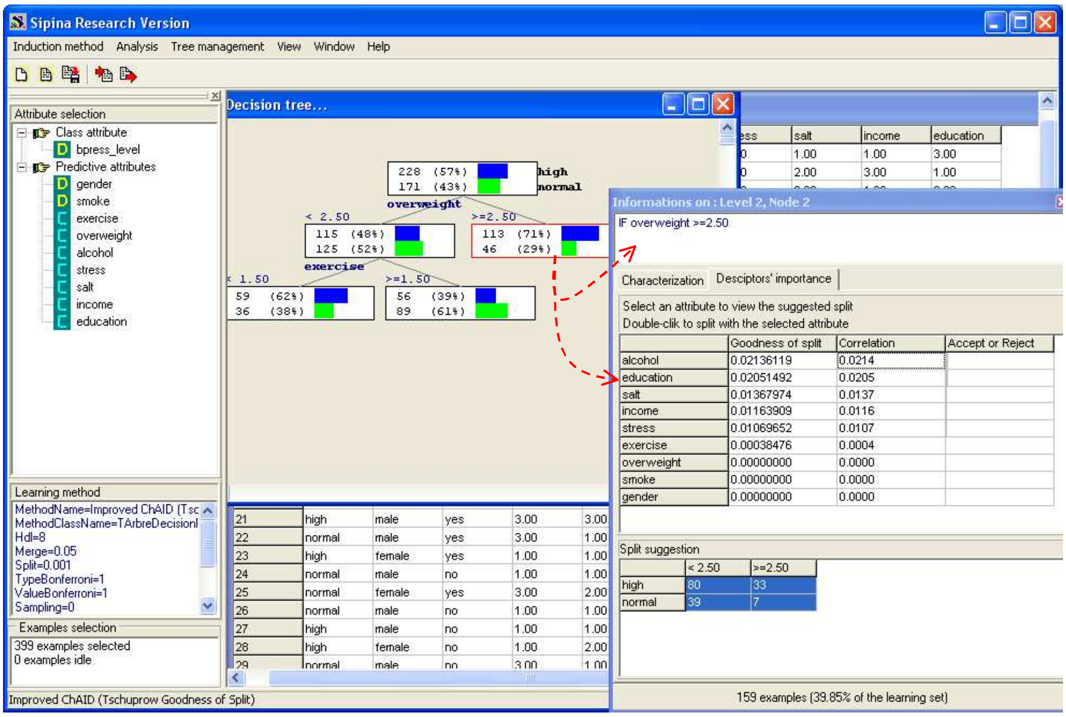
	< 2.50	>=2.50
high	142	86
normal	128	43

399 examples (100.00% of the learning set)

The expert, according the domain knowledge, may say us if this solution is really irrelevant or not.

Description of a node (subpopulation)

Now, we want to study the node at the following level, the group of the people "OVERWEIGHT > 2.5". We select the node. It is not necessary to close the previous window. The values are automatically updated.



Learning method

MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

Examples selection

399 examples selected
0 examples idle

21	high	male	yes	3.00	3.00
22	normal	male	yes	3.00	1.00
23	high	female	yes	1.00	1.00
24	normal	male	no	1.00	1.00
25	normal	female	yes	3.00	2.00
26	normal	male	no	1.00	1.00
27	high	male	no	1.00	1.00
28	high	female	no	1.00	2.00
29	normal	male	no	3.00	1.00

ess	salt	income	education
0	1.00	1.00	3.00
0	2.00	3.00	1.00

Improved ChAID (Tschuprow Goodness of Split)

Informations on : Level 2, Node 2

IF overweight >=2.50

Characterization | Desciptors' importance

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

	Goodness of split	Correlation	Accept or Reject
alcohol	0.02136119	0.0214	
education	0.02051492	0.0205	
salt	0.01367974	0.0137	
income	0.01163909	0.0116	
stress	0.01069652	0.0107	
exercise	0.00038476	0.0004	
overweight	0.00000000	0.0000	
smoke	0.00000000	0.0000	
gender	0.00000000	0.0000	

Split suggestion

	< 2.50	>=2.50
high	80	33
normal	39	7

159 examples (39.85% of the learning set)

