Photovoltaic power analysis and prediction using machine learning methods

by

Halah Shehada

B.S, Kuwait University, 2019

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Electrical and Computer Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Mohammad Bagher Shadmand

Copyright

Abstract

The stochastic nature of Photovoltaic power directly affects the stability of the grid. PV power forecasting allows power stations to know beforehand how much PV power will be available, which ensures that the grid remains in stabilized condition. PV power from India is analyzed and predicted using machine learning methods

Table of Contents

Introduction

The natural growth in world population and industrialization reinforces the need to produce more power that can satisfy the growing demand. Global warming and environmental concerns have shifted the focus of power resources from traditional fossil fuel that produce toxins and greenhouse gases, into renewable resources of energy such as PV and wind turbine [1]. PV cells are of relatively low cost compared to some other renewable energy sources. Unlike wind turbines, they can be installed in a wide range of locations. Therefore, PV cells are considered the most promising and most prevalent form of renewable energy [1].

The stability of the power grid is directly affected by the stochastic nature of PV, which affects the stability of the power grid. Multiple approaches have been proposed in the literature to mitigate the negative effects that PV cells have on the grid. Such methods include load scheduling, installing energy storage systems, and PV power forecasting [3]. PV power forecasting provides the power grid operators with an estimation beforehand of how much PV power will be available in the future, which can assist in ensuring that the grid remains in a stabilized condition.

PV power forecasting splits into two main categories; direct PV power forecasting methods, and indirect PV power forecasting methods. The direct PV power forecasting depends mainly on the historical relation between power and weather condition. Since the relation between power and weather is not linear, soft computing techniques and time series methods are suitable for this type of prediction. On the other hand, indirect PV power forecasting depends mainly on predicting solar irradiance and using this predicted solar irradiance in computing power.

Background

For example, in [7] back propagation neural network was used to forecast PV power where genetic optimization was used to optimize the weights of the neural network. The author used correlation to determine the input features to the neural network, where he observed that using neural network without genetic optimization algorithm causes slow training speed and forces the algorithm to fall into local minimum. Where in [2], a hybrid deep learning is proposed to forecast short term photovoltaic power in a time series manner. An attention mechanism that simulates the attentions of the brain is used in the two Long Short-Term Memory neural network to focus the inputs on the features that are more important in forecasting. For indirect PV power forecasting several methods are proposed in the literature, for example in [3] the authors of the paper predicted PV power by analyzing the cloud cover since PV power depends directly on the amount of solar irradiance and since the solar irradiance that reaches the ground depends on the cloud cover so when clouds cover the sun the solar irradiance that reaches the PV panel drops and directly affects the PV power, so predicting cloud movement can help in prediction PV power. Whereas in [8], deep recurrent neural network is used to predict solar irradiance where this predicted solar irradiance will be used to calculate the predicted power. First, solar irradiance historical data set is collected and then data preprocessing is performed to clean up the data where the low sampling rate cannot capture the effect of birds passing over the panels thus increasing the sampling rate allows the detection of shallow clouds, birds and flying leaves.

Data

The data were obtained from Kaggle website. The data was acquired from two PV power plants in India for a period of one week. Each pair has one power generation dataset and one sensor reading dataset. The power generation dataset is collected at the inverter level, where the sensor data is gathered at a plant level. The data contains measurements of direct current power, ambient temperature, module temperature, solar irradiance, date, and time. The dataset for each plant contained 614 samples. An example of data reading is the following see Table 1,

Table 1. Example of Dataset

Week Day	Time	DC Power	Ambient Temp	Module Temp	Irradiation
Friday	11	6829	29.4	47.6	0.636
Friday	11.25	5969	30.2	50	0.58
Friday	11.5	6226	30	49.8	0.55
Friday	11.75	9420	30.8	47.8	0.46

The dataset was reprocessed to prepare it for machine learning analysis. Every empty cell in the dataset was deleted. The time of the day was replaced with numbers ranging from 0 to 23 increasing by 0.25 to represent the 15-minute time slot between each reading and the another. Also, the date was replaced with the day of the week so instead of having a complete date in the dataset, day of the week was inserted. Since machine learning algorithms normally do not have special treatment for the date, the continuous form allows the machine learning algorithm to better use that piece of information. I used 300 samples for training and 300 samples for testing.

