Analysis of Semantic Building Blocks via Gröbner Bases

Jerry Swan[1], Geoffrey K. Neumann[1], Krzysztof Krawiec[2]
1. University of Stirling, Stirling FK9 4LA, UK.
2. Poznań University of Technology, Piotrowo 2, 60-965 Poznań, POLAND.
jerry.swan@cs.stir.ac.uk,gkn@cs.stir.ac.uk,krawiec@cs.put.poznan.pl

1 Motivation

The Buchberger 'Gröbner Basis' algorithm (GB) [1] generalises the Euclidian algorithm for greatest common divisor from univariate to multivariate polynomials. Since its invention in 1965, it has found increasingly diverse applications across mathematics and computer science. In this article, we describe the application of Gröbner Bases to determine semantic commonality amongst a population of programs. The motivation for these experiments comes from the lack of strong support for 'building blocks' in traditional Genetic Programming (GP) [2]. This is hindered by the many-to-one mapping between a program and the value it calculates: for example, the programs $2 * x$ and $x + x$ are syntactically different but semantically equivalent.

In order to better identify building blocks in GP, the search for an equivalent to the 'schema theorem' developed for Genetic Algorithms has long been of interest [3]. In the 1980s and early 1990s, analysis of Genetic Algorithms was extensively developed via the application of Walsh transforms [4], the binary equivalent of the Fourier transform. It is well-known that Walsh functions form a basis (i.e. generating set) for the space of binary strings. In an analagous manner, we analyse the effect of various GP parameters on the presence of common semantic substructure not syntactically represented in the genotype [5], in a series of experiments applied to multivariate polynomial regression problems. Unlike regression schemes which require the polynomial degree to be determined *a priori*, GP using the function set $\{+, -, *\}$ is capable of generating any polynomial, the idea being that the evolutionary process will cause the required degree to arise under selection pressure.

2 Gröbner Bases

By using the function set above and the input variables $\{x_1, \ldots x_n\}$, a GP population can be considered to be a *generating set* for the space of multivariate polynomials in n variables, capable of expressing arbitrary sums and products. This specific type of generating set gives rise to an *ideal*, formally defined as follows: let $\mathcal{R} = \mathbb{F}[x_1, \ldots, x_n]$ be the ring of polynomials in n variables with coefficients in the field $\mathbb{F}$. Given a set of polynomials $P = \{p_1, \ldots, p_k\} \in \mathcal{R}$, then the *ideal* $\langle P \rangle$ *generated by* P is a subset of $\mathcal{R}$ given by

$$\langle P \rangle = \Big\{ \sum_{i=1}^{k} r * p_i, \forall r \in \mathcal{R} \Big\}$$

Given a generating set P, the GB algorithm calculates a basis for the ideal $\langle P \rangle$, i.e. for every set of polynomials P, there exists a GB G such that $\langle P \rangle = \langle G \rangle$. Furthermore, for a fixed total ordering on terms, the GB of an ideal is unique.

For example, if $P = \{x^2 + y^2 + z^2 - 1, x^2 + y^2 + z^2 - 2x, 2x - 3y - z\}$, then it can be shown that the GB for $\langle P \rangle$ is $\{2x - 1, 3y + z - 1, 40z^2 - 8z - 23\}$. Hence the elements of P can be re-expressed as a combination of the elements of GB.

3 Methodology

By the use of the Gröbner basis algorithm, we may uncover common substructure not syntactically represented in the genotype. If these are indeed building blocks in the 'schema theorem' sense, then increasing their prevalence in the population will lead to an increase in solution quality. It is known that the GB algorithm has exponential complexity in the worst case, and hence we are concerned primarily here with establishing its merits as an analytical tool. However, it should be noted that from the ubiquity of GB that this worst-case is not often an obstacle in practice.

To test the above hypothesis we make use of Automatically Defined Functions (ADF). ADFs can be considered to be equivalent to subroutine calls, dynamically added to the function set. Every m generations, GBs are calculated for the population and used to determine common substructure, which is then promoted to ADF status. We will present the results of our analysis of some well-known multivariate symbolic regression problems at the conference.

References

[1] B. Buchberger. *An Algorithm for Finding the Basis Elements in the Residue Class Ring Modulo a Zero Dimensional Polynomial Ideal.* PhD thesis, 3 2006.

[2] Una-May O'Reilly and Franz Oppacher. The troubling aspects of a building block hypothesis for genetic programming. In *Foundations of Genetic Algorithms*, pages 73–88, 1994.

[3] R. Poli, L. Vanneschi, W. B. Langdon, and N. F. McPhee. Theoretical results in genetic programming: The next ten years? *Genetic Programming and Evolvable Machines*, 11(3/4):285–320, 2010.

[4] Michael D. Vose and Alden H. Wright. The Simple Genetic Algorithm and the Walsh transform: Part I, theory. *Evol. Comput.*, 6(3):253–273, September 1998.

[5] Nicholas Freitag McPhee, Brian Ohs, and Tyler Hutchison. Semantic building blocks in genetic programming. In *Genetic Programming*, pages 134–145. Springer, 2008.