Growing Decision Trees Genetically, in Parallel

Robert Keller
Brian Bentow
Joseph Malone

Harvey Mudd College
August, 2002
Project pre-history

Outline

• What is a decision tree?
• Why are decision trees of interest?
• Project history
• Pre-existing software
• How is genetic programming used?
• Why/How to use parallel computing?
• Our software
• Related work
• Future directions
General Application Areas

Machine Learning

“Data Mining”

Extraction of meaningful models from collections of data
Data Mining

• Given data collected from an unknown **process** (physical, biological, economic, ...)

• Devise a model that **explains** the data

• So that accurate **predictions** can be made (predict output from input data)
Modeling/Prediction

- Input variables
- Output variables
- Unknown process or parameters

Observations:

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Modeling/Prediction

input variables

output variables

model, known parameters

predictions

<table>
<thead>
<tr>
<th>input</th>
<th>output</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Modeling/Prediction

observation

<table>
<thead>
<tr>
<th>input</th>
<th>output</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

predictions

<table>
<thead>
<tr>
<th>input</th>
<th>output</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Model Validation

- How closely do predictions made by the model agree with actual output?
Types of Models

- Statistical
  - Regression equations
  - Factor analysis

- Neural network
  - Multi-level perceptron
  - Counter propagation network
  - Support Vector Machine

- Decision tree
What is a decision tree?

- A tree of questions to be asked about the input data, resulting in a prediction about the output data.
- There are numerous varieties of questions:
  - Equality: \( x_i = c \) ?
  - Single variable inequality: \( x_i \geq c \) ?
  - Multi-variable inequality: \( a_i x_i + a_j x_j \geq c \) ?
  - Boolean combinations: \( x_i \geq c_i \land x_j \geq c_j \) ?
  - etc.
Data types

- **Continuous data:**
  - Finite or infinite set of values
  - Numeric inequality comparison meaningful

- **Categorical data:**
  - Finite set of values
  - Numeric inequality comparison not meaningful
Given a set of measurements on a banknote, determine whether the banknote is a forgery.

Input measurements (continuous):
- diagonal
- height on the right
- height on the left
- inner frame to upper border
- inner frame to lower border

Output category (categorical):
- {forgery, genuine}
Example of a Decision Tree

- Given a set of measurements on a banknote, determine whether the banknote is a forgery.
diagonal > 140.3
|   inner_frame_to_lower_border > 9.6
|   |   height_on_right > 130.8 forgery
|   |   height_on_right <= 130.8 genuine
|   inner_frame_to_lower_border <= 9.6 genuine
diagonal <= 140.3 forgery
Problem: Given data observations, determine a good decision tree

Observations from actual banknotes (partial)

<table>
<thead>
<tr>
<th>diagonal,</th>
<th>height_on_left, height_on_right, ...</th>
<th>category</th>
</tr>
</thead>
<tbody>
<tr>
<td>214.80000,131.00000,131.10000,9.00000,9.70000,141.00000</td>
<td>genuine</td>
<td></td>
</tr>
<tr>
<td>214.80000,129.70000,129.70000,8.70000,9.60000,142.20000</td>
<td>genuine</td>
<td></td>
</tr>
<tr>
<td>215.00000,129.60000,129.70000,10.40000,7.70000,141.80000</td>
<td>genuine</td>
<td></td>
</tr>
<tr>
<td>215.50000,129.50000,129.70000,7.90000,9.60000,141.60000</td>
<td>genuine</td>
<td></td>
</tr>
<tr>
<td>214.90000,129.40000,129.70000,8.20000,11.00000,141.90000</td>
<td>genuine</td>
<td></td>
</tr>
<tr>
<td>215.30000,130.40000,130.30000,7.90000,11.70000,141.80000</td>
<td>genuine</td>
<td></td>
</tr>
<tr>
<td>215.20000,130.80000,129.60000,7.90000,10.80000,141.40000</td>
<td>genuine</td>
<td></td>
</tr>
<tr>
<td>215.10000,129.90000,129.70000,7.70000,10.80000,141.80000</td>
<td>genuine</td>
<td></td>
</tr>
<tr>
<td>215.10000,130.30000,130.00000,10.60000,10.80000,139.70000</td>
<td>forgery</td>
<td></td>
</tr>
<tr>
<td>214.70000,130.60000,130.10000,11.80000,10.50000,139.80000</td>
<td>forgery</td>
<td></td>
</tr>
<tr>
<td>214.80000,130.50000,130.20000,11.00000,11.00000,140.00000</td>
<td>forgery</td>
<td></td>
</tr>
<tr>
<td>214.80000,130.30000,130.40000,10.10000,12.10000,139.60000</td>
<td>forgery</td>
<td></td>
</tr>
<tr>
<td>215.30000,130.80000,131.10000,11.60000,10.60000,140.20000</td>
<td>forgery</td>
<td></td>
</tr>
<tr>
<td>214.70000,130.50000,130.50000,9.90000,10.30000,140.10000</td>
<td>forgery</td>
<td></td>
</tr>
<tr>
<td>215.00000,130.40000,130.40000,9.40000,11.60000,140.20000</td>
<td>forgery</td>
<td></td>
</tr>
</tbody>
</table>
Our Problem

