Twenty First Century Software Effort Estimation Application Process

Jacky Keung
Department of Computing
The Hong Kong Polytechnic University
Kowloon, Hong Kong
Jacky.Keung@comp.polyu.edu.hk

Ekrem Kocaguneli, Tim Menzies
Lane Department of Computer Science and Electrical Engineering
West Virginia University
Morgantown, WV 26505, USA
ekocagun@mix.wvu.edu,
tim@menzies.us

ABSTRACT

In this paper, we are summarising our experience in the research of cost estimation, showing important aspects of successful cost estimation experiments.

Adding the research building blocks, and given that we are able to determine conclusion stability, a set of more completed general estimation application process should be used.
In the past :
Dataset -> Effort Estimator -> predictions
After a few years later, preprocessors such as Analogy-X, sensitivity analysis, feature subset selection etc surfaced
Dataset -> Preprocessor -> Effort Estimator -> predictions
Now we have a method (ASE Journal) for selecting the best combination of preprocessor + effort estimator
Dataset -> Method Selection -> Preprocessor (inc feature selection) -> Effort Estimator -> predictions
The paper is a summary and provides an up-to-date for the latest development of effort estimation research, this will help us to put our work in a leading stage.
This paper shows that we can propose a strong/weak relationship between datasets... Many datasets used by prior publications are very limited in number to distinguish *strong/weak* datasets.
Similarly, the maximum number of losses for any dataset over ninety algorithms is $89 \times 7 \times 90 = 56,070$. Figure **??** sorts all 20 data sets by their total losses in all seven performance criteria (expressed as a ratio of 50,070). For example, with the TELECOM dataset, all 90 methods rarely lost.
Figure **??** is somewhat a continuation of Figure **??**, in the sense that it deals with the stability of datasests. To test the stability, we question the mean of maximum rank change among datasets, when sorted w.r.t. win, tie, $win - loss$ over 7 error measures. Figure **??** shows that the maximum value of mean-rank change is 18, i.e. a method ranked as 2^{nd} in one scenario can rank as 20^{th} in another scenario. Therefore, with that amount of datasets, it is not healthy to propose *strong* or *weak* datasets that always attain lowest/highest performance values. If a dataset can change its position with a $+x$ or $-x$ amount, then there is a need for a window size of at least $2x$ and possibly some more datasets to actually observe how datasets would rank.
Our datasets could be sorted according to how well they can distinguish between effort estimators; for that matter, there is a need for more publicly available datasets.

1. REFERENCES