Analyzing the influence of Selection on Genetic Programming's Generalization ability in Symbolic Regression

A comparison of epsilon-lexicase Selection and Tournament Selection

Roman Hoehn, B.Sc. Wirtschaftspaedagogik

2022-06-29

Introduction

Experimental Study

Results

Conclusions

Limitations and open Questions

Introduction

Research Question

- Does the usage of ϵ-lexicase parent selection influence the generalization behaviour of genetic programming in symbolic regression if compared to tournament selection?

Genetic Programming

- A metaheuristic that searches for computer programs that solve a given problem
- Inventor: John R. Koza[1]
- Evolutionary algorithm that simulates the process of Darwinian evolution:
  1. Population based
  2. The quality of solutions is evaluated by a fitness function
  3. Selection: Solutions are selected based on their individual fitness
  4. Variation: Mutation and recombination of solutions
- Unique Features:
  - Evolve solutions of variable length and structure
  - Solutions are typically represented by recursive tree structures

[1] Koza (1992)

Parent Selection

- Operator that selects individual solutions from the population for reproduction and mutation
- Most commonly used selection operator in Genetic Programming (GP): Tournament selection[2]
- Intuition: High chance for “generalist” solutions to be selected since it is based on aggregated fitness scores

[2] Fang and Li (2010), p.181

epsilon-Lexicase Selection

- Recent alternative: Lexicase Selection and it's variation ϵ-lexicase selection
- Idea: Selection method for uncompromising, continous-valued symbolic regression problems[3]
- Increases genetic diversity inside the population[4]
- Higher chance for "specialist" solutions to be selected since it is decided on a per case basis
- Performance increases have been demonstrated in many benchmarking problems[5]

[3] Helmuth, Spector and Matheson (2015), p.12

[4] Helmuth, Spector and Matheson (2015), p.1

[5] La Cava, Spector and Danai (2016), p.744-745

Symbolic Regression

- Task: Find a mathematical model that fits a given set of datapoints
- One of the first applications of GP described by Koza (1992)
- High relevance: GP can outperform state-of-the-art machine learning algorithms like gradient boosting[6]

[6] Orzechowski, Cava and Moore (2018)

Generalization

- The ability of a model to perform well on previously unseen fitness cases
- Main objective in most supervised machine learning problems
- Challenge: Avoid overfitting to training data

Motivation

- Little attention has been paid to generalization in GP[7]
- High practical relevance of symbolic regression in many fields, e.g. financial forecasting

[7] O'Neill *et al.* (2010), Kushchu (2002)

Experimental Study

Benchmark problem

UC Irvine Machine Learning Repository: Prediction of energy efficiency in buildings[8]

Table 1: Overview - Energy Heating data set

Variable	Description
X1	Relative Compactness
X2	Surface Area
X3	Wall Area
X4	Roof Area
X5	Overall Height
X6	Orientation
X7	Glazing Area
X8	Glazing Area Distribution
y1	Heating Load
y2	Cooling Load

[8] Dua and Graff (2017)

Experiment

Single run

- Total dataset ($N = 768$) is randomly split into a training and testing dataset (50:50)
- Fitness metric: Mean squared Error (MSE)
- Train two models using GP with the training dataset only, one using tournament selection and the other ϵ-lexicase selection
- For each generation: Select elite model and compute its fitness for the testing dataset

Full experiment

- Stochastic algorithm: Repeat the basic experiment for 50 total runs
- Collect and aggregate results for training error, testing error and program length

Evolutionary Parameters

Table 2: Evolutionary Parameters

Parameter	Value
Population Size	500
Number of Generations	100
Mutation Rate	20%
Crossover Rate	80%
Tournament Size	3
Epsilon selection	automatic
Elite Size	0

Primitive Set I

Table 3: Function Set

Function	Arity
Addition	2
Subtraction	2
Multiplication	2
Negation	1
Sine	1
Cosine	1
Protected Division	2
Protected Natural Logarithm	1
Protected Square Root	1