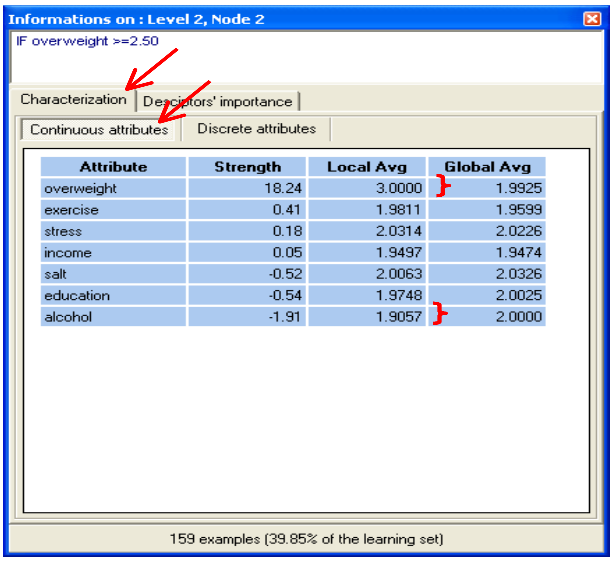
This node might be split using the variables ALCOHOL or EDUCATION. The operation was not carried out because it seems not to be numerically relevant.

The second important matter legitimates the interactive analysis in the induction trees: it is understood that the people associated with this node are overweight and hypertensive, but what about other variables, what are the other characteristics of these individuals? This is one of the drawbacks of the decision trees. The method proposes the relevant variables. But we have not visibility on the variables which were not integrated in the model. However, they can help us to deeply characterize the rules.

The interactive functionalities of SIPINA enable us to answer this requirement. The CHARACTERIZATION tab describes the groups associated with each node of the tree. SIPINA computes comparative statistics between the root node, representing the whole population, and the current node, representing the subpopulation defined by the rule.

To evaluate the importance of the difference, the value test (strength) which is a test statistic: a comparison of mean when the variable is continuous (t-test), a comparison of proportion when the variable is discrete. It is not strictly speaking a statistical test since the samples are not independent, but its interest and its flexibility are undeniable in practice.

In the "Continuous" tab, we have the values of mean on the root node (Global Avg.) and on the current node (Local Avg.).



IF overweight >=2.50

Attribute	Strength	Local Avg	Global Avg
overweight	18.24	3.0000	1.9925
exercise	0.41	1.9811	1.9599
stress	0.18	2.0314	2.0226
income	0.05	1.9497	1.9474
salt	-0.52	2.0063	2.0326
education	-0.54	1.9748	2.0025
alcohol	-1.91	1.9057	2.0000

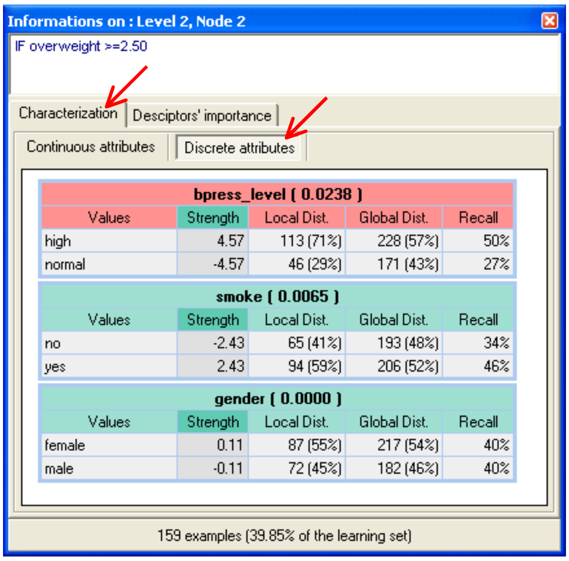
159 examples (39.85% of the learning set)

The OVERWEIGHT is really high on this subpopulation, 3.0 vs. 1.9925 on the whole population. The importance of the difference is materialized with a STRENGTH = +18.24. This result is obvious because OVERWEIGHT takes part in the decision tree construction.

But another issue, which is not obvious, is that ALCOHOL consumption seems significantly low in this subpopulation: STRENGTH = -1.91.

It is not really possible to give a threshold which enables us to decide if the difference is significant or not. Because the samples are not independent, and the group is designed in order to optimize a purity criterion. But, we can however distinguish an abnormal deviation according to the comparison based on the other variables. In this subgroup, it seems that OVERWEIGHT and ALCOHOL are interesting.

For the “Discrete” variables, we obtain the following results:



IF overweight >=2.50

bpress_level (0.0238)				
Values	Strength	Local Dist.	Global Dist.	Recall
high	4.57	113 (71%)	228 (57%)	50%
normal	-4.57	46 (29%)	171 (43%)	27%

smoke (0.0065)				
Values	Strength	Local Dist.	Global Dist.	Recall
no	-2.43	65 (41%)	193 (48%)	34%
yes	2.43	94 (59%)	206 (52%)	46%

