





ORIGINAL RESEARCH PAPER

Power transformer fault diagnosis considering data imbalance and data set fusion

Yang Zhang¹  | Hong Cai Chen²  | Yaping Du¹  | Min Chen³ | Jie Liang⁴ | Jianhong Li⁴ | Xiqing Fan⁵ | Xin Yao⁶ 

¹Department of Building Service Engineering, Hong Kong Polytechnic University, Hong Kong, China

²Academy for Advanced Interdisciplinary Studies and Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China

³Research Institute, State Grid Zhejiang Electric Power Co., Ltd, Hangzhou, China

⁴Zhejiang Huayun Information Technology Co., Ltd, Hangzhou, China

⁵State Grid Zhejiang Electric Power Co., Ltd, Lishui Power Supply Bureau, Lishui, China

⁶Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China

Correspondence

Hong Cai Chen, Academy for Advanced Interdisciplinary Studies and Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen 518057, China.
Email: hc.chen@live.com

Associate Editor: Behzad Kordi

Funding information

Science and Technology Project of State Grid Corporation of China, Grant/Award Number: 5500-202019090A-0-0-00

Abstract

Improving the accuracy of transformer dissolved gas analysis is always an important demand for power companies. However, the requirement for large numbers of fault samples becomes an obstacle to this demand. This article creatively uses a large number of health data, which is much easier to obtain by power companies, to improve diagnosis accuracy. Comprehensive investigations from the view of both data set and methodology to deal with this problem are presented. A data set consists of 9595 health samples and 993 fault samples is used for analysis. The characteristics of the data set and the influence of the health data on diagnostic accuracy are discussed. The performance of many state-of-art algorithms that handle the imbalanced problem is evaluated. Meanwhile, an efficient fault diagnosis algorithm named self-paced ensemble (SPE) is presented. In SPE, classification hardness is proposed to include the data characteristic in the classification. This method can guarantee the diversity of the data set and keep high performance. According to the experiment results, the superior of SPE is confirmed and also proves that involving more health samples can improve transformer diagnosis when fault data are limited.

1 | INTRODUCTION

Power transformer is the essential equipment in the power system, and its reliability is vital to the operation of the whole system. Power transformer is subjected to high temperature, electrical stress and mechanical stress during long-term operation [1–3]. The harsh operation conditions will gradually deteriorate transformers and eventually fails them. Monitoring the operation status of the transformer is an effective approach to identify the incipient fault thus proper maintenance can be carried out before the power transformer descends into serious failure.

There are several methods available to monitor the status of the transformer [4–7], such as oil testing, partial discharge detection, power factor testing, infrared temperature measurement and so on. Oil testing is one of the most common tests used to evaluate the condition of transformers. It identifies faults based on various gases dissolved in the transformer, or called dissolved gas analysis (DGA) [8]. This method collects the oil of power transformers periodically to analyse fault-related gases liberating from the oil. According to concentrations of fault-related gases, the status of the transformer is determined.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *High Voltage* published by John Wiley & Sons Ltd on behalf of the Institution of Engineering and Technology and China Electric Power Research Institute.

With the development of sensor technology, DGA methods have been shifting from offline to online monitoring. Compared with manual collecting and testing the oil of power transformers periodically, an online DGA monitor measures the gas concentrations in realtime and reports the condition of the transformer timely. Thus, the reliability of the transformer can be improved, and simultaneously the manual maintenance can be reduced. However, it also brings new issues. The frequent nuisance warning or mis-warning of faults due to the low diagnosis accuracy makes this technique bring new burden to the industry. Frequently reporting faults will lead to extra maintenance costs. On the other hand, mistaken fault as health will lead to severe accidents. Therefore, to implement efficient online DGA of power transformer, a diagnostic tool with high accuracy is necessary.

During years of development, DGA method has evolved into two mainstreams: interpretation methods [9, 10] and artificial intelligence (AI) based methods [11]. Interpretation methods include IEC 60,599 method [12], Key Gas method [13], Duval Triangular method [14], Doernenburg method [15] and Rogers method [16]. They identify different types of faults on the basis of gas ratios. These methods are developed empirically and no mathematical formulation can be used. According to the field experience, the diagnosis accuracy of interpretation methods is limited and cannot provide interpretation for every combination of gas ratio [17]. These methods may provide different diagnosis results for the same transformer condition [18].

