Multi-objective Software Effort Estimation: A Replication Study (Supplementary Material)

Vali Tawosi, Federica Sarro, Alessio Petrozziello, Mark Harman
{*vali.tawosi, f.sarro, a.petrozziello, mark.harman*}*@ucl.ac.uk*

Abstract

This document is supplementary to the paper entitled "Multi- Objective Software Effort Estimation: A Replication Study", which is currently under revision at IEEE Journal of Transactions on Software Engineering.

1 Supplementary Results

Table 1 shows the results of the Mean Absolute Error (MAE) for all the algorithms we investigated in our paper on five datasets. The results show that CoGEE$_{NSGAII}$ outperforms the baseline methods (i.e. Random Guessing, Mean, and Median effort) with a large difference (i.e. CoGEE$_{NSGAII}$ is almost twice as accurate as the best baseline estimator on each of the datasets). Comparing CoGEE$_{NSGAII}$ with its R counterpart from the original study, we can see that both achieved a similar accuracy for all datasets, except for Desharnais. The MAE of CoGEE$_{NSGAII}$ is better than CBR for all datasets, better than LP for all but the Desharnais dataset, and better than CART for all but the Finnish dataset. CoGEE$_{NSGAII}$ produced more accurate results than the two single objective variants (i.e., GA-SAE and GA-CI) for all the datasets, except for Desharnais for which GA-SAE is the most accurate of the three. The other two-objective benchmark, NSGAII-UO, which optimises the two components of the Sum of Absolute Errors (SAE) separately, namely, the sum of under-estimates and the sum of overestimates, is outperformed by CoGEE$_{NSGAII}$ with a comfortably large difference. All the other multi-objective variants of CoGEE, except for CoGEE$_{IBEA}$, have similar accuracy performance.

For completeness, we report in Table 2 the SAE and CI average values obtained by the algorithms compared in RQ3.[1] We can observe that CoGEE$_{NSGAII}$ obtained a lower mean SAE than the one achieved by GA-SAE in one dataset (Finnish), comparable in three datasets (China, Maxwell and

[1] We warn the reader that only looking at the average of each of the optimised measures individually, does not capture how good a prediction model is with respect to the trade-off between the objectives optimised [1, 2], and can therefore be misleading. Otherwise, the Pareto Front's quality indicators allow us to quantify the overall quality of prediction models. In our paper [3], we used such indicators to measure the trade-off balance between multiple competing objective values (in our case, SAE and CI). Pareto Front's quality indicators are well-known in the multi-objective optimisation literature [4–7] and have been extensively used in previous software engineering work (see e.g., [7–11]). Therefore, we refer the reader to our paper for a comprehensive evaluation of the results.

Miyazaki) and higher in only one dataset (Desharnais). Similarly, when comparing the CI obtained by CoGEE$_{NSGAII}$ to those obtained GA-CI, the former achieved better results for three datasets (China, Dasharnais and Maxwell), comparable results on Miyazaki and worse results on only one dataset (Finnish). This means that optimising for both SAE and CI simultaneously helps the search algorithm find more accurate estimation models with a narrower confidence interval.

Additional results (Pareto Front plots) can be found at https://solar.cs.ucl.ac.uk/os/cogee.html.

Table 1: RQs1–2: The Mean Absolute Error (MAE) values achieved by CoGEE$_{NSGAII-R}$ (original study), CoGEE$_{NSGAII}$ (this replication), the baseline (Random, Mean and Median Effort), and state-of-the-art techniques (CBR1–3, LP, and CART) for each of the five datasets. For completeness, MAE results are also included for the other three alternative evolutionary algorithms considered in answer to RQ3 (i.e. GA-CI, GA-SAE, and NSGAII-UO) and four variants of CoGEE in answer to RQ5 (i.e. CoGEE$_{NSGAIII}$, CoGEE$_{SPEA2}$, CoGEE$_{MOCell}$, and CoGEE$_{IBEA}$).

China	MAE	Desharnais	MAE	Finnish	MAE	Maxwell	MAE	Miyazaki	MAE
CoGEE$_{MOCell}$	2431.04	CoGEE$_{NSGAII-R}$	2164.94	CART	3917.90	CoGEE$_{NSGAII-R}$	3749.23	CoGEE$_{NSGAII}$	6952.33
CoGEE$_{NSGAII}$	2579.63	GA-SAE	2259.06	CoGEE$_{SPEA2}$	4438.00	CoGEE$_{MOCell}$	3782.63	CoGEE$_{NSGAIII}$	6952.33
CoGEE$_{SPEA2}$	2592.39	LP	2307.10	CoGEE$_{NSGAII}$	4481.64	CoGEE$_{NSGAIII}$	3785.27	CoGEE$_{SPEA2}$	6952.33
CoGEE$_{NSGAII-R}$	2598.74	CoGEE$_{MOCell}$	2342.78	CoGEE$_{MOCell}$	4483.23	CoGEE$_{NSGAII}$	3795.38	CoGEE$_{NSGAII-R}$	6952.38
GA-SAE	2599.64	CoGEE$_{NSGAIII}$	2352.25	CoGEE$_{NSGAII-R}$	4489.26	CoGEE$_{SPEA2}$	3798.27	GA-SAE	6952.97
LP	2612.71	CoGEE$_{SPEA2}$	2363.78	CoGEE$_{NSGAIII}$	4576.80	GA-SAE	3809.65	GA-CI	6954.04
CoGEE$_{NSGAIII}$	2648.15	CoGEE$_{NSGAII}$	2390.25	GA-CI	4596.40	LP	4088.71	CoGEE$_{MOCell}$	6965.33
GA-CI	2764.37	CART	2532.58	GA-SAE	4759.69	CART	4175.29	LP	7410.01
CART	2991.62	CoGEE$_{IBEA}$	2540.11	CBR3	4775.22	CBR3	4210.56	NSGAII-UO	8692.27
CBR3	3008.54	GA-CI	2606.96	LP	4953.46	CoGEE$_{IBEA}$	4362.80	CBR3	8813.00
CoGEE$_{IBEA}$	3054.27	CBR3	2689.23	CoGEE$_{IBEA}$	4992.85	GA-CI	4410.29	CBR2	8814.44
Median	3115.27	Median	2726.62	CBR2	5060.20	CBR2	4512.16	CBR1	9162.33
CBR2	3224.24	CBR2	2763.67	CBR1	5626.68	Median	5696.71	CoGEE$_{IBEA}$	9766.35
CBR1	3532.56	CBR1	2964.35	Mean	6710.93	NSGAII-UO	6008.14	Median	10269.00
Mean	3716.46	Mean	3010.31	Median	6945.38	Mean	6202.24	CART	10900.98
Random	4986.03	Random	4084.79	NSGAII-UO	7443.31	CBR1	6285.81	Mean	14093.08
NSGAII-UO	5883.76	NSGAII-UO	4286.18	Random	8162.29	Random	8520.99	Random	20186.92

