

Research on Short-term Photovoltaic Power Prediction Algorithm Based on Spark and Optimized RBFNN

Baoyi Wang^a, Zhen Hao^b, Shaomin Zhang^c

School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, Hebei Province, China

^aemail:wangbaoyiqj@126.com, ^bemail:zhenhao867@126.com, ^cemail:zhangshaomin@qq.com

Keywords: photovoltaic power forecasting; RBFNN; Spark; cluster of similar days

Abstract. To overcome the problem of poor accuracy of short-term photovoltaic power prediction, This paper proposes a short-term PV power prediction algorithm based on radial basis function neural network(RBFNN) after similar day clustering on solar irradiance and air temperature as the input variables; At the same time, it introduces the particle swarm optimization(PSO) algorithm to optimize the kernel function parameters of the neural network. In view of the time-consume problem of the large amount of historical data in the photovoltaic power station, Spark cloud platform which based on memory is used to realize parallel the processing algorithm. Through an example experiment, it is verified that the proposed model improves the prediction accuracy, and the parallel algorithm greatly reduces the computation time.

Introduction

With the change of global energy application, renewable clean energy (wind energy, solar energy, etc.) is being paid more attention by people. In recent years, countries continue to invest in new energy. Photovoltaic power generation has moved from early distributed off-grid mode to large grid-connected power plant models^[1]. Photovoltaic power generation is affected by factors such as irradiance, weather, time. It has the characteristics of intermittent, volatility and periodicity. Research shows that large-scale PV access will widen the peak valley gap, causes difficulty in peak shaving, which affects the safe and stable operation of power quality and power grid^[2]. Therefore, it is very important to predict the output of photovoltaic power accurately and quickly.

At present, the artificial neural network has strong robustness, memory capacity, nonlinear mapping ability and strong self-learning ability, which is widely used in photovoltaic power prediction^[3-9]. The literature [3] uses BP neural network to predict the next day's photovoltaic power generation. The literature [4] designs GA-BP and POS-BP neural network to predict short-term photovoltaic power output. The literature [5] designs similar days clustering and proposes a photovoltaic power prediction model based on Bayesian neural network. The literature [6] uses wavelet transform for historical power and combines with RBFNN to construct the network prediction model. All of the above algorithms improve the prediction accuracy, but parameters of neural networks are often lack of optimization. At the same time, photovoltaic power station has accumulated a lot of historical data after many years of operation. In order to meet the real-time requirement of power system, the computation time of a large number of data needs to be solved urgently.

Spark cloud computing technology is an open source project under the Apache foundation, which has been highly concerned and applied in the field of big data. It is based on the Hadoop Distributed File System(HDFS) and Yarn resource management framework, realizes the distributed storage and management for massive data. It also provides a highly restricted shared memory model using the flexible distributed data set, which can effectively solve the time-consume problem of data computation.

This paper first designs similar day clustering for samples based on irradiance, temperature. Then, it optimizes the kernel function parameters of the RBFNN network with PSO algorithm and establishes a 24h photovoltaic power prediction model. The Spark platform is used to realize the

parallel improvement for the time-consume problem of the prediction algorithm. Finally, the effectiveness of the model is verified by experiments on a single machine and Spark cloud platform.

Similarity Day Clustering For Samples

There is a mapping relationship between variables and PV output^[10]. The similar days can be find from the historical data to predict the PV output. The output of the photovoltaic power is mainly affected by the irradiance and temperature factors. So we use temperature sequence $t = [t_1, t_2 \dots t_i]$ and the irradiance sequence $g = [g_1, g_2 \dots g_i]$ to construct a sample vector sequence $s = [t, g]$ and use the fuzzy C means algorithm (FCM) to cluster the samples.

The objective function of clustering is:

$$J(U, C) = \sum_{k=1}^n \sum_{j=1}^q (u_{jk})^m d^2(s_k, c_j) \tag{1}$$

$$\sum_{j=1}^c u_{jk} = 1, k = 1, 2, \dots, n$$

Here, $m \in [1, \infty)$ is weighted coefficient, $u_{jk} \in [0, 1]$ is membership between s_k and c_j , $c_j = [c_{j1}, c_{j2} \dots c_{ji}]$ is a cluster center, $d_{jk} \equiv d(s_k, c_j) \equiv \sqrt{(s_{k1} - c_{j1})^2 + \dots + (s_{ki} - c_{ji})^2}$, n is the number of samples.

