Fault diagnosis and classification for photovoltaic arrays based on principal component analysis and support vector machine

L C Chen\textsuperscript{1,2}, P J Lin\textsuperscript{1,2,4}, J Zhang\textsuperscript{1,2}, Z C Chen\textsuperscript{1,2}, Y H Lin\textsuperscript{3}, L J Wu\textsuperscript{1,2} and S Y Cheng\textsuperscript{1,2}

\textsuperscript{1}College of Physics and Information Engineering, and Institute of Micro-Nano Devices and Solar Cells, Fuzhou University, 350116 Fuzhou, China
\textsuperscript{2}Jiangsu Collaborative Innovation Center of Photovoltaic Science and Engineering, 213164 Changzhou, China
\textsuperscript{3}College of Computer and Information Sciences, Fujian Agriculture and Forest University, Fuzhou 350002, China

E-mail: linpeijie@fzu.edu.cn

Abstract. In order to dig out more typical features of photovoltaic (PV) with multitudinous characteristic parameters, and realize fault diagnosis and classification for PV arrays effectively. A method based on principal component analysis (PCA) has been proposed in this paper. At first, the data set of PV array is processed by PCA and then a transform matrix is produced. Second, the processed data will be classified by supporting vector machine (SVM). Finally, a classification model will be built. Two sets of data, collected from PV simulation system and actual PV array, are adopted to examine this method. The result shows that the method is able to recognize four kinds of states accurately (normal, open circuit, short circuit and partial shadow). Consequently, the fault of PV array can be diagnosed and classified.

1. Introduction

For the threat of greenhouse effect and energy shortage, the importance of fossil energy reducing has been recognized. To protect human beings from environmental pollution and meet the growing demand for energy around the world, solar energy has become a popular research orientation among various kinds of renewable energy. Nowadays, crystalline silicon is widely used to produce the main commercial solar cell that covers 90\% of the market [1]. In addition, the rest of the market is occupied by thin film solar cell that has lower manufacturing cost but more efficient [2]. With the development of PV devices manufacturing technology and the reducing of cost, the installed capacity of PV systems has reached 2017GW in 2016. The rapid increasing of PV systems mitigate worldwide energy crisis, but the maintenance of PV systems is a challenge to technicians. For the bad working conditions of PV array, such as ultraviolet radiation, extreme weather, shadow covered and so on [3]. These working conditions will lead to numerous faults in PV array. The common faults are classified as line to line (L-L) fault, open circuit, arc fault or partial shadow [4].

To prevent the failure from affecting the operation of PV array, researchers have proposed different kinds of methods to diagnose fault of PV array. Recently, the methods based on machine learning and data mining have become the focus of this research area due to their ability of solving nonlinear
problem of PV system [5]. For example, methods based on wavelet packets are introduced to detect fault [6]. However, it’s hard to confirm which kind of fault happens. Furthermore, a six-layer detection algorithm has been used to detect different faults. The result is given by a fuzzy logic classification system and the accuracy is increased up to 98.8% [7]. A method use fractional-order color relation classifier is able to classify different faults such as mismatch fault, bridge fault, open circuit and so on [8]. In our another work, a method based on clustering by fast search and find of density peaks (CFSFDP) is introduced to detect and classify different faults [9]. Moreover, optimized kernel extreme learning machine has been used to improve the effect of classification [10].

The output of PV array contains lots of characteristic parameters. These characteristic parameters will be changed when a fault happens. Such changes will exist as long as the fault continues. By analysing the output of PV array, we suppose that different working states of PV array have their unique characteristic. In order to dig out multitudinous characteristic parameters of PV array for the more typical features, and realize fault diagnosis and classification for PV arrays effectively, the method of PCA has been used in this paper to process the dataset measured from PV array. In addition, PCA can also reduce the dimensionality of the dataset. So that the volume of dataset will be decreased as well.

In this paper, the sections are arranged as follows: Section 2 introduce the process of PCA and present the peculiarity of SVM briefly. Section 3 expound the fault diagnosis and classification for PV arrays with this method. The several working conditions of PV array, simulation results and experiment results are presented in Section 4. Finally, some conclusions are drawn in Section 5.

