Assignment 1: Part B - Cancer Diagnostic

Harmandeep Mangat hm15mx 6021109

I. Introduction

This paper will be examining the performance of the genetic program that examines the data describing an X-ray image and returns whether it is benign or malignant.

II. Experiment Details

A. Parameter Table

Duplicate Entries	100
Crossover	90%, 80%%
Mutation rate	10%, 20%
Max Depth	17
Tournament Size	7
Generation	51
Quit-on-Run-Complete	True
Subpop size	1024
Elitism	No

TABLE I

B. Fitness Evaluation

There is a 2D array that holds the training data. You loop through the array, retrieving the first eleven indexes. The first index, index 0, holds the expected result which is either M or B, and the result of the array holds the data for the x-ray image. Since the data is repeated, you only need to take the first 10 indices after the expected result. Once all the information is stored in the state, you execute the tree with the fitness function GPIndividual, which will then return the value result. It the result is greater then or equal to 0 and the expected result is M, then increase hits by one. If the result is less then 0 and and the expected result is B, then you increase hits by 1 again. If neither of those conditions meet, increase sum by one. Lastly, you need to run the KorzaFitness for each generation to retrieve the stanrdized fitness and adjusted fitness.

1) Pseudo Code:

$fori = 0$ to array size
$expected = array[i][0]$
$expected = array[i][1]$
.
.
.
$expected = array[i][10]$
$execute\_tree$
$ifresults >= 0 and expected = Mhits++$
$ifresults < 0 and expected = Bhits++$
$elsesum++$ $KorzaFitness$

C. GP Language

1) Functions: For this gp problem, I used seven functions; Add, Sub, Multiply, Division, Max, Min, and If then else.

2) Terminals: I also used two terminals which were X1 to X10, which returns one of the data points, and Ephemeral.

D. Format of Data

I read the data file, which can be found here: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Di into an 2d array of objects. I proceeded to then shuffle the array and spilt it in half, one array for training and the other for testing.

E. Variations of Experiment

1) Experiment 1: Crossover 90% and Mutation 10%

2) Experiment 2: Crossover 80% and Mutation 20%

III. Results

A. Experiment 1

After having run this experiment with both parameters, it can be concluded that the gp that used 90% crossover and 10% mutation (AVG_1), had a slower convergence but a better final fitness when compared to the parameter set of 80% crossover and 20% (AVG_2), which had a faster convergence but a lower final fitness. Overall, the better performing gp parameters were from AVG_1, as it had fewer inaccurate results when comparing the two confusion matrices.

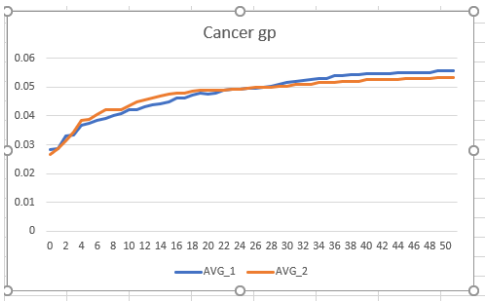

Fig. 1

	AVG_1	AVG_2
Final Fitness	0.055722095	0.053508303
Min	0.028234614	0.026670828
Max	0.055722095	0.053508303
Mean	0.04775287	0.047596381
Median	0.049674973	0.049641966
Standard Deviation	0.007373059	0.006448709

