IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 21, 2019, accepted April 1, 2019, date of publication April 9, 2019, date of current version April 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2909741

A Novel IGBT Health Evaluation Method Based on Multi-Label Classification

RUIKUN QUAN, HUI LI, (Member, IEEE), YAOGANG HU, AND PEI GAO
State Key Laboratory of Power Transmission Equipment and System Security and New Technology, School of Electrical Engineering, Chongqing University, Chongqing 400044, China

Corresponding author: Ruikun Quan (quanruikun@cqu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51675354 and in part by the Chongqing Artificial Intelligence Technology Innovation Project under Grant cstc2017rgzn-zdyf0117.

ABSTRACT The IGBT health evaluation of power semiconductor devices is usually based on the threshold evaluation method, which is usually a single characteristic parameter evaluation system. This kind of evaluation method cannot reflect the internal correlation of the change of multiple characteristic parameters in the deep level. Multi-label classification plays an important role in machine learning and can truly reflect the internal correlation principle of multi-feature parameters. Many studies have proved that multi-label classification (mlc) can effectively increase the actual classification effect of the clustering algorithm. In this paper, a clustering algorithm based on multi-label learning is applied to the health evaluation of IGBT. There are many characteristic parameters that affect each other in the actual work of IGBT, so it is difficult for a single label to reflect its actual health status. At the same time, multi-label data often belong to multiple classifications. Multi-label learning can improve the feature dependence ability of clustering method and improve the accuracy of classification. In this paper, we propose a multi-label classification learning model based on ISODATA for the multi-feature parameters of power semiconductor device IGBT, which can comprehensively consider the multi-level correlation effect of internal parameters in the multi-feature parameter extraction. The experiment results show that the algorithm model can better adapt to the IGBT health classification evaluation compared with the general clustering algorithm.

INDEX TERMS IGBT health evaluation, *ISODATA*, multi-label classification (mlc).

I. INTRODUCTION

Power semiconductor device (IGBT) is a very important core device in power electronics. It is widely used in wind power generation, photovoltaic power generation and power industry. IGBT is the core device of energy conversion and transmission, and the core component of power electronic device. Power conversion with IGBT can improve the efficiency and quality of power consumption, and has the characteristics of high efficiency, energy saving and green environmental protection [1], [2]. It is a key supporting technology to solve the problem of energy shortage and reduce carbon emissions. Power semiconductor devices generally have the following characteristics: (1) High voltage: general IGBT if the voltage of Vce is too high, it is easy to lead to device breakdown; (2) large current: the current of power device depends on W/L. The design current of IGBT module is increasing gradually under the requirement of high power supply [3]. The above conditions greatly increase the possibility of IGBT damage, so we need to study a method of IGBT health evaluation. At present, K-means clustering algorithm is the most commonly used in multi-label classification evaluation, and K-means algorithm is very sensitive to initial value [4]. K-means++ can avoid the initial value sensitivity problem by giving different random probability of sample points according to the distance from clustering center. In practice, the selection of K value can be specified by some indexes in machine learning, such as minimum loss function, better K value according to hierarchical classification, and contour coefficient of clustering, etc. In general, different data distribution is chosen in different ways, and dimensionless is needed when using distance-based algorithm, so as to prevent the sample from being too large in some dimensions to cause the distance calculation to fail [5]. K-means algorithm is a special case of Gaussian mixed clustering where the variance of mixed components is equal,

The associate editor coordinating the review of this manuscript and approving it for publication was Omid Kavehei.

2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

and only one mixed component is assigned to each sample. In step E, we fix the center of each class by selecting the nearest class for the sample to optimize the objective function, and in step M, we update the center point of each class [6]. This step can be realized by differentiating the objective function, and finally the new class center is the mean value of the sample in the class. At present, IGBT health evaluation based on junction temperature evaluation, loss evaluation, welding layer fatigue degree and other evaluation indicators, such as the general use of threshold value, but it can not provide a deep-level multi-parameter evaluation system. In this study, by using multi-label classification method to evaluate the health degree, we can discover the nonlinear and invisible mapping laws of the feature parameters by fully mining the intrinsic correlation of each characteristic parameter of IGBT. A data evaluation model with coupling of each characteristic parameter is implemented. By using iterative self-organizing (*ISODATA*) clustering algorithm to establish a health classification and evaluation method, a multi-label classification model for health evaluation of IGBT multi-feature parameters is proposed.

II. CURRENT STUDY OF IGBT HEALTH EVALUATION

A. PARAMETER CHARACTERISTICS OF POWER SEMICONDUCTOR IGBT

The power semiconductor device IGBT module is mainly composed of a number of IGBT chips, each of which is electrically connected by aluminum wires. In a standard IGBT package, a single IGBT will also have a sequel diode, and then a large amount of silica gel will be poured over the chip, and finally, the plastic case will be encapsulated, the IGBT unit stacked, and the chip from top to bottom. DBC (Directed Bonding Copper) and metal heat-plate (usually copper) are composed of three layers. DBC consists of three layers of materials, two layers are metal layers, and the middle layer is insulating ceramic layer. DBC performs better than ceramic substrates: it has lighter weight, better thermal conductivity, and better reliability, as shown in Figure 1:

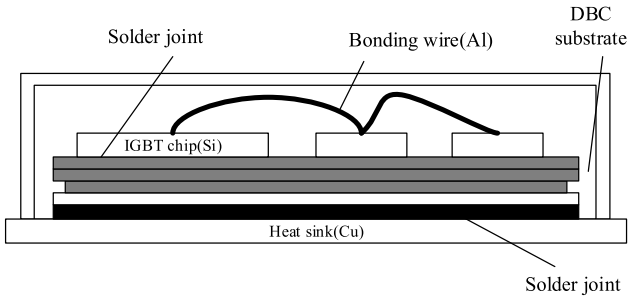

FIGURE 1. The IGBT module encapsulation structure.

