Automatica 53 (2015) 94–102

Contents lists available at ScienceDirect

Automatica

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

Brief paper

Forward and backward least angle regression for nonlinear system identification☆

Long Zhang, Kang Li

School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT9 5AH, UK

ARTICLE INFO

Article history:
Received 19 May 2013
Received in revised form
17 July 2014
Accepted 1 December 2014

Keywords:
Least angle regression (LAR)
Forward selection
Backward refinement
Nonlinear system identification

ABSTRACT

A forward and backward least angle regression (LAR) algorithm is proposed to construct the nonlinear autoregressive model with exogenous inputs (NARX) that is widely used to describe a large class of nonlinear dynamic systems. The main objective of this paper is to improve model sparsity and generalization performance of the original forward LAR algorithm. This is achieved by introducing a replacement scheme using an additional backward LAR stage. The backward stage replaces insignificant model terms selected by forward LAR with more significant ones, leading to an improved model in terms of the model compactness and performance. A numerical example to construct four types of NARX models, namely polynomials, radial basis function (RBF) networks, neuro fuzzy and wavelet networks, is presented to illustrate the effectiveness of the proposed technique in comparison with some popular methods.

1. Introduction

A large class of nonlinear dynamic systems can be described by a nonlinear autoregressive model with exogenous input (NARX) (Chen, Billing, & Luo, 1989)

$$\begin{aligned} y(t) &= f(y(t-1),\ldots,y(t-n_y),\\ &\quad u(t-1),\ldots,u(t-n_u))+\xi(t)\\ &= f(\mathbf{x}(t))+\xi(t) \end{aligned} \tag{1}$$

where the set $\{u(t), y(t)\}$ represents the real system input and output at time interval t, $t = 1, 2, \ldots, N$, N being the size of the training data set. Their largest input and output lags are n_u and n_y, respectively. $\xi(t)$ denotes the error. The set $\{\mathbf{x}(t), y(t)\}$ is the model input vector and output at time interval t. For simplicity, the model input $\mathbf{x}(t) = [y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u)]$ is rewritten as $\mathbf{x}(t) = [x_1(t), \ldots, x_r(t)]$ with the dimension $r = n_y + n_u$. $f(\cdot)$ is some unknown function.

Constructing such a NARX model involves three steps (Ljung, 1999): (1) model input selection. More specifically, the unknown lags n_y and n_u need to be determined. Statistical tests and regression methods are among the popular approaches (Haber & Unbehauen, 1990; Lind & Ljung, 2005, 2008); (2) choice of mapping function $f(\cdot)$. Polynomials (Billings & Chen, 1989), radial basis function (RBF) networks (Chen, Cowan, & Grant, 1991), neuro fuzzy networks (Harris, Hong, & Gan, 2002; Wang & Mendel, 1992) and wavelet networks (Billings & Wei, 2005; Zhang, 1997) are popular options. Though some suggestions are made on the function selection (Sjöberg et al., 1995), no unified framework is available; (3) parameter identification in function $f(\cdot)$. This requires the specific expression of the model (1). One popular NARX model structure is a linear combination of nonlinear functions whose parameters are given a priori, which is formulated as (Ljung, 1999)

$$y(t) = \sum_{i=1}^{M} p_i(\mathbf{x}(t), \mathbf{v}_i)\theta_i + \xi(t) \tag{2}$$

where p_i is some nonlinear function with pre-fixed nonlinear parameters vector $\mathbf{v}_i$, and θ_i, $i = 1, \ldots, M$, are the linear coefficients to be optimized. The model (2) is also called the linear-in-the-parameters model for pre-fixed nonlinear parameters $\mathbf{v}_i$'s. However, the number of nonlinear functions M is often large, the fixed values for these nonlinear parameters are not optimized, and some nonlinear functions are redundant. This is often referred to as an over-parametrization problem, and not all nonlinear functions are necessarily included into the final model but a good subset is desirable (De Nicolao & Trecate, 1999). Within this context, building a

☆ This work was supported in part by the U.K. Research Councils under Grants EP/G042594/1 and EP/L 001063/1, in part by the Chinese Scholarship Council, in part by the National Natural Science Foundation of China under Grants 61271347 and 61273040, and by the Science and Technology Commission of Shanghai Municipality under grant 11JC1404000. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Antonio Vicino under the direction of Editor Torsten Söderström.

E-mail addresses: zhanglonghit@gmail.com (L. Zhang), k.li@qub.ac.uk (K. Li).

http://dx.doi.org/10.1016/j.automatica.2014.12.010
0005-1098/

linear-in-the-parameters model becomes a model reduction or selection problem. This paper focuses on the model selection issue.

The main objective of model selection is to build a parsimonious model with good generalization performance. Exhaustive search to test all possible subsets is only suitable for a very small number of candidate model terms (nonlinear functions) while it is computationally too demanding when the number is large (Lind & Ljung, 2008; Mao & Billings, 1997). This is known to be an NP-hard problem. To reduce the computational effort, stepwise forward selection methods (Miller, 2002), like forward orthogonal selection (Chen et al., 1991), fast recursive algorithm (Li, Peng, & Irwin, 2005) and orthogonal matching pursuit (Pati, Rezaiifar, & Krishnaprasad, 1993), start from an empty model and add one term at a time until the model performance is satisfied. The alternative is the stepwise backward selection that begins with the full model using all the candidates, and then deletes one term at a time. All these methods are fast, greedy therefore suboptimal (Kump, Bai, Chan, Eichinger, & Li, 2012; Sherstinsky & Picard, 1996). Hence, a parsimonious model with the smallest model size is always desirable.

