Energy Conversion and Management 118 (2016) 331–344

Contents lists available at ScienceDirect

Energy Conversion and Management

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

A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting

R. Azimi [a], M. Ghayekhloo [a], M. Ghofrani [b,*]

[a] Young Researchers and Elite Club, Qazvin Branch, Islamic Azad University, Qazvin, Iran
[b] School of Science, Technology, Engineering and Mathematics (STEM), University of Washington, Bothell, USA

ARTICLE INFO

Article history:
Received 9 January 2016
Received in revised form 31 March 2016
Accepted 2 April 2016

Keywords:
Clustering
Data preprocessing
Forecasting
Solar radiation
TB K-means

ABSTRACT

Accurate forecasting of renewable energy sources plays a key role in their integration into the grid. This paper proposes a hybrid solar irradiance forecasting framework using a Transformation based K-means algorithm, named TB K-means, to increase the forecast accuracy. The proposed clustering method is a combination of a new initialization technique, K-means algorithm and a new gradual data transformation approach. Unlike the other K-means based clustering methods which are not capable of providing a fixed and definitive answer due to the selection of different cluster centroids for each run, the proposed clustering provides constant results for different runs of the algorithm. The proposed clustering is combined with a time-series analysis, a novel cluster selection algorithm and a multilayer perceptron neural network (MLPNN) to develop the hybrid solar radiation forecasting method for different time horizons (1 h ahead, 2 h ahead, . . ., 48 h ahead). The performance of the proposed TB K-means clustering is evaluated using several different datasets and compared with different variants of K-means algorithm. Solar datasets with different solar radiation characteristics are also used to determine the accuracy and processing speed of the developed forecasting method with the proposed TB K-means and other clustering techniques. The results of direct comparison with other well-established forecasting models demonstrate the superior performance of the proposed hybrid forecasting method. Furthermore, a comparative analysis with the benchmark solar radiation forecasting models shows that the proposed model gives better forecasting results.

1. Introduction

Renewable energy forecasting is a rapidly evolving field. Accurate forecasting of photovoltaic (PV) sources makes them more reliable and their implementation more widespread. PV prediction is challenging due to changing weather patterns, solar radiation, and potential mathematical modeling errors [1]. An accurate forecast requires a model accounting for all these variables. This data, if processed appropriately, significantly improves the accuracy of the forecast. Several data processing algorithms have been used to characterize irregular solar irradiance datasets and provide a data mining-based prediction technique. Cluster analysis plays an important role in solving many problems in unsupervised pattern recognition [2]. In the field of renewable energy forecasting, this technique allows handling groups of data separately, which by identifying anomalies and neglecting outliers, provides a better understanding of the collected information and improves the accuracy of the final forecast results. Several clustering methods have been used to provide a pattern-based prediction technique for solar irradiance [3–21]. In Ref. [3], a new fuzzy inference method was introduced to predict hourly global solar irradiation based on the daily clearness index. In this method, fuzzy *c*-means (FCM) clustering is established to find optimal values for the membership function parameters. Performance and limitations of the model were evaluated using data from 10 European stations with latitudes between 40° and 50° northern. The accuracy of the model was shown to be adequate for forecasting the performance of photovoltaic systems in this band of latitudes. Ref. [4] proposed a new time series clustering technique to reduce the computational complexity of smart grid optimization problems in applications involving optimum estimation and forecasting of renewable energy consumption. A hybrid forecasting technique was presented in Ref. [5] to predict the hourly power output for a photovoltaic (PV) system. The proposed hybrid method was a combination of the deterministic annealing (DA) clustering technique, an improved version of the particle swarm optimization

* Corresponding author at: UWBB Room 227, 18807 Beardslee Blvd, Bothell, WA 98011, USA.
E-mail address: mghofrani@uwb.edu (M. Ghofrani).

