Chapter 17

PARADOX-FRONT EXPLOITATION IN SYMBOLIC REGRESSION

Guido F. Smits and Mark Kotanchek

1 Dow Benelux, Terneuzen, NV; 2 Dow Chemical, Midland, MI USA

Abstract
Symbolic regression via genetic programming (hereafter, referred to simply as symbolic regression) has proven to be a very important tool for industrial empirical modeling (Kotanchek et al., 2003). Two of the primary problems with industrial use of symbolic regression are (1) the relatively large computational demands in comparison with other nonlinear empirical modeling techniques such as neural networks and (2) the difficulty in making the trade-off between expression accuracy and complexity. The latter issue is significant since, in general, we prefer parsimonious (simple) expressions with the expectation that they are more robust with respect to changes over time in the underlying system or extrapolation outside the range of the data used as the reference in evolving the symbolic regression.

In this chapter, we present a genetic programming variant, ParetoGP, which exploits the Pareto front to dramatically speed the symbolic regression solution evolution as well as explicitly exploit the complexity-performance trade-off. In addition to the improvement in evolution efficiency, the Pareto front perspective allows the user to choose appropriate models for further analysis or deployment. The Pareto front avoids the need to a priori specify a trade-off between competing objectives (e.g. complexity and performance) by identifying the curve (or surface or hyper-surface) which characterizes, for example, the best performance for a given expression complexity.

Keywords: genetic programming, Pareto front, multi-objective optimization, symbolic regression, ParetoGP

1. Introduction

Unlike normal regression in which a model structure (e.g., second-order polynomials) is hypothesized and fit to available data, symbolic regression involves the discovery of the structure as well as the coefficients within that structure. One way to accomplish this is to use genetic programming (GP)
techniques to evolve expressions which match the observed system behavior. In this section, we briefly review the practical motivations for symbolic regression as well as the classical problems characteristic to the GP approach. Finally, we outline a variant of GP which addresses some of the classical issues and has resulted in a significant improvement in the speed and robustness of symbolic regression. This variant focuses on the evolutionary effort on improving the Pareto front (which captures the trade-offs between competing objectives) rather than optimizing a single composite criteria. The rest of the paper is devoted to exploring the algorithm, its benefits and its performance.

In this chapter we assume that the reader has a working knowledge of GP concepts (Banzhaf et al., 1998, Jacob, 2001)) as well as its application to symbolic regression (Kotanchek et al., 2003).

Motivations for Symbolic Regression

In addition to the real-world benefits of empirical modeling for system modeling, emulation, monitoring and control, symbolic regression has several unique contributions. These contributions, which are especially important when faced with multivariate data from a nonlinear but unknown system, include:

- **Human insight** — examination of the evolved expressions can be indications of underlying physical mechanisms as well identification of *metavariables* (combinations or transforms of variables) which can simplify subsequent empirical modeling efforts. Additionally, examining the structure of an evolved model can be comforting in the sense that the model behavior, variables and metavariables agree with human expectation; this explainability helps to instill trust in the model(s).

- **Compact models** — generally solutions can be identified which perform well and are parsimonious with respect to structure complexity and/or number of input variables. Such models are attractive because they can be interpreted more easily and deployed easily in many environments. Such models may also be more robust and capture underlying fundamentals rather than system noise in the data.

- **Limited a priori assumptions** — unlike traditional regression which assumes a model structure for the data, symbolic regression allows the data to determine which structures of variables, functions and constants are appropriate to describe the observed behavior. Of course, appropriate functional building blocks as well as the pertinent data variables must be supplied for the evolution.

- **Natural variable selection** — the evolutionary processes of GP have a remarkable ability to focus on the driving variables necessary to capture