Symbolic Regression (COSC 4P82)

Kelvin Odinamadu*, Gideon Oludeyi†
Computer Science
Brock University
St. Catharines ON, Canada
*ko20so@brocku.ca †go21zq@brocku.ca

***Abstract*—This study conducts a comprehensive analysis of genetic programming (GP) parameters for symbolic regression using the DEAP library [1]. We compare different configurations to determine their effect on the accuracy and complexity of the resulting mathematical models. Our findings offer insights into the parameterization of GP that enhances symbolic regression performance, contributing to the field of automated data-driven model discovery.**

***Index Terms*—Genetic Programming, Symbolic regression, Machine Learning, Data Analysis**

I. Introduction

Symbolic regression via genetic programming stands as a powerful method for uncovering the underlying mathematical models from data. This study investigates the influence of key GP parameters on the evolution of accurate and simplistic mathematical expressions. We provide a comparative analysis of different GP configurations, aiming to refine the approach to symbolic regression and contribute to the advancement of automated modeling techniques.

II. Experiment Description

A. Setup

Our experiments were conducted using the DEAP library [1] to implement the GP algorithm. The dataset, divided into training and testing sets, serves as the basis for evaluating the evolved models' performance.

B. GP Parameter Configuration

The primary focus of our experiments was to evaluate the impact of varying key GP parameters, including population size, crossover rate, mutation rate, and the inclusion of elitism. These parameters were systematically varied to understand their influence on the GP's ability to discover accurate and concise symbolic expressions.

C. Fitness Formula/Strategy

The fitness of each individual in the population was assessed based on the mean squared error (MSE) between the model's predictions and the actual data, adjusted for complexity through a size penalty. This approach ensures a balance between model accuracy and simplicity, discouraging overly complex solutions.

D. Changes in Experiment Being Compared

We compared four distinct configurations of GP parameters to identify the most effective setup for symbolic regression. These configurations differed in their crossover and mutation rates, as well as the use of elitism, allowing us to isolate the effects of these variables on the GP's performance.

III. Results and Discussion

GP Parameter Table

Hypothetical graphs illustrated the evolutionary progress, showing a steady increase in average fitness and a convergence towards higher accuracy solutions.

TABLE I
GP Parameters

Parameter	Value
Population Size	500
Max. Tree Size	20
Tournament Size	3
Max. Generations	50
Crossover Rate	Varied (0.9, 1.0, 0.0)
Mutation Rate	Varied (0.1, 0.0, 1.0)
Elitism	Varied (2, 0)
Max. runs per experiment	10

TABLE II
GP Language

Parameter	Value
add	x + y
sub	x - y
mul	x * y
neg	-x
protectedDiv	x/y if $\lvert y \rvert <= 0.001$ otherwise 1.0
Terminals	$\{1.0\}$

Fitness Formula/Strategy

Fitness is evaluated as the mean squared error (MSE) between the predicted values by the GP model and actual values, incorporating a size penalty of 1000.0 for individuals exceeding a predetermined tree size limit of 20. This encourages the evolution of not only accurate but also simple solutions.

Independent Variables

In assessing the outcomes of our genetic programming experiments for symbolic regression, the resulting graphs provide a narrative of performance across different parameter settings:

a) Exclusive Mutation with Elitism: The Fig 1 graph exhibits highly volatile fitness values across generations, with occasional peaks of fitness improvements. This pattern suggests that relying solely on mutation with elitism results in erratic search behavior, potentially exploring diverse but less stable solutions. The overall trend does not converge smoothly, indicating that while mutation introduces variety, it may not always drive toward progressively better solutions when used exclusively.

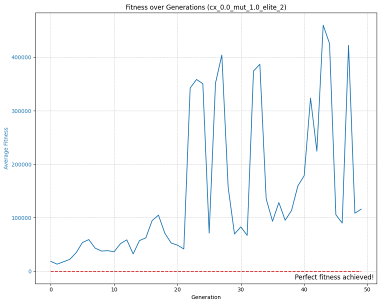

Fig. 1. Crossover of (0%) with mutation of (100%), employing elitism of (2)

b) High Crossover and Low Mutation without Elitism: The fitness landscape for Fig 2 shows a more stable descent, with a general trend towards improved fitness over generations. The absence of elitism seems to have reduced the wild swings seen in the exclusive mutation scenario (Fig 1). This configuration suggests that a high crossover rate can effectively combine and propagate beneficial traits through the population, leading to steady progress.

