Online photovoltaic fault detection method based on data stream clustering

Yuqiao Cai1,2, Peijie Lin1,2, Yaohai Lin3, Qiao Zheng1,2, Shuying Cheng1,2, Zhicong Chen1,2, and Lijun Wu1,2

1 School of Physics and Information Engineering, and Institute of Micro-Nano Devices and Solar Cells, Fuzhou University, Fuzhou, China
2 Jiangsu Collaborative innovation Center of Photovoltaic Science and Engineering, Changzhou, China
3 College of Computer and Information Sciences, Fujian Agriculture and Forest University, Fuzhou, China
E-mail: linpeijie@fzu.edu.cn

Abstract. As the core component of solar power station, PV array is particularly important for safe and stable operation of the entire system. The existence of PV array faults for a long time can lead to potential danger of the entire PV system. Since the PV data is greatly affected by the environment, the continuous data stream generated during the operation of PV arrays can form clusters of arbitrary shape. When a PV fault occurs, new data streams can form a cluster that is different from the one under normal operation. Accordingly, this paper presents a model for online fault detection of PV arrays faults using data stream clustering approach. The real-time data stream of the PV arrays is transmitted to the diagnostic system through RabbitMQ server for online detection and data storage. The online density-based spatial clustering of applications with noise (DBSCAN) algorithm is used for clustering the data. Then, the faults are detected by judging whether new clusters are formed. The experiment result shows the effectiveness of the proposed method in grid-connected PV system.

1. Introduction

PV array is composed of several PV panels connected in series and parallel. The average life of PV panels is 20-30 years [1]. Solar power stations are often located in places of harsh environment. Therefore, PV arrays are easily affected by environmental factors, such as high temperature, strong wind, heavy rain and hail, which can cause corrosion or damage to the surface of PV components. Being exposed to ultraviolet light for a long time will accelerate the aging process of PV components. Faults in PV array include: short circuit fault, partial shading, aging degradation, open circuit fault, arc fault, hot spot, etc. In order to eliminate the impact of faults in PV array, conventional protection devices are usually installed on the DC side of PV array [2], such as overcurrent protection device (OCPD), ground fault detection and interrupter (GFDI), arc fault circuit interrupter (AFCI) and so on. However, it is difficult for conventional PV array protection devices to detect early-stage fault and adapt to a wide variety of fault types. Moreover, the MPPT technique implemented in inverter can also affect the effectiveness of these protection devices. Thus, an online PV fault detection system is one of the key issues that is needed to be solved urgently.
Recently, researchers have proposed some solutions to online fault diagnosis of PV arrays. Infrared thermal imaging equipment is applied to visually reflect the thermal image of the PV array. Specific image segmentation and feature recognition technology are used to find the temperature anomaly in the thermal image to define the fault type and location [3]. The advantage of the infrared image is that it has no effect on the structure of the PV array. However, for the limitations of the infrared image, such as high noise and uneven imaging, it is impossible to effectively distinguish the fault types when the PV array is complex [4]. When a component in PV arrays fails, the output voltage and output current of the entire array are affected. Comparing the actually measured IV curve under working conditions with the theoretical analysis results, the presence or absence of faults in the array can be accurately detected. The method of PV fault diagnosis based on machine learning (ML) is widely used in recent years. Generally, this method aims to apply artificial intelligence algorithm to process the electrical and environmental dataset detected by the sensors in PV array. After this, an accurate model is established to map the relationship between fault features and fault types. Then the faults are easily identified. The typical supervised ML algorithms applied in this field includes Decision Tree (DT) [5], Random Forest (RF) [6], Artificial Neural Network (ANN) [7], Support Vector Machine (SVM) [8], Extreme Learning Machine (ELM) [9], Probabilistic Neural Network (PNN) [10] and so on. The supervised ML methods rely on labeled data to build model, but it is too expensive to obtain the labels of large number of data. For data stream, the effectiveness of the supervised diagnostic method depends on the quality of the training data. Semi-supervised usually only needs to acquire a small number of labeled samples and unsupervised ML algorithms can work without labeled data. The more common semi-supervised and unsupervised ML algorithms are: Clusters based on density peaks [11], K-means clustering Class [12], Unsupervised probabilistic neural network [13]. Because of the continuous time sequence and variability of of PV arrays data stream, unsupervised method is more suitable for fault detection of data stream. Generally, the PV array in operating is affected by environmental factors, thus the output under different environmental conditions represents a wide range of differences. Therefore, it is difficult to collect a large number of labeled samples for supervised online detection in practice. In this paper, an unsupervised online clustering detection method is proposed, which includes the transmission, storage and online detection part. The RabbitMQ is applied to design the transmission and storage part. This part transmits the collected PV data stream to the online detection system and stores it in the local hard disk or database. The online DBSCAN algorithm is adopted to cluster the normalized PV data in the online detection part. Then a fault is detected by judging whether the PV data stream forms a new cluster.

