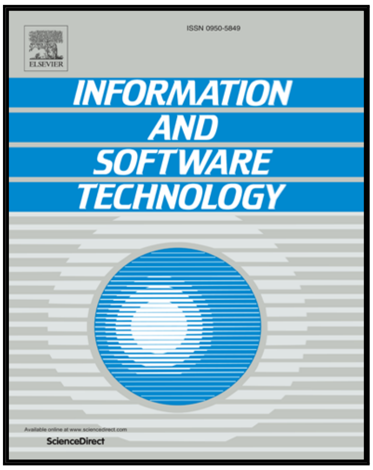
Accepted Manuscript



Cross Project Defect Prediction Using Class Distribution Estimation and Oversampling

Nachai Limsettho, Kwabena Ebo Bennin, Jacky W. Keung,
Hideaki Hata, Kenichi Matsumoto

PII:	S0950-5849(18)30053-3
DOI:	10.1016/j.infsof.2018.04.001
Reference:	INFSOF 5977
To appear in:	*Information and Software Technology*
Received date:	5 July 2017
Revised date:	9 February 2018
Accepted date:	3 April 2018

Please cite this article as: Nachai Limsettho, Kwabena Ebo Bennin, Jacky W. Keung, Hideaki Hata, Kenichi Matsumoto, Cross Project Defect Prediction Using Class Distribution Estimation and Oversampling, *Information and Software Technology* (2018), doi: 10.1016/j.infsof.2018.04.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Cross Project Defect Prediction Using Class Distribution Estimation and Oversampling

Nachai Limsettho[a], Kwabena Ebo Bennin[b,*], Jacky W. Keung[b], Hideaki Hata[a], Kenichi Matsumoto[a]

[a] *Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan*
[b] *Department of Computer Science, City University of Hong Kong, Hong Kong*

Abstract

Context: Cross-project defect prediction (CPDP) which uses dataset from other projects to build predictors has been recently recommended as an effective approach for building prediction models that lack historical or sufficient local datasets. Class imbalance and distribution mismatch between the source and target datasets associated with real-world defect datasets are known to have a negative impact on prediction performance.

Objective: To alleviate the negative effects of class imbalance and distribution mismatch on performance of CPDP models by using Class Distribution Estimation and Synthetic Minority Oversampling Technique. A novel approach called Class Distribution Estimation with Synthetic Minority Oversampling Technique (CDE-SMOTE) is proposed to optimize and improve the CPDP performance and avoid excessive oversampling.

Method: The proposed CDE-SMOTE employs CDE to estimate the class distribution of the target project. SMOTE is then used to modify the class distribution of the training data until the distribution becomes the reverse of the approximated class distribution of the target project. Four comprehensive experiments are conducted on 14 open source software projects.

[*] Corresponding author

Email addresses: nachai.limsettho.nz2@is.naist.jp (Nachai Limsettho), kebennin2-c@my.cityu.edu.hk (Kwabena Ebo Bennin), jacky.keung@cityu.edu.hk (Jacky W. Keung), hata@is.naist.jp (Hideaki Hata), matumoto@is.naist.jp (Kenichi Matsumoto)

ACCEPTED MANUSCRIPT

Download English Version:

https://daneshyari.com/en/article/6948018

Download Persian Version:

https://daneshyari.com/article/6948018

Daneshyari.com