Power semiconductor devices (IGBT) generally work at a high operating frequency. At present, the high switching characteristics of most of them are tens to hundreds of KHz, which is an important factor affecting their health. The operation process of the switch is as follows: IGBT turns off at T1 (T1 to describe IGBT turn-off behavior), at T2, IGBT is turned on again (T2 is used to describe the on-off behavior of IGBT), if the load resistance and diode voltage drop are ignored, The IGBT receives all DC bus voltages before opening the U_{DC}, IGBT process as shown in Figure 2:

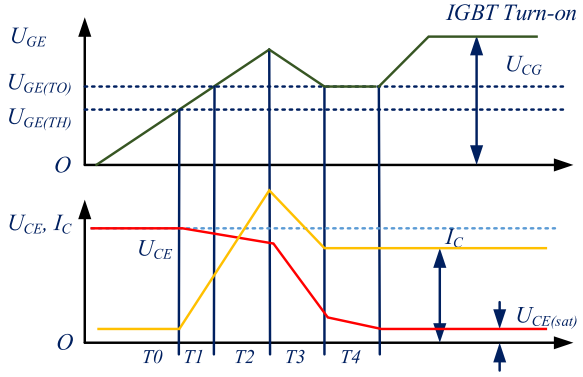

FIGURE 2. The loss of the IGBT switch model.

After the IGBT is turned on, the gate-collector voltage U_{CE} begins to rise, and at time T3, the U_{CE} rises to the threshold voltage $U_{CE(to)}$, At this time, the collector current I_C begins to rise, the collector current increases to produce the current change rate, at the same time because of the stray inductance in the internal path of the converter, the collector emitter voltage U_{CE} drops rapidly, namely:

$$U_{CE} = U_{DC} - \Delta U_{CE} = U_{DC} - L_\sigma \bullet \frac{di_C t_4 - di_C t_3}{dt}$$

At the time of T4, the collector current I_C has risen to a rating determined by the size of the inductor. At this point, however, the diode begins to turn off, and the I_C continues to increase due to the diode's reverse recovery characteristics. Transfer characteristics of Power Semiconductor device IGBT: IGBT transfer characteristic refers to the relationship between the output collector current (I_C) and the gate emitter voltage (U_{GE}). The switching voltage $U_{GE(th)}$ is the lowest gate voltage on IGBT. The IGBT is turned off when the gate voltage is less than the open voltage $U_{GE(th)}$. There is a linear relationship between I_C and U_{GE} in the range of collector current after IGBT, and the maximum gate emitter voltage is limited by the maximum collector current, which is generally about 15v. In IGBT converter, IGBT is switched back and forth in the forward blocking region and saturation region. The on-state saturation voltage drop $U_{GE(sat)}$ of IGBT is a function of junction temperature T_j, collector current I_C and gate emitter voltage U_{GE}. The saturation voltage drop of IGBT is lower than that of MOSFET and is close to that of GTR. The saturation voltage drop decreases with the increase of gate voltage, and the increase of U_{GE} will increase the conductivity of channel, thus reducing the $U_{GE} \geq 15$v. $U_{ce(sat)}$ also increases with the increase of collector current I_C: when I_C is small, $U_{CE(sat)}$ decreases with the increase of junction temperature, that is, it has a negative temperature coefficient. Once I_C exceeds a certain value, it becomes a positive temperature coefficient. The positive temperature coefficient of $U_{CE(sat)}$ is favorable to the parallel connection

of IGBT [7], [8]. The static switching characteristics of the power semiconductor device IGBT: the ideal static switching characteristics of the IGBT, including the on-on process and the turn-off process. The static switching characteristic of the IGBT refers to the relationship between collector current and collector voltage, and the IGBT is in the on-state. Because its PNP transistor is a wide-base transistor, although the equivalent circuit is Darlington structure, the current flowing through MOSFET becomes the main part of the total current of IGBT. At this point, the on-state voltage $U_{DS(on)}$ can be represented by the following formula:

$$U_{DS(on)} = U_{J1} + U_{DR} + I_C R_{OH}$$

U_{J1} is the forward voltage of the JI junction with a value of 0.7V, U_{DR} is the voltage on the spreading resistor R_{DR}, and R_{OH} is the channel resistance. Because of the conductance modulation effect in N+ region, the on-state voltage drop of IGBT is small. When the on-state voltage drop of 1000V 2∼3v.IGBT is off, there is only a very small leakage current. When IGBT is on, I_C is determined by external circuit. When IGBT is turned off, I_C is zero. When the IGBT is on, the conduction modulation of the N- region is carried out from the hole injected into the N- region from the P region, which reduces the R_{DR}, in the N- region and makes the IGBT have lower on-state voltage drop. Figure 3 shows the inverter circuit of the power semiconductor device IGBT:

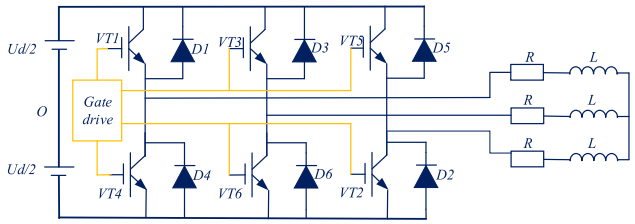

FIGURE 3. The inverter circuit of the power semiconductor device IGBT.

