1. Introduction

Working with Sipina under Linux using Wine.

In a recent tutorial (http://data-mining-tutorials.blogspot.com/2009/01/tanagra-under-linux.html), we show that it is possible to work with Tanagra under Linux using Wine. In this document, we implement Sipina with the same framework i.e. we install and use Sipina in a Linux environment. We use the Ubuntu distribution (French version 8.10). All the functionalities of Sipina are available, especially the interactive tools which allows us to explore deeply the subpopulation into a node of the tree.

In this tutorial, we implement the following steps:

1. Installing Sipina under Linux;
2. Launching the software;
3. Loading a dataset (text file with tab separator);
4. Choosing the class attribute and the predictive variables;
5. Partitioning the dataset in a train set and test set;
6. Computing the tree on the train set;
7. Evaluation the tree on the test set e.g. computing the confusion matrix, the error rate, etc.;
8. Exploring a subpopulation related to a node of the tree;
9. Launching a new analysis on a subpopulation related to a node of the tree.

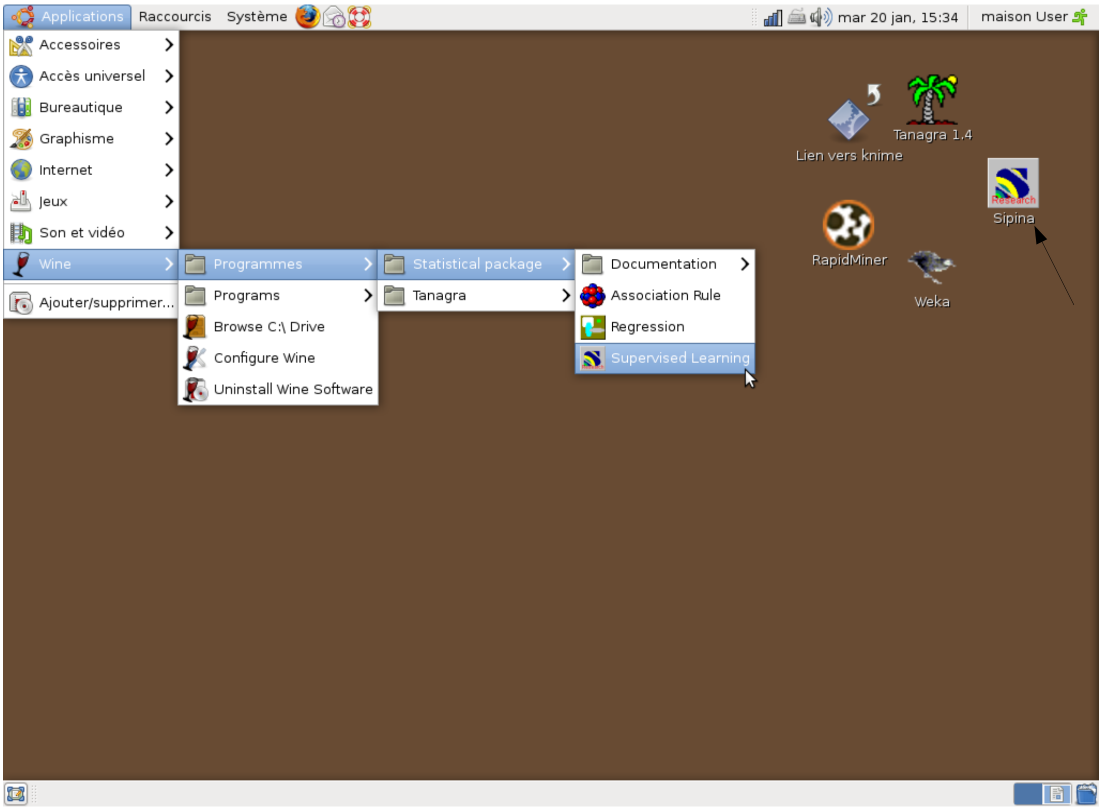
We will describe quickly the various features of the software in this tutorial. They are already presented in several documents available online (http://eric.univ-lyon2.fr/~ricco/sipina.html, see the DOWNLOAD section). Our main goal here is to show the capabilities of Sipina under Linux.

We use the French Ubuntu 8.10 distribution (http://www.ubuntu.com/); we have installed also Wine [http://en.wikipedia.org/wiki/Wine_(software)], a program which allows to Windows programs to run under Linux. We show in the tutorial that Sipina is fully functional in this environment.

