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

21^{st} Century Software Effort Estimation Application Process

Jacky W. Keung · Ekrem Kocaguneli ·
Tim Menzies

Received: 24 December 2010 / Accepted: Feburary 2011

Abstract 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 1 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 2 is somewhat a continuation of Figure 1, 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 2 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.

References

Jacky W. Keung
Department of Computing
The Hong Kong Polytechnic University
Kowloon, Hong Kong
E-mail: Jacky.Keung@comp.polyu.edu.hk

E. Kocaguneli and T. Menzies
Lane Department of Computer Science and Electrical Engineering
West Virginia University
Morgantown, WV 26505, USA
E-mail: ekocagun@mix.wvu.edu, tim@menzies.us

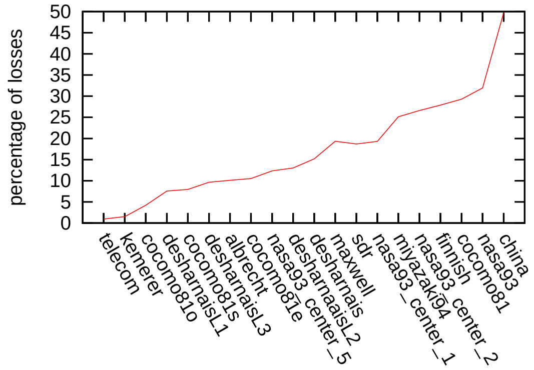

Fig. 1: Total losses seen in 20 datasets, expressed as a percentage of the maximum number of possible losses seen for one datasets (so 100%=50,070).

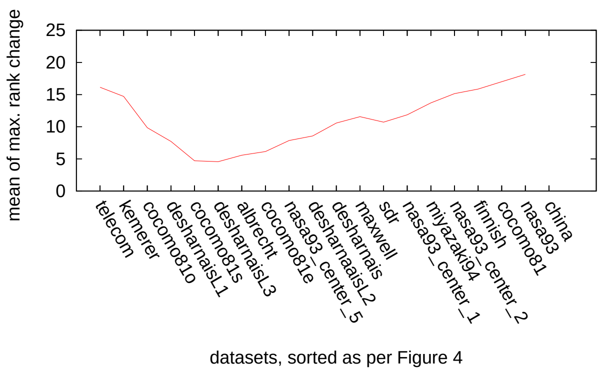

Fig. 2: Datasets and the mean of their maximum rank changes over all performance measures w.r.t. win, $loss$ and $win - loss$ values. Some datasets have lower rank-changes. However, the maximum mean-rank change is around 18 and we need more than $2 \times 18 = 36$ datasets to claim an order, hence strong/weak rekationship, between datasets.