Optimal Sensor Placement and Fault Diagnosis Model of PV Array of Photovoltaic Power Stations Based on XGBoost

Hai Wang* and Fengjie Sun
School of Electrical and Electronic Engineering, North China Electric Power University, Beijing, China

*Correspondence author email: 1185482398@ncepu.edu.cn

Abstract. As an important part of photovoltaic power stations, daily monitoring and maintenance of photovoltaic array are quite necessary. In order to be able to accurately locate the faulty module and diagnose fault types. The fault diagnosis model which based on XGBoost optimized by GridSearchCV on optimal sensor placement is proposed in this article. First, the change laws of external electrical characteristics of photovoltaic modules under the control of MPPT technology are analyzed in different fault states. On this basis, the parameters which can locate the faulty PV modules and the input of the XGBoost PV fault diagnosis model are obtained. Finally, during the process of the simulation and experiment, the failure data measured by the multi-sensor method can be used as a positioning quantity to locate the faulty module. The comparison results with the other three algorithms (LR, RF, XGBoost) prove that the performance of GS-XGBoost algorithm has great advantages in judging PV fault types (short circuit, open circuit, aging).

Keywords: The voltage weight matrix; Optimal sensor placement; XGBoost algorithm; GridSearchCV; PV fault diagnosis model.

1. Introduction
As an important part of photovoltaic power stations, in order to prevent unexpected serious faults in the PV system, several conventional fault protection devices are proposed in reference [1]. In the early years, offline detection methods were proposed to monitor and distinguish the faults in the photovoltaic array, mainly including TDR [2] and ECM [3]. But the two methods will affect the normal operation of photovoltaic power stations.

In recent years, various machine learning algorithms have been widely applied in fault monitor and diagnosis of PV array [4]. Existing supervised learning algorithms based on photovoltaic array fault diagnosis include the following: The BP neural network in reference [5] and classification algorithm based on SVM in reference [6]. The above algorithms have the risk of overfitting, and the generalization ability is also greatly limited due to its principle. Then with the development of the Internet of Things, photovoltaic faults can be diagnosed by multi-sensor method. But this method needs a large number of sensors. Therefore, in order to reduce the number of sensors and wiring complexity. The author in reference [7] proposes a photovoltaic array voltage weight matrix theory, which can greatly reduce the number of sensors. But this method will increase the complexity of sensor placement.

Based on the above problems, the multi-sensor method and ensemble learning algorithm (eXtreme Gradient Boosting-XGBoost) in this article are combined to diagnose PV fault types. The improved multi-sensor method in this paper can accurately locate faults without increasing the difficulty of installation. XGBoost algorithm can overcome the problems of low accuracy and poor generalization.
ability. The results of simulation and experiment prove that the photovoltaic fault diagnosis method proposed in this article can accurately locate faulty components and judge fault types.

2. Sensor Placement Method and Fault Location Strategy

2.1. Sensor Placement Method Based on the SP Structure

Photovoltaic module, as the basic unit of photovoltaic array, is connected in series and then connected in parallel to form photovoltaic array. Actual photovoltaic power stations are mostly installed in accordance with the SP structure. Under the premise of satisfying the maintenance of actual photovoltaic power station, this paper conducts research on the basis of SP structure.

Different sensor placement methods all need to meet the following principles: it can be applied in large-scale photovoltaic power stations; the positioning accuracy is 1; The number of sensors is as small as possible. According to analysis and verification, the sensor layout method which based on PV voltage weight matrix mentioned in reference [7] can greatly reduce the number of sensors under the premise of ensuring positioning accuracy. But the wiring difficulty has increased. Then in order to reduce wiring difficulty, traditional voltage sensor is replaced by the multi-channel voltage sensor (MCVS) in this article. The placement method based on multi-channel voltage sensor is shown in figure 1. On this basis, the resolution of locating faulty modules has not decreased, the number of sensors is greatly reduced, there is no possibility of cross wiring, the operability of the actual installation is greatly improved.