To address these issues, various AI techniques have been developed based on dissolved gas concentrations such as fuzzy logic [19], artificial neural networks [20], support vector machine [21], self-organizing map [22] and others [23–25]. In contrast to interpretation methods, AI methods learn the underlying relationships between gas concentrations and transformers faults using quantitative evidence-based theory. They can provide a diagnosis with higher accuracy and are widely adopted in practical operation. However, AI based methods still have limitations on the transformer diagnosis.

1.1 | Shortage of high-quality data

AI tools are data-driven that they require large amounts of samples to cover the full range of expected variation and null cases. The performance of AI methods is highly dependent on the quality of the data set. However, the occurrence of a transformer fault is an event with a relatively small probability. It is hard for a utility to collect enough fault samples from its own transformer to perform fault diagnosis. Plenty of reported works use public data sets [14] and data published in the literature to train the AI model, consequently result in a model with poor generalization ability. Nowadays, with the wide application of online monitoring technology, the availability of data grows rapidly. However, nearly all the data corresponds to the healthy condition of the transformer. Therefore, fault data are still lacking.

1.2 | Lack of investigations on data set fusion

To relieve the data deficiency, data set fusion is an efficient method. As mentioned above, DGA data can be collected from many sources such as public data sets, literature and private companies. Data sets combining different sources can bring diversity to the training of AI algorithms, and benefit AI algorithms to obtain better generality. However, such a data set may also bring difficulties because the characteristics and distributions of the data set from different sources may be inconsistency and discrepancy [26]. As a result, the capability of AI algorithms on data fusion is questionable.

1.3 | Lack of discussion on imbalance data set

As mentioned previously, the health data is accumulated rapidly by the utility with the increasing of operating time, however, fault data are still rare. This leads to an imbalanced data set and the imbalance ratio between health and fault is continuously increasing. In this scenario, AI algorithms perform poorly because they are generally designed to handle a balance datasheet. To avoid this problem, previous studies always train AI algorithms using a balance datasheet or with a low imbalance ratio [26, 27]. However, due to lots of health samples were abandoned, it will lose useful information in fault diagnosis. Theoretically, if the health data is collected sufficiently large and includes all the gas combinations of healthy state, the diagnosis can be completely correct. Therefore, we try to solve this problem in a converse way that we try to utilize the large number of health samples that we collect to improve the accuracy of diagnosis.

To address these limitations, this article presents a comprehensive investigation from the view of both data set and methodology. The first contribution of this article is to first show the influence of health samples on the performance of the transformer diagnosis. Previous studies mainly focus on fault samples. They train a classifier using a data set contains a large number of fault samples which is not the case for most utilities. The data set selected in this article is obtained from more than 100 transformers thus is very practical. The other contribution is providing an efficient algorithm for an imbalanced transformer diagnosis. As a huge number of health samples are involved, the data set is highly imbalanced. Traditional imbalanced classification algorithms are insufficient in this situation. Meanwhile, this article also provides a thorough discussion of data characteristics and reasons for performance differences for various algorithms. According to the results of the experiments, the above limitations are solved by using the method proposed in this article. The method proposed in this article helps power companies to improve the diagnosis accuracy from an entirely different perspective. Since this method does not excessively dependent on fault data, it is feasible and practical.

It should be noted that after the condition of the power transformer is determined, various AI algorithms [28] can further identify the type of faults easily since the data-sheet is now relatively balanced. Classification of this balance data has already been discussed elaborately in the literature and is not the focus of this article and thus will not be discussed.

The remainder of this article is structured as follows. Section 2 describes the data set fusion in this article and the data pre-processing techniques for classification. Data characteristic is also discussed in detail. In the following section, AI methods dealing with imbalanced data are briefly introduced. In Section 4, a novel algorithm for highly imbalanced data is presented. In Section 5, the experiment is carried out using different amounts of samples and different algorithms. The performance of oversampling algorithms and ensemble algorithms are compared and discussed in detail. Finally, the conclusion is drawn in Section 6.

2 | DATA SET AND PRE-PROCESSING

2.1 | Introduction of data set

A fusion dissolved gas data set is constructed in this article. Five different data sets are mixed together and used to establish the fault diagnosis model and test its performance. It includes a public data set IEC TC10 data set [14], the data collected by the Ministry of Electricity and Energy of Egypt [10], the data published in [11], two private data sets provided by North China Grid and Zhejiang Grid. In these two private data sets, the data was collected from more than 100 power transformers of different manufacturers at different voltage levels. The data set consists of 9595 health samples (gathered from transformers in a healthy status) and 993 fault samples. The details of these data sets are listed in Table 1.

