Multiobjective Evolutionary Clustering to Enhance Fault Detection in a PV System

Luciana Yamada, Priscila Rampazzo, Felipe Yamada, Luís Guimarães, Armando Leitão, and Flávia Barbosa

Abstract Data clustering combined with multiobjective optimization has become attractive when the structure and the number of clusters in a dataset are unknown. Data clustering is the main task of exploratory data mining and a standard statistical data analysis technique used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. This project analyzes data to extract possible failure patterns in Solar Photovoltaic (PV) Panels. When managing PV Panels, preventive maintenance procedures focus on identifying and monitoring potential equipment problems. Failure patterns such as soiling, shadowing, and equipment damage can disturb the PV system from operating efficiently. We propose a multiobjective evolutionary algorithm that uses different distance functions to explore the conflicts between different perspectives of the problem. By the end, we obtain a non-dominated set, where each solution carries out information about a possible clustering structure. After that, we pursue a-posteriori analysis to exploit the knowledge of non-dominated solutions and enhance the fault detection process of PV panels.

Keywords Multiobjective · Clustering · Photovoltaic systems · Fault detection

1 Introduction

In the last years, the Paris Agreement has defined the necessary targets to limit global warming using renewable energy. Despite decreasing solar panels' costs, the panels' efficiency is still reduced compared with other energy sources [1]. The industry has been working to improve the overall performance of photovoltaic (PV) systems. However, issues still need to be addressed concerning reliability, unforeseen outages,

L. Yamada (✉) · F. Yamada · L. Guimarães · A. Leitão · F. Barbosa
INESC TEC, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal
e-mail: luciana.o.yamada@inesctec.pt

P. Rampazzo
Faculdade de Ciências Aplicadas, Universidade Estadual de Campinas, Limeira, Brazil

J. P. Almeida et al. (eds.), *Operational Research*, Springer Proceedings in Mathematics & Statistics 437, https://doi.org/10.1007/978-3-031-46439-3_16

and high operation and maintenance (O&M) costs, hindering a lean integration in the electrical grid.

A PV system comprises one or more solar panels, connected in series or parallel, combined with an inverter, and a utility grid, among other components. PV panels convert sunlight into electrical energy, and faults may affect performance. A PV fault decreases the system's performance (reduces the output) or interrupts the system's operation. Faults can occur by different factors such as errors in the Maximum Power Point Tracking, shadowing, degradation, and electrical disconnection. Such failures can reduce the instantaneous power generated by the PV plant or permanently degrade the overall asset. In this sense, implementing advanced analytical tools to diagnose failures is crucial in guaranteeing the asset's reliability and performance. In addition, detecting and diagnosing potential failures is also decisive to reduce the costs associated with O&M and the system unavailability.

Although different techniques have been reported in the literature, the fault detection process must deal with some problems [2]. Regarding data availability, accurately representing the electrical characteristics under different fault conditions is challenging, and most fault diagnosis models require the precise division of the fault samples. Fault detection methods based on classification models can effectively reveal and categorize faults. However, specialized human knowledge or complex and costly equipment are needed to establish the diagnostic model and fault samples. Effective clustering of fault samples is a prerequisite for establishing an efficient detection model, but the scarcity of fault samples in real operational data makes this task difficult [3].

Considering as challenges of the fault detection problem in PV systems: (i) the difficulty of categorizing faults and having data labeled according to their different types, and (ii) the lack of information regarding the distribution of data in high-dimension, this paper proposed a Multiobjective Evolutionary Clustering Algorithm (MECA) to detect faults in PV systems through the following approach:

i Application of a multiobjective algorithm for the clustering task–to capture the different clustering perspectives through different objective functions to overcome the lack of knowledge about the data distribution format due to the diversity of failures and high-dimensional spaces.
ii Proposal of a-posteriori analysis–associating data with known labels to the results of the multiobjective algorithm for the clustering task; the aim is to obtain a cluster-sharing matrix to enhance the task of attributing labels to initially unlabeled data.
iii Validation of the method–use the multiobjective approach for clustering and enhance the assignment of missing labels in databases with few known labels–through the measurement of accuracy and confusion matrix.

The main contribution of this paper is the proposal of a-posteriori analysis that demonstrated to be able to label data and explore complementary information for data clustering.

This paper is organized as follows. Section 2 gives a description of fault detection techniques. Section 3 presents the Multiobjective Evolutionary Clustering algorithm. Section 4 reports the experimental results. Section 5 summarizes the conclusion of this work and future research.

