Sunspots Modelling: Comparison of GP Approaches

Katya Rodríguez-Vázquez
IIMAS-UNAM
Circuito Escolar, CU
04510 Coyoacán, México
katya.rodriguez@iimas.unam.mx

ABSTRACT
This paper presents a comparative study of the Multi-Branches Genetic Programming (MBGP), GP-NARMAX model approach and Standard Genetic Programming (SGP) for modelling problems. Sunspots data have been considered as study case in order to performance this comparison. The main point is to generate mathematical models in a polynomial form; thus the root node for MBGP has been set as the addition operator. Results show that MBGP rapidly evolves towards good mathematical models which are also easily to translate as well as the GP-NARMAX approach represented in its polynomial form in contrast to SGP.

Categories and Subject Descriptors
I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, Search.

General Terms
Algorithms, Performance, Design.

Keywords
Genetic Programming, Symbolic Regression, Modelling.

1. INTRODUCTION
Symbolic regression problems have traditionally been benchmark tests in the field of Genetic Programming (GP). Diverse approaches have been proposed from standard genetic programming SGP [1] and linear genetic programming [2] to others extensions such as the accelerated genetic programming introduced by Nikolaev and Iba [3] that considers the use of transfer polynomials combined with a recursive least squares algorithm, the robust symbolic regression with affine arithmetic proposed by Pennachin, et al. [4] and GP for symbolic regression using nearest neighbor indexing proposed by Mc Ree [5] amongst others.

This paper then details with the use of the previous proposed Multi-Branches Genetic Programming (MBGP) [6] for sunspots data modeling, and results are compared against standard genetic programming (SGP) and GP-NARXMAX model [7] in terms of the generation of simple models possessing a good performance.

GECCO'13 Companion, July 6–10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

2. GP APPROACHES

2.1 SGP
Standard Genetic Programming (SGP) introduced by Koza [1] is used in order to compare the performance of the MBGP and GP – NARMAX for modeling sunspots data.

2.2 MBGP
Multi-Branches Genetic Programming consists of a root node which is fixed and whose content is also fixed. This is defined as the addition function. This is with the purpose to evolve polynomial models. *N+1* coefficients are also defined corresponding to coefficients of the *N* branches plus the constant term. The *N* branches are randomly created from a defined primitives set as it is the case in SGP. However, the maximum depth or the maximum number of nodes of these branches are much lower than the maximum depth of the individuals in SGP. All branches have also the same maximum depth and their associated coefficients are estimated by means of a Least Square Algorithm. However, coefficients can be also evolved by applying the ephemeral mutation operator used in SGP and in this case, if a coefficient is set to zero, a neutral effect is presented in that particular individual. The MBGP representation is shown in Figure 1. The two main genetic operators are crossover and mutation. Crossover works by randomly selecting a branch in each of the parents and exchanging these branches between them. In the case of mutation, a branch is randomly selected and deleted. This deleted branch is replaced by a new branch randomly generated. It is clear seen than both crossover and mutation operators work in an easier manner in comparison to SGP. This is due to the fact that the whole branch is taken to crossover or mutate and there is any need to randomly select a node and check if the selected sub-branch is syntactically valid.

2.3 GP-NARMAX
The well-known NARMAX (Non-linear AutoRegressive Moving Average with eXogenous inputs) model is an extended ARMAX description for representing non-linear systems. This model is given by a non-linear function of the output y(k), the input u(k) and the possible noise disturbance e(k). Thus,

$$y(k) = F^{\ell}\left(y(k-1),\ldots,y(k-n_y),u(k-d),\ldots,u(k-d-n_u+1),e(k-1),\ldots,e(k-n_e)\right) \quad (1)$$

where n_y, n_u and n_e are the maximum lags considered for the output, input and noise terms, respectively, d is the delay and ℓ is the degree of non-linearity of the model structure. The polynomial NARMAX model is the most common expression which works well in practical applications. Equation (1) can be written in polynomial form as,

$$y(k)=\theta_0+\Bigg[\sum_{i=1}^{n_y}\theta_i y(k-i)+\sum_{i=1}^{n_u}\theta_{n_y+i}u(k-i)+\sum_{i=1}^{n_y}\sum_{j=1}^{n_y}\theta_{i,j}y(k-i)y(k-j)$$

$$+\sum_{i=1}^{n_y}\sum_{j=1}^{n_u}\theta_{i,n_y+j}y(k-i)u(k-j)+\sum_{i=1}^{n_u}\sum_{j=1}^{n_u}\theta_{n_y+i,n_y+j}u(k-i)u(k-j)$$

$$+ \textit{ higher order terms up to degree } l\ \Bigg] \qquad (2)$$

Thus, generation of polynomial NARMAX models by means of GP is performed by defining addition and product of monomials and polynomials. These are defined by the elements of the terminal set (delayed input, output and noise terms) as shown in Figure 2.