2.2 | Data characteristic analysis

DGA data typically present highly skewed distributions. As the variance range of DGA data is presented in Table 2, the dissolved gas can have low concentrations of zero ppm (parts per million), but also can attain tens of thousands of ppm for others. The extreme variances can be a source of numerical imprecisions and overflow in AI algorithms [29].

The data from different companies might present different characteristics. To reveal the distribution variation between different data sets, histograms of gas concentrations from IEC data set and Zhejiang data set are displayed in Figure 1. Due to the skewed distribution is difficult to visualise, logarithmic transformation is adopted on the data set. It shows that the concentrations of C_2H_6 show opposite distribution centres for two data sets. In IEC data set, the distribution centre locates near five for health samples, on the contrary, this distribution centre corresponds to the fault samples in Zhejiang data set. Therefore, it brings the generalization issue when training the data using

TABLE 1 Details of dissolved gas data sets

Data set	IEC	Egypt	Paper [11]	North China	Zhejiang
Health	49	0	271	50	9225
Fault	99	240	117	440	97

TABLE 2 Gas concentrations variance range of the data set

Gas (ppm)		H_2	CH_4	C_2H_6	C_2H_4	C_2H_2
Health	Max	92,600	64,064	72,128	95,650	57,000
	Min	0	0	0	0	0
Fault	Max	8288	1120	784	1344	19,712
	Min	0	0	0	0	0

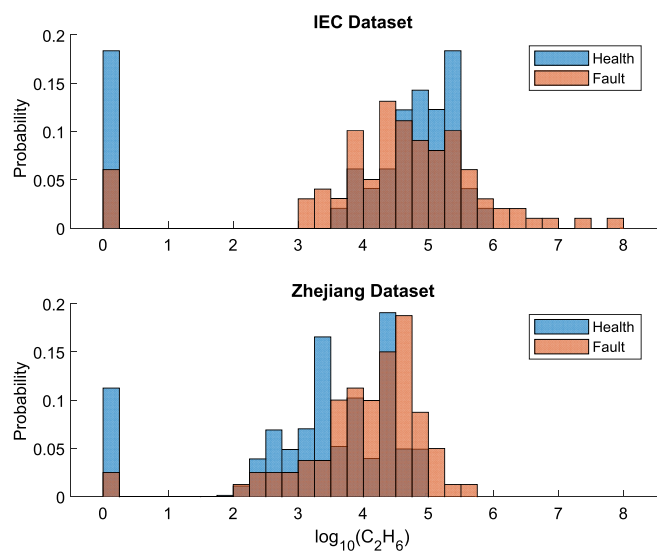


FIGURE 1 Histogram of \log_{10} concentrations of C_2H_6 for health and fault classes in IEC data set and Zhejiang data set. (Brown bars are overlaps of two classes)

different data sets. Meanwhile, although some fault samples are far from data clusters, they cannot be regarded as noise.

2.3 | Pre-processing DGA data

As the data characteristic presented in the previous section, raw DGA data have sharp variations among different concentrations of dissolved gases. Using data of different scales will cause some issues in the accuracy and convergence of AI model. To ensure accurate, efficient, or meaningful analysis, raw DGA data requires to be first transformed into the same scale.

Data transformation methods, such as normalization, and logarithm transformation [11], have been applied to rescale DGA data. In this article, the logarithm transformation of base 10 is adopted. It not only can rescale data that extreme values are avoided, for example, $\log_{10}(92,600) = 4.97$, but also transforms the gas ratios into linear relation, for example, $\log_{10}(CH_4/H_2) =$

$\log_{10}(\text{CH}_4) - \log_{10}(\text{H}_2)$. Thus, avoids overflows in AI algorithms and simplifies the relation among different gases. To avoid the infinity caused by the logarithm of zero, zero is replaced with a negligible quantity of 0.001. Finally, the data is shifted to the positive by minus its minimum, resulting in

$$x_{\text{new}} = \log_{10}(x) - \log_{10}(0.001) \quad (1)$$

where, x and x_{new} are values of gas concentrations before and after transformation.

