Environmental Parameters Analysis and Power Prediction for Photovoltaic Power Generation Based on Ensembles of Decision Trees

Shuai Zhang [1,2,3], Hongwei Dai[1], AizhouYang [1] and Zhongzhi Shi [2]

[1] Dongfang Electronics Co. Ltd., Yantai 264000, China
[2] Chinese Academy of Sciences, Beijing 110010, China
[3] Shandong Technology and Business University,Yantai 264005, China

Abstract. Due to the influence of solar irradiation, temperature and other environmental factors, the output power of photovoltaic power generation has great randomness and randomness discontinuity. In this paper, a method for analyzing environment data related photovoltaic power generation based on ensembles of decision trees algorithm is studied. Firstly, the characteristics of environmental factors of photovoltaic power generation are analyzed by K-means clustering. And then the corresponding cluster label is assigned. Furthermore, the Radom Forests is combined to build a model. Finally, the method is validated by given data above from a real project. The results show that the proposed method can provide reference for the forecasting of photovoltaic power.

Keywords: K-means Clustering, Ensembles of Decision Trees, Photovoltaic Power Generation, Environmental Data, Feature Analysis and Prediction

1 Introduction

With the replace old growth drivers with new ones strategy being implemented [1-2], new energy industry has become a strategic and pioneering one in the world. Photovoltaic power generation has been developing rapidly due to its advantages in safety, reliability, less geographic restrictions and short construction period [3]. At present, photovoltaic generation related technology have matured, but the output power has great randomness and randomness discontinuity by the influence of solar irradiation, temperature and other environmental factors [4-5]. Therefore, it is difficult to integrate into the power grid, and which is also disadvantage to rational planning for using of local energy. How to applying environmental data to analysis and predict solar power generation will become a major energy in the future [6-7].

The development of artificial intelligence, big data, data mining and other new generation of information technology have provided a good solution for solving the problem[8-9]. The reference [10] applied the K-means clustering to the actual operation data processing of a PV station in the city of Foshan, Guangdong Province, and achieved the operation state pattern recognition. A short-term forecasting method for photovol-

taic（PV）power is proposed in the reference [11], which established an SVM forecasting model and uses leave-one-out algorithm to optimize the Kernel parameter and penalty parameter to achieve the forecasting of PV power. The reference [12] proposed a novel model called forest for photovoltaic power generation (FPPG), which is an assembly predict model composed by multi regression tree and can perform better generated in power forecasting. However, photovoltaic power generation is influenced by environmental factors greatly, such as temperature, humidity, irradiation and so on, which are varies dramatically from one region to another. So how to make better those data and improve data quality astill a hot topic.

In this paper, an approach which combined K-means clustering with random forests, an ensemble of decision trees method, are researched and analyzed by some measured data from a photovoltaic power station in Shandong province. Firstly, some environmental data related photovoltaic power generation are analyzed. At the same time, measured data from a photovoltaic power station are given. And then, the algorithm principle of K-means clustering and random forests are introduced briefly. Moreover, an algorithm which combined K-means clustering with random forests and its flow chart are proposed. Finally, the method is validated by given data above from a real project. The results show that the proposed method can provide reference for the forecasting of PV power.

2. Some environmental data related photovoltaic power generation

The operating state of the photovoltaic power generation system is not only related to the working state of the system internal components, but also related to the changes of environmental parameters. There are many factors influencing the output power of photovoltaic system. Furthermore, these factors also have a complex relationship with the output power of photovoltaic power generation, especially the environmental data, mainly including temperature, humidity, wind speed, air pressure, irradiation and so on, which are the objective factors beyond human controlling but the key. Therefore, this method has an experimental application and a research meaning to analyze the environmental characteristics and find out the relationship between each factor and actual power output for further prediction.

In this paper, the measured data samples of a 40MW photovoltaic power station system in Shandong province are given. The samples including the wind speed, wind direction, temperature, humidity, air pressure, irradiation factor and power at the same sampling time. The sampling data are sampled at a 15-minute interval, 96 sampling points per day. A total of 7,488 samples were collected from 78 consecutive days. Part of the samples are shown in Fig. 1.