B. IGBT HEALTH EVALUATION BASED ON THERMAL RESISTANCE MODEL

Thermal resistance model is a health assessment model based on the circuit topology and heat dissipation of power semiconductor devices. The model takes into account the implementation loss and transient thermal impedance of power semiconductor devices in real time. Thermal resistance model can be used in high power and high current IGBT. Generally, the thermal characteristics of power semiconductor devices can be simulated and modeled by considering the packaging structure of the device, including silicon substrate, DBC and copper substrate [9]. Then according to the circuit topology and the operation condition of the device, the junction temperature loss of the device in real time operation is calculated by simulation. The model requires high accuracy of device operating condition and thermal resistance analysis. In general, the device will be aged to varying degrees in practice, which will bring error to the health assessment of power semiconductor device. In addition to heat transfer and heat storage, the device packaging structure has the ability to store heat. Heat capacity C_{th} represents the relationship between heat Q_{th} and T, and heat capacity can be used to describe the ratio of heat change to temperature difference [10], [11]. In general, T_j can be used to express the junction temperature of IGBT chip, and T_c to represent the temperature of copper bottom shell of the module, $T_{s(th)}$ to indicate the temperature of radiator, $R_{th(j-c)}$ to indicate the thermal resistance of, $R_{th(c-s)}$ to the shell to indicate the thermal resistance of the shell to the radiator. Shell temperature T_c is the temperature of the bottom surface of the copper substrate directly below the chip. Based on the definition and measurement of steady-state R_{th}, the dynamic effect of thermal resistance should be considered in practice. Using thermal resistance R_{th} and heat capacity C_{th}, a simulated thermal model can be constructed, which can be represented by transient thermal resistance or thermal impedance Z_{th}. The thermal impedance of dynamic heat transfer characteristics is reflected by the equivalent circuit of thermal resistance. The parameters of thermal impedance Z_{th} (R_i and τ_i) are generally given in the reference document, that is:

$$\tau_i = R_{thi}{}^* C_{thi}$$

Finally, the relevant junction temperature and the healthy state of the power semiconductor device can be obtained by the corresponding dynamic response curve. Figure 4 shows the thermal resistance model of power semiconductor devices:

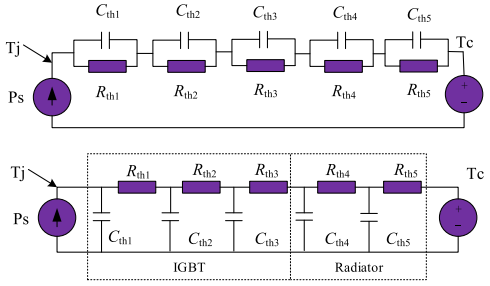

FIGURE 4. The thermal resistance model of power semiconductor devices.

C. IGBT HEALTH EVALUATION BASED ON POWER LOSS MODEL

Although the power semiconductor device (IGBT) mainly works in the switching state, IGBT is still a power electronic device with high power loss, and with the increase of the switching frequency of the device, the switching power loss of the device will also increase. The performance of power semiconductor device IGBT module is closely related to its switching frequency [12]–[14]. The switching characteristics of IGBT directly determine the switching power loss of the device, and the switching loss will restrict the improvement of the device working efficiency. At the same time, the switching power loss of the power semiconductor device IGBT will produce a lot of heat, which will cause a great temperature rise, and also has a great impact on the reliability of the

IGBT device. The switching power loss of the IGBT module is determined by the switching characteristics of the IGBT. It is related to collector-emitter voltage U_{CE} and collector current I_C. The formula for calculating loss is as follows:

$$P_{on} = \frac{1}{t_{on}} \int_0^{t_{on}} U_{CE}(t) I_C(t) dt$$
$$P_{off} = \frac{1}{t_{off}} \int_0^{t_{off}} U_{CE}(t) I_C(t) dt$$

Among them, P_{on} is turn-on power loss, P_{off} is turn-off power loss, t_{on} is turn-on time, t_{off} is turn-off time, U_{CE} is collector to emitter voltage, I_C is collector current.