Methodology

Random forest was used for classification and regression. Random forest is a mature machine learning algorithm that has demonstrated high efficiency, especially in cases where the number of samples are limited, and therefore methods such as neural networks are limited [10]. Random forest can also perform multivariate regression, allowing it to estimate continuous values rather than distinct classes. Random forest consists of training and testing stages see figure 1.

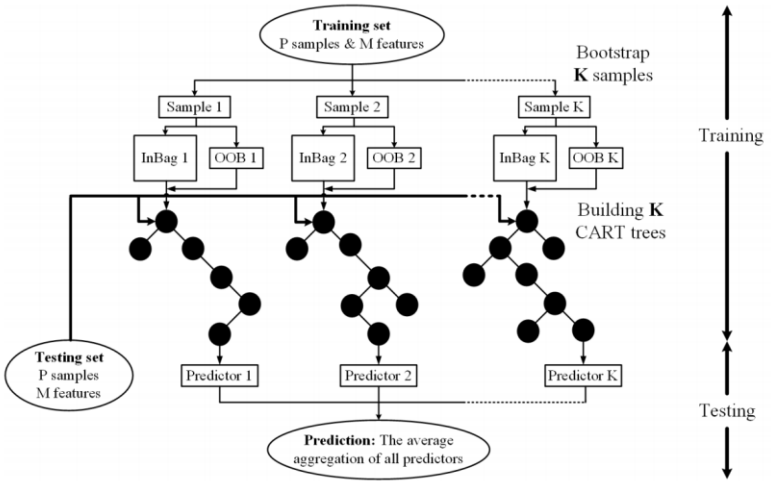

Figure 1. Random Forest Algorithm Structure.

The training dataset is used to create decision trees in the forest using bootstrapping technique. Bootstrapping technique splits the data into patterns to form the random forest regression model. Finally, the created model is tested using the testing dataset where each decision tree gives a prediction. Other classification algorithms used in this study were AdaBoost, Gaussian Bayesian Networks, and the simple linear discriminant analysis and quadratic discriminant analysis. In addition to supervised machine learning, one-class support vector machine (SVM) was also used.

One-class SVM is a mature proven technique for novelty detection and was applied to automatically identify outliers in the data as will be discussed in Section~\ref{results}.

The performance of the classification was evaluated by using a confusion matrix. One metric to measure the performance of a classifier is the classification accuracy. The classification accuracy can be defined as the number of times the classifier makes a correct classification divided by the total number of samples being classified, as shown by Equation 1,

$$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN}$$

Another metric that was used to measure the performance of the classifier is the recall. Recall can be defined as the true positive rate, as define by Equation 2,

$$TPR = \frac{TP}{TP + FN}$$

Also, precision is used where it can be defined as the true negative rate, as shown by Equation 3,

$$TNR = \frac{(TN)}{TN + FP}$$

For feature selection, the Pearson correlation coefficient between each feature and the power was computed, and the features that had the highest correlation coefficient were used, ensuring that features with lower correlation are excluded, as these features could not be directly related to the prediction of the power.

