



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

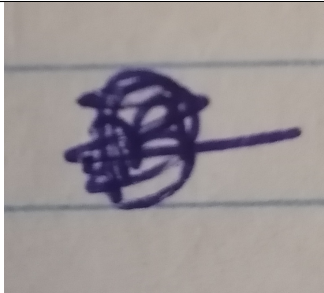
Faculty of Engineering, Built Environment and
Information Technology

COS314

ARTIFICIAL INTELLIGENCE

ASSIGNMENT 2: META-HEURISTICS: GP & SA

CONTRIBUTORS

Name and Surname	Student Number	Signature
B.M. Dzimati	20456078	

By submitting this assignment I confirm that I have read and am aware of the University of Pretoria's policy on academic dishonesty and plagiarism and I declare that the work submitted in this assignment is my own as delimited by the mentioned policies. I explicitly declare that no parts of this assignment have been copied from current or previous students' work or any other sources (including the internet), whether copyrighted or not. I understand that I will be subjected to disciplinary actions should it be found that the work I submit here does not comply with the said policies.

Contents

1	Parameter List	2
1.1	Genetic Program Parameters:	2
1.1.1	Descriptions Of Constant Parameters:	2
1.1.2	Descriptions Of Data Set Based Parameters:	2
1.1.3	Breast Cancer Dataset:	3
1.1.4	TicTac-Toe Dataset:	3
1.2	Simulated Annealing Parameters:	3
1.2.1	Descriptions Of Data Set Based Parameters:	3
1.2.2	Breast Cancer Dataset:	3
1.2.3	TicTac-Toe Dataset:	4
2	Tests & Structure of Best Rule	5
2.1	Some of the best TicTac Toe results:	5
2.2	Some of the best Breast Cancer results:	5
2.3	Example of Output:	6
3	Seed Values	8
4	Critical Analysis	9

List of Figures

1	This is an image showing the structure of the tree in postfix and infix notation as well as the accuracy for a run on the TicTac Toe data set for the GP and the SA programs	5
2	This is an image showing the structure of the tree in postfix and infix notation as well as the accuracy for a run on the Breast Cancer data set for the GP and the SA programs	6
3	This is an image showing the structure of the tree in postfix and infix notation as well as the accuracy for a run on the Breast Cancer data and the TicTac Toe set for the GP and the SA programs	7

1 Parameter List

1.1 Genetic Program Parameters:

1.1.1 Descriptions Of Constant Parameters:

- Genetic Operator rates and Replacement Rate:
 - The replacement rate used was 50% as the top 50% of the population are mutated and re-placed into the next generation.
 - The top 50% reproduces with each other through a crossover function, with the fittest trees crossing over with each other.
 - This sets the mutation to rate to being 50% while the crossover rate being 50%.
- Mutation:
 - The mutation used is a mixture of a shrink and grow mutation with the pruning of the tree being based on the max number of terminal nodes, thus effectively limiting the trees to a certain height.
- Tree Generation:
 - The generation of tress was through a mixture of half ramped and grow as the trees are recursively made up until a certain number of terminal nodes are placed on the tree to make the tree full.
 - This results in the production of Trees which are sometimes perfect and sometimes imperfect.

1.1.2 Descriptions Of Data Set Based Parameters:

- populationSize : int - - Stores the population size of each generation which stays constant through out all possible generations.
- maxTermnial : int - - Stores the maximum number of terminal nodes each tree can have that is part of the population. This in intern limits the height setting up a maximum height for all the trees. While allowing them to have varying depth and also allowing them to be unbalanced.
- numberGeneration : int - - This stores the maximum number of generations that are allowed to be made.
- data : int[][] - - This holds the training or testing data that will be used to train or test the system, except the answer.
- ans : int[] - - This holds the answers of each entry of the data, features column.
- isCancer : boolean - - Boolean used to store which data sets is being used to make code more general.

1.1.3 Breast Cancer Dataset:

- populationSize : 10
- maxTermnial : 30
- numberGeneration : 1000
- data : data read from breast_cancer.csv(The first 9 column)
- ans : data read from breast_cancer.csv(The last column)
- isCancer : Tree

1.1.4 TicTac-Toe Dataset:

- populationSize : 10
- maxTermnial : 30
- numberGeneration : 1000
- data : data read from tictactoe_train.csv(The first 9 column)
- ans : data read from tictactoe_train.csv(The last column)
- isCancer : False

1.2 Simulated Annealing Parameters:

1.2.1 Descriptions Of Data Set Based Parameters:

- Fittest : Tree – This is the best expression tree derived by the genetic program.
- temp : double - - This just stores the value of the initial temperature to be used
- data : int[][] - - This holds the training or testing data that will be used to train or test the system, except the answer.
- ans : int[] - - This holds the answers of each entry of the data, features column.
- isCancer : boolean - - Boolean used to store which data sets is being used to make code more general.

