Identification and Prediction using Symbolic Regression Alpha-beta: Preliminary Results

L. M. Torres-Treviño
Facultad de Ingeniería Mecánica y Eléctrica, Universidad Autónoma de Nuevo León-CIIDIT
Av. Universidad S/N, CP. 64400, San Nicolás de los Garza, Nuevo León, México.
luis.torrestv@uanl.edu.mx

ABSTRACT

A novel approach is proposed for generating equations from measured data of dynamic processes. A composition of unary (alpha) and binary (beta) functions is represented by a real vector and adapted by an evolutionary algorithm to build mathematical equations. The equations can be used for identification and prediction considering a mathematical model with specific number of inputs and outputs. Three cases are used for illustration of the approach where mathematical models and plots of theirs performance are presented with promising results.

Categories and Subject Descriptors

I.28 [**Problem Solving, Control Methods, and Search**]:

General Terms

Algorithms

Keywords

Symbolic regression; process identification; prediction

1. INTRODUCTION

Several processes in industry are dynamic non linear systems and a modeling of such systems in most of the cases it is required for identification and prediction. A model can be useful for analysis, control, optimization, supervision and fault diagnosis. Symbolic regression has been used for identification and prediction, per example, the work of Ly and Lipson [1] propose a multimodal symbolic regression to construct non-linear mathematical model for describing hybrid dynamical systems using data collection. Sharifi and Massoudieh propose an evolutionary data-driven model using a NSGA multiobjective genetic algorithm for discovering models considering a wash-off and building of suspended solids in highway run-off [2]. Chaotic system identification is a complex task and symbolic regression has been used for model generation. An hybrid genetic programming algorithm based on evolution strategies is used for generating a model from Lorenz attractor data [3]. Patelli and Ferariu propose the use of an elite based multiobjective Genetic Programming (GP) for non-linear identification of an industrial plant [5].

Traditionally, non linear autoregressive models and neural networks has been used for non linear dynamic identification; however, hybrids systems using genetic programming variants has been proposed. In the work of Coelho and Pessoa use a GP for selecting the structure for system identification and combines crossover and mutation with orthogonal least squares algorithm to estimate the contribution of the branches of the tree. This GP non linear autoregressive with exogenous inputs modelling is used for identification of a ball and tube system [6].

A similar work made by Xiao-Lei and Bay [7] where a NARX and a non linear autoregressive moving average with exogenous input (NARMAX) polynomial models with a multi-population genetic programming are used for handling complex stochastic nonlinear system identification.Time series prediction using symbolic regression has been used successfully, inclusive with the presence of multiple-time scale dynamics as is shown in the work of Cornforth and Lipson [4]. A summary of non linear model identification can be found in the work of Winkler et al. [8].

The principal problem with symbolic regression for identification and prediction problems is the use of genetic programming that requires in most of the cases special programming languages and procedures. Some authors have proposed alternatives to reduce this problem. Luo and Zhang [9] proposes the use of a Parse-matrix evolution to solve some regression problems. Kotanchek et al. proposes balancing algorithms to sort the data records before the use of a Pareto genetic programming and used for model building and identification of a data set of economic, political, social, geographic data collected [10].

The paper is organized as follows: In section 2, a description of symbolic regression alpha-beta is made including fitness function. In section 3, an illustration of the proposal approach is made considering three cases of identification and prediction are made using experimental data. Results and conclusions will be given in section 4 and 5 respectively.

2. SYMBOLIC REGRESSION α-β

In this work, a new form of symbolic regression is proposed where a core configuration based on simple operations called α β, the selection of operations and parameters is made by

GECCO'14, July 12–16, 2014, Vancouver, BC, Canada.

http://dx.doi.org/10.1145/2598394.2609859.

Table 1: α operator parameters and their related mathematical functions.