Data set in

Algorithm ??

Decision tree out

\[
\begin{align*}
\text{diagonal}, & \quad \text{height_on_left}, \text{height_on_right}, \ldots, \text{category} \\
\text{214.80000,131.00000,131.10000,9.00000,9.70000,141.00000,genuine} \\
\text{214.80000,129.70000,129.70000,8.70000,9.60000,142.20000,genuine} \\
\text{215.00000,129.60000,129.70000,10.40000,7.70000,141.80000,genuine} \\
\text{215.50000,129.50000,129.70000,7.90000,9.60000,141.60000,genuine} \\
\text{214.90000,129.40000,129.70000,8.20000,11.00000,141.90000,genuine} \\
\text{215.30000,130.40000,130.30000,7.90000,11.70000,141.80000,genuine} \\
\text{215.20000,130.80000,129.60000,7.90000,10.80000,141.40000,genuine} \\
\text{215.10000,129.90000,129.70000,7.70000,10.80000,141.80000,genuine} \\
\text{215.10000,130.30000,130.00000,10.60000,10.80000,139.70000,forgery} \\
\text{214.70000,130.60000,130.10000,11.80000,10.50000,139.80000,forgery} \\
\text{214.80000,130.50000,130.20000,11.00000,11.00000,140.00000,forgery} \\
\text{214.80000,130.30000,130.40000,10.10000,12.10000,139.60000,forgery} \\
\text{215.30000,130.80000,131.10000,11.60000,10.60000,140.20000,forgery} \\
\text{214.70000,130.50000,130.50000,9.90000,10.30000,140.10000,forgery} \\
\text{215.00000,130.40000,130.40000,9.40000,11.60000,140.20000,forgery}
\end{align*}
\]

- diagonal > 140.3
  - inner_frame_to_lower_border > 9.6
    - height_on_right > 130.8 forgery
    - height_on_right <= 130.8 genuine
      - inner_frame_to_lower_border <= 9.6 genuine
      - diagonal <= 140.3 forgery
Two Varieties of Tree

- **Classification tree:**
  - Determine one of a set of discrete categories for the output variable

- **Regression tree:**
  - Determine a predicted parameter, such as mean value, for the output variable
Value of Trees

- **Trees vs. Regression formulae:**
  - Tree provides discrete explanation, whereas regression formula is a monolithic continuous function

- **Trees vs. Neural nets:**
  - Neural nets often don’t display their “rationale” as well
Some Difficulties

- Real-world data may be noisy and have missing observations.
- Data may have conflicting outputs for the same input.
- There is no unique tree:
  - Tree classification accuracy can be increased,
  - at the expense of increasing tree complexity.
- Validation requires additional data not already used as input
Tree Pruning

- “Pruning” a tree is a process of cutting out selected sub-trees:
  - The size gets smaller, but
  - the error usually increases.

- The trick is to prune the sub-trees that have the least impact.
Algorithmic Approaches

• Based on choosing questions that give the best discrimination with the least error, e.g.

• **Greedy algorithm:** At a node in the tree, construct a question that divides the data into two sets, such that the category error is minimum. Repeat for child nodes.

