Forecast of photovoltaic power generation based on DBSCAN

Jiaying Zhang¹, Yingfan Zhang^{a1}

¹Inner Mongolia University of Technology, Hohhot 010051, China

Abstract. The power output of the photovoltaic power generation has prominent intermittent fluctuation characteristics. Large-scale photovoltaic power generation access will bring a specific impact on the safe and stable operation of the power grid. With the increase in the proportion of renewable energy sources such as wind power and photovoltaics, the phenomenon of wind abandonment and light abandonment has further increased. The photovoltaic power generation prediction is one of the critical technologies to solve this problem. It is of outstanding academic and application value to research photovoltaic power generation prediction methods and systems. Therefore, accurately carrying out the power forecast of photovoltaic power plants has become a research hot point in recent years. It is favored by scholars at home and abroad. First, this paper builds a simulation model of the photovoltaic cell based on known theoretical knowledge. Then it uses the density clustering algorithm (DBSCAN) in the clustering algorithm and classifies the original data. Finally, according to a series of problems such as the slow modeling speed of photovoltaic short-term power prediction, the bidirectional LSTM photovoltaic power prediction model, and CNN-GRU photovoltaic power prediction model based on clustering algorithm are proposed. After comparing the two models, it is concluded that the bidirectional LSTM prediction model is more accurate.

1 Introduction

With the development of today's society, energy shortages and environmental pollution have become increasingly prominent global problems. The excessive development and utilization of traditional energy sources such as oil and coal have led to a series of problems such as deterioration of environmental quality, global warming, severe energy shortages, and even the outbreak of energy wars, which have become a focus of attention. In recent years, through modern control, artificial intelligence, and other theories, some more practical and convenient photovoltaic power prediction methods have been published.

The prediction efficiency has also been improved accordingly. Literature [1] The wavelet analysis method is used to decompose the NWP meteorological variables, and the decomposed sub-sequences are used to train the particle swarm to optimize the support vector machine. However, when using this prediction method to select various meteorological factors, the influence of the meteorological factors on the photovoltaic power plant's output power ignores the difference in impact. Literature [8] proposed convolutional neural networks, long-short-term memory networks, and hybrid models based on convolutional neural networks and long-short-term memory network models, apply them to the data obtained in DKASC, Alice Springs photovoltaic system.

The main research content of this article is based on the in-depth study of the principle of photovoltaic power prediction. It proposes a photovoltaic power prediction method based on the clustering algorithm. First, use the

clustering algorithm to classify the photovoltaic data according to their similarity. Then, complete the classification under the well-classified model: set up the network architecture, build a learning model, and train to predict the output power of photovoltaics. Finally, the results are compared according to national evaluation standards, and the optimal prediction model is obtained.

2 DBSCAN density clustering

DBSCAN is a density-based clustering algorithm. This type of density clustering algorithm generally assumes that the sample distribution's tightness can determine the category. The samples of the same category are closely connected between them. In other words, there must be samples of the same category not far from any sample of this category. By classifying closely connected samples into one category, a cluster category is obtained. By dividing all closely connected samples into different categories, we get all clustering categories' final results.

Assuming my sample set is $D = (x_1, x_2, \dots, x_m)$ the specific density description of DBSCAN is defined as follows:

(1) ϵ -Neighborhood: For $x_j \in D$, Its ϵ -Neighborhood includes the sub-sample set whose distance from x_j is not greater than ϵ in the sample set D , that is $N_\epsilon(x_j) = \{x_i \in D \mid \text{distance}(x_i, x_j) \leq \epsilon\}$, the number of this sub-sample set is denoted as $|N_\epsilon(x_j)|$.

^a Corresponding author: 1195480936@qq.com

(2)Core object: For any sample $x_j \in D$, if its ϵ -Neighborhood corresponding $N_\epsilon(x_j)$ contains at least MinPts samples. That is, if $|N_\epsilon(x_j)| \geq \text{MinPts}$, then x_j is the core object.

(3)Direct density: If x_i is located in the ϵ -Neighborhood of x_j , and x_j is the core object, x_i is directly reached by the density of x_j . Note that the opposite is not necessarily true; that is, it cannot be said that x_j is directly reached by the density of x_i , unless x_i is also the core object.

(4)The density can reach: For x_i and x_j , if there are sample sequences $p_1, p_2, \dots, p_T$ satisfying $p_1 = x_i$, $p_T = x_j$, and p_{T+1} is directly reached by the density of p_T , then x_j is said to be reachable by the density of x_i . In other words, the density can be reached to meet the transitivity. At this time, the transmitted samples $p_1, p_2, \dots, p_{T-1}$ in the sequence are all core objects because only core objects can directly reach other samples' density.