In practical application, the corresponding simulation model can be built to calculate the switching loss of IGBT, and then the dynamic characteristics of the power module can be obtained by the physical modeling of power supply, reactance and power device. The instantaneous power, current and voltage waveforms of IGBT are obtained, and the power loss of the device is calculated. The advantage of switching power loss calculation based on simulation model is high precision, which can simulate the static and dynamic working characteristics accurately according to the structure and package of the device, but the disadvantage is that the physical simulation model is very complex. Moreover, the parameter structure and packaging process of the model are difficult to obtain accurately. Figure 5 shows the switching loss waveform of the device simulation model:

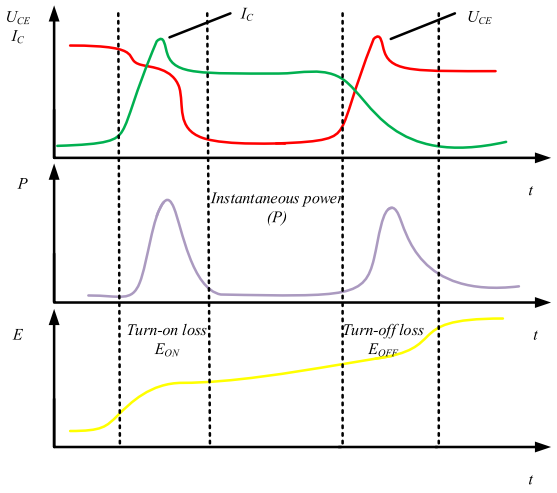

FIGURE 5. The switching loss waveform of the device simulation model.

III. MULTI-LABEL CLASSIFICATION THEORY

A. MULTI-LABEL DATA THEORY

In multi-label learning, the feature selection of information data is an important means to improve the performance of classifier [15]–[17]. In order to describe the multi-label special selection algorithm in this study, we assume that the information data feature parameters of a practical problem are : $A=\{a1, a2, a3, \ldots\ldots, an\}$, At the same time, all variables are independent random variables, we do the following definition:

Definition 1:

Let an instance $A = \{a1, a2, a3 \ldots\ldots, an\}$, where $ai \in R$, it represents the eigenvector of a sample with a possible tag space, assume an optional set of tags $P = \{p1, \ldots\ldots, pm\}$, and the set of tags $X = \{p1, \ldots\ldots, pl\} l \leq m$ corresponding to that instance, that is, X is a subset of P, X_i represents the set of tags corresponding to the instance, then the multi-label data of the sample can be described as follows

$$D = \{(Ai, Xi) | 1 \leq i \leq t, Ai \in R, Xi \subseteq P\} \tag{1}$$

where the entropy of data information is:

$$R(A) = -\sum_{i=1}^{n} p(xi) \log_2 p(xi) \tag{2}$$

Definition 2:

If the U and V are two independent random variables, and $U = \{u1, u2, \ldots\ldots, um\}$, $V = \{v1, v2, \ldots\ldots, vn\}$, then the joint information entropy of U and V, conditional information entropy of U and V is calculated as follows:

$$R(U, V) = -\sum_{i=1}^{m} \sum_{j=1}^{n} p(ui, vj) \log_2 p(ui, vj) \tag{3}$$

$$R(U|V) = R(U, V) - R(U) = -\sum_{i=1}^{m} \sum_{j=1}^{n} p(ui, vj) \log_2 p(ui, vj) \tag{4}$$

where the joint information entropy and conditional information entropy of U and V can be obtained statistically, the correlation degree of U and V is as follows:

$$\begin{aligned} MI(U; V) &= R(U) - R(U|V) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} p(ui, vj) \log_2 \frac{p(ui|vj)}{p(ui)} \end{aligned} \tag{5}$$

Assuming a random variable $W = \{w1, \ldots\ldots, wk\}$, if the constraint is known, then the conditional mutual information of U and V is:

$$\begin{aligned} MI(U; V|W) &= MI(U; V, W) - MI(U; W) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{k} p(ui, vj, wk) \log_2 \frac{p(ui|vj, wk)}{p(ui|wk)} \end{aligned} \tag{6}$$

Assuming that there are t and $\bar{t}$ in the sample features, then

$$R(U|T) = p(t) R(U|t) + p(\bar{t}) R(U|\bar{t}) \tag{7}$$

Finally, the information gain can be obtained:

$$\begin{aligned} IG(T) = &-\sum_{i=1}^{n} p(U_i) log_2 p(U_i) \\ &+ p(t) \sum_{i=1}^{n} p(U_i|t) log_2 p(U_i|t) + p(\bar{t}) \\ &\times \sum_{i=1}^{n} p(U_i|t) log_2 p(U_i|t) \end{aligned} \tag{8}$$

B. ITERATIVE SELF-ORGANIZING (ISODATA) CLUSTERING ALGORITHM THEORY

Assuming that the input sample $\{xi, i = 1, \ldots\ldots, N\}$ is a pattern sample data, the initial cluster center $\{z1, \ldots\ldots, z_{N_c}\}$ of the N_c pattern sample is first set. In the actual situation, the number of the general N_c cannot be equal to the actual number of the required clustering center. And its initial position can be selected from the sample at any time. After setting the number of initial cluster centers, the data for each pattern sample is divided into the nearest cluster, namely:

$$D_j = min\{||x - zi||, i = 1, 2, 3 \ldots\ldots, N_C\} \tag{9}$$

Through the different cluster center allocation of the pattern sample data, the distance between the sample data and the clustering center is minimum, which is consistent with each other. After the sample data is classified by the minimum distance, if the sample number $S_j < \theta_N$, then the sample is discarded, at the same time, the value of Nc is subtracted by 1, according to the actual sample's clustering center distance, each cluster center is adjusted and corrected. After the cluster center is adjusted and corrected, the average distance between the sample data and the cluster center of each cluster and the total average distance between the whole sample data and the corresponding cluster center are calculated:

$$\overline{D} = \frac{1}{N}\sum_{j=1}^{N} N_j \overline{D_j} \tag{10}$$

According to the number of iterations, if it is the last iteration, then θ_C is set to 0, if $N_C \leq K/2$, that is the number of initial cluster centers set is less than or equal to half of the specified value K. If the number of iterations $N_C \geq 2K$, then there is no split operation, that is, to calculate the standard deviation vector of the sample data distance of each initial cluster center.

