Solar Energy 163 (2018) 189–199

Contents lists available at ScienceDirect

Solar Energy

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

A decomposition-clustering-ensemble learning approach for solar radiation forecasting

Shaolong Sun[a,b,c], Shouyang Wang[a,b,d,*], Guowei Zhang[a,b], Jiali Zheng[a,b]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China
[c] Department of Systems Engineering and Engineering Management, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
[d] Center for Forecasting Science, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Keywords:
Solar radiation forecasting
Decomposition-clustering-ensemble learning approach
Ensemble empirical mode decomposition
Least square support vector regression

ABSTRACT

A decomposition-clustering-ensemble (DCE) learning approach is proposed for solar radiation forecasting in this paper. In the proposed DCE learning approach, (1) ensemble empirical mode decomposition (EEMD) is used to decompose the original solar radiation data into several intrinsic mode functions (IMFs) and a residual component; (2) least square support vector regression (LSSVR) is performed to forecast IMFs and residual component respectively with parameters optimized by gravitational search algorithm (GSA); (3) Kmeans method is adopted to cluster all component forecasting results; (4) another GSA-LSSVR method is applied to ensemble the component forecasts of each cluster and the final forecasting results are obtained by means of corresponding cluster's ensemble weights. To verify the performance of the proposed DCE learning approach, solar radiation data in Beijing is introduced for empirical analysis. The results of out-of-sample forecasting power show that the DCE learning approach produces smaller NRMSE, MAPE and better directional forecasts than all other benchmark models, reaching up to accuracy rate of 2.96%, 2.83% and 88.24% respectively in the one-day-ahead forecasting. This indicates that the proposed DCE learning approach is a relatively promising framework for forecasting solar radiation by means of level accuracy, directional accuracy and robustness.

1. Introduction

With the continuous consumption of fossil fuels, energy and environmental problems become increasingly severe. It is urgent to find a solution to solve energy and environmental problems and achieve sustainable development. Therefore, renewable energy has attracted much attention all over the world and been rapidly developed in recent years.

The solar energy is considered to be one of the cleanest and most promising renewable energy sources, which makes solar power an important direction of exploring renewable energy. The solar power generation is widely used in developed countries and considerable developing countries at present, and has partially replaced the traditional power generation. In recent years, China has gained more than 25% growth rate every year in the development and utilization of renewable energy. The solar radiation plays an important role in solar photovoltaic power generation. With the development of solar technologies, demands for solar radiation data with high accuracy are increasing. Consequently, the solar radiation forecasting has become one of core contents of solar photovoltaic power generation.

The scholars and researchers so far have conducted in-depth research for solar radiation forecasting theory and put forward a number of feasible and efficient methods to predict actual solar radiation intensity, which has achieved satisfactory results. These methods can be divided into three categories: traditional mathematical statistics, numerical weather forecasting and machine learning. The traditional mathematical statistics includes: regression analysis (Trapero et al., 2015), time series analysis (Huang et al., 2013; Voyant et al., 2013), gray theory (Fidan et al., 2014), fuzzy theory (Chen et al., 2013), wavelet analysis (Mellit et al., 2006; Monjoly et al., 2017) and Kalman filter (Akarslan et al., 2014), etc.; numerical weather forecasting solves thermal fluid dynamics equations with weather evolution by computers with high performance to predict the solar radiation intensity of the next period mainly based on the actual situation of the atmosphere (Chow et al., 2011; Mathiesen and Kleissl, 2011; Mathiesen et al., 2013), which is complicated and time-consuming. With the rise of big data mining, machine learning techniques currently have attracted much attention, for example, artificial neural networks (ANN) (Amrouche and Le Pivert, 2014; Benmouiza and Cheknane, 2013; Niu et al., 2015; Paoli et al., 2010), support vector machines (SVM) (Gala

* Corresponding author at: Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.
E-mail address: sywang@amss.ac.cn (S. Wang).

https://doi.org/10.1016/j.solener.2018.02.006
Received 28 July 2017; Received in revised form 26 December 2017; Accepted 1 February 2018

et al., 2016; Lauret et al., 2015) and heuristic intelligent optimization algorithms (Jiang et al., 2015; Niu et al., 2017; Niu et al., 2016a; Wang et al., 2015) have been widely used in solar radiation forecasting.

According to the literature above, Trapero et al. (2015) applied dynamic harmonic regression model (DHR) to predict short-term direct solar radiation and scattered solar radiation in Spain for the first time in 2015. Huang et al. (2013) took autoregressive model to forecast the solar radiation in the framework of meteorological factors-dynamic adjustment system in 2013, which increased the accuracy by 30% than general neural network or random models. Fidan et al. (2014) predicted hourly solar radiation in Izmir, Turkey by integration of Fourier transform and neural network. Mellit et al. (2006) applied infinite impulse response function to filter time series of the total solar radiation in Algeria from 1981 to 2000, and then substituted the filtered data into adaptive wavelet neural network model to forecast the total solar radiation in 2001, where the error percentage is less than 6% performing better than conventional neural network models and classical statistical methods. However, Akarslan et al. (2014) first used the multidimensional linear predictive filtering model to forecast the solar radiation, which surpasses the two-dimensional linear predictive filtering model and the traditional statistical forecasting method by means of empirical analysis. Amrouche and Le Pivert (2014) adopted the idea of spatial modeling and artificial neural networks (ANNs) to forecast the daily solar total radiation of four sites in the United States, in which the empirical results show that the forecasting results of this proposed model satisfy the expected accuracy requirements. Benmouiza and Cheknane (2013) classified the input data by K-means, then modeled different class by using nonlinear autoregressive neural network, and finally predicted the solar radiation of test data by the corresponding model.