(5)Density connected: For x_i and x_j , if there is a core object sample x_k , both x_i and x_j are reachable by the density of x_k , then x_i and x_j are called density connected. Note that the density connection relationship satisfies symmetry.

3 Power prediction model

3.1 Bidirectional LSTM

Bidirectional LSTM is an extension of traditional LSTM, improving the model performance of sequence classification problems. In the problem where all the time steps of the input sequence are available, the bidirectional LSTM trains two LSTMs on the input sequence instead of one. The first in the input sequence is as it is, and the second is an inverted copy of the input sequence. This can provide additional context for the network and can learn faster and even more fully.

Figure 1. Bi-LSTM neural network structure

mathematical expression:

$$h_t = f(\omega_1 x_t + \omega_2 h_{t-1}) \quad (1)$$

$$h'_t = f(\omega_3 x_t + \omega_5 h'_{t+1}) \quad (2)$$

$$o_t = g(\omega_4 h_t + \omega_6 h'_t) \quad (3)$$

The basic idea is to propose that each training sequence, forward and backward are two Recurrent Neural Networks (RNN), and these two are connected to an output layer. This structure provides complete past and future contextual information for each point in the output layer's input sequence. Six unique weights are reused at each time step; the six weights correspond to Input to the forward and backward hidden layers (w_1, w_3), the hidden layer to the hidden layer itself (w_2, w_5), and the forward and backward hidden layers to the output layer (w_4, w_6).

3.2 CNN-GRU

First, extract the preliminary feature representation through the convolutional layer model. Then use the GRU module to enhance and further optimize the weather data represented by the preliminary feature. Finally, generate the final deep feature representation in the hidden layer of the GRU, and input it into the activation function to optimize. The addition of the Dropout layer (dropout layer) can effectively reduce over-fitting and achieve the effect of regularization to a certain extent. The combined prediction model of the convolutional network and the gated recurrent network is shown in Figure 2:

Figure 2. CNN-GRU neural network structure

mathematical expression:

$$r_t = \sigma(\omega_r [h_{t-1}, x_t]) \quad (4)$$

$$z_t = \sigma(\omega_z [h_{t-1}, x_t]) \quad (5)$$

$$h_t = \tanh(\omega_h [r_t * h_{t-1}, x_t]) \quad (6)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * h_t \quad (7)$$

$$y_t = \sigma(\omega_o h_t) \quad (8)$$

For z_t used to control the extent to which the state information from the previous moment is brought into the current state; r_t Used to control how much information from the previous state is written to the current candidates; $[]$ and $*$ respectively represent the connection of two vectors and the multiplication

of matrices; y_i represents the final output value of the output layer from input to output; $\omega_1, \omega_2, \omega_h, \omega_o$ indicates the parameters to be learned, which need to be segmented during the training process.

4 Results & Discussion

Firstly, 15min is used as a prediction point. Secondly, seven feature vectors, including total solar irradiance, direct irradiance, diffuse radiation intensity, temperature, humidity, wind speed, and wind direction in the meteorological data are used as the input data the model. Finally, the photovoltaic power generation is used as the output data of the model.

4.1 Evaluation Index

Adopt the evaluation criteria specially set by the National Energy Administration for photovoltaic power stations to determine the most accurate prediction model. Qualification rate (Q), it is stipulated that the percentage of the forecast deviation in the installed capacity at a particular time is considered as qualified within 25%, the expression is:

$$Q = \frac{1}{n} \sum_{i=1}^n B_i \times 100 \% \quad (9)$$

$$B_i = \begin{cases} 1, (1 - \frac{|y_i - f_i|}{c_i}) \geq 0.75 \\ 0, (1 - \frac{|y_i - f_i|}{c_i}) < 0.75 \end{cases}$$

Mean Absolute Percentage Error(MSPE), also known as mean absolute percentage error, the expression is:

$$E_{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - f_i}{y_i} \right| \times 100 \% \quad (10)$$

Mean Accuracy(MA), the expression is:

$$MA = \left[1 - \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - f_i}{y_i} \right)^2} \right] \times 100 \% \quad (11)$$

4.2 Simulation result analysis

According to the actual local weather conditions in Inner Mongolia, the weather is divided into two categories (sunny, rain, and snow) with x1, x2, x3, x4, x5, x6, and x7, the seven variables represent the total irradiance and dispersion of the day to be predicted. Irradiance, direct irradiance, temperature, humidity, wind speed, wind direction. The classification results are shown in Figure 3 (where red is the sunny category, and the remaining colors are rain and snow categories). In this paper, after clustering the original data, the data of sunny days are grouped. All the data of sunny days are used to verify the above two algorithms.