α operator	mathematical operation
1	$(k_1x + k_2)$
2	$(k_1x + k_2)^2$
3	$(k_1x + k_2)^{1/2}$
4	$(k_1x + k_2)^{-1}$
5	$\exp(k_1x + k_2)$
6	$\log(k_1x + k_2)$
7	$(k_1x + k_2)^{-2}$
8	$(k_1x + k_2)^3$
9	$(k_1x + k_2)^{-3}$
10	$(k_1x + k_2)^{1/3}$
11	$\sin(k_1x + k_2)$
12	$\cos(k_1x + k_2)$
13	$\tan(k_1x + k_2)$

an evolutionary algorithms. Connectivity and complexity is involved in the search so the use of this form of symbolic regression is very simple [11, 12].

2.1 Operators and core configuration

An α operators is defined as a unary and simple mathematical function that requires only one argument. Considering a review of several mathematical models of real processes, 13 operations are chosen as α operators (see Table 1). An α operator uses two real number parameters called $k1$ and $k2$ and an integer that describes the mathematical operation.

A β operator is defined as a binary function that requires two arguments and makes the four basic arithmetic operations. A maximum of three alpha operators per variable is used, so if there are two variables, a configuration with six alpha operators and five beta operators are defined as a core configuration as is shown in equation (1). Depending de number of input variables a core configuration is used.

$$\begin{aligned} y = \quad & \beta_5(\beta_4(\beta_3(\beta_2(\beta_1(b_1\alpha_1(x_1k_{11} + k_{21}), b_2\alpha_1(x_1k_{12} + k_{22})), \\ & b_3\alpha_1(x_1k_{13} + k_{23})), b_4\alpha_1(x_2k_{14} + k_{24})), \\ & b_5\alpha_1(x_2k_{15} + k_{25})), b_6\alpha_6(x_2k_{16} + k_{26})) \end{aligned} \tag{1}$$

The representation depends of the core configuration used or the number of variables required. The number of variables determine the core configuration as was mention above. A vector with normalized real numbers can be used to represents the connectivity or the number of alpha operators per variable, the α and β operators and the k parameters of the alpha operators. A single real number is used for connectivity because this number is converted to an integer value, then it is converted again in a binary vector and extract the b values using their corresponding position of the elements of the binary vector.

k parameters is assigned directly the real value of the representation vector.

Every α and β operators are integers, so is required the following formulation to get its value from the representation V:

$$\alpha = \lceil V(i) * 13 \rceil \tag{2}$$

$$\beta = \lceil V(i) * 4 \rceil \tag{3}$$

where $\lceil . \rceil$ is the ceiling function. There are 13 α operators defined in table 1 and four β operators (basic algebraic operations)

In this work, Evonorm is used to solve the problem of selection the suitable parameters (k's), the connectivity and integers to define α and β operations.

2.2 Evolutionary algorithm Evonorm

Evonorm is an easy way to implement an estimation of distribution algorithm [13, 14]. As a evolutionary algorithm selection of new individuals and the generation of a new population is used; however, the crossover and mutation mechanism is substituted by an estimation of parameters of a normal distribution function. The following steps are used in Evonorm:

1. Evaluation of a population P.
2. Deterministic selection of individuals from P to PS.
3. Generation of a new population using PS

A population P is a matrix of size I_p (total of individuals) and D_r (total of decision variables). A solution is a set of decision variables and this set is represented as a real vector. Every row of the population P represents a set of decision variables. The selection mechanism is deterministic because the most fittest individuals are selected. Usually the number of selected individuals are lower than the number of the original population, usually a twenty or ten percent of the total population. A random variable with normal distribution is estimated per decision variable, so a marginal distribution function is used. Two parameters are estimated, the mean and the standard deviation, that is determined using the values of the selected individuals. The population of selected individuals is a matrix Ps of size I_s (total of individuals selected) and D_r. The equations (4, 5) are used to calculate the mean and standard deviation considering every vector of the population Ps.

