Are Mutation Scores Correlated with Real Fault Detection?

Abstract :

This study critically evaluates the effectiveness of mutation testing by investigating its correlation with actual fault detection in software testing. It highlights a common assumption in software testing research—that mutation scores, which are derived from artificial faults (mutants), closely correlate with the detection of real faults. However, by controlling for test suite size, the study discovers that the supposed strong correlation between mutation scores and real fault detection weakens, suggesting that the effectiveness of mutation scores as a predictor of real fault detection is overestimated. The study utilizes two significant datasets, CoreBench and Defects4J, encompassing large C and Java programs with real faults, to demonstrate that while mutation scores do influence fault detection, this influence is significantly less pronounced when the size of the test suite is accounted for. Nonetheless, achieving higher mutation scores is shown to improve fault detection capabilities, indicating that mutants can still provide useful guidance for enhancing test suites.

Chapter-by-Chapter Summary:

Introduction :

The introduction sets the stage by discussing the relevance of mutation testing in software quality assurance and the prevalent assumption of a strong correlation between mutation scores and the detection of real software faults. Despite widespread reliance on mutants for empirical validation in software testing studies, the exact nature of their relationship with real faults remains unclear and, to some extent, controversial. The section raises critical questions about the validity of using mutants as a proxy for real faults and outlines the motivations behind the study, which aims to clarify the relationship between mutation scores, test suite size, and real fault detection. This inquiry is positioned as essential for understanding the practical value of mutation testing in real-world software development scenarios.

2 Mutation Analysis

2.1 Mutants and Real Faults :

This subsection reviews previous studies on the relationship between mutants and real faults, summarizing findings that suggest both strong and weak connections. The analysis reveals mixed results across different studies, with some reporting strong correlations independent of test suite size, while others find correlations to weaken significantly when controlling for this variable. The literature review highlights the lack of consensus on the matter and sets the stage for a more comprehensive investigation.

2.2 Mutants and Hand-Seeded Faults :

Discusses the use of hand-seeded faults in studies due to the scarcity of real faults, highlighting the potential limitations of such an approach. The subsection compares mutation testing with other testing criteria like data-flow testing, uncovering inconsistencies in their effectiveness and cost.

2.3 Mutants and Defect Prediction :

Explores the emerging use of mutation scores as indicators of error-proneness and their role in defect prediction. It examines studies that found mutation scores to improve the accuracy of defect prediction methods, suggesting a valuable application of mutation testing beyond test suite evaluation.

2.4 Test Suite Size and Test Effectiveness :

Addresses the critical factor of test suite size in evaluating test effectiveness. It reviews studies that established a relationship between test suite size and fault detection capability, pointing out the non-linear and complex nature of this relationship. The discussion underscores the importance of considering test suite size when assessing the value of mutation scores and other testing criteria.

3 Experimental Procedure

3.1 Test Subjects :

Describes the selection of test subjects from the CoREBench and Defects4J datasets, chosen for their diversity and the presence of real faults. The section details the characteristics of these subjects, including program size and the number of faults, providing a solid foundation for the experimental analysis.

3.2 Fault Datasets (Defects4J and CoREBench) :

Explains the criteria for selecting the Defects4J and CoREBench fault datasets, emphasizing their relevance due to the inclusion of real faults. The procedure for isolating and reproducing these faults is outlined, ensuring the reliability of the experimental setup.

3.3 Test Suites :

Details the composition of test suites used in the study, combining developer-written tests with automatically generated tests to create a robust test pool. This mix aims to simulate real-world testing scenarios and enhance the validity of the study's findings.

3.4 Mutation Testing Tools :

Introduces the mutation testing tools employed in the study, chosen for their compatibility with the test subjects and their use in prior research. The tools' functionality and the mutation operators they support are briefly discussed.

3.5 Evaluation Metrics :

Defines the evaluation metrics used to assess the effectiveness of mutation testing, including mutation scores and fault detection capabilities. The methodology for applying these metrics in the study is outlined, clarifying how the research questions will be addressed.

3.6 Data Analysis :

Describes the statistical methods and analytical processes used to examine the data collected during the experiment. This includes regression analysis and correlation measures, which are crucial for interpreting the results and drawing meaningful conclusions.

4 Results

4.1 Visualisations :

Presents visual data illustrating the relationship between mutation scores, test suite size, and fault detection. The visualizations help clarify the impact of controlling for test suite size on the observed correlations.

4.2 Regression Models :

Reports the findings from regression analysis, showing how mutation scores and test suite size independently and jointly affect fault detection. The analysis provides statistical evidence supporting the study's main hypotheses.

4.3 Correlation Analysis with Test Suite Size Controlled/Uncontrolled :

Compares correlation results with and without controlling for test suite size, highlighting the significant difference in observed correlations. This comparison validates the study's premise that the relationship between mutation scores and fault detection is influenced by test suite size.

4.4 Fault Detection Probabilities :

Examines the actual fault detection capabilities of test suites with varying mutation scores, revealing that higher mutation scores correlate with improved fault detection, despite the weak overall correlation with real fault detection.

5 Discussion

5.1 Behavioral Similarity between Mutants and Real Faults :

Explores the extent to which mutants simulate the behavior of real faults, suggesting that while a small percentage of mutants closely resemble real faults, the majority do not. This discrepancy explains the weak correlation between mutation scores and actual fault detection.

5.2 Correlations vs. Fault Detection :

Differentiates between the correlation of mutation scores with fault detection and the actual fault detection capabilities of test suites. This distinction is crucial for interpreting the study's findings and understanding the practical implications for software testing research and practice.

6. Threats to Validity

Acknowledges potential limitations and biases in the study, including the generalizability of the results and the representativeness of the fault datasets. The section discusses measures taken to mitigate these threats and ensure the reliability of the findings.

7.Conclusions :

Summarizes the key findings of the study, emphasizing the nuanced relationship between mutation scores, test suite size, and real fault detection. The conclusion reiterates the value of mutation testing for improving test suites but cautions against overreliance on mutation scores as a direct measure of fault detection capability. The study calls for further research to develop more representative sets of mutants and refine mutation testing practices.