Figure 3.DBsACN clustering result graph

4.2.1 Bi-LSTM prediction model

The output of two parallel LSTMs, one with forwarding processed input and one backward processed output. Take a recursive layer (for example, the first LSTM hidden layer) as a parameter and combine the forward and backward output before passing it forward to the next layer. (By default, the outputs are connected to provide two cups of output for the next layer.)

Table 1. Parameter configuration table

name	size
units	50
rule	sigmoid
loss	mae
optimizer	sgd、adam、nadam
epochs	500
batch_size	72
shuffle	False
verbose	2

This paper will classify the classified data of sunny days and obtain the overall data trend after 500 times of training. The real value and the predicted value have a slight error in the peak period. Some of the prediction results are exposed to the most strong sunlight at noon and afternoon.

Figure 4.DBsACN clustering result graph.

4.2.2 CNN-GRU prediction model

CNN has the characteristics of pooling operation, local connection, and weight sharing. The time-series data as the network's input effectively reduces the complexity of feature extraction and data reconstruction. The CNN model uses the gradient descent method to train

parameters, and the trained model can learn the features in the time series data. After 500 pieces of training, the overall trend chart is shown in Figure 5.

Figure 5.DBSACN clustering result graph.

It can be clearly seen from the above figure that the results obtained using the CNN prediction model have a large error between the actual value and the predicted value at the trough. That is to say, some prediction results are affected by irradiance fluctuations in the morning and night when the sun is weakest, and the prediction results in the rest of the period are more accurate.

4.3 Evaluation

To According to national evaluation indicators, comparing and analyzing the Qualification rate (Q), Mean Accuracy(MA), and Mean Absolute Percentage Error(MAPE), the specific values are as follows.

Table 2. Evaluation index table

Neural Networks	Pass rate	Accuracy	Relative error
Bi-LSTM	0.9973	0.9587	0.0018
CNN-GRU	0.9964	0.9402	0.0049

The pass rate is 0.9973 for Bi-LSTM and 0.9964 for CNN; the accuracy is 0.9587 for two-way LSTM and 0.9402 for CNN; the relative error is 0.0018 for Bi-LSTM and 0.0049 for CNN. Comprehensive comparative analysis can be obtained through the above content: the Bi-LSTM prediction model has better indicators than the CNN prediction model. For these two models, the Bi-LSTM prediction model is the optimal prediction model.

5 Conclusions

The data source for this subject is China Electric Power Station, and the data is collected every 15 minutes. More than 30,000 data a year guarantee the accuracy of the algorithm. The density-based clustering method uses local data characteristics as the criterion for clustering. Clustering is based on the density and sparseness of data objects in the area. The data is suitable for forecasting and has great value for improving photovoltaic power generation forecasting accuracy. Using a cluster analysis algorithm to process photovoltaic data can get more

accurate data for model building and algorithm realization.

The improved deep learning algorithm used in this article can save training time and improve prediction accuracy. Although the combined model cannot guarantee that each prediction result is better than all other models, it makes its results as close as possible to the best accuracy model. Therefore, the combined forecasting model has better stability and avoids a single model's low flexibility. According to the prediction results, the two-way LSTM prediction model is better than the prediction model combined with CNN and GRU.

References

1. EseyeAT, ZhangJH, ZhengDH. Short-term photovoltaic solar power forecasting using a hybrid Wavelet-PSO-SVM model based on SCADA and Meteorological information[J].Renewable energy, 2018,118(4):357-367.
2. Sun Hui, Leng Jianwei. Short-term photovoltaic power prediction based on PSO-DBN combined with chaotic vertical and horizontal crossover[J/OL].Electrical measurement and instrumentation,2019,9(05):1-7.
3. Jing Bo, Tan Lunnong, Qian Zheng, et al. Overview of research progress in short-term forecasting of photovoltaic power generation [J].Electrical measurement and instrumentation,2017,8(12):1-6.
4. Li Lianbing, Zhang Jia, Han Jingnan, etc. Research on short-term power prediction of photovoltaic array based on the Elman algorithm[J]. Acta Solar Energy,2017,13(06):1560-1566.
5. Ge Le, Lu Wenwei, Yuan Xiaodong, etc. Power prediction of photovoltaic power station based on improved similar days and ABC-SVM[J]. Acta Solar Energy,2018,22(03):775-782.