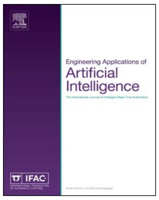
Engineering Applications of Artificial Intelligence 55 (2016) 292–306

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Inference of compact nonlinear dynamic models by epigenetic local search

William La Cava [a],*, Kourosh Danai [a], Lee Spector [b]

[a] Department of Mechanical and Industrial Engineering, University of Massachusetts, Amherst, USA
[b] School of Cognitive Science, Hampshire College, Amherst, USA

ARTICLE INFO

Article history:
Received 3 December 2015
Received in revised form
25 June 2016
Accepted 19 July 2016

Keywords:
System identification
Genetic programming
Dynamical systems
Differential equations
Symbolic regression

ABSTRACT

We introduce a method to enhance the inference of meaningful dynamic models from observational data by genetic programming (GP). This method incorporates an inheritable epigenetic layer that specifies active and inactive genes for a more effective local search of the model structure space. We define several GP implementations using different features of epigenetics, such as passive structure, phenotypic plasticity, and inheritable gene regulation. To test these implementations, we use hundreds of data sets generated from nonlinear ordinary differential equations (ODEs) in several fields of engineering and from randomly constructed nonlinear ODE models. The results indicate that epigenetic hill climbing consistently produces more compact dynamic equations with better fitness values, and that it identifies the exact solution of the system more often, validating the categorical improvement of GP by epigenetic local search. The results further indicate that when faced with complex dynamics, epigenetic hill climbing reduces the computational effort required to infer the correct underlying dynamics. We then apply the method to the identification of three real-world systems: a cascaded tanks system, a chemical distillation tower, and an industrial wind turbine. We analyze its solutions in comparison to theoretical and black-box approaches in terms of accuracy and intelligibility. Finally, we analyze population homology to evaluate the efficiency of the method. The results indicate that the epigenetic implementations provide protection from premature convergence by maintaining diversity in silenced portions of programs.

1. Introduction

A major goal of science is to characterize analytically the dynamic behavior of natural phenomena associated with biological, ecological, social, and economic systems, as well as the dynamics of artifacts such as wind turbines, robots, and aircraft. Dynamic behaviors are usually characterized by differential equations which in aggregate represent the dynamic model of the system. These dynamic models are the essence of the simulations that estimate/predict system behavior for policy decisions, design, optimization, control, and/or automation. This paper presents a method for construction of concise and mechanistically meaningful dynamic models from observations.

Dynamic models are preferably formulated according to first principles, to embody the knowledge of the process. However, first-principles models cannot often fully characterize the nonlinear dynamics of the process, as represented by process observations. In regress, first-principles models may be abandoned in favor of empirical models such as neural networks (Narendra and Parthasarathy, 1990; Gregorčič and Lightbody, 2008), linear or nonlinear autoregressive moving average (ARMAX) models (Ljung, 1999; Billings, 2013), or others (Ni et al., 1996; Sadollah et al., 2015), that have the structural flexibility to accommodate the measured process observations. Although these empirical models provide an effective basis for estimation/prediction, they have two major drawbacks. One is their 'black-box' format which obscures the knowledge of the process acquired through adaptation. The second is their case-specificity which makes them potentially deficient in representing the process under conditions (inputs) not encompassed by the measured observations. To remedy the black-box nature of these empirical models, dynamic models consisting of differential equations can be defined in algebraic form by symbolic regression (Gray et al., 1998; Cao et al., 2000; Bongard and Lipson, 2007), wherein both the structure (topology) and parameters (constants) are inferred from measured observations. Since these symbolic models are intelligible, they have the capacity to elucidate the process physics. Symbolic regression is typically conducted using genetic programming (GP) (Koza, 1992), which is a bio-inspired machine learning technique that

* Corresponding author.
E-mail address: wlacava@umass.edu (W. La Cava).

http://dx.doi.org/10.1016/j.engappai.2016.07.004
0952-1976/© 2016 Elsevier Ltd. All rights reserved.

constructs candidate models from mathematical building blocks and proceeds with selection, recombination and mutation over several generations before converging on a model that best fits the process observations.

In comparison to system identification methods that presume fixed model structures, symbolic regression can be computationally expensive because of its expanded search space. Furthermore, when guided solely by an error metric, it can yield unwieldy equations that are elusive to physical interpretation. To remedy these shortcomings, this paper introduces a new method of symbolic regression that fine-tunes candidate model structures by local search (La Cava et al., 2015). This fine tuning is enabled by the addition of an epigenetic layer for selection of program components (consisting of variables and instructions) to be included in the model. The incorporation of this epigenetic layer is motivated by two hypotheses: first, that the benefits of epigenetic regulation observed in biology may confer analogous improvements on GP systems; and second, that generalized local search methods enabled by epigenetics may improve the ability of GP to find correct model structures.

As to the first hypothesis, despite the highly regulated nature of biological genes, the role of epigenetics in regulating gene expressions is traditionally ignored in GP (with some exceptions, e.g. (Ferreira, 2001)). However, epigenetic processes may provide several evolutionary benefits. For example, because epigenetic processes allow the underlying genotype to encode various expressions and lead to neutral variation through crossover and mutation of non-coding segments, they may allow populations to avoid evolutionary bottlenecks or let them respond to changing evolutionary pressures (Jablonka and Lamb, 2002). Also, because they provide for phenotypic plasticity that enables gene expression to change in response to environmental pressure (Dias and Ressler, 2013), they may allow gene expression adaptations to be inherited in offspring without explicit changes to the genotype. This property legitimizes, via epigenetic processes, once discredited ideas of Lamarck pertaining to the inheritability of lifetime adaptations (Jablonka and Lamb, 2002; Holliday, 2006).