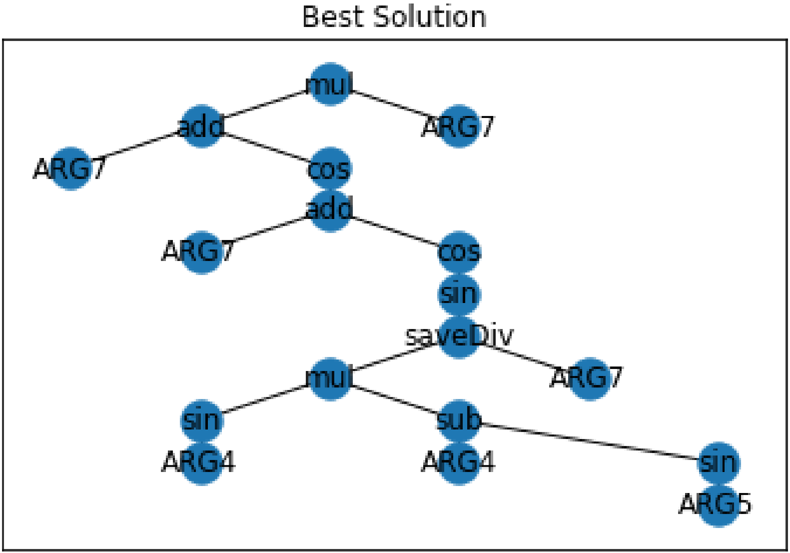
Primitive Set II

Table 4: Terminal Set

Terminal	Description
X1	Relative Compactness
X2	Surface Area
X3	Wall Area
X4	Roof Area
X5	Overall Height
X6	Orientation
X7	Glazing Area
X8	Glazing Area Distribution
random_int	Ephemeral Constant (integer)
random_float	Ephemeral Constant(float)

Example

Model evolved by tournament selection after 100 generations:



Research Question

- Does the usage of ϵ-lexicase parent selection influence the generalization behavior of genetic programming in symbolic regression if compared to tournament selection?

Statistical Testing Strategy

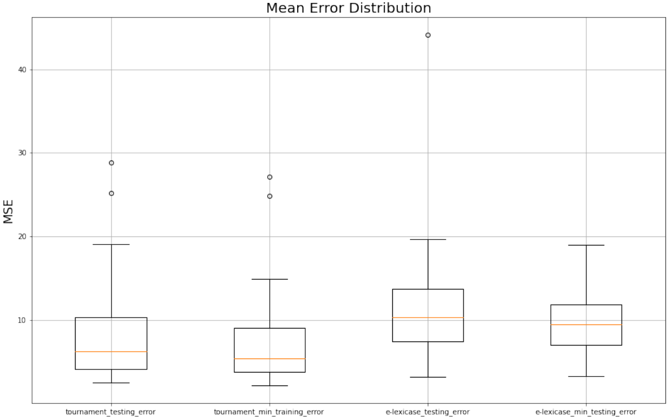
1. Differences in average fitness for both algorithms on both datasets?
2. Differences in fitness for training and testing data?

Results

Finding 1

- The differences in average fitness of the final solutions between tournament selection and ϵ-lexicase selection are highly statistical significant ($\alpha = 0.01$)
- Tournament selection-based GP achieves a higher fitness for both training and testing data
- Unexpected results based on the reviewed literature (La Cava, Spector and Danai, 2016), (La Cava *et al.*, 2017)

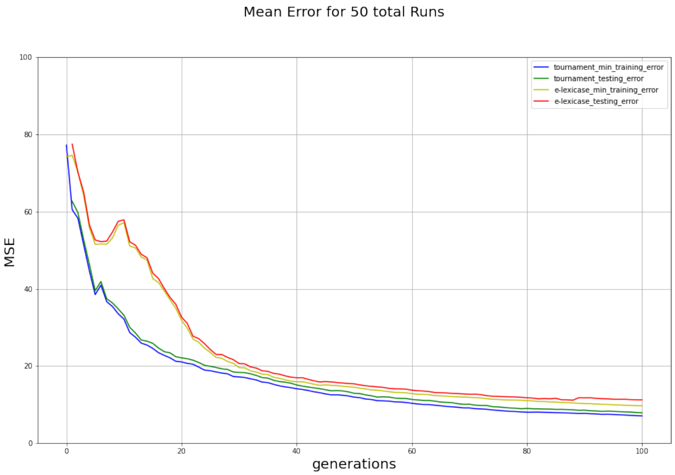
Distribution of Fitness



Finding 2

- The gap between training and testing error is not statistically significant for both selection algorithms
- Both algorithms achieve a slightly better performance for the training data
- Good generalization: No evidence of overfitting

Evolution of Fitness



Statistical Test

Table 5: Mean Error - P-Values (MWU)

