Apply Machine Learning to Detect and Diagnose Faults in Multi-Array PV Plants

Chung-Chian Hsu, Jia-Long Li, Arthur Chang, and Yu-Sheng Chen*

Big Data Research Center
National Yunlin University of Science and Technology
Douliu, Yunlin, 640 Taiwan
hsucc@yuntech.edu.tw
*Reforecast Co., Ltd
Douliu, Yunlin, 640 Taiwan

ABSTRACT—Faults in photovoltaic arrays can reduce energy production and lead to significant revenue loss if the produced power is for sell. It is important faulty arrays can be detected as soon as possible. Existing detection approaches usually require additional sensors leading to extra cost or need to estimate several parameters of the equipment via simulation which requires periodically re-estimation due to equipment aging. In this project, we propose an approach based on machine learning techniques for real-time detection and diagnosis of faults in photovoltaic (PV) arrays. The proposed approach is inexpensive which does not require additional sensors and rely on collected data only: the produced power and irradiance. In addition, to monitor massive multi-array plants, we propose to deploy the developed monitoring system on a Spark-cluster distributed platform such that the detection and diagnosis process can be finished within 5 minutes for handling over 7000 PV arrays.

Keywords: Fault detection and diagnosis, *k*-nearest neighbors, machine learning, regression, distributed cluster computing.

1. INTRODUCTION

Maintenance of solar panels is important which influences power production. One of the problems is that solar systems can sometimes fail or reduce power output due to various reasons such as short circuit, shading, inverter breakdown, etc. It is extremely crucial to detect and diagnose the faults in real-time so that repairing can be arranged as soon as possible.

In the literature, several authors proposed to use single diode model [1][2] or Sandia model [3] to estimate theoretical values of current and voltage under irradiance and module temperature. In such case, five and seven parameters need to be extracted from training data respectively for the two models in advance by numeric methods. Alternatively, many fault detection and diagnosis systems [4] require additional sensors or hardware which increase cost and not suitable for companies which manage massive number of PV plants.

To maximize the revenue for a company managing a large number of PV plants and selling the produced power, an inexpensive and real-time fault detection and diagnosis system is indispensable such that arrays faults can be detected in real-time and maintenance visits can be dispatched immediately.

We propose a detection and diagnosis framework based on machine learning techniques which do not require additional sensors except for a pyrheliometer. Specifically, we develop a fault detection approach which exploits historical irradiance and power data of the PV arrays. The range of array ratio (RA) in normal operation mode for each PV array is estimated by using a nonlinear regression algorithm from its own data collected in the past 3 months. To detect fault in real-time, the RA of the array is calculated based on the data read every 5 minutes and compared. If the number of consecutive RAs which exceed the normal range is larger than a threshold, the array is considered faulty. In case of a new plant, there is not enough historical data for the regression, the *k*-nearest-neighbor (*KNN*) technique is used to estimate the expected power of the target point. For diagnosis of the fault, we compare the current and the voltage with the data from a normal array of the multi-array plant. The prototype is deployed on a Spark-cluster distributed platform.

2. METHOD

The fault detection process at the array level is proposed as shown Figure 1. The irradiance and power output for each PV array is collected and transmitted every 5 minutes. If no signals are read due to, say, transmission problem, an alarm will be issued to a maintenance engineer. Low irradiance due to cloudy days cannot generate enough power to turn on the inverter. Thus, if irradiance is lower than 250, we ignore the data points. The read data of a PV array is compared with its pre-calculated RA model to check whether the RA of the target point falls in the normal range. If not, the point is considered faulty.

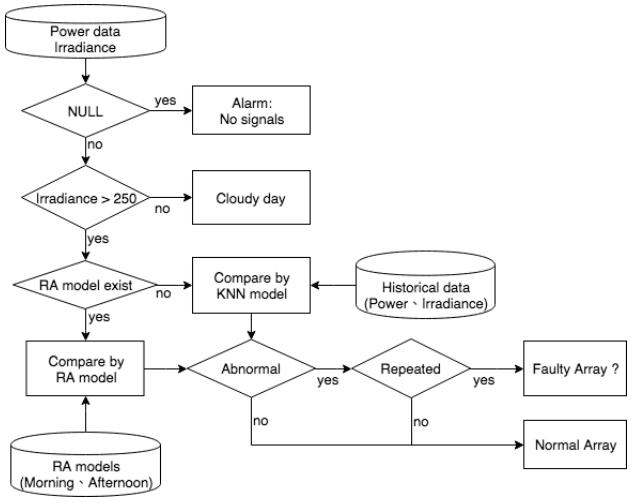

Fig. 1. Flowchart of fault detection of PV arrays.

False alarms can occur due to variations. To reduce false alarms, an array must induce at least a predefined *n* consecutive faulty data points in time to be considered possibly faulty. A diagnosis process is followed to confirm the fault and determine its type. In case that there is not pre-stored RA model due to a new plant, the *KNN* technique is used to estimate the expected power with respect to the irradiance of the target point. RA is calculated by Eq. (1).

$$RA_t = \frac{P_t/P_0}{IRR_t/IRR_0} \quad (1)$$

where P_t and IRR_t denote measured power and irradiance at time t, and P_0 and IRR_0 represent nominal capacity of the array and irradiance under standard test condition, i.e., 1000W/m^2.

To estimate the normal range, the RA of the array with respect to irradiance is computed by quadratic regression with training data collected from the array in the past 3 months, as shown in Algorithm 1. Low irradiance or abrupt change leads to instability. We ignore those data points. The range is then set to RA±θ, for instance, say, θ = 10%. The regression is defined by Eq. (2). If the measured RA at time point t is out of the normal range, the data point is considered faulty.

$$RA = w_0 + w_1 r + w_2 r^2 \quad (2)$$

where r denotes irradiance and w_i is regression coefficient.