![Figure 1. The sensor placement method in this article.](image)

2.2. Location Strategy of the Faculty PV Modules

In the figure 1, the electric potential difference between node 4 and node 1 is $U_1$, the electric potential difference between node 2 and node 3 is $U_2$, $U_1$ and $U_2$ can be measured by the multi-channel voltage sensor. The string current $I_1$ and $I_2$ can be measured by the current sensor. When a faulty module occurs in the detection unit, the abnormal value of $U_1$, $U_2$, $I_1$, $I_2$ can be used to get the location of the faulty module. The location of the faulty module can be obtained by querying the location reference table of PV faulty module obtained through calculation and summary. Table 1 is the location reference table of faulty PV module.

| PV module | $U_1$                | $U_2$                | $I_1$          | $I_2$          |
|-----------|----------------------|----------------------|----------------|----------------|
| PV11      | $1/3U_{array}$ < $U_1$ < $2/3U_{array}$ | $1/6U_{array}$ < $U_2$ < $1/3U_{array}$ | Lower than normal value | Normal value   |
| PV12      | $1/6U_{array}$ < $U_1$ < $1/3U_{array}$ | $1/6U_{array}$ < $U_2$ < $1/3U_{array}$ | Lower than normal value | Normal value   |
| PV13      | $1/6U_{array}$ < $U_1$ < $1/3U_{array}$ | $1/6U_{array}$ < $U_2$ < $1/3U_{array}$ | Lower than normal value | Normal value   |
| PV21      | $1/6U_{array}$ < $U_1$ < $1/3U_{array}$ | $1/6U_{array}$ < $U_2$ < $1/3U_{array}$ | Normal value    | Lower than normal value |
| PV22      | $1/6U_{array}$ < $U_1$ < $1/3U_{array}$ | $1/6U_{array}$ < $U_2$ < $1/3U_{array}$ | Normal value    | Lower than normal value |
| PV23      | $1/3U_{array}$ < $U_1$ < $2/3U_{array}$ | $1/6U_{array}$ < $U_2$ < $1/3U_{array}$ | Normal value    | Lower than normal value |
3. Input Parameters of PV Array Fault Diagnosis Model

3.1. Electrical Characteristics of PV Module in Different Fault State
In order to effectively diagnosis fault types, it is necessary to analyze the output characteristics of the modules under different fault states. This paper mainly studies three PV module fault types: short-circuit, open-circuit and abnormal aging. The U-I curve of the external electrical characteristics of the faculty module is shown in figure 2, there are two important parameters (the maximum power current $I_m$, the maximum power voltage $U_m$) in this figure. It can be found from figure 1 that the fault type of photovoltaic modules can be judged by the change of the maximum power point. When the short-circuit module occurs, $I_m$ remain unchanged and $U_m$ decreases; when the open-circuit module occurs, $I_m$ decreases and $U_m$ remain unchanged; When an aging module occurs, $U_m$ and $I_m$ all both decrease. The fault types can be judged by the different change of parameters $U_m$ and $I_m$ of the photovoltaic modules.

![Electrical output characteristics of the faculty PV module.](image)

3.2. Input Parameters of PV Array Fault Diagnosis Model
Since $U_m$ and $I_m$ are also difficult to measure in daily monitoring, the main content of this section is to obtain parameters that can not only replace $U_m$ and $I_m$, but also can be measured by sensor. In the actual photovoltaic power station, in order to ensure the photovoltaic power station to operate efficiently, the photovoltaic module power is always at the maximum power point under the control of the MPPT technology. Therefore, the electric potential difference $U_1$ and $U_2$ obtained in this article is the voltage difference of the maximum power voltage of the PV module, $I_1$ and $I_2$ are the maximum power current of each string. When the faculty PV modules occur, the change of parameters $U_m$ and $I_m$ can be converted into the change of these four parameters ($U_1$, $U_2$, $I_1$, $I_2$) in the article, which will be used as input parameters of the photovoltaic fault diagnosis model. The change of the maximum power point is not only affected by the internal state of the module, but also related to the external environment (temperature and the light radiation). Since the principle of fault diagnosis in this article is based on the change of the maximum power point. When a faculty PV module occurs, in order to ensure the reliability and accuracy of the fault diagnosis model as much as possible, the light radiation($lr$) and the temperature ($T$) are also applied as input parameters of the fault diagnosis model. In conclusion, the input parameters of the photovoltaic fault diagnosis model in this article are $U_1$, $U_2$, $I_1$, $I_2$, $lr$, $T$.

