Clustering Test Cases to Achieve Effective & Scalable Prioritisation
Incorporating Expert Knowledge

Yuening Hu
University of Maryland, College Park
ynhu@cs.umd.edu
Nov. 24, 2009
Outline

- Background
- Motivation
- Framework
  - Clustering
  - Clustering-based Prioritisation
  - Analytic Hierarchy Process
  - Evaluation
- Experiments & Analysis
- Related Work
- Conclusions
Background

- Test case prioritisation
  - Regression test
  - An efficient ordering of test cases
- Ideal ordering
  - Reveal faults earliest
  - Not known in advance
Background

- Available criteria
  - Structural coverage
  - Requirement priority
  - Mutation score

- Powerful expert judgement
  - Human tester
  - Rich domain knowledge
  - Human guidance to avoid bias
  - Techniques: Analytic Hierarchy Process
Motivation

- Analytic Hierarchy Process (AHP)
  - Assumption: Human involvement
    → prioritisation improvement
  - Pair-wise comparison
  - Scalability challenges: 100 meaningful comparisons
  - Usually much more than 100

AHP–based prioritisation
Clustering to control the number of comparisons
Expert–guided prioritisation
Outline

- Background
- Motivation
- Framework
  - Clustering
  - Clustering-based Prioritisation
  - Analytic Hierarchy Process
  - Evaluation
- Experiments & Analysis
- Related Work
- Conclusions
Framework

Clustering

Intra prioritisation

Inter prioritisation

Generate best order

Evaluation
Clustering

- Ideal clustering criterion
  - Similarity of detected faults
- Used clustering criterion
  - One bit per statement: 1/0
  - Binary string of each test cases
  - Hamming distance
Clustering

$k = 1 \{ABCDEF\}$

$k = 2 \{AB, CDEF\}$

$k = 3 \{AB, CDE, F\}$

$k = 4
\{AB, CD, E, F\}$

$k = 5
\{AB, C, D, E, F\}$

$k = 6
\{A, B, C, D, E, F\}$
Framework

Clustering

Intra prioritisation

Inter prioritisation

Generate best order

Evaluation
Clustering Based Prioritisation

- Interleaved Clusters Prioritisation (ICP)
  - Intra-cluster prioritisation
  - Inter-cluster prioritisation
  - Comparison limit: 100 pairs
Clustering Based Prioritisation

- Interleaved Clusters Prioritisation (ICP)
  - $n$ test cases, $k$ clusters
  - Pairs: $k(k-1)/2 + k(n/k)(n/k-1)/2$

Intra: coverage-based
Inter: human comparison

$P = 381 > 100$

$C(n,k), n=100, k=1,100$

$P = 91$

$k=14$
Analytic Hierarchy Process

- Analytic Hierarchy Process, AHP
  - Not transitive
  - Ratio-based

<table>
<thead>
<tr>
<th>$p_{ij}$</th>
<th>Preference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$i$ is equally preferable to $j$</td>
</tr>
<tr>
<td>3</td>
<td>$i$ is slightly preferable over $j$</td>
</tr>
<tr>
<td>5</td>
<td>$i$ is strongly preferable over $j$</td>
</tr>
<tr>
<td>7</td>
<td>$i$ is very strongly preferable over $j$</td>
</tr>
<tr>
<td>9</td>
<td>$i$ is extremely preferable over $j$</td>
</tr>
</tbody>
</table>
Analytic Hierarchy Process

- Comparison Matrix M

\[ \forall i (1 \leq i \leq n) \forall j (1 \leq j \leq n \land i \neq j), M(i, j) = p_{ij} \]

\[ M(i, i) = 1 (1 \leq i \leq n). \]

- Column normalized M

\[ M'(i, j) = \frac{M(i, j)}{\sum_{1 \leq k \leq n} M(i, k)} \]

- Priority weighting vector

\[ E_i = \frac{\sum_{1 \leq k \leq n} M(k, i)}{n} \]
Analytic Hierarchy Process

- **Ideal User Model**
  - $t_A$ detected $n_A$ faults
  - $t_B$ detected $n_B$ faults

