To Adapt or Not to Adapt, or The Beauty of Random Settings

João E. Batista, Nuno M. Rodrigues, Sara Silva
LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
{jebatista,nmrodrigues,sara}@fc.ul.pt

ABSTRACT

This work concerns the automatic adaptation of the probabilities of occurrence of the genetic operators in Genetic Programming. We experiment with different adaptation methods, different types of problems, and different tree-based Genetic Programming flavors with a variable number of genetic operators. Based on the published literature and on our own results, we claim that operator probabilities should be neither fixed nor carefully adapted, but instead they should be constantly and randomly changed during the evolution.

CCS CONCEPTS

• **Computing methodologies → Genetic programming**;

KEYWORDS

operator probabilities, automatic adaptation, negative results

ACM Reference Format:

João E. Batista, Nuno M. Rodrigues, Sara Silva. 2019. To Adapt or Not to Adapt, or The Beauty of Random Settings. In *Genetic and Evolutionary Computation Conference Companion (GECCO '19 Companion), July 13–17, 2019, Prague, Czech Republic.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3319619.3321994

1 INTRODUCTION

The automatic adaptation of parameters has been on the agenda of Evolutionary Algorithms (EAs) for a long time. This work concerns the automatic adaptation of the probabilities of occurrence of the genetic operators in Genetic Programming (GP). We have studied the application of different adaptation methods, and tested them in two different tree-based GP systems with a varying number of operators, on a wide array of problems of different types.

2 A CRITICAL VIEW OF RELATED WORK

A number of surveys have been published on adaptive methods for EAs, but they include very few studies on the adaptation of operator probabilities in GP, and none using standard tree-based GP. We have looked for articles on automatic parameter adaptation in GP published in the 10 years since the last major survey. Narrowing our search to the adaptation of operator probabilities, we found only five studies. From these, only two use tree-based GP [1, 3], both report truly unconvincing results, and one [3] even recognizes that random probabilities behave just as well as the adaptation methods.

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3321994

3 EXPERIMENTAL SETUP

We have used five different methods for dealing with the operator probabilities, all described below.

FIXED. As the name suggests, this method establishes fixed probabilities for the entire run. In some of the results reported in Section 4 we have used traditional GP probabilities, i.e., 90/10 for crossover/mutation. In other cases we did an attempt to find the "correct" probabilities to use, based on the results of the DAVIS adaptation method (described next).

DAVIS. A method adapted from [2], a classical procedure for adapting operator probabilities in GAs that is fairly sophisticated and highly parameterizable, and also naturally compatible with GP.

POP. An ad-hoc method developed for this work, this procedure updates the probabilities every time an offspring is created. If the offspring has better fitness than the average of the parents, the probability of the method that created it is increased, otherwise it is decreased. For decreasing the probability of an operator, its current probability P is multiplied by a penalty coefficient p ($p \in [0, 1]$, a parameter of the method), and the new probability becomes $P \cdot p$. For increasing the probability of an operator, it is the complementary of its current probability $(1 - P)$ that is multiplied by the same coefficient, and the new probability becomes $1 - ((1 - P) \cdot p)$. After each modification of a probability, the array of all probabilities is normalized so their sum is 1.

IND. This method works exactly like POP, except that each individual has its own set of probabilities. Whenever a new offspring is created, it inherits the probabilities from its parents (unchanged in case of mutation, averaged in case of more than one parent), and these probabilities are then updated and normalized following the same procedure described for POP.

RAND. As the name suggests, this method adopts random probabilities for the operators. In the beginning of each generation, each operator is assigned a random probability, normalized so that the sum for all operators is 1. This is a population-level method, since all the individuals in the population use the same random probabilities.

We have tested the methods in 17 different problems of different types (Table 1), including several symbolic regression and (binary and multiclass) classification problems, as well as the classical Artifical Ant, Even-3 Parity and 11-Multiplexer benchmarks. We have used a standard tree-based GP system (STDGP) with the two standard subtree crossover and subtree mutation operators, and also with an extended set of operators, namely homologous crossover, point mutation, swap mutation (swaps two independent subtrees within the same individual) and shrink mutation (replaces a subtree with one of its subtrees). For the classification problems we have also used the M3GP system [4] (one of the few GP systems capable of doing competent multiclass classification) with its five genetic operators, i.e., subtree crossover and mutation, plus the operators to add, remove, and swap dimensions.

Table 1: List of problems, their characteristics and the (Exp)eriments that used them (see Section 4). The Type of problem can be (R)egression, (C)lassification, or (O)ther.

Name	Samples	Features	Type	Classes	Exp
QUARTIC	21	1	R		1
ANT	n/a	n/a	O		1
PARITY	8	3	O		1
PLEXER	2048	11	O		1
F50	252	241	R		2 , 4
HEART	270	13	C	2	3 , 5
BRAZIL	4872	7	C	2	3 , 5
ADMISSION	500	7	R		4
CONCRETE	1030	8	R		4
WINE	1599	12	R		4
MCD3	322	6	C	3	5
MCD10	6798	6	C	10	5
MOVL	360	90	C	15	5
SEG	2310	19	C	7	5
VOWEL	990	13	C	11	5
WAV	5000	40	C	3	5
YEAST	1484	8	C	10	5

4 RESULTS

Here we split our body of work in different experiments and analyse the fitness of the best individual, on both training and test sets. Statistical significance of the null hypothesis of no difference is determined with the non-parametric Kruskal-Wallis test at $p = 0.01$.

