Noname manuscript No.
(will be inserted by the editor)

Finding GENERAL Defect Prediction Models Within Hundreds of Software Projects

**Suvodeep Majumder · Rahul Krishna ·
Tim Menzies**

Received: date / Accepted: date

Abstract Managers and practitioners become dubious about software analytics when its conclusions keep changing as we look at new projects. GENERAL is a new approach for quickly finding conclusions that generalize across hundreds of projects. This algorithm (a) removes spurious attributes via feature selection; (b) fixes training data imbalance via synthetic instances; (b) recursively clusters the project data; (c) finds the best model within any cluster, then promotes it up the cluster tree; (d) returns the model promoted to the top. GENERAL is much faster than prior methods (45 minutes versus 1294 minutes for our case studies) and theoretically scales better ($O(N^2/m)$ versus $O(N^2)$, which is a large reduction since often we find $m > 20$ clusters).

When tested on 756 Github projects, a single defect prediction model generalized over all those projects while also being useful and insightful and generalizable; i.e. that model worked just as well as 756 separate models learned from each project; and that model succinctly show what key factors most contributed to defects. Hence, when exploring hundreds of projects, we endorse GENERAL reasoning.

S. Majumder
Department of Computer Science North Carolina State University,
Raleigh, USA
E-mail: smajumd3@ncsu.edu

R. Krishna
Department of Computer Science,
North Carolina State University,
Raleigh, USA
E-mail: rkrish11@ncsu.edu

T. Menzies
Department of Computer Science,
North Carolina State University,
Raleigh, USA
E-mail: tim@ieee.org