2. The proposed method of online detection

In this section, the design of RabbitMQ, online DBSCAN algorithm, the normalization method of PV data and the fault detection method proposed in this paper are introduced respectively.

2.1. RabbitMQ

AMQP (Advanced Message Queuing Protocol) is an application-level advanced message queuing protocol based on asynchronous processing. It is designed for message-oriented middleware and it is independent from language and development platform [14]. AMQP can provide a standard messaging service that enables communication between qualified client applications and messaging middleware agents. The AMQP model is shown in the figure 1, where $P$ is the producer of the message, $X$ is the message switch, and $C$ is the recipient of the message. Since AMQP is Internet protocol, the producer, consumer, and message broker of the message are capable of existing on different platforms or devices in the process.
RabbitMQ is a message broker software based on AMQP and written in Erlang. RabbitMQ can receive and deliver messages. As a medium for messaging systems, RabbitMQ can provide a common messaging and receiving platform for users' applications and ensure the security of messages transmission.

To realize the transmission and storage of real-time PV data, RabbitMQ is used to build a data stream transmission and storage system by its high availability cluster model. The platform for transmitting and storing PV data stream is presented in figure 2. Three virtual hosts are established on experimental PC, one for the producer and two for consumers. The real-time data stream of the PV array enters the RabbitMQ system, and the producer is responsible for transmitting the data stream to consumer 1 and consumer 2 at a certain set rate. Consumer 1 receives the PV data from the normalization process of the producer (see 2.3 for the normalization method), and then performs online detection of the normalized data stream using the online DBSCAN algorithm (online DBSCAN see 2.2). Consumer 2 receives the PV data transmitted from the producer, and the data can be stored in the local hard disk or database as needed.

2.2. Online DBSCAN algorithm
As one of the popular density-based clustering algorithm, the density-based spatial clustering of applications with noise (DBSCAN) is broadly used [15]. The main working mechanism is to collect samples with a certain density, storing them in a cluster, through which the degree of similarity among samples in the same cluster can reach a level that is as high as possible. The degree of difference between samples in different clusters is as large as possible. The DBSCAN algorithm is especially useful for processing large data sets with noise. The DBSCAN algorithm has been proven to be effective for analyzing large amounts of heterogeneous complex data [16]. In order to be adapted on the online clustering of data stream of PV data, the online DBSCAN algorithm is proposed. Compared with DBSCAN algorithm, the online DBSCAN algorithm can quickly analyze the existing clustering
results when processing data stream. Online DBSCAN only updates the clusters in each clustering process by searching the clusters closest to the current data. Therefore, it reduces the time complexity and improves the efficiency of online clustering. In dealing with the problem of outliers, there may be data that were previously distinguished as outliers. After new data entering the system, some outliers may have the same data distribution as the currently existing clusters. After each clustering, if an outlier is found, it will iterate all current outliers to reaffirm whether there are new clusters formed by the outliers. So online DBSCAN algorithm can dynamically identify outliers and new clusters.

