Efficient Interleaved Sampling of training Data in Genetic Programming

R. Muhammad Atif Azad
CSIS Department
University of Limerick
Ireland
atif.azad@ul.ie

David Medernach
CSIS Department
University of Limerick
Ireland
david.medernach@ul.ie

Conor Ryan
CSIS Department
University of Limerick
Ireland
conor.ryan@ul.ie

ABSTRACT

The ability to generalize beyond the training set is important for Genetic Programming (GP). *Interleaved Sampling* is a recently proposed approach to improve generalization in GP. In this technique, GP alternates between using the entire data set and only a single data point. Initial results showed that the technique not only produces solutions that generalize well, but that it so happens at a reduced computational expense as half the number of generations only evaluate a single data point.

This paper further investigates the merit of *interleaving* the use of training set with two alternatives approaches. These are: the use of random search instead of a single data point, and simply minimising the tree size. Both of these alternatives are computationally even cheaper than the original setup as they simply do not invoke the fitness function half the time. We test the utility of these new methods on four, well cited, and high dimensional problems from the symbolic regression domain.

The results show that the new approaches continue to produce *general* solutions despite taking only half the fitness evaluations. Size minimisation also prevents bloat while producing competitive results on both training and test data sets. The tree sizes with size minisation are *substantially* smaller than the rest of the setups, which further brings down the training costs.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Miscellaneous; I.2.6 [**Artificial Intelligence**]: Learning—*Parameter learning, performance measures*

Keywords

Genetic Programming, Over-fitting, Interleaved Sampling, Computational Efficiency, Speedup technique, Robustness of solutions

1. INTRODUCTION

GECCO'14, July 12–16, 2014, Vancouver, BC, Canada.
ACM 978-1-4503-2881-4/14/07.
http://dx.doi.org/10.1145/2598394.2598480 .

While *Interleaved Sampling* [1] is certainly useful if it reduces over-fitting, it is also useful because it reduces the computational expense of a GP run. Even so, a question arises as to whether there is merit in using a single data point. Also, a related question is: what other measures can we use instead of using a single data point in order to further reduce the computational expense while still retaining the merits of the original Interleaved Sampling?

To answer the questions raised above, this paper takes a two pronged approach. First, *Interleaved-Random*, in order to ascertain the efficacy of using a single data point in interleaved generations, we compare the results with those from using only random search in the interleaved generations.

Next in a bid to reduce the computational expense even further we specifically reduce tree size in the interleaved generations using a method we term *Interleaved-Size*.

2. EXPERIMENTS

We compare the performance of Interleaved Sampling with Interleaved-Random, Interleaved-Size and normal GP. In Interleaved-Size we minimise the size to 15 instead of 0 (we minimize $|15 - S_i|$ with S_i size of an individual).

We note the following statistics to compare performance: the median test fitness of the best individuals (best on training) in the final generations; the median over-fitting of the best individual in the final generation. As in [1], we measure over-fitting as the absolute difference between the training and the test fitness of an individual; and the average tree size of the evolving individuals as an indicator of the computational overhead of each setup.

We compute the statistical significance of the performance difference using the *Mann-Whitney U* test at $p = 0.05$.

2.1 Problem Suite

We consider four high dimensional problems (7 to 241 input variables); the legends are `Toxicity`, `Concrete Strength`, `Bioavailability` and `Yacht`.

At the beginning of each run we randomly split the data set into two sets of identical size. One is used as the training data set and the other one is used as the testing data set.

2.2 Results

In testing fitness we can not see a *single* consistent winner however, interleaved setups in *some* form perform the best throughout. Moreover, the test set performance appears to improve as the population size increases from 50 to 500. Interleaved-Size and Interleaved Random perform

Table 1: Testing Fitness

Population	50	250	500
Toxicity			
None	204.748	199.402	209.679
Rand	**196.276**	192.208*	201.343
Size	206.139	**195.690**	**190.409**
1-pt	206.892	207.957	212.608
Bioavailability			
None	**3.27848**	2.94743	2.88460
Rand	**3.26689**	**2.85854**	**2.70530**
Size	3.34733	3.17819	3.02524
1-pt	3.96089	3.34340	3.22562
Concrete			
None	1.32205	0.70387*	**0.60527**
Rand	1.27259	**0.68247**	**0.65586**
Size	1.13441	1.01797	0.98263
1-pt	1.30947	0.75832	0.77909
Yacht			
None	**0.88911**	**0.42583**	**0.34942**
Rand	**1.00206**	**0.39604**	**0.33085**
Size	1.17339	0.83782	0.73834
1-pt	1.36252	1.02317	0.71299