$$\mu_{pr} = \sum_{k=1}^{I_s} (Ps_{pr,k})/I_s \tag{4}$$

$$\sigma_{pr} = \sqrt{(\sum_{k=1}^{I_s} (Ps_{pr,k} - \mu_{pr})^2)/I_s} \tag{5}$$

where $pr = 1..D_r$

A new population is generated using the estimated normal random variables. This is a stochastic process;, however, an heuristic is used to maintain an equilibrium between exploration and exploitation, so new solutions can be found not necessarily near of the mean calculated. The best solution found Ix at the moment is involved in the generation so in the 50% percent of the times the mean is used in the calculations and in the other 50% percent of the time the best solution found Ix is used as a mean as is shown in the following equation:

$$P_{i,pr} = \begin{cases} N(\mu_{pr}, \sigma_{pr}) & U() > 0.5 \\ N(Ix_{pr}, \sigma_{pr}) & otherwise \end{cases} \tag{6}$$

The random variable $U()$ has a uniform distribution function, $N()$ is a random variable with a normal distribution function.

2.3 Fitness function

An evaluation value of every individuals depends on complexity, and accuracy of the model. Complexity is determinate adding all the α and β operators used in the chosen configuration. Mean square error (Equation 7) is used as second objective; however, both complexity and MSRES are weighted to use only one objective function. The criteria is both the minimization of complexity and mean square error.

$$MSRES = \sum_{i=1}^{n} (y(i) - \hat{y}(i))^2 \tag{7}$$

where $y(i)$ is the response of the process and $\hat{y}(i)$ is the predicted response generated by the regression model proposed and n is the total of testing data.

3. IDENTIFICATION AND PREDICTION OF NONLINEAR DYNAMICAL PROCESSES

Identification and prediction of the non-linear dynamic systems requires similar architectures. The number of inputs depends on the number of delay used on input and output signals treating the problem as a regression one. A typical architecture is shown in Figure 1.

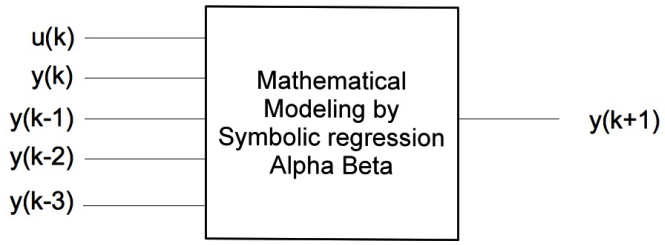

Figure 1: A common architecture used in symbolic regression α β for identification and prediction.

3.1 Cases

Three case can be found in literature, first it is the identification of the systems modelled by (8) and (9) and third it is the prediction on Box-Jenkins furnace.

Process 1

$$y(k+1) = \frac{y(k)y(k-1)[y(k)-0.5]}{1+y^2(k)+y^2(k-1)} + x(k) \tag{8}$$

Process 2.

$$y(k+1) = 0.3y(k) + 0.6y(k-1) + f(x(k)) \tag{9}$$

These two discrete processes have an specific input signal. In (8), $x(k) = \sin(2\pi k/25)$. In (9), $f(x(k)) = 0.6\sin(\pi x(k)) + 0.3\sin(3\pi x(k)) + 0.1\sin(5\pi x(k))$ with its input $x(k) = \sin(2\pi k/200)$. The initial condition in all cases is $y(0) = 0$. For identification of the first two cases a core configuration of three variables is used (10). In third case, a core configuration of four inputs is used (11).

$$y(k+1) = SRAB(u(k), y(k), y(k-1)) \tag{10}$$

$$y(k+1) = SRAB(u(k), y(k), y(k-1), y(k-2)) \tag{11}$$

4. RESULTS

An Evonorm algorithm with 100 individuals where 10 of them are selected during 50 generations. These setting are used in all three cases. A set of building and testing data is generated or recorded where a high percentage of data are used for building the model and the rest are used for testing. Ten runs are used and the solution with lower mean square error was chosen.

