Neurocomputing 148 (2015) 551–560

Contents lists available at ScienceDirect

Neurocomputing

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

Elastic net orthogonal forward regression

Xia Hong [a,*], Sheng Chen [b,c]

[a] School of Systems Engineering, University of Reading, UK
[b] School of Electronics and Computer Science, University of Southampton SO17 1BJ, UK
[c] Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 24 March 2014
Received in revised form
4 June 2014
Accepted 4 July 2014
Communicated by K. Li
Available online 12 July 2014

Keywords:
Elastic net
Forward regression
Linear-in-the-parameters model
Regularisation
Leave one out errors
Cross validation

ABSTRACT

An efficient two-level model identification method aiming at maximising a model's generalisation capability is proposed for a large class of linear-in-the-parameters models from the observational data. A new elastic net orthogonal forward regression (ENOFR) algorithm is employed at the lower level to carry out simultaneous model selection and elastic net parameter estimation. The two regularisation parameters in the elastic net are optimised using a particle swarm optimisation (PSO) algorithm at the upper level by minimising the leave one out (LOO) mean square error (LOOMSE). There are two elements of original contributions. Firstly an elastic net cost function is defined and applied based on orthogonal decomposition, which facilitates the automatic model structure selection process with no need of using a predetermined error tolerance to terminate the forward selection process. Secondly it is shown that the LOOMSE based on the resultant ENOFR models can be analytically computed without actually splitting the data set, and the associate computation cost is small due to the ENOFR procedure. Consequently a fully automated procedure is achieved without resort to any other validation data set for iterative model evaluation. Illustrative examples are included to demonstrate the effectiveness of the new approaches.

1. Introduction

A large class of nonlinear models including some types of neural networks can be classified as linear models which include statistically linear or linear-in-the-parameters models [1,2]. These models have provable learning and convergence conditions and are well suited to be used for adaptive learning. They are amenable to parallel implementations, and have clear applications in many engineering applications [3–5]. A basic principle in practical nonlinear data modelling is the parsimonious principle that ensures the smallest possible model for the explanation of the observational data. For linear models, the forward orthogonal least squares (OLS) algorithm efficiently constructs parsimonious models [6,7], and has been a popular tool in associative neural networks such as fuzzy/neurofuzzy systems [8,9] and wavelet neural networks [10,11]. The algorithm has also been utilised in a wide range of engineering applications, e.g. aircraft gas turbine modelling [12], fuzzy control of multi-input multi-output (MIMO) nonlinear systems [13], power system control [14] and fault detection [15].

The main purpose of model construction is to produce good generalisation (capability to approximate system output for new input data that are not used in estimation), through two important aspects in system identification, i.e. choosing parsimonious model structure and deriving robust model parameter estimates for a smooth prediction surface (e.g. parameter control via regularisation). Fundamental to the evaluation of model generalisation capability is the concept of cross-validation (CV) [16], which can be used either in parameter estimation (e.g. tuning regularisation parameter [17,18], forming new parameter estimates [19]), or to derive model selection criteria based on information theoretic principles [20], which regularises model structure in order to produce parsimonious models, since a parsimonious model is favoured by these criteria. The regularisation assisted OLS (ROLS) approaches have been proposed based on minimising the leave one out criteria for regression, classification and probability density estimation [21]. In particular each radial basis function (RBF) unit has a tunable centre vector as well as an adjustable diagonal covariance matrix [21]. Specifically, at each forward regression stage of the model construction procedure one RBF unit's centre vector and diagonal covariance matrix are optimised using a particle swarm (PSO) algorithm. The PSO [22,23] constitutes a population based stochastic optimisation technique, which was inspired by the social behaviour of bird flocks or fish schools. The algorithm commences with random initialisation of a swarm of individuals, referred to as particles, within the specific problem's search space. It then endeavours to find a globally optimum solution by gradually adjusting the trajectory of each particle

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2014.07.008
0925-2312/

towards its own best location and towards the best position of the entire swarm at each optimisation step. The PSO method is popular owing to its simplicity in implementation, ability to rapidly converge to a "reasonably good" solution and to "steer clear" of local minima. It has been successfully applied to a wide range of optimisation problems [24–28].

Regularisation methods are developed to carry out parameter estimation and model structure selection simultaneously [29,30]. It has been shown [31,32] that the parameter regularisation is equivalent to a maximised *a posterior* probability (MAP) estimate of parameters from Bayesian viewpoint by adopting a Gaussian prior for parameters. The regularisation [17,18] uses a penalty function on l^2 norms of the parameters. From the powerful Bayesian learning view point, a regularisation parameter is equivalent to the ratio of the related hyperparameter to the noise parameter, lending to an iterative evidence procedure for solving the optimal regularisation parameters [29,32]

Alternatively the model sparsity can be achieved by minimising the l^1 norm of the parameters. The l^1 norm minimisation is fundamental to the basis pursuit or least absolute shrinkage and selection operator (LASSO) [33,34]. The least angle regression (LAR) procedure [35] is developed for solving the problem efficiently. The Bayesian interpretation for LASSO is simply by adopting an Laplacian prior for parameters. The advantage of LASSO is that it can achieve much sparser models by forcing more parameters to zero, than models derived from the minimisation of the l^p norm, as most l^p norms will produce small, but nonzero, values. Unfortunately introducing nondifferentiable l^1 norm in the cost function brings difficulties of model parameter estimation and finding an appropriate l^1 regulariser.