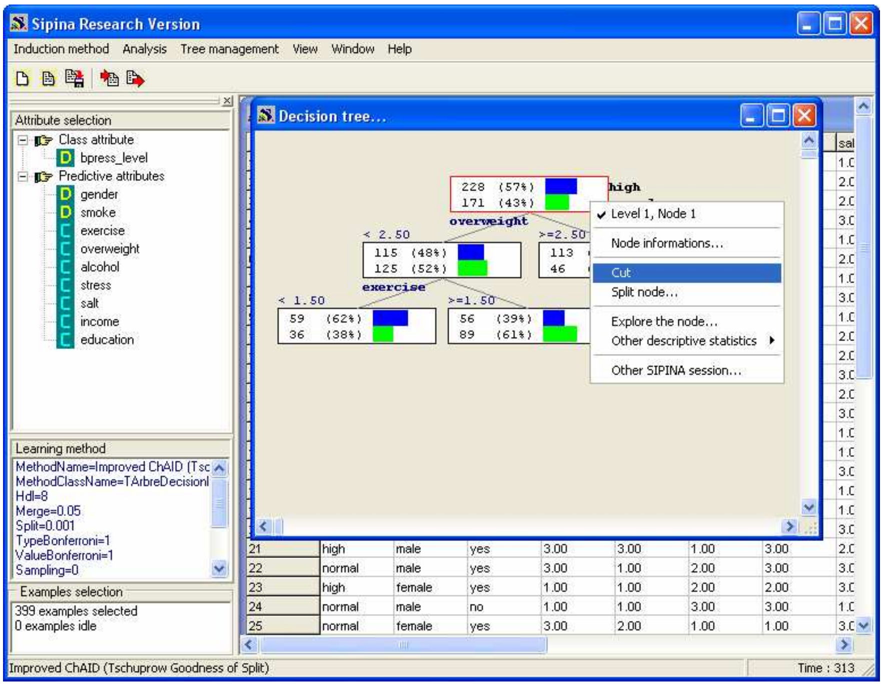
gender (0.0000)				
Values	Strength	Local Dist.	Global Dist.	Recall
female	0.11	87 (55%)	217 (54%)	40%
male	-0.11	72 (45%)	182 (46%)	40%

159 examples (39.85% of the learning set)

Now, STRENGTH is the statistical test of the comparison of proportions (frequencies). We observe that there is over representation of BPRESS = HIGH. But the proportion of SMOKE = YES is also high in this subgroup, compared with the whole population. This last characteristic is not visible in the decision tree.

User-driven induction tree

The possibility for the user of guiding the exploration is certainly one of the most desirable aspects of the decision trees. Let us take again the variable SMOKE which seems very important finally. We decide to insert it as first split variable of the root node. We select the root of the tree. In order to prune it manually, we carry out a click with the right button of the mouse, in the contextual menu we select the CUT option.



MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

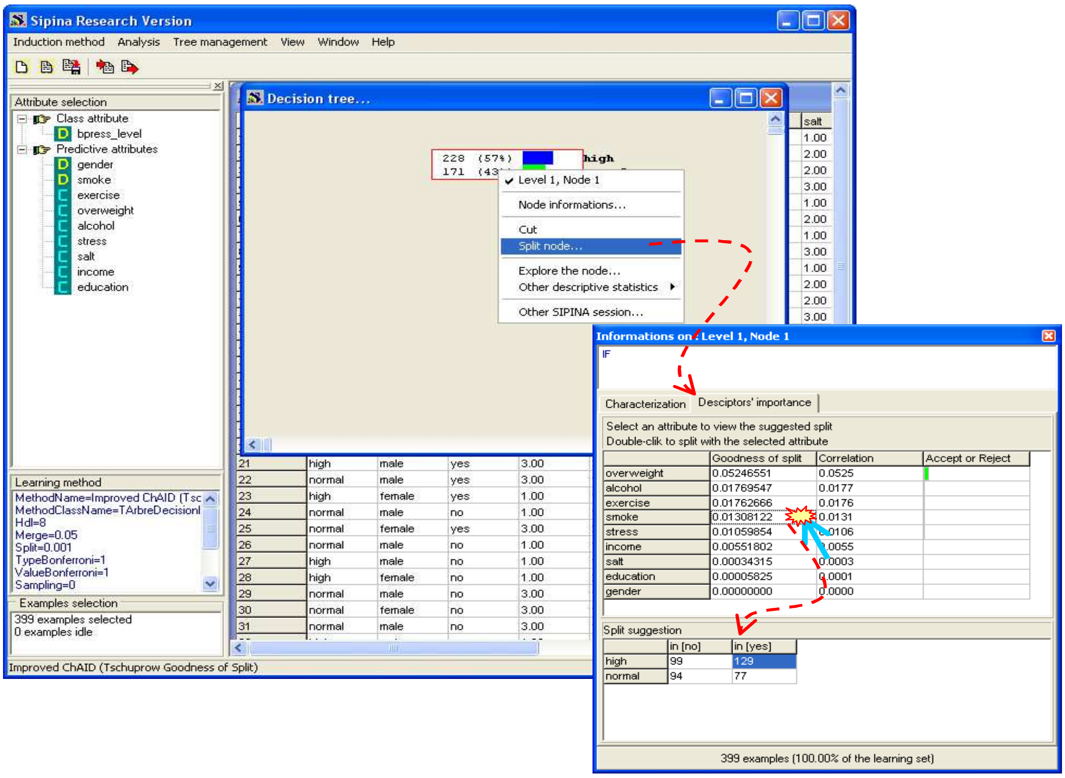
399 examples selected
0 examples idle

21	high	male	yes	3.00	3.00	1.00	3.00	2.C
22	normal	male	yes	3.00	1.00	2.00	3.00	3.C
23	high	female	yes	1.00	1.00	2.00	2.00	3.C
24	normal	male	no	1.00	1.00	3.00	3.00	1.C
25	normal	female	yes	3.00	2.00	1.00	1.00	3.C

Improved ChAID (Tschuprow Goodness of Split) Time : 313

Always with the contextual menu, we activate the option SPLIT NODE, we obtain the same window as for NODE INFORMATION option. We select DESCRIPTORS' IMPORTANCE tab. In the list of the candidate descriptors, we click on SMOKE (Figure 2).

In order to perform a split, we double-click on the box containing the value of GOODNESS OF SPLIT for the selected variable. In this case, splitting is carried out even if the conditions of acceptance (significance test etc.) are not met (Figure 3).



MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

399 examples selected
0 examples idle

Improved ChAID (Tschuprow Goodness of Split)

21	high	male	yes	3.00
22	normal	male	yes	3.00
23	high	female	yes	1.00
24	normal	male	no	1.00
25	normal	female	yes	3.00
26	normal	male	no	1.00
27	high	male	no	1.00
28	high	female	no	1.00
29	normal	male	no	3.00
30	normal	female	no	3.00
31	normal	male	no	3.00

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

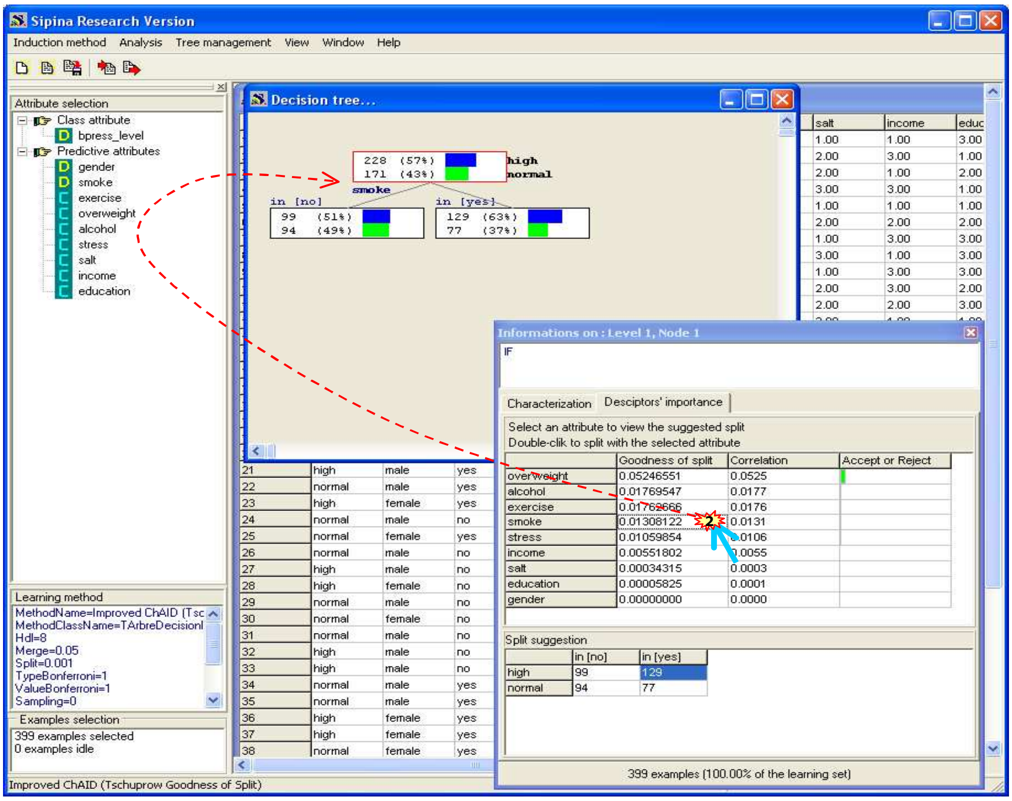
	Goodness of split	Correlation	Accept or Reject
overweight	0.05246551	0.0525	🟩
alcohol	0.01769547	0.0177	
exercise	0.01762666	0.0176	
smoke	0.01308122	0.0131	
stress	0.01059854	0.0106	
income	0.00551802	0.0055	
salt	0.00034315	0.0003	
education	0.00005825	0.0001	
gender	0.00000000	0.0000	

	in [no]	in [yes]
high	99	129
normal	94	77

399 examples (100.00% of the learning set)

