New Geometric Semantic Operators in Genetic Programming: Perpendicular Crossover and Random Segment Mutation

Qi Chen, Mengjie Zhang, Bing Xue
School of Engineering and Computer Science Victoria University of Wellington
PO Box 600, Wellington, 6140, New Zealand
Qi.Chen,Mengjie.Zhang,Bing.Xue@ecs.vuw.ac.nz

ABSTRACT

Various geometric search operators have been developed to explore the behaviours of individuals in genetic programming (GP) for the sake of making the evolutionary process more effective. This work proposes two geometric search operators to fulfil the semantic requirements under the theoretical framework of geometric semantic GP for symbolic regression. The two operators approximate the target semantics gradually but effectively. The results show that the new geometric operators can not only lead to a notable benefit to the learning performance, but also improve the generalisation ability of GP. In addition, they also bring a significant improvement to Random Desired Operator, which is a state-of-the-art geometric semantic operator.

CCS CONCEPTS

•Computing methodologies → Search methodologies;

KEYWORDS

Genetic Programming, Symbolic Regression, Geometric Semantic Operators

ACM Reference format:
Qi Chen, Mengjie Zhang, Bing Xue. 2017. New Geometric Semantic Operators in Genetic Programming: Perpendicular Crossover and Random Segment Mutation. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3067695.3076008

1 INTRODUCTION

Different from traditional Genetic Programming (GP) [1], Semantic Genetic Programming (SGP) [3], which is a recently developed variant of GP, makes use of semantic-aware search operators to produce offsprings that are highly correlated with their parents in behaviour. In GP for symbolic regression, the semantics of a program is defined as a vector, the elements of which are the corresponding outputs of the program given the input samples [3]. One particular category of SGP, Geometric Semantic GP (GSGP) [2], which searches directly in the semantic space, opens a new direction to utilise the semantics of GP individuals. However, over-grown offsprings in GSGP, which are caused by the linear combination of parent(s), are expensive to execute in both memory and time. There is a need for the development of an algorithm to fulfil the GSGP theory rather than the linear combination. Recent research has proposed many variants of GSGP to overcome the limitation [4, 6]. However, many of these geometric operators, such as Random Desired Operator (RDO) [4], can eliminate over-grown individuals but never consider the GSGP theory, which is the principle for the success of GSGP. This work aims to fulfil this gap to some extent. The overall goal of this work is to develop two new geometric semantic operators including crossover and mutation to fulfil specific semantic requirements under the theoretical framework of GSGP. A comparison between the proposed geometric operators and RDO will be conducted to investigate the effect of the new operators.

2 GEOMETRIC SEMANTIC GENETIC PROGRAMMING

The theoretical framework of GSGP is defined as follows [2] :

Definition 2.1. Geometric Semantic Crossover: Given two parent individuals with semantics $S(P_1)$ and $S(P_2)$, a geometric semantic crossover generates offspring $O_j (j \in 1, 2)$ having semantics $S(O_j)$ in the segment between the semantics of their parents, i.e., $\|S(P_1), S(P_2)\| = \|S(P_1), S(O_j)\| + \|S(O_j), S(P_2)\|$.

Definition 2.2. Geometric Semantic Mutation: Given a parent P with semantics $S(P)$, r-geometric semantic mutation produces offspring O in a ball of radius r centered in P, i.e., $\|S(P), S(O)\| \leq r$.

3 THE PROPOSED OPERATORS

This work aims to propose two new geometric operators, which are called *perpendicular crossover* and *random segment mutation*, to fulfil new semantics requirements under the theoretical framework of GSGP, which are more specific than the originally desired semantics for the offspring in both GSGP and RDO. In GSGP, the semantics of the new generation rely on the parent(s), while RDO only considers the semantics of the target. The new geometric operators utilise the semantics of both the target and the parent(s). For presentation convenience, GP with the two new geometric operators is named *NGSGP*.

