Energy and Buildings 127 (2016) 571–579

Contents lists available at ScienceDirect

Energy and Buildings

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

Daily natural gas consumption forecasting based on a structure-calibrated support vector regression approach

Yun Bai*, Chuan Li

National Research Base of Intelligent Manufacturing Service, Chongqing Technology and Business University, Chongqing 400067, China

ARTICLE INFO

Article history:
Received 16 January 2016
Received in revised form 3 June 2016
Accepted 6 June 2016
Available online 7 June 2016

Keywords:
Natural gas
Consumption
Forecast
Extended Kalman filter
Structure-calibrated support vector regression

ABSTRACT

An accurate forecast of natural gas (NG) consumption is of vital importance for economical and reliable operation of the distributive NG networks. In this paper, a structure-calibrated support vector regression (SC-SVR) approach is proposed to forecast the daily NG consumption, which is correlated with the past time series using the SVR model. To better accommodate the dynamic nature of the NG consumption, the structural parameters of the SVR model are online calibrated in response to the receding horizon of the NG consumption series. The calibration of the structural parameters for the next-day forecast is performed by extended Kalman filter. The proposed SC-SVR approach is evaluated using real data collected from a NG company in the period from January to December 2012. The results indicate that the mean absolute percentage error and the root mean squared error are 2.36% and 3913.88 m^3/d, respectively. To show the applicability and superiority of the SC-SVR approach, two peer methods, i.e., least squares SVR model and dynamic back propagation neural network are also employed for comparison. The results show that, thanks to nonlinear mapping capability of the SVR and dynamic nature of the online calibration for the model structure, the proposed SC-SVR method is capable of improving the forecast accuracy for the daily NG consumption.

1. Introduction

As one of the most important energy sources in the world, natural gas (NG) is regarded as a clean and safe fuel. In view of its advantages, the NG has been widely used in such applications as power generation, transportation, aviation, chemical engineering, and residential use. Hence the NG has become one of the essential factors for social and economic developments in recent years.

In the NG supply systems, the main objective is efficient and reliable distribution from the source to the client, which often needs forecast future demands in early time. Therefore, accurate forecast of the NG consumption is of vital importance for economical and reliable operation of the distributive NG networks. Different methods have been developed to fill in gaps in this field. Literatures of the NG consumption forecasting have been reviewed by Soldo [1]. In the published papers, several different time horizons of the NG consumption were investigated. For annual basis, Erdogdu [2] used autoregressive integrated moving average (ARIMA) model, Kumar and Jain [3] applied grey model, Azadeh et al. [4] employed an adaptive network-based fuzzy inference system-stochastic frontier analysis approach, Bianco et al. [5] proposed a scenario analysis method, Goncu [6] established a logistic-based approach, and Wu et al. [7] proposed a new grey model to forecast yearly NG consumption. For monthly basis, Kizilaslan and Karlik [8] used a hybrid neural network method, and Karimi and Dastranj [9] applied artificial neural networks to estimate the monthly NG consumption. For daily basis, Zhou et al. [10] proposed a new OIHF-Elman network involving factors such as weather, temperature and daily data type, Farahat and Talaat [11] used curve fitting approach based on genetic algorithm and Jain et al. [12] built a sensor-based forecasting model using support vector regression (SVR) for forecasting daily residential NG consumption. For hourly basis, Dombayci [13] applied neural networks to estimate heating energy consumption.

Much attention has been paid in the daily NG consumption forecasting. Lee and Tong [14] found that the forecasting accuracy of Box-Jenkins model (e.g., ARIMA) was poor for the small-sample or nonlinear data. This limits its applications, and thus calling for nonlinear approaches to deal with real forecasting cases. Artificial neural network (ANN) is one of significant nonlinear models, which has been examined by diverse fields including the NG consumption forecasts. The intrinsic drawbacks of the ANN, e.g., over-fitting, convergence to local minimum and slow learning performance, result in difficulties for dealing with the complex NG consumption series [15]. However, the ANN still exhibits its good forecasting performance. Demirel et al. [16], Taşpınar et al. [17], and Soldo et al. [18] used the ANN to forecast the daily NG consumption. To improve

* Corresponding author.
E-mail address: yunbai@foxmail.com (Y. Bai).

http://dx.doi.org/10.1016/j.enbuild.2016.06.020
0378-7788/

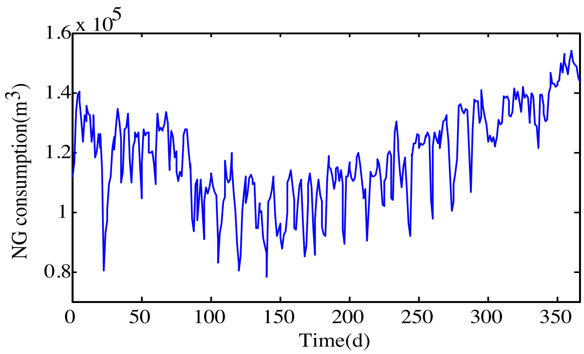

Fig. 1. NG consumption dataset from January to December 2012.