Figure 2 - Splitting with SMOKE attribute

We note that there is indeed an over representation of hypertensive among smokers (63% compared with 57% in the global population). The proportions are balanced for non smokers.



Attribute selection

- Class attribute
  - D bpress_level
- Predictive attributes
  - D gender
  - D smoke
  - C exercise
  - C overweight
  - C alcohol
  - C stress
  - C salt
  - C income
  - C education

Learning method

MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

Examples selection

399 examples selected
0 examples idle

21	high	male	yes
22	normal	male	yes
23	high	female	yes
24	normal	male	no
25	normal	female	yes
26	normal	male	no
27	high	male	no
28	high	female	no
29	normal	male	no
30	normal	female	no
31	normal	male	no
32	high	male	no
33	high	male	no
34	normal	male	yes
35	normal	male	yes
36	high	female	yes
37	high	female	yes
38	normal	female	yes

salt	income	educ
1.00	1.00	3.00
2.00	3.00	1.00
2.00	1.00	2.00
3.00	3.00	1.00
1.00	1.00	1.00
2.00	2.00	2.00
1.00	3.00	3.00
3.00	1.00	3.00
1.00	3.00	3.00
2.00	3.00	2.00
2.00	2.00	3.00

Informations on : Level 1, Node 1

IF

Characterization | Desciptors' importance

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

	Goodness of split	Correlation	Accept or Reject
overweight	0.05246551	0.0525	
alcohol	0.01769547	0.0177	
exercise	0.01762666	0.0176	
smoke	0.01308122	0.0131	
stress	0.01059854	0.0106	
income	0.00551802	0.0055	
salt	0.00034315	0.0003	
education	0.00005825	0.0001	
gender	0.00000000	0.0000	

Split suggestion

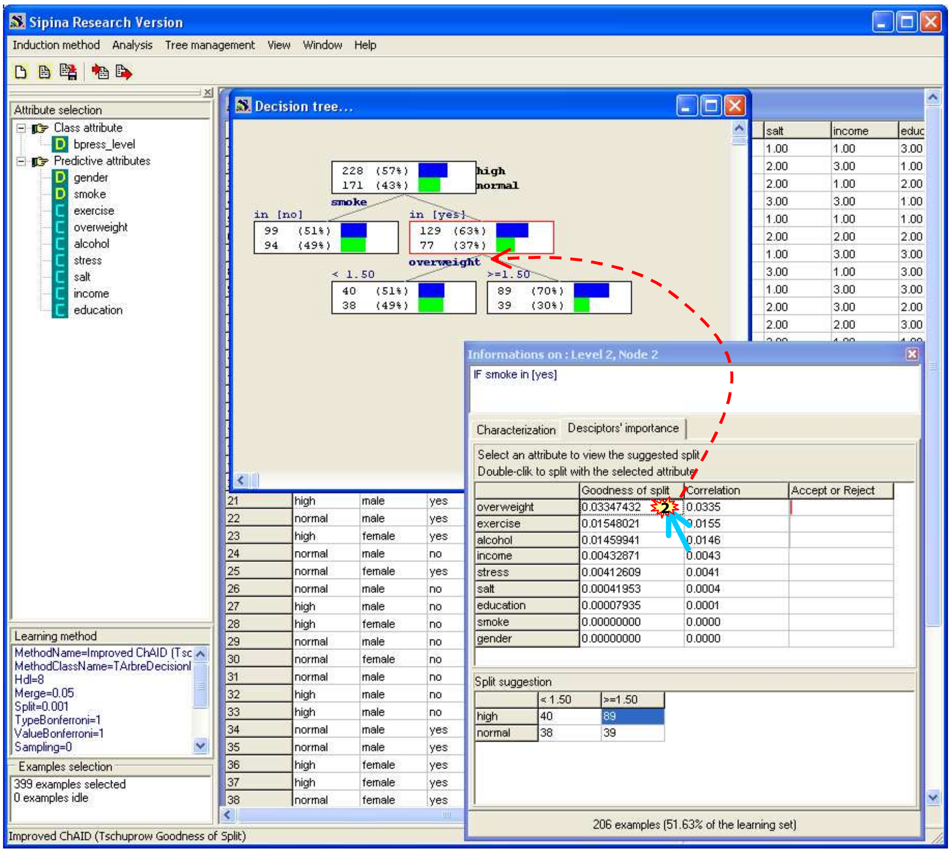
	in [no]	in [yes]
high	99	129
normal	94	77

