Detecting False Alarms from Automatic Static Analysis Tools: How Far are We?

Wonjong Kim , Tianyun Liu

Abstract

ASATs (Automatic Static Analysis Tools) : It helps find problems without running code.

A lot of warnings and 91% of False Positive

They tried to use the Machine Learning using “**Golden Features**” to detect False Alarms

However, they found that have critical drawbacks.

Data Leakage & Data Duplication

Abstract

Data Leakage :

It is when information about results to be predicted is unintentionally provided in advance.

Data Duplication :

It occurs when the training set and test set match.

→ The authors initiated this study to propose and evaluate alternative strategies that could overcome the inherent limitations of the Golden Feature and improve the filtering process of ASAT alerts.

The problem addressed in the paper

- **Problem 1** : The experiments that concluded the near-perfect performance of the Golden Features are problematic due to data leakage and duplication issues.

- **Problem 2** : The closed warning heuristic method used in the study for dataset creation has flaws. These flaws include a lack of demonstration of the robustness of the closed warning heuristic, and the labels derived from the closed warning heuristic are not always consistent with human judgment.

Motivation behind solving that problem

- **False Positives in ASATs**
- **Golden Features' Limitations**
- **Refining ASAT Accuracy**
- **Evaluating Closed Warning Heuristic**
- **Enhancing ASAT Filtering Process**
- **Improving Software Quality Assurance**

Proposed solution #1

- **Golden Features**

1. Try Golden Feature after removing Data Leakage and Data Duplication

Proposed solution #2

- **Closed Warning Heuristics**

1. Using revisions at different time intervals as reference versions
2. Filtering out undetermined actionable warnings and unknown false positives
3. Obtaining a clean dataset by removing these unknown and undetermined data for testing again

Evaluation in terms of the research questions addressed

- **Evaluation Metrics**: The study utilized the same metrics as prior studies for analyzing the performance of machine learning approaches that identify actionable Findbugs warnings.
- **Precision and Recall Calculation**: Precision and Recall were computed to measure the effectiveness of the model.
- **F1 Score**: Harmonic mean of precision and recall, crucial for balancing precision and recall in imbalanced datasets.
- **Area Under the Curve (AUC)**: This is used as a measure of the predictive power of a machine learning approach to distinguish between true and false alarms. It ranges from 0 (worst discrimination) to 1 (perfect discrimination).

Assumptions made

- For RQ1

  → Why are golden features effective?

- For RQ2

  → How suitable is the shutdown warning heuristic as a warning predictor?
  (The longer the time interval between test and reference revision, the greater the proportion of warnings that are closed.)

Novelty of the proposed solution considering state-of-the-art

A major innovation of the solution proposed by the authors is the creation of a more realistic data set to more accurately evaluate the performance of Golden Features SVM.

1. Implemented data deduplication
2. Reduces the possibility of data leakage

Limitations of the proposed solution

- Dependence on historical data
- Incomplete resolution of false alert

Practical significance of the proposed solution

- The study highlights the need for deeper study of the warning oracle to determine ground-truth labels.
- The study reveals how Golden Features SVM’s real performance
- The experiments show that the Golden Features SVM had improved performance on cleaner data.
- The study indicates opportunities and challenges for future work.

Discussion Point 1

- Challenges and strategies for ensuring the long-term maintenance and effectiveness of ASATs in a rapidly evolving software development environment.

Discussion Point 2

- A better way to handle "unknown actionable warnings" and "unknown false alarms" in experiments

Discussion Point 3

- Common issues between static analysis tools and dynamic analysis tools. Especially the Common issues between the study of the paper and Daikon.

Q & A

Thank You