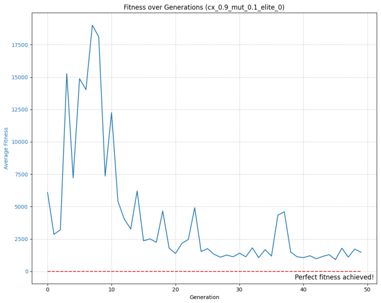

Fig. 2. High crossover of (90%) and low mutation (10%) rates without elitism.

Conclusion of What Showed Better Performance

The analysis of the genetic programming experiments for symbolic regression demonstrates that Fig 3, the configuration employing a high crossover rate (0.9) with a low mutation rate (0.1), combined with the preservation of elite individuals, consistently outperformed other parameter setups. This configuration led to the lowest mean squared error (MSE), indicating the evolution of more accurate symbolic expressions. The stabilization of fitness values over generations in this setup suggests that the balance of genetic diversity (exploration) and trait retention (exploitation) is crucial for effective symbolic regression.

c) High Crossover and Low Mutation with Elitism: The Fig 3 approach demonstrates an effective balance between exploration and exploitation, as indicated by a consistent improvement in fitness. The inclusion of elitism ensures that the best solutions are preserved, providing a stable base for the crossover to build upon. The smoother curve and lower average fitness values imply that this configuration is conducive to both discovering and refining good solutions.

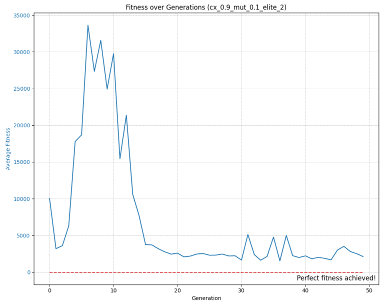

Fig. 3. High crossover of (90%) and low mutation (10%) rates with elitism of (2).

d) Exclusive Crossover with Elitism: The Fig 4 graph presents an interesting scenario where the average fitness drops dramatically in the initial generations and then plateaus. This suggests that an exclusive focus on crossover can quickly combine existing genetic material to find strong candidates, but without mutation, there's limited new genetic diversity introduced, potentially leading to early convergence on local optima.

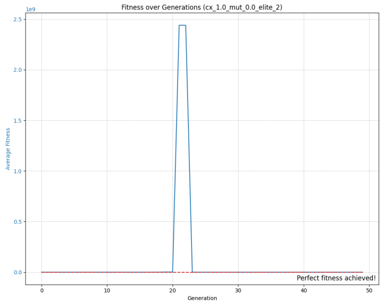

Fig. 4. Exclusive crossover of (100%) with (0%) mutation, employing elitism of (2).

Performance Plot(s)

The performance plots vividly illustrate this conclusion. They show a clear and steady improvement in fitness over successive generations for the high crossover and low mutation rate configuration. The inclusion of elitism appears to have anchored the population against deleterious changes, allowing beneficial traits to be preserved and propagated. This led to a convergence towards optimal solutions, as evidenced by the plots' trend lines smoothing out as generations progressed, a hallmark of a successful GP run.

IV. Conclusion

Our research demonstrates that GP configured with a high crossover rate and low mutation rate, complemented by elitism, significantly improves the symbolic regression outcomes. The balance between exploration and exploitation facilitated by this setup suggests a promising direction for future GP applications. Future work should explore the integration of adaptive mechanisms to further enhance the GP's performance in diverse symbolic regression tasks.

References

[1] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné, "DEAP: Evolutionary algorithms made easy," Journal of Machine Learning Research, vol. 13, pp. 2171–2175, jul 2012. https://deap.readthedocs.io/en/master/