$$\sigma_j = (\sigma_{1j}, \sigma_{2j}, \sigma_{3j} \ldots\ldots, \sigma_{nj})^T \tag{11}$$

The components of a vector are:

$$\sigma_{ij} = \sqrt{\frac{1}{N_j}\sum_{k=1}^{N_j}(x_{ik} - z_{ij})^2} \tag{12}$$

By calculating the sample standard deviation vector of each cluster center, the maximum component is obtained $\sigma_{j\max}, j = 1, 2, 3 \ldots\ldots, N_C$. If any of the maximum components, $\sigma_{j\max} > \theta_S$, and $\overline{D_j} > \overline{D}$, $N_j > 2(\theta_N + 1)$, $N_C \leq K/2$ is satisfied at the same time, then Z_j will be split into two new clustering centers. And the number of N_c plus 1.

Calculate the total cluster center distance with the following: $D_{ij} = ||Z_i - Z_j||, i = 1, 2, 3 \ldots\ldots, N_C - 1, j = i + 1, \ldots\ldots, N_C$, and then by comparing the values of the D_{ij} and θ_C, the $D_{ij} < \theta_C$ sample values are arranged incrementally in order of minimum distance, that is: $\{D_{i1j1}, D_{i2j2}, D_{i2j2} \ldots\ldots, D_{iLjL}\}$, and thereinto $D_{i1j1} < D_{i2j2} < D_{i2j2} \ldots\ldots < D_{iLjL}$, By combining the two clustering centers Z_{ik} and Z_{jk} with distance D_{ikjk}, the new cluster centers are as follows:

$$z_k^* = \frac{1}{N_{ik} + N_{jk}}[N_{ik} z_{ik} + N_{jk} z_{jk}], \quad k = 1, 2, 3 \ldots\ldots, L \tag{13}$$

Among them:

K is the expected number of cluster centers;

θ_N is the minimum number of samples for each cluster, if the sample number is less than this value, it will not be a cluster center;

θ_S is the standard deviation of the sample distribution of each cluster center;

θ_C is the minimum distance of any two clustering centers. If the distance value of the two clustering centers is less than the value in practice, then the two clusters need to be merged into one cluster;

L is the maximum logarithm of cluster centers that can be merged in iterative operations;

I is the number of iterations in actual operations.

IV. IGBT HEALTH CLASSIFICATION BASED ON *ISODATA* CLUSTERING MODEL

The *ISODATA* algorithm needs to specify more parameters, so it is generally possible to specify a more reasonable value according to the actual sample data. The main difference between the *ISODATA* algorithm and K-means, K-means++ clustering algorithm is whether the clustering center K is fixed or not. The clustering center K of the traditional clustering algorithm, K-means, K-means++ is fixed [18]. The *ISODATA* algorithm can adjust the number of cluster centers by two operations according to the actual situation of each category of sample data: (1) splitting operation, corresponding to increase the number of cluster centers; (2) merging operation, corresponding to reducing the number of cluster centers.

The input of *ISODATA* algorithm is as follows: (1) the expected number of clustering centers K0: although the number of cluster centers is variable in the *ISODATA* running process, it is still necessary for the user to make an initial number of cluster centers K0 according to the actual sample data. The range of the number of cluster centers is also determined by K0, and the final output range of cluster centers is [K0 / 2, 2K0]; (2) N_{min} is the minimum number of samples required for each sample data class in a cluster center. This value is used to determine whether a split operation can be performed when the sample contained in a class is highly dispersed. If the number of samples included in a subcategory is less than N_{min}, after splitting, the class will not be split; (3) the maximum variance sigma: it can measure the dispersion of samples in a certain category. When the sample dispersion of the cluster center exceeds this value, If the two classes are very close to each other (that is, the distance between the two classes corresponds to the cluster center is very small), it is possible to split the cluster operation.

(4) The minimum allowable distance D_{min}: between any two classes in a cluster corresponding to a cluster center, If the two classes are very close (that is, the distance between the two classes corresponding to the clustering center is very small), the two categories need to be merged.