2. Methodology
PCA is a kind of linear conversion which is used for simplifying dataset. By analysing a data set composed of interrelated quantitative correlation variables, the purpose of PCA is to extract the important information from the dataset and represent it as a set of new orthogonal variables, which is called the main component [11]. The implementation method of PCA is to project the dataset into a new coordinate system, the maximum variance of the input dataset falls on the first coordinate (called the first principal component), the second largest variance falls on the second coordinate (the second principal component), and so on. According to this principle, PCA will come in handy when we try to reduce the dimensions of the dataset and preserve the characteristics which contribute most to variance. After being processed by PCA, the data set will be sent to SVM to classify.

2.1. Implementation of PCA
Assume there is a \( N \times M \) matrix \( X \) which contains \( M \) characteristic parameters \( x_i \), and each characteristic parameter contains \( N \) measurements. The \( x_i \) is defined as equation (1).

\[
x_i = [x_{i1}, x_{i2}, \ldots, x_{iN}]^T \quad (i = 1, 2, \ldots, M)
\]  

Moreover, the \( N \times M \) matrix \( X \) is expressed as equation (2).

\[
X = [x_1, x_2, \ldots, x_M]
\]

To process the matrix \( X \) by PCA, the covariance matrix of \( X \), which indicated as \( C_X \), should be obtained. First, we should get rid of the mean value of the matrix \( X \) and obtain the zero-mean-value matrix \( \bar{X} \) as equation (3)

\[
\bar{X} = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_M] \\
= [x_{11} - \bar{x}_1, x_{21} - \bar{x}_2, \ldots, x_{M1} - \bar{x}_M]
\]

where \( \bar{x}_i \) is the mean value of \( x_i \), it is calculated as equation (4).
\[ \bar{x}_i = \frac{1}{N} \sum_{j=1}^{N} x_{i,j} \quad (i = 1, 2, \cdots, M) \]  

(4)

Now the covariance matrix \( C_x \) is calculated as equation (5).

\[ C_x = \frac{1}{M-1} [ \bar{X}^T \bar{X} ] \]  

(5)

The covariance matrix \( C_x \) is \( M \times M \) matrix. Its principal diagonal elements are the variance of each characteristic parameter. The other elements of matrix \( C_x \) are the covariance between any two characteristic parameters. The covariance reflects the noise and redundancy in the measurements [12].

After the covariance matrix \( C_x \) has been calculated, it will be performed Eigen-decomposition. Assume there are matrix \( Q \) and matrix \( \Sigma \), the relationship between \( C_x, Q \) and \( \Sigma \) is shown as equation (6)

\[ C_x = Q \Sigma Q^{-1} \]  

(6)

where the matrix \( Q \) is constituted by the eigenvectors of \( C_x \), and the matrix \( \Sigma \) is a diagonal matrix constituted by the eigenvalues of \( C_x \). There is a one-to-one correspondence between eigenvectors and eigenvalues. Sorting eigenvalues in descending order, we choose the first \( k \) eigenvalues (\( k \leq M \)) and the eigenvectors which are corresponding to the chosen eigenvalues. The \( M \times k \) projection matrix \( P \) is built by these chosen eigenvectors. The method to select value of \( k \) will be introduced in section 3. In addition, the result of PCA is expressed as equation (7).

\[ A = XP \]  

(7)

Furthermore, when \( k = M \), the covariance matrix of matrix \( A \) is identical to the matrix \( \Sigma \). Commonly, the eigenvalues of matrix \( C_x \) reflect the variance of each characteristic parameter in matrix \( A \).

2.2. Support Vector Machine (SVM)

SVM is a powerful and universal machine learning method which is widely used in classification and regression. For the nonlinear classification, the main idea is to map the input vector to a high-dimensional feature space non-linearly. In this feature space, a linear decision surface is constructed. The high generalization ability of SVM is ensured by the special properties of the decision surface [13]. After the classification has finished in high-dimensional space, the result will be mapped into the original space.

