Chapter 81
Combining Data Preprocessing Methods With Imputation Techniques for Software Defect Prediction

Misha Kakkar
Amity University Uttar Pradesh, Noida, India

Sarika Jain
Amity Institute of Information Technology, Amity University Uttar Pradesh, Noida, India

Abhay Bansal
Department of Computer Science and Engineering, Amity University Uttar Pradesh, Noida, India

P.S. Grover
KIIT Group of Colleges, Gurgaon, India

ABSTRACT

Software Defect Prediction (SDP) models are used to predict, whether software is clean or buggy using the historical data collected from various software repositories. The data collected from such repositories may contain some missing values. In order to estimate missing values, imputation techniques are used, which utilizes the complete observed values in the dataset. The objective of this study is to identify the best-suited imputation technique for handling missing values in SDP dataset. In addition to identifying the imputation technique, the authors have investigated for the most appropriate combination of imputation technique and data preprocessing method for building SDP model. In this study, four combinations of imputation technique and data preprocessing methods are examined using the improved NASA datasets. These combinations are used along with five different machine-learning algorithms to develop models. The performance of these SDP models are then compared using traditional performance indicators. Experiment results show that among different imputation techniques, linear regression gives the most accurate imputed value. The combination of linear regression with correlation based feature selector

DOI: 10.4018/978-1-7998-3016-0.ch081

outperforms all other combinations. To validate the significance of data preprocessing methods with imputation the findings are applied to open source projects. It was concluded that the result is in consistency with the above conclusion.

INTRODUCTION

Software Defect Prediction (SDP) models enable quality support teams to predict defect prone software artifacts in advance, which in-turn helps in effective resource allocation and utilization. The development of SDP model starts with data collection from various software repositories. This collected data is expected to be complete, however, it is sometimes incomplete and noisy. This incomplete data may be the result of cross-company nature of dataset or some computational errors (García et al., 2015; Lakshminarayan et al., 1999). The performance of SDP model developed from an incomplete dataset is questionable and also, many machine-learning algorithms won't be able to process such datasets.

There are many ways to deal with an incomplete dataset like - deletion techniques, toleration techniques, and imputation techniques. Deletion techniques recommend the deletion of all the instances, which include missing values, thus resulting in loss of important data. In toleration techniques, missing values are replaced by mean/ mode values, which is also not the best alternative method. Imputation is the most appropriate technique, which estimates missing values by analyzing the observed/available data. Researchers have proposed various imputation algorithms that are based on the accuracy of classifiers, which are trained using imputed values in the training dataset (Batista and Monard, 2003; Farhangfar et al., 2007; Saar-Tsechansky and Provost, 2007). Most of the imputation algorithms (Ma et al., 2006; Song et al., 2011) are trained under supervised learning, which uses complete dataset as the training dataset to compute missing values in the test dataset. Thus, complete dataset's quality affects the performance of the imputation technique, which in turn affects the performance of SDP model.

Data preprocessing techniques are used to deal with the other issue related to collected data, i.e. noisy data. Instance selection and feature selection are two significant data preprocessing steps, which aim to eliminate noisy data and reduce the size of data set by filtering out non-relevant software metrics (Gupta, 2013; Gao and Khoshgoftaar, 2014 and García et al., 2015; Kale). Feature Selection methods select most relevant software metrics, which contribute maximum to the prediction process. Instance selection methods select the most relevant instances, which contribute to the prediction process.

In this study, the authors investigate prediction capability of SDP model if either instance selection or feature selection is performed as an additional step in combination with the imputation technique. In other words, we examine which one of the two- instance selection or feature selection is more advantageous in the process of SDP model building with missing value dataset.

SDP dataset generally consist of software metrics which are numerical in nature; therefore, our focus will be on finding an imputation technique which work best for numerical datasets. There exist two similar studies (Huang et al., 2016; Tsai and Chang, 2016) with similar objectives. However, the study of Huang et al., 2016 deals with medical datasets and Tsai and Chang, 2016 used datasets from USI machine learning repositories. USI datasets used in these studies comprised of categorical, numerical as well as mixed attributes. Huang et al. (2016) concluded that imputation after instance selection gives better classification than imputation alone. Also, imputation after feature selection does not have positive impact on classification. Tsai and Chang (2016) concluded that Instance selection followed by