International Journal of Forecasting 30 (2014) 402–406

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Wind power forecasting using the *k*-nearest neighbors algorithm

E. Mangalova [a,*], E. Agafonov [b]

[a] *Siberian State Aerospace University, Russia*

[b] *Siberian Federal University, Russia*

ARTICLE INFO

Keywords:
Cross-validation
Data mining
Energy forecasting*
Forecasting competitions*
Feature selection
Nonparametric models
Regression tree

ABSTRACT

The paper deals with a modeling procedure which aims to predict the power output of wind farm electricity generators. The following modeling steps are proposed: factor selection, raw data pretreatment, model evaluation and optimization. Both heuristic and formal methods are combined to construct the model. The basic modeling approach here is the *k*-nearest neighbors method.

1. Introduction

The effective operation of wind power plants involves the optimization of their operating modes within the integrated energy system. In particular, it demands a prediction of the power output of each single plant. The problem statement and corresponding raw data were taken from the Global Energy Forecasting Competition 2012 (http://www.kaggle.com/c/GEF2012-wind-forecasting).

The data supplied include measurements of the following parameters from seven wind farms: weather forecasts with the meridional and zonal components of the wind, the angle of the wind, and the date and time of the measurement. The data cover a four-year period and consist of 18,757 measurements. Wind forecasts are available twice a day, with each one providing a prognosis for the next two days. Thus, we deal with multiple forecasts with different accuracies.

The data sets supplied are of two kinds: training and validation sets. The former serve as information for the predictive model development, the latter are provided for model quality estimation purposes.

The functioning of wind power plants tends to include irregular but frequent periods of downtime or reduced power delivery, probably as a result of routine maintenance, extraordinary weather, etc. The reasons for such periods are unknown, which leads to difficulties in the data analysis.

The solution to the problem of predicting power plants' electricity output includes the following steps: factor selection, raw data pretreatment, model evaluation and optimization. Both heuristic and formal methods are combined in order to construct the predictive model. The core modeling approach is the *k*-nearest neighbors method.

2. Model factors selection

The weather in the training data sets is represented by retrospective sequences of four forecasts. Starting with the selection of the model factors, we assume that the latest forecast is the most accurate, and therefore we simply remove older ones from the training data set.

The amount of electricity produced by a wind farm depends to a great extent on the air flow parameters. The power of air flow, in its turn, depends not only on the velocity of the flow, but also on the density of the air (Grogg, 2005). Unfortunately, we have no access to density-related parameters (such as temperature, humidity etc.) as such, though they may be expressed indirectly by some day/time information.

Finally, the following set of model factors is proposed for analysis: year, month, day of the month, day of the year, hour, zonal and meridional wind components, wind direction, and wind speed.

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

0169-2070/$ – see front matter
http://dx.doi.org/10.1016/j.ijforecast.2013.07.008

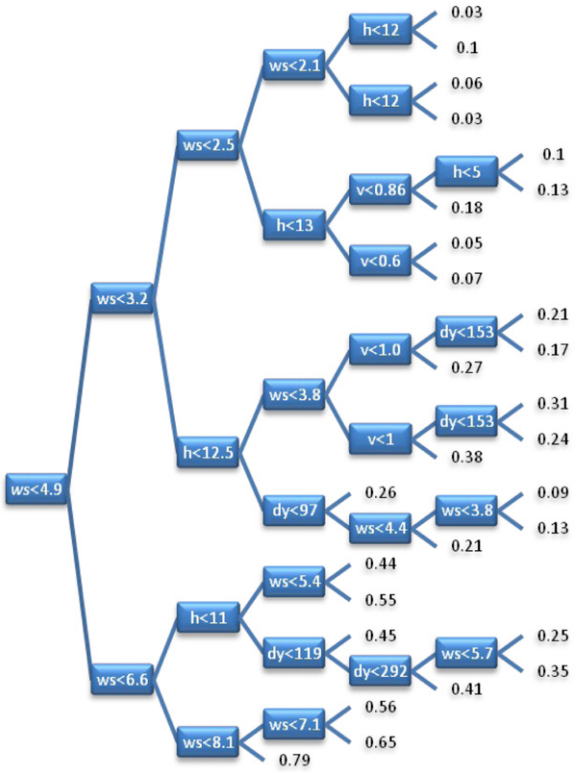

Fig. 1. The CaRT for wind farm 1: The lower alternative is for breaking the corresponding rule, the upper one is for satisfying the rule; *ws* is wind speed, *h* is hour, *dy* is day of the year, and v is the meridional wind component.

