Statistical Pattern Recognition based Technique for PV Array Fault Diagnosis

Zhicheng Xu1*, and Xiaozhao Xu2

1 State Grid Energy Research Institute CO., Ltd., Beijing, 102209, China
2 Hangzhou Silan Microelectronics CO., Ltd., Hangzhou, Zhejiang, 310000, China
*Corresponding author’s e-mail: xu-zhicheng2008@163.com

Abstract. A reliable operation of photovoltaic (PV) systems should be guaranteed by implementing an effective fault detection program so that the maintenance schedule can be arranged to repair or replace the faulty component in time. This paper presents a PV array fault diagnosis method by applying the statistical pattern recognition technique. Proper feature parameters are selected first to characterize the operation condition of a PV array. The fault data set is then collected under varying operating conditions. Once the fault categories are defined, the genetic algorithm-based fuzzy C-means (GAFCM) clustering algorithm is employed to obtain the clustering center of each fault category so that the fuzzy mapping relationship between chosen feature parameters and its corresponding fault type is established. In this way, the fault type of a PV array can be judged immediately once the values of characteristic parameters are given. Finally, a Gaussian distribution-based membership function is developed to calculate the similarity between test samples and the defined fault categories. The fault type of the test sample can be recognized by judging from the largest similarity.

1. Introduction
The photovoltaic (PV) power generation is playing a more and more important role in modern power systems with its ever-expanding share in power systems [1]. It is a challenge to ensure efficiency and safe operation of PV systems. Unfortunately, their performance is usually affected by various failures which may result in catastrophic events such as fire accidents [2]. In all the failures, PV array faults lead to the majority of the malfunctions of PV systems due to the operation of PV arrays is usually influenced by external environment (such as shading, soiling, disastrous weather) and the component failure (such as open fault, short fault, ground fault, line-to-line fault, arc fault, and polarity mismatch and so on) [3]. In order to improve the efficiency and reliability of a PV system, a real-time and effective fault detection technique for a PV array is urgently required to ensure the performance of PV systems [4]. At present, some achievements on PV array fault diagnosis methods, which mainly focused on the detection of the hard fault, have been made in [5–7]. These methods can be classified into two categories: hardware-based fault detection methods (HFD) and algorithm-based fault detection methods (AFD). The HFD and AFD methods are usually designed for specific fault type, so they cannot determine the type and location of a failure at the same time in the most situations. Moreover, they may have weak robustness when the size or topology of a PV array changed.

In this paper, a fault diagnosis method based on statistical pattern recognition is developed to detect the fault type and location of a PV array in PV module level with high accuracy in real-time and online. The genetic algorithm-based fuzzy C-means (GAFCM) clustering method is applied for the
statistical pattern recognition. In the proposed fault diagnosis method, the fuzzy relationship between kinds of fault categories is firstly established, and the characteristic parameters is selected by GAFCM technique. After getting the fault categories cluster centers, a membership function based-on normal distribution is applied to describe the memberships between the test samples and the predefined fault types. Finally, the fault category is determined by sorting the memberships.

2. GAFCM-based PV array fault diagnosis process

2.1. PV array feature extraction and fault data collection

Proper characteristic parameters should be selected first to reasonably describe the different operating conditions of a PV array, and a minimum number of characteristics parameters should be determined in order to guarantee a low dimensionality of the feature space. In order to determine the operation of a PV array output, three key points should be given, that is $(0, I_{sc}), (V_{oc}, 0)$ and $(V_m, I_m)$. The $I-V$ and $P-V$ characteristic curves of the PV array can be roughly plotted after getting these three points. Therefore, $I_{sc}, V_{oc}, V_m, and I_m$ are selected as the characteristic parameters and they are expressed in a vector form $(V_m, I_m, I_{sc}, V_{oc})$. In this paper, five kinds of fault modes are mainly discussed: normal operation, module open circuit, module short circuit, shading effect, and multiple failures. The multiple failures here indicate that two or more failures exist simultaneously in a PV array. For each fault type, the corresponding original data set is collected by operating the PV array under different ambient temperature and irradiance levels. After all fault types are simulated, the fault data set $X$ can be established by features extraction and the training set can be expressed as follows:

$$X = \{x_1, x_2, \ldots, x_n\} = \begin{bmatrix} V_{n1} & I_{n1} & I_{sc1} & V_{oc1} \\ V_{n2} & I_{n2} & I_{sc2} & V_{oc2} \\ \vdots & \vdots & \vdots & \vdots \\ V_{nm} & I_{nm} & I_{scn} & V_{ocm} \end{bmatrix} \tag{1}$$

Where, $x_i = (V_{mi}, I_{mi}, I_{sci}, V_{oci})$ represents the $i$-th fault sample characteristics vector, and the dimension of $x_i$ equals the number of fault characteristic parameters. In this way, data sets for each type of faults were recorded under varying temperature and irradiance levels. The typical value of a feature vector for each fault type can be obtained by applying the statistical pattern recognition from many samples. It means that if the value of the feature vector is known, the possible fault type of PV array may be inferred.

