1 Introduction

Answer some questions about Genetic Programming.

1. How does population size relate to performance?
2. How much need is there for mutation?
3. What sorts of problems can be solved?

2 Abalone

`https://archive.ics.uci.edu/ml/datasets/abalone`

> Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope – a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age....
>
> Abalone (via Spanish abulón, from Rumsen aulón) is a common name for any of a group of small to very large sea snails, marine gastropod molluscs in the family Haliotidae.

`https://en.wikipedia.org/wiki/Abalone`

2.1 Attributes

2.1.1 Sex

The sex of the fish is a nominal value in $\{M, F, I\}$. Given that the Genetic Programming system uses floating point data the sex will have to be either ignored or re-coded.

Here two approaches will be used:

1. As a factor. Three mutually exclusive flags with one of three variables set to 1 and the other two to 0
2. Encode sex as $F \Rightarrow -1,\ I \Rightarrow 0, M \Rightarrow 1$

Name	Data Type	Meas.	Description
Sex	nominal		M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer		+1.5 gives the age in years

Table 1: The attributes. The last one, *Rings*, is the objective to predict

2.1.2 Objective

The objective is the age of the fish. In some studies[1] the age is divided into three classes making this a classification problem.

Rings	Age
0-8	Young
9-14	Adult
15-29	Old

Encode as

-1	$\Rightarrow$	Young
0	$\Rightarrow$	Adult
1	$\Rightarrow$	Old

Here both approaches will be used

2.2 Experimental Setup

Given two encodings of sex, and two of the objective function there will be four simulation runs.

Abalone1C Continuous objective, sex as factor. See table **??** for a sample

Abalone2C Objective as classes, sex as factor. See table **??** for a sample

[1]E.g., see `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.2321&rep=rep1&type=pdf`

F	I	M	Length	Diameter	Height	W.Weight	S.Weight	V.Weight	Sh.Weight	Age
0	0	1	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15
0	0	1	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7
1	0	0	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9

Table 2: Sample data for **Abalone1C**

F	I	M	Length	Diameter	Height	W.Weight	S.Weight	V.Weight	Sh.Weight	O2
0	0	1	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	1
0	0	1	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	-1
1	0	0	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	0

Table 3: Sample data for **Abalone2N**

Abalone1N Continuous objective, sex encoded. See table **??** for a sample

Abalone2N Objective as classes, sex encoded. See table **??** for a sample

2.2.1 Common Setup

For this simulation all

```
num_generations 2000
initial_population 1000
max_population 10000
crossover_percent 50
training_percent 80
plot_xlab Age
data_file data2N.in
```

Sex	Length	Diameter	Height	W.Weight	S.Weight	V.Weight	Sh.Weight	O1
1	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15
1	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7
-1	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9

Table 4: Sample data for **Abalone1N**

Sex.1	Length	Diameter	Height	W.Weight	S.Weight	V.Weight	Sh.Weight	O2
1	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	1
1	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	-1
-1	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	0

Table 5: Sample data for **Abalone2C**

```
model_data_file Abalone2N.txt
plot_file Abalone2N.png
r_script_file Abalone2N.R
generations_file AbaloneGenerations2N.txt
birthsanddeaths_file AbaloneBirthsAndDeaths2N.txt
```

The first six parameters are constant over the four simulations. The final six will be specific to each individual simulation

2.3 Results

Each simulation took between 116 and 138 minutes real time[2].

None of the simulations produced a model that did a very good job estimating the age of the fish.

In figure **??** it can be seen that the evaluation of the best models in the simulation very quickly reached a maxima. Simulations 1C and 1N reached a higher evaluation plateau than 2C and 2N. 1N and 2N both did a little bit better than 1C and 2C but the results are close in that case.

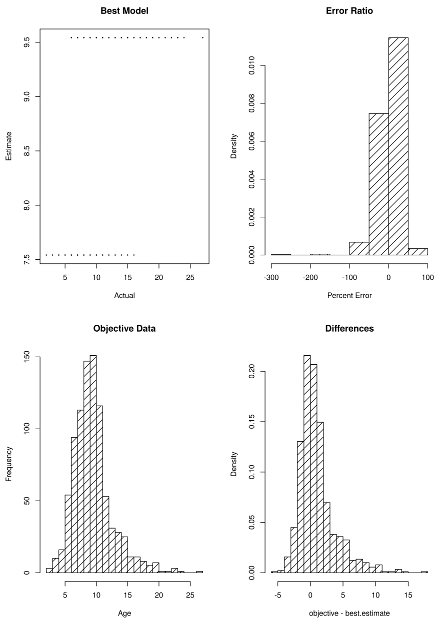
A tentative observation is that the continuous objective is easier for the Genetic Programming simulation to find than a classification. There is less evidence that encoding the Sex parameter in $-1..1$ is better than using factors.