4. Principle of XGBoost Algorithm and GridSearchCV

4.1. Principle of XGBoost Algorithm
The principle of ensemble learning is to analyze the samples through several basic learners, and then the results obtained from all base learners are collected as the final result. XGBoost is an efficient and essential algorithm of ensemble learning. Its basic function is to apply a few weak classifiers to iterate to predict the new classification membership degree. The classifier which misclassifies the sample will get higher weight in the next step, and finally Several weak classifiers obtained after several trainings
are combined into a strong classifier according to their different weights. The base learner of XGBoost is the classification and regression trees (CART). Its core principle is to continuously split the features, and then update the function by continuously adding trees to fit the residual error of the previous prediction. When K trees are obtained after training, the score of a sample can be predicted. According to the characteristics of the sample, there is a corresponding leaf node in each tree, and each leaf node corresponds to a score. Finally, the scores corresponding to each tree are accumulated to obtain the predicted value of the sample.

The implementation steps of XGBoost algorithm are as follows:

There is a set of data set, the data set contains n samples and m features. Then we need to construct an objective function for the resulting model, namely:

$$L(\phi) = \sum_{i} l(y_{i}, y_{i}^{(t)}) + \sum_{k} \Omega(f_{k})$$  \hspace{1cm} (1)

The above equation includes a loss function and a regularization term, in order to achieve good robustness and facilitate large-scale application, the model needs to have a low complexity. The expression of regularization term is:

$$\Omega(f) = \gamma T + \frac{1}{2} \|w\|^2$$  \hspace{1cm} (2)

$\gamma$ can control the number of leaf nodes, $T$ represents the number of leaf nodes and $w$ represents the score of leaf nodes. The traditional gradient boosting tree (GBDT) relies on the first derivative during the solution process, however, the second-order Taylor expansion of the loss function in XGBoost makes the solution more accurate. After the objective function is obtained, the model can be trained. The objective function can be optimized by the likelihood function, likelihood function is following:

$$L^{(t)} = \sum_{i} \left[ l\left(y_{i}, y_{i}^{(t)} + f_{i}(X_{i})\right) + \Omega(f_{i}) \right]$$  \hspace{1cm} (3)

In order to achieve fast optimization of the objective function and find its approximate optimal value, the second-order Taylor expansion is applied to constrain the likelihood function. The equation is following:

$$L^{(t)} \approx \sum_{i} \left[ l\left(y_{i}, y_{i}^{(t)} + f_{i}(X_{i}) + \frac{1}{2} f_{i}^{2}(X_{i})\right) + \Omega(f_{i}) \right]$$  \hspace{1cm} (4)

Since the calculation process of equation (4) is complicated, the constant term can be deleted to get the following simplified equation:

$$L^{(t)} \approx \sum_{i} \left[ g_{i} f_{i}(X_{i}) + \frac{1}{2} f_{i}^{2}(X_{i})\right] + \Omega(f_{i})$$  \hspace{1cm} (5)

Through extending regularization term, rewrite equation (5), namely:

$$L^{(t)} \approx \sum_{i} \left[ g_{i} f_{i}(X_{i}) + \frac{1}{2} f_{i}^{2}(X_{i})\right] + \gamma T + \frac{1}{2} \|w\|^2 \approx \sum_{j} \left[ \sum_{i \in I(j)} g_{i} w_{i} + \frac{1}{2} \sum_{i \in I(j)} h_{i} w_{i}^{2}\right] + \gamma T$$  \hspace{1cm} (6)

Equation (6) is the optimal objective function, and the optimal weight needs to be calculated:

$$w_{j}^{*} = \frac{\sum_{i \in I(j)} g_{i}}{\sum_{i \in I(j)} h_{i} + \lambda}$$  \hspace{1cm} (7)

The optimal weight is substituted into (6) to solve the optimal solution of the objective function:
The equation (8) is the structural score of the tree, in order to get a good structure, the score needs to be as small as possible. XGBoost applies the greedy algorithm to traverse the segmentation points of all features. And setting a threshold to avoid too deep trees at the same time. Because the traditional gradient boosting tree model lacks the regularization term, overfitting is very easy to occur during the training process. XGBoost overcomes this shortcoming by providing a regularization term to limit overfitting. XGBoost applies parallel processing to achieve higher training speed. Therefore, it has been widely used in financial risk control [8], user behavior prediction and other fields.

