Chapter 8

ACCURACY IN SYMBOLIC REGRESSION

Michael F. Korns

Korns Associates, 98 Perea Street, Makati 1229, Manila Philippines

Abstract

This chapter asserts that, in current state-of-the-art symbolic regression engines, accuracy is poor. That is to say that state-of-the-art symbolic regression engines return a champion with good fitness; however, obtaining a champion with the correct formula is not forthcoming even in cases of only one basis function with minimally complex grammar depth.

Ideally, users expect that for test problems created with no noise, using only functions in the specified grammar, with only one basis function and some minimal grammar depth, that state-of-the-art symbolic regression systems should return the exact formula (or at least an isomorph) used to create the test data. Unfortunately, this expectation cannot currently be achieved using published state-of-the-art symbolic regression techniques.

Several classes of test formulas, which prove intractable, are examined and an understanding of why they are intractable is developed. Techniques in Abstract Expression Grammars are employed to render these problems tractable, including manipulation of the epigenome during the evolutionary process, together with breeding of multiple targeted epigenomes in separate population islands.

A selected set of currently intractable problems are shown to be solvable, using these techniques, and a proposal is put forward for a discipline-wide program of improving accuracy in state-of-the-art symbolic regression systems.

Keywords: Abstract Expression Grammars, Differential Evolution, Grammar Template Genetic Programming, Genetic Algorithms, Particle Swarm, Symbolic Regression.

1. Introduction

The discipline of Symbolic Regression (SR) has matured significantly in the last few years. There is at least one commercial package on the market for several years http://www.rmltech.com/. There is now at least one well documented commercial symbolic regression package available for
There is at least one very well done open source symbolic regression package available for free download http://ccsl.mae.cornell.edu/eureqa. In addition to our own ARC system (Korns, 2010), currently used internally for massive (million row) financial data nonlinear regressions, there are a number of other mature symbolic regression packages currently used in industry including (Smits et al., 2010) and (Castillo et al., 2010). Plus there is an interesting work in progress by (McConaghy et al., 2009).

During the process of enhancing our ARC system with the latest thinking in published symbolic regression papers, we ran across several test problems for which our system failed to return the correct formula. Normally this is not surprising in large scale regression with much noise; however, these test problems were generated with no noise and were fairly simplistic formulas of only one basis function with minimally complex grammar depth.

After further study it is now apparent that there are very large numbers of simple test formulas against which current state-of-the-art symbolic regression systems suffer poor accuracy. For these intractable problems state-of-the-art symbolic regression engines fail to return a champion with the correct formula.

This is a serious issue for several reasons. First, users expect to receive a correct formula when inputting a simple test case. When a correct formula is not forthcoming, user interest and trust in the symbolic regression system wanes. Second, if symbolic regression cannot return a correct formula in even simplistic test cases then symbolic regression loses its differentiation from other black box machine learning techniques such as support vector regression or neural nets. Third, from its very inception (Koza, 1992) symbolic regression has been represented as a technique for returning, not just coefficients, but a correct formula. If this claim cannot be fulfilled by independent scientific review, a serious reputational issue will develop and research money will flow in other directions.

This chapter begins by outlining the accuracy issue. A simple symbolic regression grammar of fifteen obvious mathematical functions is established. All test cases are limited to a single basis function of no more than three grammar nodes deep. For all test cases the data is limited to ten thousand sample points, of five features each, and absolutely no noise. Even with these very severe limitations, large numbers of simple formulas are shown to be intractable.

The chapter continues with an examination of techniques which allow ARC to solve these previously intractable problems. The chapter closes with a proposal for a discipline wide approach to solving our symbolic regression accuracy issues.