Abstract

The rising complexity of software systems and agile development methods makes it increasingly difficult to control the quality of a software project. This makes a system for defect prediction desirable. We assume that machine learning offers the potential to realise such a solution. The goal of this bachelor thesis is to expand existing approaches by incorporating concepts originating in text analysis, especially N-Grams. Furthermore, a foundation for future work on this topic should be created. For this, a comprehensive and modular toolset was developed. It is able to analyse the Git repository of arbitrary Java projects. The extracted data can then be used as the basis for training a machine learning algorithm. With the resulting model, we try to predict how many bugfixes a file version will receive in the coming months. The implemented solution shows a significant correlation between the features used and the error-proneness of Java files. However, the results of our experiments could not conclusively prove the usefulness of N-Grams in defect prediction.

Keywords: Defect prediction, Machine Learning, Regression, N-Grams, Repository Mining, Software Metrics, Feature Design