• The problem of constructing minimal decision tree has been shown **NP-complete.** Hyafil, R. and Rivest, R.L.. Constructing optimal binary trees is NP-complete, Information processing letters, 5, 15-17, 1976.

• Therefore resort to **heuristic** approaches.
Pre-existing software

- c4.5 (formerly c3.0)
  Ross Quinlan, Machine learning background
  Now a commercial product c5.

- CART (Classification-and-Regression-Trees)
  Leo Breiman, et al., Statistics background
  Also a commercial product

- R-PART (Recursive Partitioning) library in S-plus
  Public domain implementation of CART methods
  Terry M Therneau and Beth Atkinson, Mayo Clinic

- Numerous others
What is Genetic Programming?

- **GP** is a form of genetic algorithm in which the **genomes are programs**, rather than, say, bit strings.

- **Genetic operators** such as *mutation* and *crossover* are defined on programs.

- **Fitness** of a program is how well it carries out a specific function.
Trees as Genomes

- Every program in a given language has a syntax tree.

- The genetic operators are expressed based on syntax trees.

- A decision tree is a primitive form of program.
Mutation Operator

diagonal > 140.3?
  yes
  inner frame/lower border > 9.6?
    yes
    genuine
diagonal > 140.3?
  no
  forgery

right height > 130.8?
  yes
  forgery
genuine
  no
  genuine

left height > 130.6?
  yes
  diagonal > 140.3?
    yes
    forgery
diagonal > 140.3?
    no
    genuine
  no
  forgery

right height > 130.8?
  yes
  forgery
genuine
  no
  genuine

right height > 130.8?
  yes
  forgery
genuine
  no
  genuine
Crossover Operator

inner frame/lower border > 9.6?

inner frame/upper border > 10.3?

diagonal > 140.3?

right height > 130.8?

genuine

forgery

diagonal > 140.1?

left height > 130.8?

genuine

forgery

forgery
Project Objectives

- We set out to design a program that would evolve decision trees based on input data.
- We chose to use the c4.5 data format.
- We used c4.5 as a performance benchmark.
- We wanted to be able to use parallel processors to speed up the computation.
First Milestone: gentree

- We produced a program that evolves both classification and regression trees (c4.5 only does classification).

- In some cases, our program produced better quality results than c4.5, in other cases, lower quality. We are still exploring ways to improve the results.

- Our program invariably runs longer than c4.5, especially on large data sets.

- Our program can display lots of options (accuracy vs. size tradeoffs).
Malone’s Optimization

- Rather than completely build a tree, then evaluate its fitness, only build it down to the nodes before the leaves.

- Then assign categories to the leaves that minimize the overall error.
Gentree output, example mbl-factor, run ID 102

Run parameters were:
1000 Chromosomes. Training dataset size: 310
Dataset is: /research/keller/gentree/data/factor/factor.data
Fitness function was: % correct - 0.001 * treeSize - 0.005 * nVars
The top 10% were replicated each generation
Target error was: 0   Target size was: 0

Evolution progress (Only changes of most-fit individuals are shown):