Some basic concepts about DBSCAN that will be used below are as follows:

- **Data vector**: Assuming that the data stream arrives at time \( t \), the \( n \)-dimensional data vector received by the system at time \( t \) is represented as \( x_t = \{x_{1t}, x_{2t}, \ldots, x_{nt}\} \).

- **eps neighborhood**: A user-specified parameter \( \text{eps} > 0 \) is used to specify the radius of the field for each object. The neighborhood of object \( o \) is a space centered on \( o \) and radius with \( \text{eps} \).

- **minPts**: The number of samples in the \( \text{eps} \) neighborhood of the object, it is also called the density threshold of the \( \text{eps} \) neighborhood of object.

- **Core point**: For the specified object \( p \), if the number of points contained in the \( \text{eps} \) neighborhood of \( p \) is greater than or equal to \( \text{minPts} \), the object \( p \) is referred to as a core point.

The clustering process of the online DBSCAN algorithm is as follows:

1. When \( x_t \) is entered the system, the DBSCAN compares the distance between \( x_t \) and the current cluster centers, and finds the cluster center closest to \( x_t \);
2. Calculate the number of samples contained in the \( \text{eps} \) neighborhood of \( x_t \), and compare it with the \( \text{minPts} \) set by the system. If the number is more than \( \text{minPts} \), then integrate the sample as a core point into the current cluster, and then update the cluster center of the cluster;
3. If the value is smaller than \( \text{minPts} \), first determine whether the core point is included in the \( \text{eps} \) neighborhood of \( x_t \). If it is included, the sample is integrated into the cluster where the core point is located, and the core point and cluster center of the cluster are updated. If not included, the sample is classified as an outlier;
4. When a new outlier is found, the system iterates through all the outliers that currently exist, and determines whether there are a certain number of outliers to form a new cluster.

### 2.3. Normalization of PV data

The schematic diagram of a typical grid-connected PV system is shown in figure 3. The system usually includes \( s \times p \) PV array, inverters, protection devices (OCPD and GFPD) and connection wires. In order to better analyze the PV data, the concept of a reference module, which is a PV panel independent from the PV array being used for testing, is introduced.

The component works in the same operating environment as the PV array under test (the same temperature and solar irradiance, etc.). It has the same component parameters with the PV array under test. The reference module is used to collect open circuit voltage (\( V_{\text{OC}} \)) and short circuit current (\( I_{\text{SC}} \)). The \( V_{\text{OC}} \) and \( I_{\text{SC}} \) of the entire PV array under test are estimated, and \( G \) indicates the irradiance on PV panel. In standard test condition (STC), \( G_{\text{STC}} = 1000 \text{ W/m}^2 \). Through the \( V_{\text{OC}} \) and \( I_{\text{SC}} \) of the array under test, the \( V_{\text{MPP}} \) and \( I_{\text{MPP}} \) can be normalized. The normalization method [17] is shown in equation (1).

\[
\begin{align*}
    V_{\text{NORM}} &= \frac{V_{\text{MPP}}}{s \times V_{\text{OC-REF}}} \\
    I_{\text{NORM}} &= \frac{I_{\text{MPP}}}{p \times I_{\text{SC-REF}}} \\
    G_{\text{NORM}} &= \frac{G}{G_{\text{STC}}}
\end{align*}
\]

Where \( V_{\text{NORM}} \) is the normalized voltage of the PV array under test, \( I_{\text{NORM}} \) is the normalized current, \( G_{\text{NORM}} \) is the normalized irradiance. \( V_{\text{OC-REF}} \) is the open circuit voltage of the reference module, \( I_{\text{SC-REF}} \)
is the short circuit current of the reference module, and $V_{MPP}$ and $I_{MPP}$ are respectively voltage and current at which the PV array under test is operating at the maximum power point currently. $s$ indicates the number of components in series, and $p$ indicates the number of parallel strings of the PV array. The normalization method is simple in calculating, fast in processing, and the normalized PV data has strong discriminability.

