Information and Software Technology 55 (2013) 1496–1511

Contents lists available at SciVerse ScienceDirect

Information and Software Technology

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

Analyzing and handling local bias for calibrating parametric cost estimation models

Ye Yang [a,*], Zhimin He [a,b], Ke Mao [a,b], Qi Li [c], Vu Nguyen [c], Barry Boehm [c], Ricardo Valerdi [d]

[a] Lab for Internet Software Technology, Institute of Software, Chinese Academy of Sciences, Beijing, China
[b] University of Chinese Academy of Sciences, Beijing, China
[c] Center for Systems and Software Engineering, University of Southern California, Los Angeles, USA
[d] Department of Systems and Industrial Engineering, University of Arizona, Tucson, USA

ARTICLE INFO

Article history:
Available online 20 March 2013

Keywords:
Parametric model
Effort estimation
Local bias
Weighted sampling
Model maintenance
COCOMO II

ABSTRACT

Context: Parametric cost estimation models need to be continuously calibrated and improved to assure more accurate software estimates and reflect changing software development contexts. Local calibration by tuning a subset of model parameters is a frequent practice when software organizations adopt parametric estimation models to increase model usability and accuracy. However, there is a lack of understanding about the cumulative effects of such local calibration practices on the evolution of general parametric models over time.
Objective: This study aims at quantitatively analyzing and effectively handling local bias associated with historical cross-company data, thus improves the usability of cross-company datasets for calibrating and maintaining parametric estimation models.
Method: We design and conduct three empirical studies to measure, analyze and address local bias in cross-company dataset, including: (1) defining a method for measuring the local bias associated with individual organization data subset in the overall dataset; (2) analyzing the impacts of local bias on the performance of an estimation model; (3) proposing a weighted sampling approach to handle local bias. The studies are conducted on the latest COCOMO II calibration dataset.
Results: Our results show that the local bias largely exists in cross company dataset, and the local bias negatively impacts the performance of parametric model. The local bias based weighted sampling technique helps reduce negative impacts of local bias on model performance.
Conclusion: Local bias in cross-company data does harm model calibration and adds noisy factors to model maintenance. The proposed local bias measure offers a means to quantify degree of local bias associated with a cross-company dataset, and assess its influence on parametric model performance. The local bias based weighted sampling technique can be applied to trade-off and mitigate potential risk of significant local bias, which limits the usability of cross-company data for general parametric model calibration and maintenance.

1. Introduction

Parametric cost estimation models need to be continuously calibrated and improved to assure more accurate software estimates and reflect changing software development contexts. In practice, typical parametric models for effort estimation are calibrated over a broad range of industry data, and many well-known models such as COCOMO II (CII), SEER-SEM, and PRICE-S also advocate local calibration to improve accuracy of model estimates [1–3]. In most cases, the estimation accuracy is noticeably increased after performing local calibration. Furthermore, it has become one of the best practices in existing industry/organization standards to conduct local calibration against local data of each individual organization [4,5]. Typical local calibration practices refer to the tuning of model coefficients and exponential constants against local historical data [1,6]. The tuning takes into account local characteristics in terms of: process activities, phases, measures of software size, and person-hours per person-month, and produces local constant parameters for future estimation usage. However, there is a lack of understanding about the cumulative effects of such local calibration practices on the performance of the parametric model over time. As time goes by, general model calibrated from old historical data may not well fit temporal projects, thus model maintainers need to continuously adjust model parameters by introducing newly collected data. The maintained general model is necessary by the industry especially for organizations that do not have sufficient historical data to conduct local calibration.

One disadvantage of local calibration is that organizations that adopt it are less bound to reach full compliance with the general

* Corresponding author. Tel.: +86 10 62661806 1196.
E-mail address: ye@itechs.iscas.ac.cn (Y. Yang).

0950-5849/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.infsof.2013.03.002

model [7]. Each organization has its own unique development environment and its own usage of estimation model, which is reflected by the organization's local historical data [1]. Such inconsistencies place great uncertainty on the usability of local data in calibrating the general model and make productivity comparisons between companies impossible. For example, as a success-proven parametric model, the recent calibration experiments of CII model have faced many issues [7,19]. Typically, they are counter-intuitive, i.e. negative values of regression coefficients and the limited prediction accuracy improvement. Such coefficients make little statistical or practical sense; for example, a higher level of programmer capability (PCAP) should lead to a decrease in the calculated project effort from general sense, but a negative coefficient that resulted from the statistical analysis would indicate that higher PCAP, in fact, the higher the effort level. Before reasonable explanation and consensus can be reached, such calibration results could hardly be acceptable.