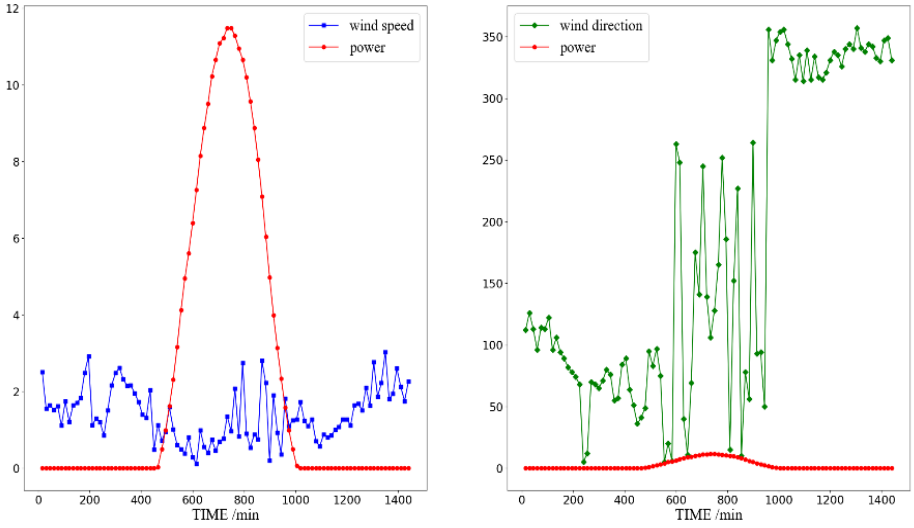
Where 'SDATE' stands for date marking, 'TIME' stands for sampling time, 'SYCG_SUMP' stands for power and 'QX000'-'QX005' stands for: wind speed, wind direction, temperature, humidity, air pressure and irradiation，respectively.

	SDATE	TIME	QX000	QX001	...	QX003	QX004	QX005	SYCG_SUMP
0	17532	15	2.5100	112	...	0.0919	1021.679993	5.4794	0.0
1	17532	30	1.5599	126	...	0.0929	1021.679993	6.3926	0.0
2	17532	45	1.6400	113	...	0.0919	1021.690002	5.4794	0.0
3	17532	60	1.5200	96	...	0.0909	1021.570007	6.3926	0.0
4	17532	75	1.6200	114	...	0.0889	1021.479980	6.3926	0.0
...	...	...	...	...	...	...	...	...	...
7483	17614	1380	1.1699	178	...	0.1509	1018.219971	0.9132	0.0
7484	17614	1395	2.0999	168	...	0.1500	1018.210022	0.0000	0.0
7485	17614	1410	2.2999	159	...	0.1489	1018.219971	0.0000	0.0
7486	17614	1425	2.0499	175	...	0.1489	1018.200012	0.0000	0.0
7487	17614	1440	2.2199	183	...	0.1469	1018.099976	0.0000	0.0

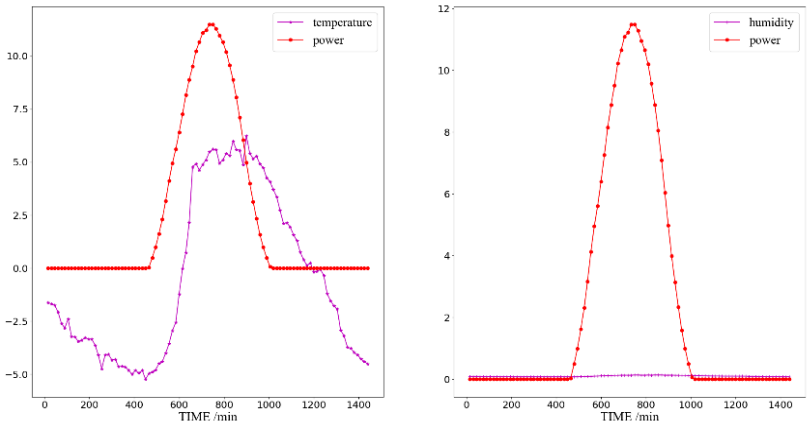
[7488 rows x 9 columns]