3 | IMBALANCED CLASSIFICATION ALGORITHMS

The fault detection procedure essentially involves a process of pattern recognition. AI algorithms attempt to learn the inherent correlation in the data and, can provide an accurate and reliable transformer diagnosis. However, most of AI algorithms are developed for balanced data sets. It performs poorly on imbalanced data sets that the separating boundary can be shifted toward the minority class. This shift can cause the generation of more false negative predictions, which lowers the model's performance on the minority positive class.

To deal with the imbalance, three categories [30, 31] have been developed, namely resampling method, cost-sensitive method and ensemble method.

Resampling methods [32, 33] are designed to reduce between-class imbalance by either undersampling or oversampling. Distance thresholding or clustering is the essential criterion in the resampling methods to determine adding or removing samples. However, undersampling may discard instances that contain potentially useful information, as well as, reduces the total number of training instances. On the other hand, oversampling of the minority instances may lead to overfitting, and it also suffers from large computational cost.

Compared with resampling, cost-sensitive methods [34, 35] modify the existing balanced AI algorithms to modify its classification preference for most classes. They assign a higher cost to the misclassification of minority class during the training process. Consequently, more emphasis is put on the generalization of the minority class. A cost-sensitive classification technique takes the cost matrix into consideration during model building. A cost matrix encodes the penalty of classification from one class to another. However, the cost matrix on a specific task is given by domain expert before-hand, which is usually unavailable in many real-world problems.

Ensemble-based classifiers [36] are constructed by multiple classifiers and try to improve the generalization ability of classification by combining them to obtain a new classifier. The basic idea is to construct several classifiers from the original data and then aggregate their predictions when unknown instances are presented. General ensemble methods include bagging or boosting [37] and negative correlation learning [38]. Ensemble-based classifiers are also integrated with resampling

methods such as SMOTEBoost [39], RUSBoost [40], OOB, and UOB [41].

For the above reasons, none of the prevailing methods can perfectly handle the imbalanced, large-scale and noisy classification tasks. These methods are sensitive to different factors. Their capacity for transformer diagnosis requires to be studied, especially when the data fusion is involved.

4 | SELF-PACED ENSEMBLE ALGORITHM

The primary reason for the inefficiency of the above imbalanced classification algorithms is lacking full consideration of data distribution for the training. Though new algorithms are constantly emerging, seldom consider the characteristics of the minority class distribution and its influence on classification performance. As claimed in Ref. [42], collecting the information about local characteristics of the minority class and distinguishing between safe, borderline, rare, and outlier examples is useful to differentiate the performance of basic classifiers.

In this section, we present an ensemble algorithm named Self-Paced Ensemble (SPE) [43]. It considers the distribution of classification based on the concept of 'classification hardness' and iteratively selects the most informative samples according to the hardness distribution.

4.1 | Classification hardness

To integrate the data distribution characteristics, the concept of 'classification hardness' is introduced [43]. As the name implies, hardness indicates samples that hard to predict for classifiers. The classification hardness function H is defined as the overall error that is calculated by summing errors of individual classifiers. Any loss function can be used as a classified hardness function. The simplest form of the hardness function is Absolute Error, which is used in this work.

In the transformer fault diagnosis, ensemble method divides majority (health data) into many bins and trains a sequence of classifiers. The classification is implemented by average of all these classifiers. Suppose F is the trained ensemble classifier which is composed by n individual classifiers f_i . We use $F(\mathbf{x}_i)$ to denote the classifier's output probability of \mathbf{x}_i . Then the hardness of sample (\mathbf{x}, \mathbf{y}) with respect to F is given by the function $H(\mathbf{x}, \mathbf{y}, F)$ as

$$H(\mathbf{x}, \mathbf{y}, F) = \sum |F(\mathbf{x}) - \mathbf{y}| = \frac{1}{n} \sum_{i=1}^n |f_i(\mathbf{x}_i) - \mathbf{y}_i| \quad (2)$$

where, \mathbf{x}_i is the vector of input features in i th classifier and \mathbf{y}_i is the corresponding vector of ground truth.

