Pruning Decision Trees: the settings

A pruned decision tree

```
26
20 : 10 : 5
```

```
27
15 : 5 : 0
```

```
28
3 : 3 : 5
```

```
29
2 : 2 : 0
```

```
30
15 : 2 : 0
```

```
31
0 : 3 : 0
```

Domain Dependent Pruning

DATA + BACKGROUND KNOWLEDGE

- cost; weights of attributes
- expert specified attribute-dependencies

Domain Independent Pruning

... because "uncertainty" influences given data

source of uncertainty:

- Incorrect perception, measurements ...
- NOISE
- Extraneous factors

CONTINUE SPLITING or NOT?

decide (prune) to stop and assign majority class (C3: 5)

distribution of examples to classes

class assigned
Error Rate of a node:

\[ r(t) = \frac{\text{# of examples misclassified in node}}{\text{# of all examples in node}} \]

Probability of Occurrence of a node:

\[ p(t) = \frac{\text{# of examples in node}}{\text{# total examples}} \]

Error Cost of a node:

\[ R(t) = r(t) \times p(t) = \frac{\text{# of examples misclassified in node}}{\text{# total examples}} \]

If node \( t \) was not pruned then error cost of subtree, \( T \) rooted at \( t \):

\[ R(T) = \sum_{i \text{ in } T_t} R(i) \]

Error Complexity

\[ a(t) = \frac{R(t) - R(T)_{t}}{\text{# leaves} - 1} \]
Error-Complexity pruning: operation

a: measures the VALUE of corresponding subtree

Forming Pruned Trees with Error-Complexity

1. a is computed for each node
2. the minimum a node is pruned
3. the above is repeated and a FOREST of PRUNED TREES is formed
4. the tree with BEST ACCURACY is selected

... need for a set of test cases

MANY TREES ARE FORMED: computational cost

Example:

\[
\begin{align*}
R(26) &= \frac{15}{35} \times \frac{35}{200} = \frac{15}{200} \\
R(T_{26}) &= \frac{2}{17} \times \frac{17}{200} + 0 + \frac{6}{11} \times \frac{11}{200} + \frac{2}{4} \times \frac{4}{200} = \frac{10}{200} \\
\text{# leaves for subtree}_{T_{26}} &= 4 \\
a(26) &= \frac{15/200 - 10/200}{4 - 1} = \frac{5}{600}
\end{align*}
\]
Minimum Error Pruning

(Niblett & Bratko, 1986)

\[ k: \text{# of classes} \]
\[ n_t: \text{# of examples in node } t \]
\[ n_{t,c}: \text{# of examples assigned to class } c \text{ in node } t \]

**intuition:**

*If we predict that all future examples are in class } c \text{ what is the proposition of wrong classifications?*}

**Expected error rate of pruning in node } t**

\[ E(t) = \frac{n_t - n_{t,c} + k - 1}{n_t + k} \]
Minimum Error Pruning: operation

Go to the leaves of the sub-tree rooted in a node t and compute weighted Error Rates for each of the leaves. If error rate of not pruning is less than the corresponding error rate when pruning then DONT PRUNE otherwise PRUNE

Example:

\[ E(27) = \frac{(20 - 15 + 3 - 1)}{(20 + 3)} = 0.304 \]

\[ E(27) = \frac{(17/20) \left[ \frac{(17-15+3-1)}{(17+3)} \right] + (3/20)\left[\frac{(3-3+3-1)}{(3+3)}\right]} = 0.220 \]

\[ E(27) > \overline{E(27)} \]

DONT PRUNE
Other Pruning approaches

► Convert decision trees to a set of rules
  (Quinlan, 1987)
  Convert each (positive) path to a rule, then prune redundant rules and tests

► Constructive induction
  (Matheus & Rendell, 1989; Pagallo, 1989)
  After generating a tree, select logical combinations along paths, add them to the set of attributes, and regenerate. Repeat until no new ones generated

► Multiple attributes per test
  (Breiman, 1984; Utgoff & Brodley, 1990; Salzberg et al., 1993)
  ... these approaches tend to have greatly increased complexities ....
**Selection/Spliting Measures in Decision Trees**

*An example domain: recur/not recur of breast cancer*

<table>
<thead>
<tr>
<th>RADIATION</th>
<th>MENOPAUSE</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>&lt;60</td>
<td>R</td>
</tr>
<tr>
<td>N</td>
<td>&gt;=60</td>
<td>R</td>
</tr>
<tr>
<td>N</td>
<td>&lt;60</td>
<td>R</td>
</tr>
<tr>
<td>N</td>
<td>NOT</td>
<td>R</td>
</tr>
<tr>
<td>Y</td>
<td>&gt;=60</td>
<td>NR</td>
</tr>
<tr>
<td>Y</td>
<td>&lt;60</td>
<td>NR</td>
</tr>
<tr>
<td>Y</td>
<td>&gt;=60</td>
<td>NR</td>
</tr>
<tr>
<td>N</td>
<td>NOT</td>
<td>NR</td>
</tr>
<tr>
<td>N</td>
<td>&lt;60</td>
<td>NR</td>
</tr>
<tr>
<td>N</td>
<td>&lt;60</td>
<td>R</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RADIATION</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>R</td>
</tr>
<tr>
<td>N</td>
<td>NR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MENOPAUSE</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;=60</td>
<td>R</td>
</tr>
<tr>
<td>&lt;60</td>
<td>NR</td>
</tr>
<tr>
<td>NOT</td>
<td>NR</td>
</tr>
</tbody>
</table>