Fig. 1. Part of the samples

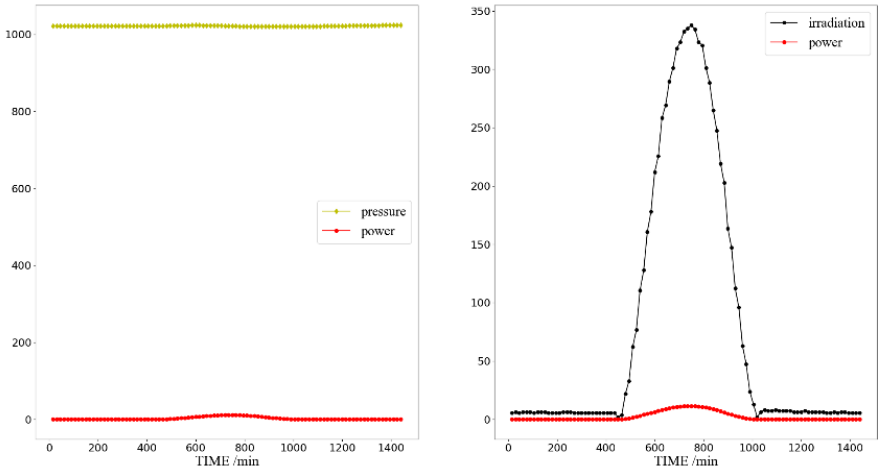
The relationship between environmental factors and power are visualized by using line graphs. For the clarity of the data presentation, take only diurnal variation as an example, which are shown in Fig. 2(a)-(c).



(a) The relationship between wind speed/wind direction and power



(b) The relationship between temperature/humidity and power



(c) The relationship between air pressure /irradiation and power

Fig. 2(a)-(c). Various environmental factors and power line chart / day

From Fig.2, we can see that power output has obvious correlation with temperature, irradiation and humidity, but poor correlation with wind speed, wind direction and air pressure, which can apply to preprocessed for subsequent cluster analysis to accurately cluster data ranges.

3. The Algorithm Principle of K-means Clustering and Random Forests

3.1 K-means Clustering [14]

The goal of cluster analysis is to partition the observations into groups ("clusters") so that the pairwise dissimilarities between those assigned to the same cluster tend to be smaller than those in different clusters. The *K*-means algorithm is one of the most popular iterative descent clustering methods. It is intended for situations in which all variables are of the quantitative type, and squared Euclidean distance equation (1) is chosen as the dissimilarity measure.

$$d(x_i, xi') = \sum_{j=1}^{p}(x_{ij} - x_{i'j}) = \|x_i - x_{x'}\| \tag{1}$$

Note that weighted Euclidean distance can be used by redefining the x_{ij} values.

The within-point scatter can be written as

$$\mathrm{W}(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k}\|x_i - x_{i'}\|^2 = \sum_{k=1}^{K} N_k \sum_{C(i)=k}\|x_i - \bar{x}_k\|^2 \tag{2}$$

Where $\bar{x}_k = (\bar{x}_{1k}, \cdots, \bar{x}_{pk})$ is the mean vector associated with the kth cluster, and $N_k = \sum_{i=1}^{N} I(C(i) = k)$. Thus, the criterion is minimized by assigning the $\boldsymbol{N}$ observations to the *K* clusters in such a way that within each cluster the average dissimilarity of the observations from the cluster mean, as defined by the points in that cluster, is minimized.

3.2 Ensembles of Decision Trees-- Random Forests [13, 14]

Decision trees are a widely used models for some machine learning tasks. But a main drawback of decision trees is that they tend to overfit the training data. Although there are some measures for prevent it, such as pre-pruning and post-pruning, even with the use of pre-pruning, decision trees tend to overfit, and provide poor generalization performance.

Ensembles are methods that combine multiple machine learning models to create more powerful models, which are one way to address this problem. Random Forests are the model that belong to this category. Random forests are essentially a collection of decision trees, where each tree is slightly different from the others. To implement this strategy, we need to build many decision tree. The specific algorithm is as follow table 1:

Firstly**,** we first take what is called a bootstrap sample of our data. A bootstrap sample means from our n_samples data points, we repeatedly draw an example randomly with replacement (i.e. the same sample can be picked multiple times), n_samples times. This will create a dataset that is as big as the original dataset, but some data points will be missing from it, and some will be repeated.

Next, a decision tree is built based on this newly created dataset. The bootstrap sampling leads to each decision tree in the random forest being built on a slightly different

dataset. Because of the selection of features in each node, each split in each tree operates on a different subset of features. Together these two mechanisms ensure that all the trees in the random forests are different.