Results

Table 2 shows the confusion matrix for the first plant (plant ID 4135001). As the confusion matrix shows, 64 samples from Friday were correctly classified by the automatic classifier as Friday. Among the samples that were misclassified, 13 samples were classified as Saturday, and five were classified as Sunday. The classification accuracy of this confusion matrix is 64%, mean squared error is 0.26, the precision is 0.637 and the recall is 0.641. Table 3 shows the confusion matrix of the random forest of the second plant used in this study (plant ID 4136001). As the confusion matrix shows, 58 cases were predicted correctly by the classifier, where the classifier predicted that those cases were Friday and indeed the actual day was Friday. The classification accuracy is 64%, mean squared error 0.26, the precision around 0.64 and the recall around 0.64. In addition to random forest, several other classifiers were applied to the data from both datasets. The accuracy of the classifiers of plant one is shown in Table 4. As you can see from Table 4 some classifiers have 100 percent accuracy which is impossible and other classifiers have very low accuracy so, I can conclude that random forest classifier is the best classifier to use since its accuracy is neither 100 percent nor very low, so random forest is the most suitable classifier for the dataset of power plant one

Table 2. Confusion Matrix of Plant One

a	b	c	d	e	f	g	Classified as
64	13	5	1	8	0	1	a=Friday
5	56	2	3	3	7	12	b=Saturday
10	4	59	5	10	3	5	c=Sunday
0	4	4	81	4	2	1	d=Monday
6	2	9	6	65	4	1	e=Tuesday
1	5	5	1	4	49	15	f=Wednesday
5	16	3	0	2	23	19	g=Thursday

Table 3. Confusion Matrix of Plant Two

a	b	c	d	e	f	g	Classified as
58	2	13	1	10	2	6	a=Friday
4	48	14	9	6	3	4	b=Saturday
11	13	52	4	5	3	8	c=Sunday
0	10	1	74	5	6	0	d=Monday
5	2	2	10	67	5	2	e=Tuesday
2	3	6	3	1	60	5	f=Wednesday
5	6	16	0	3	2	36	g=Thursday

Table 4. Classification Accuracy of Multiple Classifiers When Applied To Plant 1

AdaBoost	Gradient Boosting	Gaussian NB
0.526	1	0.203

Linear Discriminant Analysis	Quadratic Discriminant Analysis
0.146	0.16

Table 5. Classification Accuracy of Multiple Classifiers When Applied To Plant 2

AdaBoost	Gradient Boosting	Gaussian NB
0.4076	1	0.3

LinearDiscriminantAnalysis	QuadraticDiscriminantAnalysis
0.216	0.136

Now, regarding plant number two the multiple classifiers are applied on plant two dataset and the accuracy of the classifiers is the following see in Table 5.

Then, I computed the Pearson correlation coefficient between power and features of plant one dataset, the following are the correlation coefficients of plant one DC power with other features see in Table 6. Also, I computed the Pearson correlation coefficient between power and features of plant two dataset, the following are the correlation coefficient of plant two DC power with other features see in Table 7. The relation between DC power and irradiation of plant one can be visualized see figure 2. As you can see that when irradiation increases the DC power also increases which means the correlation between DC power and irradiation is High. As you can see that when the ambient temperature increases the DC Power also increases which means the correlation between DC power and ambient temperature is high see figure 3. Moreover, the relation between DC power and module temperature of plant one can be visualized see figure 4.

Table 6. Pearson Correlation Coefficient Between DC Power and Plant 1 Features.

Power And Irradiation	Power And Ambient Temprature
0.96	0.604

Power And Module Temperature
0.891

Table 7. Pearson Correlation Coefficient Between DC Power and Plant 2 Features.

Power And Irradiation	Power And Ambient Temprature
0.99	0.71

Power And Module Temprature
0.96

Table 8. Predicted DC Power and Actual DC Power.