4.2. Principle of GridSearchCV Algorithm

There are mainly two methods when selecting parameters during the machine learning process. One method is to fine-tune the parameters based on experience, the other method is to select parameters of different size and bring them into the model to select the parameters of best performance. One method of fine-tuning is to manually debug the parameters until the good parameter combination is found. This process will take a long time and we may not have enough time to explore multiple parameter combinations. Then the GridSearchCV algorithm is selected to do this search jobs in this article.

The name of GridSearchCV can actually be divided into two parts, GridSearch and CV, that is, grid search and cross-validation. The principle of GridSearch is that within the specified parameter range, the parameters are adjusted according to a fixed step length and the learner is trained with the adjusted parameters. During the operation of the GridSearch algorithm, each parameter combination corresponds to a model, then the parameter combination with the best performance on the validation set will be selected to establish the optimal classification model. It is a process of training and comparison. The basic principle of CV is to split the available data into two parts according to a specific ratio (generally 8:2). The more part is training set and the less part is validation set. The classification accuracy of the classifier on the validation set will be taken as the important indicator for evaluating the classifier performance.

5. The optimal Fault Diagnosis Model Based on XGBoost Algorithm

5.1. The Failure Data of Photovoltaic Modules

The sensor placement method proposed in this article needs simulation experiment to prove the reliability of the method. The simulation software is Simulink2017 version or above, integrated photovoltaic module modules can be called directly. During the simulation process, the short-circuit condition of the single module was simulated by short-circuiting the photovoltaic modules in the detection unit; The open circuit of the single module was simulated by disconnecting; the aging condition of single module was simulated by connecting the resistance; The unstable external environment was simulated by changing the value of the light irradiation and temperature. A total of 921 groups, photovoltaic module failure data, were obtained in this simulation. The failure data of short-circuit, the failure data of open-circuit, the failure data of aging are respectively 300 groups, the rest is the data of the photovoltaic modules in normal operation. The photovoltaic fault diagnosis model can be trained through these failure data.

5.2. Optimal PV Array Fault Diagnosis Model

The implementation of XGBoost algorithm needs pycharm software and related software packages such as SKlearn, xgboost and other basic packages. The research calculation and simulation of all fault diagnosis models in this section are implemented in pycharm software. The data obtained from the simulation experiment is divided into a training set and a validation set according to the certain proportion (8:2). The performance of different fault diagnosis models is shown by the accuracy rate for different data sets.

In this article, GridsearchCV is applied to optimize the parameters of XGBoost algorithm, the parameters that need to be optimized in XGBoost mainly include the following parameters: ’n_estimators’
and 'gamma'. The respective functions of ‘n_estimator’ and 'gamma' are shown in the following table 2. This new algorithm model obtained through optimization is called GS-XGBoost in this article. GS-XGBoost algorithm will be compared with the random forest algorithm, Logistic regression and the unoptimized XGBoost algorithm. In order to facilitate observation and debugging, the curve graphs are utilized to indicate the trend of the performance of different fault diagnosis models. The optimal value of the parameter is obtained on the basis of the performance improvement. The following are the optimal results of the parameters to be optimized:

Table 2. The respective functions of parameters

| Parameters    | Functions of parameters                                      |
|---------------|--------------------------------------------------------------|
| n_estimator   | The total number of iterations, which is the number of decision trees |
| gamma         | Penalty coefficient, which specifies the minimum loss function drop value required for node splitting |

Figure 3 and figure 4 show that the optimal parameter value of n_estimator is 681 and the optimal parameter value of gamma is 0.02. These parameters are substituted into the XGBoost algorithm to obtain the GS-XGBoost optimal fault diagnosis model. The comparison of the accuracy rate of different algorithms on the validation set and the comparison of the accuracy rate of the same algorithm on the training set and the validation set can be used as index to evaluate the performance of different fault diagnosis model. The comparison results are shown in table 3.

