Paul Pilane | u21448150

COS 314 | Artificial Intelligence

Assignment 3

Introduction

Classification of a breast-cancer data is performed through various models, GP and a WEKA tool (using a Decision Tree), and the analysis of each model is thus presented in the later sections of this paper.

Genetic Programming

Experimental Setup

Problem Representation

A node in the syntax tree, in the GP, refers to a feature on the dataset instance. As such combination of the nodes comprise a syntax tree (individual). Therefore, different individuals, or syntax trees, can be joined together to grow the tree; form a larger tree. The decision tree can thus be generated from the generated syntax tree from the GP. The GP will thus evolve the decision tree due to the genetic operators since such operators influence the tree.

Fitness Function

The fitness function used evaluates the overall performance of a node on a dataset by counting the number of instances where the node produces positive results, providing an indication of the node's capability and suitability, allowing for comparative analysis, and guiding the selection and evolution of nodes

Configuration Description

POPULATION_SIZE	MAX_GENERATIONS	MUTATION_RATE	CROSSOVERR_RATE
100	50	0.05	0.6

Initial Population Generation

The initial population is generated using the Grow Method. In the Grow Method, the program trees are grown from a random set of terminals and functions until a maximum depth is reached. The idea is to allow the trees to grow freely while still ensuring that they don't become too large and complex.

Crossover Operator

The crossover operator creates two new offspring which are formed by taking parts (genetic material) from two parents. The operator selects two parents from the population based on a selection method, Tournament Selection. A crossover point is then randomly selected in both trees,

say point p1 and p2, from tree t1 and t2 respectively – Subtree Crossover. The crossover then happens as follows: the subtree rooted at p1 is removed from t1 and inserted into the position p2 in t2. The same logic applies to the point p2; the subtree root at the point is removed from t2 and inserted into the place of p1 in t1. Crossover promotes convergence and is a local search operator

Selection Type

The selection type for parent selection is the Tournament-Selection. Such a selection type can help preserve diversity in the population, as it allows individuals with different fitness levels to compete against each other. This can prevent premature convergence, which occurs when the population converges to a suboptimal solution too quickly, such that it does not really have the best solution.

Mutation Operator

A mutation point, p, is randomly selected in the parent, and the subtree rooted at the point is removed and replaced by a randomly created terminal node; Shrink mutation is used.

Reproduction Operator

The reproduction operator copies a parent across to the next generation by simply duplicating the individual and making no alterations to it. The reproduction operator is a local operator since it makes no alterations to the parent being copied across.

Termination Criteria

The algorithm will terminate after running MAX_GENERATIONS number of times, or after a best solution has been found.

C4.5 Decision Tree

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         216               75.5245 %
Incorrectly Classified Instances        70               24.4755 %
Kappa statistic                          0.2826
Mean absolute error                      0.3676
Root mean squared error                  0.4324
Relative absolute error                 87.8635 %
Root relative squared error             94.6093 %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC     ROC Area  PRC Area  Class
               0.960    0.729    0.757      0.960   0.846      0.339   0.584     0.736     no-recurrence-events
               0.271    0.040    0.742      0.271   0.397      0.339   0.584     0.436     recurrence-events
Weighted Avg.  0.755    0.524    0.752      0.755   0.713      0.339   0.584     0.647

=== Confusion Matrix ===

   a   b   <-- classified as
 193   8 |   a = no-recurrence-events
  62  23 |   b = recurrence-events
```
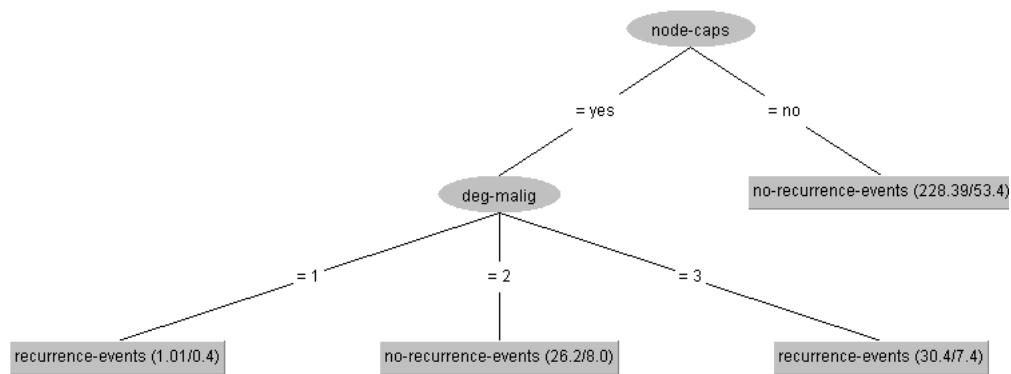
The Visualization of the tree

node-caps

= yes = no

deg-malig   no-recurrence-events (228.39/53.4)

= 1   = 2   = 3

recurrence-events (1.01/0.4)   no-recurrence-events (26.2/8.0)   recurrence-events (30.4/7.4)

Model	Accuracy (%)	f-measure
Genetic Programming	73.78	0.825
(WEKA tool) Decision Tree	75.52	0.713