Since PCA can only process and improve the classification of dataset. In order to further classify the different working conditions of PV array, a classification model is needed. SVM adopts the Structure Risk Minimization to achieve the optimal generalization ability and avoid over-fitting by balancing the error of training set and maximizing the classification interval. It can solve practical problems such as small samples and non-linearity well. Therefore, SVM is applied in this paper to achieve this purpose. In this paper, the classification model will be trained by the toolbox LIBSVM which compiled in MATLAB [14].

3. Fault diagnosis and classification for PV arrays

To diagnose the failure of PV array by PCA, the output data of PV array should be collected from the very beginning. According to the dataset measures from PV array, we find that the measurements of some characteristic parameters change a lot in different working states of PV array (such as different faults). But they have tiny variation in same working state. Specially, some of working states have similar characteristics. As shown in figure 1, maximum power point current (\( I_{MPP} \)) and maximum...
power point voltage ($V_{MP}$) of PV array are difficult to classify in normal state, open circuit, short circuit and shade. Therefore, plenty of characteristic parameters are necessary to classification. Then the PCA will be used to find out the most expressive way of transformation. Further, the separability and robustness will be improved as well.

In this case, seven characteristic parameters as $I_{MP}$, $V_{MP}$, $I_{SC}$, $V_{OC}$, $P_{MP}$, $T_c$ and $G$ are collected to build the original dataset. Where $I_{MP}$ is the current of PV array at the maximum power point, $V_{MP}$ is the voltage of PV array at the maximum power point, $I_{SC}$ is the short-circuit current of PV array, $V_{OC}$ is the short-circuit voltage of PV array, $P_{MP}$ is the power of PV array at the maximum power point, $T_c$ is the temperature of PV module and $G$ is the irradiance. Moreover, $I_{SC} \cdot V_{OC}$ is added into the original dataset, which contains $N$ samples, as a characteristic parameter. Then the original dataset will be randomly divided into the training dataset, matrix $X$, and the testing dataset, matrix $Y$, on a proportional basis.

Refer to the process of PCA introduced in section 2.1, normally we make $k<8$ and produce an $M \times k$ projection matrix $P$. For $k$ is the number of eigenvalues and corresponding eigenvectors that will be applied. Relatively, the rest of eigenvectors and eigenvalues will be removed. In theory, the value of $k$ should be as small as possible to obtain a lower-dimensions result. However, some of the eigenvectors with small eigenvalues still contribute to classification. These eigenvectors and eigenvalues should be remained as well. Hence, to determine the value of $k$ better, the classification effect under different values of $k$ is checked via SVM and the $k$ with the highest classification accuracy is selected.

It should be noticed that PCA project the dataset only but never change the location of the data in the matrix. Hence we make the training matrix $X$ and the testing matrix $Y$ multiply by the projection matrix $P$ in turn and will get matrix $A$ and matrix $B$. They are expressed as equation (7) in section 2.1 and equation (8).

$$B = YP$$

The matrix $A$ will be imported into the SVM to train a classification model after normalization. To verify the effectiveness of the classification model, it’s used to classify the matrix $B$ which has been normalized. The flow chart of fault diagnosis and classification for PV array is shown in figure 2.

In above process, seven classification models and corresponding projection matrixes will be produced. We will select a model and the corresponding projection matrix with the best classification effect. After getting the best classification model and projection matrix, the new dataset of PV array will be processed directly by multiplying the projection matrix and do not have to redo the process of PCA. It will simplify the operation and achieve the purpose of fault detection by means of the
classification model. Moreover, if there are several models have the same classification accuracy, the model with the smallest $k$ value is selected as the best solution.

![Flow chart of fault diagnosis and classification for PV array](image)

**Figure 2.** Flow chart of fault diagnosis and classification for PV array.

4. **Simulation and experiment**

The result of simulation and experiment will be displayed in this section. Furthermore, the detail of simulation system and experimental platform will be listed as well.

4.1. **Simulation system and experimental platform**

The PV array simulation model established by MATLAB/Simulink is built for testing. The model has 5×10 PV modules, of which 10 modules are installed in a series as a string, and 5 identical strings are connected in parallel to form an array. Under STC, the maximum output power of the simulation photovoltaic array is 2750 W, the open-circuit voltage is 217 V and the short-circuit current is 17A.

