Empirical Software Engineering manuscript No.
(will be inserted by the editor)

On the Relative Value of Cross-Company and Within-Company Data for Defect Prediction

Burak Turhan · Tim Menzies · Ayse Bener · Justin Distefano

Received: date / Accepted: date

Abstract We propose a practical defect prediction approach for companies that do not track defect related data. Specifically, we investigate the applicability of cross-company (CC) data for building localized defect predictors using static code features.

Firstly, we analyze the conditions, where CC data can be used as is. These conditions turn out to be quite few. Then we apply principles of analogy-based learning (i.e. nearest neighbor (NN) filtering) to CC data, in order to fine tune these models for localization. We compare the performance of these models with that of defect predictors learned from within-company (WC) data. As expected, we observe that defect predictors learned from WC data outperform the ones learned from CC data. However, our analyses also yield defect predictors learned from NN-filtered CC data, with performance close to, but still not better than, WC data. Therefore, we perform a final analysis for determining the minimum number of local defect reports in order to learn WC defect predictors. We demonstrate in this paper that the minimum number of data samples required to build effective defect predictors can be quite small and can be collected quickly within a few months.

Hence, for companies with no local defect data, we recommend a two-phase approach that allows them to employ the defect prediction process instantaneously. In phase one, companies should use NN-filtered CC data to initiate the defect prediction process and simultaneously start collecting WC (local) data. Once enough WC data is collected (i.e. after a few months), organizations should switch to phase two and use predictors learned from WC data.

Burak Turhan
Department of Computer Engineering, Bogazici University, Istanbul, Turkey
E-mail: turhanb@boun.edu.tr

Tim Menzies
Lane Department of Computer Science and Electrical Engineering, West Virginia
E-mail: tim@menzies.us

Ayse Bener
Department of Computer Engineering, Bogazici University, Istanbul, Turkey
E-mail: bener@boun.edu.tr

Justin Distefano
Lane Department of Computer Science and Electrical Engineering, West Virginia
E-mail: jdistefano@ismwv.com

KEYWORDS: defect prediction; learning; metrics (product metrics); cross-company; within-company; nearest-neighbor filtering