Neurocomputing 189 (2016) 135–144

Contents lists available at ScienceDirect

Neurocomputing

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

Time series wind power forecasting based on variant Gaussian Process and TLBO

Juan Yan [a], Kang Li [a,*], Erwei Bai [a,b], Zhile Yang [a], Aoife Foley [c]

[a] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT7 1NN, UK
[b] Department of Electrical and Computer Engineering, University of Iowa, Iowa, IA 52242, USA
[c] School of Mechanical and Aerospace Engineering, Queen's University Belfast, BT7 1NN, UK

ARTICLE INFO

Article history:
Received 10 November 2014
Received in revised form
5 November 2015
Accepted 25 December 2015
Communicated by Ling Wang
Available online 7 January 2016

Keywords:
Gaussian Process
Model consistency
TLBO
Wind power forecasting

ABSTRACT

Due to the variability and stochastic nature of wind power, accurate wind power forecasting plays an important role in developing reliable and economic power system operation and control strategies. As wind variability is stochastic, Gaussian Process regression has recently been introduced to capture the randomness of wind energy. However, the disadvantages of Gaussian Process regression include its computation complexity and incapability to adapt to time varying time-series systems. A variant Gaussian Process for time series forecasting is introduced in this study to address these issues. This new method is shown to be capable of reducing computational complexity and increasing prediction accuracy. The convergence of the forecasting results is also proved. Further, a teaching learning based optimization (TLBO) method is used to train the model and to accelerate the learning rate. The proposed modelling and optimization method is applied to forecast both the wind power generation in Ireland and that from a single wind farm to demonstrate the effectiveness of the proposed method.

1. Introduction

As power systems in many countries and regions are penetrated with increasing wind power, it is imperative to forecast wind power generation accurately in advance for reliable and effective power system operation and control. Currently, wind energy time series forecasting has shown to be an effective technique for short term forecasting [1]. Unlike the numerical weather prediction (NWP) methods, which employ weather information such as temperature, wind speed and wind direction, time series models employ historical measurement data solely, to make short term predictions from several minutes to several hours ahead, which could be very useful for short term load balancing and energy storage decisions [2]. Although such forecasting horizon is relatively short in comparison with NWP, time series methods have demonstrated their great computation efficiency. Existing time series forecasting techniques include traditional methods such as ARMA [3], Persistence, Neural Network [4–6] and Neural-fuzzy [7]. Besides, methods such as Kalman filters [8] and Gaussian Process (GP) [9–11] have also been recently introduced.

Although the GP approach was first used in the statistics community in 1964 [12] and applied to curve fitting in 1978 [13], this stochastic process did not attract much attention until a comparison between GP and other well known methods was carried out by Rasmussen in 1996 [14]. Since then, the implementing and application of GP have been further researched and extended. Initially the GP learning process was studied and simplified [15], then GP was applied to system regression [16,17] and classification [18,19]. GP is a global non-parametric method that assumes that all the variables follow one joint Gaussian distribution and all the available data are employed in the prediction procedure. It is a special case of Bayesian inference, where all the priori are assigned to be Gaussian. Moreover, GP is viewed as similar to kernel estimators, because a covariance function is used to describe the correspondence between two outputs. The main difference between GP and kernel estimators lies in that the sum of the weights does not have to be unity. Alternatively, GP can be viewed as a form of basis function approach, differing from those normal ones due to the flexible coefficients [20]. One significant advantage of GP is that besides giving the most probable estimation, GP describes the distribution of the new prediction which can be quite beneficial in developing model based control strategies in industry. Secondly, the global property of GP guarantees robust estimation even when the number of available data is limited or imbalanced. Moreover, its covariance function contains less

* Corresponding author.
E-mail address: k.li@qub.ac.uk (K. Li).

http://dx.doi.org/10.1016/j.neucom.2015.12.081
0925-2312/

hyperparameters in comparison with other advanced machine learning methods such as Neural Network and Fuzzy Logic, and thus avoids the curse of dimensionality as the dimension of input increases.

Due to these advantages, GP regression began to be applied in a variety of fields, from multi-sensor networks [21,22] to image processing [23–25], from semiconductor industrial process [26] to medical health [27] and biological observation [28]. However, drawbacks still exist. First, in GP all available data is assumed to follow one joint Gaussian distribution and employed further to make new predictions. Such mechanisms generate expensive computational demand caused by the matrix inversion in GP modelling. Especially a large number of irrelevant data are used which causes unnecessary computational burden. Secondly, its ability to reflect the local property of a system remains to be an issue. To tackle these problems, sparse approximation techniques for full GP [29,30] and methods of local GP mixtures [31] have been proposed recently. Moreover, variants of GP appeared. In [32], a temporally local GP (TLGP) method has been developed and applied in forecasting wind power of the whole Ireland. Further, deeper study on evaluating the forecasting accuracy and the residual distribution of TLGP has been implemented in [33]. In this paper, TLGP for time series systems is presented first and then the model consistency is proved using a test theory. Moreover, the optimization techniques are investigated and a new optimization technique, namely the teaching-learning based optimization (TLBO) is used, to overcome the limitations of some conventional optimization techniques such as linear programming and quadratic programming [34–38]. The proposed TLGP-TLBO method shows great accuracy and convergence property in the real application to both the all-island wind generation and a small farm output.

