Energy Conversion and Management 127 (2016) 208–225

Contents lists available at ScienceDirect

Energy Conversion and Management

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

A hybrid wind power forecasting model based on data mining and wavelets analysis

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

[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 May 2016
Received in revised form 31 August 2016
Accepted 1 September 2016

Keywords:
Data mining
Data preprocessing
Forecasting
HANTS
Wavelet decomposition
Wind power

ABSTRACT

Accurate forecasting of wind power plays a key role in energy balancing and wind power integration into the grid. This paper proposes a novel time-series based K-means clustering method, named T.S.B K-means, and a cluster selection algorithm to better extract features of wind time-series data. A hybrid of T.S.B K-means, discrete wavelet transform (DWT) and harmonic analysis time series (HANTS) methods, and a multilayer perceptron neural network (MLPNN) is developed for wind power forecasting. The proposed T.S.B K-means classifies data into separate groups and leads to more appropriate learning for neural networks by identifying anomalies and irregular patterns. This improves the accuracy of the forecast results. A cluster selection method is developed to determine the cluster that provides the best training for the MLPNN. This significantly accelerates the forecast process as the most appropriate portion of the data rather than the whole data is used for the NN training. The wind power data is decomposed by the Daubechies D4 wavelet transform, filtered by the HANTS, and pre-processed to provide the most appropriate inputs for the MLPNN. Time-series analysis is used to pre-process the historical wind-power generation data and structure it into input-output series. Wind power datasets with diverse characteristics, from different wind farms located in the United States, are used to evaluate the accuracy of the hybrid forecasting method through various performance measures and different experiments. A comparative analysis with well-established forecasting models shows the superior performance of the proposed forecasting method.

1. Introduction

Wind power is one of the fastest growing renewable energy generations due to its mature technology as well as its environmental and societal benefits [1]. Highly variable and intermittent nature of wind energy causes reliability and stability issues for the power system operation.

Accurate wind power forecasting enhances the value of wind energy by improving the reliability and economic feasibility, and by reducing the integration and operational costs of wind power production [2].

Wind power is subject to several different meteorological parameters, such as air temperature, wind speed and relative humidity, which make its forecasting difficult and complicated [3]. An inaccurate forecast may result in a grid imbalance between supply and demand [4]. Wind forecasting is broadly categorized as very short-term, short-term and long term based on the prediction horizon considered for the study.

1.1. Related works

Several approaches have been proposed for short-term wind speed and power forecasting [5–19].

A general spatio-temporal time-series prediction methodology was proposed in [5]. Different machine learning techniques including random walk, auto-regressive integrated moving average (ARIMA), Regression Trees (RT), Support Vector Machines (SVM) and Random Forest (RF) were used for hourly wind speed estimation.

Ref. [6] proposed two collaborative methods for wind power forecasting. The collaboration was in the form of information sharing in a network of wind farms. The first method presented a centralized management where the data points are exchanged between nodes. ARIMA models of the correlated wind farms were shared in the second method. However, ARIMA models have limited capabilities to forecast time series with sudden changes.

* 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.2016.09.002
0196-8904/Published by Elsevier Ltd.

Data mining algorithms have been widely used in hybrid forecasting methods for renewable energy generation. Different data mining techniques have been used for wind speed and power prediction.

Ref. [7] provided models for short term wind speed forecasting models based on linear and non-linear autoregressive moving average (ARMA and NARMA). The parameters of the linear models were provided using the Gauss–Newton method and the non-linear autoregressive models were developed using data mining methods including support vector machine regression (SMOreg), bagging and M5 model rules (M5R). However, both ARMA and NARMA models require exhaustive time series observation which limits their applications for short term forecasting.

A new classification, named least-squares support vector machine (LSSVM) was used in [8] for short term wind speed forecasting. Linear, Gaussian, and polynomial kernels were proposed as three SVM kernels. The training sample size, SVM order, regularization parameter, and kernel parameters were adjusted to obtain accurate forecast results. LSSVM and gravitational search algorithm (GSA) were combined in Ref. [9] to provide a hybrid model (LSSVM–GSA) for wind power prediction. The best parameter of the model was estimated by GSA and the optimal kernel function was selected. Proper adjustment of the SVM parameters and selection of the kernel function is the major limitation of the methods proposed in [8,9]. In addition, increasing the number of features degrades the SVM performance. Furthermore, high-computational complexity of LSSVM (n^3 order) significantly reduces the speed as the sample size (n) increases.

A new hybrid model was developed in Ref. [10] to predict wind speed for 1–5 h ahead. Wavelet decomposition was used to decompose time series data into subseries. Each subseries were then predicted with neural network which is trained by crisscross optimization algorithm. Eventually, the final result was created by an aggregate calculation.