399 examples (100.00% of the learning set)

Improved ChAID (Tschuprow Goodness of Split)

Figure 3 – Tree with SMOKE as first splitting variable

From now, we can build the tree according to our domain knowledge. We select the group SMOKE = YES. We observe the Tschuprow's t of predictive variables. OVERWEIGHT seems again a relevant attribute. We double-click on the value in order to split the node.



Attribute selection

- Class attribute
  - D bpress_level
- Predictive attributes
  - D gender
  - D smoke
  - C exercise
  - C overweight
  - C alcohol
  - C stress
  - C salt
  - C income
  - C education

Learning method

MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

Examples selection

399 examples selected
0 examples idle

21	high	male	yes
22	normal	male	yes
23	high	female	yes
24	normal	male	no
25	normal	female	yes
26	normal	male	no
27	high	male	no
28	high	female	no
29	normal	male	no
30	normal	female	no
31	normal	male	no
32	high	male	no
33	high	male	no
34	normal	male	yes
35	normal	male	yes
36	high	female	yes
37	high	female	yes
38	normal	female	yes

salt	income	educ
1.00	1.00	3.00
2.00	3.00	1.00
2.00	1.00	2.00
3.00	3.00	1.00
1.00	1.00	1.00
2.00	2.00	2.00
1.00	3.00	3.00
3.00	1.00	3.00
1.00	3.00	3.00
2.00	3.00	2.00
2.00	2.00	3.00

Informations on : Level 2, Node 2

IF smoke in [yes]

Characterization | Desciptors' importance

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

	Goodness of split	Correlation	Accept or Reject
overweight	0.03347432	0.0335	
exercise	0.01548021	0.0155	
alcohol	0.01459941	0.0146	
income	0.00432871	0.0043	
stress	0.00412609	0.0041	
salt	0.00041953	0.0004	
education	0.00007935	0.0001	
smoke	0.00000000	0.0000	
gender	0.00000000	0.0000	

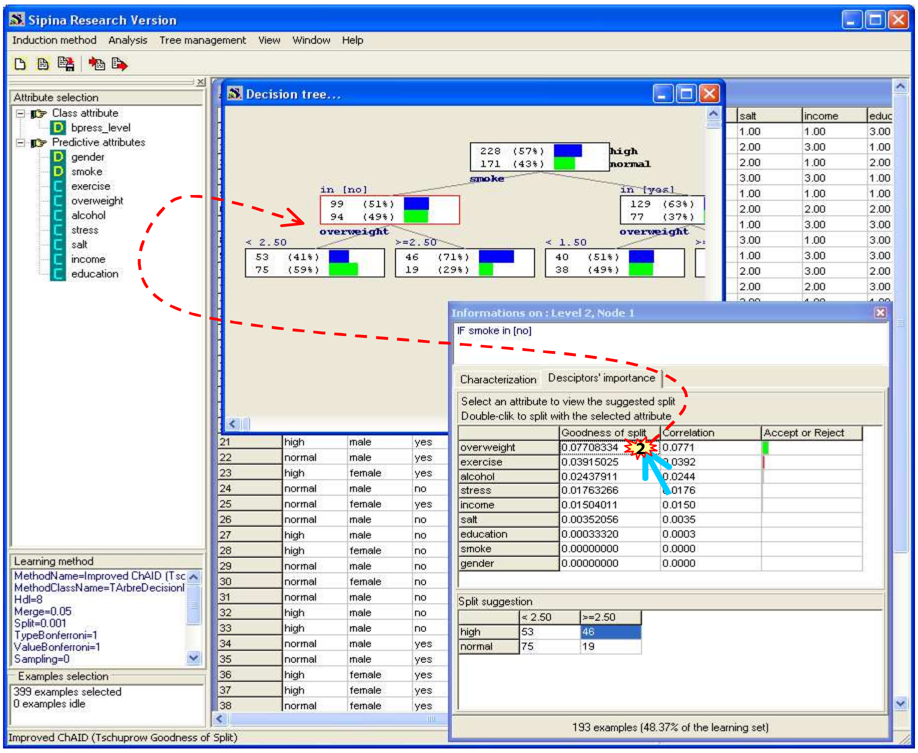
Split suggestion

	< 1.50	>=1.50
high	40	89
normal	38	39

206 examples (51.63% of the learning set)

