Hindawi
International Journal of Photoenergy
Volume 2017, Article ID 4903613, 14 pages
https://doi.org/10.1155/2017/4903613

Research Article

A Density Peak-Based Clustering Approach for Fault Diagnosis of Photovoltaic Arrays

Peijie Lin,[1] Yaohai Lin,[2] Zhicong Chen,[1] Lijun Wu,[1] Lingchen Chen,[1] and Shuying Cheng[1]

[1]*Institute of Micro/Nano Devices and Solar Cells, College of Physics and Information Engineering, Fuzhou University, Fuzhou 350116, China*
[2]*College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China*

Correspondence should be addressed to Shuying Cheng; sycheng@fzu.edu.cn

Received 28 November 2016; Accepted 6 February 2017; Published 28 March 2017

Academic Editor: Cheuk-Lam Ho

Fault diagnosis of photovoltaic (PV) arrays plays a significant role in safe and reliable operation of PV systems. In this paper, the distribution of the PV systems' daily operating data under different operating conditions is analyzed. The results show that the data distribution features significant nonspherical clustering, the cluster center has a relatively large distance from any points with a higher local density, and the cluster number cannot be predetermined. Based on these features, a density peak-based clustering approach is then proposed to automatically cluster the PV data. And then, a set of labeled data with various conditions are employed to compute the minimum distance vector between each cluster and the reference data. According to the distance vector, the clusters can be identified and categorized into various conditions and/or faults. Simulation results demonstrate the feasibility of the proposed method in the diagnosis of certain faults occurring in a PV array. Moreover, a 1.8 kW grid-connected PV system with 6×3 PV array is established and experimentally tested to investigate the performance of the developed method.

1. Introduction

The rapid increase in the amount of grid-connected photovoltaic (PV) systems has put forward a significant research topic, that is, operating condition analysis and fault diagnosis of PV systems. As one of the most important components, the performance of PV arrays (DC side) usually affects the operation of the entire system. However, due to complex outdoor working environments, the PV array is susceptible to thermal cycling, humidity, ultraviolet light, hard shadows, and other environmental factors that cause various faults such as cracking, hot spots, modules' short circuit, and PV strings' open circuit. As a result, these will lead to power losses and even fire hazards [1]. The overcurrent protection devices (OCPDs) and ground fault detection interrupters (GFDIs) are usually installed as the traditional fault detection and protection for the PV arrays [2]. However, due to the nonlinear output characteristics of the PV array, various faults remain and cannot be eliminated by the protection devices [3, 4].

To address these problems, various fault diagnosis approaches for PV arrays have been studied, including thermal imaging [5–7], earth capacitance measurement (ECM), time-domain reflectometry (TDR) [8, 9], power loss analysis [10–12], current and voltage indicators evaluation [13–16], and machine learning [17–23]. The infrared thermal imaging method is applied to detect and identify the hot spot and degradation fault in PV modules according to the temperature characteristics of the PV module. The ECM is presented to detect the location of open-circuit faults in PV strings, and the TDR is applied to identify the degradation of a PV array. Power loss analysis method is proposed to detect various types of faults occurring in solar PV systems by comparing the measured and theoretical output power of the PV array. The automatic supervision and fault detection procedure that based on evaluation of current and voltage indicators in grid-connected PV systems is proposed to identify the short circuits and open circuits in PV arrays [13] as well as inverter disconnection and partial shading conditions [14]. Moreover, the procedure

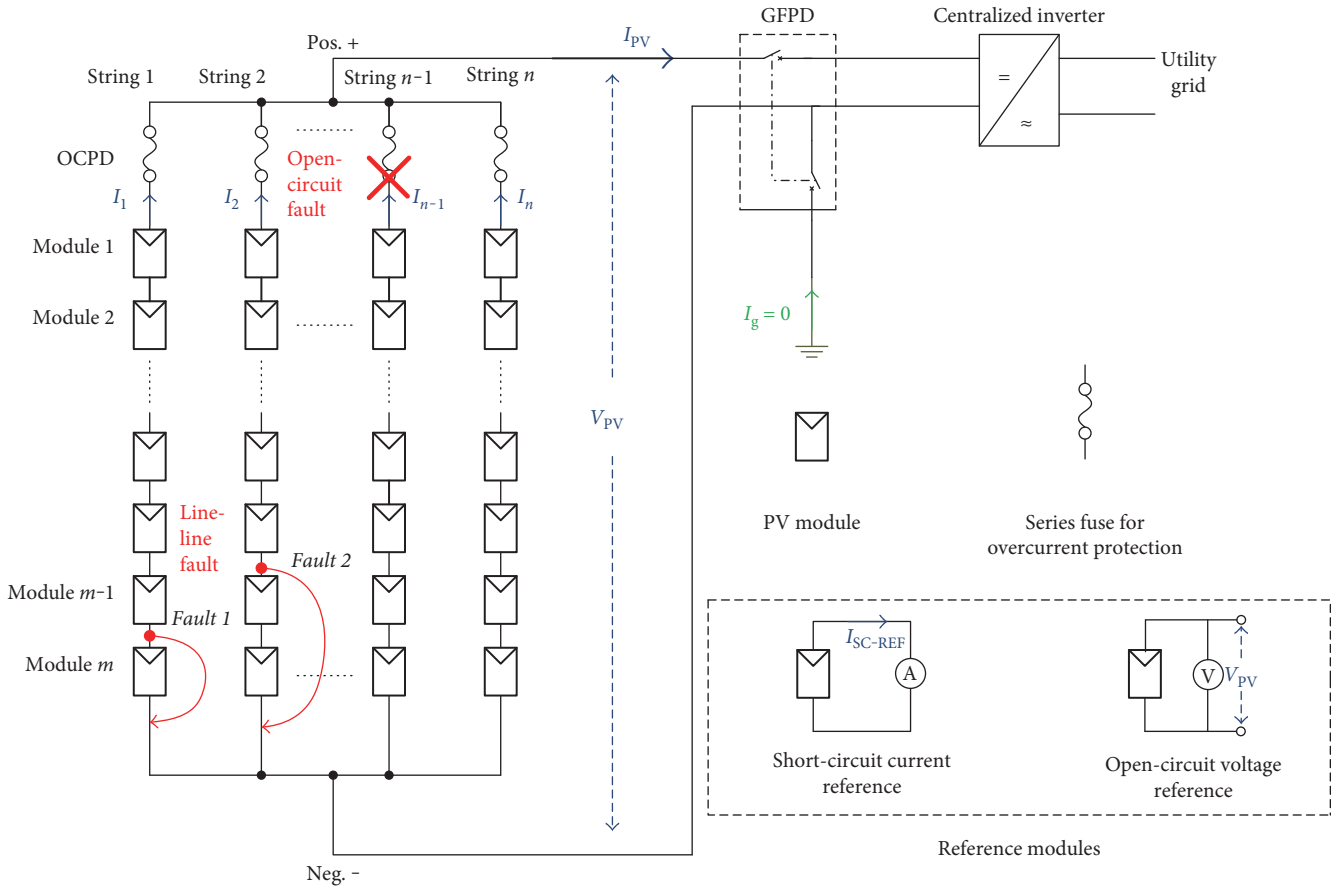

Figure 1: Schematic diagram of series-parallel grid-connected PV system.

is combined with an OLE (Object Linking and Embedding) for Process Control (OPC) monitoring for remote supervision and diagnosis of grid-connected PV systems [15]. Furthermore, the analysis of current and voltage indicators is applied to detect, in real time, the faults related to bypassed PV modules, open-circuit strings and partial shading for a PV plant connected to a single-phase grid [16].

Furthermore, to better detect and classify PV faults, machine learning algorithms are widely carried out. A fault detection and classification model based on decision tree is presented to deal with the line-line, open-circuit, and partial shade faults in PV arrays [17]. Artificial neural network technique is applied to monitor the health status, measure degradation, and indicate maintenance schedules of a PV system [18]. The study in [19] proposed a method to identifying the short-circuit location of PV modules in one string by using three-layered feed-forward neural network. An online PV modules' fault diagnosis model is established based on back propagation neural network [20]. The Bayesian neural network and polynomial regression models are researched for the evaluation of soiling effects on PV plants [21]. A new artificial neural network approach is implemented in a field-programmable gate array (FPGA) and has the ability to identify eight types of fault occurring in a PV array [22]. A semisupervised learning model is employed for line-line and open fault detection and classification in PV arrays [23].

In practice, daily operational data from various PV systems are stored in the monitoring systems, enabling the working condition estimation of PV arrays and fault diagnosis based on the data [24–26]. According to the distribution characteristics of PV data analyzed in this paper, a density peak-based clustering approach for fault diagnosis in PV arrays is proposed. The approach diagnoses the PV faults by clustering and classifying the daily operational data. The advantage of the proposed approach is that a larger amount of training data and tedious training process are not needed and only few labeled reference data obtained from a simulated PV system is required to identify clusters.

The rest of this paper is organized as follows: Section 2 depicts the distribution characteristics of PV data and the process of the proposed method. The simulation results are presented in Section 3, and several working conditions of PV array are studied. In Section 4, experiments and result analysis are carried out. Finally, some conclusions are drawn in Section 5.