A novel ANN-Markov chain (MC) model was proposed in Ref. [11] to forecast the second-scale and hour scale wind speed. MC was applied to forecast the long-term wind speed while ANN was applied to forecast the short-term wind speed. The major disadvantage of the MC model is the large data set that is required for its training.

A statistical-based wind power forecasting method, named AWNN (adaptive wavelet neural network), was proposed in Ref. [12] without using numerical weather prediction (NWP) inputs. The forecasting method included two phases. In the first phase, wavelet decomposition of wind series was performed; then an adaptive wavelet neural network was used to apply regression on each decomposed signal and predict the wind speed. In the second phase, a feed-forward neural network was used to map between the output wind speed and wind power data. However, this method has poor performance for predicting sudden changes in wind power generation.

Ref. [13] proposed a neighborhood-based estimation approach using a spatio-temporal lazy learning algorithm based on K-Nearest-Neighbor (KNN) technique for very short term wind speed forecasting. This method was applied for monitoring a grid of wind farms, which collaborate by sharing information. The lazy learning has some disadvantages including high memory/storage requirements for large problem domains and time-consuming search for similar examples. The large memory/storage requirement is to store the entire training dataset including the noisy training data which unnecessarily increases the case base as no abstraction is made during the training phase. Another disadvantage is that lazy learning methods are usually slower to evaluate. In particular, KNN method lacks an efficient algorithm to determine the value of parameter K (number of nearest neighbors). This method has also high computational complexity because it requires calculating distance of each instance query with respect to all training samples.

Clustering is one of the main branches of data mining. The main goal of clustering is to generate compact groups of objects or data that share similar patterns within the same cluster, and isolate these groups from those which contain elements with different characteristics. In the field of renewable energy forecasting, this technique allows handling groups of data separately, which provides a better understanding of the collected information and improves the accuracy of the final forecast results. Several clustering methods have been used to identify patterns and provide a wind prediction technique [14–19].

Ref. [14] proposed a novel multi-step wind forecasting method, based on a novel Fuzzy System, weather research and forecasting (WRF) model, and a Cuckoo Search (CS) optimization algorithm to correct the results obtained using physical laws. However, CS optimization algorithm has drawbacks including weak local searching ability, slow convergence rate and low optimization accuracy.

Ref. [15] proposed a hybrid machine learning-based model, using Bayesian clustering by dynamics (BCD) and support vector regression (SVR), to provide short-term forecasting for the power generation of a wind farm. However, BCD method is limited to the univariate time series, that is, the algorithm is able to cluster the behaviors of only one variable at a time.

Ref. [16] introduced several models for wind turbine power curve classification based on parametric and nonparametric methods. The parametric models were developed based on logistic expression and least mean-square algorithms such as genetic algorithm, evolutionary programming, particle swarm optimization, and differential evolution approach. Fuzzy C-means (FCM) clustering algorithm, data mining techniques and neural networks were used to develop the nonparametric models. However, limitations of least mean-square algorithms including unstable convergence and easily trapping in regional optimum may lead to non-optimal solutions.

Ref. [17] proposed a hybrid method based on K-means clustering algorithm, linear machine classifier method and MLPNN for wind power forecasting.

A fuzzy weighted clustering algorithm was proposed in Ref. [18] for wind speed and wind direction data preprocessing. A dynamic Elman neural network was then used for short-term wind power forecasting. The proposed fuzzy clustering method has the disadvantages of sensitivity to random initial values, easily falling into local optimum, and heavy calculations for the multidimensional data.

A hybrid of Bayesian information criterion, K-means clustering and neural network was used in Ref. [19] to provide a statistical short-term wind power forecasting model. Bayesian information criterion was used for identifying the bad data. K-means algorithm clustered numerical weather prediction data and neural network predicted the wind power. However, K-means algorithm is very sensitive to the initial cluster centroid that is randomly selected in the first phase of the algorithm. The random selection may provide incorrect results for clustering [20]. Different variants of K-means algorithm have been proposed to address this limitation [21–25]. Ref. [21] proposed a global K-means algorithm for finding optimal solutions. However, this search–based clustering method is not suitable for medium and high-volume data, due to its high computational load. FCM algorithm was proposed in [22] to enhance the accuracy of the K-means clustering. However, it requires more computational processing and therefore is slower than the K-means algorithm. K-means++ algorithm was proposed in Ref. [23] to obtain an appropriate initial set of centroids for initializing the K-means algorithm. However, the inherent sequential nature of this method exponentially increases its processing time