Information and Software Technology 000 (2017) 1–16

Contents lists available at ScienceDirect

Information and Software Technology

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

Automated change-prone class prediction on unlabeled dataset using unsupervised method

Meng Yan[b], Xiaohong Zhang[a,b,*], Chao Liu[b], Ling Xu[b], Mengning Yang[b], Dan Yang[b]

[a] Key Laboratory of Dependable Service Computing in Cyber Physical Society Ministry of Education, Chongqing 400044, PR China
[b] School of Software Engineering, Chongqing University, Chongqing 401331, PR China

ARTICLE INFO

Article history:
Received 20 September 2016
Revised 17 February 2017
Accepted 4 July 2017
Available online xxx

Keywords:
Software maintenance
Change-prone prediction
Unlabeled dataset
Unsupervised prediction

ABSTRACT

Context: Software change-prone class prediction can enhance software decision making activities during software maintenance (e.g., resource allocating). Researchers have proposed many change-prone class prediction approaches and most are effective on labeled datasets (projects with historical labeled data). These approaches usually build a supervised model by learning from historical labeled data. However, a major challenge is that this typical change-prone prediction setting cannot be used for unlabeled datasets (e.g., new projects or projects with limited historical data). Although the cross-project prediction is a solution on unlabeled dataset, it needs the prior labeled data from other projects and how to select the appropriate training project is a difficult task.
Objective: We aim to build a change-prone class prediction model on unlabeled datasets without the need of prior labeled data.
Method: We propose to tackle this task by adopting a state-of-art unsupervised method, namely CLAMI. In addition, we propose a novel unsupervised approach CLAMI+ by extending CLAMI. The key idea is to enable change-prone class prediction on unlabeled dataset by learning from itself.
Results: The experiments among 14 open source projects show that the unsupervised methods achieve comparable results to the typical supervised within-project and cross-project prediction baselines in average and the proposed CLAMI+ slightly improves the CLAMI method in average.
Conclusion: Our method discovers that it is effective for building change-prone class prediction model by using unsupervised method. It is convenient for practical usage in industry, since it does not need prior labeled data.

1. Introduction

Software maintenance has been regarded as one of the most expensive and tough tasks in the whole software lifecycle [1]. Change is fundamental for software maintenance according to the technological advancements and new requirements. Managing and controlling change in software maintenance is one of the significant concerns of the software industry [2]. A change could be made because of existence of bugs, new features or refactoring [3,4]. It is the source of defects and modifications. Understanding the knowledge about change-prone classes in a software is significant for developers and mangers [5]. A change-prone class means that the class is likely to change with a high probability after a product release. It can represent the weak part of a software system [2]. Thus, software change-prone class prediction contributes to better allocation of software resources (e.g., time and staff) in the software maintenance process [6]. This technique aids to support maintenance related decision making by identifying change-prone classes in advance. As a result, the quality assurance teams or testers can determine the critical parts of the software where the quality assurance or testing activities should pay more attention and track rigorously.

In order to predict change-prone classes in advance, various categories of software metrics which indicate various characteristics have been proved to correspond to the change-proneness, such as OO metrics (e.g., cohesion, coupling, inheritance, etc.) [7], code smells [8], design patterns and [9] evolution metrics [10,11]. Based on these metrics, a number of studies which use machine learning techniques have been proposed for building change-prone class prediction models, such as Bayesian networks [12], neural networks [13], and ensemble methods [6]. A typical prediction model based on machine learning is designed by learning from historical labeled data within a project in a supervised way as Fig. 1(a) shows. This manner is referred as supervised within-project pre-

* Corresponding author at: School of Software Engineering, Chongqing University, Huxi Town, Shapingba, Chongqing, PR. China 401331.
E-mail address: xhongz@cqu.edu.cn (X. Zhang).

