An Empirical Study on Clustering for Isolating Bugs in Fault Localization

Yanqin Huang¹, Junhua Wu², Yang Feng¹, Zhenyu Chen¹, Zhihong Zhao¹
Outline

- Introduction
- Assumption
- Our Approach
- Experimental Design
- Experimental Results
- Conclusion
Introduction

- SBFL: Spectrum-Based Fault Location
  - program spectrum
  - risk evaluation formulas
- Clustering: group tests
Assumption

- negligible effect of multiple faults
- test cases have some similar behaviors if they due to the same fault
Our Approach

- SBFL Concept

Figure: An example for SBFL
## Our Approach

### Current SBFL Techniques

<table>
<thead>
<tr>
<th>Group</th>
<th>Name</th>
<th>Formula Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER1</td>
<td>Naish2</td>
<td>$a_{ef} - \frac{a_{ep}}{a_{ep} + a_{np} + 1}$</td>
</tr>
<tr>
<td>ER2</td>
<td>Jaccard</td>
<td>$\frac{a_{ef}}{a_{ef} + a_{np} + a_{ep}}$</td>
</tr>
<tr>
<td>ER3</td>
<td>Tarantula</td>
<td>$\frac{a_{ef}}{a_{ef} + a_{nf}} / \left( \frac{a_{ef}}{a_{ef} + a_{nf}} + \frac{a_{ep}}{a_{ep} + a_{np}} \right)$</td>
</tr>
<tr>
<td>ER4</td>
<td>Wong2</td>
<td>$a_{ef} - a_{ep}$</td>
</tr>
<tr>
<td>ER5</td>
<td>Wong1</td>
<td>$a_{ef}$</td>
</tr>
<tr>
<td>ER6</td>
<td>Rogot1</td>
<td>$\frac{1}{2} \left( \frac{a_{ef}}{2a_{ef} + a_{nf} + a_{ep}} + \frac{a_{np}}{2a_{np} + a_{nf} + a_{ep}} \right)$</td>
</tr>
</tbody>
</table>
Our Approach

- Cluster Algorithms
  - Hierarchical clustering Algorithms
  - Partitioning clustering Algorithms e.g: simple k-means clustering
Our Approach

- Framework
  - Executing Test Cases
  - Collecting Profiles
  - Extracting BugHit Matrix
  - Applying Clustering
  - Using SBFL for Fault Localization
## Experimental Design

### Subject Programs

<table>
<thead>
<tr>
<th>Program name</th>
<th>Number of functions</th>
<th>Version</th>
<th>Test suite size</th>
</tr>
</thead>
<tbody>
<tr>
<td>flex</td>
<td>205</td>
<td>2.0</td>
<td>567</td>
</tr>
<tr>
<td>grep</td>
<td>188</td>
<td>3.0</td>
<td>809</td>
</tr>
</tbody>
</table>
Experimental Design

- Cluster Procedure
  - Measurement: \( n \)-dimensional Euclidean distance
    \[
    D(X, Y) = \sqrt{\sum_{i=1}^{n} d_i^2}
    \]

- Evaluation Metric
  - EXAM score method
Experimental Design

- Experimental Steps
  - Inject faults
  - Distinguish failures
  - Get execution profile
  - Cluster
  - Locate faults
Experimental Results

- Experimental Results on the flex program
Experimental Results on the grep program
Experimental Results

- Observations
  - simple k-means clustering mostly performs better than hierarchical clustering
  - Wong1 picked from ER5 always has the most excellent performance comparing to other techniques
Experimental Results

- Comparison of Wong1 from ER5 with Clustering Algorithms
Experimental Results

- Comparison of Wong1 and Naish2

The formula Naish2 is defined as follows:

\[ R_T(f_i) = a_{ef} - \frac{a_{ep}}{a_{ep} + a_{np} + 1} \]

The formula Wong1 is defined as:

\[ R_T(f_i) = a_{ef} \]
Conclusion

- ER5 (Wong1) achieves the best results of fault localization with clustering.
- K-means outperforms hierarchical clustering for isolating bugs in fault localization.
Thank you!