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 for GEFCom2014 probabilistic wind power forecasting

Ekaterina Mangalova *, Olesya Shesterneva

Siberian State Aerospace University named after academician M. F. Reshetnev, Krasnoyarsk, Russia

ARTICLE INFO

Keywords:
Probabilistic forecasting
Nonparametric smoothing
Quantile estimation
Distance metric
Optimization

ABSTRACT

The paper deals with a forecasting procedure that aims to predict the probabilistic distribution of wind power generation. The k-nearest neighbors algorithm is adapted for this probabilistic forecasting task. It allows quantiles to be estimated without requiring assumptions as to the probability distribution. The influences of several factors (wind speed, wind direction and hour) on the normalized wind power are investigated. The feasibility of the approach is demonstrated through the probabilistic wind power forecasting track of the Global Energy Forecasting Competition 2014.

1. Introduction

The importance of wind energy has increased in recent years, as wind power capacity has grown rapidly, and wind energy production is around 4% of the total worldwide electricity usage. At least 83 countries around the world use wind power to supply their electricity grids (REN21 Renewables, 2011).

The effective operation of wind power plants involves the optimization of their operating modes within the integrated energy system. In particular, predictions of the power production of each individual wind farm are crucial for energy management.

The aim of the probabilistic wind power forecasting track is to forecast the probabilistic distribution (in quantiles) of the wind power generation of 10 wind farms, all located in the same region of the globe. The target variable is power generation, normalized by the respective nominal capacity of each wind farm. The explanatory variables are the past power measurements and input weather forecasts, given as u and v components (zonal and meridional), which can be transformed to wind speeds and directions. These forecasts are given at two heights, 10 and 100 m above ground level (Hong et al., in this issue).

This paper consists of eight sections. Section 2 justifies the application of the k-nearest neighbors algorithm to probabilistic forecasting. Section 3 defines significant features and introduces the notation. Section 4 describes the k-nearest neighbors algorithm for wind power probabilistic forecasting. Section 5 describes the optimization procedure and shows the influence of the parameters. Section 6 shows the influence of the zonal and meridional components on normalized wind power, and describes the disadvantages of using the wind direction instead of wind components. Section 7 presents the results of the k-nearest neighbors algorithm implementation. The paper is concluded in Section 8.

2. Justification of algorithm

There are two density estimation strategies: parametric and nonparametric. The parametric approach assumes that a distribution family is chosen, e.g., the Gaussian distribution, then the parameters of the distribution are estimated from the available data. The nonparametric approaches are more flexible than the parametric ones,

* Corresponding author.
E-mail address: e.s.mangalova@hotmail.com (E. Mangalova).

http://dx.doi.org/10.1016/j.ijforecast.2015.11.007

Please cite this article in press as: Mangalova, E., & Shesterneva, O., K-nearest neighbors for GEFCom2014 probabilistic wind power forecasting. International Journal of Forecasting (2016), http://dx.doi.org/10.1016/j.ijforecast.2015.11.007

since no distribution of the response has to be assumed. In this case, the distribution is estimated from the data directly (Silverman, 1998).

Power generation is normalized by the respective nominal capacity of each wind farm. Thus, the range of the target variable is between zero and one, and typical parametric distribution assumptions (e.g., Gaussian) are incorrect. Moreover, the distributions of the target variable are obviously different for different wind speeds (Fig. 1). For this reason, a nonparametric approach is chosen.

The k-nearest neighbors algorithm is highly efficient for spot wind power forecasting (Mangalova & Agafonov, 2014), and has several advantages. First, the k-nearest neighbors algorithm allows its results to be interpreted by experts. Second, the significance of the input variables can be estimated, provided that a weighted distance metric is used. Third, no multiple learning is needed when new data are obtained.

The goal of this work is to adapt the k-nearest neighbors algorithm to probabilistic forecasting.

3. Significant factors and data pretreatment

Based on research devoted to GEFCom2012 (Mangalova & Agafonov, 2014), significant factors have been chosen. The day of the year is excluded from the list of significant factors, because the training set in GEFCom2012 is longer than that in GEFCom2014, and an effective use of this factor in GEFCom2014 is impossible.

A list of the variables and their notations is provided in Table 1.