Results archived as Abalone20180501

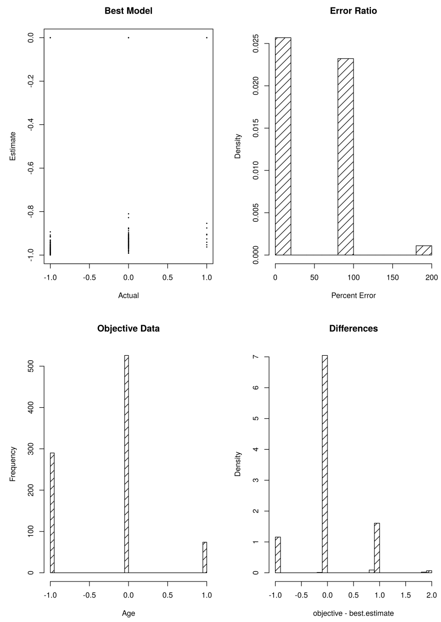
2.4 Increase Population

Hypothesising that increasing the population size in the simulations will improve the results the population size increased by a factor of 10.

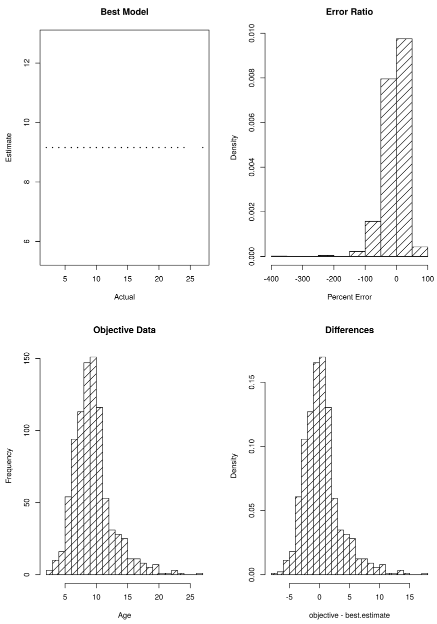
[2]On a Intel NUC: Intel Core i3-3217U CPU @ 1.80GHz with 15G RAM



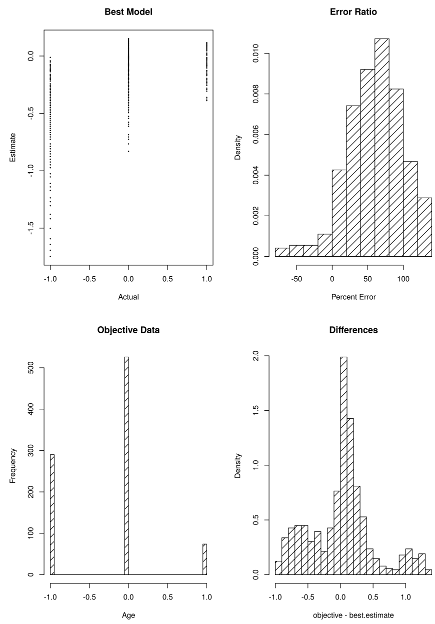
(a) Results of simulation for Abalone1C



(b) Results of simulation for Abalone2C

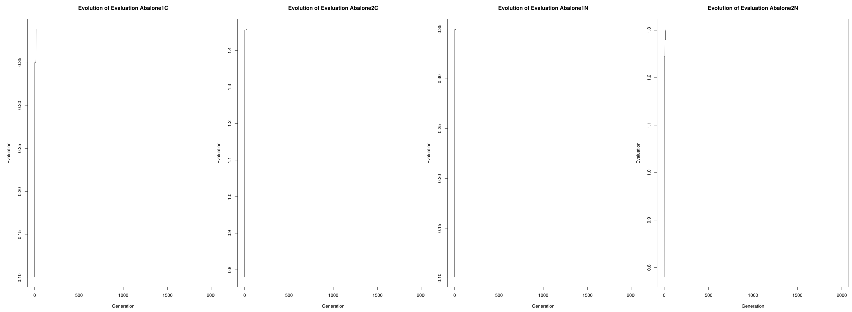


(c) Results of simulation for Abalone1N



(d) Results of simulation for Abalone2N

Figure 1: Results of Abalone Age Estimation Simulation



(a) Abalone1C (b) Abalone2C (c) Abalone1N (d) Abalone2N

Figure 2: Evolution of Evaluation for Abalone Age Estimation Simulation

`initial_population 1000`

`max_population 10000`