Actual DC Power	Predicted DC Power	Mean Squared Error
1204	1205	1.42

As you can see that when the module temperature increases the DC Power also increases which means the correlation between DC power and module temperature is high. Then, I applied random forest regression to predict the next 15 minutes of DC power of plant one. As you can see in Table 8, this is a sample of the actual and predicted power. Also, the mean squared error is calculated for each prediction. The actual and predicted power can be visualized see figure 5. As you can see that the predicted values are almost identical to the actual power values which means that the accuracy of the random forest regression is high. Finally, I detected the outliers of the dataset using one class svm algorithm of plant one and two. For plant one I took 93 samples from Friday and 10 samples from Saturday. The one class svm detected some of the outliers correctly as an outlier and assigned -1 to the outlier sample however, one class svm also detected wrongly some of the inliers as an outlier and assigned -1 to the inlier. The true positive rate is 0.7 and the true negative rate is 0.43 for this outlier detection. Similarly, for plant two dataset, when I applied one class svm to the dataset that includes 93 samples from Friday and 10 samples from Saturday the one class svm correctly detected some of the outliers and assigned -1 to the outlier sample on the other hand, one class svm also detected wrongly some of the inliers as an outlier and assigned -1 to the inliers. The true positive rate is 0.4 and the true negative rate is 0.27 for this outlier detection. Then, I calculated the P-Value of irradiation feature from two different days; Friday and Thursday to see if the data is statistically significant or not. The mean of the data taken from Friday is 0.206 and the standard deviation is 0.255. The mean of the data taken from Thursday is 0.346 and the standard deviation is 0.356. The two tailed P value equals 0.0043

so, it is considered as statistically significant. The main objective and assumption of the paper is that DC power has high correlation with ambient temperature, module temperature and irradiation and the results support these previous assumptions. One way to improve the regression accuracy is to collect larger dataset. Also, recurrent neural network can be used to predict power instead of using random forest where the weights of the neural network can be optimized using genetic algorithm.