2. Proposed Models

In this section, the features of PV data are analyzed, such as data distribution, cluster shape, and cluster number. Then, the procedure of the proposed approach is described in detail.

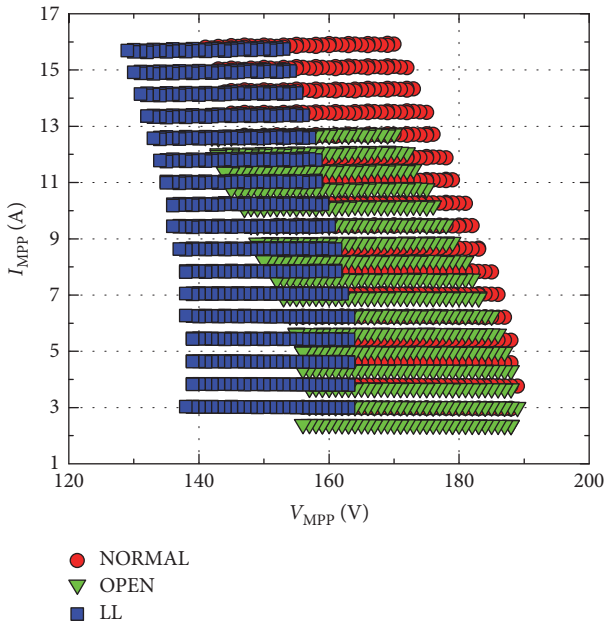

Figure 2: The V_{MPP} versus I_{MPP} of PV array over a range of irradiance and temperature.

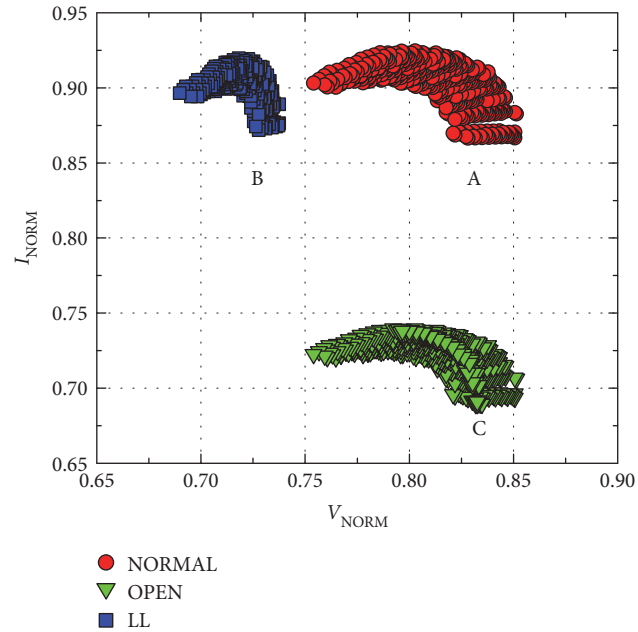

Figure 3: The distribution of PV data over a range of irradiance and temperature.

2.1. Photovoltaic Data Distribution. The schematic diagram of a typical series-parallel grid-connected PV system is shown in Figure 1. The system generally is comprised of $m \times n$ PV array, a centralized inverter, protection devices (such as OCPD and GFPD), and connection wires [27]. Usually, the PV array can output maximum power under variable environment due to the maximum power point tracking (MPPT) technology of inverters. When faults occur, however, the MPPT is possible to keep the optimal power output if the PV array can reach the inverter's working voltage. As a result, the current of the PV array may be significantly reduced, leading to the failure of the OCPD to clear the fault [3].

In every daily operation cycle, the voltage (V_{MPP}) and current of PV array at MPPs change due to the variations of solar irradiance and atmospheric temperature. In order to investigate the changes of V_{MPP} and I_{MPP} under different conditions, a normal (NORMAL) and two common faults of a specific PV array are considered. As shown in Figure 1, the two faults are line-line (LL) fault and open-circuit (OPEN) fault, which may be difficult to be cleared by conventional OCPD. The simulated V_{MPP} versus I_{MPP} over a range of irradiance and temperature is shown in Figure 2. Obviously, part of the V_{MPP} and I_{MPP} overlaps, causing difficulties for the PV fault diagnosis.

To make better visualization and identification of PV faults, the approach proposed in [13, 23] is applied to normalize the V_{MPP} and I_{MPP}. The normalization formula can be expressed as follows:

$$V_{NORM} = \frac{V_{MPP}}{m \times V_{OC\text{-}REF}}, \quad I_{NORM} = \frac{I_{MPP}}{n \times I_{SC\text{-}REF}}, \tag{1}$$

where V_{NORM} and I_{NORM} are the normalized PV voltage and PV current, respectively; $V_{OC\text{-}REF}$ is the open-circuit voltage of reference PV module; $I_{SC\text{-}REF}$ is the short-circuit current of reference PV module (as shown in Figure 1); m is the number of modules in series in each PV string; and n is the number of strings in parallel in the array. Hereafter, the data set of V_{NORM} and I_{NORM} is simply referred to as PV data, which is the input data of the proposed model. The PV data distribution of a PV array over a range of irradiance and temperature is shown in Figure 3.

It is clearly demonstrated that the PV data have good data clustering and the clusters are nonspherical in shape. In each cluster, data from the bottom to the upper-left indicate the data from low irradiance to high irradiance. In daily operation, the PV system generally runs under NORMAL condition and the corresponding PV data are distributed in only a cluster, that is, cluster A in Figure 3. When fault occurs, such as LL fault, the data distribution is changed from cluster A to cluster B. Furthermore, the data may vary from cluster B to cluster C if another fault happens, such as OPEN fault. Hence, the number of clusters cannot be predefined. Moreover, the center of each cluster has a relatively large distance from any points with a higher local density. Therefore, the PV data can be clustered by using an appropriate clustering algorithm and then further analyzed for PV array faults.

2.2. Procedure of the Proposed Approach. There are two phases in our proposed approach. Firstly, the daily PV operation data are recorded and assigned into several clusters by using a clustering algorithm. Each cluster represents a kind of work conditions of the PV array. Secondly, with the aid of the labeled reference data, each cluster will be identified, respectively. Thus, the recorded PV data can be divided into

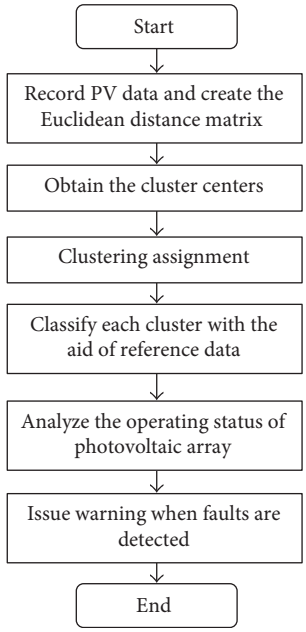

Figure 4: Flowchart of the proposed approach.

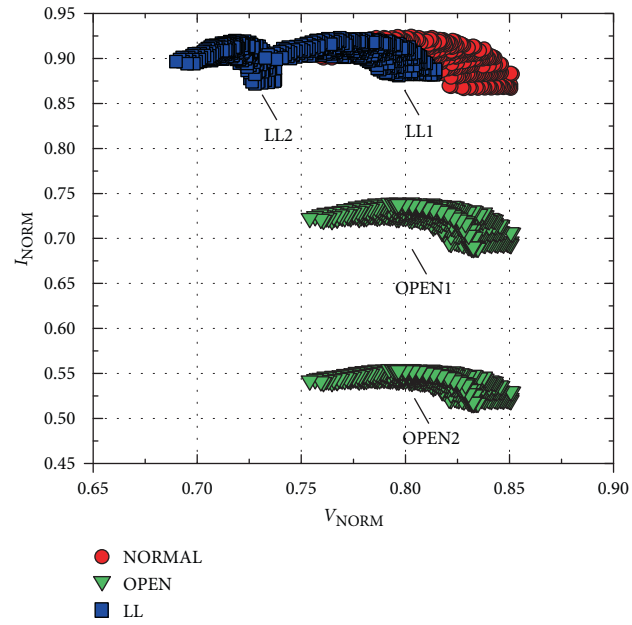

Figure 5: The distribution of PV data.

Table 1: Main parameters of SM55.

Parameters	Values
Maximum power P_{MPP} (W)	55
Maximum power current I_{MPP} (A)	3.15
Maximum power voltage V_{MPP} (V)	17.4
Short-circuit current I_{SC} (A)	3.45
Open-circuit voltage V_{OC} (V)	21.7

the aforementioned work conditions, that is, NORMAL, LL, OPEN, or their combinations.

2.2.1. Phase 1 PV Data Clustering. Recently, an algorithm implementing clustering by fast search and find of density peaks (CFSFDP) published on *Science* is proposed by Rodriguez and Laio [28]. This method is based on two assumptions: the cluster centers must have the highest local density and they have relatively large distance to the points with higher density. It has an excellent ability to analyze arbitrary shape clusters as well as different dimensional cases and to find cluster centers. As discussed in Section 2.1, PV data have some features, such as non-spherical, cluster centers have a relatively large distance from any points with a higher local density, and cluster number cannot be predefined. Therefore, the CFSFDP algorithm is very suitable for the analysis of the PV data.