2 Fault Detection

According to [4], fault detection techniques are divided into visual and thermal, and electrical methods. Visual and thermal imaging methods help in situations where faults are challenging to detect in the visual inspection process. These methods analyze images collected by cameras that can detect temperature differences in the PV module and recognize the failures' exact location. The electrical methods use the measurements of the electrical output such as the current and voltage of the PV system, to diagnose the faults. Among the electrical methods, some of the techniques found are statistical and signal processing approaches, current-voltage characteristics analysis, power losses analysis, and artificial intelligence (AI) techniques [5]. In this work, we will focus on AI techniques.

The AI techniques has shown significant advances and contributions to several scientific areas. One of the main uses of fault detection is related to supervised methods. In [6], a Monitoring System (MS) is presented to measure the electrical and environmental variables to produce instantaneous and historical data. Integrated with the MS, an Auto-Regressive with an Exogenous input model is used to detect faults in the system, such as short-circuit, open-circuit, partial shadowing, and degradation. Regarding the classification of the faults, different supervised models were compared.

Unsupervised methods, such as data clustering, are applied to group data with similar characteristics, differentiating data with failures and data in regular operation. In [7], the K-means algorithm is used to cluster thermal images to detect and localize damage. Elbow method and mean silhouette method are used to define the best number of clusters. K-means proved to detect faults in images and can be integrated into the thermal drone system. In [8], an unsupervised method, density-based spatial clustering of applications with noise algorithm, is used for clustering faults in a PV system. The normalization proposed in [9] is used to avoid overlap of the clusters. The approach identifies the faults of open-circuit and short-circuits.

In addition to the mentioned methods, semi-supervised approaches are applied in the literature. The main objective of these algorithms is to combine many labeled samples with a few unlabeled ones to develop more efficient models [10]. These algorithms try to improve the performance of supervised or unsupervised learning using a combination of methods and information generated by each other [11]. In [9], a Graph-Based Semi-supervised Learning (GBSSL) method is proposed for fault detection and classification in PV arrays. The normalization based on the panels' standard test condition was proposed and proved to help avoiding overlap clusters under normal and fault conditions. The GBSSL demonstrated that the faults could

be correctly detected during weather changes or PV arrays degradation. In this work, the authors consider the short and open-circuit faults.

The Fuzzy C-Means (FCM) [12] algorithm is a clustering algorithm based on the fuzzy division of an objective function that groups samples of failures by the degree of membership through the Euclidean distance between the sampling point and the center of the corresponding cluster, which makes the algorithm limited to processing datasets with spherical shapes. The authors of [3] propose the function of the Gaussian kernel (GK) in the FCM algorithm to map data in the high-dimensional space and transform the non-linear information of the original space into a linear problem, indicating a significant improvement in the applicability and the algorithm clustering accuracy. The GK-FCM algorithm performs unsupervised clustering and labeling fault samples under typical fault conditions. The labeled fault samples serve as input to a probabilistic neural network based model to facilitate intelligent diagnosis of PV array faults.

Although works with similar intentions have been proposed in the literature, the approach presented here is new: treating the fault detection PV system by analyzing the results of a multiobjective clustering algorithm. Unlike the existing works, we extract additional information from clustering in a-posterior analysis, considering a few labeled samples, and use this information to classify the failure types for unlabeled data.

3 Methodology

In this section, an overview of MECA is presented. Based on the algorithm presented in [13], we propose an algorithm to cluster PV system data and find possible labels to unlabeled data in a-posteriori analysis. Our algorithm has some changes from the original algorithm regarding the initialization and the mutation operators. The general framework of MECA is outlined in Algorithm 1.

The individual of MECA represents a way of grouping the status of a PV system for a specific irradiance range. Each status can be associated with one type of fault or a normal condition. Status with similar characteristics should be clustered together. The individual was coded using a locus-based adjacency coding. The methods used in initialization introduce individuals with characteristics explored in the two objective functions: Compactness and Connectivity. K-means aims to minimize the sum of squared errors of each sample to the nearest centroid. The centroids represent the average of the samples in each cluster. Therefore, this algorithm highlights structures of compact clusters. Similarly, K-medoids prioritize compact clusters with less sensitivity to outliers. Kruskal Algorithm introduces individuals with smaller distances between the samples, highlighting the connectivity of the data.

Next, we use the NSGA-II [14] to sort and select individuals from the population in the binary tournament to compose the parents population P_g. The selected set of parents go through crossover and mutation to create an offspring population Q_g. The uniform crossover was used to combine genes from two parents. The mutation oper-

ator proposed in this work emphasizes the characteristics of the objective functions. Relying on the data distribution, one of the objective functions or a mix of the two can better represent the data. For example, K-means and K-medoids can not create a proper individual due to the spherical characteristic in data with the elongated or connected format. In contrast, Kruskal Algorithm prioritizes connectivity, and it is hard to deal with data with different types of distribution or even in compact or spherical shapes. Therefore, the neighborhood mutation was developed to change the assignment of a sample to the cluster of its nearest neighbor, exploring the connection between samples. Similarly, the centroid mutation explores the characteristic of compact clusters, changing a sample to another closest centroid. These designed mutation operators help maintain the genetic diversity of the population. In our fault detection problem, as we can not know the data distribution and consider different irradiance ranges, the proposed objective function and mutation operators is a proper choice to capture distinct patterns.