1.2.2 Breast Cancer Dataset:

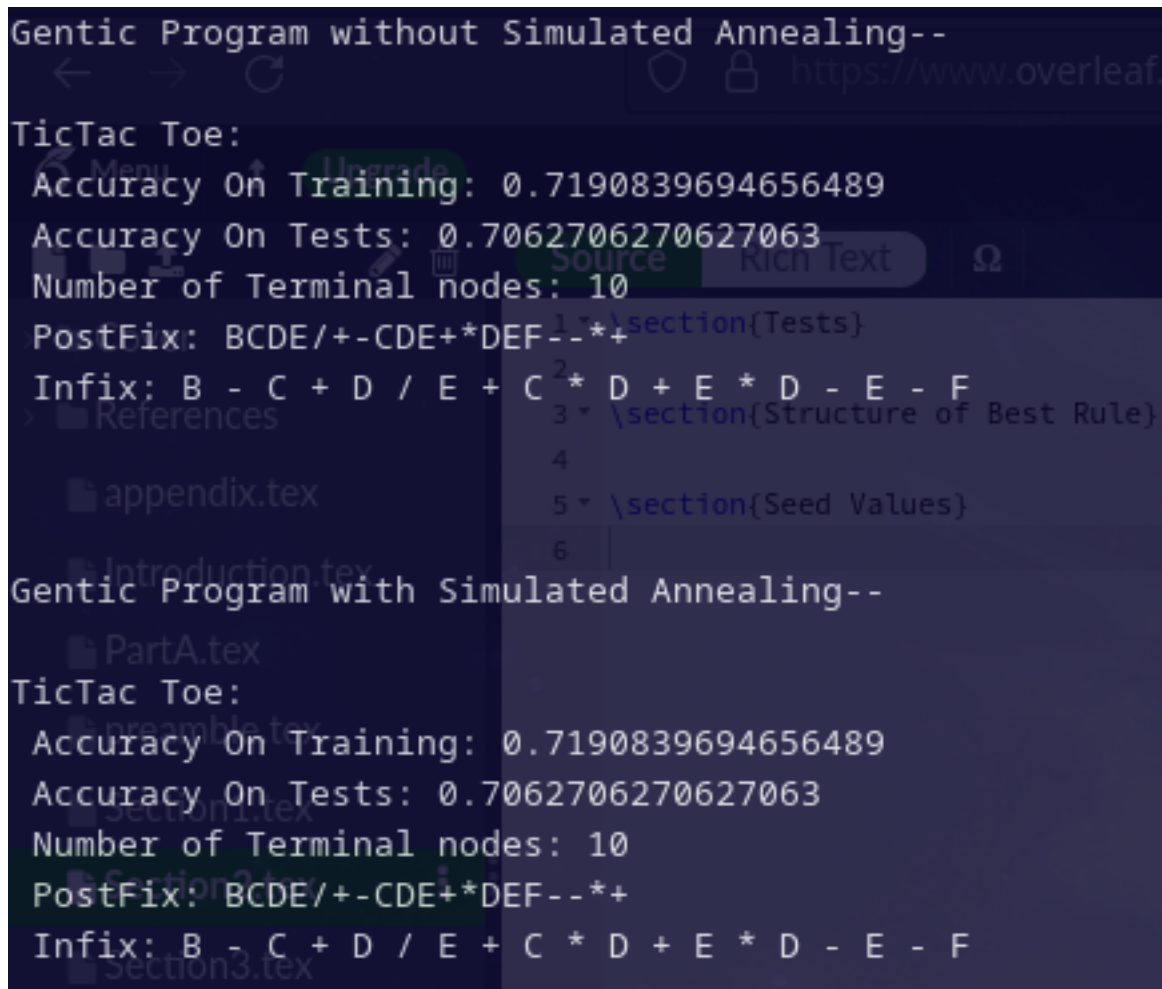
- Fittest : Best/Fittest tree from the GP on the Breast cancer Dataset.
- temp : 0.989
- data : data read from breast_cancer.csv(The first 9 column)
- ans : data read from breast_cancer.csv(The last column)
- isCancer : Tree

1.2.3 TicTac-Toe Dataset:

- Fittest : Best/Fittest tree from the GP on the TicTac Toe Dataset.
- temp : 0.689
- data : data read from tictactoe_train.csv(The first 9 column)
- ans : data read from tictactoe_train.csv(The last column)
- isCancer : False