4.1 First case

The performance of the model using testing data as is shown in Figure (2). Mathematical model is shown in Equation 12. A total of 5000 data were used for building and 200 for testing. Mean square error on testing data is 0.0007307 considering testing data.

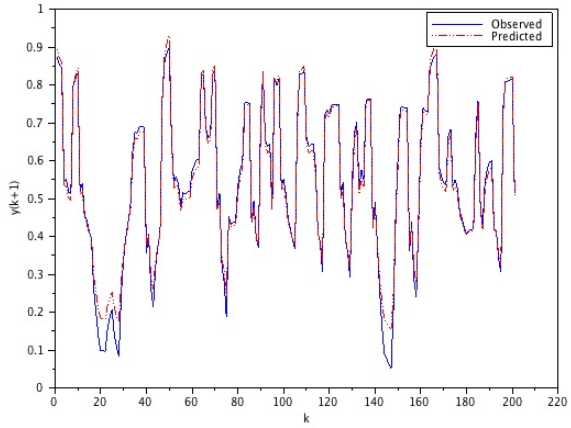

Figure 2: Performance of the best model found versus the real output response.

$$\begin{aligned} y = \; & ((k_{11}u(k) + k_{21}) - (k_{13}u(k) + k_{23})) \\ & (k_{14}y(k) + k_{24})^2 + (k_{15}y(k) + k_{25})^2 - \\ & (k_{18}y(k-1) + k_{28}) + (k_{19}y(k-1) + k_{29}) \end{aligned} \tag{12}$$

where k_{11}=0.9809215, k_{21} = 0.6328762, k_{13}= 0.4296275, k_{23} =0.0722073, k_{14} = 0.3753558, k_{24} = 0.0120735, k_{15} = 0.4772979, k_{25} = 0.5380167, k_{18} = 0.5925161, k_{28} = 0.4077315, k_{19} =0.6232973, and k_{29} = 0.2440728.

4.2 Second case

The performance of the model using testing data as is shown in Figure (3). A total of 3000 data were used for building and 200 for testing. Mathematical model is shown in Equation 13. Mean square error on testing data is 0.0018459 considering testing data.

$$\begin{aligned} y = \; & ((k_{11}u(k) + k_{21})^{1/2} - \exp(k_{14}y(k) + k_{24}) \\ & -(k_{15}y(k) + k_{25})^{-1} + \exp(k_{16}u(k) + k_{26}) \\ & +(k_{17}y(k-1) + k_{27})^{1/2} - (k_{18}y(k-1) + k_{28})^{1/2} \end{aligned} \tag{13}$$

where k_{12} = 0.4232765, k_{22}= 0.9903058, k_{14}= 0.1049508, k_{24}= 0.6088263, k_{15}= 0.3934277, k_{25}= 0.7465172, k_{16} = 0.2367586, k_{26}= 0.7177606, k_{17} = 0.3027999, k_{27}= 0.5093251, k_{19}= 0.1463342, and k_{29} = 0.4300170.

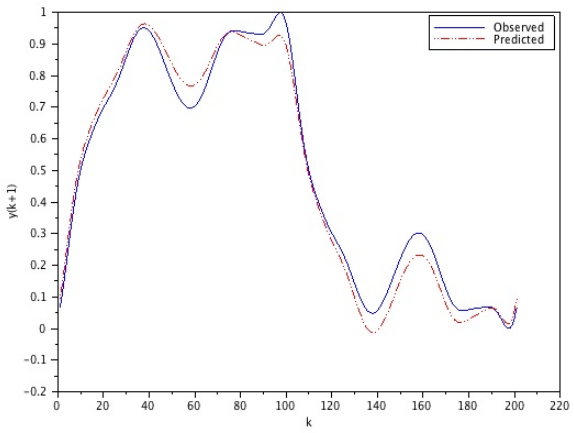

Figure 3: Performance of the best model found versus the real output response.

