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-line¹. 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.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Category</th>
<th>Informations</th>
</tr>
</thead>
<tbody>
<tr>
<td>bpress_level</td>
<td>Discrete</td>
<td>2 values</td>
</tr>
<tr>
<td>gender</td>
<td>Discrete</td>
<td>2 values</td>
</tr>
<tr>
<td>smoke</td>
<td>Discrete</td>
<td>2 values</td>
</tr>
<tr>
<td>exercise</td>
<td>Continue</td>
<td>-</td>
</tr>
<tr>
<td>overweight</td>
<td>Continue</td>
<td>-</td>
</tr>
<tr>
<td>alcohol</td>
<td>Continue</td>
<td>-</td>
</tr>
<tr>
<td>stress</td>
<td>Continue</td>
<td>-</td>
</tr>
<tr>
<td>salt</td>
<td>Continue</td>
<td>-</td>
</tr>
<tr>
<td>income</td>
<td>Continue</td>
<td>-</td>
</tr>
<tr>
<td>education</td>
<td>Continue</td>
<td>-</td>
</tr>
</tbody>
</table>

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](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](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](http://www.math.yorku.ca/Who/Faculty/Ng/ssc2003/BPMain.htm)
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.

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.
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.

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

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

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.

<table>
<thead>
<tr>
<th>Rule (If Premise Then Conclusion)</th>
<th>Confidence</th>
<th>Lift</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>If overweight (\geq) 2.5 Then Blood pressure = high</td>
<td>71%</td>
<td>1.25</td>
<td>40%</td>
</tr>
<tr>
<td>If overweight &lt; 2.5 ET Exercise &lt; 1.5 Then Blood Pressure = high</td>
<td>62%</td>
<td>1.08</td>
<td>24%</td>
</tr>
<tr>
<td>If overweight &lt; 2.5 ET Exercise (\geq) 1.5 Then Blood pressure = low</td>
<td>61%</td>
<td>1.40</td>
<td>36%</td>
</tr>
</tbody>
</table>

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

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.

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 ...

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.
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.
The expert, according to 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.

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.).

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:

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.
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).

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.
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.
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 smokers. The threshold is higher (> 2.5) for the whole population. For non smokers, the cut point (> 2.5) is the same one.

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

**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.