3.1 Perpendicular Crossover

Given two parent individuals, perpendicular crossover is a semantic search operator that generates offsprings having two geometric properties. The first property is that the offsprings need to stand in the vector defined by their parents. The second is that the offspring should make the vector, which is defined by the target semantics and the offspring point, perpendicular to the given vector of their parents. Suppose the target semantic is $\vec{T}$, and the semantics of the two parents are $\vec{P}_1$ and $\vec{P}_2$. α refers to the angle between the relative

GECCO '17 Companion, Berlin, Germany

DOI: http://dx.doi.org/10.1145/3067695.3076008

Table 1: Benchmark Problems

Name	# Features	#Total Instances	#Training Instances	#Test Instances
LD50	626	234	163	71
DLBCL	7399	240	160	80

semantics of $\vec{P}_2$ and $\vec{T}$ to $\vec{P}_1$, while β is its counterpart to $\vec{P}_2$. The angle α and β are defined as follows:

$$\alpha = \arccos\left(\frac{(\vec{P}_1 - \vec{T}) \cdot (\vec{P}_1 - \vec{P}_2)}{\|\vec{P}_1 - \vec{T}\| \cdot \|\vec{P}_1 - \vec{P}_2\|}\right) \quad \beta = \arccos\left(\frac{(\vec{P}_2 - \vec{T}) \cdot (\vec{P}_1 - \vec{P}_2)}{\|\vec{P}_2 - \vec{T}\| \cdot \|\vec{P}_1 - \vec{P}_2\|}\right) \tag{1}$$

where $(\vec{P}_1 - \vec{T}) \cdot (\vec{P}_1 - \vec{P}_2) = \sum_{i=1}^{n}(p_{1i} - t_i) \cdot (p_{1i} - p_{2i})$, $\|\vec{P} - \vec{T}\| = \sqrt{\sum_{i=1}^{n}(p_i - t_i)^2}$ and $\|\vec{P}_1 - \vec{P}_2\| = \sqrt{\sum_{i=1}^{n}(p_{1i} - p_{2i})^2}$. p_{1i}, p_{2i} and t_i are the values of $\vec{P}_1$, $\vec{P}_2$ and $\vec{T}$ in the ith dimension, respectively.

The parametric equation is used to express a line in the semantic space. Specifically, suppose L is the line given by the two parents $\vec{P}_1$ and $\vec{P}_2$ in an n dimensional space, the semantics of the offspring program O in the line L is given in Equation (2).

$$\vec{O} = \vec{P}_1 + k \cdot V_L \tag{2}$$

where $V_L = \vec{P}_1 - \vec{P}_2$ is a vector, the elements of which are defined as $\{p_{11} - p_{21}, p_{12} - p_{22}, \ldots, p_{1n} - p_{2n}\}$. $k = \|\vec{P}_1 - \vec{O}\| / \|\vec{P}_1 - \vec{P}_2\|$ is a real number parameter. When $0 < k < 1$ ($\alpha < 90$ and $\beta < 90$), $\vec{O}$ is a point on the segment between P and Q. Further, if $k < 0$ ($\alpha > 90$), $\vec{O}$ is outside the segment on the $\vec{P}_1$ side, while if $k > 1 (\beta > 90)$, $\vec{O}$ is outside on the $\vec{P}_2$ side.

3.2 Random Segment Mutation

Random segment mutation (RSM) is a kind of geometric mutation, on which the desired semantics of the offspring is standing in the segment of the parent and the target point in the semantic space. Firstly, RSM needs to find the segment between the target semantic $\vec{T}$ and the semantics of the parent $\vec{P}$. Then a random point is obtained along this segment, which is treated as the desired semantics of the offspring $\vec{O}$. RSM makes a small but very important change to RDO, i.e. RDO treats the target semantics as the desired semantics for all the offspring, while RSM utilises the target semantics in an implicit way.

When implementing the perpendicular crossover and RSM in NGSGP, the semantic backpropagation [4] and semantic library search are applied to the parent(s) to obtain the desired semantics.

