

The Influence of Population Size on Geometric Semantic GP

Mauro Castelli^{*1}, Luca Manzoni^{†2}, Sara Silva^{‡3}, and Leonardo Vanneschi^{§4}

¹*ISEGI, Universidade Nova de Lisboa, 1070-312 Lisboa, Portugal*

²*Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy*

³*LabMAg, FCUL, Universidade de Lisboa, 1749-016 Lisboa, Portugal*

⁴*ISEGI, Universidade Nova de Lisboa, 1070-312 Lisboa, Portugal*

Abstract

In this work we study the influence of the population size on the learning ability of Geometric Semantic Genetic Programming (GSGP) for the task of symbolic regression. The results show that having small populations results on a better training fitness with respect to the use of large populations after the same number of fitness evaluations. However, models obtained with large populations show a better performance on unseen data.

Introduction and Basic Notions. Geometric Semantic Genetic Programming (GSGP) is a new kind of genetic programming that has been recently introduced [1]. GSGP has been shown to be effective for solving many different problems (see, for example [2]). Here we investigate the influence of the population size on the dynamics of GSGP. Since it is equivalent to the onemax on the semantic space, GSGP does not need a population to converge to an optimal solution, i.e., one individual is sufficient. However, we are interested in understanding the influence of the population size on the convergence speed and on the generalization ability of GSGP.

Experiments and Results. We have compared different population sizes of GSGP on three real-world symbolic regression problems: %F, %PPB, and LD50. We refer the reader to [2] for a description of these datasets. In all cases we use the root means square error (RMSE) as the fitness and the data were split as 70% training data and 30% test data. The population sizes tested were 1, 2, 5, 10, 20, 50, and 100. In all cases a limit of 10^5 fitness evaluations was enforced. The crossover probability was fixed to 0.9 and the mutation probability to 0.5. The

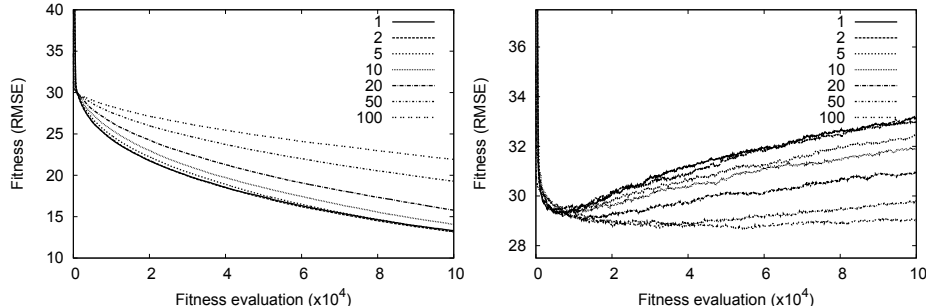
*mcastelli@isegi.unl.pt

†luca.manzoni@disco.unimib.it

‡sara@fc.ul.pt

§lvanneschi@isegi.unl.pt

selection phase was done using tournament selection with a tournament size of 4, the mutation step was 1, the maximum initial depth and the depth of the random trees was fixed to 6, the initialization was done with a ramped-half-and-half method, the functional symbols used were +, −, ×, and protected division, and random constants between −100 and 100 were allowed. We have performed 100 independent runs for each of the considered population size value.



Here we present only two plots. Both plots show, for all the considered population sizes and the %F problem, the median of the RMSE of the best individual across all the runs on the y axis, and the number of fitness evaluations on the x axis. On the left the results on the training set, and on the right the ones on the test set. The order in which the different population sizes appear in the plots is reversed: the best performer on the training set is the worst performer on the test set. Also in the %PPB dataset, having only one individual produces the best results on the training data, but a poor performance on unseen data, i.e., a small population does not generalize well. On the other hand, on the LD50 dataset, the most “difficult” problem, using a small population size produces the best results on both training and test data. To summarize, it seems that there exists a trade-off between speed of convergence and ability to generalize. This trade-off does not seem to hold for dataset in which it is difficult to generalize and, in that case, small populations seem to be preferable than larger ones.

Conclusions and Future Works. We plan to complete this study by carefully investigating the obtained results. In particular, we want to understand under what conditions it is better to rely on a small population instead of a bigger one. Furthermore, we want to investigate other possible effects related the use of different population sizes (e.g., variance of the performance between runs).

References

- [1] A. Moraglio, K. Krawiec, and C. G. Johnson. Geometric semantic genetic programming. In *Parallel Problem Solving from Nature - PPSN XII*, volume 7491 of *LNCS*, pages 21–31. Springer, 2012.
- [2] L. Vanneschi, S. Silva, M. Castelli, and L. Manzoni. Geometric semantic genetic programming for real life applications. In *Genetic Programming Theory and Practice XI*, pages 191–209. Springer, 2014.