The training set is represented by $T = (x_i, y_i)$, $i = 1, 2, \ldots, n$, where

$$x_i = \left\{ws_i^{\langle 1\rangle}, ws_i^{\langle 2\rangle}, uv_i^{\langle 1\rangle}, uv_i^{\langle 2\rangle}, uv_i^{\langle 3\rangle}, uv_i^{\langle 4\rangle}, h_i, wsn_i^{\langle 1\rangle}, wsn_i^{\langle 2\rangle}\right\}. \quad (1)$$

Neighboring wind speeds $wsn^{\langle 1\rangle}$ and $wsn^{\langle 2\rangle}$ are added to the model if doing so makes a noticeable improvement in model accuracy. It is important to note that neighboring wind speeds can be more useful than wind speeds at the wind farm under consideration. For example, when predicting the wind power at wind farm #1, the wind forecasts for wind farm #7 are more significant than those for wind farm #1.

Weather forecasts contain the most significant information for the model process, but they are stochastic spot forecasts. Mangalova and Agafonov (2014) propose the following procedures for pretreating weather forecasts:

Simple moving average:

Step 1: Assign

$$ws'^{\langle m\rangle} = ws^{\langle m\rangle} = \left(ws_1^{\langle m\rangle}, \ldots, ws_n^{\langle m\rangle}\right). \quad (2)$$

Step 2: Calculate

$$ws_i^{\langle m\rangle} = \frac{\sum_{j=-2}^{1} ws'^{\langle m\rangle}_{i+j}}{4}, \quad i = 3, \ldots, n-1. \quad (3)$$

Modified moving average:

The modified moving average is calculated consistently for each i:

$$ws_i^{\langle m\rangle} = \frac{\sum_{j=-2}^{1} ws^{\langle m\rangle}_{i+j}}{4}, \quad i = 3, \ldots, n-1. \quad (4)$$

Pretreatment is used for wind speeds and zonal and meridional components at two heights: 10 and 100 m above ground level.

4. K-nearest neighbors for probabilistic forecasting

The k-nearest neighbors algorithms can be used for estimating probabilistic distributions. Such algorithms estimate quantiles using the outputs of the nearest neighbors.

The use of different distance metrics and methods of quantile estimation results in different versions of the k-nearest neighbors algorithm for probabilistic forecasting. We also describe the algorithm for the wind power forecasting track of GEFCom2014.

4.1. Distance metrics

A distance metric is required for calculating the similarity of two observations. The weighted cityblock distance

$$D_1\left(x_i, x_j\right) = \begin{cases} \overbrace{w_1\left|ws_i^{\langle 1\rangle} - ws_j^{\langle 1\rangle}\right| + (2-w_1)\left|ws_i^{\langle 2\rangle} - ws_j^{\langle 2\rangle}\right|}^{\text{wind speed components}} \\ \quad + \overbrace{w_2\sum_{m=1}^{4}\left|uv_i^{\langle m\rangle} - uv_j^{\langle m\rangle}\right|}^{u \text{ and } v \text{ components}} \\ \quad + \overbrace{w_3\sum_{m=1}^{2}\left|wsn_i^{\langle m\rangle} - wsn_j^{\langle m\rangle}\right|}^{\text{neighboring wind speed components}}, & \Delta h_{i,j} < H, \\ \infty, \quad \Delta h_{i,j} \geq H, \end{cases} \quad (5)$$

$$\Delta h_{i,j} = \min\left(\left|h_i - h_j\right|, 24 - \left|h_i - h_j\right|\right), \quad (6)$$

and the weighted squared Euclidean distance

$$D_2\left(x_i, x_j\right) = \begin{cases} \overbrace{w_1\left(ws_i^{\langle 1\rangle} - ws_j^{\langle 1\rangle}\right)^2 + (2-w_1)\left(ws_i^{\langle 2\rangle} - ws_j^{\langle 2\rangle}\right)^2}^{\text{wind speed components}} \\ \quad + \overbrace{w_2\sum_{m=1}^{4}\left(uv_i^{\langle m\rangle} - uv_j^{\langle m\rangle}\right)^2}^{u \text{ and } v \text{ components}} \\ \quad + \overbrace{w_3\sum_{m=1}^{2}\left(wsn_i^{\langle m\rangle} - wsn_j^{\langle m\rangle}\right)^2}^{\text{neighboring wind speed components}}, & \Delta h_{i,j} < H, \\ \infty, \quad \Delta h_{i,j} \geq H, \end{cases} \quad (7)$$

$$\Delta h_{i,j} = \min\left(\left|h_i - h_j\right|, 24 - \left|h_i - h_j\right|\right), \quad (8)$$

Please cite this article in press as: Mangalova, E., & Shesterneva, O., K-nearest neighbors for GEFCom2014 probabilistic wind power forecasting. International Journal of Forecasting (2016), http://dx.doi.org/10.1016/j.ijforecast.2015.11.007