This paper is organized as follows. First, the standard GP for time series wind power forecasting is presented in Section 2. In Section 3, the variant efficient GP is presented in detail, together with the computational analysis and the model convergence proof. Following the model description, the learning and teaching procedure of TLBO is introduced in Section 4. In Section 5, the wind power generation for the whole island of Ireland and from a small farm on it are used as case studies to confirm the effectiveness of the proposed method. Finally, Section 6 concludes the paper.

2. Gaussian Process for time series forecasting

2.1. Standard Gaussian Process

For a multiple-input–single-output (MISO) nonlinear system, let (X,Y) denote a set of input–output data $\mathcal{D}$, and suppose the kth ($k \in [1,N]$) sample $(\boldsymbol{x}(k), y(k))$ satisfies Eq. (1), where v is an i.i.d. random sequence of white noise with zero mean and finite variance $\sigma_v{}^2$, which in the case of wind power forecasting refers to wind power measurement noise. Here $\boldsymbol{x} \in R^D$, which is the input vector:

$$y(k) = f(\boldsymbol{x}(k)) + v(k) \tag{1}$$

GP is a stochastic process where an indexed collection of random variables follow joint Gaussian distribution [39]. Generally, the mean function could be assumed to be zero if the data are properly scaled and de-trended [40] as shown in the following equation:

$$P(Y \mid C_Y, X) = \mathcal{N}(0, C_Y) \tag{2}$$

For a given new output $y_0 = f(\boldsymbol{x}_0)$, if it follows joint Gaussian distribution with the available data Y in (2), then the joint distribution could be written as (3) in a partitioned form where A, B, C is shown in (4). Here, $cov(a,b)$ denotes the covariance between two variables a and b, and its value is decided by the so-called covariance function:

$$\begin{bmatrix} y_0 \\ Y \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} A & B \\ B^\top & C_Y \end{bmatrix}\right) \tag{3}$$

$$\begin{aligned} &A = cov(y_0, y_0) \\ &B(i) = cov(y_0, y(i)), \quad y(i) \in Y \quad \text{and} \quad i \in (1,N) \\ &C_Y(i,j) = cov(y(i), y(j)), \quad y(i), y(j) \in Y \;\; i,j \in (1,N) \end{aligned} \tag{4}$$

There exist many forms of covariance functions [14]. The square exponential function shown in (5) is one of the most popular ones due to its infinite differentiability:

$$cov(y(i), y(j)) = \Phi(\boldsymbol{x}(i), \boldsymbol{x}(j)) = s \cdot \exp\left[-\frac{1}{2}\sum_{d=1}^{D} \omega_d (x_d(i) - x_d(j))^2\right] + v \cdot \delta_{ij} \tag{5}$$

Here, D refers to the dimension of model input $\boldsymbol{x}$ and δ_{ij} refers to the Kronecker delta representing the observation noise for each sample, and Φ represents the covariance function. The hyperparameters involved could be denoted as $\theta = [s, v, \omega_1, \ldots, \omega_D]$:

$$\begin{aligned} \ln P(Y \mid X, \theta^*) &= \ln\left[\frac{1}{(2\pi)^{N/2} \mid C_Y \mid^{1/2}} \exp\left(-\frac{1}{2} Y^\top C_Y^{-1} Y\right)\right] \\ &= -\frac{1}{2} Y^\top C_Y^{-1} Y - \frac{1}{2} \ln \mid C_Y \mid - \frac{N}{2} \ln 2\pi \end{aligned} \tag{6}$$

Standard GP employs the gradient based methods to optimize the marginal likelihood function, due to its fast convergence rate and satisfactory accuracy. The gradient form of the log marginal density is shown in (7), where θ_j^* represents the jth element of the hyperparameter vector:

$$\frac{\partial P(Y \mid X, \theta^*)}{\partial \theta_j^*} = \frac{1}{2} Y^\top C_Y^{-1} \frac{\partial C_Y}{\partial \theta_j^*} C_Y^{-1} Y - \frac{1}{2} \mathrm{Tr}\left[C_Y^{-1} \frac{\partial C_Y}{\partial \theta_j^*}\right] \tag{7}$$

2.2. Time series wind power forecasting

Time series prediction is an effective way for short term wind power forecasting. It employs only the historical measurement data and neglects the potential exogenous inputs, thus a time series system could be expressed in (8) where L represents the time lag:

$$y(t) = f(y(t-1), y(t-2), \ldots, y(t-L)) + v(t) \tag{8}$$

Denote $\boldsymbol{x}(t) = [y(t-1), y(t-2), \ldots, y(t-L)]^\top$, which represents the state vector at time t. In this case, L in equivalent to D in the previous section. Given a sequence of data Y as training data, for time instant t, the output could be predicted with (9) where $B(t)$ describes the covariance between $y(t)$ and Y, and C_Y denotes the self-covariance of data sequence Y:

$$\hat{y}(t) = B(t) C_Y^{-1} Y \tag{9}$$

$$B(t) = (\Phi(\boldsymbol{x}(t), \boldsymbol{x}(1)), \Phi(\boldsymbol{x}(t), \boldsymbol{x}(2)), \ldots, \Phi(\boldsymbol{x}(t), \boldsymbol{x}(N))) \tag{10}$$

$$C_Y(i,j) = \Phi(\boldsymbol{x}(i), \boldsymbol{x}(j)) \;\; i,j \in [1,N] \tag{11}$$

Here N represents the dimension of Y. As Y describes a sequence of data in time series, the elements of Y could be sampled long while ago. When training the model to identify the hyper-parameters using (6) and making new predictions using (9), it can be seen that the computation complexity is $O(N^3)$. So it could be quite computationally expensive when large training data is used.