A 1.8 kW grid-connected photovoltaic system is applied to test the performance of the proposed method under the real working conditions, as shown in figure 3. This PV array consists of 18 PV modules, of which 6 modules are installed in a series as a string, and 3 strings are connected in parallel to an array. In addition, two separate PV modules are used as the reference modules, one for collecting the open circuit voltage and the other for collecting short-circuit current. The parameters of the experimental photovoltaic array are shown in table 1.

![Experiment platform of PV system](image)

**Figure 3.** Experiment platform of PV system.
Table 1. Parameters of the experimental photovoltaic array.

| Devices       | Model | Parameters (STC)       |
|---------------|-------|------------------------|
| PV Array      | GL-100| \(V_{MPP}=105V\)       |
|               |       | \(I_{MPP}=17.1A\)      |
|               |       | \(V_{OC}=129V\)        |
|               |       | \(I_{SC}=18.1A\)       |

4.2. Verification by simulation data
The daily working states simulated by the simulation system are include: (1) normal, (2) one of the strings open circuit (Open1), (3) one module in a string short circuit (Short1), (4) one module in a string shadow covered (Shade1). We mix data from four states and define labels for each state so that they can be trained and classified by SVM. In addition, a certain amount of data is randomly selected from each state as the test data, and these test data are mixed to form the data matrix. And the rest is the training data, which consists of the training data matrix. The simulation irradiance is 200-1000 W/m². Each working state has 310 samples for training and 59 samples for testing. The accuracy of classification under different \(k\) is shown in table 2.

Table 2. Classification accuracy of simulation.

| \(k\) | Correct/Total | Accuracy       |
|-------|---------------|----------------|
| 2     | 184/236       | 77.9661%       |
| 3     | 188/236       | 79.661%        |
| 4     | 201/236       | 85.1695%       |
| 5     | 225/236       | 95.339%        |
| 6     | 230/236       | 97.4576%       |
| 7     | 236/236       | 100%           |
| 8     | 236/236       | 100%           |

According to table 2, the highest classification accuracy is 100% when \(k=7\) or \(k=8\). Therefore, the value of \(k\) is selected as 7. The classification accuracy of simulation proves the validity of the proposed approach. After being processed by PCA, the feature vector of the training data becomes 7 dimensions, which is illustrated in figure 4 by using the boxplots. The figure 4 has 7 sub-figures, each sub-figure represents a dimension of the feature vector. The horizontal label represents the working states of PV array (1: normal, 2: Open1, 3: Short1, 4: Shade1). From figure 4, the following phenomenon can be observed. For the second dimension, the data sample of Open1 state is larger than...
other states. For the fifth dimension, the data sample of each state is significantly different. For the sixth dimension, the data sample of Short1 state is the largest. Therefore, these dimensions contribute a lot to enhance the classification accuracy due to their outstanding separability.

4.3. Verification by experiments
To further test the performance of the proposed method, an experimental work is conducted. The daily working states and the specific operation are the same as the simulation verification. The measured irradiance is 200-1000 W/m². Each working state has 269 samples for training. Normal state and Open1 state have 53 samples for testing. Shade1 and Short1 has 52 samples for testing. The accuracy of classification under different \( k \) is shown in table 3.

| \( k \) | Correct/Total | Accuracy   |
|-------|---------------|------------|
| 2     | 123/210       | 58.5714%   |
| 3     | 143/210       | 68.0952%   |
| 4     | 160/210       | 76.1905%   |
| 5     | 180/210       | 85.7143%   |
| 6     | 210/210       | 100%       |
| 7     | 210/210       | 100%       |
| 8     | 210/210       | 100%       |

Following the classification results which are enumerated in table 3, the classification accuracy reaches 100% when \( k=6 \), \( k=7 \) or \( k=8 \), respectively. Thus, the value of \( k \) is selected as 6 and the feature vector of the training data is 6 dimensions. The boxplots of feature vector are shown in figure 5. Similar to simulation, the horizontal label represents the working states of PV array and the 6 sub-figures represent 6 dimensions of the feature vector. For the first and second dimensions, the data sample of Op1en1 state can be distinguished from other states. For the third dimension, the data sample of normal state is the largest. For the fifth dimension, the data sample of Shade1 state is the smallest. Moreover, in the sixth dimension, the data sample of Short1 state is the largest. The experimental results also prove the efficiency of the proposed model.