Table 2: Median Of Over-fitting of Best Individuals

Population	50	250	500
Toxicity			
None	22.6673	24.1284	43.325
Rand	20.2454	20.9253*	**34.284**
Size	22.0368	**20.2185**	**19.824**
1-pt	22.4657	20.3742	**25.227**
Bioavailability			
None	0.16730	0.22574	0.44006
Rand	0.14764	**0.17124**	0.22994*
Size	0.13925	**0.17914**	**0.11040**
1-pt	0.16503	0.19926	0.29134
Concrete			
None	0.03018	0.02215	**0.01844**
Rand	0.02720	0.02121	0.02384*
Size	0.03169	0.02821	0.02082*
1-pt	0.03095	0.02789	0.03361
Yacht			
None	**0.08174**	**0.05788**	**0.04359**
Rand	**0.08616**	**0.06225**	**0.04078**
Size	**0.09642**	**0.06370**	0.08677
1-pt	0.14529	0.09217	0.09136

at least as well as Interleaved Sampling. In terms of over-fitting, rather surprisingly, Interleaved-Sampling does not consistently outperform normal GP. However, as in Table 1, interleaved methods in some form perform at least as well as standard GP. Interestingly, Interleaved-Size performs at least as well as the more *informed* Interleaved-Sampling. In terms of average size, Interleaved-Size clearly outperforms the rest of the setups.

2.3 Discussion

The results show that the interleaved use of the training set in *some form* performs at least as well as standard GP on testing fitness and overfitting thus substantiating the idea earlier introduced in [1].

The statistics for Interleaved-Random are often closer to the standard GP than the other two counterparts. Although, the tree sizes with Interleaved-Random are no smaller than normal GP, the approach still gains over standard GP due to savings in data processing.

Table 3: Average Size

Population	50	250	500
Toxicity			
None	98.61	86.39	98.65
Rand	89.17	81.52	83.57
Size	**16.38**	**16.05**	**15.71**
1-pt	84.60	79.86	82.02
Bioavailability			
None	65.09	89.45	98.61
Rand	64.06	79.98	89.17
Size	**17.25**	**16.53**	**16.38**
1-pt	56.15	75.88	84.60
Concrete			
None	34.50	70.93	89.30
Rand	36,96	72.91	82.44
Size	16.35	**16.88**	**16.74**
1-pt	36.87	70.33	70.13
Yacht			
None	72.43	92.58	97.03
Rand	61.10	82.80	82.16
Size	**16.44**	**17.19**	**16.79**
1-pt	21.87	**46.62**	48.12

It is remarkable that Interleaved-Size and Interleaved-Random that do not use any data at all in the interleaved generations are competitive with respect to the other two methods.

To break the tie, we consider the results on tree sizes. We see a clear result in that Interleaved-Size consistently produces much smaller individuals than the rest of the setups. Therefore, Interleaved-Size not only saves the effort in data processing but also successfully utilises the interleaved generation to counter bloat which can be a limiting factor for GP runs.

3. CONCLUSIONS

The results indicate that while interleaved use of the training data set is indeed a useful idea, the originally proposed Interleaved Sampling by no means is the optimal approach. Instead, Interleaved-Random and Interleaved-Size perform just as well across on a range of problems without calling the fitness function *at all* in the interleaved generations. This is particularly useful because a fitness evaluation can be expensive even with a single data point in situations such as when an expensive simulation is needed to evaluate even a single data point.

4. REFERENCES

[1] I. Goncalves and S. Silva. Balancing learning and overfitting in genetic programming with interleaved sampling of training data. In K. Krawiec, A. Moraglio, T. Hu, A. S. Uyar, and B. Hu, editors, *Proceedings of the 16th European Conference on Genetic Programming, EuroGP 2013*, volume 7831 of *LNCS*, pages 73–84, Vienna, Austria, 3-5 Apr. 2013. Springer Verlag.