Improved ChAID (Tschuprow Goodness of Split)

BLOOD PRESSURE is high when OVERWEIGHT is high. But compared with the first tree (Figure 1), we note that the cut point is different for the SMOKE = YES group. A moderate OVERWEIGHT (> 1.5) is detrimental when the people are smoker. The threshold is higher (> 2.5) for the whole population. For non smokers, the cut point (> 2.5) is the same one.



salt	income	educ
1.00	1.00	3.00
2.00	3.00	1.00
2.00	1.00	2.00
3.00	3.00	1.00
1.00	1.00	1.00
2.00	2.00	2.00
1.00	3.00	3.00
3.00	1.00	3.00
1.00	3.00	3.00
2.00	3.00	2.00
2.00	2.00	3.00

IF smoke in [no]

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

	Goodness of split	Correlation	Accept or Reject
overweight	0.07708334	0.0771	
exercise	0.03915025	0.0392	
alcohol	0.02437911	0.0244	
stress	0.01763266	0.0176	
income	0.01504011	0.0150	
salt	0.00352056	0.0035	
education	0.00033320	0.0003	
smoke	0.00000000	0.0000	
gender	0.00000000	0.0000	

Split suggestion

	< 2.50	>=2.50
high	53	46
normal	75	19

193 examples (48.37% of the learning set)

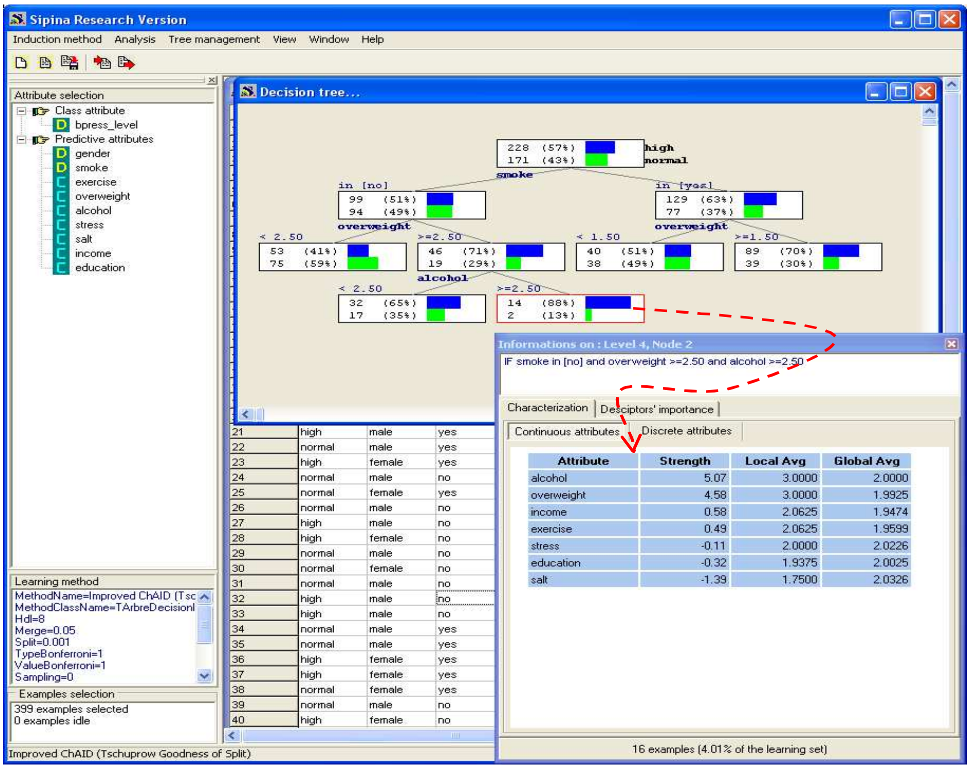
Learning method
MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

Examples selection
399 examples selected
0 examples idle

21	high	male	yes
22	normal	male	yes
23	high	female	yes
24	normal	male	no
25	normal	female	yes
26	normal	male	no
27	high	male	no
28	high	female	no
29	normal	male	no
30	normal	female	no
31	normal	male	no
32	high	male	no
33	high	male	no
34	normal	male	yes
35	normal	male	yes
36	high	female	yes
37	high	female	yes
38	normal	female	yes

Improved ChAID (Tschuprow Goodness of Split)

When we continue the analysis, we may obtain (among various solutions) the following tree.