2.2. Application of GAFCM-based pattern recognition on PV array fault detection

The GAFCM-based technique is applied to establish the relationship between a fault type and its typical value of feature vector. The standard FCM algorithm converges to local minimum easily and it is sensitive to initial values. In order to overcome these shortcomings, a genetic algorithm based FCM (GAFCM) is introduced in this paper. The standard FCM is implemented to obtain the local optimal classification results. Then the genetic algorithm is used to get rid of individuals which may fall into their local optimum. This search process is repeated until an optimal solution is found. The detailed GAFCM process can be found in [8,9]. In this way, the PV array fault diagnosis process is implemented as following steps.

If we define the sub-categories as $F_1, F_2, \ldots, and FC$, where $C$ is the number of FCM clusters, it means the PV fault data set $X$ will be divided it into $C$ sub-fault categories by applying the GAFCM approach. A fuzzy mapping relationship between fault categories and corresponding feature parameters is established by this analysis. The clustering centers $c_j = (V_{mj}, I_{mj}, I_{sci}, V_{oci}), j=1,2,\ldots,C$ can be used as typical values of these fault modes. If a test sample is given, we math its feature values to each clustering centers $c_j$ and judge which one has the highest match degree. The possible fault type of the test sample can be determined by choosing the fault type with highest match degree.
Let \( \mathbf{x} = (V_m, I_m, I_s, V_{oc}) \) be the test sample for diagnosis. In order to judge which fault category it belongs to, the memberships between \( \mathbf{x} \) and all kinds of fault clustering centers \( c_j \) should be computed. A Gaussian distribution-based membership function is developed for PV array fault pattern recognition. For fault type \( F_j \), the Gaussian distribution function of characteristic parameter \( c \) of the test sample \( \mathbf{x} \) can be expressed in (2):

\[
\mu_{F_j} (c) = \exp \left[ -\frac{(c - c_o)}{2\sigma^2} \right] \times 100\%
\]

(2)

Where, characteristic parameter \( c \) is one of \( V_m, I_m, I_s, \) and \( V_{oc} \), \( \mu_{F_j} (c) (j=1,2,…,C) \) denotes the membership (namely the similarity) between \( c \) and the corresponding component \( c_o \) of clustering center vector \( c_j \) obtained by GAFCM; \( \sigma \) is the standard deviation of characteristic parameter \( c \), and it can be determined based on historical data set as follows

\[
\sigma = \frac{y_{\text{max}} - y_{\text{min}}}{6}
\]

(3)

Where, \( y_{\text{max}}, y_{\text{min}} \) represent the maximum and minimum values of characteristic parameters \( c \). (3) is based on the Gaussian distribution “3σ rule” which indicates that about 99.7% of values drawn from a normal distribution are within three standard deviations away from the mean. After getting the Similarity of each component using (2), the similarity vector \( \mu_{F_j} (\mathbf{x}) \) of test sample \( \mathbf{x} \) is shown as follows:

\[
\mu_{F_j} (\mathbf{x}) = (\mu_{F_j} (V_m), \mu_{F_j} (I_m), \mu_{F_j} (I_s), \mu_{F_j} (V_{oc}))
\]

(4)

The total Similarity between all feature parameters of the test sample \( \mathbf{x} \) and fault type \( F_j \) is denoted as \( \mu_{\text{total}} \), and it is computed by taking a weighted average of each component as the following equation:

\[
\mu_{\text{total}} = \sum \alpha \times \mu_{F_j} (c), c \in \{V_m, I_m, I_s, V_{oc}\}
\]

(5)

Where \( \alpha \) is the weight of component \( c \) of the feature vector \( \mathbf{x} \) and it meets \( \sum \alpha = 1 \). The total similarities \( \mu_{\text{total}} (j=1, 2,…, C) \) between \( \mathbf{x} \) and all \( C \) kinds of fault types is computed, and they are arranged in descending order. Then the fault category of the test sample \( \mathbf{x} \) can be judged by choosing the fault type with maximum similarity. Figure 2 shows the flowchart of the PV array fault diagnosis based on GAFCM.
3. Results and discussion

