First order decision tree for the Predictive Toxicology Challenge 2001

Hendrik Blockeel, Kurt Driessens, Nico Jacobs,
Raymond Kosala, Stefan Raeymaekers, Jan Ramon,
Jan Struyf, Wim Van Laer, Sofie Verbaeten

Katholieke Universiteit Leuven, Department of Computer Science
Celestijnenlaan 200A, B-3001 Leuven, Belgium

July 24, 2001

Abstract

This report describes the models obtained by the Machine Learning group of the Computer Science department of the Katholieke Universiteit Leuven related to the Predictive Toxicology Challenge 2001. More information about how the results were obtained can be found in the full paper submitted for this challenge.

1 Resulting model

From our submissions to the PTC challenge, five models were selected by the organisers with a ROC analysis. In this report we describe these models and how they have to be read. More information about how we obtained these results can be found in the paper submitted to the PTC workshop.

The model we submit is the decision tree in figure 5. This should be interpreted as follows: each node in an oval (or round corner rectangle) is a test, and each node (leaf) in a rectangle (containing 4 numbers) are 4 activity levels (one for each animal and gender:female mice, male mice, female rats and male rats) — the higher this value, the more active this type of molecules is and if the value is above the activity threshold, we consider the molecule positive (active, carcinogenic) . You read the tree by starting at the main node `atom(A,br)`. This is a test, specifying "is there — in the molecule we are looking at — at least one atom A that is an bromine atom?". If the answer to this question is yes, we take the left branch, otherwise the right branch. Suppose there is a bromine atom and so we take the left branch. We then arrive in the rectangle with four numbers. These numbers are the predicted activity levels for respectively female mice, male mice, female rats and male rats. If these numbers are above the threshold (for that specific animal and gender) we predict that the

```
if contains bromine
then if FMthreshold<=0.4265
     then POS
     else NEG
else if contains oxygen
     then if oxygen is bond to hydrogen
          then if FMthreshold<=-0.3095
               then POS
               else NEG
          else if contains sulphur
               then if FMthreshold<=-0.4621
                    then POS
                    else NEG
               else if FMthreshold<=-0.0108
                    then POS
                    else NEG
     else if contains chlorine
          then if FMthreshold<=-0.2414
               then POS
               else NEG
          else if FMthreshold<=-0.0231
               then POS
               else NEG
```

Figure 1: Decision tree for female mice

element will be active (carcinogenic), otherwise we predict that the element will be inactive (not carcinogenic). The difference between the different models we submitted is the threshold level for the different species and gender (see figure 6 for the exact values used in our different models).

A last aspect that is important when reading the decision tree is that it is a relational decision tree: it can express relations between tests. Suppose for instance that there is a molecule which does contain no bromine atom, but does contain an oxygen atom. We then arrive in the node with as test `bond(B,C),atom(C,h)`. This test must be read as: "is there at least one atom C that has a bond with the atom B, such that the C atom is an hydrogen atom?". Note that the B was used in the previous test (the oxygen test) as well, and so this refers to this atom as well. As a result, a molecule can only end up in the rectangle left from this node if it does not contain a bromine, but does contain at least one oxygen atom that is bond to an hydrogen atom.

So the tree can be read as 4 sets of nested if-then-else rules, one for each animal and gender. Such a tree for female mice is shown in figure 1.

When we analyse the model we see that the activity level of the bromine test is so high we always predict POS when a bromine is present. The other

```
if contains bromine
then POS
else NEG
```

Figure 2: Pruned tree for female mice and male mice (model 1)

```
if contains bromine
then POS
else if contains oxygen
     then if oxygen is bond to hydrogen (e.g. alcohol, ..)
          then NEG
          else if contains sulphur
               then NEG
               else POS
     else POS
```

Figure 3: Pruned tree for male mice (model 2)

rectangles have lower values, so that dependent on the threshold more molecules are predicted positive if the threshold drops. We see that the presence of sulphur when oxygen is present but not bound to hydrogen, results in lower activity levels, indicating that it reduces the chance that the molecule is active (carcinogenic).

We can prune this tree by taking the actual thresholds (figure 6) into account. These pruned trees contain less information than the tree, but are easier to read (and classify the examples in the same way). In this way we obtain 3 trees: figure 2 for male and female mice (model 1), figure 3 for male mice (model 2) and figure 4 for male and female rats (model 3).

```
if contains bromine
then POS
else if contains oxygen
     then if oxygen is bond to hydrogen
          then POS
          else if contains sulphur
               then NEG
               else POS
     else POS
```

Figure 4: Pruned tree for female rats and male rats (model 3)

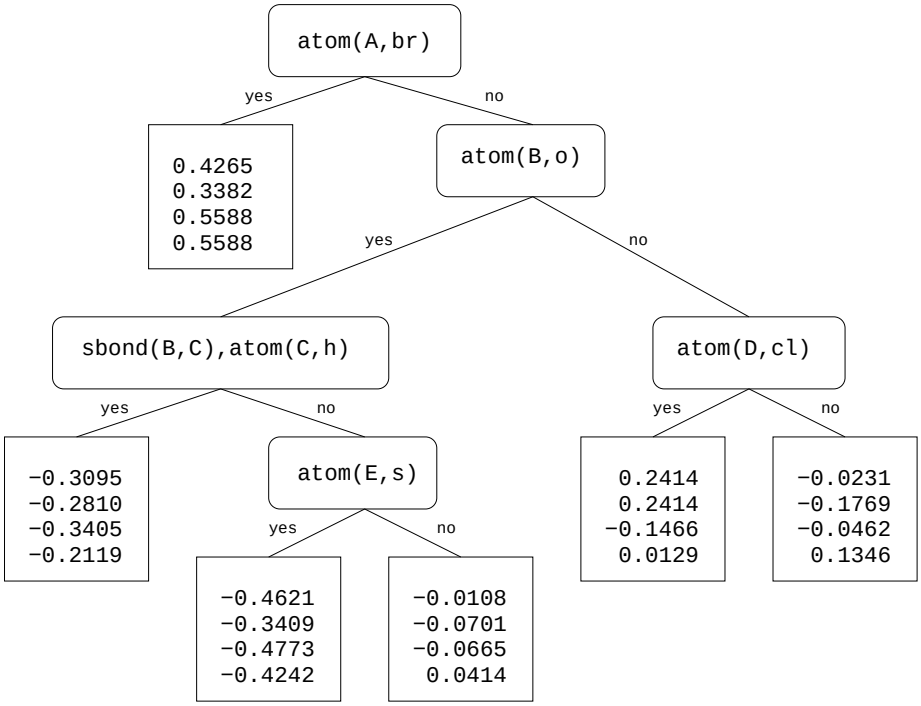

Figure 5: Decision tree for female mice, male mice, female rats and male rats.

	Model 1		Model 2		Model 3
FM	0.4265			FR	-0.3405
MM	0.3382	MM	-0.1769	MR	-0.2119

Figure 6: Thresholds for activity levels for female mice (FM), male mice (MM), female rats (FR) and male rats (MR)