FCM updates u_{jk} and c_j based on formula (2) (3) and minimize the value of the objective function through multiple iterations.

$$u_{jk} = \frac{1}{\sum_{l=1}^c (d_{jk} / d_{lk})^{2/(m-1)}} \tag{2}$$

$$c_j = \frac{\sum_{k=1}^n (u_{jk})^m P_k}{\sum_{k=1}^n (u_{jk})^m} \tag{3}$$

When the number of iterations reaches the preset limit or the difference of the objective function is convergent, the clustering is stopped, and the cluster samples and the cluster centers are obtained according to the membership degree matrix.

Radial Basis Function Neural Network

Radial basis function neural network (RBFNN) is a local approximation of the feed-forward neural network, with 3 layer structure (Figure 1): the input layer, the hidden layer, the output layer. The input layer consists of a plurality of sensing neurons and connects with the external environment; the hidden layer performs a nonlinear mapping for feature extraction, and then the effect function is used to generate a response to the input signal; the output of the network is linear. Compared with the traditional BP neural network, the number of the hidden layer nodes of RBFNN can be determined according to the needs. To some extent, it overcomes the shortcomings of slow convergence speed and local extremum of BP neural network.

The neural network inputs is $x = [x_1, x_2 \dots x_n] = s$; $y = [y_1, y_2 \dots y_i] = p$ is the output of the network. If the Gauss activation function is used in the hidden layer nodes; the output of the hidden layer node m is expressed as h_m .

$$h_m = \exp\left[-\frac{(x - \theta_m)^T (x - \theta_m)}{\delta_m^2}\right] \tag{4}$$

$$H = [h_1, h_2 \dots h_m] \tag{5}$$

The mapping from the hidden layer to the output layer is y .

$$y = WH + b \tag{6}$$

θ_m and δ_m are the data center vector and extension constant of the Gauss kernel function of node m respectively.

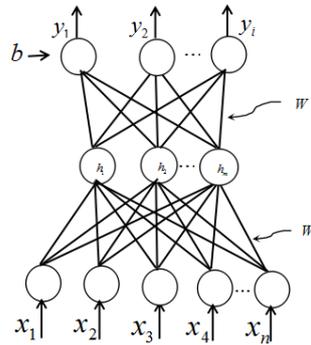


Figure 1 RBFNN structure

Optimized RBFNN Prediction Model

In order to improve the prediction accuracy, the parameters of RBFNN model need to be optimized. First we design the structure of RBFNN:

A. the input layer nodes: The number of nodes in the input layer is associated with the input variables. The input variables are temperature, irradiance.

B. the hidden layer nodes: The number of hidden nodes need to be determined according to the actual situation. If the network nodes use the Gauss kernel function, the initial value of the kernel function data center will be set by the cluster centers of the similar days. Expansion constant is the maximum difference between data centers.

C. the output layer node: The output layer corresponds to the power output of the forecast day.

The data center and the extended constants of Gauss kernel function of traditional RBFNN are confirmed usually, will not change, which is not suitable for incremental change of dynamic system^[11,12]. In this paper, the particle swarm optimization algorithm is introduced to optimize the parameters of the vector which is composed of the data center and the expansion constant.

Particle swarm optimization (PSO) is a typical swarm intelligence optimization algorithm. The particle has position vector and velocity vector, and the different particles have the individual fitness related to the objective function.

The root mean square error (RMSE) is used as the objective function of the particle optimization.

$$f(L, p) = RMSE = \frac{1}{L} \sum_{l=1}^L \sqrt{[\sum_{i=1}^I (p'_{li} - p_{li})^2] / I} \quad (7)$$

p'_{li} is the output power predictive value, p_{li} is the actual value of power, I is the number of power acquisition time of the day, L is the number of training samples.

The implementation process of similar days clustering and optimization of RBFNN prediction algorithm is as follows:

Step1: All the data are normalized, and the samples are randomly divided into training samples and test samples; the number of training samples is far greater than the number of test samples.

Step2: Try to carry out a number of similar days clustering for all samples. Determine the best number of clusters, and get the final cluster and the cluster center.

Step3: Initialize multiple particles by clustering samples and cluster centers.

Step4: Calculate the fitness for each particle, and update the local optimal position of each particle and the global best position.

Step5: Update a generation of individual particles. If $\|f^{(t+1)} - f^t\| < \varepsilon$ or the maximum number of iterations is met, the algorithm ends; otherwise, turn step (4).

Step6: Test the predictive performance of the network model by the test samples.

In this paper, the least square method is used to determine the network connection weights between the hidden layer and the output layer. The specific algorithm can be find in literature [14].

Parallel Implementation Of Algorithm Under Spark Platform

Although the algorithm has good prediction effect after parameter optimization, but due to the photovoltaic power plant has a various types of photovoltaic panels, and the meteorological data acquisition scale is small, the amount is large, which caused the calculation is more time consuming. In this ,we use Spark^[13] to process the algorithm in parallel.

