Assignment 2

James Halladay

Programming Languages

Dr Seif. Azghandi

Categorizing Bugs

The authors of this paper study different programming languages and seek to find what features of the languages contribute to the problems that often appear in code written in the language. They did this by studying the largest public code repository on the internet, github. Using commit logs, the authors created a dataset of several hundred thousand bugs based on what caused the bug and the impact it had in the program it was found in. To accomplish this task, the team employed the use of a supervised learning classifier called a Support Vector Machine or SVM after tuning the dataset and removing noisy values that might confuse the classifier. After training the classifier, they found it averaged an accuracy of 84% while even reaching 100% when classifying bugs caused by concurrency. Applying the classifier to the entire database provided by github then allowed the team to create larger databases of bugs and the languages associated with them to draw larger conclusions about what bugs certain features of a language may encourage.