PSXO – Population-Wide Semantic Crossover

Leonardo Vanneschi,
Mauro Castelli
NOVA IMS, Universidade Nova de Lisboa, 1070-312 Portugal
lvanneschi,mcastelli@novaims.unl.pt

Luca Manzoni
DISCo, Universitá di Milano Bicocca
Milano, Italy 20126
luca.manzoni@disco.unimib.it

Krzysztof Krawiec
Institute of Computing Science
Poznan University of Technology
Poznan, Poland
krawiec@cs.put.poznan.pl

Alberto Moraglio
Department of Computer Science
University of Exeter, UK
A.Moraglio@exeter.ac.uk

Sara Silva
BioISI – Biosystems & Integrative Sciences Institute, Departamento de Informática Faculdade de Ciências, Universidade de Lisboa, 1749-016 Portugal
sara@fc.ul.pt

Ivo Gonçalves
NOVA IMS, Universidade Nova de Lisboa, 1070-312 Portugal
igoncalves@novaims.unl.pt

ABSTRACT

Since its introduction, Geometric Semantic Genetic Programming (GSGP) has been the inspiration to ideas on how to reach optimal solutions efficiently. Among these, in 2016 Pawlak has shown how to analytically construct optimal programs by means of a linear combination of a set of random programs. Given the simplicity and excellent results of this method (LC) when compared to GSGP, the author concluded that GSGP is "overkill". However, LC has limitations, and it was tested only on simple benchmarks. In this paper, we introduce a new method, Population-Wide Semantic Crossover (PSXO), also based on linear combinations of random programs, that overcomes these limitations. We test the first variant (Inv) on a diverse set of complex real-life problems, comparing it to LC, GSGP and standard GP. We realize that, on the studied problems, both LC and Inv are outperformed by GSGP, and sometimes also by standard GP. This leads us to the conclusion that GSGP is not overkill. We also introduce a second variant (GPinv) that integrates evolution with the approximation of optimal programs by means of linear combinations. GPinv outperforms both LC and Inv on unseen test data for the studied problems.

KEYWORDS

Semantics, Population-Wide Crossover, Inverse Matrix, Real-Life Problems

ACM Reference format:
Leonardo Vanneschi,
Mauro Castelli, Luca Manzoni, Krzysztof Krawiec, Alberto Moraglio, Sara Silva, and Ivo Gonçalves. 2017. PSXO – Population-Wide Semantic Crossover. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3067695.3076003

1 INTRODUCTION

In 2016 a paper published by Pawlak [4] claimed that the same excellent performance as GSGP [3] can be obtained, in much shorter running time, by means of a simple linear combination of some particular random programs. That work has cast a shadow on the real usefulness of GSGP, concluding that "geometric semantic genetic programming is overkill".

The objective of this paper is to take the work of Pawlak and provide a more informed opinion on whether GSGP (and more generally GP) is still in demand, or it is really "overkill". More specifically, we build on the same idea and introduce improvements in three different aspects: contrarily to what happens in the algorithm proposed by Pawlak, in this work the initial random trees used to build the linear combination (1) can be of any cardinality; (2) can use any predefined set of primitive functions; (3) can have any format or shape.

2 THE PSXO METHOD AND ITS VARIANTS

Let $p_1, p_2, \ldots, p_m$ be a set of random programs, generated with any initialization method. The objective of the method that we propose in this paper (that we call Inv) is to find the vector of weights such that:

$$[s(p_1), s(p_2), \ldots, s(p_m)]\ \mathbf{w} = \mathbf{t} \tag{1}$$

where, for each $i = 1, 2, \ldots, m$, $s(p_i)$ is the semantics of program p_i and $\mathbf{t}$ is the target vector. As for the LC method, the globally optimal solution is:

$$p^* = \sum_{i=1}^{m} w_i p_i \tag{2}$$

As we can see, p^* is a combination of the individuals p_i in a population. In contrast to LC, the set of programs $p_1, p_2, \ldots, p_m$ can have any cardinality and the programs can be totally random, use any primitive operators and have any form. Equation (2) can be interpreted as a particular type of crossover involving all the individuals in a population. We call this operator Population-Wide Semantic Crossover, and name our method after it.

In this paper, the following variants of PSXO have been studied:

GECCO '17 Companion, Berlin, Germany

DOI: http://dx.doi.org/10.1145/3067695.3076003

(1) Inv. The method described above.

(2) Inv-mod. This variant works like Inv, with the only difference that the population contains k additional individuals, each being a single-node program returning the value of one of k input variables.

(3) GPinv. The objective of this method is to integrate evolution with the Inv method. A population is randomly initialized and an approximation of a globally optimal solution is generated using Equation (2). Then, the population is evolved using GSOs and, at each generation, the current population is used to find another approximation of a globally optimal solution, again using Equation (2).