The CaRT (Classification and Regression Tree) approach (Breiman, Friedman, Olshen, & Stone, 1984) is applied for selecting the most significant factors from those listed above. The CaRT procedure splits the data set stepwise into subsets along those factors that allow for the best MSE improvement in the corresponding piecewise constant approximation. Thus, splitting along some factor indicates that the tree model has some dependence on this factor. This factor should be selected to appear in the final predictive model.

In addition to the above-mentioned factors, we also include the wind speed for neighbor wind farms. It is proposed that this be done later for individual wind farm models, while tuning them. Applying CaRT with the neighbor factors at the very beginning may lead to uncertainty in the selection of significant factors, due to the strong correlations between some of them, such as the weather forecast data (zonal and meridional wind components, wind direction, and wind speed).

As part of introducing the CaRT procedure, we also define a stopping rule. We require the number of data points in the leaves to be not less than 500. This stopping rule prevents the selection of factors which influence only small portions of the data set (less than about 5% of the training sample).

Fig. 1 depicts the CaRT for wind farm 1.

Table 1 contains plus signs for the factors and wind farms that have been split during the regression tree construction, and have therefore proved their significance.

Table 1
Significant factors for wind farms, found using CaRT.

	1	2	3	4	5	6	7
Zonal wind component		+	+	+		+	+
Meridional wind component	+	+	+	+	+	+	+
Wind direction		+			+		+
Wind speed	+	+	+	+	+	+	+
Year					+		
Month							
Day of the month							
Hour	+	+	+	+	+	+	+
Day of the year	+	+			+	+	+

Any factors which were significant for five or more wind farms were chosen for inclusion in the final predictive model.

Let us denote the basic set of significant factors as follows: x^1 is the zonal wind component, x^2 is the meridional wind component, x^3 is the wind speed, x^4 is the hour, and x^5 is the day of the year. The wind speeds for neighboring wind farms become the complimentary factors for individual wind farm models. We add neighboring wind speeds into the model sequentially: the wind speed x^6 should provide the greatest improvement in model accuracy, and x^7 is chosen so as to proceed with the improvement.

In addition, let y denote the normalized wind power measurements and n the sample size.

3. Raw data pretreatment

Weather forecasts contain the most significant information for the model process, but they are stochastic. We propose the following procedures for weather forecast pretreatments:

(a) Algorithm of the simple moving average:

Step 1:

Assign

$$\widehat{x}^j = x^j = \left(x_1^j, x_2^j, \ldots, x_n^j\right), \quad j = 1, 2, 3, 6, 7. \tag{1}$$

Step 2:

$$x_p^j = \frac{\sum_{i=-t_1}^{t_2} \widehat{x}_{p+i}^j}{t_1 + t_2 + 1}, \quad j = 1, 2, 3, 6, 7, \quad p = t_1 + 1, t_1 + 2, \ldots, n - t_2, \tag{2}$$

where $t_1, t_2 \geq 0$, $[t_1, t_2]$ is the smoothing interval.

(b) Algorithm of the weighted moving average:

Step 1:

Assign

$$\widehat{x}^j = x^j = \left(x_1^j, x_2^j, \ldots, x_n^j\right), \quad j = 1, 2, 3, 6, 7. \tag{3}$$

Step 2:

$$x_p^j = \frac{\sum_{i=-t_1}^{t_2} \omega_{t_1+i+1} \widehat{x}_{p+i}^j}{\sum_{i=-t_1}^{t_2} \omega_{t_1+i+1}}, \quad j = 1, 2, 3, 6, 7, \quad p = t_1 + 1, t_1 + 2, \ldots, n - t_2, \tag{4}$$

where ω are weights.