Parsimony Measures in Multi-objective Genetic Programming for Symbolic Regression

Bogdan Burlacu
Gabriel Kronberger
Michael Kommenda
Josef Ressel Centre for Symbolic Regression
Heuristic and Evolutionary Algorithms Laboratory
University of Applied Sciences Upper Austria
Hagenberg, Austria
bogdan.burlacu@fh-hagenberg.at
gabriel.kronberger@fh-hagenberg.at
michael.kommenda@fh-hagenberg.at

Michael Affenzeller
Institute for Formal Models and Verification
Johannes Kepler University
Linz, Austria
Heuristic and Evolutionary Algorithms Laboratory
University of Applied Sciences Upper Austria
Hagenberg, Austria
michael.affenzeller@fh-hagenberg.at

ABSTRACT

We investigate in this paper the suitability of multi-objective algorithms for Symbolic Regression (SR), where desired properties of parsimony and diversity are explicitly stated as optimization goals. We evaluate different secondary objectives such as length, complexity and diversity on a selection of symbolic regression benchmark problems. Our experiments comparing two multi-objective evolutionary algorithms against standard GP show that multi-objective configurations combining diversity and parsimony objectives provide the best balance of numerical accuracy and model parsimony, allowing practitioners to select suitable models from a diverse set of solutions on the Pareto front.

CCS CONCEPTS

• **Computing methodologies** → **Search methodologies**; • **Theory of computation** → *Random search heuristics*; *Theory of randomized search heuristics*; • **Applied computing** → *Computer-aided design*;

KEYWORDS

genetic programming, symbolic regression, multi-objective optimization, parsimony, diversity

ACM Reference Format:
Bogdan Burlacu, Gabriel Kronberger, Michael Kommenda, and Michael Affenzeller. 2019. Parsimony Measures in Multi-objective Genetic Programming for Symbolic Regression. In *Proceedings of the Genetic and Evolutionary Computation Conference 2019 (GECCO '19).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3319619.3322087

1 INTRODUCTION

Symbolic regression (SR) is a grey-box modeling technique where an appropriate mathematical structure of the regression model is found by exploring the space of all possible expressions, usually by employing genetic programming to evolve an initially-random population of expression tree solution candidates.

Since the model structure is derived from data, SR typically tends to produce large, complex models that are hard to interpret and prone to overfitting. Model simplicity and interpretability are main requirements in industrial applications of symbolic regression, thus justifying approaches where these goals are explicitly stated as optimization objectives. In this context, we explore the possibility of using combinations of secondary objectives (eg., parsimony and diversity) to improve desired model characteristics.

2 METHODOLOGY

We employ the NSGA-II [2] and MOEA/D [7] algorithms together with a set of parsimony and diversity objectives. The algorithms differ from one another in the basic concept they employ for the search of Pareto optimal solutions.

The two algorithms are implemented in HeuristicLab [5] and utilize the same objective functions and genetic operators, differing only in their specific search logic.

The Multi-objective Evolutionary Algorithm based on Decomposition (MOEA/D) by Zhang and Li [7] decomposes a MOP into N scalar optimization subproblems, formulated via a scalarization approach using uniformly distributed weight vectors. Different decomposition methods are possible; we employ the Chebyshev approach with objective scaling as suggested in [7].

The Non-dominated Sorting Genetic Algorithm (NSGA-II) [2] uses the crowding distance between ranked non-dominated solutions to guide selection towards a uniformly spread Pareto front. It employs elitism by filling a new population each generation with the best solutions from both parent individuals and generated offspring. We adopt the NSGA-II algorithm with the adaptations for symbolic regression proposed by Kommenda et al. [3].

Secondary objectives such as length and complexity are intended to complement the usual fitness measure and help the algorithm to (i) evolve solutions faster by not having to process overly-large trees, and (ii) increase solution parsimony, leading to better interpretability and lower risk of overfitting.

We additionally use the standard GP algorithm as a baseline for comparison. All algorithms are configured with a population size

GECCO '19, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3322087

of 1000 individuals, 500 generations, 100% crosssover rate and 25% mutation rate. Tree individuals are limited to maximum depth 100 and maximum length 50.

We use a collection of three parsimony measures (tree length, visitation length and complexity) and combine them with a distance-based diversity measure.

- Tree length, to bias the search towards smaller models.
- Visitation length [4] was introduced by Smits and Kotanchek as a way to simultaneously favor smaller, flatter and more balanced structures
- Recursive complexity [3] by Kommenda et al. aims to produce simpler expressions by penalizing nesting of symbols inside the tree structure, as well as non-linear symbols
- Tree diversity [1] by Burlacu et al. employs tree hashing to identify isomorphic subtrees and defines a distance based on the degree of overlap between two trees. It promotes average distance within the population as a secondary objective.

We employ the listed objectives both individually and in pairs, using the Pearson's R^2 correlation coefficient as a main objective. We then use the hypervolume indicator H [8] to characterize multi-objective performance. Since the tested algorithmic configurations use different numbers of objectives, the resulting pareto fronts are mapped to the same two-dimensional objective spaces defined by (quality, length) and (quality, complexity). The goal is to identify Pareto fronts containing small, simple and numerically accurate solutions. A final value $H = \frac{H_L + H_C}{2}$ is aggregated from the (quality, length) and (quality, complexity) hypervolumes.

