Genetic Programming Needs Better Benchmarks

Web: http://gpbenchmarks.org — Twitter: @gpbenchmarks #gpbenchmarks

McDermott et al.
“What say you? Hence, Horrible villain! or I’ll spurn thine eyes Like balls before me; I’ll unhair thy head: Thou shalt be whipp’d with wire, and stew’d in brine, Smarting in lingering pickle.”

“Gracious madam, I that do bring the news made not the match.”
This is a Position Paper

We **want** to spark debate and gather the opinions of researchers and practitioners.

We **don’t** want to devise new benchmarks.

We **don’t** want to impose our ideas on the community.

We **want** to establish a community consensus.
This is a Position Paper

We **want** to spark debate and gather the opinions of researchers and practitioners.

We **don’t** want to devise new benchmarks.

We **don’t** want to impose our ideas on the community.

We **want** to establish a community consensus.

...with your help.
Perspectives
The Toy Problem Problem
De Facto Benchmarks

Web: http://gpbenchmarks.org — Twitter: @gpbenchmarks #gpbenchmarks

John R. Koza

Genetic Programming

On the Programming of Computers by Means of Natural Selection
Challenges and Ideas to Drive Innovation
Current Practice
Published Use of Benchmarks


<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbolic Regression</td>
<td>32</td>
</tr>
<tr>
<td>Classification</td>
<td>27</td>
</tr>
<tr>
<td>Path Finding and Planning</td>
<td>10</td>
</tr>
<tr>
<td>Boolean Functions</td>
<td>9</td>
</tr>
<tr>
<td>Traditional Programming</td>
<td>8</td>
</tr>
<tr>
<td>Predictive Modelling</td>
<td>7</td>
</tr>
<tr>
<td>Constructed Problems</td>
<td>3</td>
</tr>
<tr>
<td>Control Problems</td>
<td>1</td>
</tr>
<tr>
<td>Others</td>
<td>4</td>
</tr>
</tbody>
</table>

Limited variety e.g. 26% of papers involving symbolic regression used the quartic equation.

Web: http://gpbenchmarks.org — Twitter: @gpbenchmarks #gpbenchmarks
A Few Previous Benchmark Suite Efforts

**Machine Learning**
UCI Machine Learning Repository (1985–)

**Evolutionary Computation**
De Jong Test Suite (1975)
Evolutionary Strategy Test Functions (1975)
Evolutionary Computation Benchmarking Repository (2006)
Black-Box Optimization Benchmarking Workshop (2009–)

**Genetic Programming**
Koza-I (1992) and Koza-II (1994)
GP-Beagle (2000)

Web: http://gpbenchmarks.org — Twitter: @gpbenchmarks #gpbenchmarks McDermott et al.
Defining and Distributing Benchmarks
What makes for a good benchmark?

Tunably Difficult
Varied
Relevant (Real World? Constructed?)
Fast (?)
Accommodating to Implementors
Supports good empirical method (e.g. problem generation)
Easy to interpret and compare
Representation Independent
Precisely Defined (to an extent!)
Known global optimum?

Web: http://gpbenchmarks.org — Twitter: @gpbenchmarks #gpbenchmarks
McDermott et al.
How should a Benchmark Suite be Defined?

**Standardised Code**
Lock-in to given languages, toolkits etc.

**Specifications**
Will people implement them?

**Specifications + Reference Implementation**
People have at least one reference implementation to rally around
Library implementers can copy code or use the specs

Try before you buy - 53 regression problems and more - see the website for details.
A Cautionary Tale

Focus on **scientific progress** more than benchmark performance.

Recall the problem behind the data.

Problem solving is more than a performance metric for a given algorithm/

See “Machine Learning that Matters”
K. Wagstaff, ICML 2012.
Next Steps
Time for Engaging Discussion!

What next?

1. Please complete our survey.
2. Try out the initial regression benchmarks code.
3. Further discussion via the mailing list.
4. We will analyse and report the results of the survey.
5. A draft benchmark suite will be published (if demanded!).

GP Benchmarks.org