Overview

Genetic Programming: Syntax & Semantics

1. Setting the Stage
   - What is Natural Computing?
   - What is Evolutionary Computation?
   - An Introduction to Genetic Programming (GP)

2. Grammar-based GP

3. Semantic methods & Open Issues in GP
Semantic Methods & Open Issues in GP
Semantic Methods

- Over dependence on fitness (single point)
- Credit Assignment
- Semantic analysis of evolving populations
- Semantic-aware program construction
- Semantic-aware search operators
Attribute Grammars - Adding Semantics to Solution Construction

\[
\text{maximise } \sum_{j=1}^{n} p_j x_j \quad (1)
\]

subject to \[\sum_{j=1}^{n} w_{ij} x_j \leq c_i, \quad i = 1 \ldots m \quad (2)\]

\[x_j \in \{0, 1\}, \quad j = 1 \ldots n \quad (3)\]

---

Where, \(p_j\) refers to the profit, or worth of item \(j\), \(x_j\) refers to the item \(j\), \(w_{ij}\) refers to the relative-weight of item \(j\), with respect to knapsack \(i\), and \(c_i\) refers to the capacity, or weight-constraint of knapsack \(i\). There are present \(j = 1 \ldots n\) items, and \(i = 1 \ldots m\) knapsacks.
Semantic-aware search operators

- Crossover
- Mutation
- Semantic Locality & Diversity
**Algorithm 1**: Semantic Similarity based Crossover

select Parent 1 $P_1$;  
select Parent 2 $P_2$;  
Count=0;  
**while** $Count < Max\_Trial$ **do**  
- choose a random crossover point $Subtree_1$ in $P_1$;  
- choose a random crossover point $Subtree_2$ in $P_2$;  
- generate a number of random points ($P$) on the problem domain;  
- calculate the SSD between $Subtree_1$ and $Subtree_2$ on $P$  
  **if** $Subtree_1$ is similar to $Subtree_2$ **then**  
    execute crossover;  
    add the children to the new population;  
    return true;  
  **else**  
    Count=Count+1;  
**if** $Count = Max\_Trial$ **then**  
- choose a random crossover point $Subtree_1$ in $P_1$;  
- choose a random crossover point $Subtree_2$ in $P_2$;  
- execute crossover;  
- return true;
Sampling Semantic Distance

Based on SS, we define a Sampling Semantic Distance (SSD) between two subtrees. It differs from that in [24] in using the mean absolute difference in SS values, rather than (as before) the sum of absolute differences. Let $U = (u_1, u_2, ..., u_N)$ and $V = (v_1, v_2, ..., v_N)$ represent the SSs of two subtrees, $S_1$ and $S_2$; then the SSD between $S_1$ and $S_2$ is defined in equation 1:

$$SSD(S_1, S_2) = \frac{1}{N} \sum_{i=1}^{N} |u_i - v_i|$$

(1)

We follow [24] in defining a semantic relationship, Semantic Similarity (SSi), on the basis that the exchange of subtrees is most likely to be beneficial if they are not semantically identical, but also not too different. Two subtrees are semantically similar if their SSD lies within a positive interval. The formal definition of SSi between subtrees $S_1$ and $S_2$ is given in the following equation:

$$SSi(S_1, S_2) = \text{TruthValue}(\alpha < SSD(S_1, S_2) < \beta)$$

where $\alpha$ and $\beta$ are two predefined constants, the lower and upper bounds for semantics sensitivity. In general, the best values for these semantic sensitivity bounds are problem dependent. In this work we set $\alpha = 10^{-4}$ and several values of $\beta$ were tested.

<table>
<thead>
<tr>
<th>Functions</th>
<th>Training Data</th>
<th>Testing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1 = x^4 + x^3 + x^2 + x$</td>
<td>20 random points $\subseteq [-1,1]$</td>
<td>30 points $\subseteq [0:0.05:1.5]$</td>
</tr>
<tr>
<td>$F_2 = x^3 - x^2 - x - 1$</td>
<td>20 random points $\subseteq [-1,1]$</td>
<td>30 points $\subseteq [0:0.05:1.5]$</td>
</tr>
<tr>
<td>$F_3 = \text{arcsin}(x)$</td>
<td>20 random points $\subseteq [-1,0]$</td>
<td>30 points $\subseteq [-1:0.67:1]$</td>
</tr>
<tr>
<td>$F_4 = \sqrt{x}$</td>
<td>20 random points $\subseteq [0,2]$</td>
<td>30 points $\subseteq [0:0.1:3]$</td>
</tr>
<tr>
<td>$F_5 = 0.3\sin(2\pi x)$</td>
<td>20 random points $\subseteq [-1,1]$</td>
<td>30 points $\subseteq [0:0.05:1.5]$</td>
</tr>
<tr>
<td>$F_6 = \cos(3x)$</td>
<td>20 random points $\subseteq [-1,1]$</td>
<td>30 points $\subseteq [0:0.05:1.5]$</td>
</tr>
</tbody>
</table>
### Semantic vs. Syntactic Locality

**Table 3.** Comparison of the effects of SC, SSC and SySC on GP performance (mean of the best fitness). The values are scaled by $10^2$.

<table>
<thead>
<tr>
<th>Xovers</th>
<th>$F_1$</th>
<th>$F_2$</th>
<th>$F_3$</th>
<th>$F_4$</th>
<th>$F_5$</th>
<th>$F_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>1.51</td>
<td>3.07</td>
<td>0.37</td>
<td>0.96</td>
<td>4.36</td>
<td>1.48</td>
</tr>
<tr>
<td>SySC6</td>
<td>1.63</td>
<td>3.20</td>
<td>0.46</td>
<td>1.06</td>
<td>4.42</td>
<td>1.46</td>
</tr>
<tr>
<td>SySC8</td>
<td>1.49</td>
<td>3.50</td>
<td>0.43</td>
<td>0.99</td>
<td>4.36</td>
<td>1.98</td>
</tr>
<tr>
<td>SySC10</td>
<td>1.56</td>
<td>3.08</td>
<td>0.39</td>
<td>1.18</td>
<td>4.41</td>
<td>2.04</td>
</tr>
<tr>
<td>SSC04</td>
<td>0.78</td>
<td>1.30</td>
<td>0.20</td>
<td>0.58</td>
<td>3.36</td>
<td>0.67</td>
</tr>
<tr>
<td>SSC05</td>
<td>0.85</td>
<td>1.40</td>
<td>0.21</td>
<td>0.61</td>
<td>3.28</td>
<td>0.81</td>
</tr>
<tr>
<td>SSC06</td>
<td>0.87</td>
<td>1.70</td>
<td>0.22</td>
<td>0.38</td>
<td>3.44</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Open Issues

1. Identifying appropriate representations for GP
2. Fitness landscapes & problem difficulty in GP
3. Static vs. Dynamic Problems
4. The influence of biology on GP
5. Open-ended evolution in GP
6. Generalisation in GP
7. GP Benchmarks
8. GP and Modularity
9. The Complexity of GP
Open Issues

10 Miscellaneous issues:
   - The Halting Problem
   - How much Domain Knowledge?
   - GP Theory
   - Constants in GP
   - Bloat
   - Distributed GP
   - The Elephant in the Room!
Sample Literature

Semantic Methods & Open Issues


Sample Literature (continued)

Semantic Methods & Open Issues


Thank You