<table>
<thead>
<tr>
<th>Condition</th>
<th>$p_{AB}$</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_A = n_B$</td>
<td>1</td>
<td>Equal</td>
</tr>
<tr>
<td>$n_A &gt; 0$ and $n_B = 0$</td>
<td>7</td>
<td>Very Strongly prefer $t_A$</td>
</tr>
<tr>
<td>$n_A &gt; 0$, $n_B &gt; 0$, $n_A \geq 3n_B$</td>
<td>9</td>
<td>Extremely prefer $t_A$</td>
</tr>
<tr>
<td>$n_A &gt; 0$, $n_B &gt; 0$, $n_A \geq 2n_B$</td>
<td>7</td>
<td>Very Strongly prefer $t_A$</td>
</tr>
<tr>
<td>$n_A &gt; 0$, $n_B &gt; 0$, $n_A \geq n_B$</td>
<td>5</td>
<td>Strongly prefer $t_A$</td>
</tr>
</tbody>
</table>

$$p_{BA} = \frac{1}{p_{AB}}$$
Analytic Hierarchy Process

- Human Error Model
  - Only type 1 ~ 6 considered

<table>
<thead>
<tr>
<th>Type</th>
<th>Original</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$p_{AB} &gt; 1$</td>
<td>$p'_{AB} = 1$</td>
</tr>
<tr>
<td>2</td>
<td>$p_{AB} &lt; 1$</td>
<td>$p'_{AB} = 1$</td>
</tr>
<tr>
<td>3</td>
<td>$p_{AB} &gt; 1$</td>
<td>$p'_{AB} &lt; 1$</td>
</tr>
<tr>
<td>4</td>
<td>$p_{AB} &lt; 1$</td>
<td>$p'_{AB} &gt; 1$</td>
</tr>
<tr>
<td>5</td>
<td>$p_{AB} = 1$</td>
<td>$p'_{AB} &gt; 1$</td>
</tr>
<tr>
<td>6</td>
<td>$p_{AB} = 1$</td>
<td>$p'_{AB} &lt; 1$</td>
</tr>
<tr>
<td>7</td>
<td>$p_{AB} &gt; 1$</td>
<td>$p'<em>{AB} &gt; 1$ and $p'</em>{AB} \neq p_{AB}$</td>
</tr>
<tr>
<td>8</td>
<td>$p_{AB} &lt; 1$</td>
<td>$p'<em>{AB} &lt; 1$ and $p'</em>{AB} \neq p_{AB}$</td>
</tr>
</tbody>
</table>
Analytic Hierarchy Process

- Pair-wise comparison

<table>
<thead>
<tr>
<th>Test Case</th>
<th>Branch 1 (Fault 1)</th>
<th>Branch 2 (Fault 2)</th>
<th>Branch 3 (Fault 3)</th>
<th>Branch 4 (Fault 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>$t_2$</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t_3$</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

- $(t_1, t_2, t_3)$ or $(t_1, t_3, t_2)$?
Analytic Hierarchy Process

- **Single criterion hierarchy: ICPs**
  - Pair-wise comparison from the human expert

- **Multi criteria hierarchy: ICPm**
  - Pair-wise comparison
  - Coverage-based prioritisation: scale of 3
  - Preference Value: \{9, 7, 5, 3, 1, 1/3, 1/5, 1/7, 1/9\}
Framework

Clustering

Intra prioritisation

Inter prioritisation

Generate best order

Evaluation
Evaluation

- Average Percentage of Fault Detection (APFD)

\[
APFD = 1 - \frac{TF_1 + \ldots + TF_m}{nm} + \frac{1}{2n}
\]

- T: n test cases; F: m faults
- T': the ordered T
- TF_i: the order of the first test case reveal the ith fault
Outline

- Background
- Motivation
- Framework
  - Clustering
  - Clustering-based Prioritisation
  - Analytic Hierarchy Process
  - Evaluation
- Experiments & Analysis
- Related Work
- Conclusions
# Experimental Setups