Algorithm implementation steps in Spark platform are as follows:

Step1: Get the PV sample data from the HDFS on the Spark platform; translate it into RDD and load it into memory; perform similar day clustering on a single spark node.

Step2: Obtain the clustering and cluster centers and constitute a number of particles. The particles are distributed to a number of Spark nodes. Meanwhile, samples are divided into training set and test set, each node has a copy of training samples for training.

Step3: Calculate the fitness of the objective function for particles at each node with the training set. The master node coordinates the optimization strategy of each particle according to the fitness value.

Step4: When the difference of the global minimum of the particle obtained by the two iteration is convergent or the maximum number of iterations is reached, the algorithm terminates and the best particle is returned by the Spark's actions operator.

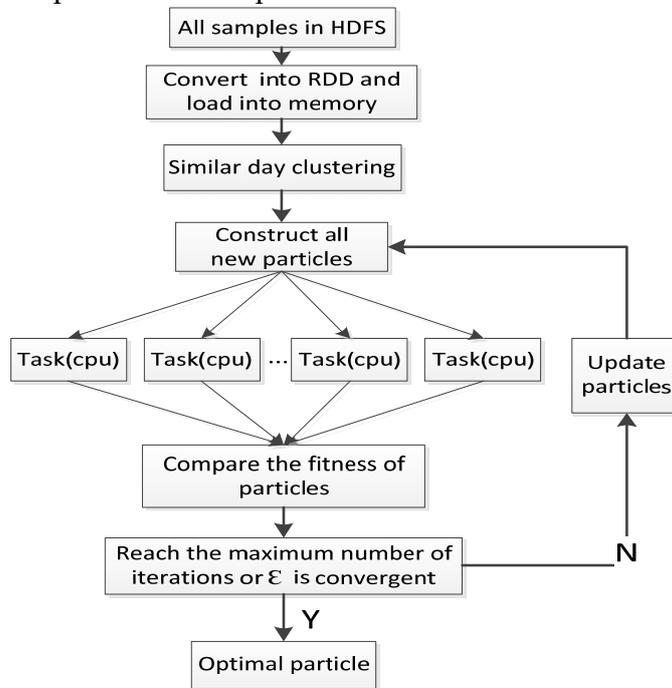


Figure 2 Samples clustering and parallel optimization of RBFNN parameters on Spark platform

Experiment And Result Analysis

With the above method, select a photovoltaic power station 2014-10, 2015-09 real data as the sample data. They includes Power(kW), irradiance(W/m2), temperature(°C) data; data sampling time interval is 5min. Forecast the power of each day in February 2016.

All data are normalized to eliminate influence of dimension. The formula is shown as the following.

$$G'_i = \frac{G_i - G_{\min}}{G_{\max} - G_{\min}} \tag{8}$$

Here, G_i is the original data; G_{\min} is the minimum value in the original data; G_{\max} is the maximum in the original data; G'_i is the corresponding data after processing.

All samples are clustered and are divided into training set and test set. Training set train neural

network, test set test network prediction error.

Forecast the power output of all the forecast-days. And the RMSE value is the index to evaluate the model. The lower the RMSE value is, the higher the accuracy of prediction is.

From the table 1, it can be seen that the proposed model has a better forecasting results.

Table 1 RMSE mean contrast

Prediction algorithm	RMSE
RBFNN ^[9]	0.3592
RBFNN optimized by PSO	0.2535

The Spark cloud computing platform in the laboratory is composed of 20 nodes, each node is configured as Intel (R) Core (TM) i5-2400, 2 GB RAM. Hadoop version is 2.6.4. Spark version is spark-1.5.0-bin-hadoop2.6.

The algorithm uses the same data to run on 1 node and 5 node. Then we compare their time-consuming.

From the table 2, it can be seen that the time difference between parallel on one node and stand-alone is small. However, the computation time is significantly reduced on multiple node.

Table 2 Time contrast after parallelization

Types	Non-parallel	Spark-parallel	
Node	stand-alone	1 node	5 nodes
Node	337	310	98

Speedup is used to test the parallel performance of the algorithm. The algorithm is run in 5, 10, 20 nodes with different quantity of data.

$$R_{\text{speedup}} = T_s / T_c \tag{9}$$

T_s is the time-consuming by the stand-alone operation, T_c is the time-consuming by the Spark cloud platform.

