Computational Model for Photovoltaic Solar Energy Forecasting Based on the K-Nearest Neighbor Method

Oberdan Pinheiro Rocha[1]*, Alexandre Menezes da Silva[1], Alex Álisson Bandeira Santos[1]
[1]SENAI CIMATEC University Center; Salvador, Bahia Brazil

Integrating PV technologies into power systems requires precise planning of PV performance. The ability to predict solar photovoltaic generation is a challenge for its integration into electrical systems. Improvements in forecasting models with more accurate results and fewer errors are necessary for the future development of microgrid projects and the dispatch of the economic sector. This research presents a computational machine learning model to predict the PV output power using historical PV output power data from a 960 kWP grid-connected PV system in southern Italy. The results showed agreement between the predicted and actual values, with errors ranging from 5% to 12%. We concluded that using machine learning techniques makes it feasible to predict the photovoltaic output power.
Keywords: Photovoltaic Power Forecast. Machine Learning. Regression.

Introduction

Solar energy is a friendly environment renewable energy source, making it a potential source of energy for industrial development [1]. Between 2017 and 2018, global photovoltaic power generation increased from 405 GW to 505 GW, representing 2.4% of total renewable energy [2]. In recent years, solar energy has experienced rapid growth in both developing and developed countries. There are several solar applications, such as photovoltaics, solar thermal, and other solar projects; however, solar radiation data is critical in each of them [3]. Although photovoltaic technology is a technological advancement, there is still a need for more research and development to improve the system's varied components.

Due to the complex nature of the operating conditions of photovoltaic systems, comprehending these elements is a significant challenge. In this sense, improving the techniques for measuring the power produced by photovoltaic generators is a strategy that can help the energy generation sector. According to IEA [4], there is an exponential increase in new renewable energies, such as wind and photovoltaic solar energy, in a Brazilian electrical matrix, presenting relevant technological evolution. The generation of photovoltaic solar energy depends fundamentally on solar irradiation at ground level, incident on a horizontal plane that is influenced by meteorological factors (cloud cover, rainfall, ambient temperature, atmospheric pressure, wind direction, humidity). These characteristics have promoted a broad spectrum of studies and the development of methods, techniques, and forecasting models.

This research aims to present a Computational Model for Photovoltaic Solar Energy Forecasting based on the k-nearest neighbor method. We organized two basic modules to propose the computational model: (i) data processing; and (ii) prediction model.

Studies

There are many studies on predicting the power generated by a photovoltaic module in the literature. Techniques involving machine learning, deterministic and hybrid techniques are examples [5]. Machine learning approaches investigated for power estimation of photovoltaic plants include multiple linear regression, support vector machine, adaptive neuro-fuzzy inference, and artificial neural networks. The research attempts to understand the relationship between outputs by adequately evaluating the data set, including acquired output and input parameters. Among them, ANN has been

Received on 26 June 2021; revised 22 August 2022.
Address for correspondence: Oberdan Pinheiro Rocha. Av. Orlando Gomes, 1845 - Piatã, Salvador - BA- Brazil. Zipcode: 41650-010. E-mail: oberdan.pinheiro@fieb.org.br. DOI 10.34178/jbth.v5i3.222.

widely applied [6]. Al-Amoudi and Zhang [7] developed a radial basis function neural network for peak power point prediction and solar panel modeling. The proposed method saves energy and accurately measures the maximum power without searching for the optimal power point.

Almonacid and colleagues applied an artificial neural network to estimate the photovoltaic output power. The results of this study showed that the artificial neural network technique produces significant improvements over traditional approaches. For example, artificial neural network errors range from 6% to 8%, while conventional approaches have errors ranging from 6% to 30% (taking into account only the effect of temperature and solar irradiance).

Mellit and Kalogirou used adaptive neuro-fuzzy inference to model and simulate photovoltaic plants. The proposed model predicts and simulates diverse electrical data from a photovoltaic energy system using solar irradiance, ambient temperature, and clarity index. Fonseca and colleagues [10] investigated the application of a support vector machine to predict the annual photovoltaic output power of a 1 MW photovoltaic plant. In addition, cloudiness was numerically predicted to examine how it influenced PV forecasts. Six variables were used as input vectors in this process to estimate the output power of the photovoltaic plant: low-level cloudiness, relative humidity of extraterrestrial insolation, normalized temperature, upper-level cloudiness, and mid-level cloudiness.

