Decision Trees
Implementierungstechniken für Hauptspeicherdatenbanksysteme
Dominik Vinan
Outline

- Introduction
- CART/TDIDT
  - ID3
  - C4.5
- Advantages & Disadvantages
- Implementation ID3
# Example Database

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temp</th>
<th>Humid</th>
<th>Wind</th>
<th>Playtennis?</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>overcast</td>
<td>hot</td>
<td>high</td>
<td>weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>rain</td>
<td>mild</td>
<td>high</td>
<td>weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>rain</td>
<td>cool</td>
<td>normal</td>
<td>weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>rain</td>
<td>cool</td>
<td>normal</td>
<td>strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>overcast</td>
<td>cool</td>
<td>normal</td>
<td>strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>sunny</td>
<td>mild</td>
<td>high</td>
<td>weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>sunny</td>
<td>cool</td>
<td>normal</td>
<td>weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>rain</td>
<td>mild</td>
<td>normal</td>
<td>weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>sunny</td>
<td>mild</td>
<td>normal</td>
<td>strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>overcast</td>
<td>mild</td>
<td>high</td>
<td>strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>overcast</td>
<td>hot</td>
<td>normal</td>
<td>weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>rain</td>
<td>mild</td>
<td>high</td>
<td>strong</td>
<td>No</td>
</tr>
</tbody>
</table>
Introduction

Example:
(Order = sunny,
Temp = mild,
Hmd = normal,
Wind = weak)
CHAID & CART/TDIDT

- Top-Down Induction of Decision Trees (CART/TDIDT)
  - Non-Incremental approach i.e. needs to start over after change of training data.
  - Needs Pruning as Trees can become overly complex.
  - Examples: CART( Algorithm), ID3, C4.5, C5.0.

- Chi-square Automatic Interaction Detectors (CHAID)
  - Main Difference to CART:
    - Tree growth is limited => avoids pruning.
Information

- Impurity (Entropy)
- Information Gain
  - Expected gain of information after splitting.

\[ \text{gain}(A) = I(p, n) - H(A) \]

Classification- vs. Regression Trees

Classification Tree:

- Classifies categorial target values.
- E.g.: Response {'Yes', 'No'}.

Regression Tree:

- Finds splitting value for continuous target values.
- E.g.: ExpectedTemp { 20.5, 10.7, 30.0, 17.7, ... }.

=> Predictors can be either numeric or categorial.
Iterative Dichotomiser 3 (ID3)

- By J. Ross Quinlan
- Selection criterion: **Information Gain** or **Gain Ratio**.
- **Information Gain**
  - Bits of gained information.
  - Has a bias towards variables with multiple values.
- **Gain Ratio**
  - Takes number and size of branches into account.
  - Is not always defined.
# Example Database

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temp</th>
<th>Humid</th>
<th>Wind</th>
<th>Playtennis?</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>overcast</td>
<td>hot</td>
<td>high</td>
<td>weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>rain</td>
<td>mild</td>
<td>high</td>
<td>weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>rain</td>
<td>cool</td>
<td>normal</td>
<td>weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>rain</td>
<td>cool</td>
<td>normal</td>
<td>strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>overcast</td>
<td>cool</td>
<td>normal</td>
<td>strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>sunny</td>
<td>mild</td>
<td>high</td>
<td>weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>sunny</td>
<td>cool</td>
<td>normal</td>
<td>weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>rain</td>
<td>mild</td>
<td>normal</td>
<td>weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>sunny</td>
<td>mild</td>
<td>normal</td>
<td>strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>overcast</td>
<td>mild</td>
<td>high</td>
<td>strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>overcast</td>
<td>hot</td>
<td>normal</td>
<td>weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>rain</td>
<td>mild</td>
<td>high</td>
<td>strong</td>
<td>No</td>
</tr>
</tbody>
</table>
**ID3 Pseudocode**

1. Select primary key, target attribute, and the dataset
2. If not (pure or stopping criterion met): Call ID3
   a. Calculate *Entropy* & *Information Gain* for every attribute.
   b. Select attribute $X$ with $\text{MAX}(\text{Information Gain})$.
   c. Make a tree node using attribute $X$.
   d. Split dataset into subsets for every value of $X$.
   e. Recursive call of ID3 for every subset.
Example calculation ID3 (Root)

\[
i(Sunny) = \frac{5}{14} \cdot \left( -\frac{2}{5} \cdot \log\left(\frac{2}{5}\right) - \frac{3}{5} \cdot \log\left(\frac{3}{5}\right) \right) = 0.346768
\]

\[
i(Rain) = \frac{5}{14} \cdot \left( -\frac{3}{5} \cdot \log\left(\frac{3}{5}\right) - \frac{2}{5} \cdot \log\left(\frac{2}{5}\right) \right) = 0.346768
\]

\[
i(Overcast) = \frac{4 \cdot (-1 \cdot \log(1))}{14} = 0
\]

\[
E(Outlook, D) = (i(Sunny) + i(Overcast) + i(Rain)) = 0.693486
\]

\[
gain(Outlook) = I(D) - E(Outlook, D) = 0.940286 - 0.693486 = 0.2468
\]

\[
gain(Humid) = I(D) - E(Humid, 0) = 0.1518
\]

\[
gain(Wind) = I(D) - E(Wind, 0) = 0.0481
\]

\[
gain(Temp) = I(D) - E(Temp, 0) = 0.0292
\]
C4.5/C5.0

- Improvements over ID3
  - Handles continuous and categorial variables.
  - Handles missing values.
  - More efficient pruning.

  => C4.5 makes ID3 applicable in practice.

- C5.0 - Commercial implementation
  - Increased performance.
  - Less memory.
  - More precise.
Advantages & Disadvantages

+ Human readable rules.
+ Limited computation power (for application).
+ Handles continuous and categorial values.

- Not optimal for predicting specific values.
- Growing and pruning trees is computationally complex.
- Inefficient for non-rectangular regions.
## Implementation

<table>
<thead>
<tr>
<th>Day</th>
<th>Overcast</th>
<th>Temp</th>
<th>Humid</th>
<th>Wind</th>
<th>Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Thank You!