4.3 Third case

The performance of the model using testing data as is shown in Figure (4). 200 data were used for building 98 for testing. Mathematical model is shown in Equation 14. Mean square error on testing data is 0.0002214 considering testing data.

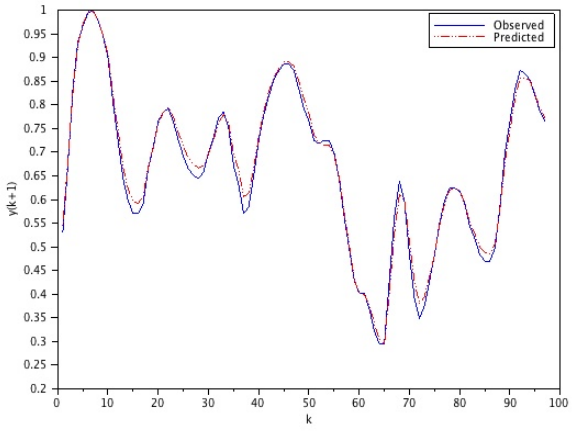

Figure 4: Performance of the best model found versus the real output response.

$$\begin{aligned} y = \quad & -(k_{13}u(k)+k_{23})^2 + (k_{14}y(k)+k_{24}) \\ & -\log(k_{16}y(k)+k_{26})^2 + \log(k_{18}y(k-1)+k_{28}) \\ & -(k_{110}y(k-2)+k_{210}) + (k_{112}y(k-2)+k_{212}) \end{aligned} \tag{14}$$

where k_{13}=0.1070782, k_{23} = 0.2730927, k_{14} = 0.9421514, k_{24} = 0.5979528, k_{16}=0.0842617, k_{26} = 0.0292099, k_{18}= 0.0568625, k_{28} = 0.2968381, k_{110}= 0.4361714, k_{210} = 0.7237257, k_{112}=0.7237689, k_{212} = 0.1376476.

5. CONCLUSION AND FUTURE WORK

These preliminary results illustrate the possibility to use symbolic regression α β for identification and prediction problems. The approach selects the parameters, complexity and connectivity generating feasible mathematical models; however, the core configuration most be given depending of the number of input variables that must be initially give. Improvement considering fitness and complexity requires multi-objective optimization and a weighting mechanism was used for simplicity and generate practical models; however, it is possible to use related algorithms to generate a Pareto front considering different alternatives of models by complexity and connectivity and let the final selection to the user. Finally, a comparison with other approaches will open an opportunity to improve the proposal. All of these stuff is considered as a future work.

6. REFERENCES

[1] Ly, Daniel L.; Lipson, Hod Learning Symbolic Representations of Hybrid Dynamical Systems Journal of machine learning research Volume: 13 (2012) 3585-3618

[2] Zelinka, Ivan; Skanderova, Lenka; Chadli, Mohammed; et al. Evolutionary Identification and Synthesis of Predictive Models. Conference: Nostradamus Conference Location: Ostrava, Czech Republic, SEP, 2012 Sponsor(s): IT4 Innovat; VSB Tech Univ Ostrava; MIR Lab; Ctr Chaos & Complex Networks; Journal Unconvent Comp Nostradamus: Modern methods of prediction, modeling and analysis of nonlinear systems. Advances in Intelligent Systems and Computing. Volume: 192 (2013) 261-272

[3] Brandejsky, Tomas Symbolic regression of deterministic chaos Book Editor(s): Matousek, R. 17th International Conference on Soft Computing Meldel 2011 Location: Brno, Czech Republic Date: JUN 15-17, 2011 Mendel 2011 - 17th International conference on soft computing book Series:: Mendel (2011) 90-93