We compared the predicted values produced by the support vector machine and those produced with the conventional technique using the mean absolute error, mean absolute percent error, and the root means square error. The results revealed that the support vector machine algorithm offered accurate projections of photovoltaic energy production. Furthermore, the selection of internal parameters significantly impacts the PV output power. A good selection of support vector machine parameters is, therefore, a critical step in increasing the overall efficiency of the model. We did not well investigate the estimation of energy generated by the photovoltaic output power using the k-nearest neighbor machine learning algorithm.

Materials and Methods

Data Processing

We used the history of photovoltaic solar generation measurements to propose a forecast model. For the quality of the prediction-model-performance, the data must have possible quality. The data must be free of errors and disturbances caused by equipment and measurement sensor failures or events in the photovoltaic solar generation system or other unnatural causes that could affect the output power path. Errors and perturbations lead to outliers, discontinuities, and data gaps, compromising the model's fit and the quality of its predictions. There were applied filters for pre-processing data: (i) treatment of null values and (ii) filtering of overestimated records.

Forecast Model

We used the k-nearest neighbor (kNN) algorithm to develop the forecast model [11]. kNN uses distance functions to find a set of k samples closest to unknown samples. The k most similar instances closest to the current data point used a labeled dataset. As a result, the algorithm predicts how similar the recently received observations are to the training observation. During the learning phase, this algorithm maintains the complete training set. Unknown samples (i.e., new input data) have their labels (classes) compared to each instance in the training set, and by finding the mean of the response variables, we can predict them.

Regarding prediction, kNN is a good algorithm [12] since it uses local knowledge and highly adaptive behavior. However, one of the limitations of kNN is that a large amount of historical data is needed to build a model to search for the k nearest neighbor. When making predictions about regression problems, KNN will average the most similar instances in the training dataset. The parameter k controls the size of the neighborhood. For example, if set to 1, predictions are made using the single training instance most similar to a given new pattern for which a prediction is requested. Common values for k are 3, 7, 11 and 21 [12]. Another important parameter is the distance measure, which controls how training data is stored and searched. The studies used Euclidean distance to calculate the distance between instances, which is good for numerical data with the same scale. Manhattan distance is good to use if its attributes differ in measurements or type.

Data Used

We used the dataset collected from the photovoltaic system [13], located on the campus of the University of Salento, in Monteroni di Lecce (LE), Apulia, Italy (40°19’32” 16N, 18°5’52” 44E) developed within the scope of the European project “7th Framework Program Building Energy Advanced Management Systems (BEAMS)”. The collected data are related to the 960kW P photovoltaic system. We installed the photovoltaic modules in shelters used as parking lots. The first section of the system has 1,104 modules and an effective area of 1,733.3 m2, and the second has 1,896 modules and an effective area of 2,976.7 m2. It also has inverters for injecting the energy generated into the grid, an ambient temperature sensor, two other sets of sensors (temperature of the modules and global solar irradiation), one for each section of the system, and a SCADA system for the collection and storage of data. In this research, we used the following data sets: measurements of the University of Salento system for 21 months (from April 2012 to December 2013), covering measurements, on an average hourly basis, of the global solar irradiance in the two sections of panels and the total generation of the panels.

Results and Discussion

We used the cross-validation technique (10-fold cross-validation) to evaluate the computational model based on the kNN method for photovoltaic power prediction. We configured the model with k = 3, 5, 7, 9, 11, and 21. The distance measure used was the Euclidean distance since the numerical data present in the data set were on the same numerical scale.

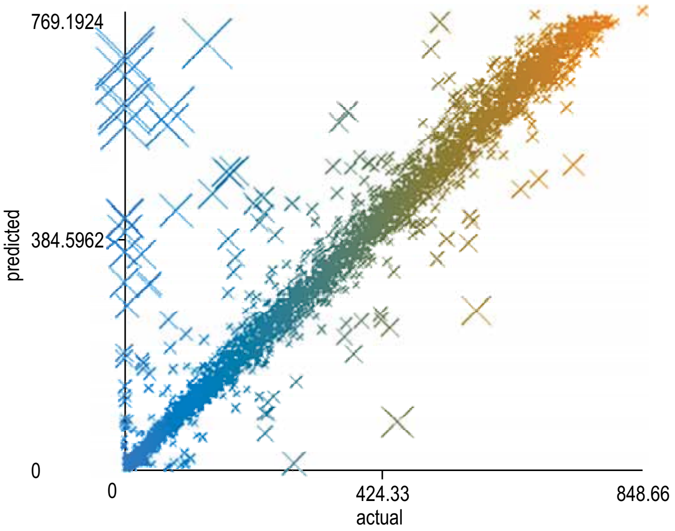
The number of instances present in the original dataset was 11,919. After applying the filters in the pre-processing step, the number of instances was reduced to 9,426. Table 1 presents the relative absolute error and the root relative squared error in percentage for the different k parameters.