Algorithm 1: Estimating the normal range of $\widehat{RA}_{IRR}$ for each PV array

1. Retrieve power P_t and irradiance IRR_t of the PV array for the past 3 months.
2. Discard noisy data points: IRR_t < 250 or abrupt-change > 100.
3. Calculate RA_t with P_t and IRR_t.
4. Retain only RA in the 2nd and 3rd quantile with respect to each IRR.
5. Perform nonlinear regression for $\widehat{RA}_{IRR}$ on the set $\{(RA, IRR)\}$.
6. Set the normal range to $\widehat{RA}_{IRR}$ ± 10%.

For a new plant, there is not enough historical data for estimating the range. In such case, we use the *KNN* method to estimate the expected power of the target point with respect to the read irradiance. The k closest irradiances in the past x days were retrieved, and the average of their corresponding powers is taken as the estimation, as defined by Eq. (3). If the measured power is less than the average $\hat{p}$ by y%, the data point is considered faulty. We will conduct extensive analysis on historical data to determine proper values for k, x and y.

$$\hat{p} = \frac{1}{k} \sum_{i \in R_k} p_i \quad (3)$$

where R_k is the set of k data points which irradiances closest to the irradiance of the target point, p_i is its corresponding power, and $\hat{p}$ is the estimated power.

To determine the fault type, a rule-based procedure is used as shown in Fig. 3. The current ratio *IR* and the voltage ratio *VR* between the target array and a normal array from the plant are calculated and used for determining the fault type. The rationale behind the idea is that we shall be able to find a normal array in a multi-array plant for the comparison. Moreover, since we did not install additional sensors in individual panels or strings, some of the fault-type diagnoses cannot be specific. However, as long as the accuracy of fault detection is high, the system will not result in many maintenance visits in vain and can satisfy the operational need.

To monitor massive multi-array plants in real-time, the developed system will be deployed on a Spark distributed platform on the cloud platform Amazon Web Services (AWS) as shown in Fig. 4. Power and irradiance data are sampled every 5 minutes. The detection and the diagnosis process must be finished in 5 minutes to avoid congestion.

Fig. 4 demonstrates a RA regression result for a PV array of a plant. As can be seen, the data fits the model reasonably well. By comparing with the result, we can determine whether the output of a PV array lies in the normal range. If not, the diagnosis procedure will be activated, and an alarm will be issued to maintenance engineer.

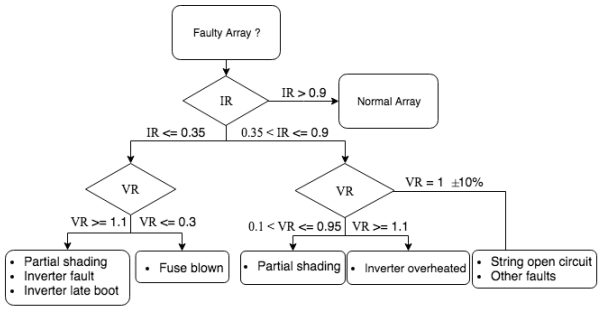

Fig. 2. Flowchart of fault diagnosis.

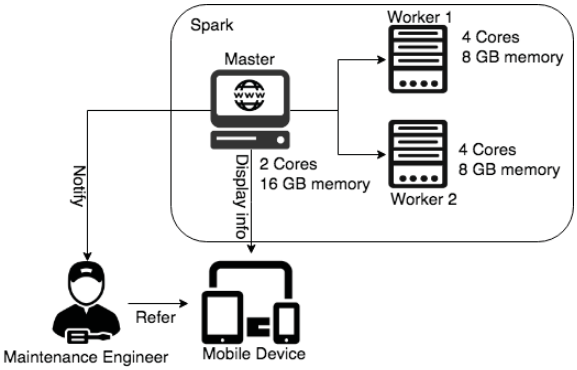

Fig. 3. The developed system will be deployed on a Spark distributed platform of Amazon Web Services.

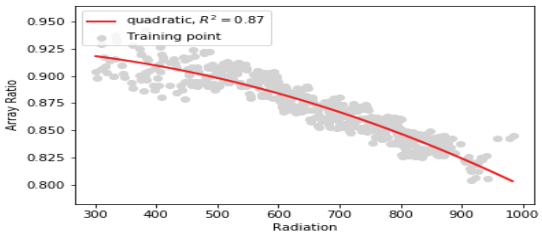

Figure 4. Regression result of the RA model with respect to irradiance.

3. REFERENCES

[1] M. Dhimish, V. Holmes, Fault detection algorithm for grid-connected photovoltaic plants, Solar Energy 137 (2016) 236-245.

[2] A. McEvoy, L. Castaner, T. Markvart, Solar Cells: Materials, Manufacture and Operation. Academic Press, 2012.

[3] R. Benkercha, S. Moulahoum, Fault detection and diagnosis based on C4.5 decision tree algorithm for grid connected PV system, Solar Energy 173, 2018, pp. 610-634.

[4] A. Mellit, G.M. Tina, S.A. Kalogirou, Fault detection and diagnosis methods for photovoltaic systems: a review, Renewable and Sustainable Energy Reviews, 91, 2018, pp. 1-17.

[5] D. Sera, R. Teodorescu, P. Rodriguez, PV panel model based on datasheet values, 2007. http://dx.doi.org/10.1109/ISIE.2007.437498.