The combined population $R_g = P_{g-1} \cup Q_g$, is sorted according to non-dominated classification and crowding distance to choose exactly TP population members to the next population P_g [14].

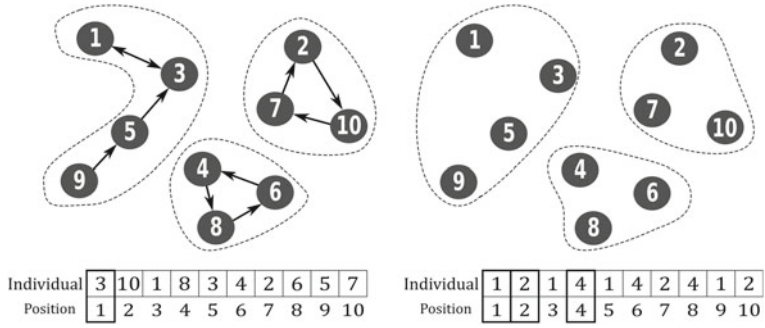
After obtaining the non-dominated frontier, a-posteriori analysis is applied to exploit the knowledge of the solutions and enhance the fault detection process of PV panels. For this purpose, we combine the information obtained from the non-dominated solutions with a small percentage of labeled samples.

More details of the implementations of the coding, evaluation, operators and a-posteriori analysis are described in the following.

Algorithm 1: MECA

```
Initialize population P_0 with size TP = Q_Kmeans + Q_Kmedoids + Q_MST
for g = 1,..., G do
    Select the sets of parents, p_1 and p_2, through binary tournament
    Create F_c offsprings from p_1 and p_2 (uniform crossover)
    Create F_m offsprings from p_1 and p_2 (centroid and neighborhood mutation)
    Q_g = F_c ∪ F_m Create R_g = P_{g-1} ∪ Q_g
    Evaluate R_g in Compactness and Connectivity
    Select TP individuals of R_g through rank and crowding distance to form the next
    population P_g
end
Return the non-dominated solutions
```

All the parameters were defined experimentally. The population size (TP) was set to 100, and the number of generations (G) to 100. The initial population is generated using the following K-means [15], K-medoids [16], and the Kruskal Algorithm [17]. We define the number of solutions of each method, Kruskal (Q_{MST}), K-means (Q_{Kmeans}), and K-medoids ($Q_{Kmedoids}$), respectively to 24, 38 and 38. For each method, an individual will be created considering the chosen range of the cluster where $[K_{min}, K_{max}] = [2, 10]$.



(a) The locus-based adjacency coding.

(b) The decoding process.

Fig. 1 Illustration of the locus-based adjacency coding and the decoding process

Coding

In our dataset, each observation corresponds to a status of a PV system, measured by attributes related to weather conditions and electrical variables. The status indicates whether the system operates in a normal condition or has any faults. Thus, each individual in the population corresponds to a way of grouping the N observations of the dataset. To represent the individual, we used locus-based adjacency coding [13], where an individual is a vector $[a_1, a_2, ..., a_N]$, where a_i is associated with sample i-th and represents one of the nodes of the graph. The value of an element $a_i = n$ indicates that sample i is linked to sample n. These links create the node's connection and, in the end, generate the final clusters in the decoding process. With this representation, the number of clusters in individuals can be flexible during the search process. In Fig. 1a, we represent an individual with locus-based adjacency coding. For example, position one is associated with node and sample 1. The value 3 indicates that sample 1 has a link with sample 3, as we can see in the graph. In Fig. 1b, the individual is decoded using one of the samples as the root (smallest value) to group samples assigned to the same cluster. In the decoding process, the roots are 1, 2 and 4.