2 Tests & Structure of Best Rule

2.1 Some of the best TicTac Toe results:



```

Gentic Program without Simulated Annealing--
TicTac Toe:
Accuracy On Training: 0.7190839694656489
Accuracy On Tests: 0.7062706270627063
Number of Terminal nodes: 10
PostFix: BCDE/+-CDE+*DEF--*+
Infix: B - C + D / E + C * D + E * D - E - F

Gentic Program with Simulated Annealing--
TicTac Toe:
Accuracy On Training: 0.7190839694656489
Accuracy On Tests: 0.7062706270627063
Number of Terminal nodes: 10
PostFix: BCDE/+-CDE+*DEF--*+
Infix: B - C + D / E + C * D + E * D - E - F

```

Figure 1: This is an image showing the structure of the tree in postfix and infix notation as well as the accuracy for a run on the TicTac Toe data set for the GP and the SA programs

2.2 Some of the best Breast Cancer results:

```

[momo@livebookhome Implementation]$ java main
Gentic Program without Simulated Annealing--

Breast Cancer:
Accuracy On Training: 0.9560669456066946
Accuracy On Tests: 0.9560975609756097
Number of Terminal nodes: 20
PostFix: BC-C*C/CD+-CDEF*F-**+CD+D/D+D+D+D+D/-
Infix: B - C * C / C - C + D + C * D * E * F - F - C + D / D + D + D + D + D + D / D

Gentic Program with Simulated Annealing--

Breast Cancer:
Accuracy On Training: 0.9288702928870293
Accuracy On Tests: 0.9609756097560975
Number of Terminal nodes: 11
PostFix: BC+CD+D/D+D+D+D+D+D/-
Infix: B + C - C + D / D + D + D + D + D + D / D

```

Figure 2: This is an image showing the structure of the tree in postfix and infix notation as well as the accuracy for a run on the Breast Cancer data set for the GP and the SA programs

2.3 Example of Output:

```

[momo@vivoBookMomo Implementation1]$ java Main
Genetic Program without Simulated Annealing--

Breast Cancer:
Accuracy On Training: 0.9602510460251046
Accuracy On Tests: 0.9512195121951219
Number of Terminal nodes: 17
PostFix: BC/CDE/E/EF/+E/EF/+EF/+E/EFG*/+ -*
Infix: B / C * C - D / E / E + E / F / E + E / F + E / F / E + E / F * G

TicTac Toe:
Accuracy On Training: 0.751145038167939
Accuracy On Tests: 0.7029702970297029
Number of Terminal nodes: 15
PostFix: BC/CDEF--EFGH+++//DE-E-+DE/++
Infix: B / C + C / D - E - F / E + F + G + H + D - E - E + D / E

Genetic Program with Simulated Annealing--

Breast Cancer:
Accuracy On Training: 0.9602510460251046
Accuracy On Tests: 0.9512195121951219
Number of Terminal nodes: 17
PostFix: BC/CDE/E/EF/+E/EF/+EF/+E/EFG*/+ -*
Infix: B / C * C - D / E / E + E / F / E + E / F + E / F / E + E / F * G

TicTac Toe:
Accuracy On Training: 0.751145038167939
Accuracy On Tests: 0.7029702970297029
Number of Terminal nodes: 15
PostFix: BC/CDEF--EFGH+++//DE-E-+DE/++
Infix: B / C + C / D - E - F / E + F + G + H + D - E - E + D / E

```

Figure 3: This is an image showing the structure of the tree in postfix and infix notation as well as the accuracy for a run on the Breast Cancer data and the TicTac Toe set for the GP and the SA programs

3 Seed Values

Seed Values weren't used as the function used to generate random numbers cannot to be seeded. But while testing a pattern emerged. The Breast Cancer data sets always produces a tree with 95% accuracy, so the Genetic Program is set to repeat it's self up until it produces a tree with an accuracy of 95%. The same is done for the TicTac Toe but the accuracy was lowered to 67%, the program is set to repeat the Genetic Program up until it returns a tree of accuracy 95% for the Breast Cancer data sets, and another tree with accuracy of 67% for the TicTac Toe data sets.

4 Critical Analysis

The Tree that is found by the GP is usually the global optimum as applying the SA doesn't usually offer a better solution. This means that the global optimum tree has an accuracy of 95% on the test data for the Breast Cancer data sets and around 70% on the test data for the TicTac Toe data sets. This means that the Breast Cancer dataset offers a more a pattern based and structured data sets which can be structured into a graph that is able to almost always correctly predict the value given the features. What can be said about the TicTac Toe dataset is that it doesn't follow a more mathematical based pattern as strict as the Breast Cancer data set, this could be due to the presence of many anomalies in the training data thus resulting in dirty data or this could be that the feature set by nature is more varied and not as patterned based. Or it could be that a poor program was written in the context of the tictac toe data sets, e.g high mutation rate or high replacement rate.