- **Subjects**
  - From Software Infrastructure Repository (SIR)

<table>
<thead>
<tr>
<th>Program</th>
<th>Test Suite</th>
<th>(Avg.) TS Size</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>printtokens</td>
<td>4</td>
<td>317.00</td>
<td>726</td>
</tr>
<tr>
<td>schedule</td>
<td>4</td>
<td>225.25</td>
<td>412</td>
</tr>
<tr>
<td>space</td>
<td>4</td>
<td>158.50</td>
<td>6,199</td>
</tr>
<tr>
<td>gzip</td>
<td>1</td>
<td>212</td>
<td>5,680</td>
</tr>
<tr>
<td>sed</td>
<td>1</td>
<td>370</td>
<td>14,427</td>
</tr>
<tr>
<td>vim</td>
<td>1</td>
<td>975</td>
<td>122,169</td>
</tr>
<tr>
<td>bash</td>
<td>1</td>
<td>1061</td>
<td>59,846</td>
</tr>
</tbody>
</table>
Results & Analysis

- RQ1: Effectiveness: ICP V.S. OP, SC

<table>
<thead>
<tr>
<th>Subject</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Suite</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>OP</td>
<td>0.991</td>
<td>0.995</td>
<td>0.993</td>
<td>0.993</td>
</tr>
<tr>
<td>ICP_s</td>
<td>0.824</td>
<td>0.917</td>
<td>0.952</td>
<td>0.913</td>
</tr>
<tr>
<td>SC</td>
<td>0.806</td>
<td>0.865</td>
<td>0.782</td>
<td>0.844</td>
</tr>
</tbody>
</table>

- OP: Optimal Ordering
- SC: Statement Coverage
- ICPs: ICP with single criteria

OP > ICPs > SC
## Results & Analysis

### RQ1: Effectiveness: ICP V.S. OP, SC

<table>
<thead>
<tr>
<th>Subject</th>
<th>schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test Suite</strong></td>
<td>1</td>
</tr>
<tr>
<td><strong>OP</strong></td>
<td>0.991</td>
</tr>
<tr>
<td>9</td>
<td>0.825</td>
</tr>
<tr>
<td>7</td>
<td>0.825</td>
</tr>
<tr>
<td>5</td>
<td>0.825</td>
</tr>
<tr>
<td><strong>ICPM</strong></td>
<td>3</td>
</tr>
<tr>
<td><strong>p[H][C]</strong></td>
<td>1</td>
</tr>
<tr>
<td>1/3</td>
<td>0.823</td>
</tr>
<tr>
<td>1/5</td>
<td>0.820</td>
</tr>
<tr>
<td>1/7</td>
<td>0.821</td>
</tr>
<tr>
<td>1/9</td>
<td>0.821</td>
</tr>
<tr>
<td><strong>SC</strong></td>
<td>0.806</td>
</tr>
</tbody>
</table>

- **OP**: Optimal Ordering
- **SC**: Statement Coverage-based ordering
- **ICPM**: ICP with multi criteria
Results & Analysis

- RQ2: Configuration: human V.S. coverage

<table>
<thead>
<tr>
<th>Subject</th>
<th>schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Suite</td>
<td>1</td>
</tr>
<tr>
<td>OP</td>
<td>0.991</td>
</tr>
<tr>
<td>9</td>
<td>0.825</td>
</tr>
<tr>
<td>7</td>
<td>0.825</td>
</tr>
<tr>
<td>5</td>
<td>0.825</td>
</tr>
<tr>
<td>ICPm</td>
<td>3</td>
</tr>
<tr>
<td>p[H][C]</td>
<td>1</td>
</tr>
<tr>
<td>1/3</td>
<td>0.823</td>
</tr>
<tr>
<td>1/5</td>
<td>0.820</td>
</tr>
<tr>
<td>1/7</td>
<td>0.821</td>
</tr>
<tr>
<td>1/9</td>
<td>0.821</td>
</tr>
<tr>
<td>SC</td>
<td>0.806</td>
</tr>
</tbody>
</table>