<table>
<thead>
<tr>
<th>gen</th>
<th>size</th>
<th>error</th>
<th>nCorrect</th>
<th>fitness</th>
<th>mean</th>
<th>stddev</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>21</td>
<td>13.55%</td>
<td>268/310</td>
<td>0.84</td>
<td>0.63</td>
<td>0.042</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>13.55%</td>
<td>268/310</td>
<td>0.85</td>
<td>0.67</td>
<td>0.067</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>13.23%</td>
<td>269/310</td>
<td>0.86</td>
<td>0.70</td>
<td>0.088</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>11.29%</td>
<td>275/310</td>
<td>0.86</td>
<td>0.72</td>
<td>0.102</td>
</tr>
<tr>
<td>4</td>
<td>23</td>
<td>11.61%</td>
<td>274/310</td>
<td>0.86</td>
<td>0.73</td>
<td>0.111</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>9.68%</td>
<td>280/310</td>
<td>0.88</td>
<td>0.73</td>
<td>0.113</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>9.68%</td>
<td>280/310</td>
<td>0.89</td>
<td>0.74</td>
<td>0.116</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>10.00%</td>
<td>279/310</td>
<td>0.89</td>
<td>0.75</td>
<td>0.112</td>
</tr>
<tr>
<td>10</td>
<td>19</td>
<td>9.03%</td>
<td>282/310</td>
<td>0.89</td>
<td>0.75</td>
<td>0.117</td>
</tr>
<tr>
<td>11</td>
<td>23</td>
<td>9.03%</td>
<td>282/310</td>
<td>0.89</td>
<td>0.74</td>
<td>0.124</td>
</tr>
<tr>
<td>12</td>
<td>25</td>
<td>8.39%</td>
<td>284/310</td>
<td>0.89</td>
<td>0.74</td>
<td>0.124</td>
</tr>
<tr>
<td>14</td>
<td>29</td>
<td>8.06%</td>
<td>285/310</td>
<td>0.90</td>
<td>0.76</td>
<td>0.118</td>
</tr>
<tr>
<td>15</td>
<td>43</td>
<td>7.74%</td>
<td>286/310</td>
<td>0.90</td>
<td>0.77</td>
<td>0.116</td>
</tr>
</tbody>
</table>
## Examples

<table>
<thead>
<tr>
<th>Example</th>
<th>method</th>
<th>tree size</th>
<th>training error %</th>
<th>test error %</th>
</tr>
</thead>
<tbody>
<tr>
<td>iris</td>
<td>unpruned C4.5</td>
<td>5</td>
<td>2.6</td>
<td>5.4</td>
</tr>
<tr>
<td>iris</td>
<td>pruned C4.5</td>
<td>5</td>
<td>2.6</td>
<td>5.4</td>
</tr>
<tr>
<td>iris</td>
<td>gentree</td>
<td>5</td>
<td>2.6</td>
<td>5.4</td>
</tr>
<tr>
<td>mbl-factor</td>
<td>unpruned C4.5</td>
<td>289</td>
<td>1.6</td>
<td>(test = training)</td>
</tr>
<tr>
<td>mbl-factor</td>
<td>pruned C4.5</td>
<td>53</td>
<td>20.9</td>
<td></td>
</tr>
<tr>
<td>mbl-factor</td>
<td>gentree</td>
<td>25</td>
<td>1.6</td>
<td></td>
</tr>
</tbody>
</table>
# Other Comparisons

<table>
<thead>
<tr>
<th>Sample</th>
<th>C4.5 Size</th>
<th>C4.5 Error</th>
<th>Gentree Size</th>
<th>Gentree Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>crx</td>
<td>23</td>
<td>11.7</td>
<td>9</td>
<td>12.2</td>
</tr>
<tr>
<td>monk1</td>
<td>18</td>
<td>16.1</td>
<td>11</td>
<td>8.87</td>
</tr>
<tr>
<td>monk2</td>
<td>31</td>
<td>23.7</td>
<td>9</td>
<td>28.7</td>
</tr>
<tr>
<td>monk3</td>
<td>12</td>
<td>6.6</td>
<td>5</td>
<td>6.56</td>
</tr>
<tr>
<td>herman</td>
<td>59</td>
<td>15.4</td>
<td>9</td>
<td>24</td>
</tr>
<tr>
<td>post operative</td>
<td>1</td>
<td>22</td>
<td>11</td>
<td>9.76</td>
</tr>
</tbody>
</table>
Comparison Trees: iris

gentree

5 nodes  generation 0  error 2.632%  correct 74/76  fitness 0.9712
a4 > 0.6  (51 of 76)
  |  a4 > 1.6  (25 of 51)  category Iris-virginica  (24/1)
  |  a4 <= 1.6  (26 of 51)  category Iris-versicolor  (25/1)
a4 <= 0.6  (25 of 76)  category Iris-setosa  (0/25)

c4.5

a4 <= 0.6 : Iris-setosa  (25.0)
a4 > 0.6 :
  |  a4 <= 1.6 : Iris-versicolor  (26.0/1.0)
  |  a4 > 1.6 : Iris-virginica  (25.0/1.0)

Evaluation on training data (76 items):