In CFSFDP, two important indicators are defined and computed: ρ_i and δ_i, which represent the local density of a data point and the distance from data points of higher density, respectively. In the proposed approach, for each PV data point i, the procedure for calculating its ρ_i and δ_i is as follows:

Firstly, the PV data are recorded and organized as $X = [x_1, x_2, \ldots, x_N]$, where $x_i = [V_{NORMi}\, I_{NORMi}]^T$ and N is the number of PV data points. The distance matrix of data points should be calculated. Let d_{ij} represent the Euclidean distance between x_i and x_j; then

$$d_{ij} = \left\| x_i - x_j \right\|_2, \tag{2}$$

where $\|\cdot\|$ denotes the 2-norm operator.

Then ρ_i is calculated by using the Gaussian kernel function, as follows:

$$\rho_i = \sum_{i=1}^{N} \exp\left(-\frac{d_{ij}}{2d_c^2}\right), \tag{3}$$

where d_c is the cutoff distance, which represents the neighborhood range of data point i. The CFSFDP algorithm suggests that one can choose d_c so that the average number of neighbors is around 1% to 2% of the total number of points in the PV data set and 2% is applied in this study. And δ_i is computed as follows:

$$\delta_i = \min_{j:\rho_j > \rho_i} \left(d_{ij}\right). \tag{4}$$

For the point with the highest density, the δ_i is defined as $\max_j(d_{ij})$. It is obvious that points with local or global maxima density have large δ_i. According to ρ_i and δ_i, there are some characteristics that can be obtained as follows: a point has high ρ and low δ, which means that the point i is close to the clustering center; a point has low ρ and low δ, which indicates that the point is located in the boundary of the clustering; a point has low ρ and high δ, which implies that the point is far away from each clustering and can be noise or outliers. So only the points with both high ρ and high δ are the clustering centers. Therefore, the product of ρ_i and δ_i is applied to measure the probability of cluster centers, which is denoted as γ_i [28].

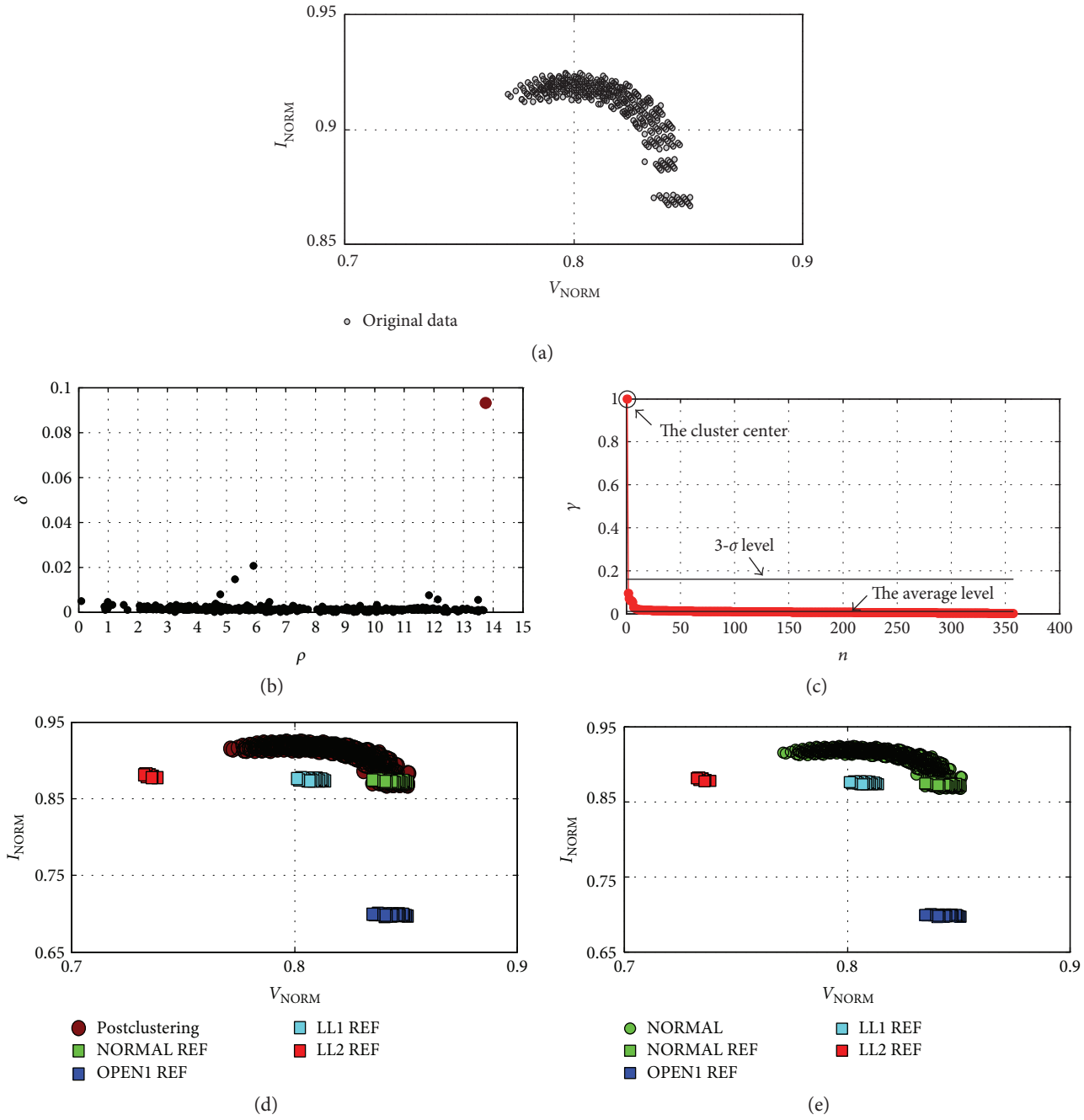

FIGURE 6: Analysis for the NORMAL case: (a) original data, (b) decision graph, (c) the value of γ in decreasing order, (d) data after clustering, and (e) cluster after identifying.

In this study, ρ_i and δ_i are normalized and employed to calculate γ_i as follows:

$$\gamma_i = \frac{\rho_i}{\rho_{\max}} \cdot \frac{\delta_i}{\delta_{\max}}. \quad (5)$$

Thus, only the data points with large γ can be selected as cluster centers. In our study, each cluster corresponds to an operational condition of the PV systems and the number of daily conditions is much smaller than the total amount of data. Therefore, the 3-sigma (3-σ) rule is applied as the criterion to automatically select the large γ and then determine the cluster centers [29].

Finally, after the cluster centers have been found, the CFSFDP algorithm constructs clusters by assigning other points to the same cluster as its nearest neighbor of higher density. The cluster assignment is performed in a single