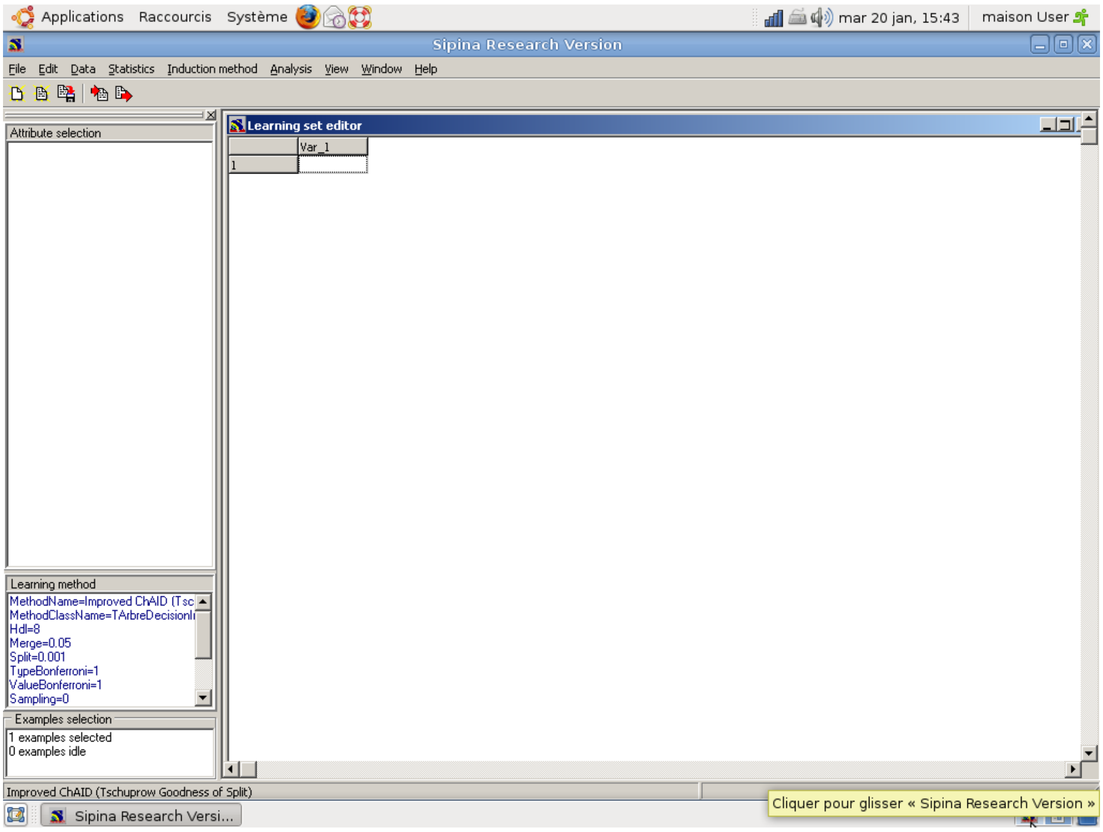
2. Installation of Sipina

After we download the setup file « setup_statp_ackage.exe » on our server (http://eric.univ-lyon2.fr/~ricco/sipina.html), we start the installation using wine. The simplest way is to enter the following command line on a terminal window: *wine /your path/setup_stat_package.exe*

The installation process is started. We can validate each step and use the default installation path (c:\program files\...). At the end of the installation, the "Statistical Package" is now available in the "Applications" menu as we can see in the screenshot below. Sipina corresponds to the "Supervised Learning" item. We can also create a shortcut on the desktop.



3. Using Sipina under Linux



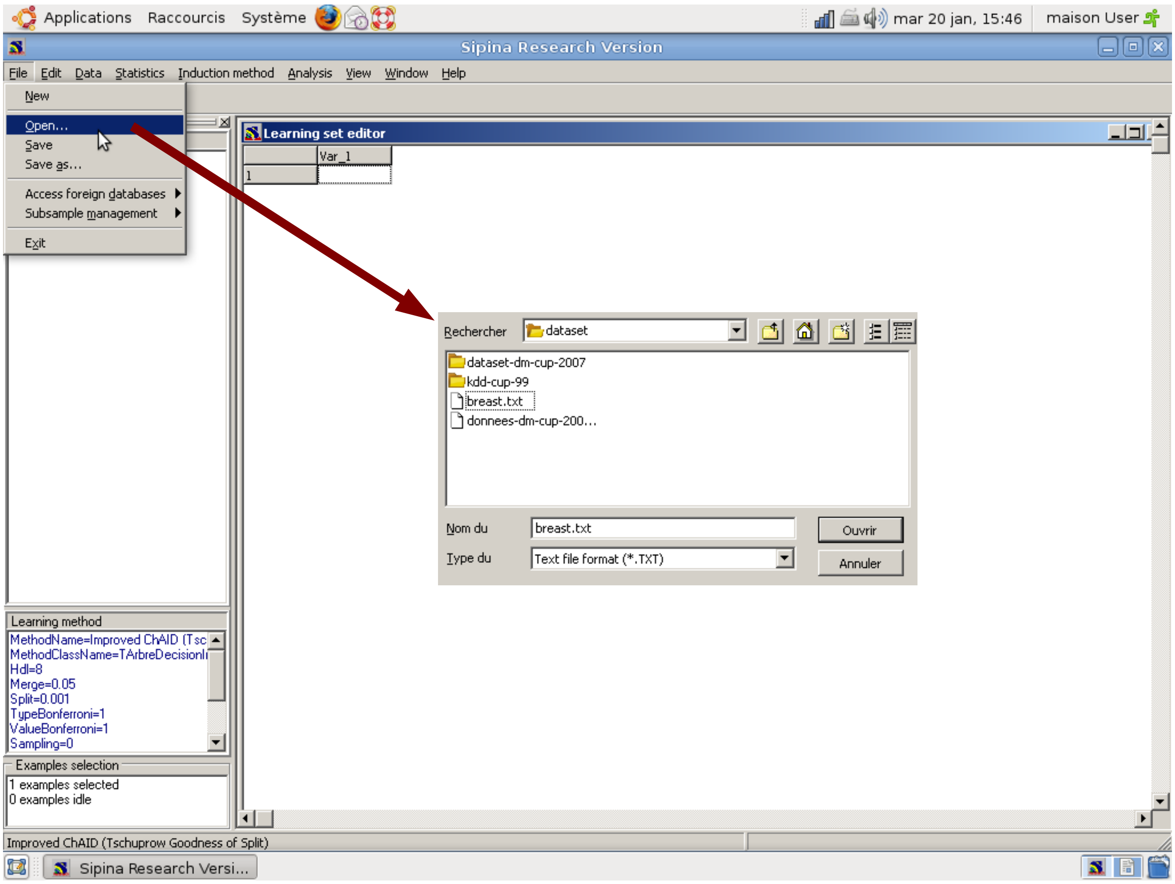
```
MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0
```

```
1 examples selected
0 examples idle
```

Improved ChAID (Tschuprow Goodness of Split)

When we start Sipina, we find the usual interface. Nothing distinguishes it from its running under Windows.

Data importation. We use the BREAST.TXT in this tutorial (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/breast.txt ; *Note: the decimal separator is always "." for Sipina, whatever your system configuration*). We want to predict the characteristics of cells extracted from a tumor ("malignant" or "benign") from their description (shape, size, etc.). To import the dataset, we click on the FILE/OPEN menu. We select the text file format.



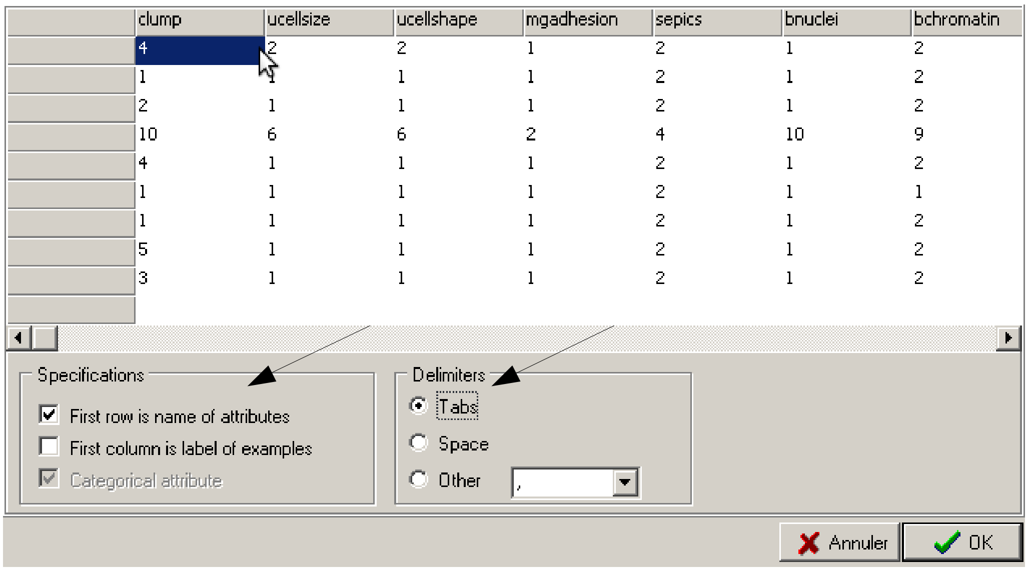
Nom du: breast.txt
Type du: Text file format (*.TXT)