The distribution of classification hardness contains information highly related to task difficulty, such as outliers, etc. Note that the classification hardness function uses the current model as one of the inputs to the function. Intuitively, the classification hardness gives the difficulty of classifying a specific sample for a

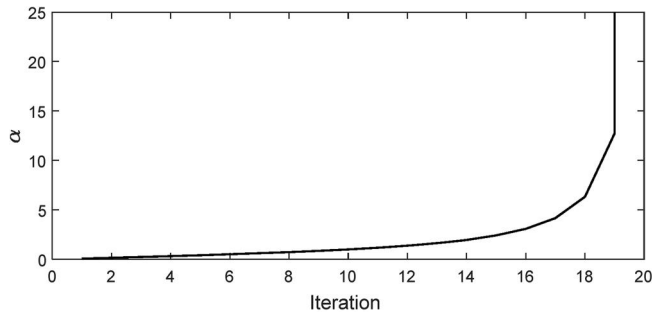


FIGURE 2 The change of balancing factor α with iterations

specific classifier. By observing the hardness distribution, we can get the fit of the model on the current data set.

4.2 | Balancing factor

At the beginning of training, SPE tries to equalize the classification hardness of each bin. However, as the training process evolves, the population of ‘simple’ samples grows rapidly because the ensemble classifier will gradually fit the training set. In this situation, a lot of fitted samples will be retained if the selection is based on keeping hardness the same, leading to a classifier lacking of diversity. Therefore, a balancing factor α is introduced to decrease the probability of those samples along with the iteration.

We use the tan function in Equation (3) to control the growth of balancing factor α . Thus we have in the first iteration $\alpha = 0$ and in the last iteration $\alpha \rightarrow \infty$, as shown in Figure 2.

$$\alpha = \tan(i\pi/2n) \quad (3)$$

where, i the index of iteration, and n is the total number of the iteration which is also equal to the number of classifiers in the ensemble.

When α goes large, we focus more on the harder samples instead of the simple hardness contribution. Through this mechanism, SPE gradually focuses on the harder data samples, while still keeps the knowledge of easy sample distribution in order to prevent overfitting.

4.3 | Self-paced ensemble

Integrating the concept of classification hardness and balancing factor, an ensemble algorithm SPE [43] is developed. Similar to boosting algorithms [37], SPE builds classifiers sequentially using undersampling, and obtains the final predictor by the summation of all classifiers. The main difference compared with other ensemble algorithms is the undersampling strategy.

To demonstrate the difference, boosting algorithm is briefly introduced. Boosting algorithm build classifiers sequentially as shown in Figure 3. It works by resampling a subset of majority and adjusting the weights on the training

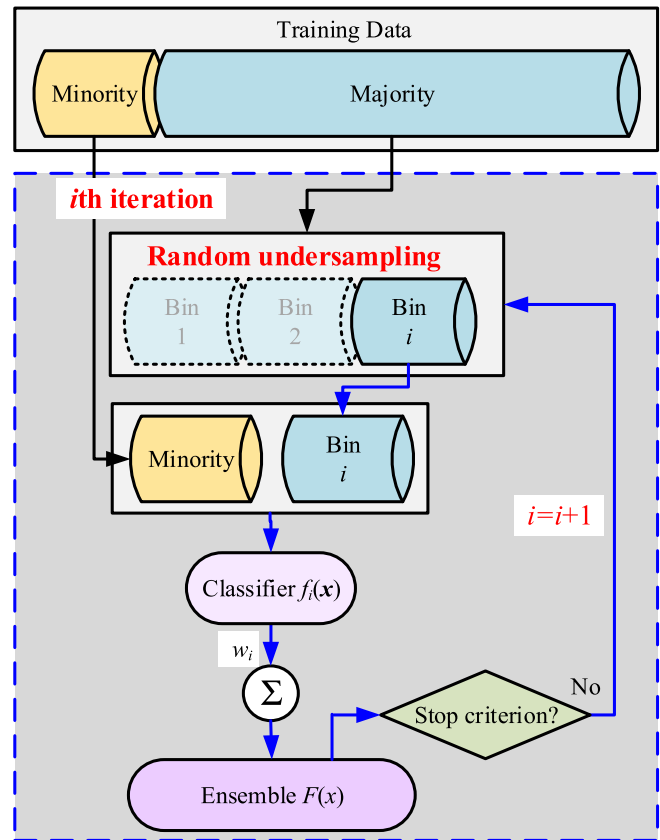


FIGURE 3 Flowchart of the boosting algorithm. Classifiers are ensembled sequentially

instances adaptively according to the performance of the previous classifiers. Higher weights are assigned to each classifier for wrongly classified examples. The outputs are then updated using the weighted average approach. The final predictor is obtained by combining all classifiers. Boosting algorithm will be seriously affected by outliers in the late training period. These outliers are overemphasized, and even disturb the existing classification boundary, which makes the performance of the model worse.