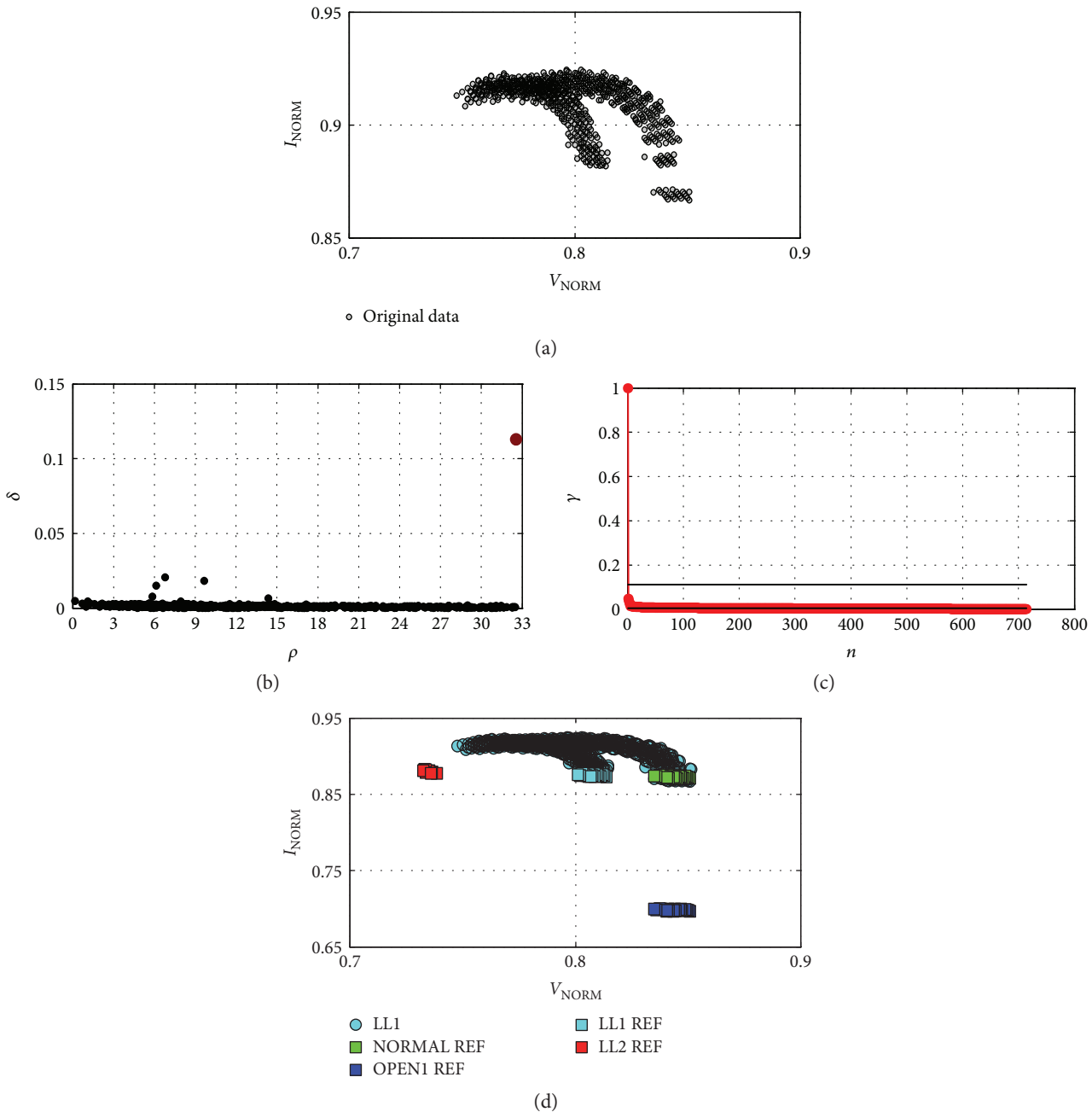

Figure 7: Analysis for the combination of NORMAL and OPEN1: (a) original data, (b) decision graph, (c) the value of γ in decreasing order, and (d) data after clustering and identifying.

step and does not require optimizing any objective function iteratively.

2.2.2. Phase 2 Cluster Classification. To identify the class of each cluster, a set of labeled reference data should be created first. From Section 2.1, PV data have a relatively great distance among different work conditions at low irradiation. Therefore, the labeled PV data obtained under low irradiation is adopted as the reference data. In addition, the reference data are obtained based on PV simulation models to avoid shortcomings that may be caused by experimental method, such as the potential safety issue and additional labor cost.

Subsequently, the minimum distance between the labeled reference data and the clusters is applied to define their correlation. Let N_R represent the number of the reference data categories and $r \in [1, N_R]$ the id of the reference data categories. Let N_C represent the number of clusters and $c \in [1, N_C]$ the id of cluster. For cluster c, the minimum distance between it and each reference data category can be expressed as a row vector:

$$D_M^c = \left[d_{c,1}, \ldots, d_{c,r}, \ldots, d_{c,N_R} \right]. \quad (6)$$

Then each element in the vector is compared with the cutoff distance d_c, respectively. If $d_{c,r} < d_c$, this illustrates that

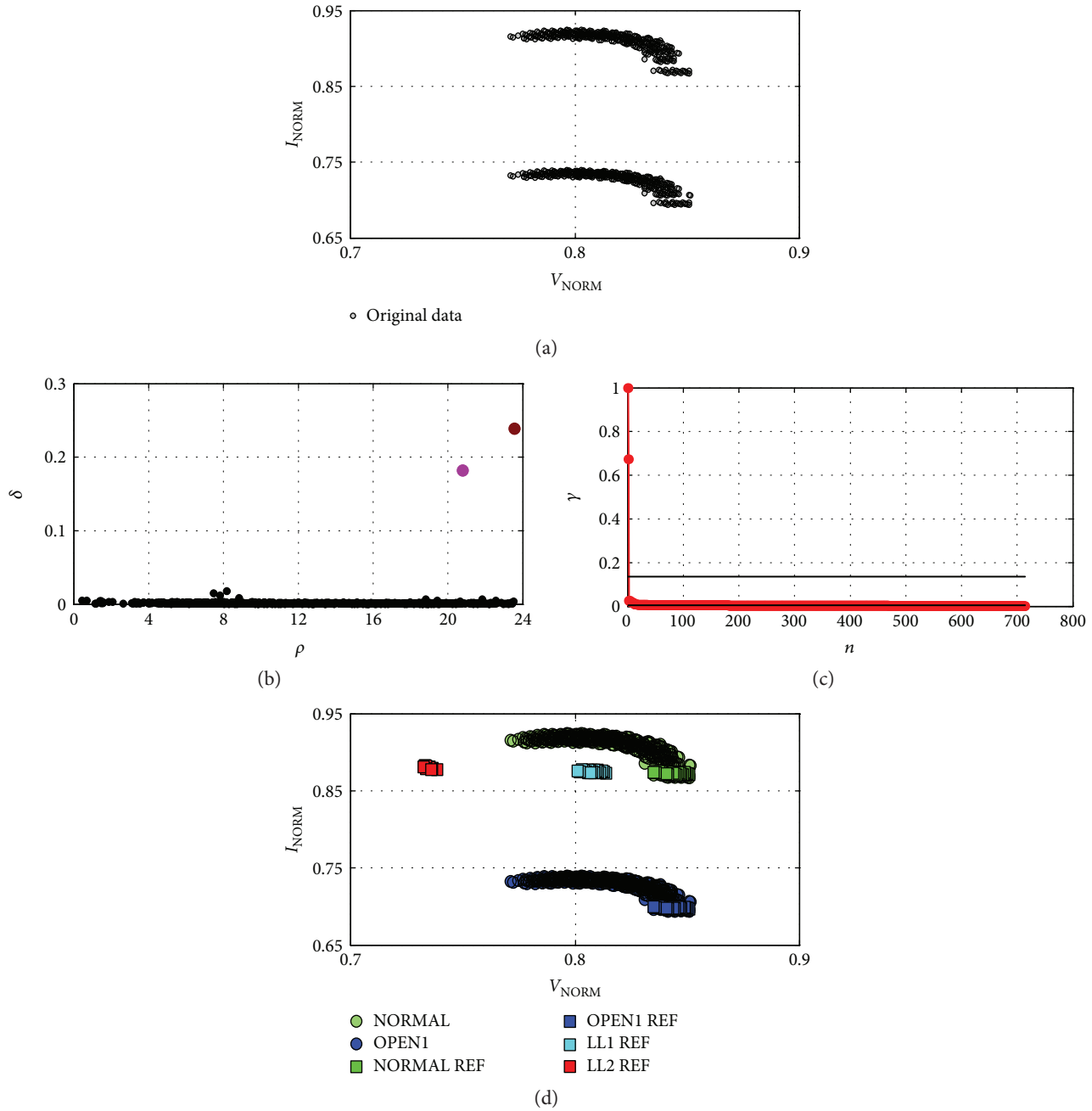

FIGURE 8: Analysis for the combination of NORMAL and LL1: (a) original data, (b) decision graph, (c) the value of γ in decreasing order, and (d) data after clustering and identifying.

the reference data of r category can be assigned to cluster c. In other words, cluster c can be labeled as r category. If all the elements are bigger than d_c, then the category of the smallest elements will be found and used to label cluster c.

Consequently, the flowchart of the proposed approach for PV array analysis is shown in Figure 4. First, the daily PV running data, that is, $X_i = [V_{NORM1}, V_{NORM2}, V_{NORM3}, \ldots, V_{NORMN}]$ and $Y_i = [I_{NORM1}, I_{NORM2}, I_{NORM3}, \ldots, I_{NORMN}]$, are recorded, and the Euclidean distance matrix is created. Subsequently, the neighborhood range of data points is selected to calculate the local density and the minimum distance between a point and any other point with higher density, namely, ρ_i and δ_i, respectively. Cluster centers are obtained based on the product of ρ_i and δ_i and then followed by the cluster assignment of all data points. Finally, clusters are classified by investigating the minimum distance between the data of each reference category and that of each cluster. According to the labeled cluster, the operating status of PV array can be identified. When a fault is detected, the alarm will be sent out if necessary.