Table 1. The algorithm of random forests [14]

Random Forest
1. For b = 1 to B: (a) Draw a bootstrap sample Z^* of size N from the training data. (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size nmin is reached. i. Select m variables at random from the p variables. ii. Pick the best variable/split-point among the m. iii. Split the node into two daughter nodes. 2. Output the ensemble of trees $\{T_b\}_1^B$. To make a prediction at a new point x: Regression: $\hat{f}_{rf}^B(x)=\frac{1}{B}\sum_{b=1}^B T_b(x)$ Classification: Let $\hat{C}_b(x)$be the class prediction of the bth random-forest tree. Then $\hat{C}_{rf}^B(x)$ = majority vote $\{\hat{C}_b(x)\}_1^B$

4. The Algorithm Combined K-means Clustering with Random Forests and Engineering Testing

4.1 The Algorithm Combined K-means Clustering with Random Forests

In Random Forests Algorithm，critical parameter in this process is max_features. If we set max_features to n_features, that means that each split can look at all features in the dataset, and no randomness will be injected. If we set max_features to one, that means that the splits have no choice at all on which feature to test, and can only search over different thresholds for the feature that was selected randomly[13]. So, we applied K-means Clustering to preprocessing the data, thus, we can obtain a cluster label for every sampling point. Next, we let this label as the target function to establish a random forest to achieve forecast. The specific algorithm flow is shown in Fig. 3.

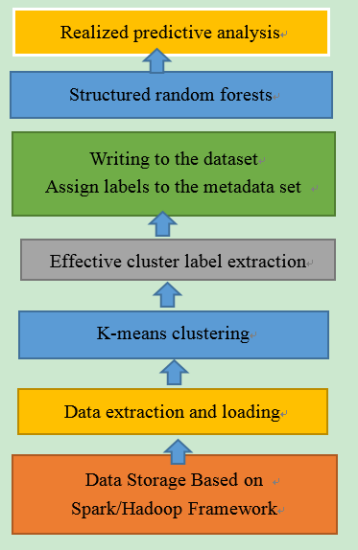
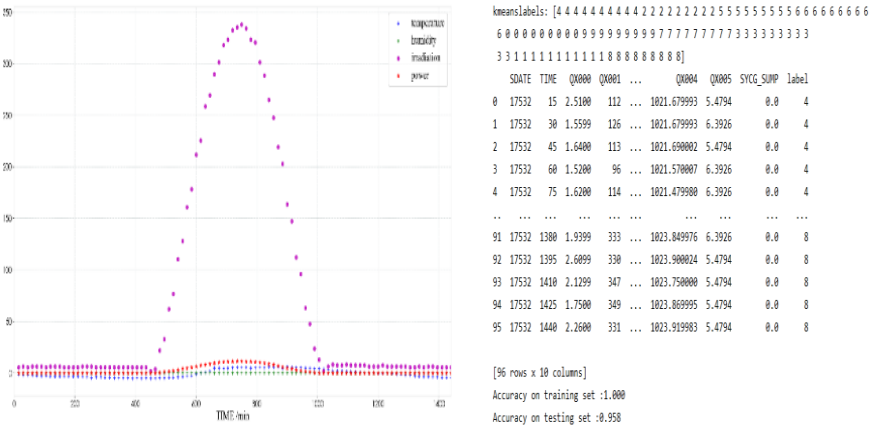


Fig. 3. Algorithm flow

4.2 Engineering Testing

Based on Spark/Hadoop framework and python language, the data of the second part above are applied and analyzed. Due to limited space, only 1 day's data sampling point is taken as a legend and 7 days' data analysis is taken as a table example in Fig.4 and Table 2. It can be seen that K-means clustering has well realized self-analysis of data, and given cluster labels. In the last, a stable prediction accuracy is obtained.



kmeanslabels: [4 4 4 4 4 4 4 4 4 4 2 2 2 2 2 2 2 2 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6
6 0 0 0 0 0 0 0 0 0 9 9 9 9 9 9 9 9 7 7 7 7 7 7 7 7 3 3 3 3 3 3 3 3 3
3 3 1 1 1 1 1 1 1 1 1 8 8 8 8 8 8 8 8 8]