http://dx.doi.org/10.1016/j.infsof.2017.07.003
0950-5849/

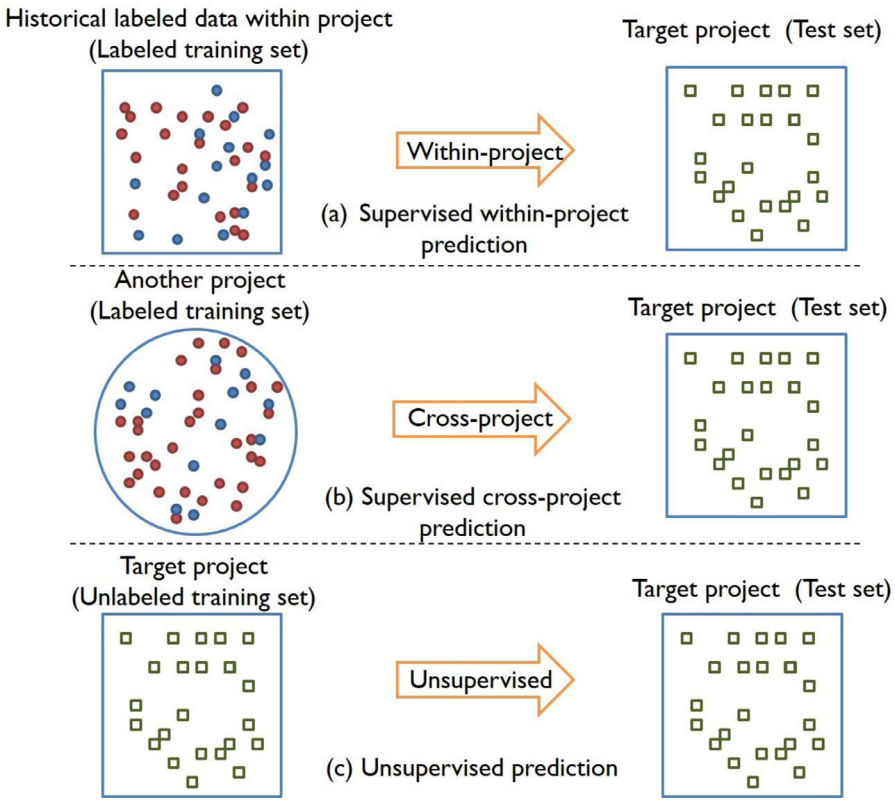

Fig. 1. Illustration of three prediction manners. Manner (a) is the supervised within-project prediction which training on historical labeled data and testing on target data within a project. Manner (b) is the cross-project prediction which training on another labeled project data and testing on the target project. Manner (c) is the unsupervised prediction which directly learning from the target project data.

diction [14]. Namely, the key idea is to train the model on historical labeled data within a project and then predict the target data. We refer the dataset which have historical labeled data as "*labeled dataset*". However, in practice, it is often time-consuming and expensive to collect labeled data. Furthermore, this manner is difficult to apply on new projects or projects with limited historical data whose label information are unavailable *(referred to as "unlabeled dataset"),* since it is difficult to collect label information for training a prediction model.

Cross-project change-prone class prediction method has been proposed to address the above-mentioned issue [14] as Fig. 1(b) shows. The cross project technique is motivated by the similar techniques in defect prediction [15,16]. It enables change-prone class prediction on unlabeled projects by learning from other projects which are already labeled. However, one issue which remains is that training set and testing set in cross-project prediction come from different project which possess different distributions [17]. The distribution similarity of training set and testing set is important for building a prediction model [18,19]. As a result, the success rate (ratio of combination whose performance is greater than a certain threshold) of cross-project prediction reported in the work [14] is generally poor (30%). Moreover, the cross-project change-prone prediction may not be effective and it depends on the selection of the source project [14].

To address the above-mentioned limitation, we propose to tackle this task by using unsupervised method as Fig. 1(c) shows. Compared with supervised models, unsupervised method does not need the prior labeled data to build prediction models which are more desirable in practice. It has been widely used in software quality prediction [17,20,21]. In detail, we apply a state-of-art unsupervised method (CLAMI: Clustering, Labeling, Metric selection and Instance selection) to the change-prone class prediction which has been successfully used in another field [17]. The key idea is to conduct the prediction on unlabeled dataset by learning from itself. Strictly, it is a special case of within-project manner. In this work, we use unsupervised refers to as the prediction without the need of historical labeled data particularly. Concretely, clustering is to group the instances, labeling is to estimate the label of groups by using a unsupervised way, metric selection and instance selection is to select more informative features and training sets. By the following, we predict the target set by training on the selected features and training sets.

The detailed process of the unsupervised method can be interpreted by dividing three phases as Fig. 2 shows. Each phase consists of two steps. The clue of the whole process is to build the prediction model by learning on selected informative metrics and instances from the target dataset itself. In detail, the first phase is clustering and labeling. In this phase, an unlabeled dataset is clustered into groups according to the difference between metric value and metric threshold. Subsequently, we estimate the labels of the dataset according to the magnitude of metric values [17]. The goal of this phase is to provide the estimated labels of all the instances. However, the estimated labels of all the instances might not be correct enough. In our unsupervised method, part of them will be automatically selected as final training set according to our criteria in the following phase. The second phase is to conduct the metric selection and instance selection from the labeled instances in the first phase. As a result, an informative training set of metrics and instances are generated. The third phase is modeling and prediction. The prediction model is built by learning from the selected instances and features in the second phase.

In particular, the labeling step in the first phase of the adopted method CLAMI is conducted by measuring the count of violation (i.e., a metric value is greater than a certain threshold) of an in-

Please cite this article as: M. Yan et al., Automated change-prone class prediction on unlabeled dataset using unsupervised method, Information and Software Technology (2017), http://dx.doi.org/10.1016/j.infsof.2017.07.003