Solution Evaluation

To evaluate each individual ($k = 1, ..., TP$) of the population and express different characteristics of the data, we use two objective functions based on the MOCK [13]: Compactness and Connectivity. As an objective reflecting the cluster compactness (1), we minimize the Euclidean distance $\delta(i, \mu_j)$ between each sample i of the cluster C_j with its centroid μ_j:

$$\text{Compactness} = \sum_{C_j \in C} \sum_{i \in C_j} \delta(i, \mu_j). \tag{1}$$

To represent the cluster connectivity (2), we apply a penalty factor that considers the distance from the nearest neighbors. The nearest neighbors are calculated using the euclidean distance between data points. In the objective function, $nn_{i,j}$ is the j-th nearest neighbor of a sample i, L is the number of neighbors that are considered to compute the metric, and N is the size of the samples in the dataset. For the parameter L, we set it to 10, as suggested in [13]. This objective function should be minimized.

$$\text{Connectivity} = \sum_{i=1}^{N} \sum_{j=1}^{L} p_{i,nn_{i,j}}. \tag{2}$$

The penalty factor use the following rules:

$$p_{i,nn_{i,j}} = \begin{cases} \frac{1}{j}, & \text{if } nn_{i,j} \text{ does not belong to the same cluster of the sample } i \\ 0, & \text{otherwise.} \end{cases}$$

Non-dominated Sorting Genetic Algorithm II (NSGA-II)

As defined in [14], in NSGA-II each individual ($k = 1, ..., TP$) is associated with two attributes: $rank_k$ and $distance_k$ (non-dominated classification and crowding distance). If two solutions have different non-domination levels (different non-dominated frontiers), we choose the solution k with the lower $rank_k$. Otherwise, if two k_1 and k_2 solutions belong to the same frontier ($rank_{k1} = rank_{k2}$), then we prefer the solution that is located in a less crowded region (that is, higher $distance_k$) [18]. The crowding distance represents the sum of the normalized distances of the nearest neighbors of the solution along each objective. Large distance values are assigned for the extreme solutions, and the crowding distance is calculated for the rest. Formally, given a solution m, the distance d_m is defined as:

$$d_m = \sum_{f=1}^{F} \frac{g_f(m+1) - g_f(m-1)}{g_f^{max} - g_f^{min}} \tag{3}$$

Where F is the number of objective functions, $m + 1$ and $m - 1$ are the nearest neighbors of a solution m, g_f^{max} and g_f^{min} the maximum and minimum values of each function g_f, considering individuals from the same frontier (same *rank*).

Non-dominated classification and crowding distance are used in the binary tournament selection algorithm to create the parents' population and to select the population for the next generations.

Crossover and Mutation

A population of parents is built through a binary tournament. Pairs of parents from this population are selected to generate pairs of offspring. To compose the offspring population, we define the solutions generated by the crossover (F_c) to 20 and by the mutation (F_m) to 20. The uniform crossover is applied to each pair of selected solutions, thus generating two offspring. The binary crossover mask with a uniform distribution is created; each individual has a 50% chance to copy the gene from one and 50% from the other.

The mutation process complements the crossover, allowing a more extensive search space to be explored. In the literature, mutation, and crossover operators for the clustering problem are presented in [19]. The mutation is based on the nearest neighbors in the algorithms that use locus-based adjacency graph encoding. In the present work, two types of mutation are proposed to emphasize the characteristics of objective functions: centroid mutation and neighborhood mutation.

The centroid mutation aims to assign the selected sample to the cluster with the nearest centroid. For each offspring:

1. Select randomly one sample i of the offspring.
2. Calculate the Euclidean distance between sample i and all the other centroids of the offspring.
3. Assign the sample i to a new cluster, considering the nearest centroid.

The neighborhood mutation aims to assign the selected sample to the cluster of the nearest neighbor. For each offspring:

1. Select randomly one sample i of the offspring.
2. Identify the L nearest neighbor of the sample i through the k-nearest neighbors algorithm.
3. Assign the sample i to a new cluster, considering the nearest neighbor.

A-posteriori analysis

Following the MECA's execution, a-posteriori analysis examines the non-dominated solutions to gather insights into the clustering process. This analysis aims to determine whether we can leverage the acquired information to identify potential labels for unlabeled data. In the first step of a-posteriori analysis, we divide the non-dominated frontier into three portions, as illustrated in Fig. 2.

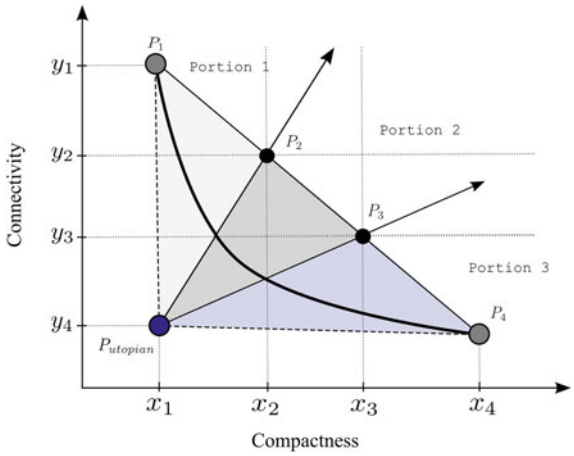


Fig. 2 Division of the non-dominated frontier into three portions. The gray solutions represent the extremes of the frontier, and the blue solution represents the Utopian solution