4 THE EXPERIMENT

To investigate the effectiveness of the two new geometric search operators, a set of experiments have been conducted to compare NGSGP with RDO. In addition, the performance of GP using only the perpendicular crossover (PC) and GP with only random segment mutation (RSM) are also examined. Standard GP is used as a baseline for comparison. The methods are tested on two real-world datasets [5, 6], as shown in Table 1. Each GP method has been conducted for 100 independent runs on each problem. For comparison, the root mean square error (RMSE) of the best-of-run model on the training set and its corresponding test error are recorded.

The mean and standard deviation of RMSEs achieved by the 100 best-of-run programs on the training set and the test sets are shown in Table 2. The minimum values among the five methods are marked in bold. The four geometric semantic GP methods generally have much smaller RMSEs than standard GP on the training datasets. Compared with RDO, NGSGP has a higher training error on LD50 but a smaller training error on DLBCL. On LD50 and DLBCL, RDO generalises worse than standard GP. On LD50, RDO achieves the best training performance but the worst generalisation (test) performance among the methods, which indicates its overfitting to the training set. However, on the two test sets, NGSGP can generalise better than standard GP and significantly better than RDO.

Table 2: Training and Test Errors of the 100 Best Programs

	GP	RDO	RSM	PC	NGSGP
	Mean±Std	Mean±Std	Mean±Std	Mean±Std	Mean±Std
			Training		
LD50	1950.94±67.66	**1692.2±317.1**	1888.59±108.26	1952.63±51.68	1844.52±88.5
DLBCL	0.65±0.02	0.63±0.07	0.65±0.04	0.62±0.03	**0.57±0.08**
			Test		
LD50	2007.5±67.1	4354.9±9236.7	1996.8±79.4	2020.7±83.3	**1987.88±87.5**
DLBCL	0.7±0.04	0.71±0.04	0.7±0.04	0.69±0.05	**0.62±0.07**

NGSGP is guided by many intermediate semantic targets (different desired semantics for each offspring) under the theoretical requirement of GSGP, which can help maintain the semantic diversity of the population better than utilising only one target (i.e. the target semantics) in RDO. Higher semantic diversity will lead to a better exploration ability of GP and has a positive effect on enhancing the effectiveness of the evolutionary search. On the other hand, when tackling the real-world data containing noise, the property of less greedy to the target semantics in NGSGP will lead to the reduction of overfitting, thus generalise better.

5 CONCLUSIONS

This work develops two new geometric semantic operators to fulfil better desired semantics, which are under the theoretical requirement of GSGP, for the offspring programs in GP. The effect of the proposed operators has been confirmed by the notable improvement on learning and generalisation performance over standard GP and GP using RDO.

For future work, we are interested in speeding up the new operators by improving the algorithm on semantic library search and introducing bloat free mechanism to the new operators. Moreover, instead of approximating the semantic requirement by semantic backpropagation and library search, we also plan to fulfil the semantic requirement in a more accurate way.

REFERENCES

[1] John R Koza. 1992. *Genetic programming: on the programming of computers by means of natural selection.* Vol. 1. MIT press.

[2] Alberto Moraglio, Krzysztof Krawiec, and Colin G Johnson. 2012. Geometric semantic genetic programming. In *International Conference on Parallel Problem Solving from Nature.* Springer, 21–31.

[3] Quang Uy Nguyen, Hoai Nguyen Xuan, and Michael O...Neill. 2009. Semantic Aware Crossover for Genetic Programming: The Case for Real-Valued Function Regression. In *Genetic Programming, European Conference, Eurogp 2009, Tbingen, Germany, April 15-17, 2009, Proceedings.* 292–302.

[4] Tomasz P Pawlak, Bartosz Wieloch, and Krzysztof Krawiec. 2015. Semantic backpropagation for designing search operators in genetic programming. *IEEE Transactions on Evolutionary Computation* 19, 3 (2015), 326–340.

[5] Andreas Rosenwald, George Wright, Wing C Chan, Joseph M Connors, Elias Campo, Richard I Fisher, Randy D Gascoyne, H Konrad Muller-Hermelink, Erlend B Smeland, Jena M Giltnane, and others. 2002. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine* 346, 25 (2002), 1937–1947.

[6] 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 *European Conference on Genetic Programming.* Springer, 205–216.