Table 3. The accuracy rate of different algorithms for simulation data.

| Algorithm     | Accuracy rate of training set | Accuracy rate of validation set |
|---------------|-------------------------------|-------------------------------|
| LR            | 91.7%                         | 88.6%                         |
| RF            | 100.0%                        | 99.5%                         |
| XGBoost       | 90.8%                         | 91.9%                         |
| GS-XGBoost    | 98.1%                         | 99.5%                         |

Whether the fault diagnosis model has the risk of overfitting can be obtained by comparing the accuracy rate of the training set and the accuracy rate of the validation set. According to the results of table 3, the accuracy rate of the Logistic Regression algorithm is lower than other algorithms. Though the accuracy of the validation set of the random forest algorithm is as high as 99.5%, the accuracy of the training set is 100%, so the algorithm may have the risk of overfitting. The accuracy of the training set of XGBoost algorithm is 90.8% and the accuracy of the validation set is 91.9%. After the XGBoost algorithm is optimized, the accuracy of the training set of the GS-XGBoost algorithm is 98.1% and the accuracy of the validation set is 99.5%. Compared with the XGBoost algorithm, the new GS-XGBoost algorithm has lower risk of overfitting and a 2% increase in accuracy rate. It can be concluded that the performance of GS-XGBoost algorithm on photovoltaic fault diagnosis is better than other algorithms.
5.3. Experimental Simulation of PV Array Fault Diagnosis Model

In the previous section, it is concluded that the GS-XGBoost algorithm has higher accuracy rate and lower risk of overfitting through comparison with other algorithms. In order to meet the requirements of actual photovoltaic power stations, the generalization ability of the GS-XGBoost algorithm must be tested. In this section, the 2×3 photovoltaic array is established in the photovoltaic power station of North China Electric Power University. After the MCVS (the silver box is multi-channel voltage sensor) is installed according to the sensor placement method in this article, the experimental platform is shown in figure 5.

In order to meet the actual situation, the experiment was carried out under different weather conditions. A total of 627 sets of data were obtained, the failure data of short-circuit, the failure data of open-circuit, the failure data of aging are respectively 198 groups, the rest is the data of the photovoltaic modules in normal situation.

Table 4. The part data of testing set.

| PV module | Module situation | Light irradiation(W/M²) | Temperature(℃) | U₁(V) | U₂(V) | I₁(V) | I₂(V) |
|-----------|------------------|-------------------------|----------------|-------|-------|-------|-------|
| PV₁₁      | Short circuit    | 956                     | 28             | 35.75 | 30.39 | 7.51  | 8.41  |
| PV₁₁      | normal           | 867                     | 28             | 32.08 | 32.08 | 7.68  | 7.68  |
| PV₁₂      | Open circuit     | 802                     | 27             | 31.48 | 31.48 | 0.00  | 7.10  |
| PV₁₂      | aging            | 925                     | 24             | 30.77 | 30.77 | 7.76  | 8.14  |
| PV₁₃      | Short circuit    | 841                     | 26             | 30.13 | 35.46 | 6.56  | 7.43  |
| PV₁₃      | aging            | 791                     | 23             | 30.61 | 33.40 | 6.65  | 7.00  |
| PV₂₁      | Open circuit     | 841                     | 23             | 30.90 | 32.64 | 7.44  | 0.00  |
| PV₂₁      | Short circuit    | 791                     | 23             | 29.79 | 35.10 | 7.00  | 6.16  |
| PV₂₂      | aging            | 841                     | 29             | 31.37 | 31.37 | 7.42  | 6.75  |
| PV₂₂      | Open circuit     | 903                     | 30             | 31.68 | 31.68 | 8.01  | 0.00  |
| PV₂₃      | Short circuit    | 956                     | 27             | 35.63 | 30.25 | 8.42  | 7.58  |
| PV₂₃      | aging            | 925                     | 28             | 34.11 | 31.32 | 8.12  | 7.79  |

Through observation, it can be concluded that the failure data in the table 4 meets the fault location strategy in this article. So when faulty module occur in actual PV station, the four parameters U₁, U₂, I₁, I₂ can be used to locate the faulty module. Then uploading these data into the pycharm software and getting the accuracy rate under different diagnosis models (LR, RF, XGBoost, GS-XGBoost), as shown in table 5.