From the extreme solutions of the frontier (P_1 and P_4), we find the corresponding Utopian solution (P_{utopian}) and determine the equation of a line through the two extreme points. Then, we divide the frontier into three portions, finding lines that intersect the Utopian solution and the points P_2 and P_3. This division aims to group similar data distribution in the non-dominated frontier. Portion 1 corresponds to solutions with better values in Compactness, while Portion 3 corresponds to solutions with better values in Connectivity. On the other hand, Portion 2 has solutions with a mix of characteristics of the two objective functions. MECA creates individuals with different numbers of clusters. As presented in [13], solutions with a minimum value in Compactness result in solutions with a high number of clusters, while a minimum value in Connectivity corresponds to solutions with a low number of clusters. The idea of dividing the frontier is to maintain the distributions of each portion as much as possible, not mixing the structure of a solution with a high number of clusters with another with a low number of clusters. Furthermore, as we do not know the data distribution of the fault detection problem, we want to know what portion better represents our data.

Besides the information provided by the non-dominated solutions, we used a small percentage of samples with known conditions (labeled data) to guide data labeling and extract complementary information regarding the clustering. These samples could be seen as faults and normal data identified by the specialists and can be used as a reference to help the decision-maker in new unknown cases.

For each portion of the non-dominated frontier, we construct a cluster-sharing matrix with dimension $n \times q$, where n is the total number of samples in the dataset and q are the labeled samples. The matrix elements $c_{i,j}$ represent the number of times a determined sample i was allocated in the same cluster as the sample j. If two samples are frequently grouped, they can be considered similar and in the same cluster. From this cluster-sharing matrix (4), we can extract probability information as the chance of a sample being assigned in the same cluster as a sample with a

known condition. To find the probability that a sample n belongs to the same group as a sample q, we divide each element of the cluster-sharing matrix by the total sum of each row.

$$MP_{n,q} = \begin{pmatrix} \frac{c_{1,1}}{\sum c_{1,q}} & \frac{c_{1,2}}{\sum c_{1,q}} & \cdots & \frac{c_{1,1}}{\sum c_{1,q}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{c_{n,1}}{\sum c_{n,q}} & \frac{c_{n,2}}{\sum c_{n,q}} & \cdots & \frac{c_{n,q}}{\sum c_{n,q}} \end{pmatrix}. \tag{4}$$

Accuracy (ACC) is used to measure the correct assignment of the labels for the unlabeled data. In our experiments, part of the samples are considered unlabeled. Thus, we use the actual label as a reference value to evaluate the performance of a-posteriori analysis. To calculate the ACC, we consider the number of observations a-posteriori analysis correctly assigned the labels (U_c) to the total number of unlabeled samples (T_u).

$$ACC = \frac{U_c}{T_u}. \tag{5}$$

4 Experimental Results

In this section, we will present the photovoltaic fault dataset and the computation results obtained from the experiments.

4.1 Photovoltaic Fault Dataset

We used a real-world dataset in these experiments, publicly available[1] [6]. The dataset was generated from a PV plant simulator that describes the system behavior. The data was collected and labeled, including faults (degradation, short-circuit, open-circuit, and shadowing) and normal condition. In this work, we consider the Normal data (No), Short-circuit (Sc), and Open-circuit faults (Oc). The Normal condition is associated with data without any fault. Open-circuit fault generates an interruption in the circulation of electric current due to a disconnection in the system [6]. Short-circuit fault occurs when a low impedance path appears along the system [6]. In this work, shadowing is not considered because it can cause bad data and lead to incorrect fault detection [9].

The dataset contains six attributes: irradiance, PV module temperature, and voltage and current output for both PV strings. To analyze the applicability of a multiobjective evolutionary clustering algorithm to fault detection, we divide the irradiance into six intervals between 801.116–936.475 W/m^2. It is essential to mention that in

[1] https://github.com/clayton-h-costa/pv_fault_dataset.