<table>
<thead>
<tr>
<th></th>
<th>Before Pruning</th>
<th>After Pruning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Size</td>
<td>Errors</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2( 2.6%)</td>
</tr>
</tbody>
</table>
**Comparison Trees: mbl-factor**

**c4.5**

Simplified Decision Tree:

- **pos1 = T:**
  - Pos5 = H: 1 (1.0/0.8)
  - pos5 = F: 1 (10.0/2.4)
  - pos5 = P: 0 (5.0/2.3)
  - pos5 = Y: 1 (13.0/3.6)
  - pos5 = M: 1 (0.0)
  - pos5 = N: 1 (0.0)
  - pos5 = L: 1 (2.0/1.0)
  - pos5 = V: 0 (3.0/1.1)
  - pos5 = A: 1 (1.0/0.8)
  - pos5 = I: 0 (2.0/1.8)
  - pos5 = G: 0 (1.0/0.8)
  - pos5 = D: 0 (1.0/0.8)
  - pos5 = E: 0 (1.0/0.8)
  - pos5 = S: 0 (1.0/0.8)
  - pos5 = K: 1 (0.0)
  - pos5 = Q: 1 (0.0)
  - pos5 = R: 0 (1.0/0.8)
  - pos5 = W: 1 (0.0)
  - pos5 = T: 1 (0.0)
  - pos5 = C: 1 (0.0)

- **pos1 = S:** 1 (149.0/20.4)
- **pos1 = Y:** 1 (18.0/2.5)
- **pos1 = V:** 0 (3.0/2.1)
- **pos1 = Q:** 0 (7.0/3.4)
- **pos1 = A:** 0 (9.0/3.5)
- **pos1 = N:** 0 (4.0/2.2)
- **pos1 = L:** 0 (12.0/1.3)
- **pos1 = G:** 0 (4.0/1.2)
- **pos1 = F:** 0 (5.0/1.2)
- **pos1 = D:** 0 (11.0/1.3)
- **pos1 = R:** 0 (8.0/1.3)
- **pos1 = P:** 0 (21.0/1.3)
- **pos1 = H:** 0 (3.0/1.1)
- **pos1 = I:** 0 (3.0/1.1)
- **pos1 = C:** 0 (4.0/1.2)
- **pos1 = E:** 0 (4.0/1.2)
- **pos1 = K:** 0 (3.0/1.1)
Some Challenges

- Fitness is based on a combination of size and accuracy. What is the “best” way to weight the two?

- If just one tree has to be recommended, which one?

- What is the best way to validate the results?
gentree trade-off display: iris

<table>
<thead>
<tr>
<th>gen</th>
<th>size</th>
<th>error</th>
<th>nCorrect</th>
<th>fitness</th>
<th>mean</th>
<th>stddev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>65.79%</td>
<td>26/76</td>
<td>0.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>32.89%</td>
<td>51/76</td>
<td>0.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>2.63%</td>
<td>74/76</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>1.32%</td>
<td>75/76</td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>gen</th>
<th>size</th>
<th>error</th>
<th>nCorrect</th>
<th>fitness</th>
<th>mean</th>
<th>stddev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>66.22%</td>
<td>25/74</td>
<td>0.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>32.43%</td>
<td>50/74</td>
<td>0.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>5.41%</td>
<td>70/74</td>
<td>0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>4.05%</td>
<td>71/74</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Examples run by gentree

agaricus-lepiota
anneal
balance-scale
banknotes
breast-cancer
breast-cancer-wisconsin
bupa
car
clean1
crx
dsilk
factor
german
glass
golf
hayes-roth
hepatitis
imports-85
income-food
ionosphere
iris
krkopt
kyphosis
labor-neg
lung-cancer
lymphography
monk1
monk2
monk3
parity3
parity4
pima-indians-diabetes
post-operative
primary-tumor
segal
servo
soybean
tic-tac-toe
vote
wine
yeast
zoo
Cross-Validation

- We implemented n-way cross validation (user selects n):
  - Data set is divided into n disjoint sets.
  - n training runs are done, with \((n-1)/n\) of the data used as training and \(1/n\) as test.
  - The average error vs. size is plotted.
  - We expect that the “knee” in the curve indicates the most robust error value.
  - Given the target error value, a tree of corresponding size can be generated.
n-Way Cross-Validation

error

recommended target error

tree size
Traditionally, cross-validation has been used to prune down the tree size,
but we aren’t pruning in the traditional sense.