3.1. Fault data collection
A simple PV array system is developed by Matlab/Simulink to extract the values of characteristic parameters under different irradiance and temperature levels. The PV module model in [10] is employed, and the detailed meaning of each parameter can be found in [10]. A SunTech STP270-24/Vd PV module is simulated using this model. 72 solar cells are connected in series to get $P_{sc} = 270\text{Wp}$, $V_{oc} = 35.0\text{V}$, $I_{sc} = 7.71\text{A}$, $V_{mp} = 44.5\text{V}$ and $I_{mp} = 8.2\text{A}$. In the simulation, each PV module is installed with a bypass diode and a blocking diode. The PV module is established using the $S$-function in Matlab software. Figure 3 illustrates respectively the $I$-$V$ and $P$-$V$ characteristics of simulated PV module under 1000W/m$^2$, 800W/m$^2$ and 600W/m$^2$ at 25°C cell temperature. It is proved that this model is accurate enough and effective to model the actual PV module by the data tags in figure 3. In order to model the distributed rooftop PV system, a PV array system demonstrated in figure 4 with $3 \times 2$ the series-parallel connection is constructed in Matlab/Simulink. It is noted that the irradiance and temperature of each module can be set independently, and the temperature port is encapsulated in the single module. The PV system generates a maximum power of 1619.8W at 105V rated output voltage and 15.43A rated output current at standard test condition (STC) condition (cell temperature 25°C and irradiance 1000W/m$^2$).
Based on this PV array system, the fault categories can be divided into 13 categories which are described in Table 1.

| Fault mode          | Fault categories | Description                                                                 |
|---------------------|------------------|-----------------------------------------------------------------------------|
| Normal operation    | F1               | All modules are under normal operation with uniform temperature and irradiance |
| short fault         | F2               | One module short circuit in any one branch                                  |
|                     | F3               | Two modules short circuit in any one branch                                 |
|                     | F4               | One module short circuit in each one branch                                 |
| open fault          | F5               | One branch open circuit                                                     |
| Shading             |                  |                                                                             |
| bimodal             | F6               | One branch under uniform temperature and irradiance; another has two different irradiances but the same temperature |
|                     | F7               | Each branch has two different irradiances                                   |
| Three peaks         | F8               | One branch with two irradiances, another with three irradiances             |
|                     | F9               | Each branch with three irradiances                                          |
| Multiple failures   | F10              | One branch open circuit, another with two irradiances                       |
|                     | F11              | One branch with two irradiances, another under uniform irradiance but one module short circuit |
|                     | F12              | Each branch with two irradiances and one module short circuit in each branch |
|                     | F13              | Each branch with three irradiances and one module short circuit in each branch |

The output voltage, current, and power will vary at the different ambient environment with respect to all fault categories. Consequently, this case is studied in the range of cell temperature of 25°C, 60°C and irradiance 700W/m², 800W/m². In this simulation, the maximum power point (MPP) of the PV system under partial shading operations can be recorded using a ramp-based controlled voltage source shown in Figure 4. Whereas in a practical engineering, the MPP under partial shading can be calculated by the PSO approach [11]. In order to simulate the different operation states of each fault type under different conditions, four varying levels in 25°C, 60°C and five different levels in 700W/m², 800W/m² are evenly spaced selected to get 20 scenarios. In this way, 20 samples for each fault category are simulated in the varied atmosphere to get a data set with 13×20=260 elements.
3.2. PV fault diagnosis process

GAFCM is implemented to cluster the dataset \( \mathbf{X} \) in (1) into 13 distinct groups with the GAFCM parameters setting as follows: clusters number \( C = 13 \), population size \( n = 30 \), crossover probability \( P_c = 0.8 \), mutation probability \( P_m = 0.01 \), maximum number of iterations \( L = 1000 \), and minimum amount of improvement \( \varepsilon = 10^{-3} \), and the weight value of each component \( \alpha \) is set equal to 1/4 because the four characteristic parameters have the same degree of importance. Running the simulation model under different operation conditions and the feature parameters are recorded as test samples for diagnosis. Clustering center vectors (\( V_{m\alpha}, I_{m\alpha}, I_{c\alpha}, V_{oc\alpha} \)) of all kinds of fault categories are shown in table 2.