Table 1 The number of data points for each irradiance range. "T" = total number of labeled or unlabeled data, "No" = Normal data, "Sc" = Short-circuit faults, and "Oc" = Open-circuit faults

Range	Irradiance (W/m^2)	Number of data points							
		Labeled				Unlabeled			
		T	No	Sc	Oc	T	No	Sc	Oc
1	801.116–858.668	301	146	118	37	1201	583	473	145
2	858.668–893.252	301	202	15	84	1201	805	61	335
3	893.252–911.252	300	130	169	1	1201	519	679	3
4	911.252–922.624	301	101	196	4	1201	403	782	16
5	922.624–929.616	300	58	35	207	1202	232	142	828
6	929.616–936.475	301	57	22	222	1200	228	85	887

this dataset, the short-circuit and open-circuit faults occurred at a high irradiance level. In real situations, there is much more unlabeled data than labeled because the labeling process requires specific technical knowledge, is time-consuming, and can be costly [2]. A database was adapted to apply the a-posteriori analysis associated with data with known labels and evaluate the possibility of assigning labels to unlabeled data through this proposal. We randomly selected 1500 samples for each irradiance range and considered 80% of the samples unlabeled to test our data labeling approach. The labeled data is considered as the samples with known conditions, whereas the condition of the unlabeled data will be determined through a-posteriori analysis. Table 1 presents the number of data points for each irradiance range and each condition.

4.2 Computation Results

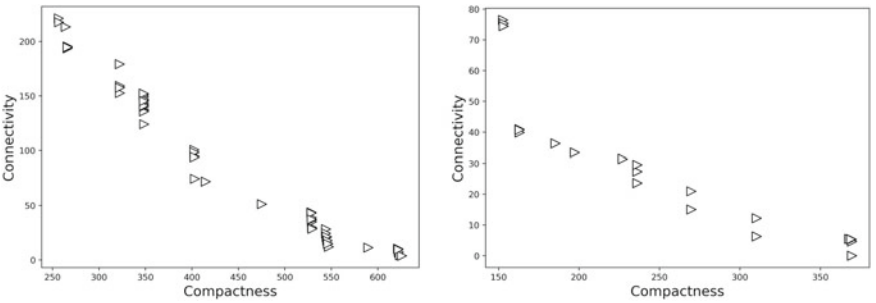
Due to the genetic algorithm's stochastic nature, each irradiance range's dataset is run ten times. Then, a-posteriori analysis was applied considering the non-dominated frontier information and the samples with the known conditions (labeled samples), which are records of the PV system with some fault or normal data. In a-posteriori analysis, we identify the unknown operation condition based on the probability information extracted from the cluster-sharing matrix.

In Table 2, we reported the average, the standard deviation, and the minimum and maximum obtained accuracy values of the data labeling from 10 runs based on a-posteriori analysis.

Since we had small deviations for all intervals, we selected two intervals to analyze the non-dominated frontier, extraction of complementary information, data labeling, and confusion matrix. To evaluate a-posteriori analysis, we calculate the accuracy with Eq. (5), presented in Sect. 3. We chose round 7 from range 3, with the lowest accuracy of 0.8102. To represent the best accuracy, we chose round 7 from range 5

Table 2 The average, standard deviation, minimum and maximum accuracy values from 10 runs

Range	Accuracy (ACC)			
	Average	Standard deviation	Minimum	Maximum
1	0.9902	0.0017	0.9883	0.9925
2	1	0	1	1
3	0.9082	0.0429	0.8102	0.9484
4	0.9960	0.0012	0.9942	0.9983
5	0.9760	0.0163	0.9575	1
6	1	0	1	1



(a) 893.252-911.252 W/m^2 (range 3). (b) 922.624-929.616 W/m^2 (range 5).

Fig. 3 Non-dominated frontier of round 7 for the irradiance range 3 and 5

with a precision of 1. Although ranges 2 and 6 achieved the highest accuracy across all 10 runs, we did not choose them to demonstrate the following results. These two ranges obtained the same maximum probability value ($p = 1$) for all conditions in the three portions, so we do not observe different probability information.

In Fig. 3, we present the non-dominated frontiers of round 7 for ranges 3 and 5. We can see that the objective functions are conflicting, and the irradiance ranges can reach different values in Compactness and Connectivity. We analyzed the correlations between Compactness and Connectivity and observed a negative correlation for almost all ranges. Only for interval 4 we observed a low correlation. Then, the objective functions are not conflicting, and using a mono-objective algorithm for this range would be enough.