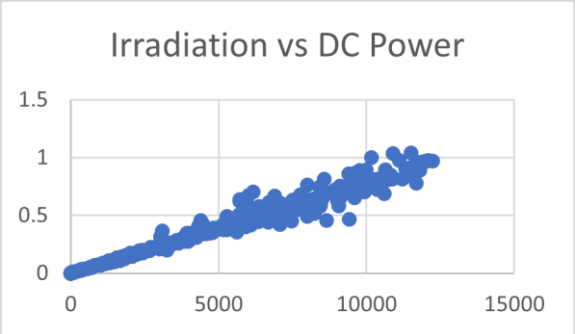

Figure2. Irradiation Vs DC Power

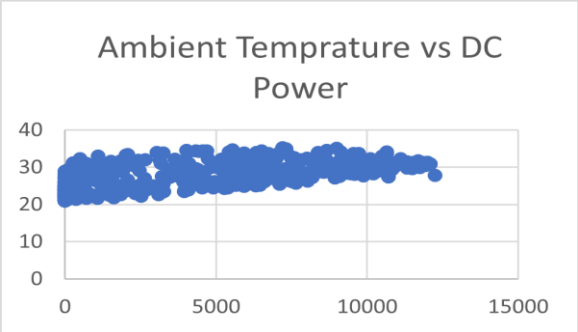

Figure 3. Ambient Temperature Vs DC Power

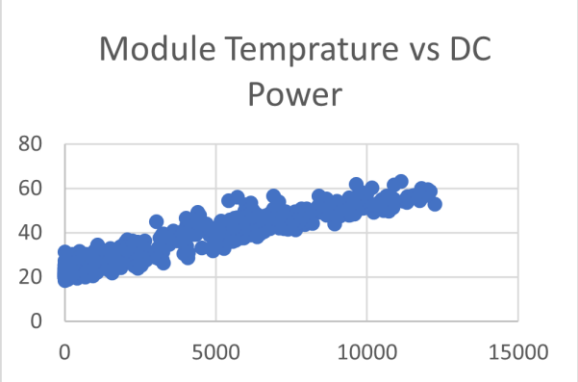

Figure 4. Module Temperature Vs DC Power

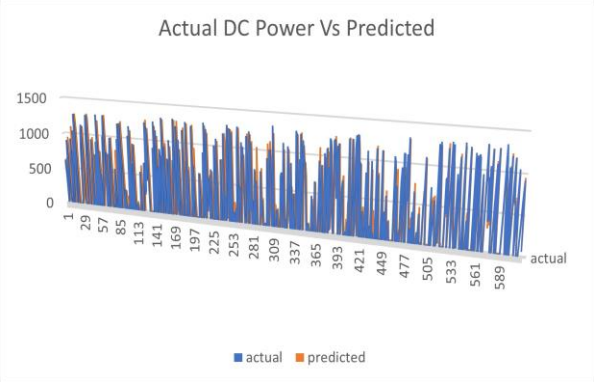

Figure 5. Actual DC Power Vs Predicted

Conclusion

The results show that the irradiation directly affects DC power, which proves the theoretical formula of calculating DC power. The random forest regression that is proposed in the paper has lower mean squared error than other prediction methodologies proposed in the literature. The analysis and prediction accuracy can be improved by collecting a larger dataset with more features. Thus, one direction for future work is to collect a larger dataset with more features and apply the recurrent neural network algorithm on the dataset to predict the future readings of power while optimizing the parameters of the recurrent neural network using genetic optimization.

References

[1] Aleem, Sk & Hussain, Suhail & Ustun, Taha Selim. (2020). A Review of Strategies to Increase PV Penetration Level in Smart Grids. Energies. 13. 636. 10.3390/en13030636

[2] H. Zhou, Y. Zhang, L. Yang, Q. Liu, K. Yan and Y. Du, "Short-Term Photovoltaic Power Forecasting Based on Long Short-Term Memory Neural Network and Attention Mechanism," in IEEE Access, vol. 7, pp. 78063-78074, 2019, doi: 10.1109/ACCESS.2019.2923006.

[3] S. Tiwari, R. Sabzehgar and M. Rasouli, "Short Term Solar Irradiance Forecast based on Image Processing and Cloud Motion Detection," 2019 IEEE Texas Power and Energy Conference (TPEC), College Station, TX, USA, 2019, pp. 1-6,doi:10.1109/TPEC.2019.8662134.

[4] A. Yona, T. Senjyu and T. Funabashi, "Application of Recurrent Neural Network to Short-Term-Ahead Generating Power Forecasting for Photovoltaic System," 2007 IEEE Power Engineering Society General Meeting, Tampa, FL, 2007, pp. 1-6, doi:10.1109/PES.2007.386072.

[5] C. A. Severiano, P. C. L. Silva, H. J. Sadaei and F. G. Guimarães, "Very short-term solar forecasting using fuzzy time series," 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, 2017, pp. 1-6, doi: 10.1109/FUZZ-IEEE.2017.8015732

[6] P. Luo, S. Zhu, L. Han and Q. Chen, "Short-term photovoltaic generation forecasting based on similar day selection and extreme learning machine," 2017 IEEE Power & Energy Society General Meeting, Chicago, IL, 2017, pp. 1-5, doi: 10.1109/PESGM.2017.8273776.

[7] Zhengqiu Yang, Yapei Cao and Jiapeng Xiu, "Power generation forecasting model for photovoltaic array based on generic algorithm and BP neural network," 2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems, Shenzhen, 2014, pp. 380-383, doi: 10.1109/CCIS.2014.7175764

[8] A. Alzahrani, P. Shamsi, M. Ferdowsi and C. Dagli, "Solar irradiance forecasting using deep recurrent neural networks," 2017 IEEE 6th International Conference on Renewable Energy Research and Applications (ICRERA), San Diego, CA, 2017, pp. 988-994, doi: 10.1109/ICRERA.2017.8191206.

[9] Jie Shi, Wei-Jen Lee, Yongqian Liu, Yongping Yang and Peng Wang, "Forecasting power output of photovoltaic system based on weather classification and support vector machine," 2011 IEEE Industry Applications Society Annual Meeting, Orlando, FL, 2011, pp. 1-6, doi: 10.1109/IAS.2011.6074294.

[10] A. Lahouar, A. Mejri and J. Ben Hadj Slama, "Importance based selection method for day-ahead photovoltaic power forecast using random forests," 2017 International Conference on Green Energy Conversion Systems (GECS), Hammamet, Tunisia, 2017, pp. 1-7, doi: 10.1109/GECS.2017.8066171.