Experiment 1 – Four classical benchmarks. In the first experiment we used the classical four benchmarks (QUARTIC, ANT, PARITY, PLEXER) and obtained results with the methods FIXED, DAVIS, and RAND. The FIXED method used the average probabilities of crossover/mutation found by DAVIS (see Section 3), i.e., 90/10, 45/55, 50/50 and 50/50 on the four benchmarks, respectively. No significant differences were found in terms of fitness.

Experiment 2 – Multi-operator Regression. Given the lack of any significant results in the previous experiment, we tried the same adaptation method (DAVIS) on a harder problem (F50), using a set of six operators (see Section 3). Here we used two variants of the FIXED method, the first using all probabilities equal (FIXED-equal) and the second one using the probabilities found by DAVIS (FIXED-found), that were 10/10/20/40/10/10 for the six genetic operators, by the order specified in Section 3. Only on training fitness the FIXED-found method was borderline significantly better than DAVIS.

Experiment 3 – Binary Classification. We used two binary classification problems (HEART, BRAZIL) and obtained results with four methods (FIXED, IND, POP, RAND). The FIXED method used the traditional 90/10 setting for the crossover/ mutation probabilities. Both IND and POP methods used 0.95 as the penalty coefficient (see Section 3). On HEART, the FIXED method was the worst, particularly on the training set, and RAND was the best, particularly on the test set. On BRAZIL, no significant differences were observed.

Experiment 4 – All Regression. We used four symbolic regression problems (ADMISSION, F50, CONCRETE, WINE), where one of them (F50) had already been used with DAVIS and multiple operators in Experiment 2. We obtained results with four methods (FIXED, IND, POP, RAND), where FIXED had two variants (FIXED 50/50 and FIXED 90/10, corresponding to equal probabilities and traditional probabilities, respectively), IND used a penalty coefficient of 0.75 (IND 75) and POP used a penalty coefficient of 0.95 (POP 95). These differences reflect our attempts at finding statistically significant results. FIXED 90/10 was again the worst method, while FIXED 50/50 was not significantly different from the others. The exception was the atypical F50 problem, exhibiting tremendous overfitting, where FIXED 90/10 was the best on the test set.

Experiment 5 – Classification with M3GP. We used nine classification problems (HEART, BRAZIL, MCD3, MCD10, MOVL, SEG, VOWEL, WAV, YEAST), where only two of them (HEART, BRAZIL) are binary classification problems already used with STDGP in Experiment 3. We used the M3GP system and obtained results with four adaptation methods (FIXED, IND, POP, RAND), where FIXED used the equal probabilities setting (20% for each of the five M3GP operators) and both IND and POP used a penalty coefficient of 0.99. On all nine problems, the only significant differences were that RAND was better than other methods in a few cases.

5 DISCUSSION

The results described above are very easy to summarize: from fixing the probabilities to carefully adapting them, and finally setting them randomly in every generation, most of the time this choice does not make any difference on either the training or the test fitness. The exceptions to the rule are the fixed traditional settings 90/10, that normally produces worse results, and the random settings changed in every generation, that tends to produce better results. Still, only in a handful of cases. The F50 problem is a clear outlier where the 90/10 setting yields better results on the test set. This is probably caused by factors related to bloat control that affect the occurrence of overfitting. Studying these it outside the scope of this work, however it provides new ideas for the future.

6 CONCLUSIONS

Given what we found in the literature and in our own results, we conclude that in tree-based GP the automatic adaptation of operator probabilities is a waste of computational resources that does not result in easier learning or better generalization. We also conclude that randomly assigning new probabilities in every generation is apparently the best option.

ACKNOWLEDGMENTS

UID/CEC/00408/2019, PTDC/CCI-INF/29168/2017, DSAIPA/DS/0022/2018, PTDC/ASP-PLA/28726/2017, PTDC/CTA-AMB/30056/2017, PTDC/EMS-SIS/ /0642/2014 and PTDC/CCI-CIF/29877/2017 from the FCT, Portugal.

REFERENCES

[1] N. Al-Madi and S. A. Ludwig. 2012. Adaptive genetic programming applied to classification in data mining. In *Proceedings of NaBIC 2012*. IEEE Press, 79–85. https://doi.org/doi:10.1109/NaBIC.2012.6402243

[2] Lawrence Davis. 1989. Adapting Operator Probabilities in Genetic Algorithms. In *Proceedings of ICGA 1989*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 61–69. http://dl.acm.org/citation.cfm?id=645512.657242

[3] Jeannie Fitzgerald and Conor Ryan. 2013. Individualized self-adaptive genetic operators with adaptive selection in Genetic Programming. In *Proceedings of NaBIC 2013*. IEEE Press, 232–237. https://doi.org/doi:10.1109/NaBIC.2013.6617868

[4] Luis Muñoz, Sara Silva, and Leonardo Trujillo. 2015. M3GP: Multiclass Classification with GP. In *Proceedings of EuroGP 2015 (LNCS)*, Vol. 9025. Springer, Copenhagen, 78–91. https://doi.org/doi:10.1007/978-3-319-16501-1_7