TABLE II: Figure 1 Table

	Positive	Negative
Positive	64	11
Negative	3	191

TABLE III: Confusion Matrix: 90% crossover, 10% Mutation

```
(* (- (/ (- (- x10 0.0) x8) (+ (+ (+ (Max
 (Min x5 x4) (* x5 x4)) (+ (- x8 (* x5 x4))
 (Min (+ x6 x10) (- (* (- (+ (- x8 x10) x6)
     (- x3 x9)) (+ (Min (- (- x10 0.0) (Min x7
     x8)) (Min x6 0.0)) (+ (- x7 x10) x10))) (-
     (/ (- (- x10 0.0) (Min x7 x8)) (+ (+ (Min
         x5 x4) x6) (+ (- x10 x6) (+ (- x8 x2) (*
         x10 x2))))) (* (- (+ (Min x6 x3) (* x1 x2))
     (Max (- (Min x6 0.0) (+ x7 x9)) (/ x10 x9)))
     (/ (* x5 x8) (Max x8 x10)))))))) (/ x9 x7))
 (Max (Min x5 x4) (+ (* x7 x1) (+ (Max (Min
     x5 x4) (* x5 x4)) (+ (- x8 x2) (* x10 x2)))))))
 (* (- (+ (Min x6 x3) (* x1 x2)) (/ (- x2
     x3) (- (Min (+ x7 x2) (- x10 x6)) (Max (Max
     x8 x6) (/ x10 x9))))) (/ (* x5 x8) (Max x8
     x4)))) (Min (Min (Min (+ x6 x10) (Min (Min
 (Min (+ x6 x10) (- (+ (- x1 x3) (- x6 x4))
     (+ (Min (* x10 x2) (Min x6 0.0)) (+ (- x7
         x10) x10)))) (+ (- x1 x3) (- x6 x4))) (+
 (Max (- (+ (* x5 x9) (* x1 x2)) (/ (- x2
     x3) (- (- (+ 0.0 x6) (- x3 x9)) (Max (Max
     x8 x6) (/ x10 x9))))) (* x5 x4)) (+ (* (-
 x6 x4) (/ (Min x6 0.0) (Max x8 x4))) (* x7
 x1))))) (+ (- x1 x3) (- x6 x4))) (+ (Max
 (- (/ (- (- x8 (* x5 x4)) x8) (+ (+ x7 (/
     x9 x7)) (Max (Min x5 x4) (+ (* x7 x1) (+
     (Max (Min x5 x4) (* x5 x4)) (+ (Min x5 x4)
     (* x10 x2)))))) (* (- (+ (- x10 x6) (* x1
     x2)) (/ (- x2 x3) (- (Min (+ x7 x2) (- x10
     x6)) (Max (Max x8 x6) (/ x10 x9))))) (/ (*
     x5 x8) (Max x8 x4)))) (* x5 x4)) (+ (- x8
 x2) (* x10 x2)))))
```

	Positive	Negative
Positive	66	32
Negative	1	186

TABLE IV: Confusion Matrix: 80% crossover, 20% Mutation

```
 (+ (- (/ (* x8 x2) (Max x10 x5)) (/ (+ (Min      (+ x4 x1)
(/ x4 x3)) x3) (+ x2 x7))) (Min
   (* (- x9 x4) (/ x9 x3)) (* (Max (Min (Max
   (/ x2 (/ (* x8 x2) (Max (/ (* x10 x9) (Min
       (- x8 x10) x8)) (/ (Max x7 (Min (+ x4 x1)
   (/ x4 x3))) (/ x2 x8))))) x2) (+ (Max (Min
   x10 x8) (Min 1.0 x4)) (/ (* x8 x2) (+ x7
   x10)))) (- (* (Max (* x8 x2) (Min 1.0 x4))
   x2) (Min x2 x2))) (Max (* (Max (Min x6 (/
   (/ (- (- x4 x6) (Max x10 x8)) (+ x9 x3))
   (- x6 x10))) (Max (/ (* x10 x9) (- (/ (*
   x8 x2) (Max x10 x5)) (/ (- (/ (* x8 x2) (Max
   x10 (Min (+ x8 x5) (+ x1 1.0)))) (/ (+ x8
   x5) x2)) (/ (- x8 x5) (- x8 x10))))) (+ (-
   x8 x10) (- (+ x4 x5) (/ x2 x8))))) (Max (*
   (- (* (Max (* x8 x2) (* (- x10 1.0) (- x8
       x9))) x2) (Min x2 x2)) (+ x1 x1)) (+ (Min
   (+ x8 x5) (+ x1 1.0)) (- (+ x4 x5) (/ x2
   x8))))) (+ (- x8 x10) (- (+ x4 x5) (/ x2
   x8)))))))
```

IV. Conclusions

In this paper, I have compared to set a parameter with respects to the problem of solving whether the person has cancer by running a gp of a set of x-ray data. The first set of parameters used where 90% crossover and 10% mutation. It had a slower convergence but a better final fitness. The second set of parameter used was 80% crossover and 20% mutation, which had a faster convergence but a lower final fitness. When comparing the confusion matrices, to can be concluded that the first set of parameter give better results has it has less errors. In order to improve these results, more functions as a different parameter set could be used to get fewer errors when running the testing data set.