Parameter Tuning for Search-Based Test-Data Generation Revisited
Support for Previous Results

Anton Kotelyanskii
Gregory M. Kapfhammer
Software Testing

Test Suites
Automatic Generation
Confronting Challenges
Evaluation Strategies
Empirical Studies
Challenges
Importance
Replication
Rarity
EvoSuite

Amazing test suite generator
Uses a genetic algorithm

Input: A Java class
Output: A JUnit test suite

http://www.evosuite.org/
Parameter Tuning

*RSM:* Response surface methodology
*SPOT:* Sequential parameter optimization toolbox
Successfully applied to many diverse problems!
Defaults or Tuned Values?
Experiment Design

- Eight EvoSuite parameters
- Ten projects from SF100
- 475 Java classes for subjects
- 100 trials after parameter tuning
- Aiming to improve statement coverage
<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Size</td>
<td>5</td>
<td>99</td>
</tr>
<tr>
<td>Chromosome Length</td>
<td>5</td>
<td>99</td>
</tr>
<tr>
<td>Rank Bias</td>
<td>1.01</td>
<td>1.99</td>
</tr>
<tr>
<td>Number of Mutations</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Max Initial Test Count</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Crossover Rate</td>
<td>0.01</td>
<td>0.99</td>
</tr>
<tr>
<td>Constant Pool Use Probability</td>
<td>0.01</td>
<td>0.99</td>
</tr>
<tr>
<td>Test Insertion Probability</td>
<td>0.01</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Experiments

184 days of computation time estimated
Cluster of 70 computers running for weeks
Identified 139 "easy" and 21 "hard" classes
Mann-Whitney U-test and
Vargha-Delaney effect size
# Results

<table>
<thead>
<tr>
<th>Category</th>
<th>Effect Size</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results Across Trials and Classes</td>
<td>0.5029</td>
<td>0.1045</td>
</tr>
<tr>
<td>No &quot;Easy&quot; and &quot;Hard&quot; Classes</td>
<td>0.5048</td>
<td>0.0314</td>
</tr>
</tbody>
</table>

Using *lower-is-better* inverse statement coverage.

Effect size *greater* than 0.5 means that tuning is *worse*.

Testing shows we do not *always* reject the null hypothesis.

Additional empirical results in the QSIC 2014 paper!
Discussion

Tuning improved scores for 11 classes
Otherwise, same as or worse than defaults
A "soft floor" may exist for parameter tuning
Additional details in the QSIC 2014 paper!
Practical Implications

Fundamental Challenges
Tremendous Confidence
Great Opportunities
Important Contributions

Comprehensive Experiments
Conclusive Confirmation
For EvoSuite, *Defaults = Tuned*