ARTICLE IN PRESS

International Journal of Forecasting ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

International Journal of Forecasting

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

K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting

Yao Zhang, Jianxue Wang *

School of Electrical Engineering, Xi'an Jiaotong University, Xi'an, China

ARTICLE INFO

Keywords:
Wind power
Probabilistic forecasting
Point forecasting
k-nearest neighbors
Kernel density estimator
Coordinate descent algorithm

ABSTRACT

Probabilistic forecasts provide quantitative information in relation to energy uncertainty, which is essential for making better decisions on the operation of power systems with an increasing penetration of wind power. On the basis of the k-nearest neighbors algorithm and a kernel density estimator method, this paper presents a general framework for the probabilistic forecasting of renewable energy generation, especially for wind power generation. It is a direct and non-parametric approach. Firstly, the k-nearest neighbors algorithm is used to find the k closest historical examples with characteristics similar to the future weather condition of wind power generation. Secondly, a novel kernel density estimator based on a logarithmic transformation and a boundary kernel is used to construct wind power predictive density based on the k closest historical examples. The effectiveness of this approach has been confirmed on the real data provided for GEFCom2014. The evaluation results show that the proposed approach can provide good quality, reliable probabilistic wind power forecasts.

1. Introduction

The random and intermittent nature of wind power is a big challenge for the optimal use of renewable energy. Accurate predictions of wind power generation are effective tools for satisfying the requirements of power system planning and operation (Pinson, 2013). However, most of the wind power forecasting approaches that are used at present only give the expected wind power output, which is a deterministic (point) prediction. Recently though, a new kind of forecasting method, probabilistic forecasting, has arisen as an active topic. Unlike point predictions, probabilistic forecasts can produce quantitative information on the associated uncertainty of wind power generation (Zhang, Wang, & Wang, 2014). Thus, the focus of GEFCom2014 is on probabilistic energy forecasting (Hong et al., this issue).

In this paper, a general framework of probabilistic forecasting based on the k-nearest neighbors algorithm (k-NN) and the kernel density estimator (KDE) method is implemented for forecasting the wind power generation. Firstly, the k-NN algorithm was used to find the k historical examples with features most similar to the future weather condition. Then, the point forecast of the wind power output is the weighted average of the values of k nearest neighbors. Secondly, the KDE method based on a logarithmic transformation and a boundary kernel was proposed for constructing the wind power predictive density. The proposed model ranked fifth in the wind power track of GEFCom2014. The evaluation results confirmed that a combination of the k-NN algorithm and the KDE method can provide good quality, reliable probabilistic wind power forecasts.

* Corresponding author. Tel.: +86 29 82668868; fax: +86 29 82665489.
E-mail addresses: zy.3110161019@stu.xjtu.edu.cn (Y. Zhang), jxwang@mail.xjtu.edu.cn (J. Wang).

http://dx.doi.org/10.1016/j.ijforecast.2015.11.006
0169-2070/

Table 1
Candidate input variables for building forecasting model.

Name	Example	Explanation
YEAR	2012/2013/...	Calendar year with the century
MONTH	01/02/...	Calendar month as decimal number
DAY	001/002/...	Day of year as decimal number
HOUR	00/01/...	Hours as decimal number
WS100	4.66903 m/s	Wind speed forecasting at 100 m
WD100	216.25525°	Wind direction forecasting at 100 m
WP100	0.05488 p.u.	Wind power output (spot prediction)

2. Data preparation

The quality of the historical data has an impact on the wind power prediction accuracy. The original wind power data included several outliers which may be due to wind turbine maintenance. In the original GEFCom2014 dataset, the wind power output at some farms was found to be zero for many hours in succession. If the value is zero for too long, it is reasonable to suspect that the corresponding data points may be outliers. In this paper, if the wind power output is zero for 48 h or more in succession, it is defined as a maintenance event. Outliers caused by maintenance events are removed from the dataset. In this paper, the definition of a maintenance event is arrived at mainly from the point of view of industrial experience. Although this definition may be a bit too strong, it is still effective for detecting some maintenance events correctly.