In recent years, the integrated model has grown rapidly. Paoli et al. (2010) used the integrated model to predict the total solar radiation in three sites in France. First, they pretreated the original total solar radiation sequence by using the seasonal index adjustment method. And then they used the multi-layer perceptron neural network (MLPNN) for daily solar radiation prediction. The results show that the mean absolute percentage error of the multi-layer perceptual neural network model is about 6%, which is superior to the ARIMA, the Bayesian, Markov chain model and K-nearest neighbor algorithm. Lauret et al. (2015) used three different methods which are artificial neural network (ANN), Gaussian process (GP) and support vector machine (SVM) to forecast global horizontal solar irradiance (GHI). Through analyzing the actual data in three different places in France, the three machine learning algorithms proposed in this paper are found to be better than the linear autoregressive (AR) and persistence model. Gala et al. (2016) used a combination of support vector regression (SVR), gradient boosted regression (GBR), and random forest regression (RFR) to predict three-hour accumulated solar radiation of 11 regions in Spain. Wang et al. (2015) proposed a new integrated model to predict hourly solar radiation in 2015. First, they used the network structure of multi-response sparse regression (MRSR), leave-one-out cross validation, and extreme learning machine (ELM), then used the cuckoo search (CS) to optimize its weight and threshold, and finally analyzed six sites in the US. The prediction results show that this combination model is stronger than the ARIMA, BPNN and optimally pruned extreme learning machine (OP-ELM).

The main innovation of this paper is to propose a novel decomposition clustering ensemble (DCE) learning approach integrating EEMD, K-means and LSSVR to improve the performance of solar radiation forecasting by means of forecasting accuracy and robustness, and to compare its forecasting power with some popular existing forecasting models. The rest of the paper is organized as follows. Section 2 describes the formulation process of the proposed DCE learning approach in detail. Related methodologies are illustrated in Section 3. The empirical results and effectiveness of the proposed approach are discussed in Section 4. Finally, Section 5 provides some conclusions and indicates the direction of further research.

2. Decomposition-Clustering-Ensemble (DCE) learning approach

According to the work of TEI@I methodology (Wang et al., 2005; Wang, 2004), which is based on integration (@ Integration) of text mining (T) plus econometrics (E) plus intelligence techniques (I). Yu et al. (2008) proposed a decomposition ensemble learning approach for crude oil price forecasting. Recently, this learning approach has been applied in many fields, including financial time series forecasting (Yu et al., 2009), nuclear energy consumption forecasting (Tang et al., 2012; Wang et al., 2011), hydropower consumption forecasting (Wang et al., 2011), crude oil price forecasting (He et al., 2012; Yu et al., 2014; Yu et al., 2015; Yu et al., 2016), etc.

By analyzing the literature above in detail, there are three main steps involved in the decomposition and ensemble learning approach, i.e. decomposition, single forecasting and ensemble forecasting. Firstly, some decomposition algorithms can be used to decompose the original time series into a number of meaningful components. Secondly, some forecasting methods are applied to forecast all components respectively. Finally, these forecasting results of each component are combined to generate an aggregated output as the final forecasting result using some ensemble methods. It can be concluded that ensemble learning is critical to the final forecasting results. Sub-component forecasting has different attributes in different time. If ensemble weights are the same all the time, different attributes cannot be captured. Therefore, DCE learning approach is proposed in this paper, which employs clustering scheme to cluster the sub-component forecasting results. By using different ensemble weights in different forecasting time, a better performance can be obtained by compared with the fixed ensemble weights.

The framework of DCE learning approach is shown in Fig. 1. It can be seen from Fig. 1, DCE learning approach contains four steps:

(1) Decomposition. Decomposition method is introduced to decompose the original time series into some relatively simple and meaningful component series.
(2) Individual forecasting. Various forecasting models are employed to forecast each component series.
(3) Clustering. A clustering method is used to cluster the individual forecasting results.
(4) Ensemble forecasting. An ensemble method is applied to calculate the ensemble weights of different cluster. Then the corresponding clusters' ensemble weights are used for component forecasts to obtain the ensemble forecasting results.

3. Related methodologies

3.1. Ensemble empirical mode decomposition

The empirical mode decomposition (EMD) is an adaptive signal decomposing method which is proposed by Huang et al. (1998). It can be used to decompose the signal into several intrinsic mode functions (IMFs) and a residual series.

Each IMF has a zero local mean value and its number of extreme values and zero crossings are equal or differ at most by one. Different IMFs have different frequency ranges and represents different kinds of natural oscillation modes embedded in the original signal. Thus, different IMFs may highlight different details of the signal. Compared with other methods, the key advantage of EMD is that the basis function is derived directly from the original signal rather than a priori fixed basis function. It is especially suitable for analyzing non-linear and non-stationary signals.

The ensemble empirical mode decomposition (EEMD) is proposed as an improved version of EMD which overcomes the mode mixing problem of EMD. The main idea of EEMD is to add white noise into the original signal with many trials. Then the EMD is applied to the noisy