- **OP**: Optimal Ordering
- **SC**: Statement Coverage-based ordering
- **ICPm**: ICP with multi criteria
RQ3: Tolerance: highest tolerated error rate

- Higher APFD than SC till 0.5
- Higher APFD than SC till 1.0
Clustering with 14 clusters works
Any prioritisation better than random → improvement

HCRP: hierarchical clustering random prioritisation
Suitability Test

- **Suitability Test – Automated ICP**
  - Fault set: AR (Already Revealed)
  - TBR (To Be Revealed)
  - Intra & Inter cluster prioritisation on AR set
    - structural coverage
    - Fault information in AR
  - Result $\geq$ traditional way
  - Pair-wise comparison will do better on TBR
Suitability Test

Suitability Test configuration

<table>
<thead>
<tr>
<th>Program</th>
<th>Size of AR</th>
<th>Size of TBR</th>
<th>Mult. Ver.</th>
</tr>
</thead>
<tbody>
<tr>
<td>printtokens</td>
<td>3</td>
<td>4</td>
<td>No</td>
</tr>
<tr>
<td>schedule</td>
<td>4</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>space</td>
<td>18</td>
<td>20</td>
<td>No</td>
</tr>
<tr>
<td>gzip</td>
<td>2</td>
<td>3</td>
<td>Yes</td>
</tr>
<tr>
<td>sed</td>
<td>6</td>
<td>4</td>
<td>Yes</td>
</tr>
<tr>
<td>vim</td>
<td>4</td>
<td>3</td>
<td>Yes</td>
</tr>
<tr>
<td>bash</td>
<td>4</td>
<td>9</td>
<td>Yes</td>
</tr>
</tbody>
</table>
**Suitability Test**

- **RQ4: Suitability:** how accurately does the automated suitability test predict the successful result of ICP?

<table>
<thead>
<tr>
<th>Subject</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Suite</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>OP</strong></td>
<td>0.991</td>
<td>0.995</td>
<td>0.993</td>
<td>0.993</td>
</tr>
<tr>
<td><strong>NCS P AR</strong></td>
<td>0.899</td>
<td>0.974</td>
<td>0.922</td>
<td>0.949</td>
</tr>
<tr>
<td><strong>HCSP AR</strong></td>
<td>0.984</td>
<td>0.970</td>
<td>0.972</td>
<td>0.986</td>
</tr>
<tr>
<td><strong>NCS P TBR</strong></td>
<td>0.831</td>
<td>0.880</td>
<td>0.854</td>
<td>0.883</td>
</tr>
<tr>
<td><strong>ICP&lt;sub&gt;M&lt;/sub&gt; TBR</strong></td>
<td>0.994</td>
<td>0.992</td>
<td>0.992</td>
<td>0.992</td>
</tr>
</tbody>
</table>

- **OP:** Optimal Ordering
- **NCSP:** No clustering/Statement Prioritisation
- **HCSP:** Hierarchy clustering with Statement Prioritisation
- **ICP<sub>m</sub>:** ICP with multi criteria
Experiment summary

- Effectiveness
- Configuration
- Tolerance
- Suitability

Successful
Outline

- Background
- Motivation
- Framework
  - Clustering
  - Clustering–based Prioritisation
  - Analytic Hierarchy Process
  - Evaluation
- Experiments & Analysis
- Related Work
- Conclusions
Related Work

- Other prioritisation techniques -- Rothermel
  - Branch–total/additional, Statement–total/additional
  - Fault–Exposing Potential–total/additional
  - No single dominating criterion

- Other prioritisation + clustering usage -- Leon
  - Prioritizing by clustering execution profile
  - Better than coverage–based

- Other AHP applications – human preference
  - Karlsson: requirement prioritisation
  - Finnie: project management
  - Douligeris: Quality of Service
  - Tonella: Case–Base Ranking in test case prioritisation
Conclusion

Contributions
- A novel use of clustering
- A novel AHP–based prioritisation technique
- A more realistic user model by an error model
- An automated process of verifying effectiveness

Future work
- Different clustering criteria
Thanks for your attention!

Questions?