[4] Cornforth, Theodore W.; Lipson, Hod Symbolic Regression of Multiple-Time-Scale Dynamical Systems Book Editor(s): Soule, T Conference: 14th International Conference on Genetic and Evolutionary Computation Conference (GECCO) Location: Philadelphia, PA JUL 07-11, 2012 Proceedings of the fourteenth international conference on genetic and evolutionary computation conference. DOI: 10.1145/2330163.2330266 (2012) 735-740

[5] Patelli, Alina; Ferariu, Lavinia Elite Based Multiobjective Genetic Programming in Nonlinear Systems Identification Advances in Electrical and Computer Engineering. Volume: 10, Issue:1. DOI: 10.4316/AECE.2010.01017. (2010) 94-99

[6] Coelho, Leandro dos Santos; Pessoa, Marcelo Wicthoff Nonlinear model identification of an experimental ball-and-tube system using a genetic programming approach Mechanical Systems and Signal Processing. Volume: 23, Issue: 5 DOI: 10.1016/j.ymssp.2009.02.005 (2009) 1434-1446

[7] Yuan Xiao-lei; Bai Yan Stochastic Nonlinear System Identification Using Multi-objective Multi-population Parallel Genetic Programming. IEEE Conference: 21st

Chinese Control and Decision Conference Location: Guilin, PEOPLES R CHINA Date: JUN 17-19, 2009 Sponsor(s): NE Univ; IEEE Ind Elect Singapore Chapter; Guilin Univ Elect Technol; IEEE Control Syst Soc; IEEE Ind Elect Soc Source: CCDC 2009: 21ST Chinese Control and Decision Conference, vols 1-6, (2009) 1148-1153.

[8] Winkler, Stephan; Affenzeller, Michael; Wagner, Stefan; et al. Using Genetic Programming in Nonlinear Model Identification Editors: Alberer, D; Hjalmarsson, H; DelRe, L Workshop on Identification for Automotive Systems Location: Johannes Kepler Univ Linz, Linz, AUSTRIA Date: JUL 15-16, 2010 Identification for Automotive Systems Book Series: Lecture Notes in Control and Information Sciences. Volume: 418 (2012) 89-109

[9] Luo, Changtong; Zhang, Shao-Liang Parse-matrix evolution for symbolic regression Engineering Applications of Artificial Intelligence. Volume: 25 Issue: 6 DOI: 10.1016/j.engappai.2012.05.015 (2012) 1182-1193

[10] Brandejsky, Tomas and Matousek, R Symbolic Regression of Deterministic Chaos 17th International Conference on Soft Computing MENDEL 2011 Location: Brno, CZECH REPUBLIC Date: JUN 15-17, 2011 Sponsor(s): B & R Automat CZ Ltd; Humusoft Ltd; AutoCont CZ Ltd Mendel 2011 - 17th International conference on soft computin. Book Series: Mendel (2011) 90-93

[11] Luis M. Torres-Treviño Symbolic Regression Using alpha, beta Operators and Estimation of Distribution Algorithms: Preliminary Results 3rd symbolic regression and modeling workshop for GECCO 2011 July 12-16, 2011, Dublin, Ireland. Distributed on CD-ROM at GECCO-2011.

[12] L. M. Torres-Treviño , I. Escamilla, B. González, R. Praga-Alejo & P. Pérez-Villanueva Modeling cutting machining process using symbolic regression alpha -beta The International Journal of Advanced Manufacturing Technology Noviembre 2012. ISSN 0268-3768 ISSN 0268-3768. DOI 10.1007/s00170-012-4655-5 The International Journal of Advanced Manufacturing Technology: Volume 67, Issue 9 (2013) 2351-2366

[13] L. Torres-T., (2006) Evonorm, a new evolutionary algorithm to continuous optimization, Workshop on Optimization by Building and Using Probabilistic Models (OBUPM) Genetic and Evolutionary Computation Conference (GECCO).

[14] L. Torres-Treviño , Evonorm: Easy and effective implementation of estimation of distribution algorithms,Journal of Research in Computing Science, 23, (2006) 75–83.