**BETER:** values are more unequal distributed into classes

**SELECTION/SPLITING MEASURES TRIES TO CAPTURE and MEASURE SUCH DISTRIBUTIONS**
**The concept of a Contingency Matrix**

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>Cc</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A1</strong></td>
<td>X11</td>
<td>X12</td>
<td>....</td>
<td>X1c</td>
</tr>
<tr>
<td><strong>A2</strong></td>
<td>X21</td>
<td>X22</td>
<td>....</td>
<td>X2c</td>
</tr>
<tr>
<td><strong>Ar</strong></td>
<td>Xr1</td>
<td>Xr2</td>
<td>....</td>
<td>Xrc</td>
</tr>
<tr>
<td><strong>X.1</strong></td>
<td>X.1</td>
<td>X.2</td>
<td>....</td>
<td>X.c</td>
</tr>
</tbody>
</table>

**Xij = # examples of class j in which Ai occur**

We are looking metrics to measure the association power between classes and attributes.

George Potamias

PRUNE_SELECT_MEASURE 8/14
\( x^2(\text{chi-square}) \) selection measure

\[
\sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}
\]

From contingency matrix:

\[
\begin{align*}
\text{Observed} & : X_{ij} \\
\text{Expected} & : \frac{X_i \times X_j}{N}
\end{align*}
\]

\[
x^2(A_i) = \frac{(X_{ij} - E_{ij})^2}{E_{ij}}
\]

e.g.

RADIATION: ........ 4.29 \( \rightarrow \) it is selected
MENOPAUSE: ...... 0.53
G (Joint Entropy) selection measure

G: computes the JOINT ENTROPY of an attribute & a class

\[ G = E(A_i, C_j) \]

Minger, 1987:

G measure reduces to Information Gain measure (Quinlan’s)

\[ G(A_i) = 2N \times \text{Information\_gain}(A_i) \]

N: # of examples

e.g.

RADIATION: \( 2 \times 10 \times 0.27436 = 5.49 \)
MENOPAUSE: \( 2 \times 10 \times 0.02740 = 0.548 \)

it is selected
GINI index selection measure

GINI measure captures the IMPURITY of an attribute with respect to classes. It is a measure of DIVERSITY

Form contingency matrix:

GINI(total) = GINI(Class_Totals) = 1-(X.1/N) - (X.2/N) - ... - (X.c/N)

GINI(Ai) = GINI(Ai_row) = 1-(Xi1/N) - (Xi2/N) - ... - (Xic/N)

A take values A1, A2, ..., Am

\[
\text{GINI}(A) = \text{GINI(totals)} - \sum_{i=1}^{m} \frac{X_i}{N} * \text{GINI}(A_i)
\]

e.g.

RADIATION: ........ 0.215
MENOPAUSE: ....... 0.027

it is selected
**GAIN RATIO** selection measure

Quinlan)

GAIN_RATIO is a modification on the original INFORMATION GAIN measure which takes in consideration the distribution of attribute values

...the less unequal the distribution the less informative the attribute ....

*an Information Theoretic measure*

Form contigency matrix

\[
\text{Information Value}(A) = - \sum_{i=1}^{|A|} \frac{X_i}{N} \log \frac{X_i}{N}
\]

Favors atts with small range of values

\[
\text{GAIN RATIO}(A) = \frac{\text{IG}(A)}{\text{IV}(A)}
\]

e.g.

RADIATION: ....... 0.45  \[\text{it is selected}\]
MENOPAUSE: ....... 0.026

George Potamias  PRUNE_SELECT_MEASURE 12/14
ECONOMIC INDUCTION selection measure  
(Nunez, 1988)

ECONOMIC INDUCTION measure assigns a weight to an attribute related to the cost of it. The cost of an attribute is a function specified by domain (in)dependent factors.

specified by economic/statistical models and experts in the field

\[ f(\text{cost}(A)) \]

\[
\text{INFORMATION COST}(A) = \frac{f(\text{cost}(A))}{\text{IG}(A) }
\]

Quinlan’s Information Gain

e.g.  
if we assign a COST to RADIATION: 50 and a COST to MENOPAUSE: 5

\[
\text{RADIATION: } \frac{50}{0.27} = 182
\]
\[
\text{MENOPAUSE: } \frac{5}{0.27} = 18
\]

it is selected because has less information cost
Criteria for Evaluating Selection Measures

- **TREE SIZE**

  \[ \text{# of leaves} \Rightarrow \text{# of rules} \]

  specificity/generality of paths/rules

  Occam’s Razor principle:
  
  **THE FEWER THE TERMS IN A MODEL THE BETTER**

- **PREDICTIVE ACCURACY**

  measure error rate

- **COMPREHENSIBILITY**

  extract knowledge from decision trees
  
  DIFFICULT to access