The algorithms are as follows:

Algorithm 1 *ISODATA* Clustering Algorithm

Step1: Randomly selects K0 sample data as the initial clustering center $C = \{c_1, c_2, c_3 \ldots\ldots, c_{k0}\}$;

Step2: Calculates the distance from each sample *xi*, in the dataset to K0 cluster centers and classifies them into the classes corresponding to the least distance cluster centers;

Step3: Determines whether the number of elements in each class is less than that of N_{min},. If it is less than this value, the classification is abandoned, $K = K - 1$, and the sample data in the class is reassigned to the class corresponding to the cluster center with the smallest distance;

Step4: Recalculates the cluster center of each cluster center for the classification ci, corresponding to that cluster center

$$c_i = \frac{1}{|c_i|} \sum_{x \in c_i} x$$

Step5: If the number of cluster centers $K \leq K0/2$, do the split operation;

Step6: If the number of cluster centers $K \leq 2K0$, do the merge operation;

Step7: Terminates if it reaches the maximum number of iterations, otherwise it returns to step2 to continue execution;

ISODATA is a kind of clustering method based on iterative self-organization, which has been widely used in clustering. *ISODATA* clustering method is aimed at high latitudes, and the massive data can estimate the value of K effectively. It improves the shortcoming of the traditional K-means clustering algorithm, which is fixed and invariant in the process of computing [19]. At the same time, the clustering algorithm can also be used in multiple datasets and discard the clustering algorithm when the number of samples belonging to a certain category is too small. The cluster is divided into two clusters when the number of samples is too large and the degree of dispersion is too large. Aiming at the high dimensional data structure of IGBT health state evaluation, a clustering algorithm model based on iterative self-organization is proposed to improve the clustering reliability. Figure 6 shows the basic process of IGBT health state classification based on the iterative self-organized (*ISODATA*) clustering algorithm model:

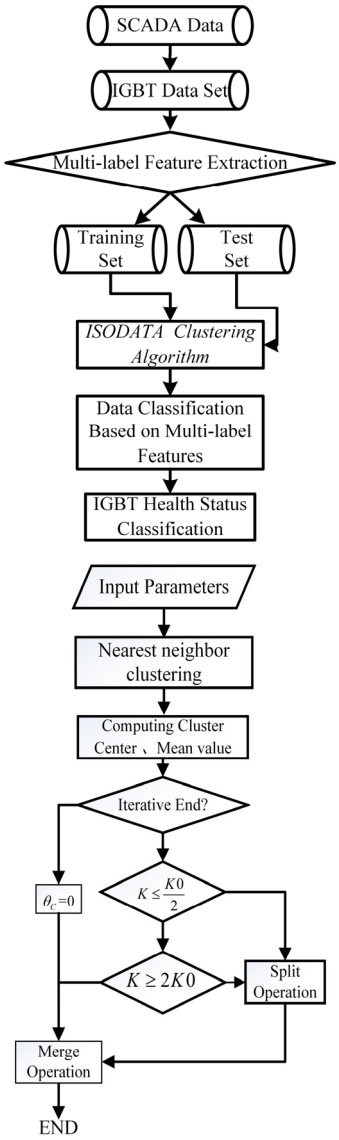

FIGURE 6. The iterative self-organizing (*ISODATA*) clustering algorithm.

V. EXPERIMENTS AND RESULTS

A. DATA SET

The data collected by SCADA system of wind turbine in a wind farm are randomly selected. The main data types are wind speed, hub speed, generator operating frequency, generator current, generator voltage, generator side power of inverter, etc. We select the wind speed of the weather station, the power of the generator side of the frequency converter, the temperature of the frequency converter as the data characteristic label for the wind turbine unit 1. We select the wind speed of the weather station, the generator side power of the frequency converter, the temperature of the frequency converter for the wind turbine unit 2. The ambient temperature is the data feature label, and after data cleaning, the data set structure is shown in Table 1:

TABLE 1. Experiment data set.

Data Set	Sample Number	Characteristic Dimension	Label Dimension	Label Cardinality
Unit 1	972	6	3	2.685
Unit 2	1071	6	4	1.312

B. EXPERIMENTAL SETTING

Clustering evaluation is to evaluate the feasibility of clustering on the dataset and the quality of the results produced by the clustering method, including: estimating the clustering trend, determining the number of clusters in the data set, and determining the clustering quality. Estimation of clustering trends means that for a given data set, the existence of a non-random structure of the data set is evaluated. If the clustering method is used blindly on the dataset, the resulting clustering may be misleading. Clustering trend assessment determines whether a given data set has a non-random structure that can lead to meaningful clustering, a data set that does not have any non-random structure, such as a uniformly distributed point in a data space. Although the clustering algorithm can return clusters for the data set, these clusters are random and meaningless, and the clustering requires the non-uniform distribution of the data. In the experiment, the *ISODATA* clustering algorithm is used to classify the health status of IGBT, and the multi-label feature data in Table 1 are used for clustering analysis. By using the concept of ''confusion matrix'', we establish the classifier evaluation index: true positive (true positive, *TP*): to predict positive class to positive class number; True negative (true negative, *TN*): predicts negative class to negative class; false positive (false positive, *FP*): that is negative class to positive class; false negative (false negative, *FN*): the positive class is predicted to be negative class number; Table 2 shows its structure:

TABLE 2. Confusion matrix structure.