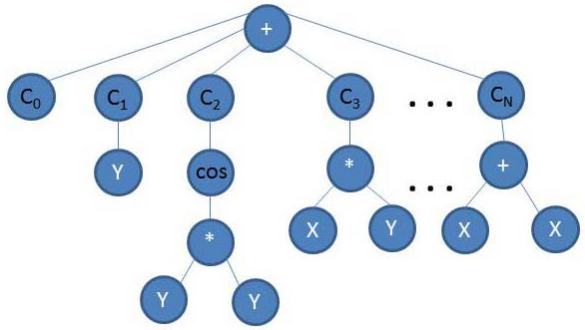

Figure 1. MBGP Encoding.

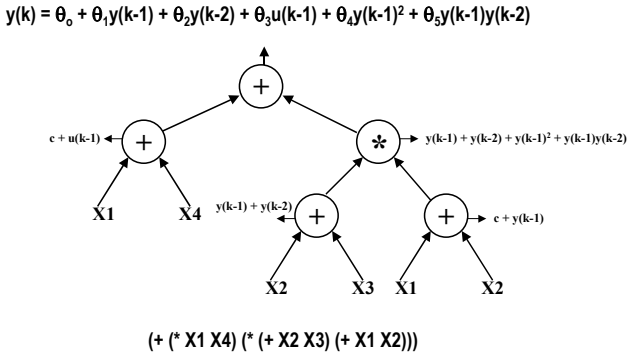

Figure 2. NARMAX-GP encoding (X1=constant term; x2=y(k-1); X3=y(k-2); X4=u(k-1))

3. RESULTS AND DISCUSSIONS

Record data of the annual sunspots from the year 1500 to 2010 was used. In the case of SGP, crossover is defined by the exchanging of selected sub-branches of the parent-individuals. Mutation selects a random node, deletes the associated sub-branch and replaces it by a branch randomly generated.

In the case of MBGP, crossover and mutation operator are defined as in section II. For both approaches, same set of functions are defined and also the same number of nodes are considered. Parameter setting is detailed in Table I. Genetic operators for GP-NARMAX approach were defined as in the case of SGP, but note that maximum number of nodes for GP-NARMAX has been defined as 40; individuals trees of this size can generate complex polynomial models as shown in Table II, where individuals consisting of only 39 nodes correspond to polynomials models of hundreds of terms. This is due to the product of polynomials. In the case of MBGP, it performed better than the SGP producing simple models consisting of only 40 nodes even though a maximum number of branches of 10 of 16 nodes size were defined. A second experiment shown in Table II as MBGP2 was performed. Here, the number of branches increased to 30; however, improvement in performance was not significant. GP-NARMAX showed the best performance, however models of high complexity (hundreds of terms of polynomial) corresponded to these results. Considering a {+, *} function set for SGP and MBGP cases, the mean square errors increase. In order to reduce complexity of NARMAX models, a multi-objective formulation is an alternative, in contrast to MBGP that generates simple models based on a single objective approach.

Table I. SGP and MBGP Parameter Setting

	SGP	***MBGP***	***GP-NARMAX***
Function Set	+, -, *, %, cos, sin, sqrtp, pow2	+, -, *, %, cos, sin, sqrtp, pow2	+, *
Population Size	100	100	100
Number of Branches	N/A	10	N/A
Max.Nodes Branches	N/A	16	N/A
Max. Nodes	160	N/A	40
Pmut	0.05	0.05	0.05
Pxover	0.95	0.95	0.95
Max. Generations	1000	1000	1000

Table II. Comparative Results

	SGP	***MBGP1***	***MBGP2***	***NARMAX***
Avg	17.322	14.567	14.519	13.873
Best	16.434	14.516	14.132	12.806
Nodes	128	41	172	39
Terms		11	31	203

4. REFERENCES

[1] Koza, J.R. (1992) Genetic programming: On the programming of Computers by Means of Natural Selection, MIT Press

[2] Banzhaf, W., P. Nordin, R.E. Keller and F. Francone (1997) Genetic Programming: An Introduction, Morgan Kaufmann.

[3] Nikolaev,N., and Iba,H. (2002). Genetic Programming of Polynomial Models for Financial Forecasting. In: Shu-Heng Chen (Ed.), Genetic Algorithms and Genetic Programming in Computational Finance, Chapter 5, Kluwer Academic Publ., Boston, MA, pp.103-123.

[4] Pennachin, C.L., M. Looks and J.A. Vasconcelos (2010) Robust Symbolic Regression with Affine Arithmetic, Proc. Of the 12th GECCO'10, pp. 917-924.

[5] McRee, R.K (2010) Symbolic Regression Using Nearest Neighbor Indexing, Proc. Of the 12th GECCO'10, pp. 1983-1990.

[6] Oliver-Morales, C. and K. Rodriguez (2004) Symbolic regression Problems by Genetic Programming with Multi-Branches, MICAI 2004 (Monroy et al. Eds.), Springer-Verlag, pp. 717-726.

[7] Rodriguez-Vázquez, K., C.M. Fonseca and P.J. Fleming (2004) Identifying the Structure of Nonlinear Dyamic Systems Using MultiObjective Genetic Programming, IEEE Trans. System, Man and Cybernetics, Part A, 34(4), pp. 531-545.