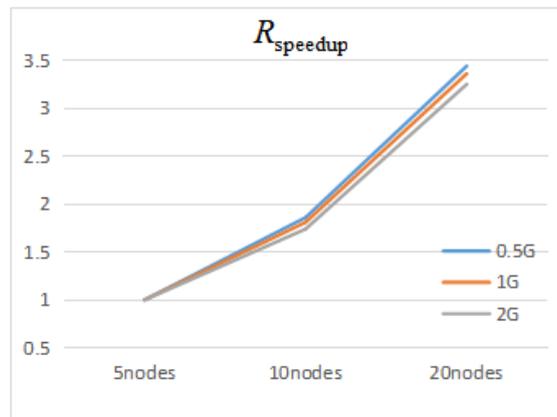


Figure 3 Speedup ratio

From Figure 3, it can be seen that the model has better performance after parallelization on the Spark platform.

Conclusion

Photovoltaic industry has developed rapidly. The rapid and accurate PV output prediction can effectively reduce its impact on the large power grid. In order to improve the prediction accuracy, this paper establishes a 24h photovoltaic power prediction model based on the radial basis function neural network, which is optimized by the particle swarm optimization algorithm. At the same time, the Spark cloud platform is used to parallel the proposed algorithm, which solves the time-consume problem of large amount of data. The validity and practicability of the model are verified by experiments. In the latter part of the work, the accuracy of prediction will be further studied.

Acknowledgement

In this paper, the research was sponsored by the National Natural Science Foundation of China (Project No. 61300040) and Scientific Research Project of Hebei Province (Project No. Z2012077).

References

- [1] GONG Yingfei, LU Zongxiang, QIAO Ying, et al. An Overview of Photovoltaic Energy System Output Forecasting Technology[J]. *Automation of Electric Power Systems*, 2016,40(4): 140-151.
- [2] DING Ming, WANG Weisheng, WANG Xiuli, et al. A Review on the Effect of Large-scale PV Generation on Power Systems[J]. *Proceedings of the CSEE*, 2014,34(1):1-14.
- [3] Chen Changsong, Duan Shanxu, Yin Jinjun. Design of Photovoltaic Array Power Forecasting Model Based on Neutral Network[J]. *Transactions of China Electrotechnical Society*, 2009,24(9): 153-158.
- [4] YAO Zhongmin, PAN Fei, SHEN Yuhui, et al. Short-term prediction of photovoltaic power generation output based on GA-BP and POS-BP neural network[J]. *Power System Protection and Control*, 2015,43(20):83-89.
- [5] JI Ling, NIU Dong-xiao, WANG Peng. Photovoltaic Load Forecasting Based on the Similar Day and Bayesian Neural Network[J]. *Chinese Journal of Management Science*, 2015, 23(3): 118-122.
- [6] Zhang Liying, Liu Zhiyu, Meng Lingjia, et al. Photovoltaic output power prediction approach based on wavelet transform and neural network[J]. *Renewable Energy Resources*, 2015,33(2): 171-176.
- [7] LUO Jian-chun, CHAO Qin, LUO Hong, et al. PV short-term output forecasting based on LVQ-GA-BP neural network[J]. *Power System Protection and Control*, 2014,42(13):89-94.
- [8] LI Yuancheng, WANG Bei, WANG Xufeng. Photovoltaic power forecasting based on harmony search and Gaussian process algorithms[J]. *Electric Power Automation Equipment*, 2014, 34(8):13-18.
- [9] WANG Xiaolan, GE Pengjiang. PV array output power forecasting based on similar day and RBFNN[J]. *Electric Power Automation Equipment*, 2013,33(1):100-103+109.
- [10] MO Weiren, ZHANG Boming, SUN Hongbin, et al. Method to select similar days for short-term load forecasting [J]. *Journal of Tsinghua University: Science and Technology*, 2004, 44(1): 106-109.
- [11] LIU Zixi, ZHANG Hang. Optimal design of RBF NN based on QPSO algorithm[J]. *Journal of Central South University (Science and Technology)*, 2013. 44(S1):27-30.
- [12] YE Lin, CHEN Zheng, ZHAO Yongning, et al. Photovoltaic Power Forecasting Model Based on Genetic Algorithm and Fuzzy Radial Basis Function Neural Network[J]. *Automation of Electric Power Systems*, 2015,39(16):16-22.
- [13] WANG Baoyi, WANG Dongyang, ZHANG Shaomin. Distributed short-term load forecasting algorithm based on Spark and IPPSO_LSSVM[J]. *Electric Power Automation Equipment*, 2016,36(1):117-122.
- [14] MOODY J, DARKEN C J. Fast learning in networks of locally tuned processing units[J]. *Neural Computation*, 1989,1(2):281-294.