Table 2. Clustering centers and test sample memberships.

| fault categories | \( V_{m\alpha} \) | \( I_{m\alpha} \) | \( I_{c\alpha} \) | \( V_{oc\alpha} \) | \( \mu(V_m) \) | \( \mu(I_m) \) | \( \mu(I_c) \) | \( \mu(V_{oc}) \) | \( \mu_{total} \) |
|-----------------|-----------------|-----------------|-----------------|-----------------|----------------|----------------|----------------|----------------|----------------|
| F1              | 96              | 10.88           | 11.66           | 120             | 0.016          | 0.787          | 0.723          | 0.036          | 0.390          |
| F2              | 66              | 12.42           | 13.42           | 87              | 0.971          | 0.200          | 0.138          | 0.390          | 0.575          |
| F3              | 33              | 11.1            | 11.66           | 42              | 0.030          | 0.697          | 0.723          | 0.006          | 0.364          |
| F4              | 66              | 12.1            | 13.33           | 81              | 0.971          | 0.294          | 0.156          | 0.956          | 0.594          |
| F5              | 96              | 5.44            | 5.83            | 120             | 0.016          | 0.006          | 0.008          | 0.036          | 0.017          |
| F6              | 95.63           | 11.37           | 13              | 119.8           | 0.018          | 0.581          | 0.233          | 0.037          | 0.217          |
| F7              | 101.5           | 10.08           | 12.35           | 120.4           | 0.004          | 0.993          | 0.447          | 0.033          | 0.369          |
| F8              | 99.78           | 9.23            | 10.74           | 120.3           | 0.006          | 0.889          | 0.982          | 0.034          | 0.478          |
| F9              | 100.1           | 10.06           | 11.85           | 122.7           | 0.005          | 0.994          | 0.647          | 0.021          | 0.417          |
| F10             | 97.44           | 5.24            | 6.29            | 119.4           | 0.011          | 0.004          | 0.020          | 0.040          | 0.019          |
| F11             | 64.75           | 10.7            | 11.24           | 87.77           | 0.991          | 0.853          | 0.872          | 0.980          | 0.924          |
| F12             | 63.8            | 10.73           | 12.37           | 78.45           | 0.999          | 0.842          | 0.439          | 0.887          | 0.792          |
| F13             | 63.64           | 12.77           | 14.34           | 80.47           | 0.999          | 0.124          | 0.033          | 0.944          | 0.525          |

The maximum and minimum values of all feature parameters are selected from the data set and they are given in table 3. The standard deviations of each characteristic parameter are respectively 11.420, 1.400, 1.488, and 13.537 which can be easily calculated by (3) according to the maximum and minimum values.

Table 3. Maximum and minimum values of feature parameters.

| \( V_m/N \) | \( I_m/A \) | \( I_{c\alpha}/A \) | \( V_{oc}/N \) |
|------------|------------|-------------------|--------------|
| Min        | 33         | 4.43              | 5.41         | 42           |
| Max        | 101.5      | 12.8              | 14.3         | 123.2        |

Therefore, the Gaussian distribution membership function of each feature parameter can be expressed as follows:

\[
\mu(V_m) = \exp \left[ - \frac{(V_m - V_{m\alpha})^2}{260.833} \right] \times 100\%
\]

\[
\mu(I_m) = \exp \left[ - \frac{(I_m - I_{m\alpha})^2}{3.920} \right] \times 100\%
\]

\[
\mu(I_{c\alpha}) = \exp \left[ - \frac{(I_{c\alpha} - I_{c\alpha\alpha})^2}{4.428} \right] \times 100\%
\]

\[
\mu(V_{oc}) = \exp \left[ - \frac{(V_{oc} - V_{oc\alpha})^2}{366.501} \right] \times 100\%
\]

The Gaussian distribution-based function in (2) is used to calculate the similarity metric between test samples and 12 clustering centers. The failure of the test sample can be judged by selecting the fault type with maximum similarity. Let \( x = (V_m, I_m, I_{c\alpha}, V_{oc}) = (63.23, 9.91, 10.46, 85.07) \) as a concrete example for diagnosis. The memberships of all four characteristic parameters are calculated by (4) and
and they are presented in table 2. It can be concluded from the last column of table 2 that the test sample belongs to fault type F11 due to it has the highest membership. The example \( x = (63.23, 9.91, 10.46, 85.07) \) is obtained under the situation “one PV module string with two irradiances, another branch with two module under the same irradiance and one module short circuit”, and this situation is accordance with fault mode F11 presented in table 1. The result indicates that the proposed GAFCM method is efficient to judge the fault type. In engineering application, if the practical fault is not consistent with the detected one, then the fault type ranked the second (i.e. F12) may be the possible fault. In this way, this proposed method can indicate the potential failure of a PV array.