	Positive	Negative
True	True Positive(TP)	True Negative(TN)
False	False Positive(FP)	False Negative(FN)

The two most basic indexes of cluster analysis are recall rate (Recall Rate), precision rate (Precision Rate),) recall rate also called recall rate, accuracy rate also called precision rate. Figure 7 shows the calculation of the confusion matrix:

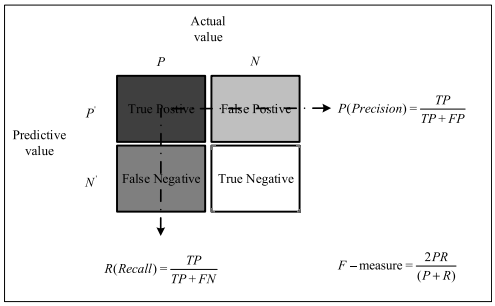

FIGURE 7. The Recall Rate and Precision Rate.

In practice, we usually use F1-Score and mAP indicators to evaluate classification performance.

C. EXPERIMENTAL RESULT

In the experiment, we used different threshold values to calculate the precision rate and recall rate of a group of different thresholds, and at the same time we used mAP (mean average precision) to solve the single point limitation of P, R, F-measure. Thus the index reflecting the global performance is obtained. Figure 8 shows the veracity-recall curve of different threshold values of different clustering algorithms:

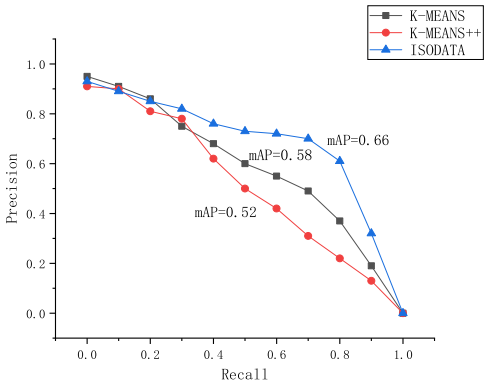

FIGURE 8. The mAP value of K-MEANS, K-MEANS++, ISODATA.

As you can see, although the verity-recall curves of the different algorithms overlap, the performance of the *ISODATA* algorithm is in most cases much better than that of the other algorithms.

In addition, we can also use ROC (receiver operating characteristic) and AUC (area under roc curve) as classifier evaluation indicators. ROC is concerned about the tradeoff between TP(real positive example) and FP(error positive example). By setting a threshold, we classify instances into positive or negative classes (for example, greater than the threshold value is divided into positive classes). AUC values generally range from 0.5 to 1.0, with larger AUC representing better classification performance. Figure 9 shows the AUC values of different clustering algorithms at different tag selection rates:

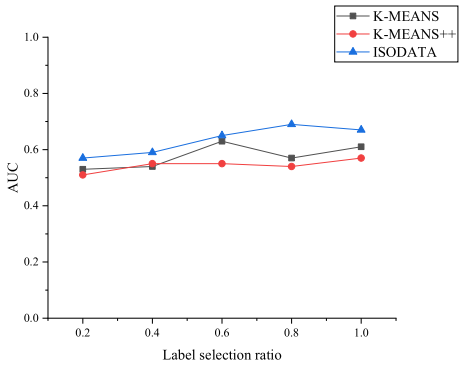

FIGURE 9. The AUC value of K-MEANS, K-MEANS++, ISODATA.

Finally, we use the iterative self-organization clustering algorithm to analyze the multi-label classification of the data set. In the experiment, we select part of the tag feature data, and then we quantify and normalize the data. Figure 10 shows

that multi-label data can be clustered with multiple classifications, and the clustering results of different health states of IGBT can be obtained:

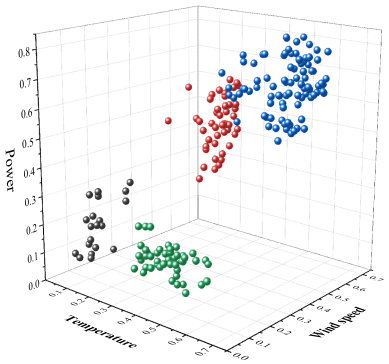

FIGURE 10. **Multi-label data can be clustered with multiple classifications.**

VI. CONCLUSION

An iterative self-organizing clustering algorithm based on multi-label classification is used to classify the IGBT health status of wind power converters, in which the multi-label data are not only related to one of them. The multi-label attribute determines that the data may also be part of other clustering, so a multi-label classification algorithm based on iterative self-organization is used to classify the data. The experimental results show that the proposed clustering algorithm has better classification performance and can be used for state clustering of multi-label data. The data driven method is used to excavate the different health states of IGBT, which provides a new reference method for further evaluation and prediction of its working condition, and also provides support for the predictive maintenance of PHM in wind farm IGBT. In the future, we will further study a multi-label classification algorithm with higher classification performance and accuracy.

REFERENCES

[1] S. Yang, D. Xiang, A. Bryant, P. Mawby, L. Ran, and P. Tavner, "Condition monitoring for device reliability in power electronic converters: A review," *IEEE Trans. Power Electron.*, vol. 25, no. 11, pp. 2734–2752, Nov. 2010.

[2] S. Yang, A. Bryant, P. Mawby, D. Xiang, L. Ran, and P. Tavner, "An industry-based survey of reliability in power electronic converters," *IEEE Trans. Ind. Appl.*, vol. 47, no. 3, pp. 1441–1451, May/Jun. 2011.