Table 5. The accuracy rate of different algorithm models for actual data.

| Algorithm | Accuracy rate | Algorithm | Accuracy rate |
|-----------|---------------|-----------|---------------|
| LR        | 69.8%         | XGBoost   | 84.1%         |
| RF        | 91.3%         | GS-XGBoost| 95.2%         |
It can be seen from table 4 that the accuracy rate of the LR algorithm is 69.8%. Due to the insufficient generalization performance of the algorithm, the accuracy rate of RF for actual PV failure data is only 91.3%. The XGBoost algorithm also does not maintain a high accuracy rate, with an accuracy rate of only 84.1%. Compared with RF and the XGBoost algorithm, the generalization ability of the GS-XGBoost algorithm has been greatly improved, the accuracy rate is the highest among the four algorithms, reaching 95.2%. The comparison of the accuracy rate of the four algorithms shows that the generalization ability of the GS-XGBoost algorithm is greatly improved. The other three algorithms are not suitable for fault detection in actual photovoltaic power stations. In conclusion, the performance of GS-XGBoost algorithm optimized by GridSearchCV has great advantages in photovoltaic array fault diagnosis, which is of great significance for the application of Ensemble Learning algorithms in photovoltaic power station maintenance.

6. Conclusion
Studying fault diagnosis in photovoltaic array is an important prerequisite for the normal operation of photovoltaic power stations. In this article, the electrical characteristics of photovoltaic modules are analyzed and studied, the electrical characteristics of the faulty modules are obtained. The parameters derived from the new sensor placement method can locate faulty modules. The six parameters (U1, U2, I1, I2, Ir, T) in this article can be used as the input of the PV array fault diagnosis model to diagnose fault types without interrupting the normal operation of the power station. The simulations and experiments in this article prove that the new GS-XGBoost algorithm has better performance (higher accuracy rate, lower risk of over-fitting, stronger generalization performance) than other algorithms (LR, RF, XGBoost). The above work is of great significance for Ensemble learning algorithms to be applied to the maintenance of photovoltaic power station.

Acknowledgments
This work was supported by the Information and Communication Company of State Grid Qinghai Electric Power Company and State Grid Corporation, entitled 'Research on Key Technologies of Carrier Wireless Convergence Communication for Distributed Monitoring of Photovoltaic Power Station’, under Grant 522814150017. This work was supported by State Grid Corporation, under Grant 52010118000Q.

References
[1] Pillai DS, Rajasekar N. A comprehensive review on protection challenges and fault diagnosis in PV systems. Renew Sustain Energy Rev 2018; 91: 18–40.
[2] Roy S, et al. An irradiance independent, robust ground fault detection scheme for PV arrays based on spread spectrum time domain reflectometry (SSTDR). IEEE Trans Power Electron 2017; PP (99):1–1.
[3] Takashima T, et al. Experimental studies of fault location in PV module strings. Sol Energy Mater Sol Cells 2009;93(6–7):1079–82.
[4] Kumar BP, et al. Online fault detection and diagnosis in photovoltaic systems using wavelet packets. IEEE J Photovoltaics 2018;8(1):257–65.
[5] Youssef A, El-Telbany M, Zekry A. The role of artificial intelligence in photo-voltaic systems design and control: a review. Renew Sustain Energy Rev 2017; (78):72–9.
[6] Sun, J.; Sun, F.; Fan, J.; Liang, Y. Fault Diagnosis Model of Photovoltaic Array Based on Least Squares Support Vector Machine in Bayesian Framework. Appl. Sci. 2017, 7, 1199.
[7] Hu Y, H; Chen H; Xu R, D; Li R. Photovoltaic (PV) array fault diagnosis strategy based on optimal sensor placement [J]. Proceedings of CSEE, 2011,31(33):19-30.
[8] Chen T; Guestrin C. XGBoost: A scalable tree Boosting system[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining.2016.