On Comparing Mutation Testing Tools through Learning-based Mutant Selection

Milos Ojdanic, Ahmed Khanfir, Aayush Garg,
Renzo Degiovanni, Mike Papadakis, and Yves Le Traon

University of Luxembourg, Luxembourg

Conference: The 4th ACM/IEEE International Conference on Automation of Software Test (AST), Melbourne, Australia, May 15-16, 2023.
https://rdegiovanni.github.io/publications/files/AST2023.pdf

Recently many mutation testing tools have been proposed that rely on bug-fix patterns and natural language models trained on large code corpus. As these tools operate fundamentally differently from the grammar-based traditional approaches, a question arises of how these tools compare in terms of 1) fault detection and 2) cost-effectiveness. Simultaneously, mutation testing research proposes mutant selection approaches based on machine learning to mitigate its application cost. This raises another question: How do the existing mutation testing tools compare when guided by mutant selection approaches? To answer these questions, we compare four existing tools – μBERT (uses pre-trained language model for fault seeding), IBIR (relies on inverted fix-patterns), DeepMutation (generates mutants by employing Neural Machine Translation) and PIT (applies standard grammar-based rules) in terms of fault detection capability and cost-effectiveness, in conjunction with standard and deep learning based mutant selection strategies. Our results show that IBIR has the highest fault detection capability among the four tools; however, it is not the most cost-effective when considering different selection strategies. On the other hand, μBERT having a relatively lower fault detection capability, is the most cost-effective among the four tools. Our results also indicate that comparing mutation testing tools when using deep learning-based mutant selection strategies can lead to different conclusions than the standard mutant selection. For instance, our results demonstrate that combining μBERT with deep learning-based mutant selection yields 12% higher fault detection than the considered tools.