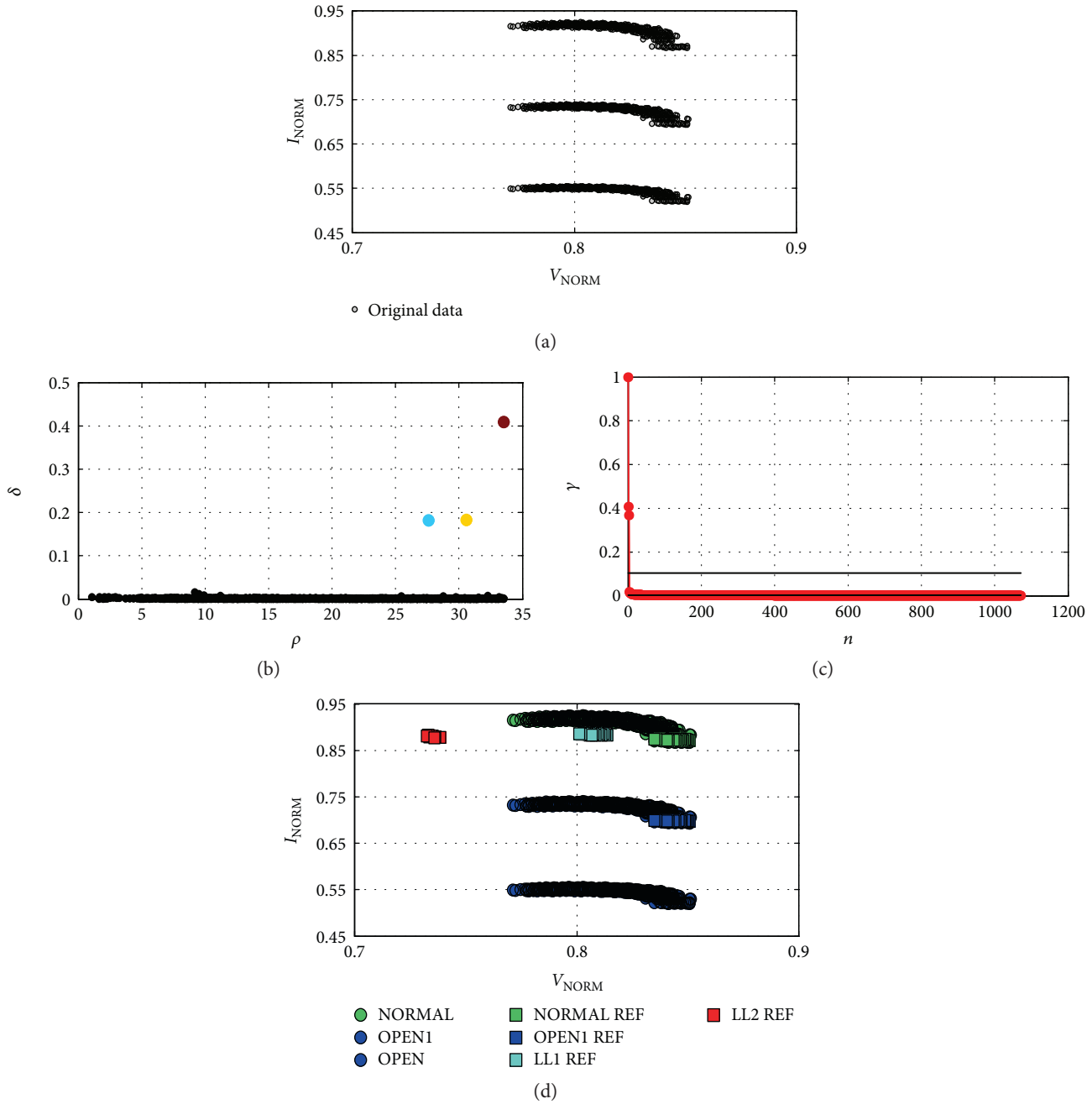

Figure 9: Analysis for the combination of NORMAL, OPEN1, and OPEN2 case: (a) original data, (b) decision graph, (c) the value of γ in decreasing order, and (d) data after clustering and identifying.

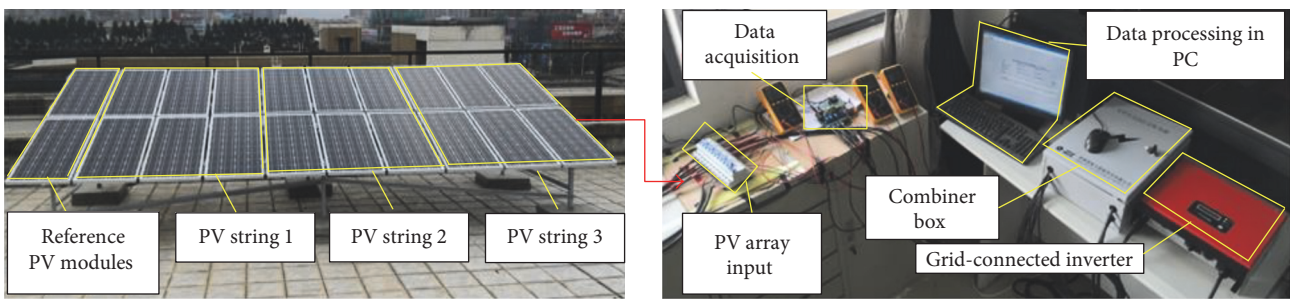

Figure 10: The experimental platform.

Table 2: Parameters of the PV components.

Components	Type	Parameters
PV modules and reference modules	Monocrystalline	At STC: P_{MPP}: 100 W, I_{MPP}: 5.71 A, V_{MPP}: 17.5 V, I_{SC}: 6.03 A, V_{OC}: 21.5 V
PV array	6×3 modules	At STC: P_{MPP}: 1.8 kW, I_{MPP}: 17.13 A, V_{MPP}: 105 V
Grid-connected inverter	Goodwe GW2500-NS	Max. output power: 2500 W; PV voltage range: 80~500 V; MPPT voltage range: 80~450 V

Table 3: Experimental environment and data.

Case	Solar irradiances	Ambient temperature	Amount of data
NORMAL	~170–1020 W/m^2	26–35°C	1380
Combination of NORMAL and LL1	~180–930 W/m^2	9–19°C	NORMAL: 710; LL1: 530
Combination of NORMAL and OPEN1	~180–920 W/m^2	5–16°C	NORMAL: 760; LL1: 500

3. Simulation and Results

In this section, several data sets are constructed to investigate the performance of the proposed method. First, the settings of simulation system are introduced. Furthermore, the test data under different conditions are simulated and briefly described. Finally, simulation results are presented.

3.1. Simulated PV System. In this study, we adopt one-diode model for PV module and apply the monocrystalline PV module SM55 to build a simulation PV system in MATLAB/Simulink [30]. The schematic diagram of the system is shown in Figure 1. The system consists of 10×5 PV modules, that is, $m = 10$ and $n = 5$. The main parameters of each PV module at standard test conditions (STC) are shown in Table 1 [31].

The module-plane solar irradiance (G_T) and ambient air temperature (T_{amb}) can be used for finding the operating solar cell temperature (T_{cell}) with the following equation [32]:

$$T_{cell} = T_{amb} + \frac{\text{NOCT} - 20^{\circ}\text{C}}{800\ \text{W/m}^2} \cdot G_T, \tag{7}$$

where NOCT is the nominal operating cell temperature of the PV module SM55 and is chosen as 45°C [31].

3.2. Simulation Data under Different Conditions. As shown in Figure 5, there are three categories in operating conditions of the PV system, that is, normal condition, line-line (LL) fault, and open-circuit (OPEN) fault. The test data are obtained by simulating a whole daily running status of the PV system. The input ambient parameters for the simulation system are as follows: the solar irradiance (G_T) widely varying from 100 to 1000 W/m^2 with step change of 50 W/m^2 and the ambient temperature (T_{amb}) changes from 0°C to 40°C with step by 1°C. The PV data (V_{NORM} versus I_{NORM}) under the three conditions are plotted in Figure 5 and analyzed as follows:

(1) Normal condition: Under the changing of solar irradiance and temperature, the PV data usually have the following operating range: V_{NORM} $(0.77,0.86)$ and $I_{NORM} \in (0.86,0.92)$.

(2) Line-line fault: The LL fault category contains two types of faults: LL1 and LL2. The LL1 fault presents that there is one-module mismatch between the fault point "Fault1" and negative conductor (Fault1-Neg) in the faulted string. Similarly, the LL2 fault is defined as two-module mismatch in the fault string. Compared with NORMAL, I_{NORM} of LL is slightly reduced, whereas V_{NORM} is observably decreased. Besides, the data of NORMAL and LL1 overlap at high solar irradiance.

(3) Open-circuit fault: the OPEN fault category consists of two kinds of faults: OPEN1 and OPEN2. They are defined as open-circuit faults on one string and two strings, respectively. It is obvious that the OPEN fault has the same V_{NORM} as the one of NORMAL condition. However, I_{NORM} is reduced in proportion according to the number of open strings.

3.3. Simulation Results. Although the daily operating temperature range of a PV system is changing, the daily normalized data of the PV system has similar data distribution. Therefore, to simulate daily running condition of the PV system, only the data obtained under a low temperature range (0°C to 20°C) is selected as the test data for analysis in this paper. The reference data are simulated under the solar irradiance of 210 W/m^2 to distinguish them from the test data. The reference data consist of four categories and are arranged in accordance with the following order: NORMAL, OPEN1, LL1, and LL2; thus $N_r = 4$ and $r \in [1, 4]$.

As discussed in Section 2.1, there may be a variety of conditions in the daily operating of the PV system. Therefore, three cases are researched, including one condition, the combination of two conditions, and the combination of three conditions. Simulation results of all cases are shown in Figures 6–9 and are discussed as follows.