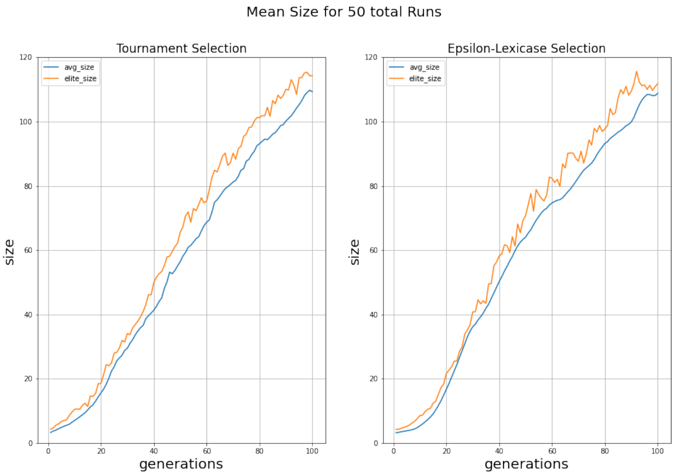
X	tournament_training_errors	tournament_testing_errors	elexicase_training_errors	elexicase_testing_errors
tournament_training_errors	1.000	0.309	0.000	0.000
tournament_testing_errors	0.309	1.000	0.002	0.000
elexicase_training_errors	0.000	0.002	1.000	0.257
elexicase_testing_errors	0.000	0.000	0.257	1.000

Program Growth

- So far: No proof of differences in generalization
- New approach: Program growth as a possible indicator for overfitting?
- Theory: Minimum description length principle (MDLP)[9]
- Downside: Growth/Bloat is no clear indicator of overfitting[10]

[9] Wang, Wagner and Rondinelli (2019), p. 268
[10] Silva and Vanneschi (2009), p. 8

Evolution of Size



Finding 3

- GP typical growth behaviour for both operators
- Solutions grow at a similiar rate in each generation
- No statistically significant differences in overall program size based on selection

Conclusions

Conclusions

- Experiment did not yield evidence for differences in generalization behavior between tournament and ϵ-lexicase selection
- The performance of tournament selection is significantly higher than that of ϵ-lexicase selection for the selected symbolic regression problem
- No evidence for differences in growth behavior between both algorithms

Limitations and open Questions

Limitations and open Questions

1. Configuration of evolutionary parameters
2. Results are based on a single symbolic regression
3. Limited by computational resources

References

References I

Dua, D. and Graff, C. (2017) 'UCI machine learning repository'. University of California, Irvine, School of Information; Computer Sciences. Available at: http://archive.ics.uci.edu/ml.

Fang, Y. and Li, J. (2010) 'A review of tournament selection in genetic programming', in Cai, Z. et al. (eds) *Advances in computation and intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 181–192.

Helmuth, T., Spector, L. and Matheson, J. (2015) 'Solving uncompromising problems with lexicase selection', *IEEE Transactions on Evolutionary Computation*, 19(5), pp. 630–643. doi:10.1109/TEVC.2014.2362729.

Koza, J.R. (1992) *Genetic programming: On the programming of computers by means of natural selection*. Cambridge, MA, USA: MIT Press. Available at: http://mitpress.mit.edu/books/genetic-programming.

References II

Kushchu, I. (2002) 'An evaluation of evolutionarygeneralisation in genetic programming', *Artificial Intelligence Review - AIR*, 18, pp. 3–14. doi:10.1023/A:1016379201230.

La Cava, W. *et al.* (2017) 'A probabilistic and multi-objective analysis of lexicase selection and epsilon-lexicase selection'. arXiv. doi:10.48550/ARXIV.1709.05394.

La Cava, W., Spector, L. and Danai, K. (2016) 'Epsilon-lexicase selection for regression', in *Proceedings of the genetic and evolutionary computation conference 2016*. New York, NY, USA: Association for Computing Machinery (GECCO '16), pp. 741–748. doi:10.1145/2908812.2908898.

O'Neill, M. *et al.* (2010) 'Open issues in genetic programming', *Genetic Programming and Evolvable Machines*, 11, pp. 339–363. doi:10.1007/s10710-010-9113-2.

References III

Orzechowski, P., Cava, W.L. and Moore, J.H. (2018) 'Where are we now?', in *Proceedings of the genetic and evolutionary computation conference*. ACM. doi:10.1145/3205455.3205539.

Silva, S. and Vanneschi, L. (2009) 'Operator equalisation, bloat and overfitting: A study on human oral bioavailability prediction', in *Proceedings of the 11th Annual Genetic and Evolutionary Computation Conference, GECCO-2009*, pp. 1115–1122. doi:10.1145/1569901.1570051.

Wang, Y., Wagner, N. and Rondinelli, J.M. (2019) 'Symbolic regression in materials science', *MRS Communications*, 9(3), pp. 793–805. doi:10.1557/mrc.2019.85.