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

1 examples selected
0 examples idle

Improved ChAID (Tschuprow Goodness of Split)

In the dialog box which appears, we set: the column separator and the variables name (in the first row).

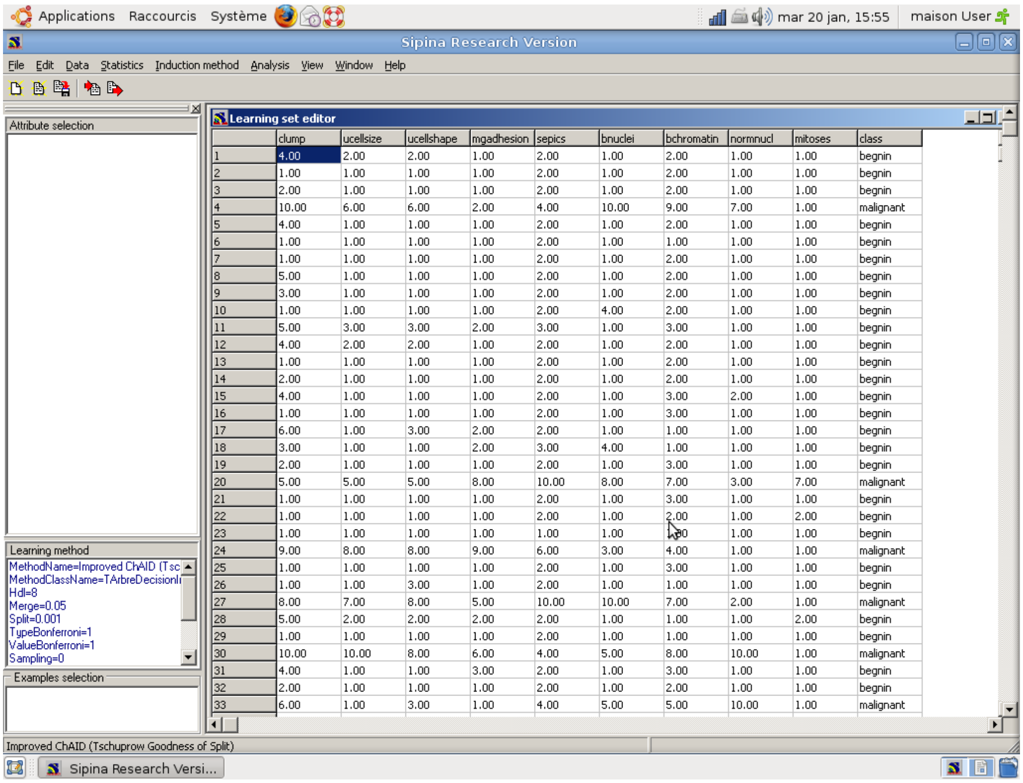


	clump	ucellsize	ucellshape	mgadhesion	sepics	bnuclei	bchromatin
	4	2	2	1	2	1	2
	1	1	1	1	2	1	2
	2	1	1	1	2	1	2
	10	6	6	2	4	10	9
	4	1	1	1	2	1	2
	1	1	1	1	2	1	1
	1	1	1	1	2	1	2
	5	1	1	1	2	1	2
	3	1	1	1	2	1	2

Specifications
☑ First row is name of attributes
☐ First column is label of examples
☑ Categorical attribute

Delimiters
● Tabs
○ Space
○ Other ,

The dataset is displayed in a grid.



