Testing GP Parameters on Symbolic Regression

Michael Wisnewski*
Aidan Larock*
mw17an@brocku.ca
al16my@brocku.ca
Brock University
St. Catharines, Ontario, Canada

Abstract

This paper will discuss the effect that parameters have on genetic programming and explain why the final fitness of an algorithm may be affected. This paper will cover the mutation, crossover, and elitism parameters, explaining their importance in a genetic program and analyze the effect that each has on the final fitness of a genetic program. For this paper, symbolic regression will be used and the genetic program will attempt to fit an individual tree to a correct formula.

Keywords: genetic programming, symbolic regression, genetic programming parameters

1 Problem Statement

Symbolic regression is the process of creating a mathematical formula that is fitted to a data set, so that an algorithm can extrapolate data beyond or within the set. For example the user may input the expression $f(x) = x^4 + x^3 - x^2 + x + 20$ and within high accuracy regress over the allotted numbers to near perfection. At which point the program must be able to take any input at each run from the user and output a tree made by the genetic program that solves for the user input. That is, the output of the regressed function produced by the genetic program.

2 GP Parameter tables

This paper will cover four distinct variations of parameters. Each variation of parameters will have similar dependent variables of depth, tournament size, population size, terminals, and number of generations. The independent variables that will be altered can be seen bellow (Table 2.). four distinct runs altering mutation and crossover rates as well as a final variation which uses elitism instead of tournament selection to repopulate the next generation.

*Both authors contributed equally to this research.

Table 1. GP parameters User and Default

Max Depth	17
Tournament Size	3
Population Size	400
Terminals	1
Number of Generations	20

Table 2. GP parameter variants

Variant	Crossover	Mutation	Elitism
i)	90%	10%	0
ii)	100%	0%	0
iii)	0%	100%	0
v)	90%	10%	2

3 Fitness and Strategy

In genetic programming, individuals start with a random makeup of terminals and functional nodes. Each individual is evaluated using a mean squared error for each data point in a training file.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(observed - actual)^2 \tag{1}$$

While using MSE, the individuals which are more fit return lower values as their observed outputs lie more closely to the actual outputs for each value. Therefore, individuals with lower MSE values are more likely to reproduce and populate future generations. As such, individuals which had a lower fitness value are considered as better, therefore the algorithm attempts to minimize the fitness in order to find the best individuals.

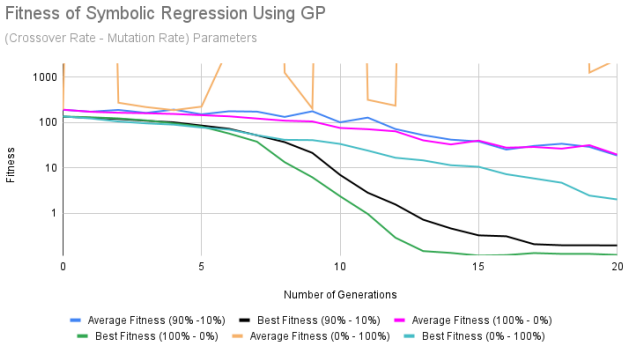
4 Results

Each variation of parameter was run a total of 10 times, each time using a different random seed. The crossover and mutation variations used variations as seen above (Table 2.). Elitism was also tested against non-Elitism. For the elitism runs, the mutation rate and crossover rate remained constant over both tests (90% crossover, 10% mutation).

4.1 Crossover and Mutation

It is clear that crossover and mutation play a large role in genetic programming. Crossovers allow for individuals to pass on genes they already have. Mutation on the other hand randomly alters subtrees within an individual.

Figure 1. Variations of crossovers and mutations)

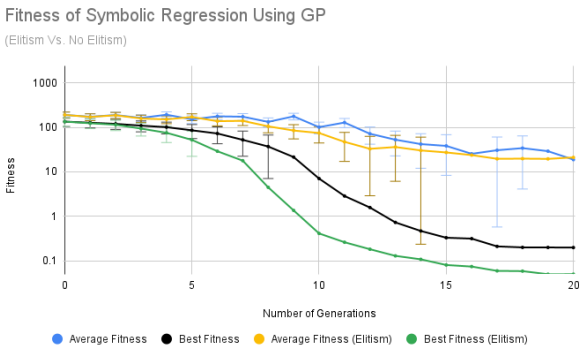


While using a variation of crossover and mutation, results were scattered. Variations which used more crossover had better and smoother fitness and transitions between generations. Variations with higher levels of mutation rates tended to have larger average fitness and large differences between generations.

4.2 Elitism

Elitism in genetic programming is simply the most fit (elite) individuals in a population are strictly selected to repopulate the next generation. While in other selection methods such as tournament or roulette wheel selection, each individual has a chance at being selected (better fitness having higher chances), with elitism the best individuals are the ones selected. The number of individuals selected for the following results are 2 individuals, which can be seen above (Table 2.).

Figure 2. Variations of Elitism and non-Elitism)



The variations with elitism applied were able to reach a peak fitness faster than their counterparts. Runs without elitism applied also had larger differences in between generations than those with. This is standard as runs with elitism would have only the best individual passing on genes.

5 Analysis

Genetic programming is unique in that many different types of user parameters are passed into the initialization phase of an algorithm. In this paper, variables were separated into dependent and independent variables in order to test the effect of independent variables on the fitness of an algorithm.

5.1 Crossover and Mutation

When changing the variables of mutation and crossover rates, runs which had higher crossover rates tended to outperform runs with higher mutation rates. However, the variation of 100% mutation rate was still improving towards the end of the run. While the variation without any mutation (100% crossover 0% mutation) was the most fit, with a more difficult data set the 90% crossover and 10% mutation rate variation would be preferred in order to prevent the algorithm at stopping at a potential local maxima. Mutation is an important factor in genetic programming as it helps individuals create solutions that would not be possible with crossover alone.

5.2 Elitism

When applying elitism to the algorithm, the runs with elitism performed better than runs without elitism. Similarly to only crossovers, elitism converges far too quickly and allows for no chance that genes which would produce better results in the future, but are on poor individuals would be passed on to future generations. This result can be seen in the error bars (figure 2), as the runs without elitism have a larger diversity in the later stages of generations. This diversity is important as it allows for future generations to have a better chance at reaching the global maximum when compared to runs with elitism.

6 Conclusion

While the top performers for symbolic regression were the results of elitism and lower mutation rates, it is important to note that the data also shows that these results converged quicker and with less diversity than results which included mutation and no elitism. This implies that while these results are superior when using simpler data with low chances of local maximas. With more complex data, mutations and non-elitism would be preferred as these genetic programs are less greedy and allow for a wider variate of solutions to progress into future generations.