Regarding the second hypothesis, although local search methods have been developed and integrated into evolutionary algorithms (Gruau and Whitley, 1993; Whitley et al., 1994; Jeong and Lee, 1996; Ross, 1999; Giraud-Carrier, 2002), especially in genetic algorithms (GAs) through prescribed changes to the genotype, the role of structure optimization in symbolic regression is typically left to the GP process. Aside from some recent developments (Arnaldo et al., 2014), local search is traditionally conducted at the genome level. More generic local search methods, like tree snipping (Bongard and Lipson, 2007), focus on improving secondary metrics like size or legibility, whereas the traditional search methods, like stochastic hill-climbing (Bongard and Lipson, 2007), linear (Iba and Sato, 1994) or non-linear regression (Topchy and Punch, 2001) are confined to constant optimization. Although these local search methods improve symbolic regression performance, they cannot aid the search for program topology.

Epigenetics, on the other hand, provide a natural basis for performing local search at the structural (i.e., program topology) level. Motivated by this benefit of epigenetics, we introduce in this paper an epigenetics-enabled GP system to conduct topological optimization of programs at the level of gene expression. The contributions of this method are twofold: first, it introduces a generic method of topological search of the space of individual genotypes via modifications to gene expression. Second, it improves programs without affecting the genotype and without discarding the acquired knowledge gained through evolution, thereby lowering the risk of premature convergence observed in previous studies (Whitley et al., 1994). These contributions are achieved by conducting local search on the epigenome rather than the genome and making these adaptations inheritable via evolutionary processes.

The proposed Epigenetic Linear Genetic Programming (ELGP) method is tested on a large array of data generated from nonlinear ordinary differential equations (ODEs), as well as from three real-world processes, to evaluate the quality of its solutions. The paper is organized as follows. We formulate in Section 2 the identification problem and describe in Section 3 the ELGP method and its application to inference of dynamic models. We also review the relevant work in the context of GP and nonlinear dynamics modeling in Section 4. We then present the experimental analysis of different epigenetic implementations on a series of increasingly complex problems in Section 5. We begin by testing the method on a large set of data obtained from simulated nonlinear ODEs in different engineering fields, in order to illustrate its breadth of application. We then perform identification on hundreds of randomly constructed nonlinear systems, varying in complexity and dimensionality, to evaluate the scalability of the method in comparison to traditional GP approaches. Finally, we apply the ELGP method to three real-world problems, including the identification of (1) a benchmark cascaded tanks system (Wigren and Schoukens, 2013), (2) a chemical distillation tower, and (3) an industrial wind turbine. The results are presented in Section 6 and include comparisons of ELGP's performance in relation to other linear and nonlinear identification methods. We finish this discussion with an analysis of population diversity to study how gene expression evolves for each ELGP implementation.

2. Problem statement

The underlying assumption of symbolic regression is that there exists an analytical model of the system that would generate the measured observations $y(t_k)$ at the sample times $t_k = t_1, \ldots, t_N$ under the input, $\mathbf{u}(t)$, as

$$y(t_k, \mathbf{u}) = \hat{y}(t_k, M^*(\mathbf{x}, \mathbf{u}, \mathbf{\Theta}^*)) + \nu; \quad k = 1, \ldots, N \tag{1}$$

where $\hat{y}$ is the model output, ν represents measurement noise in y, $\mathbf{x} = [x_1, \ldots, x_n]^T$ is the vector of state variables, and $M^*(\mathbf{x}, \mathbf{u}, \mathbf{\Theta}^*)$ is the correct model form embodied by the correct parameter values $\mathbf{\Theta}^*$, written M^* hereafter for brevity. In the search for the correct model form M^*, GP typically attempts to solve the problem

$$\text{minimize } f(M) \text{ subject to } M \in \mathfrak{G} \tag{2}$$

where $\mathfrak{G}$ is the space of possible models M, and f denotes a minimized fitness function. Given that it is impractical to exhaustively search $\mathfrak{G}$, the model found to minimize $f(M)$ may only be locally optimal. For practical purposes it is assumed that a suboptimal model can nevertheless fulfill the purpose of adequately representing the process, as depicted by the measured observations.

A common choice for estimating a candidate model output $\hat{y}(\widehat{M})$ is numerical integration or simulation of the state variables, i.e. the "output error" method (Ljung, 1999). However, given the sensitivity of simulation to different model structures (La Cava and Danai, 2015) and the computational cost of numerical integration, the alternative approach of algebraically estimating candidate model outputs is preferred for symbolic regression (Bongard and Lipson, 2007; Schmidt and Lipson, 2009). In the algebraic approach, un-measured states, denoted $\tilde{\mathbf{x}}$, are estimated from measurements via numerical differentiation together with smoothing functions. In the case of first-order differential equations with unmeasured state derivatives, the target is estimated numerically as $y(t_k, \mathbf{u}) = \tilde{\dot{x}}$, such that the prediction error of a candidate model has the form