The approach seems plausible, but the theory is unclear.
Knee-selection has not been automated.
Another Challenge: Missing Data Values

- There are several known methods for dealing with missing data values, none universally accepted.

- c4.5 uses a “fractional re-distribution” method.

- We implemented a “surrogate variable” method, which prescribes a value for the missing variable based on proximity to other data points.
Another Challenge: Maintaining Population Diversity

- A common difficulty with genetic approaches is that, because of natural selection, the population of solutions loses diversity over time.

- This means that fewer nuances are introduced, and the trend may move toward a fitness local maximum.

- We have been working on ways to prevent loss of diversity, without sacrificing quality of the most fit solutions.

- One approach involves defining an equivalence relation among the individuals and only keeping one member of each equivalence class.
Choice of Equivalence Relation

- Numerous equivalence relations are possible.
- Identity is too fine, and same-error-value may be too coarse.
- We implemented “confusion matrices” as the basis for equivalence.

\[
\begin{array}{cc}
(a) & (b) & \text{<-classified as} \\
108 & 21 & \text{(a): class 0} \\
8 & 173 & \text{(b): class 1}
\end{array}
\]
Why/How to use parallel computing?

- Parallel computing offers a means to speed up computation when other algorithmic and technology-based approaches saturate.

- Genetic programming seems to be a reasonable candidate for parallel computing: evolution in nature takes place in parallel.
Parallel version of gentree.
Population is divided into “demes”.
Each deme evolves independently for a while, then passes most fit to its immediate neighbor, called the “island model”.
We currently use a logical “ring” organization to connect the processors.
Demes can also help with the diversity problem.
Choice for Genetic Programming

- We use distributed memory
- **MPI** library (Multi-Processor Interface)
- Very portable
- Many systems use it:
  - “Beowulf” clusters (16 processors at HMC)
  - Others
- Based on Single-Program Multiple-Data Model
  - One program for all computers
  - Differentiable by processor id
## Sample Speedup Measurements

Running `mbl-factor` with aggregate population 1000 processors:

<table>
<thead>
<tr>
<th>processors</th>
<th>run time (sec.)</th>
<th>speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>580.530</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>328.650</td>
<td>1.77</td>
</tr>
<tr>
<td>4</td>
<td>167.490</td>
<td>3.47</td>
</tr>
<tr>
<td>8</td>
<td>74.230</td>
<td>7.80</td>
</tr>
</tbody>
</table>
Early Termination

- Our implementation includes a distributed load sharing and early-termination capability.

- When a processor is not “making progress” toward better solutions in its deme, it notifies its neighbor. The notification is passed to other processors, until a processor is found that is making progress, which then gives some highly fit individuals to the originating processor.

- If no such processor is found, then a message is sent out to terminate execution.
Added Challenge

- Early termination makes it difficult to determine speedup, because the amount of work done is dependent on the number of processors.

- We don’t want to delay getting answers to the user, yet we also want to be getting true speedup from the multi-processor version of the system.
Related work (1)

- There is a fair amount of work on **parallel genetic programming**.

- In 1999, almost no references to **genetic programming of decision trees** could be found. An old, cursory, one I discovered was:

At the start of the project, we discovered some more recent work, notably: H.C. Kennedy, et al., *The construction and evaluation of decision trees: A comparison of evolutionary and concept learning methods*, Springer, LNCS 1305, 1997 and


At the end of the project, we discovered some very relevant work hiding under “genetic algorithms”: Athanasios Papagelis, Dimitris Kalles, *GATree: Genetically Evolved Decision Trees*, International Conference on Tools in AI, IEEE, Nov. 2001. This work reports extraordinarily good results for tree size (6 times better than C4.5, in some cases). We are still in the process of evaluating the claims.
gatree reported results on UCI datasets