(1) Case Study I: One Condition. The NORMAL condition is studied in this case, and the original test data are plotted in

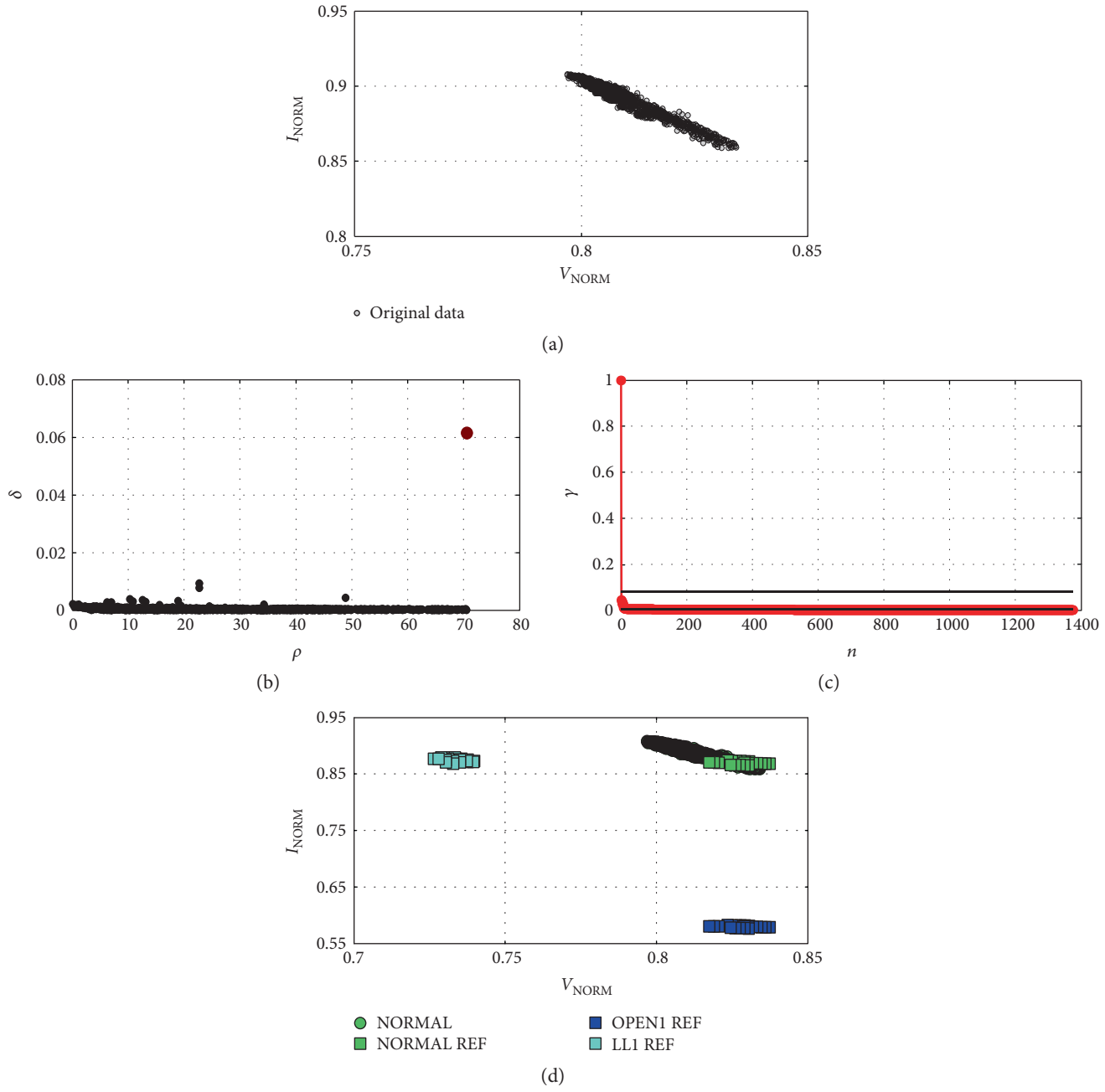

Figure 11: Experimental result of the NORMAL case: (a) original data, (b) decision graph, (c) the value of γ in decreasing order, and (d) data after clustering and identifying.

Figure 6(a) and are represented with black. According to the CFSFDP algorithm, the ρ_i and δ_i of all data points are calculated, respectively. Figure 6(b) shows the graph of δ_i as a function of ρ_i for each data point, which is called the decision graph. The γ_i in decreasing order is plotted in Figure 6(c). Compared to the 3-σ level, it is clear that only the top one can be chosen as the cluster center, indicating that there exists one cluster. Then, other points are assigned to the cluster as its nearest neighbor of higher density, as shown in Figure 6(d). The data points are colored when they belong to the cluster. It is obviously that all the test data are correctly clustered.

After the completion of the data clustering, the cluster is characterized by using the four types of reference data which are shown in Figure 6(d) with different colors. The d_c is chosen to be 0.00289 so that the average number of neighbors is around 2% of the total number of data. And the minimum distance vector D_M^1 is calculated to be [0.00013, 0.16708, 0.01753, 0.04974]. It can be concluded that the first element of D_M^1 is smaller than the d_c, so the cluster can be characterized as NORMAL and is painted with the same color of the NORMAL REF, as shown in Figure 6(e). Therefore, the test data of NORMAL condition can be accurately clustered and characterized.

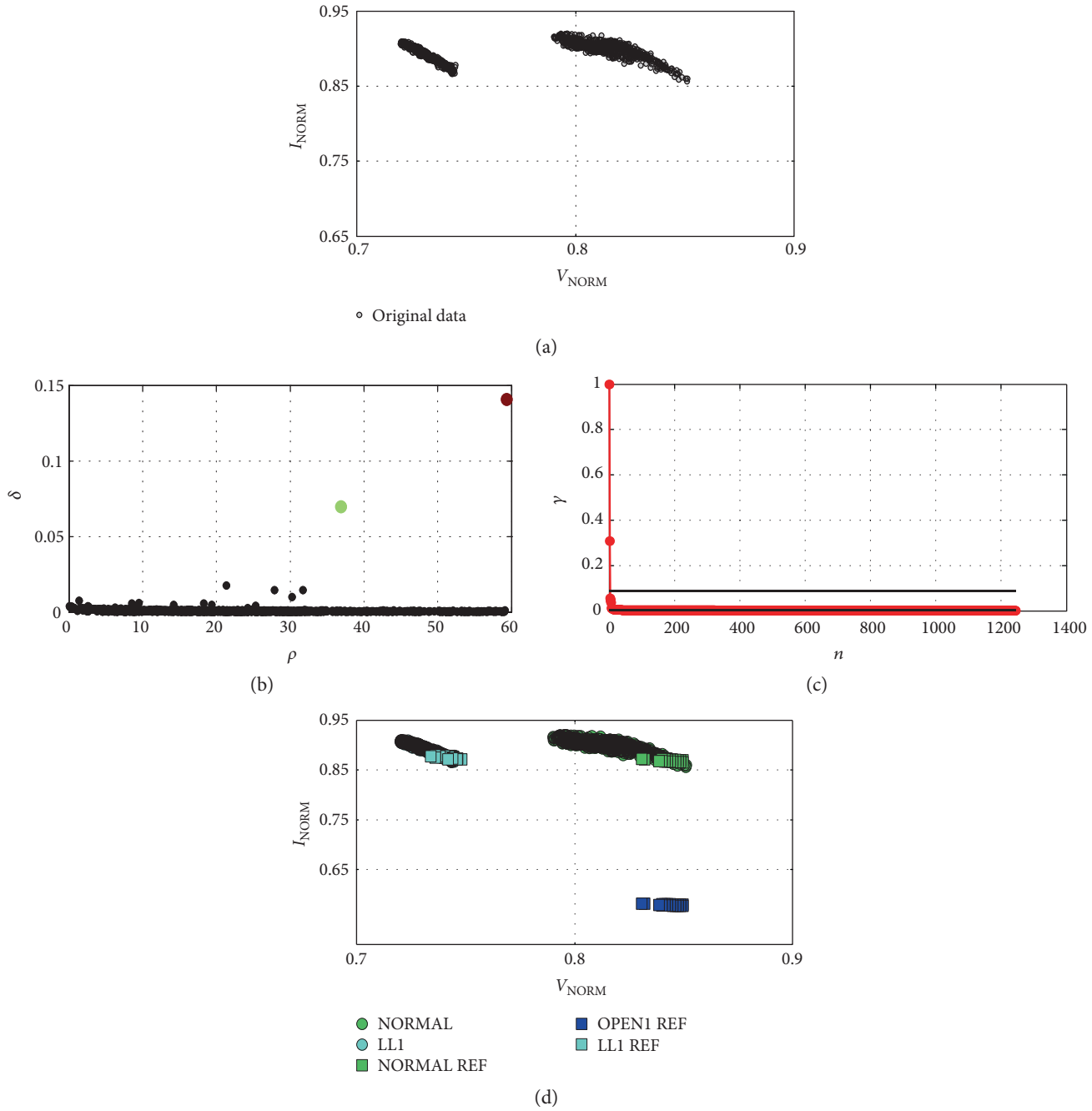

Figure 12: Experimental result of the combination of NORMAL and LL1 case: (a) original data, (b) decision graph, (c) the value of γ in decreasing order, and (d) data after clustering and identifying.