![Figure 3. Schematic diagram of grid-connected PV system](image)

During the operation of PV array, as the environmental factors change, the voltage ($V_{MPP}$) and current ($I_{MPP}$) at the maximum power point will change accordingly. The distribution of the $V_{MPP}$ and $I_{MPP}$ of the simulated PV array in the range of varying solar irradiance is shown in figure 4. The data of different states of the PV array have high overlap and cannot be effectively classified.

![Figure 4. Unnormalized PV data](image)
Figure 5. Normalized PV data

The normalized PV data are shown in the figure 5. It can be seen that the normalized PV data are clustered obviously, and the shape of the cluster is mostly columnar. Thus, density-based clustering algorithm is used to cluster such data.

2.4. Procedure of online detection

The flow of the online detection method for PV faults proposed in this thesis is shown in figure 6. The real-time data stream from the PV array enters the Rabbitmq system, which transfers the PV data into the online detection system and stores it. When outliers are defined, the system outputs outlier information; when a new cluster is recognised, it outputs new cluster information and alarms.

Figure 6. Flowchart of the proposed model
3. Experimental results and analysis
In this section, we tested the effectiveness of the proposed method by using a 1.8 kW grid-connected PV system under actual operating conditions. As shown in figure 7, the PV array platform is a 3×6 PV array consisting of 6 identical PV modules connected in series as a string and 3 strings in parallel. Single crystal solar cells are used in PV modules. The output of the PV array is connected to a GW2500-NS PV grid-connected inverter through the confluence box. The detailed parameters of the PV array and inverter are shown in Table 1.

![Figure 7. PV Experiment platform](image)

| DEVICE      | PARAMETER                     |
|-------------|-------------------------------|
| PV ARRAY    | GL-100                        |
|             | MPP=1800W, VMPP=105V, IMPP=17.1A, VOC=129V, ISC=18.1V |
|             | Rated DC Power: 2700W         |
|             | Maximum Input Voltage: 500V   |
| GRID INVERTER| GW2500-NS                     |
|             | Starting Voltage: 80V         |
|             | Maximum Power Point Voltage   |
|             | Maximum DC Current: 18A       |

Figure 8 shows how the diagnostic system distinguish outliers from a new cluster during fault detection, where the time interval between \( t_1 \) and \( t_0 \) is 51 seconds, \( eps=0.12, minPts=20 \). As shown in Figure 8(a), the diagnostic system first collects 50 normal running PV data to form an initial cluster, and then sets the OPEN1 fault (one strings of the PV array has an open circuit fault) on the running PV array after the diagnostic system runs for a period of time. The green dot and the red dot indicate the
data of the normal working state and the outlier data by online DBSCAN, respectively. Figure 8(a) is the diagnostic diagram of the PV array at time $t_0$. It can be seen from the figure that the detection system has discovered 6 outliers at time $t_0$. However, there are not enough outliers to diagnose a new cluster at time $t_0$, the system does not report the formation of a new fault, but it still outputs the outlier information. The specific information of each outlier can be obtained through the diagnostic system, including the formation time of the outliers and their corresponding array output values. Figure 8(b) shows the detection results of the diagnostic system at time $t_1$. It can be found that a new cluster has been formed at time $t_1$, some of the outliers that are diagnosed at the time $t_0$ are included in the new cluster. When a new cluster is found, the diagnostic system outputs the formation time of the cluster, the current clustering center and the size of the cluster.

The detection results of the diagnostic system for different types of PV faults are shown in figure 9.

Figure 9(a) is the detection result when two PV strings of open circuits occur in the system (OPEN2). The output voltage of the PV array basically remains unchanged compared with the normal state, but the output current is significantly reduced as the number of open circuit strings increases. Figure 9(b)-(c) are diagnostic results of faults in which the PV modules in the PV array are accidentally connected to the strings of the strings. Figure 9(b) represents a case where one PV module is short-circuited (Line-Line 1), and figure 9(c) represents a case where two PV modules are short-circuited (Line-Line 2). The experiment result shows that four typical PV faults (OPEN1, OPEN2, Line-Line 1 and Line-Line 2) can be accurately detected.