Table 1: Classification accuracy

<table>
<thead>
<tr>
<th>Dataset</th>
<th>C4.5</th>
<th>OneR</th>
<th>GATree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colic</td>
<td>83.84±3.41</td>
<td>81.37±5.36</td>
<td>85.01±4.55</td>
</tr>
<tr>
<td>Heart-Statlog</td>
<td>74.44±3.56</td>
<td>76.3±3.04</td>
<td>77.48±3.07</td>
</tr>
<tr>
<td>Diabetes</td>
<td>66.27±3.71</td>
<td>63.27±2.59</td>
<td>63.97±3.71</td>
</tr>
<tr>
<td>Credit</td>
<td>83.77±2.93</td>
<td>86.81±4.45</td>
<td>86.81±4.45</td>
</tr>
<tr>
<td>Hepatitis</td>
<td>77.42±6.84</td>
<td>84.52±6.2</td>
<td>80.46±5.39</td>
</tr>
<tr>
<td>Iris</td>
<td>92±2.98</td>
<td>94.67±3.8</td>
<td>93.8±4.02</td>
</tr>
<tr>
<td>Labor</td>
<td>85.26±7.98</td>
<td>72.73±14.37</td>
<td>87.27±7.24</td>
</tr>
<tr>
<td>Lymph</td>
<td>65.52±14.63</td>
<td>74.14±7.18</td>
<td>75.24±10.69</td>
</tr>
<tr>
<td>Breast-Cancer</td>
<td>71.93±5.11</td>
<td>68.17±7.93</td>
<td>71.03±8.34</td>
</tr>
<tr>
<td>Zoo</td>
<td>90±7.91</td>
<td>43.8±10.47</td>
<td>85.4±4.02</td>
</tr>
<tr>
<td>Vote</td>
<td>96.09±3.86</td>
<td>95.63±4.33</td>
<td>95.63±4.33</td>
</tr>
<tr>
<td>Glass</td>
<td>55.24±7.49</td>
<td>43.19±4.33</td>
<td>53.48±4.33</td>
</tr>
<tr>
<td>Balance-Scale</td>
<td>78.24±4.4</td>
<td>59.68±4.4</td>
<td>71.15±6.47</td>
</tr>
<tr>
<td><strong>AVERAGES</strong></td>
<td><strong>78.46</strong></td>
<td><strong>72.64</strong></td>
<td><strong>78.98</strong></td>
</tr>
</tbody>
</table>

Table 2: Average tree sizes

<table>
<thead>
<tr>
<th>Dataset</th>
<th>C4.5</th>
<th>GATree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colic</td>
<td>27.4</td>
<td>5.84</td>
</tr>
<tr>
<td>Heart-Statlog</td>
<td>39.4</td>
<td>8.28</td>
</tr>
<tr>
<td>Diabetes</td>
<td>140.6</td>
<td>6.6</td>
</tr>
<tr>
<td>Credit</td>
<td>57.8</td>
<td>3</td>
</tr>
<tr>
<td>Hepatitis</td>
<td>19.8</td>
<td>5.56</td>
</tr>
<tr>
<td>Iris</td>
<td>9.6</td>
<td>7.48</td>
</tr>
<tr>
<td>Labor</td>
<td>8.6</td>
<td>8.72</td>
</tr>
<tr>
<td>Lymph</td>
<td>28.2</td>
<td>7.96</td>
</tr>
<tr>
<td>Breast-Cancer</td>
<td>35.4</td>
<td>6.68</td>
</tr>
<tr>
<td>Zoo</td>
<td>17</td>
<td>10.12</td>
</tr>
<tr>
<td>Vote</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>Glass</td>
<td>60.2</td>
<td>8.98</td>
</tr>
<tr>
<td>Balance-Scale</td>
<td>106.6</td>
<td>8.92</td>
</tr>
<tr>
<td><strong>AVERAGES</strong></td>
<td><strong>43.2</strong></td>
<td><strong>7.01</strong></td>
</tr>
</tbody>
</table>
gatree's fitness function

\[
\text{payoff(tree } i) = \text{CorrectClassified}_i^2 \times \frac{x}{\text{size}_i^2 + x}
\]
Future directions

- An über-decision-tree generator can be built by seeding the population with results of C4.5 or other algorithms.

- While parallel processing looks promising, we have more work to do to cogently present its benefits.