http://dx.doi.org/10.1016/j.enconman.2016.04.009
0196-8904/

(PSO) algorithm, generalized radial basis function network (GRBFN), and weight decay (WD) method for forecasting. Ref. [6] introduced a short term PV generation forecasting by synthesis self-organizing maps (SOM), algorithm and wavelet neural networks. SOM clustered the weather data and seasonal variations and the wavelet neural network created forecasting models for each cluster sample. However, the SOM algorithms may converge to non-optimal clustering results depending on the initialization and learning rate considered for the algorithm. In addition, neighborhood violations occur if the output space topology does not match with the data shape. In Ref. [7], an adaptive model is proposed based on artificial neural networks (ANN) learning technique to forecast global, direct and diffuse solar radiation from air temperature, relative humidity and sunshine hour values. Ref. [8] developed a method based on clustering and learning vector quantity (LVQ) for predicting solar flares. K-means algorithm was used to cluster imbalanced data and provide better training datasets. The LVQ method was then utilized to organize the datasets into class labels to indicate whether the solar flare happens in the next 24 h or not. However, the accuracy of the LVQ method is highly dependent on the initialization of the model as well as the learning parameters used (learning rate, training iterations, etc.). The accuracy is also dependent on the class distribution of the training data. A good distribution of samples is required to construct useful models. Moreover, it is difficult to determine the appropriate number of codebook vectors for a given problem. Ref. [9] proposed a novel solar radiation forecasting model using an adaptive neuro-fuzzy inference system (ANFIS) based upon day of the year as the only input. It was shown that the ANFIS model plays an important role in the field of solar radiation forecasting. However, computational complexities of the ANFIS model including the determination of network parameters such as number and type of membership functions, number of rules for each input variable, number of membership functions for each output variable, and finding the optimal learning parameters limit its functionality for short term forecasting. Support vector regression (SVR) methodology was used in Ref. [10] to forecast the horizontal global solar radiation based on the daily hours of sunshine and daylight as effective inputs. The obtained results showed that the proposed method is appropriate to forecast the monthly average daily global solar radiation with high precision levels. Support vector machine was combined with wavelet transform algorithm (SVM-WT) for feature extraction and was used in conjunction with artificial neural network (ANN) and genetic algorithm (GA) to estimate the monthly average daily global solar radiation [11]. The results indicated that SVM-WT increases the forecasting accuracy. However, the major limitation of SVM lies in the choice of the kernel function. It also has several key parameters that need to be set correctly to achieve the appropriate results for any given problem. In addition, the performance of SVM degrades as the number of features increases. Ref. [12] proposed a novel multimodal framework for solar radiation forecasting. The framework was based on the assumption that there are several patterns in the stochastic component of solar radiation series. The time series data were classified into multiple subsequences. The K-means algorithm was then applied to group the subsequences into different clusters. Finally, the time-delay neural network (TDNN) was trained to model a specific pattern in each cluster. The pattern corresponding to the current time was then determined for the forecasting purpose. This process was followed by selecting the appropriate trained TDNN model. K-means clustering and auto regressive neural networks (NAR) were combined with Ref. [13] to develop a hybrid method for solar forecasting. K-means clustering algorithm was used in Ref. [14] to predict wind and PV energy production. A hybrid of K-means algorithm and cumulative distribution function was presented in Ref. [15] that uses a daily clearness index for solar forecasting. The efficiency of the K-means is highly related to the initial cluster centroids and randomly selecting the initial centroids may lead to incorrect results. K-means is also quite sensitive to noisy data, which can influence the accuracy of forecasts [17].

Ref. [18] proposed a novel hybrid method by combining satellite image analysis with exponential smoothing state space (ESSS) model and artificial neural network (ANN) for solar radiation forecasting. Geostationary satellites as a weather information provider, allows self-organizing map (SOM) method to derive and analyze a cloud cover index. The ESSS model was used for cloud cover index forecasting. ANN technique was then applied to estimate solar radiation based on the predicted cloud cover index.

Ref. [19] proposed a hybrid method by combining SOM, support vector regression (SVR) and particle swarm optimization (PSO) for hourly solar radiation forecasting. First, the input space was divided into several disjointed regions or clusters by SOM algorithm. The characteristic correlation between the actual and predicted values was then calculated by the SVR models. PSO algorithm was used to deal with the volatility of SVR with different parameters. This led to a better performance of the SVR models for solar forecasting. The SOM algorithm, if not initialized and trained appropriately, converges to a local optimum. In addition, the computational complexity of the algorithm limits its application for large-volume or high-dimensional dataset. Moreover, its neighborhood preserving feature violates when the output space topology does not match with the structure of the data in the input space.

A solar radiation estimation method was presented in Ref. [20]. This method used a combination of the satellite images, solar radiation data and a variety of other atmospheric and weather satellite data to generate solar radiation maps to study the sun and its behavior at different geographical coordinates (latitude and longitude). The method is however not applicable for regions with the view limitations of the Solar Geographical Information System (SolarGIS) such as United States, Canada, Northern European countries, Japan, and parts of Australia. Ref. [21] proposed a hybrid machine learning (ML) method by combining support vector regression (SVR), gradient boosted regression (GBR) and random forest regression (RFR) to improve the initial radiation forecasts provided by the state-of-the-art European Center for Medium range Weather Forecasting (ECMWF) model. Based on the evaluation made in this paper, the ECMWF with SVR method enhanced the solar forecasting accuracy.

The chaotic nature of the solar data disrupts the neural network learning process and presents the forecasting results with high errors, especially when a cloudy day appears after a number of successive clear days [22]. In this paper, we propose a novel clustering approach named TB K-means to group the solar time series data into clusters to better identify its anomalies and irregular patterns. The proposed clustering method splits the training data into multiple clusters. This technique allows handling groups of data separately, which by better detecting anomalous data points (outliers) and irregular patterns, provides a better characterization for the collected data. Clustering the solar data provides the most appropriate portion of the data rather than the whole data to be used for the NN training. This leads to deep learning for neural networks and improving the accuracy of solar forecasting. The proposed clustering method combines a new initialization technique, K-means algorithm and a new gradual data transformation approach to appropriately select the initial cluster centroids and move the real data into the locations of the initial cluster centroids that are closer to the actual positions of the associated data. By doing this, the data are placed in an artificial structure. The inverse transformation is then performed to gradually move back the pathological data to their original places. During this process, K-means updates the clustering centroids after any changes in the data structure. This increases the clusters' coherence during