3 RESULTS

We perform empirical testing on a set of benchmark and real-world problems: Breiman-1, Friedman 1 & 2, Poly-10, Chemical and Housing data [6]. Detailed results are available online[1].

From a performance standpoint, the MOEA/D and NSGA-II algorithms are virtually indistinguishable on the tested problems, thus only NSGA-II is used for further discussion. The standard GP algorithm ranks behind most multi-objective configurations in terms of training performance and places last in the ranking based on generalization capability on test data.

Table 1 shows that diversity and parsimony are an effective combination that consistently produces high-quality results. At the same time the best-performing configurations produce more diverse Pareto fronts as reflected in their hypervolume rank.

4 CONCLUSIONS

We have shown that explicitly optimizing for desired model characteristics using multiple secondary objectives represents a viable approach. The combination of diversity and parsimony objectives seems particularly suited for producing high quality solutions and diverse Pareto fronts from which practitioners can select models that best suit their requirements.

[1] https://dev.heuristiclab.com/trac.fcgi/wiki/AdditionalMaterial#GECCO2019

Table 1: Median R^2 quality and hypervolume H rank over all problems, on training and test (inside parentheses) data.

Algorithm	Secondary objectives	R^2 rank	H rank
NSGA-II	visitation length, diversity	4 (5)	3.5 (2.5)
NSGA-II	length, diversity	4 (5)	7.5 (4.0)
NSGA-II	complexity, diversity	8 (6)	3.5 (4.0)
NSGA-II	complexity, visitation length	12 (8)	6.5 (6.5)
NSGA-II	diversity	5 (10)	18.0 (18.0)
NSGA-II	complexity, length	13 (10)	6.5 (8.5)
NSGA-II	complexity	17 (10)	11.5 (12.0)
NSGA-II	visitation length	14 (13)	10.0 (11.5)
NSGA-II	length	15 (14)	13.0 (13.0)
Standard GA	N/A	15 (18)	N/A

ACKNOWLEDGMENTS

The authors gratefully acknowledge support by the Christian Doppler Research Association and the Federal Ministry of Digital and Economic Affairs within the *Josef Ressel Centre for Symbolic Regression*

REFERENCES

[1] Bogdan Burlacu, Michael Affenzeller, Gabriel Kronberger, and Michael Kommenda. 2019. Online Dversity Control in Symbolic Regression via a Fast Hash-based Tree Similarity Measure. *Accepted for publication in 2019 IEEE Congree on Evolutionary Computation* abs/1902.00882 (2019). arXiv:1902.00882 http://arxiv.org/abs/1902.00882 Preprint.

[2] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. 2002. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (April 2002), 182–197. https://doi.org/10.1109/4235.996017

[3] Michael Kommenda, Gabriel Kronberger, Michael Affenzeller, Stephan Winkler, and Bogdan Burlacu. 2015. Evolving Simple Symbolic Regression Models by Multi-objective Genetic Programming. In *Genetic Programming Theory and Practice XIII (Genetic and Evolutionary Computation)*, Rick Riolo, William P. Worzel, M. Kotanchek, and A. Kordon (Eds.). Springer, Ann Arbor, USA. https://doi.org/doi:10.1007/978-3-319-34223-8

[4] Guido Smits and Mark Kotanchek. 2004. Pareto-Front Exploitation in Symbolic Regression. In *Genetic Programming Theory and Practice II*, Una-May O'Reilly, Tina Yu, Rick L. Riolo, and Bill Worzel (Eds.). Springer, Ann Arbor, Chapter 17, 283–299. https://doi.org/doi:10.1007/0-387-23254-0_17

[5] S. Wagner, G. Kronberger, A. Beham, M. Kommenda, A. Scheibenpflug, E. Pitzer, S. Vonolfen, M. Kofler, S. Winkler, V. Dorfer, and M. Affenzeller. 2012. Architecture and Design of the HeuristicLab Optimization Environment. In *First Australian Conference on the Applications of Systems Engineering, ACASE (Topics in Intelligent Engineering and Informatics)*, Robin Braun, Zenon Chaczko, and Franz Pichler (Eds.), Vol. 6. Springer International Publishing, Sydney, Australia, 197–261. https://doi.org/doi:10.1007/978-3-319-01436-4_10 Selected and updated papers.

[6] David R. White, James McDermott, Mauro Castelli, Luca Manzoni, Brian W. Goldman, Gabriel Kronberger, Wojciech Jaśkowski, Una-May O'Reilly, and Sean Luke. 2013. Better GP Benchmarks: Community Survey Results and Proposals. *Genetic Programming and Evolvable Machines* 14 (2013), 3–29. Issue 1. https://doi.org/10.1007/s10710-012-9177-2

[7] Q. Zhang and H. Li. 2007. MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. *IEEE Transactions on Evolutionary Computation* 11, 6 (Dec 2007), 712–731. https://doi.org/10.1109/TEVC.2007.892759

[8] Eckart Zitzler and Lothar Thiele. 1998. Multiobjective optimization using evolutionary algorithms — A comparative case study. In *Parallel Problem Solving from Nature — PPSN V*, Agoston E. Eiben, Thomas Bäck, Marc Schoenauer, and Hans-Paul Schwefel (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 292–301.