In this paper, we suggest that the inconsistencies between the general parametric estimation model and the calibrated local models are a result of the local bias introduced by each individual organization's historical data. The aim of this paper is to contribute to a better understanding of local bias and its implications on the usage and calibration of general parametric models. We design and conduct three empirical studies to measure, analyze and address local bias in cross-company dataset, including: (1) defining a method for measuring the local bias associated with individual organization data subset in the overall dataset; (2) analyzing the impacts of local bias on the performance of estimation model; (3) proposing a weighted sampling approach to handle local bias. The studies are conducted on the latest COCOMO II 2010 calibration dataset.

To the best of our knowledge, there is no existing study on analysis and technique to address local bias and its impacts associated with the evolution cycle of parametric software cost estimation models. The main contributions of this study consist of the following:

(1) Providing a definition for consistently understanding and measuring local bias.
(2) Impact assessment and correlation analysis verify that local bias can be harmful to general model performance.
(3) Verifying the effectiveness of weighted sampling technique for handling local bias.
(4) Offering insights to ease parametric model evolution by identifying and avoiding local bias early on the data collection stage.

The rest of the paper is organized as follows: Section 2 describes related work. Section 3 describes the main research questions, basic assumptions, subject dataset, and analysis methods of our study. Section 4 presents the detailed analysis process and results on measuring local bias. Section 5 describes the process and results of assessing the impacts of local bias on model performance; Section 6 evaluates the effectiveness of weighted sampling technique for handling local bias when calibrating models on cross-company datasets. Section 7 discusses the analysis results and provides some improvement suggestions. Section 8 points out potential threats to validity. Finally, Section 9 concludes the paper with future work.

This paper extends our prior publication [35] on the 7th International Conference on Predictive Models in Software Engineering (PROMISE'11) in terms of:

(1) Updating model performance assessment approach. In this paper we employed the repeating hold-out strategy and median values to assess accuracy and stability of prediction results. Compared with performance assessment approach adopted in [35], the new approach can effectively shield effects of outliers. The new performance assessment indicators also provide an easier way to understand experiment results of our empirical studies.
(2) Adding empirical studies on handling local bias. We proposed to employ weighted sampling technique to help address the negative impacts of local bias on model calibration. Results of empirical studies show that the proposed approach can help to maintain accuracy and stability of calibration output.
(3) Providing more details and discussions about the experiment design and results. This paper provides more details of our empirical experiments on analyzing and handling local bias. Results analysis and discussion sections are also enriched compared with [35].

2. Related work

In the past decades, a variety of models have been proposed for software effort estimation. Typical models such as analogy based estimation models, parametric models, and machine learning based models have been widely adopted and validated in practice [1,31–33]. Among these models, parametric estimation models are primary formal methods for effort estimation for large projects. For example, the COCOMO II model has demonstrated its effectiveness over a great range of projects [1].

Global and local calibrations are necessary for model adoption and maintenance. However, calibrations of parametric models often introduce biases. According to opinions of Bottou, Hastie, and Wasserman et al., there are multiple sources of bias, including bias inherent in model structure, bias introduced by data sampling, and bias due to the optimization method used to improve model performance [8–10]. These biases may lead to serious problems to model adoption and maintenance, e.g., the counter-intuitive values of regression coefficients described in Section 1.

In addition to the biases described above, the mismatch of software development contexts among companies is also a primary source of bias. There has been much work on this issue. Kitchenham et al. systematically reviewed 10 such papers and concluded that models derived from within-company datasets perform significantly better than models derived from cross-company datasets [11]. Jeffery et al. compared the accuracy of estimation models derived from the ISBSG repository (a cross-company dataset) with those derived from the dataset of an Australian company (a within-company dataset) [12]. And they also concluded that models derived from the within-company dataset were more accurate.

Rather than model comparison, some studies focused on the preliminary analysis of datasets used to build an estimation model. Kitchenham [13] proposed a procedure for analyzing imbalanced datasets, and helped explain the difficult situation that happened when CII model was initially calibrated [5]. The most commonly used methods to pre-process data is eliminating factors and filtering out data points [21,23]. Based on forward pass residual analysis, the procedure identifies the most significant factors, and then produces a more statistically significant model. Another paper by Liu and Mintram [14] proposed a generic framework for preliminary analysis of cost estimation dataset. Using the framework, the analyst can systematically remove outliers and identify dominant variables.

Besides the nature of data and model calibration, model performance assessment is another incentive for obtaining and maintaining highly accurate and stable models. Port and Korte [24] and Menzies et al. [21] pointed out that the uncertainty of the predicted effort should be considered when evaluating estimation methods. Usually, confidence interval (e.g., the 90% confidence