	SDATE	TIME	QX000	QX001	...	QX004	QX005	SYCG_SUMP	label
0	17532	15	2.5100	112	...	1021.679993	5.4794	0.0	4
1	17532	30	1.5599	126	...	1021.679993	6.3926	0.0	4
2	17532	45	1.6400	113	...	1021.690002	5.4794	0.0	4
3	17532	60	1.5200	96	...	1021.570007	6.3926	0.0	4
4	17532	75	1.6200	114	...	1021.479980	6.3926	0.0	4
..	...	...	...	...	...	...	...	...	...
91	17532	1380	1.9399	333	...	1023.849976	6.3926	0.0	8
92	17532	1395	2.6099	330	...	1023.900024	5.4794	0.0	8
93	17532	1410	2.1299	347	...	1023.750000	5.4794	0.0	8
94	17532	1425	1.7500	349	...	1023.869995	5.4794	0.0	8
95	17532	1440	2.2600	331	...	1023.919983	5.4794	0.0	8

[96 rows x 10 columns]
Accuracy on training set :1.000
Accuracy on testing set :0.958

Fig. 4. The result of 1day's data sampling point by using K-means Clustering -Random Forests

Table 2. Forecast results for 7 days

Sdate	n_clus-ters	Accuracy on training set	Accuracy on testing set
17532	10	1.000	0.958
17533	10	0.986	0.917
17534	10	1.000	0.958
17535	10	1.000	0.917
17536	10	1.000	0.958
17537	10	1.000	1.000
17538	10	1.000	0.958

5 Conclusions

One method based on ensembles of decision trees is studied in this paper. By K-means clustering, the characteristics of environmental factors of photovoltaic power generation are analyzed, and data related to the power output through visualization are found out, and simultaneously, the corresponding cluster label is assigned. After that, the Radom Forests is combined to build a model, and then power generation environment data and predict the power are analyzed. The results show that this method has stable prediction accuracy and certain reference value for the environment data analysis and power prediction of photovoltaic power generation.

References

1. The Approval about the Overall Plan of Shandong the Replace Old Growth Drivers with New One Comprehensive Reform Pilot Area by the State Council, http://www.gov.cn /zhengce/ content/2018-01/10/content_5255214.htm.
2. People's Government of Shandong Province: Implementation Planning for the Key Projects of the Replace Old Growth Drivers with New Ones in Shandong Province, http://www.shandong. gov.cn/art/ 2018/ 3/ 16/art_ 2522_11096. html.
3. China Electricity Council: China electric power industry annual development report. China Electric Power Press, Beijing(2018).
4. Bird, L., Lew, D., Milligan, M.: Wind and solar energy curtailment：A review of international experience. Re‐newable and Sustainable Energy Reviews 65(577–586) 2016.
5. Zhang, C.Q., Zheng, Q.: SKBA-LSSVM Short-term Forecasting Model for PV Power Generation. Proceedings of the CSU-EPSA 31(8), 86–93 (2019).
6. Wu, J.Z.: Drivers and state-of-the-art of integrated energy systems in Europe. Automation of Electric Power Systems 40(5) , 1–7 (2016).
7. Ai, X., Han, X.N., Sun, Y.Y.: The Development Status and Prospect of Grid-connected Photovoltaic Generation and Its Related Technologies. Modern Electric Power 30(1), 1–7 (2013).
8. He, Q., Li, N., Luo, W.J., Shi, Z.Z.: A Survey of Machine Learning Algorithms for Big Data. Pattern Recognition and Artificial Intelligence 27(4), 327−335 (2014).

9. Hu, K.Y., Li, Y.L., Jiang, X., Li, J., Hu, Z.H.: Application of Improved Neural Network Model in Photovoltaic Power Generation Prediction. Computer Systems & Applications 28(12), 37−46 (2019).
10. Yang, D.Y., Ge, Q., Dong, Y.C., Tang, Y.L., He, C.X.: Research on operation state pattern recognition of PV station based on the principle of K-means clustering. Power System Protection and Control 44(14), 25–30 (2016).
11. Yu, Q.L.，Xu, C.Q.，Li, S.，Liu, H.，Song, Y.，Liu, X.O.: Application of Fuzzy Clustering Algorithm and Support Vector Machine to Short-term Forecasting of PV Power. Proceedings of the CSU-EPSA 28(12), 115–129 (2016).
12. Song, X.H., Guo, Z.Z., Guo, H.P., Wu, S.H., Wang, Z.Q., Wu, C.A.: A new forecasting model based on forest for photovoltaic power generation. Power System Protection and Control 43(2), 13–18 (2015).
13. Andreas C., M., Sarah, G.: Introduction to Machine learning with python. O'Reilly Media, Inc., Sebastopol (2016).
14. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning Data Mining, Inference, and Prediction. 2nd edn. Springer-Verlag, New York (2009).