Table 3 presents the additional information obtained for interval 3. We selected only some samples to demonstrate the information obtained in a-posteriori analysis. For each portion of the non-dominated frontier, we obtained different probability information based on the distribution of solutions. The columns No, Sc, and Oc of each portion correspond to the probability information of an unlabeled sample being assigned to the same grouping of the labeled samples of one of these conditions. The actual condition is the correct label for the samples. To analyze the performance of

Table 3 Complementary information extracted from the three portions of non-dominated frontiers about the irradiance range 3 (893.252–911.252 W/m^2) for the samples A to F

Sample	Irradiance 893.252–911.252									Actual Condition
	Portion 1			Portion 2			Portion 3			
	No	Sc	Oc	No	Sc	Oc	No	Sc	Oc	
A	**1**	0	0	0.923	0.077	0	0.598	0.402	0	No
B	0.533	0.460	0.007	0.355	**0.635**	0.009	0.4933	0.505	0.002	No
C	0.062	**0.938**	0	0.320	0.676	0	0.290	0.708	0.002	Sc
D	**0.740**	0.260	0	0.364	0.635	0.001	0.290	0.708	0.002	Sc
E	0.223	**0.766**	0.011	0.355	0.635	0.009	0.355	0.635	0.010	Oc
F	0.671	0.329	0	**0.718**	0.282	0	0.573	0.423	0.004	Oc

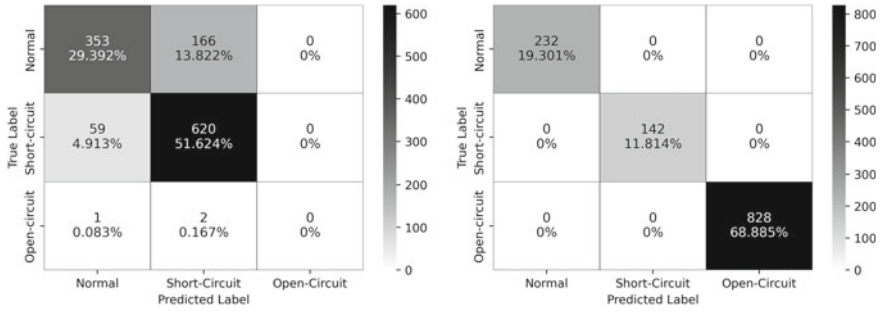
the algorithm in detecting failures, we considered as the final condition of the sample the maximum probability value obtained in one of the three portions of the frontier. In Table 3, for sample A, portion 1 indicates that the sample has a 100% chance of belonging to the same cluster of samples under normal conditions. For samples A and C, the condition suggested by a-posteriori analysis corresponds to the actual condition, but not for samples B, D, E, and F. The approach could not detect the faults for Open-circuit samples in this range. This difficulty may be associated with the small number of unlabeled and labeled samples of this condition in range 3.

Table 4 presents complementary information obtained for range 5. In this range, all unlabeled samples were correctly detected. Furthermore, for Open-circuit samples, the three portions reached the maximum probability ($p = 1$), as we can see in samples E and F of Table 4. We observed the same results for ranges 1 and 4 regarding Open-circuit faults. All portions of the non-dominated frontier obtained homogeneous results and efficiently detected Open-circuit faults.

To analyze the overall performance of data labeling, we used the Confusion Matrix to understand the potential of the proposed approach in fault detection. In the matrix, "True Label" corresponds to the actual condition of the record and "Predicted Label" to the label extracted from complementary information of a-posteriori analysis. Figure 4a presents the Confusion Matrix for irradiance range 3, with an accuracy of 0.8102. For range 3, 166 normal data (13.822%), 59 Short-circuit faults (4.913%), and 3 Open-circuit faults were mispredicted. As mentioned earlier, we had a few unlabeled and labeled Open-circuit samples for this range. Figure 4b presents the Confusion Matrix of range 5, with an accuracy of 1. In this case, all the unlabeled samples were predicted correctly.

Table 4 Complementary information extracted from the three portions of non-dominated frontiers about the irradiance range 5 (922.624–929.616 W/m^2) for the samples A to G

Sample	Irradiance 922.624–929.616									Actual condition
	Portion 1			Portion 2			Portion 3			
	No	Sc	Oc	No	Sc	Oc	No	Sc	Oc	
A	0.535	0.465	0	0.560	0.440	0	**0.858**	0.142	0	No
B	**1**	0	0	0.979	0.021	0	0.858	0.142	0	No
C	0.312	**0.688**	0	0.441	0.559	0	0.314	0.686	0	Sc
D	0.553	0.447	0	0.560	0.440	0	0.334	**0.666**	0	Sc
E	0	0	**1**	0	0	**1**	0	0	**1**	Oc
F	0	0	**1**	0	0	**1**	0	0	**1**	Oc



(a) 893.252-911.252 W/m^2 (range 3). (b) 922.624-929.616 W/m^2 (range 5).