(4) GPinv-mod. This method integrates evolution with Inv-mod. It works like GPinv, but as it happens for Inv-mod, the initial population additionally contains a single-node program for each variable in the training set, returning the value of that variable. All these variants, in the next section, are compared with: Linear Combination (LC), that is the method presented in [4]; Standard GP (StGP), as in [2]; GSGP, as in [5, 6]; Hybrid, that is a method that combines GSGP with local search, presented in [1].

3 EXPERIMENTAL STUDY

Eight different symbolic regression test problems were considered in the experimental study.

Results show that LC consistently returns the lowest error on the training set for all the studied problems, but it never returns the lowest error on the test set. More specifically, LC is the method that returns the worst results on the test set for 4 over 8 of the studied problems.

Among the proposed methods, Inv and Inv-mod are the ones that are more similar to LC, in the sense that they are not evolutionary algorithms and they try to reconstruct an optimal program by means of a "one-step" linear combinations of random programs. Specifically, Inv and Inv-mod are not the best methods on the training set for any of the studied problems. While on the training set their performance is generally comparable with the one of the other methods, except LC, for 7 over 8 of the studied problems Inv and Inv-mod are outperformed by the evolutionary techniques on the test set.

Finally, GPinv and GPinv-mod return results that are generally comparable with the best of the other evolutionary methods (StGP, GSGP and Hybrid), both on the training and on the test set.

These considerations allow us to conclude that LC, Inv and Inv-mod have returned results that are rather disappointing, which allows us to come at diametrically opposite conclusions compared to the work of Pawlak [4]. In other words, trying to replace evolution by a linear combination of random programs is never beneficial on unseen test data, at least for the problems studied here. Evolution is still appropriate, and GSGP is not overkill.

To investigate the effect of population size on the presented results. we run a set of experiments considering population sizes of 20, 100, 200, 500 and 1000 individuals. The general conclusion we are allowed to draw is that population size does not seem to have a marked influence on the results obtained on the test set. This allows us to conclude that small populations (like for instance a population size equal to 20, as in our simulations) can be used.

4 CONCLUSIONS

In this paper, we have defined an algorithm that, starting from any randomly generated population is able, in one step, to find an approximation of a perfect solution. That solution consists in a linear combination of the initial random trees and the heart of the algorithm is the method that allows us to calculate the weights of this linear combination, i.e., a method able to approximate a matrix inverse, or pseudo-inverse. In the end, the solution returned by the system is a linear combination of a population of individuals, so we could consider it as a particular type of crossover, that we call Population-Wide Semantic Crossover.

Experimental results suggest that, although approximating a perfect solution by means of a linear combination of the individuals in the population can be very useful, evolution is still in demand. In other words, (geometric semantic) genetic programming is *not* overkill.

ACKNOWLEDGMENTS

This work was partially funded by project PERSEIDS (PTDC/EMS-SIS/0642/2014) and BioISI RD unit, UID/MULTI/04046/2013, funded by FCT/MCTES/PIDDAC, Portugal. K. Krawiec acknowledges financial support from the National Science Centre, Poland, grant no. 2014/15/B/ST6/05205.

REFERENCES

[1] Mauro Castelli, Leonardo Trujillo, Leonardo Vanneschi, Sara Silva, Emigdio Z-Flores, and Pierrick Legrand. 2015. Geometric Semantic Genetic Programming with Local Search. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation (GECCO '15)*. ACM, New York, NY, USA, 999–1006. DOI: http://dx.doi.org/10.1145/2739480.2754795

[2] John R. Koza. 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* MIT Press, Cambridge, MA, USA.

[3] Alberto Moraglio, Krzysztof Krawiec, and ColinG. Johnson. 2012. Geometric Semantic Genetic Programming. In *Parallel Problem Solving from Nature - PPSN XII*, CarlosA.Coello Coello, Vincenzo Cutello, Kalyanmoy Deb, Stephanie Forrest, Giuseppe Nicosia, and Mario Pavone (Eds.). Lecture Notes in Computer Science, Vol. 7491. Springer Berlin Heidelberg, 21–31.

[4] Tomasz P. Pawlak. 2016. Geometric Semantic Genetic Programming Is Overkill. In *Genetic Programming: 19th European Conference, EuroGP 2016, Porto, Portugal, March 30 - April 1, 2016, Proceedings*, Malcolm I. Heywood, James McDermott, Mauro Castelli, Ernesto Costa, and Kevin Sim (Eds.). Springer, 246–260. DOI: http://dx.doi.org/10.1007/978-3-319-30668-1_16

[5] Leonardo Vanneschi, Mauro Castelli, Luca Manzoni, and Sara Silva. 2013. A New Implementation of Geometric Semantic GP and its Application to Problems in Pharmacokinetics. In *Proceedings of the 16th European Conference on Genetic Programming, EuroGP 2013 (LNCS)*, Vol. 7831. Springer Verlag, Vienna, Austria, 205–216.

[6] Leonardo Vanneschi, Sara Silva, Mauro Castelli, and Luca Manzoni. 2014. Geometric semantic genetic programming for real life applications. In *Genetic Programming Theory and Practice XI*. Springer New York, 191–209.