(2) Case Study II: Combination of Two Conditions. In this case, the combination of NORMAL and LL1 and the combination of NORMAL and OPEN1 are studied, respectively. For the first one, as can be seen from Figures 7(b) and 7(c), only one cluster center is found due to the NORMAL and LL1 with many data overlapping. Thus, the test data are grouped into the cluster. The minimum distance vector D_M^1 equals [0.00013, 0.16709, 0.00052, 0.03143] and the d_c equals 0.00321, which illustrates that the first and third elements in the vector are smaller than d_c. However, there is only one cluster to be identified, and the proposed approach tends to classify the cluster as LL1 fault (shown in Figure 7(d)) since the condition of PV array has changed from normal to fault.

For the second combination, as shown in Figures 8(b) and 8(c), two cluster centers are obtained. The d_c is 0.00368. For the two clusters, the D_M^1 is [0.00013, 0.16708, 0.01753, 0.04974] and D_M^2 is [0.13573, 0.00039, 0.14365, 0.14686]. Accordingly, it is clear that the first element of D_M^1 and second element of D_M^2 are smaller than d_c. Thus, the other data points are assigned to two clusters based on

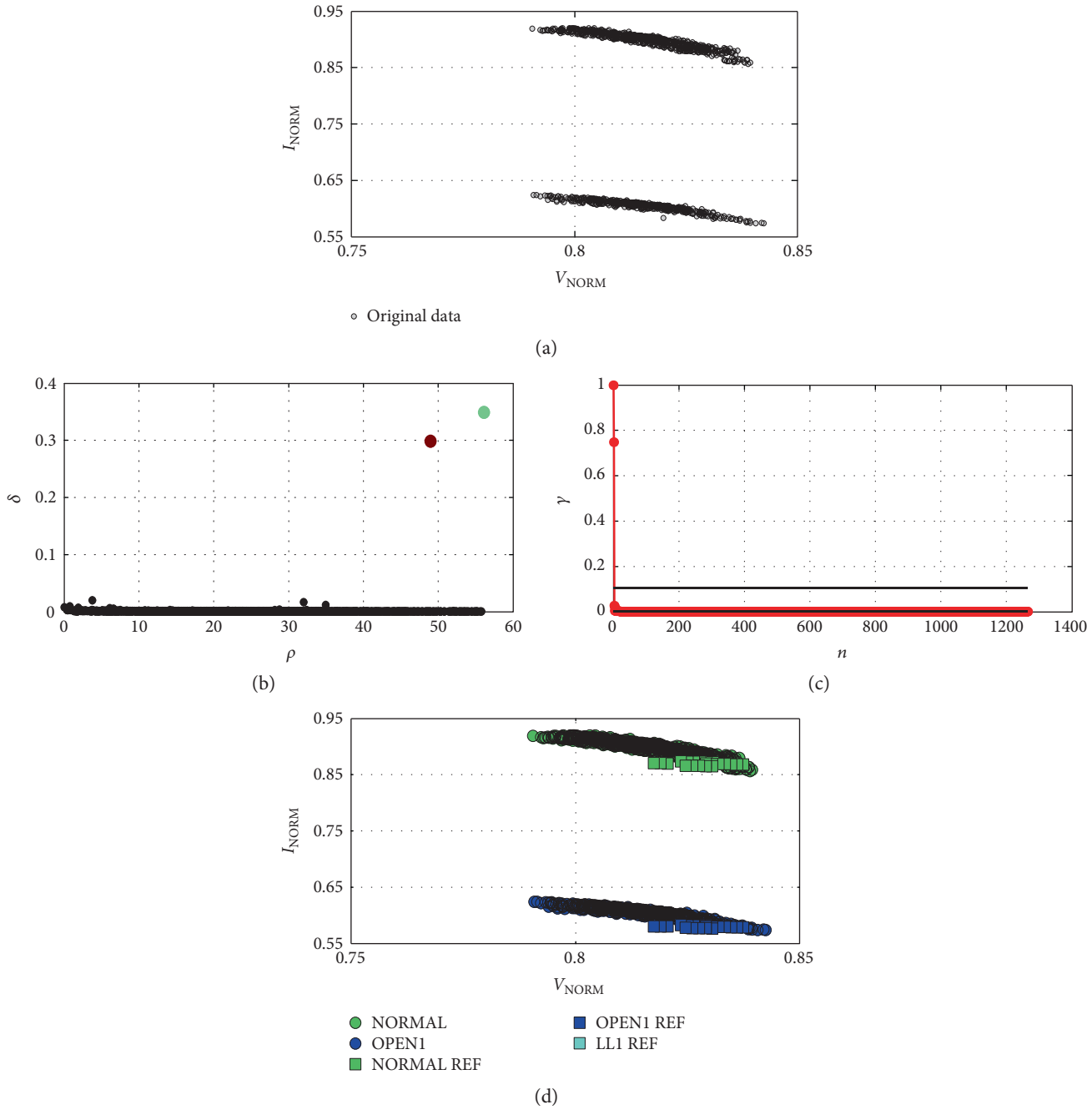

Figure 13: Experimental result of the combination of NORMAL and OPEN1 cases: (a) original data, (b) decision graph, (c) the value of γ in decreasing order, and (d) data after clustering and identifying.

the two cluster centers and recognized as NORMAL and OPEN1, respectively, as shown in Figure 8(d). Consequently, the test data of this combination can be accurately clustered and characterized.

(3) Case Study III: Combination of Three Conditions. The combination of three conditions, that is, NORMAL, OPEN1, and OPEN2, is investigated. The case represents the three conditions which successively occur in one day. Hence, there should be three data clusters. The original data is shown in Figure 9(a). From Figures 9(b) and 9(c), it is clear that three cluster centers are properly chosen, that is, $N_c = 3$. For the three clusters, the minimum distance vectors are as follows: $D_M^1 = [0.00013, 0.16708, 0.01753, 0.04974]$, $D_M^2 = [0.13573, 0.00039, 0.14365, 0.14686]$, and $D_M^3 = [0.31827, 0.14515, 0.32846, 0.32709]$. The d_c equals 0.00454; thus, it can be illustrated that the first element of D_M^1 and the second elements of D_M^2 are smaller than σ. Therefore, clusters one and two can be classified as NORMAL and OPEN1, respectively. For the third cluster, it can be found that

all the elements in D_M^3 are larger than d_c, while the second elements are the smallest. Thus, the cluster can be identified as the category of OPEN, as shown in Figure 9(d).

Consequently, the proposed approach has the ability to accurately cluster the PV data in various simulated cases and diagnoses the faults in PV arrays.

4. Experimental Results

In this section, the presented approach is tested with an experimental PV system, and the experimental platform as well as the experimental results is presented.

4.1. Experimental Platform. A 1.8 kW grid-connected photovoltaic system is applied to test the performance of the proposed algorithm under the real working conditions, as shown in Figure 10. The PV array consists of three PV strings in parallel, and each string has six modules in series. The reference PV modules have the same electrical parameters with the PV array. Moreover, it can be assumed that the PV array and the reference PV modules have the identical working environment since they are installed together. Therefore, the reference PV modules are applied real time normalizing the PV data online. The overview for parameters of components in the PV system is given in Table 2.

Three instances are implemented and studied, including NORMAL, the combination of NORMAL and LL1, and the combination of NORMAL and OPEN1. The first case is carried out in summer with a high running temperature range, and the other two cases are operated in spring with a relatively low temperature range. The detailed description about these conditions has been presented in Section 3.2. The experimental environment for the PV array and the amount of data recorded during the experiments are given in Table 3.

Besides, the reference data are obtained by using a PV simulation based on the parameters from Table 2. The reference data include three categories and are arranged in such a sequence: NORMAL, OPEN1, and LL1. The solar irradiance for the PV simulation is fixed at 200 W/m^2. According to the operating temperature range of the three cases, the ambient temperature range for the PV simulation is 21–40°C for the first case and 0–20°C for the others, respectively.

4.2. Experimental Results. Figures 11–13 illustrate the experimental results of the aforementioned three cases. It is obvious that the distribution of experimental data has remarkable clustering, which is similar to the simulated ones. For the NORMAL condition, as shown in Figure 11, only a cluster is found by the proposed approach. And D_M^1 equals [0.00011, 0.27664, 0.06441] and d_c equals 0.00132, which indicates that the cluster can be accurately categorized as NORMAL.

Second, for the second case, as can be seen from Figure 12, the data are exactly clustered into two groups. And d_c is 0.00238, D_M^1 equals [0.00018, 0.27546, 0.05649], and D_M^2 equals [0.08573, 0.29803, 0.00016]. Accordingly, it is clear that the two clusters can be recognized as NORMAL and LL1, respectively.

Finally, for the third instance, as shown in Figure 13, two clusters are exactly obtained. The d_c is 0.00242. For the two clusters, D_M^1 and D_M^2 equal [0.00055, 0.27494, 0.06779] and [0.24378, 0.00052, 0.25088], respectively. Therefore, the test data of this instance can be characterized as NORMAL and OPEN1, respectively.

Consequently, according to the experimental results, the proposed approach has the ability to cluster and classify the daily data of the PV array.