[3] B. Gadalla, E. Schaltz, and F. Blaabjerg, "A survey on the reliability of power electronics in electro-mobility applications," in *Proc. Int. Aegean Conf. Elect. Mach. Power Electron.*, Sep. 2015, pp. 304–310.

[4] G. Nasierding, G. Tsoumakas, and A. Z. Kouzani, "Clustering based multi-label classification for image annotation and retrieval," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 2009, pp. 4514–4519.

[5] M. S. Ahmed, L. Khan, and M. Rajeswari, "Using correlation based subspace clustering for multi-label text data classification," in *Proc. IEEE Int. Conf. Tools Artif. Intell.*, Oct. 2010, pp. 296–303.

[6] L. Grady and E. L. Schwartz, "Isoperimetric graph partitioning for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 469–475, Mar. 2006.

[7] M. Amiri, M. Amiri, and E. Afjei, "A novel distributed FACTS controller based on combined two half-bridge inverter," in *Proc. Electr. Power Conf.*, Oct. 2008, pp. 1–4.

[8] F. W. Fuchs, "Some diagnosis methods for voltage source inverters in variable speed drives with induction machines—A survey," in *Proc. 29th Annu. Conf. IEEE Ind. Electron. Soc.*, Nov. 2003, pp. 1378–1385.

[9] U.-M. Choi *et al.*, "Reliability improvement of power converters by means of condition monitoring of IGBT modules," *IEEE Trans. Power Electron.*, vol. 32, no. 10, pp. 7990–7997, Oct. 2017.

[10] U.-M. Choi, F. Blaabjerg, S. Munk-Nielsen, S. Jørgensen, and B. Rannestad, "Condition monitoring of IGBT module for reliability improvement of power converters," in *Proc. IEEE Transp. Electrific. Conf. Asia–Pacific*, Jun. 2016, pp. 602–607.

[11] T. J. Kärkkäinen and P. Silventoinen, "Considerations for active condition monitoring in power electronic converters," in *Proc. 15th Eur. Conf. Power Electron. Appl.*, Sep. 2013, pp. 1–5.

[12] R. Moeini, P. Tricoli, H. Hemida, and C. Baniotopoulos, "Increasing the reliability of wind turbines using condition monitoring of semiconductor devices: A review," in *Proc. IET Int. Conf. Renew. Power Gener.*, Sep. 2016, pp. 1–6.

[13] P. Ghimire, A. R. de Vega, S. Beczkowski, B. Rannestad, S. Munk-Nielsen, and P. Thogersen, "Improving power converter reliability: Online monitoring of high-power IGBT modules," *IEEE Ind. Electron. Mag.*, vol. 8, no. 3, pp. 40–50, Sep. 2014.

[14] K. Ma, A. S. Bahman, S. Beczkowski, and F. Blaabjerg, "Complete loss and thermal model of power semiconductors including device rating information," *IEEE Trans. Power Electron.*, vol. 30, no. 5, pp. 2556–2569, May 2015.

[15] C. Nieuwenhuis and D. Cremers, "Spatially varying color distributions for interactive multilabel segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1234–1247, May 2013.

[16] A. J. Asman and B. A. Landman, "Characterizing spatially varying performance to improve multi-atlas multi-label segmentation," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2011, pp. 85–96.

[17] T. Ajanthan, R. Hartley, and M. Salzmann, "Memory efficient max flow for multi-label submodular MRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 886–900, Apr. 2019.

[18] N. Iam-On, T. Boongeon, S. Garrett, and C. Price, "A link-based cluster ensemble approach for categorical data clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 3, pp. 413–425, Mar. 2012.

[19] G. Tsoumakas, I. Katakis, and L. Vlahavas, "Random k-labelsets for multilabel classification," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 7, pp. 1079–1089, Jul. 2011.

RUIKUN QUAN received the M.S. degree in computer science and technology from Southwest University, Chongqing, China, in 2016. He is currently a Research Assistant with the Department of Electrical Machinery and Electrical Apparatus, School of Electrical Engineering, Chongqing University. His research interests include machine learning, complex systems, and machine intelligence.

HUI LI received the M.Eng. and Ph.D. degrees in electrical engineering from Chongqing University, Chongqing, China, in 2000 and 2004, respectively. He was a Postdoctoral Research Fellow with the Institute of Energy Technology, Aalborg University, Aalborg, Denmark, from 2005 to 2007. Since 2008, he has been a Professor with the Department of Electrical Machinery and Electrical Apparatus, School of Electrical Engineering, Chongqing University, where he is currently a Researcher with the State Key Laboratory of Equipment and System Safety of Power Transmission and Distribution and New Technology. His current research interests include wind power generation, design, and control of electrical machines.

YAOGANG HU received the B.S. degree in electrical engineering from Yanshan University, Qinhuangdao, China, in 2008, and the M.S. degree in electrical engineering from Chongqing University, Chongqing, China, in 2011, where he is currently pursuing the Ph.D. degree in electrical engineering. His research interests include condition monitoring, fault diagnosis, and lifetime prediction for wind turbine.

PEI GAO received the B.S. degree in physics from Southwest University, Chongqing, China, in 2016, where she is currently pursuing the M.S. degree in physics. Her research interests include the machine learning and complex systems.