Investigating the Effects of Balanced Training and Testing Datasets on Effort-Aware Fault Prediction Models

**Speaker**  Mr BENNIN Kwabena Ebo  
PhD Student  
Department of Computer Science  
City University of Hong Kong  
Hong Kong

**Date**  17 May 2016 (Tuesday)  
**Time**  2:30 pm - 3:00 pm  
**Venue**  CS Seminar Room, Y6405, 6th Floor  
Yellow Zone, Academic 1  
City University of Hong Kong  
83 Tat Chee Avenue  
Kowloon Tong

**Abstract**

To prioritize software quality assurance efforts, fault prediction models have been proposed to distinguish faulty modules from clean modules. The performances of such models are often biased and irrelevant due to the skewness or class imbalance of the datasets considered. To improve the prediction performance of these models, sampling techniques have been employed to rebalance the distribution of fault-prone modules and non-fault-prone modules. The effects of these techniques have been evaluated in terms of accuracy/geometric mean/F1-measure in previous studies; however, these measures do not consider the effort needed to fix faults. To empirically investigate the effect of sampling techniques on the performance of software fault prediction models in a more realistic setting, this study employs Norm( Popt ), an effort-aware measure that considers the testing effort. We performed two sets of experiments aimed at (1) assessing the effects of sampling techniques on effort-aware models and finding the appropriate class distribution for training datasets during application of sampling techniques (2) investigating the role of balanced training and testing datasets on performance of predictive models. Of the four sampling techniques applied, the over-sampling techniques outperformed the under-sampling techniques with Random Over-sampling performing best with respect to the Norm( Popt ) evaluation measure. Also, performance of all the prediction models improved when sampling techniques were applied between the rates of 20-30% on the training datasets implying that a strictly balanced dataset (50% faulty modules and 50% clean modules) does not result in the best performance for effort-aware models. Our results also indicate that performances of effort-aware models are significantly dependent on the proportions of the two types of the classes in the testing dataset. Models trained on moderately balanced datasets are more likely to withstand fluctuations in performance as the class distribution in the testing data varies.

This paper will be presented at the 40th IEEE Computer Society International Conference on Computers, Software & Applications, June 10-14, 2016, Atlanta, USA

Supervisor: Dr Jacky Keung

Research Interests: Defect prediction; Class imbalance learning and Data quality; Bug localization

All are welcome!

In case of questions, please contact Dr Jacky KEUNG at Tel: 3442 2591, E-mail: jacky.keung@cityu.edu.hk, or visit the CS Departmental Seminar Web at [http://www.cs.cityu.edu.hk/news/seminars/seminars.html](http://www.cs.cityu.edu.hk/news/seminars/seminars.html).