Different from boosting algorithm, SPE adopts the distribution of classification hardness to resample the subset as the framework shown in Figure 4. The initial concept of ‘self-paced’ is to incrementally involve instances into learning, where easy instances are involved first and harder ones are then introduced gradually [44]. The majority class (health data) are undersampled into balanced bins by keeping the hardness contribution in each bin the same in the early stage. Thus, the undersampling is guided to select training samples that contribute the most hardness to the current iteration. Balancing factor α is added as the weight when new bins are updated. It grows along with the iteration of training and determines the decreasing level of importance for samples.

The pseudocode of SPE is described in Algorithm I. Given the training set of fault P (minority) and the training set of health N (majority), SPE firstly trains f_0 based on P and a randomly selected subset N_0 . The initial hardness is obtained

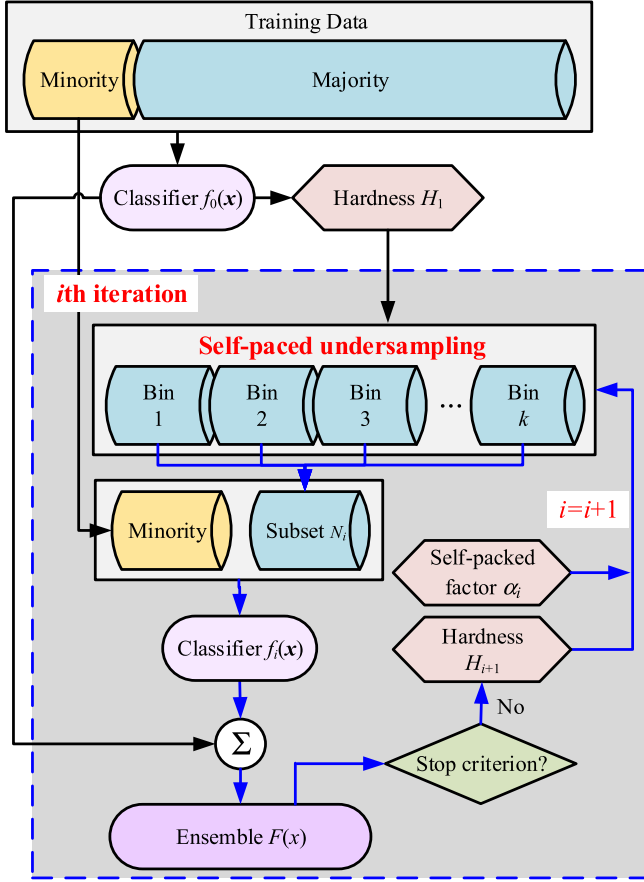


FIGURE 4 Flowchart of the self-paced ensemble. The majority is divided into bins based on hardness

according to the result of f_0 . Then, the algorithm enters the iteration. It divides the majority set N into k bins according to their hardness values. The l th bin can be obtained according to

Algorithm 1 Self-paced ensemble [43]

Input: A set of minority class examples P , a set of majority class examples N , $|P| < |N|$. Base classifier f , number of base classifiers n , number of bins k , and hardness function H .

Output: An ensemble classifier $F(x)$.

- 1) $i = 0$, train classifier f_0 using a random subset N_0 and P , where $|N_0| = |P|$.
- 2) **Repeat**
- 3) $i = i + 1$
- 4) Ensemble individual classifiers $F_i(x) = \frac{1}{i} \sum_{j=1}^{i-1} f_j(x)$
- 5) Divide majority set into k bins based on $H(x, y, F_i)$, obtaining B_1, B_2, \dots, B_k
- 6) Average hardness contribution in l th bin:

$$b_l = \sum_{s \in B_l} H(x_s, y_s, F_i) / |B_l|, l = 1, \dots, k$$
- 7) Update self-paced factor $\alpha = \tan(i\pi/2n)$
- 8) Unnormalized sampling weight of l th bin $w_l = \frac{1}{b_l + \alpha}$, $l = 1, \dots, k$
- 9) Undersample from l th bin with $\frac{w_l}{\sum_{j=1}^k w_j} |P|$ samples for all $l = 1 \rightarrow k$ and obtains N_i .