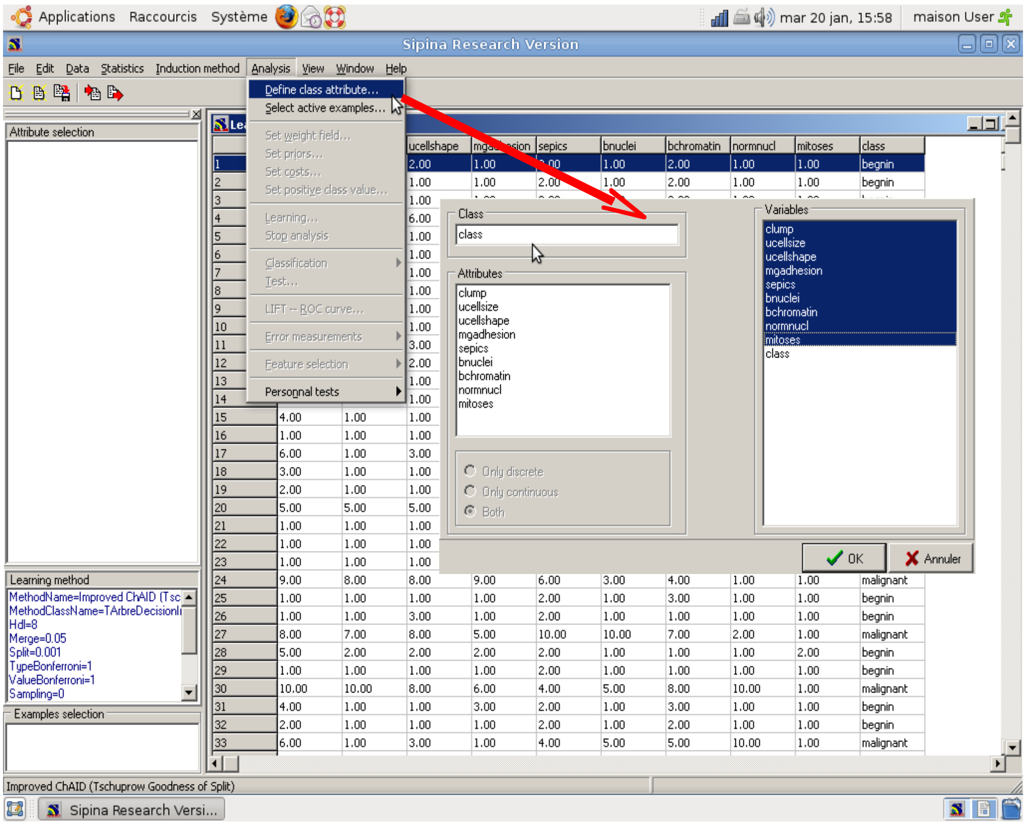
	clump	ucellsize	ucellshape	mgadhesion	sepics	bnuclei	bchromatin	normnucl	mitoses	class
1	4.00	2.00	2.00	1.00	2.00	1.00	2.00	1.00	1.00	begnin
2	1.00	1.00	1.00	1.00	2.00	1.00	2.00	1.00	1.00	begnin
3	2.00	1.00	1.00	1.00	2.00	1.00	2.00	1.00	1.00	begnin
4	10.00	6.00	6.00	2.00	4.00	10.00	9.00	7.00	1.00	malignant
5	4.00	1.00	1.00	1.00	2.00	1.00	2.00	1.00	1.00	begnin
6	1.00	1.00	1.00	1.00	2.00	1.00	1.00	1.00	1.00	begnin
7	1.00	1.00	1.00	1.00	2.00	1.00	2.00	1.00	1.00	begnin
8	5.00	1.00	1.00	1.00	2.00	1.00	2.00	1.00	1.00	begnin
9	3.00	1.00	1.00	1.00	2.00	1.00	2.00	1.00	1.00	begnin
10	1.00	1.00	1.00	1.00	2.00	4.00	2.00	1.00	1.00	begnin
11	5.00	3.00	3.00	2.00	3.00	1.00	3.00	1.00	1.00	begnin
12	4.00	2.00	2.00	1.00	2.00	1.00	2.00	1.00	1.00	begnin
13	1.00	1.00	1.00	1.00	2.00	1.00	2.00	1.00	1.00	begnin
14	2.00	1.00	1.00	1.00	2.00	1.00	2.00	1.00	1.00	begnin
15	4.00	1.00	1.00	1.00	2.00	1.00	3.00	2.00	1.00	begnin
16	1.00	1.00	1.00	1.00	2.00	1.00	3.00	1.00	1.00	begnin
17	6.00	1.00	3.00	2.00	2.00	1.00	1.00	1.00	1.00	begnin
18	3.00	1.00	1.00	2.00	3.00	4.00	1.00	1.00	1.00	begnin
19	2.00	1.00	1.00	1.00	2.00	1.00	3.00	1.00	1.00	begnin
20	5.00	5.00	5.00	8.00	10.00	8.00	7.00	3.00	7.00	malignant
21	1.00	1.00	1.00	1.00	2.00	1.00	3.00	1.00	1.00	begnin
22	1.00	1.00	1.00	1.00	2.00	1.00	2.00	1.00	2.00	begnin
23	1.00	1.00	1.00	1.00	1.00	1.00	[illegible]	1.00	1.00	begnin
24	9.00	8.00	8.00	9.00	6.00	3.00	4.00	1.00	1.00	malignant
25	1.00	1.00	1.00	1.00	2.00	1.00	3.00	1.00	1.00	begnin
26	1.00	1.00	3.00	1.00	2.00	1.00	1.00	1.00	1.00	begnin
27	8.00	7.00	8.00	5.00	10.00	10.00	7.00	2.00	1.00	malignant
28	5.00	2.00	2.00	2.00	2.00	1.00	1.00	1.00	2.00	begnin
29	1.00	1.00	1.00	1.00	2.00	1.00	1.00	1.00	1.00	begnin
30	10.00	10.00	8.00	6.00	4.00	5.00	8.00	10.00	1.00	malignant
31	4.00	1.00	1.00	3.00	2.00	1.00	3.00	1.00	1.00	begnin
32	2.00	1.00	1.00	1.00	2.00	1.00	2.00	1.00	1.00	begnin
33	6.00	1.00	3.00	1.00	4.00	5.00	5.00	10.00	1.00	malignant

Learning method

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

Improved ChAID (Tschuprow Goodness of Split)

Defining the status of the variables. In order to define the class attribute, we click on the ANALYSIS / DEFINE CLASS ATTRIBUTE menu. We use drag-and-drop to select the variables.



Class

class

Attributes

clump
ucellsize
ucellshape
mgadhesion
sepics
bnuclei
bchromatin
normnucl
mitoses

○ Only discrete
○ Only continuous
● Both

Variables

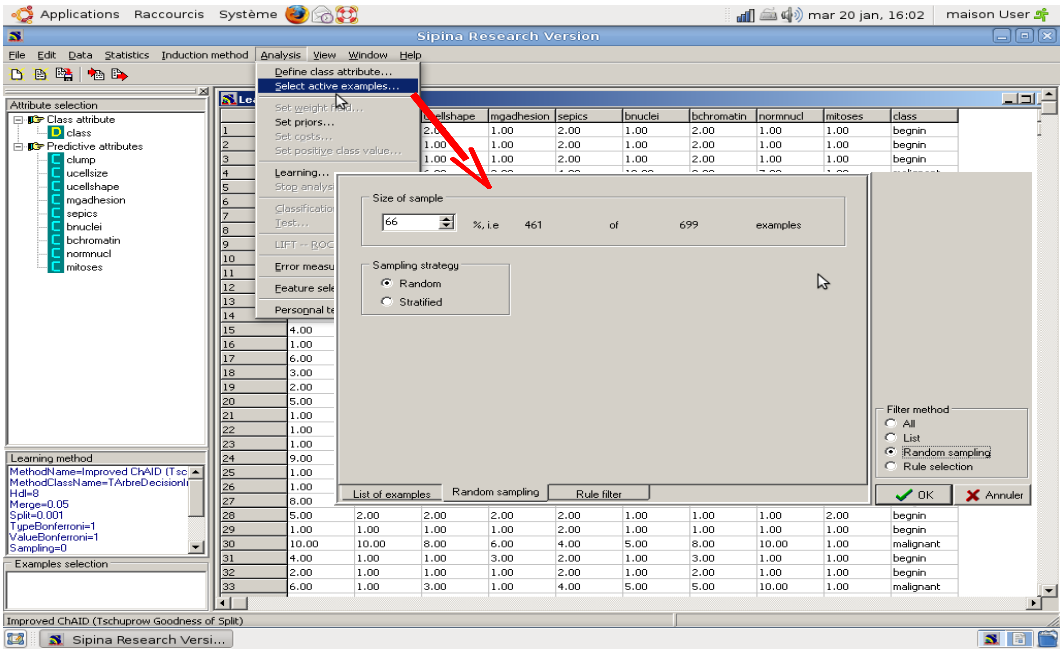
clump
ucellsize
ucellshape
mgadhesion
sepics
bnuclei
bchromatin
normnucl
mitoses
class