Figure 8. Detection of OPEN1 at different times (a) Clustering at time $t_0$; (b) Clustering at time $t_1$
4. Conclusion
In this paper, an online PV fault detection method based on data stream clustering is proposed. The effectiveness of the method is verified by a 1.8kW grid-connected PV system. This method uses RabbitMQ system to transfer and store PV data stream. The RabbitMQ system holds the advantages of high scalability and recoverability. The data collected can be processed in parallel and permanently retained. In the online clustering stage of PV data stream, the online DBSCAN algorithm can process the normalized PV data online efficiently, accurately distinguishing outliers from fault clusters. The detection system can define the typical electrical fault of PV arrays in time.

The detection system proposed in this thesis can only detects the typical electrical faults (open fault and short circuit fault) of the PV arrays. There is no further research on other fault types such as shadow fault and abnormal aging fault on PV arrays. Therefore, the future research work of this paper will comprehensively consider the online detection of various types of PV arrays faults under different working conditions, thus further improving the effectiveness of this detection method.
5. References

[1] Yang H, Wang H, et al. 2003 Status of photovoltaic industry in China Energy Policy vol 31 pp 703-707

[2] Alam MK, Khan F, Johnson J, Flicker J, et al. 2015 A comprehensive review of catastrophic faults in PV arrays: Types, detection, and mitigation techniques IEEE J Photovoltaics vol 5 pp 982–997.

[3] Pillai DS, Prasanth Ram J, Rajasekar N, et al. 2019 Extended analysis on line-line and line-ground faults in PV arrays and a compatibility study on latest NEC protection standards Energy Conversion and Management vol 196 pp 988-1001

[4] Dhoke A, Sharma R, Saha TK 2018 PV module degradation analysis and impact on settings of overcurrent protection devices Solar Energy vol 160 pp 360-367

[5] Rabah B, Samir M. 2018 Fault detection and diagnosis based on C4.5 decision tree algorithm for grid connected PV system Solar Energy vol 173 pp 610-634

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

[7] Chen Z, Han F, et al. 2018 Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents Energy Convers & Manage vol 178 pp 250-264

[8] Chine W, Mellit A, et al. 2016 A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks Renew Energy vol 90 pp 501-512

[9] Fauzan HJ, Seongmun Oh, Jaesung J Development of photovoltaic abnormal condition detection system using combined regression and support vector machine Energy vol 176 pp 457-467

[10] Chen Z, Wu L, Lin P, et al. 2017 Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and I-V characteristics Apply Energy vol 204 pp 912-931

[11] Chen Z, Lin P, et al. 2017 A density peak-based clustering approach for fault diagnosis of photovoltaic arrays vol 204 pp 912-931

[12] Liu G, Zhu L, Wu X, Wang J 2019 Time series clustering and physical implication for photovoltaic array systems with unknown working conditions Solar Energy vol 180 pp 401-411

[13] Zhu H, Lu L, Yao J, Dai S, Hu Y 2018 Fault diagnosis approach for photovoltaic arrays based on unsupervised sample clustering and probabilistic neural network model Solar Energy vol 176 pp 395-405

[14] Maciej R, Krzysztof G, et al. 2014 Conference on Computer Science and Information Systems, Warsaw, Poland vol 2 pp 879-884

[15] Januzaj E, Kriegel H, Pfeifle M 2004 European Conference on Principles of Data Mining and Knowledge Discovery Springer, Berlin, Heidelberg pp 231-244

[16] Ester M, Kriegel H, Sander J and Xu X 1996 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining Menlo Park CA pp 226 – 231

[17] Ye Z, Lehman B, Ball R, et al 2014 Graph-Based Semi-supervised Learning for Fault Detection and Classification in Solar Photovoltaic Arrays IEEE Transactions on Power Electronics vol 30 pp 2848-2858