To improve the model compactness and the generalization performance, the combination of forward selection and backward selection has been proposed in Li, Peng, and Bai (2006) and Zhang, Li, Bai, and Wang (2012), where the backward selection is used to reselect and replace those insignificant terms produced by the forward selection. Alternatively, a number of hybrid methods combining forward selection and backward elimination (instead of backward replacement) have been reported (Haugland, 2007; Soussen, Idier, Brie, & Duan, 2011; Zhang, 2011), where the backward elimination removes insignificant terms. The elimination scheme is also referred to as model pruning. For example, term clustering (Aguirre & Billings, 1995) and simulation error (Farina & Piroddi, 2010; Piroddi & Spinelli, 2003) based pruning methods have been studied for constructing polynomial NARX models.

It is noted that the subset selection may fail in the following scenarios:

- The candidate terms are highly correlated and redundant, which may lead to the ill-conditioning problem (Moussaoui, Brie, & Richard, 2005). The forward selection can avoid selecting highly correlated terms but is not entirely immune to the ill-conditioning problem. The backward selection easily suffers from this problem as it has to deal with the inversion of all the terms at the beginning.
- If the training data is severely polluted by noise, these subset selection methods may fit the models into noise which leads to the over-fitting problem (Chen, Hong, & Harris, 2010; Poggio & Girosi, 1990). The pre-filter and k cross validation are useful to provide a tradeoff between the training accuracy and generalization performance, but additional computations are incurred.
- If the training data does not contain sufficient information, a model with no or low bias but high variance may not have a satisfactory prediction accuracy. A biased model may be more desirable using a good bias/variance trade-off technique (Johansen, 1997; Poggio & Girosi, 1990).
- If small changes in the data can result in a very different model, then the model is less robust and its prediction accuracy is reduced (Tibshirani, 1996).

Given these above considerations, regularization methods are popular techniques to build sparse, robust and biased models by imposing additional penalties or constraints on the solution. A general regularization algorithm is the Tikhonov regression that adds a penalty term to sum squared error (SSE) cost function (Bishop, 1997; Johansen, 1997; Moussaoui et al., 2005), which is given by

$$CF_{Tikhonov} = \sum_{t=1}^{N} \xi^2(t) + \lambda \sum_{i=1}^{M} DF_i \tag{3}$$

where the regularization parameter λ controls the fitting smoothness and the model size. DF_i denotes the function derivatives of different orders. However, this may be computationally too demanding. More recently, the ridge regression and least absolute shrinkage and selection operator (LASSO) use additional l_2 norm and l_1 norm penalties, respectively. The cost function becomes

$$CF_{ridge} = \sum_{t=1}^{N} \xi^2(t) + \lambda \sum_{i=1}^{M} \theta_i^2 \tag{4}$$

and

$$CF_{lasso} = \sum_{t=1}^{N} \xi^2(t) + \lambda \sum_{i=1}^{M} |\theta_i|. \tag{5}$$

These two methods use simplified penalty terms on weights and they aim to minimize the sum of SSE and norms of model weights. Though the ridge regularization can shrink the large weights towards zeros but has little effects on small weights (Kump et al., 2012). Unlike the ridge regression, LASSO has the potential to shrink some weights to exact zeros and can be interpreted as a Bayesian estimator (Tibshirani, 1996). More recently, some modifications have been proposed on the penalty term, such as using the differences between adjacent coefficients (Ohlsson, Ljung, & Boyd, 2010) or differences among all the coefficients (Ohlsson & Ljung, 2013). The difficulty is to mathematically give an explicit solution for the optimal regularization parameter λ. The optimal regularization parameter can be determined by cross validation. Alternatively, it can be estimated by the Bayesian framework under Gaussian prior distributions. Though a number of algorithms have been proposed (Osborne, Presnell, & Turlach, 2000; Rosset & Zhu, 2007; Tibshirani, 1996), most of them are computationally inefficient compared to the forward selection.

As a promising regularization scheme – the least angle regression (LAR) – has been proposed and widely studied (Efron, Hastie, Johnstone, & Tibshirani, 2004). It is a variant of the forward selection as it begins with an empty model with initially no regressor and then selects one term at a time until a stop criterion is satisfied. Unlike the forward selection where the model weights (coefficients) are identical to the least squares solution, the least angle scheme is used to determine the weights. LAR has a few distinctive advantages. First, it is computationally just as fast as the forward selection and more efficient than LASSO methods due to its complete piecewise linear path. Further, it can be easily modified to produce solutions for LASSO estimator. However, the LAR is still a local method and may not produce a sparser model than the forward selection and LASSO methods.

The main objective of this paper is to improve the model sparsity and generalization performance of the LAR algorithm. This is achieved by introducing a replacement scheme using an additional refinement stage. The new method has a forward LAR stage and backward LAR stage. The forward LAR is the same as the original LAR. The backward stage compares the significance of each term in the initial model with the remaining terms in the candidate pool and then replaces insignificant ones, leading to an improved model in terms of compactness and performance. The main difference with our previous work on forward and backward methods is that the least angle scheme rather than least squares approach is employed to determine the model coefficients. Unlike other existing model pruning methods, the proposed method employs the replacement scheme instead of elimination. Further, the LAR is a computationally efficient regularization method without additional computational efforts to determine the regularization parameter. A more detailed difference analysis is given in Section 3. Extensive numerical simulations on the construction of four NARX models, including the polynomial, RBF, neuro fuzzy and wavelet models are presented to demonstrate that the new method is able to produce sparser model than the original LAR algorithm and some alternatives.