Learning method

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

Improved ChAID (Tschuprow Goodness of Split)

Partitioning the dataset. We want to use 66% of the dataset for the learning phase, the remainder (34%) for the test phase. We click on the ANALYSIS / SELECT ACTIVE EXAMPLES menu. Into the dialog box, we choose the RANDOM SAMPLING item: 66% of the dataset is selected i.e. 461 examples.



Size of sample: 66 %, i.e 461 of 699 examples

Sampling strategy
- ● Random
- ○ Stratified

Filter method
- ○ All
- ○ List
- ● Random sampling
- ○ Rule selection

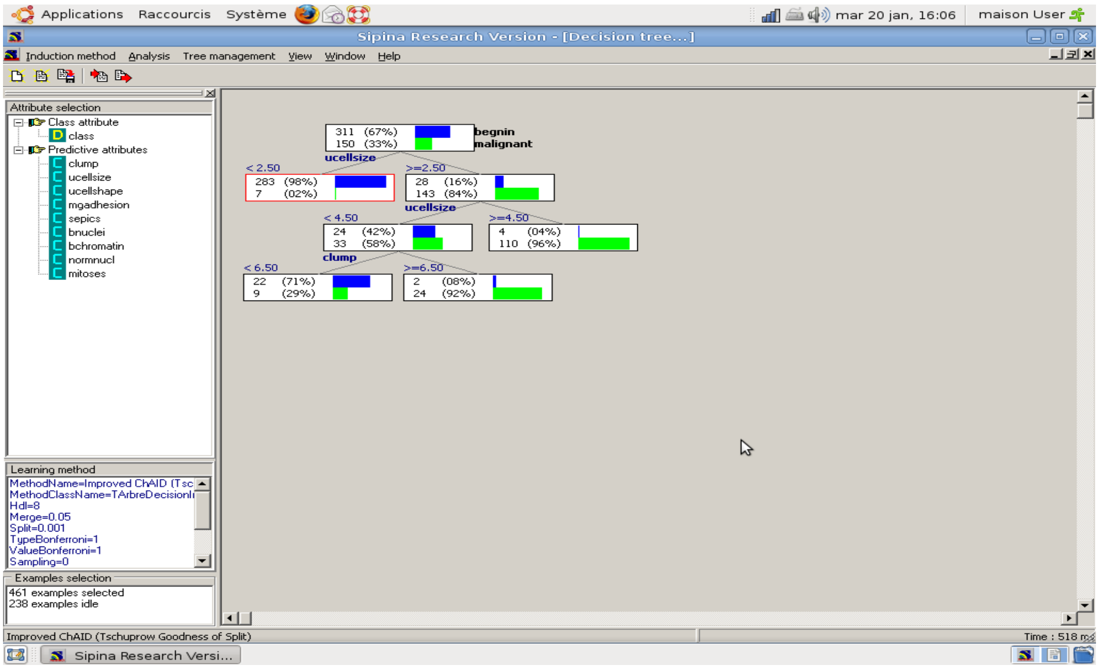
Learning method
MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

Examples selection

Improved ChAID (Tschuprow Goodness of Split)

	clump	ucellsize	ucellshape	mgadhesion	sepics	bnuclei	bchromatin	normnucl	mitoses	class
28	5.00	2.00	2.00	2.00	2.00	1.00	1.00	1.00	2.00	begnin
29	1.00	1.00	1.00	1.00	2.00	1.00	1.00	1.00	1.00	begnin
30	10.00	10.00	8.00	6.00	4.00	5.00	8.00	10.00	1.00	malignant
31	4.00	1.00	1.00	3.00	2.00	1.00	3.00	1.00	1.00	begnin
32	2.00	1.00	1.00	1.00	2.00	1.00	2.00	1.00	1.00	begnin
33	6.00	1.00	3.00	1.00	4.00	5.00	5.00	10.00	1.00	malignant

Creating the decision tree. We click on the ANALYSIS / LEARNING menu to start the analysis. The tree is displayed in a new window.



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

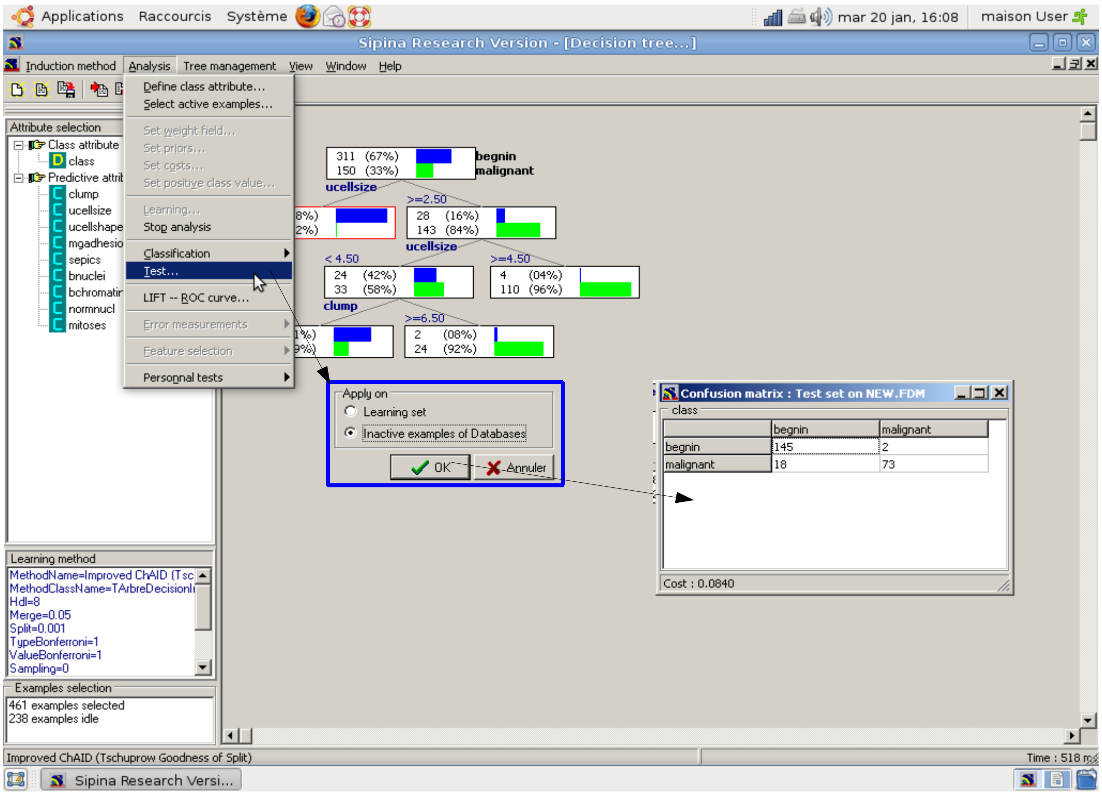
Examples selection
461 examples selected
238 examples idle

