COSC-5P71 FINAL PROJECT
(GP Alternatives of Breast Cancer Classification)

Amirmahdi Khosrvi Tabrizi
Student Number: 7278419
Email: ak21cx@brocku.ca

April 2022

1 Introduction

The purpose of this project is to investigate different possible ways of using Genetic Programming (GP) in order to find and compare solutions to classify breast cancer tumors. Not to mention, the GP system that is used in this project is DEAP.

For this project we have considered two different scenario to implement our GP evolutionary algorithm. The first and most basic way to achieve this goal is to define a single population with initial genetic operations, for instance, crossover and mutation. Another way to implement, which is more complicated, is to take advantage of island model. By using island model another alternative to reach our goal we need to include another operation called migration. There will be more explanation in following sections.

Wisconsin Breast Cancer Diagnostic Data Set is the data set used in this project. We use this data set as an input for our function, which technically is our solution, provided by symbolic regression to predict the tumor type. The tumor type will be either Malignant (M) or Benign (B).

Following sections will be started with a quick explanation of *Symbolic Regression* and *Island Model.* in section 3 the experiment setup and parameters are mentioned and the last two sections are about results and conclusion respectively.

Keywords— Genetic Programming, Island Model, Symbolic Regression, Classification

2 Symbolic Regression

Symbolic Regression (SR) is a type of regression analysis that searches the space of mathematical expressions to find the model that best fits a given dataset, both in terms of accuracy and simplicity. No particular model is provided as a starting point to the algorithm. Instead, initial expressions are formed by randomly combining mathematical building blocks such as mathematical operators, analytic functions, constants, and state variables.

It is also one of the best known problems in GP. It is commonly used as a tuning problem for new algorithms, but is also widely used with real-life distributions, where other regression methods may not work.

All symbolic regression problems use an arbitrary data distribution, and try to fit the data with the most accurate symbolic formula available.The *Wisconsin Breast Cancer Diagnostic Data Set* is the one used in this case. Usually, measures like the RMSE (Root Mean Square Error) or MSE (Mean Squared Error) are used to measure an individual's fitness. The latter is the measurement chosen for this project for which more details are available in following sections.

3 Island Model

The *Island Model* deals with the populations which is subdivided into a number of discrete finite populations (or islands). Therefore, *Island Model* works on multiple independent islands that evolve mostly isolated from each other, and only occasionally some individuals migrate from their islands to other islands. This imitates the nature in a better way, delays the premature convergence, and increases the search diversification since each island can potentially follow a different search trajectory through the search space. It is also possible that *Island Model* may simply be particularly well suited to exploiting the separable nature of the test problems.

4 Experiment and Setup

The main goal of this project is to train different classification function based on *Wisconsin Breast Cancer Diagnostic Data Set* to predict the type of tumors which are either Benign (B) or Malignant (M). We use GP to evolve and find the best classification function among the trained functions. For classification function this project is taking advantage of symbolic regression to represent it. Each of these functions is considered as an individual in the population.

There are different of implementing GP to solve this problem. in this project two of them are investigated, regular and Island mode. Beside the main difference between these two alternatives that is mentioned in section 3, By using *Island Model* another genetic operation should be added which is called *Migration*.

In this section we introduce different steps of the implementation and also the parameters and scenarios that is used for them.

4.1 Train and Test sets Selection

Just like other classification problems, splitting the data set into two different Train and Test subset is one of the first works to do. This can help our algorithm not to be over trained. There are 569 records in our data set which **398 records** (70%) of it are considered as Training set and the rest **171 records** (30%) as Test set. These data sets also get shuffled before being used as the input of our algorithm. Not to mention, the library that is used for this is Pandas.

4.2 GP Language

As it is mentioned before, our individuals in this project are functions that predict the type of the cancer tumor. These functions are going to be represented as trees. Each tree consists of nodes and leaves which in or case nodes are functions and leaves are terminals. GP language is defined by these functions and terminals and in this section we are going to introduce the GP language of this project.

The functions that are considered for this project are arithmetic functions as follow:

- Add
- Sub
- Multiply
- Protected Division (Division by zero is no allowed)
- Negation

Terminals:

- Ephemeral Constant
- 30 variables for the attributes

These functions are chosen because in assignment-1 we could get reasonable results with them. On the other hand, these functions are simple enough that you can be sure you will not end up with something complex.

4.3 Fitness Function

The fitness function is the function that determines how much good an individual is. In this project the fitness function assigns a value to each of the individuals based on the number of tumor types they have predicted correctly. It means at the bigger their fitness value are, the better they are.

4.4 GP Parameters

4.4.1 Population

Considering that there are two different alternatives in this project to implement, namely regular one and Island model, there are two different population as well.