The raw dataset only provides the (u, v)-wind prediction, i.e., (U10, V10) and (U100, V100), which is then converted into the wind speed w and wind direction ϕ. The conversion is conducted as

$$w = \sqrt{u^2 + v^2}, \tag{1}$$

and

$$\phi = \begin{cases} \dfrac{180}{\pi} \times \operatorname{atan2}(u, v) & \text{if } \operatorname{atan2}(u, v) \geq 0 \\ \dfrac{180}{\pi} \times \operatorname{atan2}(u, v) + 360 & \text{if } \operatorname{atan2}(u, v) < 0. \end{cases} \tag{2}$$

The numeric weather prediction (NWP) model may be stochastic and may result in poor quality forecasts. Thus, a moving average of order five is employed to pre-treat the NWP output. After the data preparation process, Table 1 provides the attributes that are considered as candidate input variables for building the probabilistic forecasting model. The attributes 'YEAR', 'MONTH', 'DAY' and 'HOUR' are extracted from the attribute 'TIMESTAMP'. As the hub height of a modern wind turbine is between 40 and 120 m, we chose to use U100 and V100 as input variables, instead of U10 and V10. The attribute 'WP100' represents the wind power spot prediction, and is obtained using multiple linear regression (MLR). The MLR model takes the wind speed 'WS100' and the time of day 'HOUR' as regressors. The MLR model is formulated as follows:

$$\begin{aligned} \hat{p}_{t+k|t} &= b_0 + b_1 w_{t+k|t} + b_2 w_{t+k|t}^2 + b_3 w_{t+k|t}^3 \\ &\quad + b_4 \cos\left(\frac{2\pi}{24} d_{t+k|t}\right) + b_5 \sin\left(\frac{2\pi}{24} d_{t+k|t}\right) \\ &\quad + b_6 \cos\left(\frac{4\pi}{24} d_{t+k|t}\right) + b_7 \sin\left(\frac{4\pi}{24} d_{t+k|t}\right), \end{aligned} \tag{3}$$

where $w_{t+k|t}$ is the wind speed prediction for time $t + k$ given at time t (i.e., the attribute 'WS100' in Table 1). d_{t+k} is the time of day (i.e., the attribute 'HOUR' in Table 1), which takes the values $0, 1, \ldots, 23$. Eq. (3) captures diurnal periodicity using a Fourier series of the time of day. The range of wind directions, 'WD100', is split into several sub-intervals, and the multiple linear regression model is fitted for each sub-interval.

3. Forecasting methodology

3.1. *k*-nearest neighbors algorithm

The main idea of k-NN algorithm is that whenever there is a new point to predict, its k nearest neighbors are chosen from the training data. Then, the prediction of the new point can be the average of the values of its k nearest neighbors. Wind power point prediction using the k-NN algorithm has been developed successfully in recent years (Mangalova & Agafonov, 2014), and resembles the similar-day approach for electrical load forecasting (Hong, 2014). The similar-day approach is still used by many utilities, and derives the future power load using historical days with similar temperatures and day types. In this paper, the k-NN algorithm is used to find the appropriate historical examples with characteristics similar to the future weather condition (provided by the NWP model). Then, wind power observations of these historical examples will be extracted and used to construct the wind power predictive density. The main process of probabilistic wind power forecasting using the k-NN algorithm can be boiled down to four steps:

- Calculating the pre-defined distance between the testing example and the training example;
- Choosing k nearest neighbors from the training examples with the k smallest distances;
- Predicting the wind power output based on a weighted averaging technique; and
- Predicting the wind power density based on the kernel density estimator method.

Firstly, a distance measure is required to characterize the similarity of any two instances (each record in the dataset is called an instance). The most commonly used distance metrics are the Euclidean distance, the Manhattan distance and the Mahalanobis distance. In this paper, the original Manhattan distance was enhanced by using weighting. The weighted Manhattan distance is calculated as

$$D[X, Y] = \sum_{i=1}^{n} w_i |x_i - y_i|, \tag{4}$$