Fig. 4 Confusion Matrix for the irradiance range 3 and 5. The values in the matrix represent the number of samples followed by the corresponding percentage of the total unlabeled data

5 Conclusion

This paper proposed MECA to capture the different clustering perspectives through objective functions. We proposed a-posteriori analysis to label data from the information of the non-dominated frontier. The frontier, divided into three portions, can evaluate which perspectives a cluster may have, providing information when we do not know the data distribution. Besides, a-posteriori analysis makes it possible to infer whether a cluster has more than one perspective (by analyzing the probabilities of the portions). The approach was applied in six irradiance ranges, and the results were validated through the accuracy and confusion matrix. A-posteriori analysis showed promising results regarding fault detection. In general, for all the irradiance ranges, the approach could detect the conditions of faults and normal data correctly, demonstrating a helpful approach to detecting and classifying faults in PV systems. The results also highlight the benefits of using non-dominated solutions to explore complementary information for data clustering. Relying on the irradiance range and

the type of fault, we observed that each portion of the frontier could better represent the data distribution, or we can have a homogeneous result between the three portions. This result indicates that using the division of the frontier instead of the entire non-dominated frontier could help better understand the data distribution. In future work, we would like to extend the proposed method to other problems requiring data labeling, such as semi-supervised learning problems.

Acknowledgements This work is financed by the ERDF—European Regional Development Fund, through the Operational Programme for Competitiveness and Internationalisation—COMPETE 2020 Programme under the Portugal 2020 Partnership Agreement, within project SmartPV, with reference POCI-01-0247-FEDER-068919.

References

1. Ali, H.M.: Recent advancements in PV cooling and efficiency enhancement integrating phase change materials based systems-a comprehensive review. Sol. Energy **197**, 163–198 (2020)
2. Dhimish, M., Holmes, V., Mehrdadi, B., Dales, M., Mather, P.: Photovoltaic fault detection algorithm based on theoretical curves modelling and fuzzy classification system. Energy **140**, 276–290 (2017)
3. Zhu, H., Lu, L., Yao, J., Dai, S., Hu, Y.: Fault diagnosis approach for photovoltaic arrays based on unsupervised sample clustering and probabilistic neural network model. Sol. Energy **176**, 395–405 (2018)
4. Tina, G.M., Cosentino, F., Ventura, C.: Monitoring and Diagnostics of Photovoltaic Power Plants. Springer International Publishing (2016)
5. Mellit, A., Tina, G., Kalogirou, S.: Fault detection and diagnosis methods for photovoltaic systems: a review. Renew. Sustain. Energy Rev. **91**(February), 1–17 (2018)
6. Lazzaretti, A.E., da Costa, C.H., Rodrigues, M.P., Yamada, G.D., Lexinoski, G., Moritz, G.L., Oroski, E., de Goes, R.E., Linhares, R.R., Stadzisz, P.C., Omori, J.S., dos Santos, R.B.: A monitoring system for online fault detection and classification in photovoltaic plants. Sens. (Basel, Switz.) **20**(17), 4688 (2020)
7. Et-taleby, A., Boussetta, M., Benslimane, M.: Faults detection for photovoltaic field based on k-means, elbow, and average silhouette techniques through the segmentation of a thermal image. Int. J. Photoenergy **2020**, 1–7 (2020)
8. Cai, Y., Lin, P., Lin, Y., Zheng, Q., Cheng, S., Chen, Z., Wu, L.: Online photovoltaic fault detection method based on data stream clustering. IOP Conf. Ser. Earth Environ. Sci. **431**(1), 012,060 (2020)
9. Zhao, Y., Ball, R., Mosesian, J., de Palma, J.F., Lehman, B.: Graph-based semi-supervised learning for fault detection and classification in solar photovoltaic arrays. IEEE Trans. Power Electron. **30**(5), 2848–2858 (2015)
10. Zhu, X.J.: Semi-supervised learning literature survey. Tech. Rep. TR-1530, Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA (2005)
11. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Mach. Learn. **109**(2), 373–440 (2020)
12. Mahela, O.P., Shaik, A.G.: Recognition of power quality disturbances using s-transform based ruled decision tree and fuzzy c-means clustering classifiers. Appl. Soft Comput. **59**, 243–257 (2017)
13. Handl, J., Knowles, J.: An evolutionary approach to multiobjective clustering. IEEE Trans. Evol. Comput. **11**(1), 56–76 (2007)

14. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002)
15. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. Oakland, CA, USA (1967)
16. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: an Introduction to Cluster Analysis. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, vol. 344, pp. 68–125 (1990)
17. Kruskal, J.B.: On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Am. Math. Soc. **7**(1), 48–50 (1956)
18. Rampazzo, P.C.B., Yamakami, A., de França, F.O.: Evolutionary approaches for the multiobjective reservoir operation problem. J. Control. Autom. Electr. Syst. **26**(3), 297–306 (2015)
19. Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S.: A survey of multiobjective evolutionary clustering. ACM Comput. Surv. (CSUR) **47**(4), 1–46 (2015)