We observed that the best k was 21, with an error of 5.5% and 12.7%. Figure 1 shows the predicted and actual values of all 9,426 training cases using kNN. The y-axis projects values for each PV power, while the x-axis presents the actual values. The model follows the photovoltaic power prediction closely, although it presents variations.

Table 1. Forecast error.

k	Absolute error (%)	Squared error (%)
3	5.9	13.6
5	5.8	13.1
7	5.7	13
9	5.6	12.8
11	6	12.8
21	5.5	12.7

Figure 1. Diagram example.



For the case of kNN, the algorithm searches in the feature space for the k closest samples depending on a predetermined distance. As a result, kNN is evaluated using the parameter k, implying that k is the key-setting parameter. So, based on the data set, the objective is to get the best value of k in the model.

Conclusion

Accurate power output prediction is critical for the operational planning of electrical power systems. In this study, machine learning estimated the energy generated by a solar photovoltaic system located in the Italian Mediterranean region of Italy. The results show a reasonable agreement between the actual and predicted values. This study implies that the machine learning technique, particularly the kNN, can be used to characterize other areas of the photovoltaic installation and other photovoltaic generators. Future research will analyze other databases to find appropriate indicators of merit. With this study, it may be possible to predict PV power using just a few small samples accurately.

Acknowledgments

The authors are grateful for the financial support of partner institutions SENAI CIMATEC and CHESF/ANEEL.

References

1. Fan J, Wu L, Zhang F, Cai H, Wang X, Lu X, Xiang Y, Evaluating the effect of air pollution on global and diffuse solar radiation prediction using support vector machine modeling based on sunshine duration and air temperature, Renewable and Sustainable Energy Reviews 2018;94:2090-4479.
2. Mas'ud A. Comparison of three machine learning models for the prediction of hourly PV output power in Saudi Arabia, Ain Shams Engineering Journal 2022;13:732-747.
3. Hemeida AM, El-Ahmar MH, El-Sayed AM, Hasanien HM, Alkhalaf S, Esmail MFC, Senjyu T. Optimum design of hybrid wind/PV energy system for remote area, Ain Shams Engineering Journal 2020;11:2090-4479.
4. IEA (2018), World Energy Outlook 2018, IEA, Paris https://www.iea.org/reports/world-energy-outlook-2018.
5. Graditi G, Ferlito S, Adinolfi G. Comparison of Photovoltaic plant power production prediction methods using a large measured dataset. Renew Energy 2016;90:513-519.
6. Hiyama T, Kitabayashi K. Neural network based estimation of maximum power generation from PV module using environmental information. IEEE Trans Energy Convers 1997;12(3):241-247.
7. Al-Amoudi A, Zhang L, Application of radial basis function networks for solar-array modelling and maximum power-point prediction. IEE Proc Gener Transm Distrib 2000;147(5):310.

8. Almonacid F, Rus C, Pérez-Higueras P, Hontoria L. Calculation of the energy provided by a PV generator. Comparative study: Conventional methods *vs*. artificial neural networks. Energy 2011;36(1):375-384.
9. Mellit A, Kalogirou SA. ANFIS-based modelling for photovoltaic power supply system: A case study. Renew Energy 2011;36(1):250-258.
10. Fonseca JGS, Oozeki T, Takashima T, et al. Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in Kitakyushu, Japan. Prog Photovoltaics Res Appl 2012;20(7):874-882.
11. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967;13(1):21-27.
12. Al-Qahtani FH.Multivariate k-nearest neighbour regression for time series data — A novel algorithm for forecasting UK electricity demand. Proceedings of the International Joint Conference on Neural Networks 2013.
13. Malvoni M, De Giorgi MG, Congedo PM. Data on photovoltaic power forecasting models for Mediterranean climate. Data in Brief 2019;7(1):1639-1642.