ID3 Algorithm for Building Decision Trees

Initial tree: single node containing all original data samples

Until each leaf node has samples whose value of the target attribute is as homogeneous as possible:

- Select a leaf node with an inhomogeneous sample set
- Replace leaf with a test node that divides the sample set into subsets with minimal disorder, using the following measure of disorder:

$$\text{Average disorder} = \sum_{b}\left(\frac{n_b}{n_t}\right) \times \left(\sum_{c} -\frac{n_{bc}}{n_b}\log_2 \frac{n_{bc}}{n_b}\right)$$

The *Reality* of the Real World…

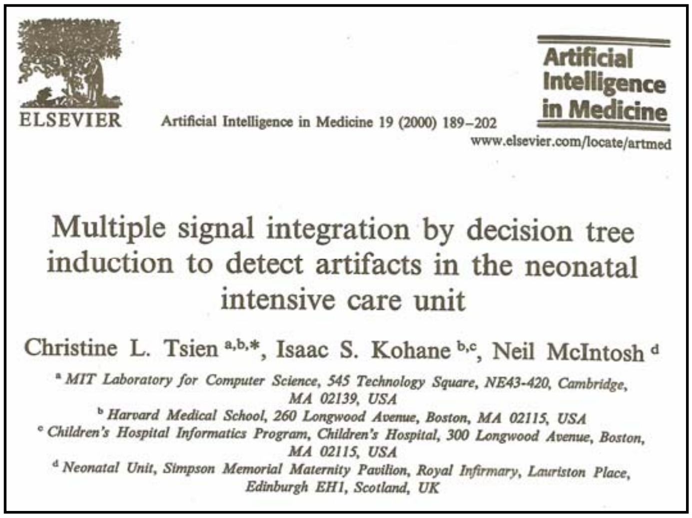
- ***Data may be incorrect*** due to errors in attribute values
- ***Data may be incomplete*** – may be impossible or infeasible to measure every attribute for every data sample
- ***Domain may be nondeterministic*** – two or more samples with the same attributes can belong to different classes

$\longrightarrow$ **May not be possible to reach a perfect classification of the data**

When *leaves* of the decision tree *contain multiple classes,* with *no more attributes* to distinguish the samples:

- Label the set with the dominant class
- Record the fraction of samples for each class

Medical Decision Making



ELSEVIER

Artificial Intelligence in Medicine 19 (2000) 189–202

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Multiple signal integration by decision tree induction to detect artifacts in the neonatal intensive care unit

Christine L. Tsien [a,b,*], Isaac S. Kohane [b,c], Neil McIntosh [d]

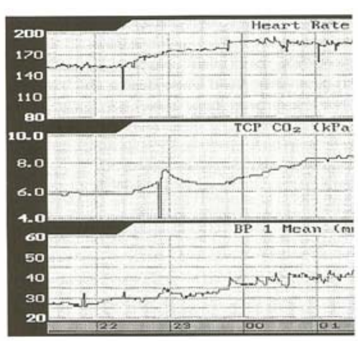
[a] MIT Laboratory for Computer Science, 545 Technology Square, NE43-420, Cambridge, MA 02139, USA

[b] Harvard Medical School, 260 Longwood Avenue, Boston, MA 02115, USA

[c] Children's Hospital Informatics Program, Children's Hospital, 300 Longwood Avenue, Boston, MA 02115, USA

[d] Neonatal Unit, Simpson Memorial Maternity Pavilion, Royal Infirmary, Lauriston Place, Edinburgh EH1, Scotland, UK

Medical Data



Physiological properties:

- Heart rate
- Blood pressure
- Partial pressures of
  - carbon dioxide
  - oxygen

Statistics measured within moving temporal windows:

- mean, median, slope, standard deviation, min/max values, range

Expert clinician assigned class labels: artifact, not artifact, transition

Data divided into training set and test set

Example Decision Tree

```
hr_low3 <= 113 :
|   hr_range5 > 78 : 1 (180.0/3.0)
|   hr_range5 <= 78 :
|   |   hr_low10 <= 30 : 1 (60.0/1.0)
|   |   hr_low10 > 30 :
|   |   |   hr_med5 <= 121 : 0 (145.0/21.0)
|   |   |   hr_med5 > 121 :
|   |   |   |   ox_low10 > 6 : 1 (41.0)
|   |   |   |   ox_low10 <= 6 :
|   |   |   |   |   hr_range3 <= 38 : 0 (30.0/10.0)
|   |   |   |   |   hr_range3 > 38 : 1 (37.0/8.0)
hr_low3 > 113 :
|   hr_std_dev3 <= 8.14 : 0 (10565.0/66.0)
|   hr_std_dev3 > 8.14 :
|   |   hr_range3 > 36 : 1 (41.0/6.0)
|   |   hr_range3 <= 36 :
|   |   |   hr_low5 > 129 : 0 (160.0/28.0)
|   |   |   hr_low5 <= 129 :
|   |   |   |   ox_low10 <= 4 : 0 (35.0/6.0)
|   |   |   |   ox_low10 > 4 :
|   |   |   |   |   bp_abs_slope10 <= 0.21 : 1 (38.0/5.0)
|   |   |   |   |   bp_abs_slope10 > 0.21 : 0 (37.0/15.0)
```

Fig. 2. Heart rate artifact detection: decision tree model.

Final class labels: 1 (artifact), 0 (no artifact)

(# samples in leaf / # samples in minority class)

Training Procedure

Approach 1: use all training data to construct decision tree, test with new data

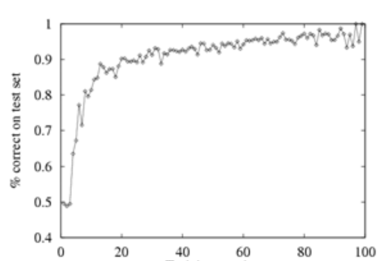
Approach 2: divide data into subsets for training and testing

(1) Use the training set to construct a decision tree

(2) Measure percentage of samples in the test set that are correctly classified using the decision tree

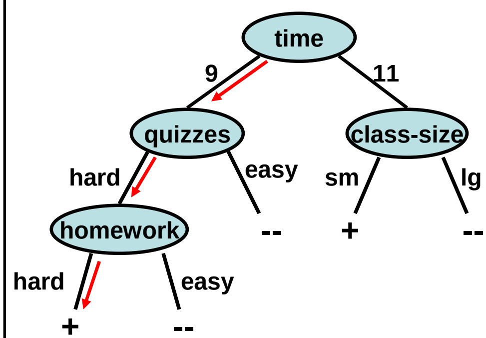
Option a: if misclassified set is small, add it to the training set and repeat steps (1) & (2)

Option b: repeat steps (1) & (2) with different sizes of the training set and different randomly selected samples for training set



From Decision Trees to Rules

Liked?	Lecturer	Homework	Time	Quizzes	Class Size
Yes	Good	Hard	11:00	Easy	Small
Yes	Bad	Hard	11:00	Hard	Small
Yes	Good	Easy	11:00	Easy	Small
Yes	Bad	Easy	11:00	Hard	Small
Yes	Good	Hard	9:00	Hard	Large
No	Bad	Easy	11:00	Hard	Large
No	Good	Hard	9:00	Easy	Small
No	Bad	Hard	9:00	Easy	Large
No	Bad	Hard	9:00	Easy	Small
No	Good	Easy	9:00	Hard	Large



Rule:

if time = 9
quizzes = hard
homework = hard
then liked

Simplified rule:

if quizzes = hard
homework = hard
then liked

Trimming rules further…

if quizzes = hard
homework = hard
then liked

if time = 11
class-size = small
then liked

Default rule:
if no other rule applies
then not liked

~~**if time = 9**~~
~~**quizzes = hard**~~
~~**homework = easy**~~
~~**then not liked**~~

~~**if time = 9**~~
~~**quizzes = easy**~~
~~**then not liked**~~

~~**if time = 11**~~
~~**class-size = large**~~
~~**then not liked**~~

Before constructing rules, we can ***prune the decision tree***