5. Conclusion
A fault diagnosis and classification method based on PCA and SVM is presented in this paper. The output data of PV array is processed by PCA to improve their separability. Moreover, the dimensions
of dataset can be reduced. After that, SVM is adopted to train a classification model. Moreover, to determine the value of $k$, the effect of classification model with different $k$ is studied and the best $k$ is chosen. Moreover, if there are several models have the same classification accuracy, the model with the smallest $k$ value is selected as the best solution. The aim of this method is to obtain a corresponding projection matrix and a model with the best classification. The simulated and experimental results indicate that the proposed method can accurately classify four types of working states of PV array, i.e. normal, open circuit, short circuit and partial shadow.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (Grant Nos. 61574038, 61601127, and 51508105), the Science Foundation of Fujian Science & Technology Department (Grant Nos. 2018J01774, 2018J01645, and 2016H6012) and the Science Foundation of Fujian Education Department (Grant No. JAT160073), and the Fujian Provincial Economic and Information Technology Commission of China (Grant Nos. 83016006 and 830020).

References
[1] Pavan A M, Vergura S, Mellit A and Lugi V 2017 Explicit empirical model for photovoltaic devices. Experimental validation Sol. Energy 155 647-53
[2] Lineykin S, Averbukh M and Kuperman A 2014 Issues in modeling amorphous silicon photovoltaic modules by single-diode equivalent circuit IEEE T. Ind. Electron. 61 6785-93
[3] Kichou S, Silvestre S, Nofuentes G, Torres-Ramirez M, Chouder A and Guasch D 2016 Characterization of degradation and evaluation of model parameters of amorphous silicon photovoltaic modules under outdoor long term exposure J. Energy 96 231-41
[4] Alam M K, Khan F, Johnson J and Flicker J 2015 A comprehensive review of catastrophic faults in PV arrays: types, detection, and mitigation techniques IEEE J. Photovolt. 5 982-97
[5] Youssef A, El-Telbany M and Zekry A 2017 The role of artificial intelligence in photovoltaic systems design and control: A review Renew. Sust. Energ. Rev. 78 72-9
[6] Kumar B P, Ilango G S, Reddy M J B and Chilakapati N 2018 Online fault detection and diagnosis in photovoltaic systems using wavelet packets IEEE J. Photovolt. 8 257-65
[7] Dhimish M, Holmes V, Mehrdadi B and Dales M 2017 Diagnostic method for photovoltaic systems based on six layer detection algorithm Electr. Pow. Syst. Res. 151 26-39
[8] Kuo C L, Chen J L, Chen S J, Kao C C, Yau H T and Lin C H 2017 Photovoltaic energy conversion system fault detection using fractional-order color relation classifier in microdistribution systems IEEE T. Smart Grid 8 1163-72
[9] Lin P J, Lin Y H, Chen Z C, Wu L J, Chen L C and Cheng S Y 2017 A density peak-based clustering approach for fault diagnosis of photovoltaic arrays Int. J. Photoenergy 2017 Article ID 4903613
[10] Chen Z C, Wu L J, Cheng S Y, Lin P J, Wu Y and Lin W C 2017 Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and IV characteristics Appl. Energ. 204 912-31
[11] Abdi H and Williams L J 2010 Principal component analysis Wiley Interdisciplinary Reviews Computational Statistics 2 433-59
[12] Shlens J 2014 A tutorial on principal component analysis Statistics 51 219-26
[13] Cortes C and Vapnik V 1995 Support-vector networks Mach Learn 20 273-97
[14] Chang C C and Lin C J 2011 LIBSVM: A library for support vector machines Acm. T. Intel. Syst. Tec. 2 1-27