IF smoke in [no] and overweight >=2.50 and alcohol >=2.50

Attribute	Strength	Local Avg	Global Avg
alcohol	5.07	3.0000	2.0000
overweight	4.58	3.0000	1.9925
income	0.58	2.0625	1.9474
exercise	0.49	2.0625	1.9599
stress	-0.11	2.0000	2.0226
education	-0.32	1.9375	2.0025
salt	-1.39	1.7500	2.0326

16 examples (4.01% of the learning set)

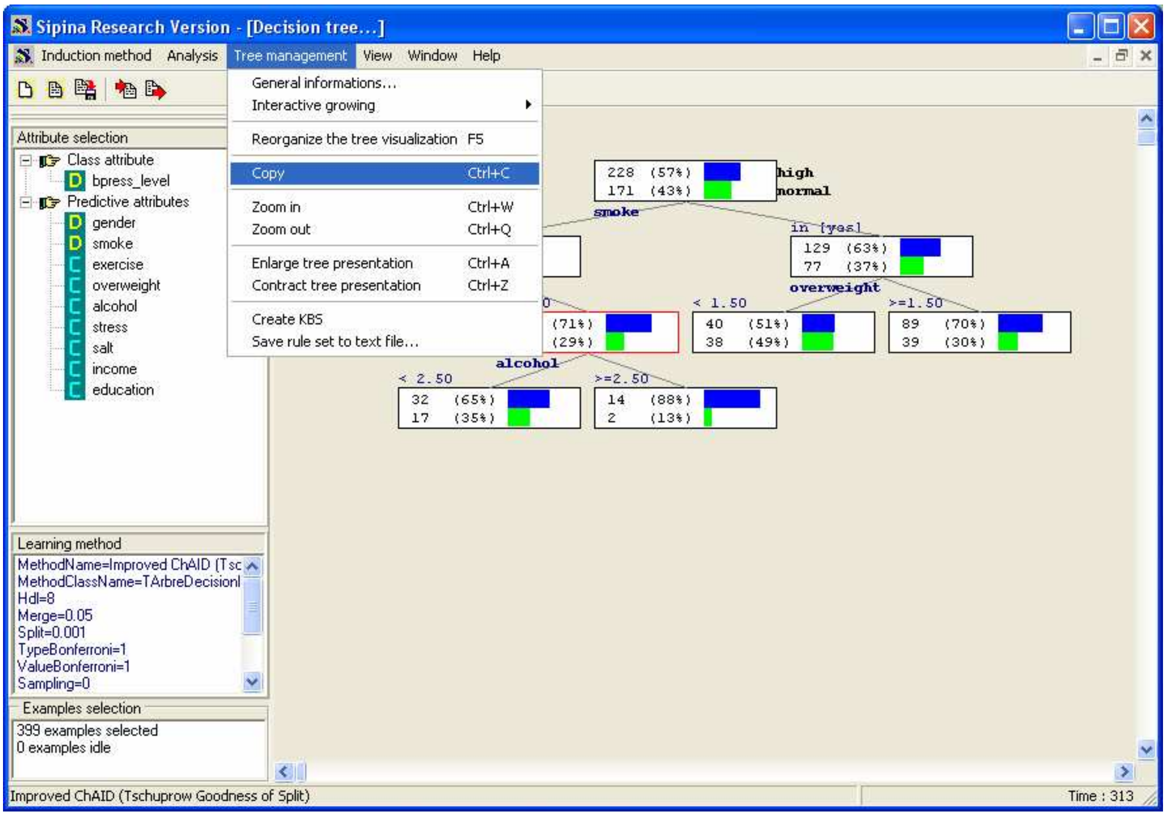
Learning method
MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

Examples selection
399 examples selected
0 examples idle

21	high	male	yes
22	normal	male	yes
23	high	female	yes
24	normal	male	no
25	normal	female	yes
26	normal	male	no
27	high	male	no
28	high	female	no
29	normal	male	no
30	normal	female	no
31	normal	male	no
32	high	male	no
33	high	male	no
34	normal	male	yes
35	normal	male	yes
36	high	female	yes
37	high	female	yes
38	normal	female	yes
39	normal	male	no
40	high	female	no

Improved ChAID (Tschuprow Goodness of Split)

Even if SIPINA is far from having all the functionalities of commercial software (SPAD Interactive Tree Decision, SPSS Answer Tree, SAS EM, etc), it nevertheless proposes some options intended to improve the presentation of the results. They are available in the TREE MANAGEMENT menu. We can, among other things, copy the graphical representation of the tree in the clipboard.



MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

399 examples selected
0 examples idle

Improved ChAID (Tschuprow Goodness of Split)

Time : 313

Conclusion

The decision trees are a popular data mining method. This popularity relies partly on the flexibility of the software which gives to the user the possibility to guide the induction process according to the domain knowledge. In this tutorial, we present the functionalities of SIPINA for interactive exploration.