Improved ChAID (Tschuprow Goodness of Split)

***Note**: Because the data is randomly partitioned, perhaps we do not obtain the same tree.*

Evaluation on the test set. We want to evaluate the classifier performance on the remainder dataset i.e. 238 examples. We click on the ANALYSIS / TEST menu. In the dialog box, we select the “Inactive

Examples". The confusion matrix is displayed in a new window. The test error rate is 8.40%.



Apply on
○ Learning set
● Inactive examples of Databases
OK Annuler

Confusion matrix : Test set on NEW.FDM

class

	begnin	malignant
begnin	145	2
malignant	18	73

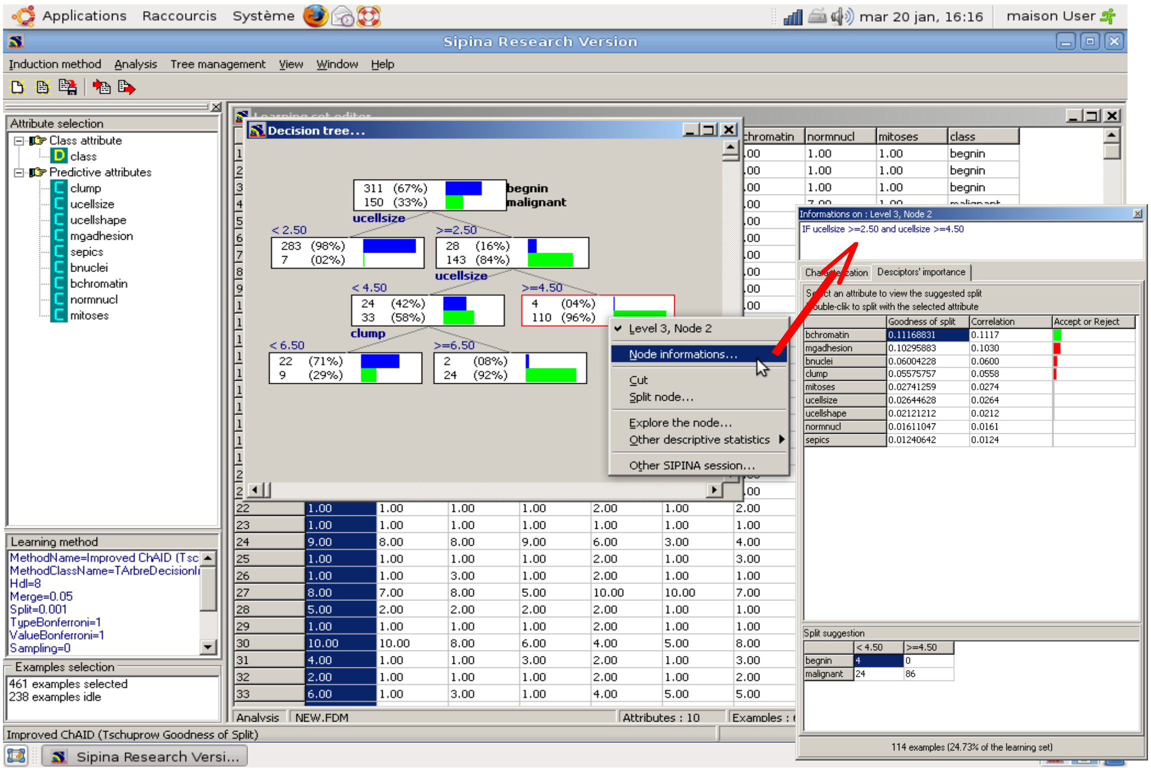
Cost : 0.0840

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

Examples selection
461 examples selected
238 examples idle

Improved ChAID (Tschuprow Goodness of Split)
Time : 518 ms

Exploration of a subpopulation. We want to highlight the characteristics of a subpopulation related to a node. We select the node; we click on the NODE INFORMATIONS item into the contextual menu. In the DESCRIPTORS IMPORTANCE tab, we have the goodness of split of each predictive variable.



chromatin	normnucl	mitoses	class
.00	1.00	1.00	begnin
.00	1.00	1.00	begnin
.00	1.00	1.00	begnin

22	1.00	1.00	1.00	1.00	2.00	1.00	2.00
23	1.00	1.00	1.00	1.00	1.00	1.00	1.00
24	9.00	8.00	8.00	9.00	6.00	3.00	4.00
25	1.00	1.00	1.00	1.00	2.00	1.00	3.00
26	1.00	1.00	3.00	1.00	2.00	1.00	1.00
27	8.00	7.00	8.00	5.00	10.00	10.00	7.00
28	5.00	2.00	2.00	2.00	2.00	1.00	1.00
29	1.00	1.00	1.00	1.00	2.00	1.00	1.00
30	10.00	10.00	8.00	6.00	4.00	5.00	8.00
31	4.00	1.00	1.00	3.00	2.00	1.00	3.00
32	2.00	1.00	1.00	1.00	2.00	1.00	2.00
33	6.00	1.00	3.00	1.00	4.00	5.00	5.00

Analysis NEW.FDM Attributes : 10 Examples :

Informations on : Level 3, Node 2

IF ucellsize >=2.50 and ucellsize >=4.50

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
bchromatin	0.11168831	0.1117	🟩
mgadhesion	0.10295883	0.1030	🟥
bnuclei	0.06004228	0.0600	🟥
clump	0.05575757	0.0558	🟥
mitoses	0.02741259	0.0274	
ucellsize	0.02644628	0.0264	
ucellshape	0.02121212	0.0212	
normnucl	0.01611047	0.0161	
sepics	0.01240642	0.0124	