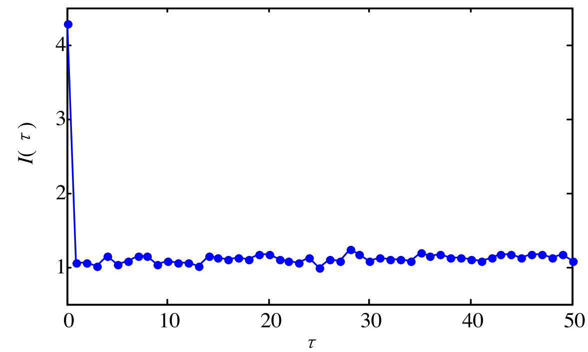

Fig. 2. The change of $I(\tau)$ over τ.

the forecasting performance of the ANN, other techniques such as fuzzy logic [19,20], particle swarm optimization [21,22], genetic algorithms [23,24] and wavelet transform [25,26] were introduced in literatures. Other nonlinear models, summarized in literature [27], are also applied in daily NG forecasts successfully.

Some of the aforementioned methods are established on the basis of the NG consumption series with other information (additional meteorological and logical data), which increases the difficulty of the data collection. To solve this problem, the authors build here a data-driven model using only the time series of the NG consumption. In addition, considering the dynamic characteristics of the real time series, soft sensor technology is introduced to calibrate the model structures in time.

The soft sensor is an inferential estimator, drawing conclusions from process observations when hardware sensors are unavailable or unsuitable [28]. With the use of the soft sensor, the objective variables can be estimated with a high degree of accuracy [29]. To make the soft sensor models adaptive to the dynamic situations, researchers have developed online calibration procedure by taking interference, time-variance nature and nonlinearity into consideration. Inspired by the online calibration idea of the soft sensors [30], a dynamic forecasting model is proposed for the daily NG consumption forecasting. Considering the nonlinear mapping capability, the least squares SVR (LSSVR) is suggested to associate the next-day NG consumption with the past time series. To track the dynamic characteristic of the real NG consumption over the time, a model structure calibration function is attached to the forecasting model. The online calibration for the model structure is realized using extended Kalman filter (EKF). In this way, a dynamic forecasting model, structure-calibrated LSSVR (SC-SVR), is constructed to improve the forecasting accuracy for the daily NG consumption.

2. Dataset and performance criterion

In the first subsection, the data of the daily NG consumptions are presented, the input-output structures are determined in the subsection 2.2, and the subsection 2.3 gives the evaluation criteria.

2.1. Data collection

The data of the NG consumption of Anqing, China are recorded as the dataset for modeling and forecasting. The city of Anqing is the former provincial capital of Anhui province, China. It locates at the southwest of Anhui province, and the north of the Yangtze River (115°46′ ∼ 117°44′E/29°47′ ∼ 31°17′N). In this study, 366 daily records of the NG consumptions from January to December 2012 are collected and plotted in Fig. 1. According to the gas company's statistics, the information of these data contain both residential and commercial consumptions, accounting for around 18% and 82% in total daily NG consumption, respectively.

The NG series are divided into two sets, the training set and the testing set. Usually the training set contains 70–90% of all data and remained data are used for testing [31]. Following this data segmentation principle, the first 300 records as shown in Fig. 1 are employed for training, while the rests for testing.

2.2. Input and output data determination

As mentioned in the Introduction section, the data-driven model has become an appropriate alternative to knowledge-driven model in the time series forecasting due to the minimum information requirements (less input variables) without taking into account the potential physical processes [32,33]. Therefore, the i-th day NG consumption can be associated with the historical data, i.e.,

$$\hat{G}(i) = F(G(i-\tau_{opt}), G(i-2\tau_{opt}), \ldots\ldots, G(i-m_{opt}\tau_{opt})), \tag{1}$$

where τ_{opt} denotes the optimal delay time (subscript "opt" indicates the optimal value), m_{opt} is the optimal embedding dimension, $F(.)$ stands for the correlation pattern between the future and the historical values, and $G(i)$ and $G(i)$ represent the real observation and the forecasted values at time i, respectively.

As shown in the above regression model (Eq. (1)), there are two input variables (τ and m) to be specified for the application. The choice of the delay time τ of each observation series may affect the quality of the input series of the regression model – a too great τ will result in the input series distortion, while a too small τ will be insufficient to reveal the input series information. Meanwhile, the appropriate embedding dimension m can not only ensure accurate calculation of space variables, but minimize the effects of amount of calculation and system noise as well. In this paper, mutual information (MI) method and false nearest neighbor (FNN) method are applied to determine τ and m, respectively.

As the measure of the dependence between two series, according to the theory of Shannon entropy, the MI method can be used to calculate the nonlinear correlation between $\mathbf{G} = \{G(i)\}$ and $\mathbf{G}_\tau = \{G(i-\tau)\}$. The mutual information variable $I(\tau)$ with different τ is described as

$$I(\tau) = H(\mathbf{G}) + H(\mathbf{G}_\tau) - H(\mathbf{G}, \mathbf{G}_\tau), \tag{2}$$

where $H(\mathbf{G})$ and $H(\mathbf{G}_\tau)$ are the marginal entropies, and $H(\mathbf{G}, \mathbf{G}_\tau)$ is the joint entropy of $\mathbf{G}$ and $\mathbf{G}_\tau$. The three variables are respectively given by

$$H(\mathbf{G}) = -\sum_{i=1}^{N} P(G(i)) \ln P(G(i)),\ \mathrm{H}(\mathbf{G}_\tau) = -\sum_{i=1}^{N} P(G(i-\tau))$$

$$\ln P(G(i-\tau)),\ \mathrm{and} \mathrm{H}(\mathbf{G}, \mathbf{G}_\tau) = -\sum_{i=1}^{N} P(G(i), G(i-\tau))$$

$$\ln P(G(i), G(i-\tau)), \tag{3}$$