1. A comprehensive comparative study of clustering-based unsupervised defect prediction models

Xu, Zhou (1, 2); Li, Li (3); Yan, Meng (1, 2); Liu, Jin (4); Luo, Xiapu (5); Grundy, John (3); Zhang, Yifeng (4); Zhang, Xiaohong (1, 2)
Source: *Journal of Systems and Software*, v 172, February 2021; **ISSN:** 01641212; **DOI:** 10.1016/j.jss.2020.110862; **Article number:** 110862; **Publisher:** Elsevier Inc.
Author affiliation: (1) Key Laboratory of Dependable Service Computing in Cyber Physical Society (Chongqing University), Ministry of Education, China (2) School of Big Data and Software Engineering, Chongqing University, Chongqing, China (3) Faculty of Information Technology, Monash University, Australia (4) School of Computer Science, Wuhan University, Wuhan, China (5) Department of Computing, The Hong Kong Polytechnic University, Hong Kong, Hong Kong
Abstract: Software defect prediction recommends the most defect-prone software modules for optimization of the test resource allocation. The limitation of the extensively-studied supervised defect prediction methods is that they require labeled software modules which are not always available. An alternative solution is to apply clustering-based unsupervised models to the unlabeled defect data, called Clustering-based Unsupervised Defect Prediction (CUDP). However, there are few studies to explore the impacts of clustering-based models on defect prediction performance. In this work, we performed a large-scale empirical study on 40 unsupervised models to fill this gap. We chose an open-source dataset including 27 project versions with 3 types of features. The experimental results show that (1) different clustering-based models have significant performance differences and the performance of models in the instance-violation-score-based clustering family is obviously superior to that of models in hierarchy-based, density-based, grid-based, sequence-based, and hybrid-based clustering families; (2) the models in the instance-violation-score-based clustering family achieves competitive performance compared with typical supervised models; (3) the impacts of feature types on the performance of the models are related to the indicators used; and (4)the clustering-based unsupervised models do not always achieve better performance on defect data with the combination of the 3 types of features. © 2020 Elsevier Inc. (0 refs)
Main heading: Predictive analytics
Controlled terms: Defects - Forecasting - Open source software - Software testing
Uncontrolled terms: Alternative solutions - Comparative studies - Competitive performance - Defect prediction methods - Defect prediction models - Empirical studies - Software defect prediction - Test resource allocation
Classification Code: 723 Computer Software, Data Handling and Applications - 723.5 Computer Applications - 951 Materials Science
Funding Details: Number: 152239/18E, Acronym: -, Sponsor: -; Number: cstc2020jcyj-bshX0114, Acronym: -, Sponsor: -; Number: FL190100035, Acronym: ARC, Sponsor: Australian Research Council; Number: 61972290,62002034, Acronym: NSFC, Sponsor: National Natural Science Foundation of China; Number: 2020M673137, Acronym: -, Sponsor: China Postdoctoral Science Foundation; Number: -, Acronym: RGC, UGC, Sponsor: Research Grants Council, University Grants Committee; Number: -, Acronym: -, Sponsor: Natural Science Foundation of Chongqing; Number: 2018YFB2101200, Acronym: NKRDPC, Sponsor: National Key Research and Development Program of China; Number: 2020CDCGRJ072,2020CDJQY-A021, Acronym: -, Sponsor: Fundamental Research Funds for the Central Universities;
Funding text: The authors would like to thank Jaechang Nam for providing the source code of CLA and CLAMI. This work is supported by the National Key Research and Development Project (No. 2018YFB2101200), the National Natural Science Foundation of China (Nos. 62002034 , 61972290), the Fundamental Research Funds for the Central Universities (Nos. 2020CDJQY-A021 , 2020CDCGRJ072), China Postdoctoral Science Foundation (No. 2020M673137), the Natural Science Foundation of Chongqing in China (No. cstc2020jcyj-bshX0114), the Hong Kong Research Grant Council Project (No. 152239/18E). Grundy is supported by Australian Research Council Laureate Fellowship FL190100035 .The authors would like to thank Jaechang Nam for providing the source code of CLA and CLAMI. This work is supported by the National Key Research and Development Project (No. 2018YFB2101200), the National Natural Science Foundation of China (Nos. 62002034, 61972290), the Fundamental Research Funds for the Central Universities (Nos. 2020CDJQY-A021, 2020CDCGRJ072), China Postdoctoral Science Foundation (No. 2020M673137), the Natural Science Foundation of Chongqing in China (No. cstc2020jcyj-bshX0114), the Hong Kong Research Grant Council Project (No. 152239/18E). Grundy is supported by Australian Research Council Laureate FellowshipFL190100035.
Database: Compendex
Data Provider: Engineering Village