Split suggestion

	< 4.50	>=4.50
begnin	4	0
malignant	24	86

114 examples (24.73% 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

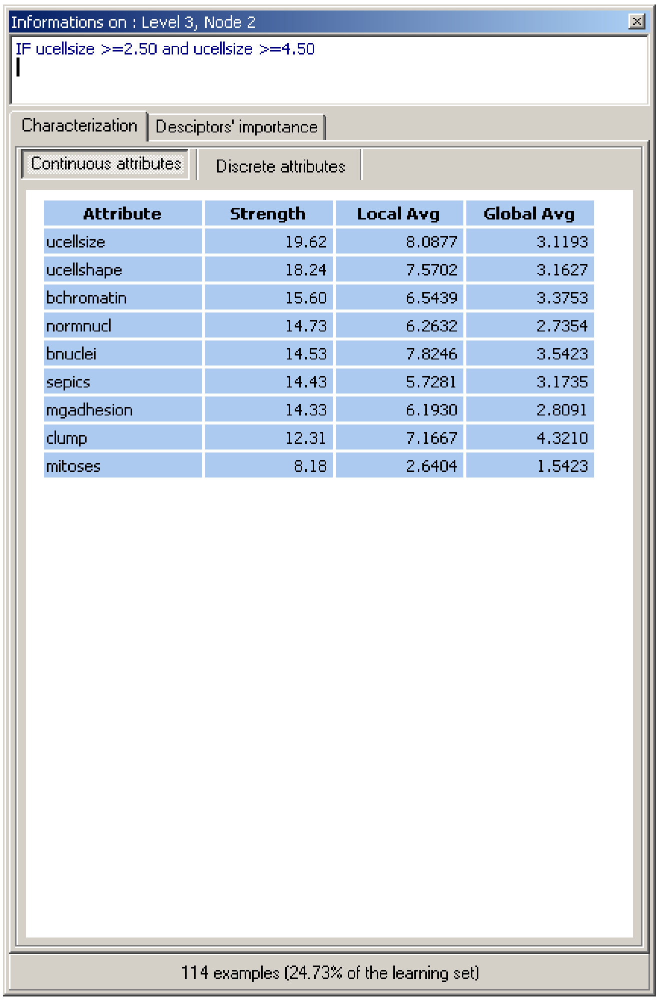
Examples selection
461 examples selected
238 examples idle

Improved ChAID (Tschuprow Goodness of Split)

Into the CHARACTERIZATION tab, we can compare the descriptive statistics computed on the whole population (the root of the tree) and the studied subpopulation (the current node of the tree).

For instance, we observe that UCELLSHAPE, which is not used in the tree, allows characterizing differently the current node: the local mean (7.57) of this variable is higher here than the global mean (3.16) computed on the whole dataset.

We observe that all the variable have significantly higher values into this subpopulation.



IF ucellsize >=2.50 and ucellsize >=4.50

Attribute	Strength	Local Avg	Global Avg
ucellsize	19.62	8.0877	3.1193
ucellshape	18.24	7.5702	3.1627
bchromatin	15.60	6.5439	3.3753
normnucl	14.73	6.2632	2.7354
bnuclei	14.53	7.8246	3.5423
sepics	14.43	5.7281	3.1735
mgadhesion	14.33	6.1930	2.8091
clump	12.31	7.1667	4.3210
mitoses	8.18	2.6404	1.5423

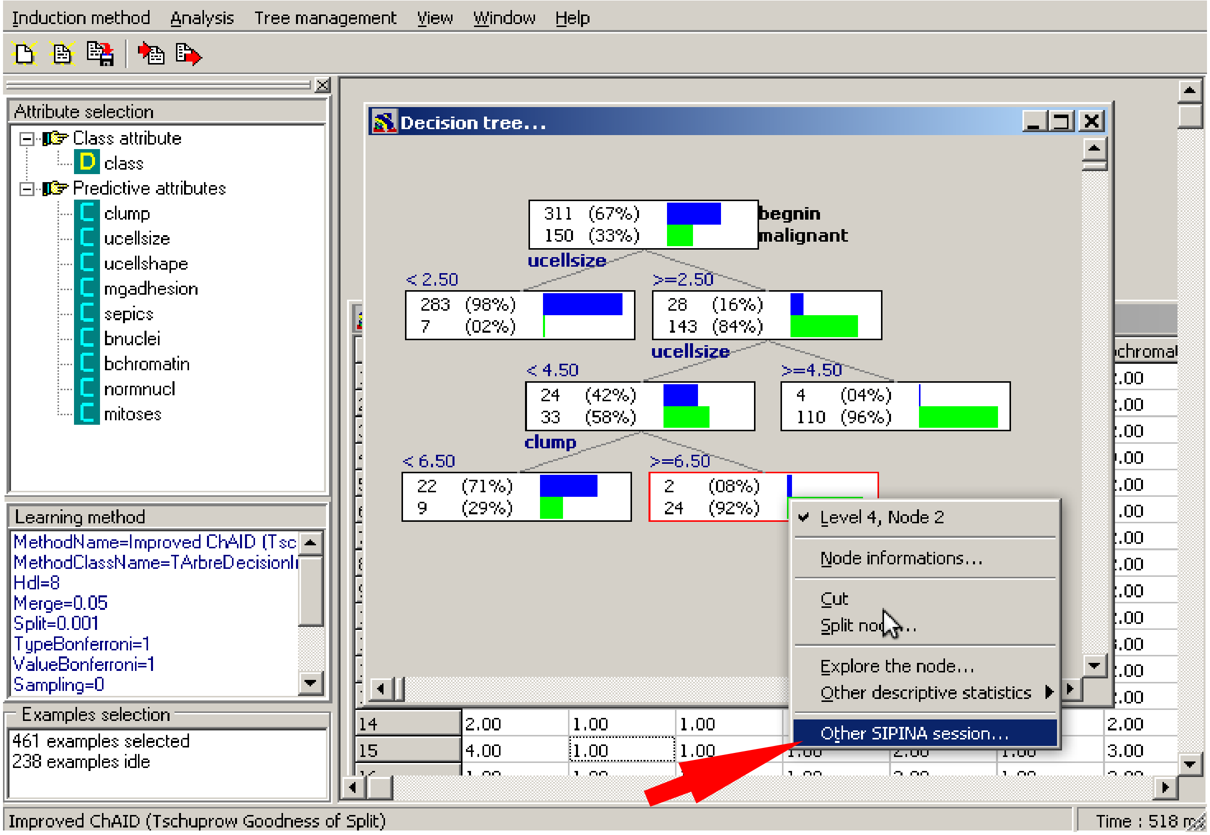
114 examples (24.73% of the learning set)

4. Exploring more deeply a subpopulation

Sometimes, we want to explore mode deeply the node by launching a new analysis on the related subset of examples. For instance, we want to study the subset of examples (26 cases) described as follows:

« UCELLSIZE >= 2.5 **and** UCELLSIZE < 4.5 **and** CLUMP >= 6.5 »

With the contextual menu, we click on the OTHER SIPINA SESSION item. A new SIPINA session is started and the corresponding subset of examples is automatically loaded.



