

An Empirical Evaluation of Performance of Machine Learning Techniques on Imbalanced Software Quality Data

Authors : Ruchika Malhotra, Megha Khanna

Abstract : The development of change prediction models can help the software practitioners in planning testing and inspection resources at early phases of software development. However, a major challenge faced during the training process of any classification model is the imbalanced nature of the software quality data. A data with very few minority outcome categories leads to inefficient learning process and a classification model developed from the imbalanced data generally does not predict these minority categories correctly. Thus, for a given dataset, a minority of classes may be change prone whereas a majority of classes may be non-change prone. This study explores various alternatives for adeptly handling the imbalanced software quality data using different sampling methods and effective MetaCost learners. The study also analyzes and justifies the use of different performance metrics while dealing with the imbalanced data. In order to empirically validate different alternatives, the study uses change data from three application packages of open-source Android data set and evaluates the performance of six different machine learning techniques. The results of the study indicate extensive improvement in the performance of the classification models when using resampling method and robust performance measures.

Keywords : change proneness, empirical validation, imbalanced learning, machine learning techniques, object-oriented metrics

Conference Title : ICITCS 2016 : International Conference on Information Technology and Computer Science

Conference Location : Venice, Italy

Conference Dates : April 11-12, 2016