5. Conclusions

According to the distribution features of the daily operating data from a PV system, a clustering approach has been presented to identify the working conditions of the PV system and further diagnose the faults in the PV array. The proposed method has the ability to cluster the PV data and identify the clusters based on the minimum distance vector between the reference data and the clusters. Three kinds of daily work cases are simulated to validate the effectiveness of the approach, that is, the normal condition, the combination of normal condition with one fault, and the combination of normal condition with two faults. The simulated results indicate that the method can accurately cluster the PV data and identify the faults in each case. Furthermore, a grid-connected PV system is built to test the experimental performance of the developed approach. Under different temperatures and irradiation ranges, three daily operating status of the PV system are implemented and the experimental results also demonstrate the usefulness of the algorithm in a practical system.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank Dr. Ye Zhao from the Power Electronics Research Group at Northeastern University for the generous offer of valuable suggestions about photovoltaic modeling and fault analysis. This work was supported by the National Natural Science Foundation of China (Grant nos. 61574038, 31300473, 61601127, and 51508105), the Science Foundation of Fujian Education Department of China (Grant no. JAT160073), the Science Foundation of Fujian Science & Technology Department of China (Grant nos. 2015H0021, 2015J05124, and 2016H6012), the Fujian Provincial Economic and Information Technology Commission of China (Grant nos. 830020 and 83016006), and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (Grant no. LXKQ201504).

References

[1] V. Sharma and S. S. Chandel, "Performance and degradation analysis for long term reliability of solar photovoltaic systems: a review," *Renewable and Sustainable Energy Reviews*, vol. 27, pp. 753–767, 2013.

[2] *Article 690—Solar Photovoltaic Systems, NFPA70, National Electrical Code*, 2014.

[3] Y. Zhao, J. F. de Palma, J. Mosesian, R. Lyons, and B. Lehman, "Line–line fault analysis and protection challenges in solar photovoltaic arrays," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 9, pp. 3784–3795, 2013.

[4] J. Flicker and J. Johnson, "Analysis of fuses for blind spot ground fault detection in photovoltaic power systems," *Sandia National Laboratories Report*, Tech. Rep., NM, USA, 2013.

[5] Y. Hu, W. Cao, J. Wu, B. Ji, and D. Holliday, "Thermography-based virtual MPPT scheme for improving PV energy efficiency under partial shading conditions," *IEEE Transactions on Power Electronics*, vol. 29, no. 11, pp. 5667–5672, 2014.

[6] Z. Zou, Y. Hu, B. Gao, W. L. Woo, and X. Zhao, "Study of the gradual change phenomenon in the infrared image when monitoring photovoltaic array," *Journal of Applied Physics*, vol. 115, no. 4, pp. 1–11, 2014.

[7] C. Buerhop, D. Schlegel, M. Niess, C. Vodermayer, R. Weißmann, and C. J. Brabec, "Reliability of IR-imaging of PV-plants under operating conditions," *Solar Energy Materials and Solar Cells*, vol. 107, pp. 154–164, 2012.

[8] T. Takashima, J. Yamaguchi, K. Otani, T. Oozeki, K. Kato, and M. Ishida, "Experimental studies of fault location in PV module strings," *Solar Energy Materials and Solar Cells*, vol. 93, no. 6, pp. 1079–1082, 2009.

[9] T. Takashima, J. Yamaguchi, and M. Ishida, "Fault detection by signal response in PV module strings," in *33rd IEEE Photovoltaic Specialists Conference, 2008 (PVSC'08), IEEE*, pp. 1–5, California, USA, 2008.

[10] S. Silvestre, A. Chouder, and E. Karatepe, "Automatic fault detection in grid connected PV systems," *Solar Energy*, vol. 94, pp. 119–127, 2013.

[11] A. Chouder and S. Silvestre, "Automatic supervision and fault detection of PV systems based on power losses analysis," *Energy Conversion and Management*, vol. 51, no. 10, pp. 1929–1937, 2010.

[12] W. Chine, A. Mellit, A. M. Pavan, and S. A. Kalogirou, "Fault detection method for grid-connected photovoltaic plants," *Renewable Energy*, vol. 66, pp. 99–110, 2014.

[13] S. Silvestre, M. A. da Silva, A. Chouder, D. Guasch, and E. Karatepe, "New procedure for fault detection in grid connected PV systems based on the evaluation of current and voltage indicators," *Energy Conversion and Management*, vol. 86, pp. 241–249, 2014.

[14] S. Silvestre, S. Kichou, A. Chouder, G. Nofuentes, and E. Karatepe, "Analysis of current and voltage indicators in grid connected PV (photovoltaic) systems working in faulty and partial shading conditions," *Energy*, vol. 86, pp. 42–50, 2015.

[15] S. Silvestre, L. Mora-López, S. Kichou, F. Sánchez-Pacheco, and M. Dominguez-Pumar, "Remote supervision and fault detection on OPC monitored PV systems," *Solar Energy*, vol. 137, pp. 424–433, 2016.

[16] I. Yahyaoui and M. E. V. Segatto, "A practical technique for on-line monitoring of a photovoltaic plant connected to a single-phase grid," *Energy Conversion and Management*, vol. 132, pp. 198–206, 2017.

[17] Y. Zhao, L. Yang, B. Lehman, J. F. de Palma, J. Mosesian, and R. Lyons, "Decision tree-based fault detection and classification in solar photovoltaic arrays," in *Twenty-Seventh Annual IEEE Applied Power Electronics Conference and Exposition (APEC), 2012, IEEE*, pp. 93–99, Florida, USA, 2012.

[18] D. Riley and J. Johnson, "Photovoltaic prognostics and health management using learning algorithms," in *38th IEEE Photovoltaic Specialists Conference (PVSC), 2012, IEEE*, pp. 001535–001539, Texas, USA, 2012.

[19] S. Syafaruddin, E. Karatepe, and T. Hiyama, "Controlling of artificial neural network for fault diagnosis of photovoltaic array," in *16th International Conference on Intelligent System Application to Power Systems (ISAP), 2011, IEEE*, pp. 1–6, Hersonissos, Greece, 2011.

[20] Y. Wang, Z. Li, C. Wu, D. Q. Zhou, and L. Fu, "A survey of online fault diagnosis for PV module based on BP neural network," *Power System Technology*, vol. 37, no. 8, pp. 2094–2100, 2013.

[21] A. M. Pavan, A. Mellit, D. De Pieri, and S. A. Kalogirou, "A comparison between BNN and regression polynomial methods for the evaluation of the effect of soiling in large scale photovoltaic plants," *Applied Energy*, vol. 108, pp. 392–401, 2013.

[22] W. Chine, A. Mellit, V. Lughi, A. Malek, G. Sulligoi, and A. M. Pavan, "A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks," *Renewable Energy*, vol. 90, pp. 501–512, 2016.

[23] Y. Zhao, R. Ball, J. Mosesian, J. F. de Palma, and B. Lehman, "Graph-based semi-supervised learning for fault detection and classification in solar photovoltaic arrays," *IEEE Transactions on Power Electronics*, vol. 30, no. 5, pp. 2848–2858, 2015.

[24] B. Fang, X. Yin, Y. Tan et al., "The contributions of cloud technologies to smart grid," *Renewable and Sustainable Energy Reviews*, vol. 59, pp. 1326–1331, 2016.

[25] T. Hu, M. Zheng, J. Tan, L. Zhu, and W. Miao, "Intelligent photovoltaic monitoring based on solar irradiance big data and wireless sensor networks," *Ad Hoc Networks*, vol. 35, pp. 127–136, 2015.

[26] K. H. Tseng, H. J. Wu, G. H. Lin, and P. T. Cheng, "Establishment and case analysis of a photovoltaic cloud management system," in *IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), 2016, IEEE*, pp. 831–836, Hefei, China, 2016.

[27] Y. Zhao, B. Lehman, J. F. de Palma, J. Mosesian, and R. Lyons, "Fault analysis in solar PV arrays under: low irradiance conditions and reverse connections," in *37th IEEE Photovoltaic Specialists Conference (PVSC), 2011, IEEE*, pp. 002000–002005, Washington, USA, 2011.

[28] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[29] Z. A. Bakar, R. Mohemad, A. Ahmad, and M. M. Deris, "A comparative study for outlier detection techniques in data mining," in *2006 IEEE Conference on Cybernetics and Intelligent Systems, IEEE*, pp. 1–6, Bangkok, Thailand, 2006.

[30] S. M. MacAlpine, R. W. Erickson, and M. J. Brandemuehl, "Characterization of power optimizer potential to increase energy capture in photovoltaic systems operating under non-uniform conditions," *IEEE Transactions on Power Electronics*, vol. 28, no. 6, pp. 2936–2945, 2013.

[31] SHELL, *Shell SM55 Photovoltaic Solar Module*, http://www.solarquest.com/microsolar/suppliers/siemens/sm55.pdf.

[32] E. Skoplaki and J. A. Palyvos, "Operating temperature of photovoltaic modules: a survey of pertinent correlations," *Renewable Energy*, vol. 34, no. 1, pp. 23–29, 2009.