Semantically-meaningful Numeric Constants for Genetic Programming

Jerry Swan[1], John Drake[2], Krzysztof Krawiec[3]

1. University of Stirling, Stirling FK9 4LA, UK.

2. University of Nottingham Ningbo China, Ningbo, 315100, CHINA.

3. Poznań University of Technology, Piotrowo 2, 60-965 Poznań, POLAND.

jerry.swan@cs.stir.ac.uk,psxjd2@nottingham.ac.uk,krawiec@cs.put.poznan.pl

Theories may be equivalent in all their predictions and are hence scientifically indistinguishable. However [...] different views suggest different kinds of modifications which might be made and hence are not equivalent in the hypotheses one generates from them.

Richard P. Feynman, Nobel Lecture 1965.

Motivation

Symbolic Regression (SR) via Genetic Programming (GP) enjoys two key advantages over other forms of model fitting:

- Model agnosticism: there is less *a priori* comittment to the specific model to be fitted than in e.g. linear or polynomial regression.
- Explanatory power: the symbolic nature of the resulting expressions means that large volumes of data can potentially be compressed into a succinct (and hopefully conceptually malleable/generalizable) form.

Ephemeral Random Constants (ERCs) are part of the standard SR toolkit. It is well-known that GP is not good at regressing numeric constants as terminal nodes [1] and ERCs thus serve a dual role: both helping to obtain a better fit in the presence of noise and finding constants that rightfully reproduce the semantics of the underlying model. A disadvantage of ERCs is their contribution to overfitting: standard GP mechanisms for introducing ERCs are 'blind' in the sense that there is no means of biasing them towards either of these roles. Previous work by Lipson et al. [2] explicitly modelled noise distribution in order to perform better in the presence of asymetric noise. We adopt a contrary approach and explicitly bias the semantically-aware role. By the use of *inverse equation solvers*, ERCs can be replaced by closed-form expressions involving rationals and fundamental constants (e.g. π, ϕ, e). With these particular biases, such an approach might be used to help discover physical laws from experimental data.

Inverse Equation Solvers

An inverse equation solver takes a real value as input and outputs one or more closed-form approximations. Recall that a closed form expression is one that can be evaluated analytically via finitely many applications of *elementary functions*, viz. rational constants,$+, -, *, \div$, n^{th} roots, the exponential function and its inverse. While the human eye is unlikely to notice that $\frac{1096.6331584}{e^3}$ has a

near-equivalent closed-form expression, by using an inverse equation solver to replace the numerator with e^7 (accurate to $2.84585e^{-8}$), the entire expression can be reduced to e^4 (accurate to $1.41686e^{-9}$). By this means, we have the potential to simultaneously increase both economy of expression and accuracy. Publicly-available inverse solvers include RIES[1] and the Inverse Symbolic Calculator[2] [3], the latter making use of the PSLQ algorithm for finding integer relations. By contrast, RIES employs a bidirectional search on the graph of closed-form expressions, using the specified real value as target and with the source chosen from some nearby transcendental constant or rational value. By using RIES, it is generally possible to find a semantically-rich closed-form expression that matches an ERC to within 10 significant digits within a few seconds.

Experiments

We used standard tree-based GP[3], with tournament selection, point mutation, subtree crossover and loss function given by RMS error. We compare this setup against a variant in which, at the end of each generation, each ERC in the population is replaced with a closed-form expression generated by RIES. There is an option for pre- and post- simplification (via Mathematica's *Simplify* function) of the GP tree. We will present two experiments intended to demonstrate the utility of this approach. Firstly, the well-known 'Buffon's Needle' experiment consists of a series of trials in which a needle is dropped on a surface marked with equally-spaced parallel lines. This allows an approximation of π to be computed as a function of the number of needles crossing lines. As a simple example of how this method can be used to discover generalizable structure in data, we demonstrate that π can be *symbolically* induced by standard GP augmented with an inverse solver. Secondly, the Fibonacci sequence is defined by the recurrence relation $F_0 = 0, F_1 = 1, F_{n+1} = F_{n-1} + F_n$. It is well-known that the ratio of successive terms of the sequence converges to the *golden ratio* $\phi = \frac{1+\sqrt{5}}{2}$.

The symbolic induction of these constants from noisy data is the simplest exemplar of the replacement of ERCs with 'semantically-meaningful priors'. Experimental results exhibiting encouraging solution quality and convergence behaviour will be presented at the workshop.

References

[1] M. Evett, T. Fernandez, Numeric mutation improves the discovery of numeric constants in genetic programming, in: Genetic Programming: 3rd Annual Proceedings, Morgan Kaufmann, Wisconsin, USA, 1998, pp. 66–71.

[2] M. D. Schmidt, H. Lipson, Learning noise, in: GECCO Proceedings, 2007, pp. 1680–1685.

[3] D. H. Bailey, S. Plouffe, Recognizing numerical constants, in: The Organic Mathematics Project Proceedings, Vol. 20, Canadian Mathematical Society, Ottawa, ON K1G 3V4, Canada, 1997, pp. 73–88.

[1] `http://mrob.com/pub/ries/`

[2] `http://oldweb.cecm.sfu.ca/projects/ISC/isc_info.html`

[3] `http://www.epochx.org/`