Running the simulation model in various irradiance and temperature levels, we can get a dataset containing 195 test samples. A comparison of K-means, standard FCM and GAFCM are shown in table 4 to verify the efficiency of the proposed method. It shows that the GAFCM method has higher accuracy and is more effective.

| Method | Misclassified samples | Diagnosis accuracy/% |
|--------|-----------------------|----------------------|
| K-means | 35 | 82.05 |
| FCM | 20 | 89.70 |
| GAFCM | 13 | 93.30 |

In this case study, a small size PV system is used to illustrate the fault detection process. Without loss of generality, it is also applicable to a large PV system.

4. Conclusions

In this paper, a GAFCM-based statistical pattern recognition approach is proposed for PV array fault detection. It is effective in detecting the type and location of a failure of the PV module at the same time. A fuzzy mapping relationship between fault categories and corresponding feature parameters is established by applying GAFCM method. Furthermore, a membership function based on Gaussian distribution can effectively represent the volatility and uncertainty of feature parameters due to the intermittent nature of PV output.

A distributed rooftop PV system is used to illustrate the effectiveness of the proposed method. Moreover, this small size system can be regarded as a fault diagnosis unit so that it can be applied to the fault diagnosis of a larger size PV system, which is composed of many units. The algorithm has good scalability; only the number of clustering C needs to be modified if fault categories change. What is more, the order of each similarity is not only able to determine the type of the fault but also can be used to predict potential failures. If the diagnosed fault type is inconsistent with the actual fault, it can be inferred that the secondary similarity fault type may occur. According to the compared results, the proposed approach is simple and accurate, and it can detect the hard failure and soft failure as well. In the near future work, the proposed method should be conducted and verified on a practical test PV system.

Acknowledgments

This work is supported by Science and Technology Project of State Grid Corporation of China (B3670118B001): Research on Key Problems of Energy Storage Development Adapting to Large-scale Integration of Renewable Energy.

References

[1] Ding M., Xu Z., Wang W., Wang X., Song Y., Chen D.(2016) A review on China's large-scale PV integration: Progress, challenges and recommendations. Renew. Sustain. Energy Rev., 53: 639–652.
[2] Falvo M. C., Capparella S. (2015) Safety issues in PV systems: Design choices for a secure fault detection and for preventing fire risk. Case Stud. Fire Saf., 3: 1–16.

[3] Alam M. K., Khan F., Johnson J., Flicker J. (2015) A Comprehensive Review of Catastrophic Faults in PV Arrays: Types, Detection, and Mitigation Techniques. IEEE J. Photovoltaics, 5(3): 982–997.

[4] Firth S. K., Lomas K. J., Rees S. J. (2010) A simple model of PV system performance and its use in fault detection. Sol. Energy, 84(4): 624–635.

[5] Hu Y., Gao B., Song X., Tian G. Y., Li K., He X. (2013) Photovoltaic fault detection using a parameter based model. Sol. Energy, 96: 96–102.

[6] Silvestre S., Chouder A., Karatepe E. (2013) Automatic fault detection in grid connected PV systems. Sol. Energy, 94: 119–127.

[7] Gokmen N., Karatepe E., Celik B., Silvestre S. (2012) Simple diagnostic approach for determining of faulted PV modules in string based PV arrays. Sol. Energy, 86(11): 3364–3377.

[8] Biju V. G., Mythili P. (2012) A genetic algorithm based fuzzy C mean clustering model for segmenting microarray images. Int. J. Comput. Appl., 52(11):42-48.

[9] Wikaisuksakul S. (2014) A multi-objective genetic algorithm with fuzzy c-means for automatic data clustering. Appl. Soft Comput., 24: 679–691.

[10] Baltas P. (1996) The Arizona University photovoltaic designer program (ASUPVD). Dep. Electr. Comput. Eng. Arizona state Univ.

[11] Miyatake M., Veerachary M., Toriumi F., Fujii N., Ko H. (2011) Maximum power point tracking of multiple photovoltaic arrays: a PSO approach. IEEE Trans. Aerosp. Electron. Syst., 47(1): 367–380.