Information and Software Technology 58 (2015) 388–402

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Software defect prediction using ensemble learning on selected features

Issam H. Laradji, Mohammad Alshayeb *, Lahouari Ghouti

Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

ARTICLE INFO

Article history:
Received 4 February 2014
Received in revised form 22 May 2014
Accepted 8 July 2014
Available online 24 July 2014

Keywords:
Defect prediction
Ensemble learning
Software quality
Feature selection
Data imbalance
Feature redundancy/correlation

ABSTRACT

Context: Several issues hinder software defect data including redundancy, correlation, feature irrelevance and missing samples. It is also hard to ensure balanced distribution between data pertaining to defective and non-defective software. In most experimental cases, data related to the latter software class is dominantly present in the dataset.
Objective: The objectives of this paper are to demonstrate the positive effects of combining feature selection and ensemble learning on the performance of defect classification. Along with efficient feature selection, a new two-variant (with and without feature selection) ensemble learning algorithm is proposed to provide robustness to both data imbalance and feature redundancy.
Method: We carefully combine selected ensemble learning models with efficient feature selection to address these issues and mitigate their effects on the defect classification performance.
Results: Forward selection showed that only few features contribute to high area under the receiver-operating curve (AUC). On the tested datasets, greedy forward selection (GFS) method outperformed other feature selection techniques such as Pearson's correlation. This suggests that features are highly unstable. However, ensemble learners like random forests and the proposed algorithm, average probability ensemble (APE), are not as affected by poor features as in the case of weighted support vector machines (W-SVMs). Moreover, the APE model combined with greedy forward selection (enhanced APE) achieved AUC values of approximately 1.0 for the NASA datasets: PC2, PC4, and MC1.
Conclusion: This paper shows that features of a software dataset must be carefully selected for accurate classification of defective components. Furthermore, tackling the software data issues, mentioned above, with the proposed combined learning model resulted in remarkable classification performance paving the way for successful quality control.

* Corresponding author.
E-mail addresses: g200790850@kfupm.edu.sa (I.H. Laradji), alshayeb@kfupm.edu.sa (M. Alshayeb), lahouari@kfupm.edu.sa (L. Ghouti).

http://dx.doi.org/10.1016/j.infsof.2014.07.005
0950-5849/