References

[1] T. Menzies, G. Gay, X. Devroey, and F. Sarro, "Proposed ACM SIGSOFT Standard for Optimization Studies in SE (including SBSE)." [Online]. Available: https://github.com/Greg4cr/sbse-sigsoft-standard

[2] G. Guizzo, F. Sarro, J. Krinke, and S. R. Vergilio, "Sentinel: A hyper-heuristic for the generation of mutant reduction strategies," *IEEE Transactions on Software Engineering*, 2020.

[3] V. Tawosi, F. Sarro, P. Alessio, and M. Harman, "Multi-objective software effort estimation: A replication study," *IEEE Transactions on Software Engineering*, vol. under review, 2021.

[4] E. Zitzler, L. Thiele, M. Laumanns, C. Fonseca, and V. da Fonseca, "Performance assessment of multiobjective optimizers: an analysis and review," *IEEE Transactions on Evolutionary Computation*, vol. 7, no. 2, pp. 117–132, 2003.

[5] C. A. C. Coello, G. B. Lamont, and D. A. V. Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems Second Edition*, 2nd ed. Springer Science, 2007.

Table 2: RQ3. Mean and standard deviation of the Sum of the Absolute Errors (SAE) and Confidence Interval (CI) values over 30 runs, for algorithms compared in RQ 3.

		SAE		CI	
Dataset	Algorithm	Mean	St. dev.	Mean	St. dev.
China	CoGEE$_{NSGAII}$	**1287234.9**	19281.96	357.7	10.68
	GA-SAE	1297212.1	17791.15	**340.1**	11.46
	GA-CI	1379418.3	19478.35	388.1	7.37
	NSGAII-UO	2935997.0	1085768.21	851.7	418.02
Desharnais	CoGEE$_{NSGAII}$	184049.1	3348.68	516.0	14.88
	GA-SAE	**173948.0**	2824.87	**495.2**	13.16
	GA-CI	200736.0	6429.21	552.9	25.62
	NSGAII-UO	330035.9	63416.07	645.6	195.55
Finnish	CoGEE$_{NSGAII}$	**170302.3**	6566.97	1184.5	45.35
	GA-SAE	180868.4	1017.65	1264.4	11.20
	GA-CI	174663.2	60.18	**1168.5**	0.51
	NSGAII-UO	282845.9	56036.45	1577.0	458.77
Maxwell	CoGEE$_{NSGAII}$	**235313.7**	2239.35	**929.9**	3.64
	GA-SAE	236198.4	3845.72	1019.1	30.40
	GA-CI	273437.8	9217.52	946.7	28.71
	NSGAII-UO	372504.7	12538.96	1151.3	38.92
Miyazaki	CoGEE$_{NSGAII}$	**333712.0**	0.00	**6327.0**	0.00
	GA-SAE	333742.4	97.21	6327.7	2.45
	GA-CI	333793.7	126.67	6328.0	3.28
	NSGAII-UO	417228.8	27075	7013.6	374.63

[6] Y. Cao, B. J. Smucker, and T. J. Robinson, "On using the hypervolume indicator to compare pareto fronts: Applications to multi-criteria optimal experimental design," *Journal of Statistical Planning and Inference*, vol. 160, pp. 60–74, 2015.

[7] M. Li, T. Chen, and X. Yao, "How to evaluate solutions in pareto-based search-based software engineering? a critical review and methodological guidance," *IEEE Transactions on Software Engineering*, pp. 1–1, 2020.

[8] G. Guizzo, G. M. Fritsche, S. R. Vergilio, and A. T. R. Pozo, "A Hyper-Heuristic for the Multi-Objective Integration and Test Order Problem," in *Proc. of GECCO'15*, 2015.

[9] G. Guizzo, S. R. Vergilio, A. T. Pozo, and G. M. Fritsche, "A multi-objective and evolutionary hyper-heuristic applied to the integration and test order problem," *Applied Soft Computing*, vol. 56, pp. 331–344, 2017.

[10] F. Ferrucci, M. Harman, J. Ren, and F. Sarro, "Not going to take this anymore: multi-objective overtime planning for software engineering projects," in *Proc. of 35th International Conference on Software Engineering, ICSE'13*, 2013, pp. 462–471.

[11] F. Sarro, F. Ferrucci, M. Harman, A. Manna, and J. Ren, "Adaptive multi-objective evolutionary algorithms for overtime planning in software projects," *IEEE Transactions on Software Engineering*, vol. 43, no. 10, pp. 898–917, 2017.