The population of the regular implementation is 300 while the total population of the Island model is 600, 100 for the first Island 200 for the second Island and 300 for the last one.

4.4.2 Tree Generation

In the GP system that is used in this project (DEAP) there are two type of trees, referred to as full trees and grown trees. Full trees are the ones that have all leaves the same distance away from the root, and grown trees can have leaves which have a varying depth from the root. To initialize the trees, the *genHalfAndHalf* method in DEAP was used. This method generates half of the trees as full trees and the other half as grown trees.

As Python puts a limit on the call stack depth, it is not possible to get trees that have higher level than 90 without getting a memory error. To avoid it, the *staticLimit* method is used. This implements a static limit on some measurement on a GP tree, as defined by Koza in [Koza1989]. It may be used to decorate both crossover and mutation operators. When an invalid (over the limit) child is generated, it is simply replaced by one of its parents, randomly selected. The limit that is considered for this project is 17.

4.4.3 Selection Method

Tournament Selection is a Selection Strategy used for selecting the fittest candidates from the current generation in a Genetic Algorithm. These selected candidates are then passed on to the next generation. In this project a K-way tournament selection is used, we select k-individuals and run a tournament among them. Only the fittest candidate amongst those selected candidates is chosen and is passed on to the next generation. The K in our case is 4.

4.4.4 Crossover Operator

The crossover method used is One-Point crossover, and is referred to as *cxOnePoint* in the DEAP system. One-Point crossover takes in two trees, and randomly selects a point in each individual. The resulting sub-trees with the crossover points used as roots are swapped. The two result- ing trees are returned.

4.4.5 Mutation Operator

mutUniform is the mutation method provided by DEAP that is used in this project. It randomly select a point in the tree individual, then replace the subtree at that point as a root by another newly generated tree.

4.4.6 Migration

Although DEAP has its own migration method, this method has not been used in this project and it has its own migration implementation.

When it comes to migration there are two main factor to be considered, the number of migrants and the way of selecting an island and its individuals to emigrate from. The former can be considered as a user parameter, but there are two different ways of selection that has been investigated, random selection and average.

In case of random selection, to generate the next generation of the island population, another island is selected randomly. In this state, based on number of migrants , a number of individuals of the selected island will randomly be selected to immigrate. This function is called *random-migration-evolution*

The way that average selection works is that first of all it checks the current island average fitness. Then, if the fitness average of that island is high, the immigrants of this island will emigrate from the island with the lowest average and vice versa. The individuals of the selected island will be chosen randomly again. This function is called *avg-migration-evolution*

Objective:	Finding a function that is able to predict the Cancer types (Benign, Malignant) with the help of Symbolic Regression.
Function Set:	(ADD), (SUB), (MULTIPLY), (PROTECTED DIVISION), (NEGATION)
Terminal Set	Ephemeral Constant, 30 variables for the attributes

Default Parameters:	Population=300, Number of Islands=3(Population: 100, 200, 300),
	Individuals in each Tournament=4, Number of runs=10,
	Tree-Depth-Limit=17

User Parameter:	crxpr(Proportion of Crossover and Mutation)=0.9 (90% crossover 10% mutation), nElit(Elitism)=2, ngen(Number of generation)=50, migpr(Migration percentage)=40%

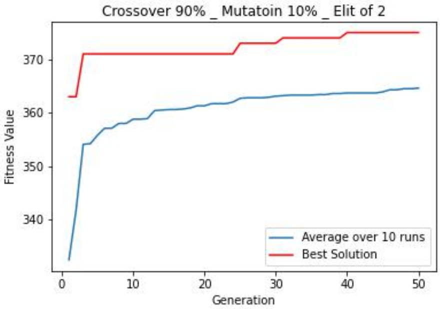
5 Result

There are three main functions which are the key to this project when it comes to running. These functions are *evolution*, *avg-migration-evolution* and *random-migration-evolution*. All these functions get the GP parameters they need as an input and then start the evolution; Finally, they return an array of the best answers or individuals for each generation.

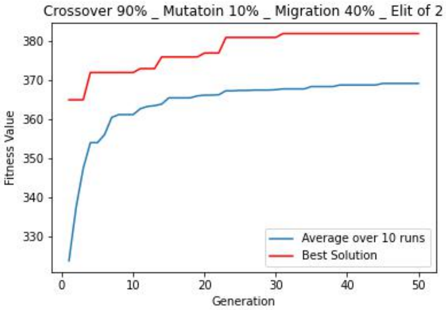
The *evolution* is the simplest among these three functions. It does the regular model while the remaining ones (*avg-migration-evolution* and *random-migration-evolution*) do the Island model in two different migration approach that are discussed above.