Another disadvantage of using l^1 optimisation is that a group of correlated terms cannot be selected together, which is not desirable for the sake of interpretability of the model in some applications. On the other hand, the use of l^2 will improve model generalisation, but cannot be used for model selection by itself. Combining a locally regularised orthogonal least squares (LROLS) model selection [36] with D-optimality experimental design enhances model robustness [31].

Recently a promising concept of the elastic net (EN) has been proposed by minimising the l^1 and l^2 norms of the parameters together [30]. The EN keeps the model sparsity of LASSO, while strongly correlated terms tend to be in or out of the model together. It is shown that the elastic net problem can be transformed into an equivalent LASSO problem on an augmented data, based on which the LAR procedure is applicable, referred to as LARS-EN [30]. Note that because there are two regularisation parameters in the elastic net, the cross validation has to be performed over a two-dimensional space. The tenfold cross validation was used in the choosing two regularisation parameters by searching over a grid of l^2 norm regularisation parameter values. Then for each setting of the l^2 norm regularisation parameter, the algorithm LARS-EN produces the entire solution path of the elastic net, which is used to select l^1 norm regularisation parameter by tenfold CV. Clearly this may not yield the optimal parameters if the grid search is set at a coarse level, but increasing the grid search at a very fine level would inevitably increase the computational cost. It would be desirable that the two regularisation parameters can be optimised simultaneously based on cross validation as well as in an efficient manner.

In this paper we propose an efficient model identification method aiming at maximising a model's generalisation capability. The paper contains two elements of novel contribution. Firstly an elastic net cost function is defined and applied based on orthogonal decomposition, which facilitates the automatic model structure selection process with no need of using a predetermined error tolerance to terminate the forward selection process. Secondly an original derivation of analytical evaluation of LOOMSE is presented based on the resultant ENOFR models without actually splitting the data set. Consequently a fully automated procedure is achieved without resort to any other validation data set for iterative model evaluation. The algorithm has a two level structure. At the upper level, the two regularisation parameters in the elastic net are optimised using PSO by minimising the LOOMSE. At the lower level are the simultaneous model selection and elastic net parameter estimation. Illustrative examples are included to demonstrate the effectiveness of the new approaches.

2. Preliminaries

Consider the general nonlinear system represented by the nonlinear model [37]:

$$y(k)=f(\mathbf{x}(k))+e(k), \tag{1}$$

where $\mathbf{x}(k)\in\Re^m$ denotes the system input vector and $y(k)$ is the system output variable, respectively. $e(k)$ is the system white noise and $f(\bullet)$ is the unknown system mapping. The system model (1) is to be identified from an observation data set $D_N=\{\mathbf{x}(k),y(k)\}_{k=1}^N$ using some suitable functional which can approximate $f(\bullet)$ with arbitrary accuracy. One class of such functionals is the kernel regression model of the form:

$$y(k)=\hat{y}(k)+e(k)=\sum_{i=1}^{n_M}\theta_i\phi_i(\mathbf{x}(k))+e(k), \tag{2}$$

where $\hat{y}(k)$ denotes the model output, θ_i are the model weights, $\phi_i(\mathbf{x}(k))$ are the regressors, and n_M is the total number of candidate regressors or model terms.

By letting $\boldsymbol{\phi}_i=[\phi_i(\mathbf{x}(1))\cdots\phi_i(\mathbf{x}(N))]^T$, for $1\le i\le n_M$, and defining

$$\mathbf{y}=\begin{bmatrix}y(1)\\ \vdots\\ y(N)\end{bmatrix},\quad \boldsymbol{\Phi}=[\boldsymbol{\phi}_1\cdots\boldsymbol{\phi}_{n_M}],$$

$$\boldsymbol{\theta}=\begin{bmatrix}\theta_1\\ \vdots\\ \theta_{n_M}\end{bmatrix},\quad \mathbf{e}=\begin{bmatrix}e(1)\\ \vdots\\ e(N)\end{bmatrix}, \tag{3}$$

the regression model (2) can be written in the matrix form

$$\mathbf{y}=\boldsymbol{\Phi}\boldsymbol{\theta}+\mathbf{e}. \tag{4}$$

Let an orthogonal decomposition of the matrix $\boldsymbol{\Phi}$ be

$$\boldsymbol{\Phi}=\mathbf{WA}, \tag{5}$$

where

$$\mathbf{A}=\begin{bmatrix}1 & a_{1,2} & \cdots & a_{1,n_M}\\ 0 & 1 & \ddots & \vdots\\ \vdots & \ddots & \ddots & a_{n_M-1,n_M}\\ 0 & \cdots & 0 & 1\end{bmatrix} \tag{6}$$

and

$$\mathbf{W}=[\mathbf{w}_1\ldots\mathbf{w}_{n_M}] \tag{7}$$

with columns satisfying $\mathbf{w}_i^T\mathbf{w}_j=0$, if $i\ne j$. The regression model (4) can alternatively be expressed as

$$\mathbf{y}=\mathbf{W}\mathbf{g}+\mathbf{e}, \tag{8}$$

where the orthogonal weight vector $\mathbf{g}=[g_1\cdots g_{n_M}]^T$ satisfy the triangular system $\mathbf{A}\boldsymbol{\theta}=\mathbf{g}$, which can be used to determine model parameters $\boldsymbol{\theta}$, given $\mathbf{A}$ and $\mathbf{g}$.