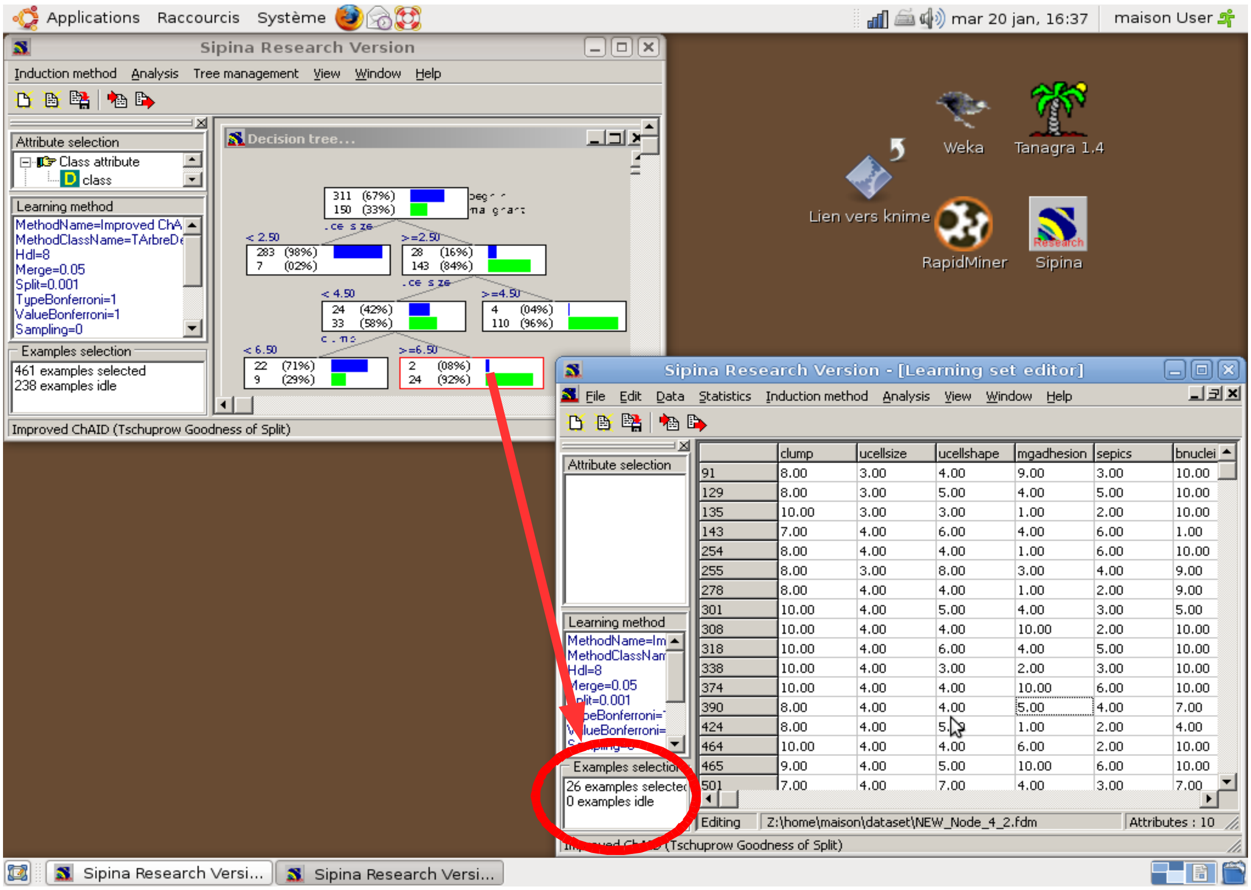
MethodName=Improved ChAID (Tsc
MethodClassName=TArbreDecisionI
Hdl=8
Merge=0.05
Split=0.001
TypeBonferroni=1
ValueBonferroni=1
Sampling=0

461 examples selected
238 examples idle

Improved ChAID (Tschuprow Goodness of Split)

Time : 518 ms

All the variables, but only the selected examples, are sent. We can start now a new analysis.



	clump	ucellsize	ucellshape	mgadhesion	sepics	bnuclei
91	8.00	3.00	4.00	9.00	3.00	10.00
129	8.00	3.00	5.00	4.00	5.00	10.00
135	10.00	3.00	3.00	1.00	2.00	10.00
143	7.00	4.00	6.00	4.00	6.00	1.00
254	8.00	4.00	4.00	1.00	6.00	10.00
255	8.00	3.00	8.00	3.00	4.00	9.00
278	8.00	4.00	4.00	1.00	2.00	9.00
301	10.00	4.00	5.00	4.00	3.00	5.00
308	10.00	4.00	4.00	10.00	2.00	10.00
318	10.00	4.00	6.00	4.00	5.00	10.00
338	10.00	4.00	3.00	2.00	3.00	10.00
374	10.00	4.00	4.00	10.00	6.00	10.00
390	8.00	4.00	4.00	5.00	4.00	7.00
424	8.00	4.00	5.00	1.00	2.00	4.00
464	10.00	4.00	4.00	6.00	2.00	10.00
465	9.00	4.00	5.00	10.00	6.00	10.00
501	7.00	4.00	7.00	4.00	3.00	7.00

Editing Z:\home\maison\dataset\NEW_Node_4_2.fdm Attributes : 10

26 examples selected
0 examples idle

Improved ChAID (Tschuprow Goodness of Split)

5. Conclusion

In this tutorial, we showed that it is possible to install and use Sipina in the Linux environment without the need for specific technical skill. We then have the whole environment of the software which is fully functional.