(Continued)

Algorithm 1 Self-paced ensemble [43]

10) Train f_i using P and a newly undersampled subset N_i .

11) **Until** $i = n$

12) **Output:** An ensemble classifier

$$F(x) = \frac{1}{n} \sum_{j=1}^n f_j(x).$$

$$B_l = \left\{ (x, y) \mid \frac{l-1}{k} \leq H(x, y, F) < \frac{l}{k} \right\} \quad H \in [0, 1] \quad (4)$$

After dividing the health data, undersampling is carried out to obtain a new subset N_i for training. The subset N_i is composed by samples that randomly select from each bin. The number of selected samples in l th bin is determined by a weight w_l which is obtain by hardness and balancing factor as

$$w_l = \frac{1}{b_l + \alpha} \quad (5)$$

Consequently, the classifier f_i for the i th iteration is trained on a balanced data set. In order to select samples that are most beneficial for the current ensemble, hardness value H and self-paced factor α are updated in each iteration. They are used to divide the majority set N in the next iteration. After n iterations, all classifiers are trained and their summation F is the final ensemble classifier. The classification is determined by F , equivalently the average of all classifiers f_i . In our work, decision tree is used for classifier f_i .

5 | EXPERIMENTAL STUDIES AND ANALYSIS

In this section, experiments using interpretation methods, state-of-art imbalanced algorithms and SPE are compared. The test data is randomly selected from the combined data set, which is composed of 320 health and 260 fault samples. In the following experiment, the test data remains the same.

5.1 | Evaluation metrics

Evaluation metrics play a crucial role in assessing the classification performance. In previous studies, precision (or accuracy) is the most commonly used metric for transformer diagnosis [11]. However, using only precision is not sufficient to evaluate for class imbalance problems since it is highly sensitive to the data distribution. Generally, it is difficult to evaluate the performance of imbalanced algorithms using a single metric. In this article, we choose minority Recall and G-mean to be the evaluation metrics due to they are insensitive to the imbalance and commonly used evaluation for class imbalance problems [41]. Minority Recall shows the performance of the minority class, but

does not reflect any performance on the other class. G-mean is an overall performance metric that reflects how well the performance is balanced between two classes. The performance of the classification algorithms can be determined by the combination of two metrics, and great classification performance is obtained when both of the two metrics are high.

Assuming the minority class to be positive class (P) and the majority class to be negative class (N), classified samples can be separated into four groups as denoted in the confusion matrix given in Table 3. A confusion matrix consists of information about actual and predicted classification returned by a classifier.

Recall is defined as the classification accuracy on correctly identifying positive class, and it can be obtained by

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

G-mean is defined as the geometric mean of recalls over both positive and negative classes. It is designed to measure the balanced accuracy between two classes. A low score for G-mean denotes a classifier that is highly biased toward one single class. It is defined as

$$\text{G-mean} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \quad (7)$$

5.2 | Tests on interpretation method

DGA is the standard method for diagnosing transformer faults based on the gases generated due to the dissolver of the transformer insulating oil. IEC 60,599 [12] provides a list of DGA methods. Five typical gases are used as the diagnosing criterion for transformers, namely hydrogen (H_2), methane (CH_4), ethane (C_2H_6), ethylene (C_2H_4) and acetylene (C_2H_2). Based on the concentration of these gases and the ratio between them, three famous DGA methods are established, namely, Rogers Ratios method [13, 45], Duval Triangular method (DTM) [14], and IEC Ratios method [12]. These three interpretation methods are used to identify the status of the transformer on our test data. As can be seen from Table 4, these methods perform poorly on health samples. The recall is as low as 40%, and G-mean is around 60%. Since the region in DTM is designated for faulty status and no region is assigned to the healthy status, DTM cannot be used to identify whether the transformer is healthy or not, resulting in a G-mean of NaN (the number of health samples is zero). Due to their insufficiency, interpretation methods are not suitable for power transformer diagnosis when the healthy condition is considered.

TABLE 3 Confusion matrix for imbalanced problems

	Positive prediction	Negative prediction
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

It should be noted that the primary task for online transformer diagnosis is to identify if the transformer is healthy or faulty, which results in binary classification. After the fault is found, interpretation methods or AI algorithms are followed to further identify the type of fault. This work focuses on the primary task that only binary classification is considered.

