Chapter 36
Ensemble Techniques-Based Software Fault Prediction in an Open-Source Project

Wasiur Rhmann
Babasaheb Bhimrao Ambedkar University, Amethi, India

Gufran Ahmad Ansari
B. S. Abdur Rehman Crescent Institute of Science and Technology, India

ABSTRACT

Software engineering repositories have been attracted by researchers to mine useful information about the different quality attributes of the software. These repositories have been helpful to software professionals to efficiently allocate various resources in the life cycle of software development. Software fault prediction is a quality assurance activity. In fault prediction, software faults are predicted before actual software testing. As exhaustive software testing is impossible, the use of software fault prediction models can help the proper allocation of testing resources. Various machine learning techniques have been applied to create software fault prediction models. In this study, ensemble models are used for software fault prediction. Change metrics-based data are collected for an open-source android project from GIT repository and code-based metrics data are obtained from PROMISE data repository and datasets kc1, kc2, cm1, and pc1 are used for experimental purpose. Results showed that ensemble models performed better compared to machine learning and hybrid search-based algorithms. Bagging ensemble was found to be more effective in the prediction of faults in comparison to soft and hard voting.

1. INTRODUCTION

Data available in software repositories can help to improve various activities of the software development cycle. Mining software repository may be helpful for bug prediction, software testing, and maintenance (Xie et al., 2007). Software quality heavily depends on software testing. Software testing aims to find bugs in software. The identification of bugs earlier in the software development life cycle can reduce the total development cost of software (Lim & Goel, 2008).

DOI: 10.4018/978-1-7998-9158-1.ch036

Software fault prediction models are created with the historical data of the software projects and help to identify faulty modules before actual testing of the products. Identification of faulty modules can help to properly allocate resources for testing and maintenance. The quality of the software heavily depends on software testing, which is an integral part of software development. Software testing consumes around 40-50% of the cost and time of software development. Identification of faulty modules early in the development cycle can be very helpful to effectively allocate testing resources. Software fault prediction studies aim to predict faulty modules using historical data related to software projects. Software fault prediction uses various characteristics of software and the characteristics of software are measured using software metrics to predict fault-prone software modules. Different software metrics were proposed and used for software defect prediction (He et al., 2015). Various studies have been done to obtain a suitable subset of efficient software metrics for software bug prediction (Malhotra et al., 2010; Emam et al., 2001; Gyimothy et al., 2005). Software metrics are used to measure the characteristics of the software project, product, and process. Different types of software metrics have been used to predict different quality attributes of the software like change proneness, defectiveness. In recent years, metrics related to developers, organizations, and networks were used in software fault prediction (Caglayan et al., 2015). Software fault prediction studies have used Object-oriented metrics more than source code metrics or process metrics (Radjenovic et al., 2013).

Various machine learning techniques named Random forest, Support vector machine, Decision tree, Neural Network, etc., have been explored to predict fault-prone modules (Malhotra, 2015). Ensemble techniques combine more than one machine learning techniques and have been applied in various fields like spam detection (Singh & Batra, 2018), accident risk assessment (Kaeeni et al., 2018), bankruptcy prediction (Verikas et al., 2010), breast cancer classification (Nagarajan et al., 2017), etc., to obtain better accuracy. The novelty of this study is that change metrics obtained from open source android projects are used for software fault prediction using ensemble techniques and compared our fault prediction performance with the available static code metrics based fault prediction and hybrid search-based algorithm based fault prediction. The motivation behind the selection of change metrics for fault prediction is that they can be easily computed from different versions of the software obtained from Git repository while computation of static code metrics is difficult. No works have been reported to compare the performance of ensemble techniques based models with Hybrid Search-based algorithms (HSBA) for software fault prediction.

The present work addresses the followings research questions:

RQ1: How Ensemble models perform in software fault prediction using change-metrics?
RQ2: Are change-metrics based ensemble-models are better compared to the change metrics based ML model for fault prediction?
RQ3: Which method is better in predicting faulty modules? Hybrid Search-based algorithms or Ensemble-based models?

The main contributions of the present work are:

1. Assessing the predictive performances of Ensemble-based techniques in the prediction of faulty modules of the software;