To get the whole project running, we have run these three functions 10 times with different ***random seeds*** (125, 85, 318, 10, 57, 701, 26, 564, 912, 487) to compare the performance of these three scenarios. Not to mention, the GP parameters that is chosen for this experiment is based on the assignment-1's result which showed that crossover of 90%, mutation of 10% and elitism of 2 is reasonable choice. In this experiment every parameters are the same except that for *Island Models* we have one more parameter (Migration rate). The reason is to see the exact difference of choosing different alternatives while all the parameters are the same.

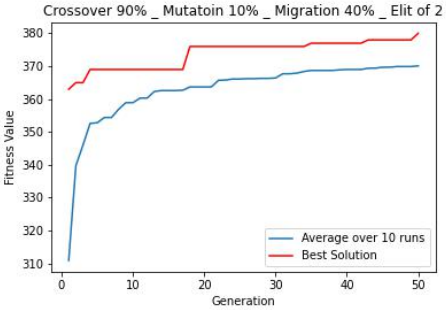
- **Regular Model**



- **Island Model with Random Island Selection**



- **Island Model with Average Island Selection**



5.1 Insight

As it can be seen from the graphs above, when it comes to the *Average over 10 runs*, all of the three alternatives reach the average fitness value around 360/398, although in first 10 generations *Regular Model* and *Island Model with Random Island Selection* have shown better performance.

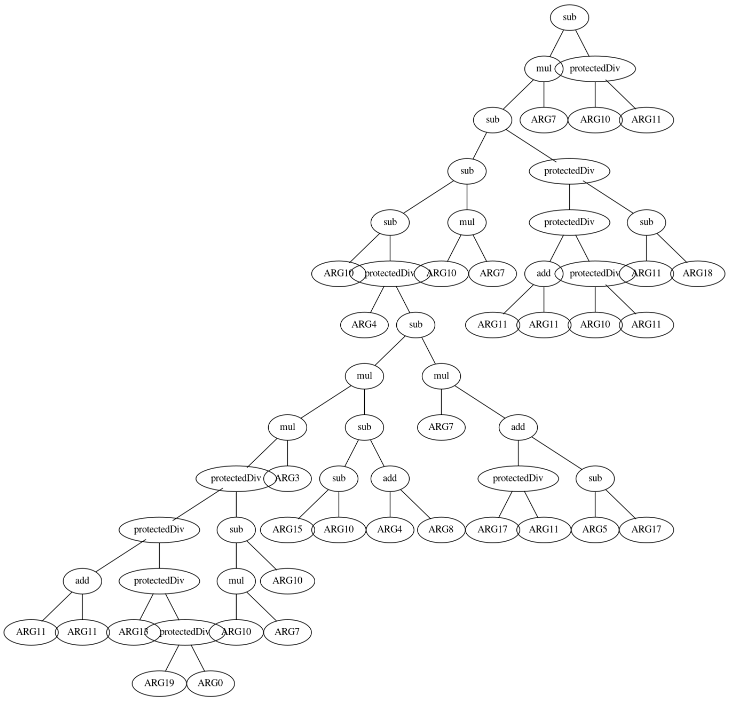
On the other hand, considering the overall *best solution*, I should say that both Island Model show better performance by ending up with higher fitness value around 380/398 at the end of the **50 generation**. The Regular model approximately reaches to 370/398 at the end. But, between all of these alternatives *Island Model with Random selection* is more reliable overall as its fitness value increases steadily over generations.

5.2 Testing

For the testing the overall best answer (Island Model with Random Selection) among all of these three alternatives has been chosen and the following table shows the confusion matrix which tested our answer with our test set that includes **171 records**.

Actual Predicted	Malignant	Benign
Malignant	60	5
Benign	2	104

- **Solution Tree**



6 Conclusion

This project examined the three alternatives of using GP to classify the cancer tumor types which can be either Malignant or Benign. One of these three alternatives is the Regular model which does not have any sub-population. The remaining two ways are taking advantage of Island model and each of them have different scenario to chose an Island to emigrate. One of the selects based on minimum and maximum average while the other one selects randomly.

When it comes to performance, it can be seen that all three of them are doing great. However, the Island Model with Random Selection shows more reliable performance over generations as its fitness value increases gradually and reached the highest fitness value of 383.

For future work, changing the number of the island and its population or trying different migration rate might be the case.

References

- DEAP 1.3.1 documentation
- A Field Guide to Genetic Programming - R. Poli, W. B. Langdon, N F. McPhee - J. R. Koza
- Breast Cancer Wisconsin (Diagnostic) Data Set