5.3 | Tests on AI imbalanced algorithms

To test various AI algorithms, state-of-art imbalanced algorithms, as well as SPE, are evaluated for comparison. We compare seven popular oversampling algorithms, namely, SMOTE [46], Borderline SMOTE [47], Safe-level SMOTE [48], ADASYN [32], MWMOTE [33], CGMOS [49] and MAHAKIL [50]. A subsequent classifier is followed after the resampling. In order to compare various types of classifiers, support vector machines (SVM) [51], K-Nearest neighbours (KNN) and decision trees (DT) are used. For comparison, cost sensitive (CS) strategy is integrated with classifiers SVM, KNN and DT with the cost matrix of the inverse of the imbalance ratio are evaluated. These classifiers are implemented in Matlab 2019b and have been well optimized to deal imbalance by Matlab Inc. Meanwhile, we also evaluate seven ensemble algorithms as Bagging, SMOTEBagging [52], AdaBoost [37], SMOTEBoost [39], RUSBoost [40], EasyEnsemble [53] and BalanceCascade [53]. The base classifier of these ensemble algorithms is DT and the number of classifiers in the ensemble is 100.

To investigate the influence of health samples on diagnosis performance, the number of health samples (N_h) for training is increased gradually from 600 to 9000, where 600 corresponds to the balanced situation. The Recall and G-mean of these algorithms are listed in Table 5 for individual classifiers, Table 6

TABLE 4 Performance of interpretation methods (Unit: %)

Method	Rec	Gm
Rogers Ratios method	41.2	64.2
IEC	41.1	60.6
Duval Triangular method	40.2	NaN

TABLE 5 Binary classification performance of cost sensitive classifiers (Unit: %)

Method	Support vector machines		K-Nearest neighbours		Decision trees	
	Rec	Gm	Rec	Gm	Rec	Gm
600	58.4	73.6	62.2	75.7	62.9	78.1
1200	60.6	74.6	64.3	76.9	64.6	79.1
1800	62.7	77.4	66.7	79.4	70.1	81.7
2400	64.5	78.4	72.2	84.3	74.8	85.0
6000	80.4	86.8	84.5	89.2	78.1	86.7
9000	88.3	90.2	88.8	90.9	89.7	92.1

TABLE 6 Binary classification performance of resampling algorithms (Unit: %)

Method		SMOTE		BorSMOTE		SafeSMOTE		ADASYN		MWMOTE		CGMOS		MAHAKIL	
Metric		Rec	Gm	Rec	Gm	Rec	Gm	Rec	Gm	Rec	Gm	Rec	Gm	Rec	Gm
600 balance	SVM	58.0	73.2	64.4	77.5	58.1	73.2	59.3	74.3	58.1	73.3	58.8	73.7	57.9	72.8
	KNN	62.2	76.9	67.9	78.2	62.2	76.8	67.2	78.0	62.7	77.1	65.1	78.6	63.5	77.7
	DT	62.8	78.1	65.4	79.5	62.7	78.1	67.9	79.1	62.7	77.7	65.1	78.4	64.7	79.8
1200	SVM	57.5	73.4	57.7	73.5	57.6	73.6	56.7	73.1	57.0	73.1	57.5	73.6	57.6	72.5
	KNN	61.1	76.9	60.7	76.2	60.7	76.3	59.9	76.2	60.9	76.3	59.6	76.2	57.7	74.0
	DT	62.1	76.0	62.7	77.3	62.3	78.1	63.2	78.8	63.4	78.6	62.8	78.9	59	75.6
1800	SVM	57.4	73.5	57.4	73.5	57.5	73.5	57.3	73.5	57.8	73.2	57.5	73.6	57.8	72.8
	KNN	64.3	78.5	64.2	78.4	64.4	78.8	64.5	79.2	66.9	79.6	63.8	78.5	59.3	74.8
	DT	67.1	80.6	68.1	81.1	69.8	82.3	60.4	76.0	70.9	82.2	72.5	83.7	60.5	76.8
2400	SVM	57.8	73.5	57.4	73.2	58.0	73.6	62.4	77.1	68.0	79.4	59.5	74.7	58.0	72.9
	KNN	69.5	82.2	70.9	82.8	69.4	81.8	69.3	82.3	72.3	83.3	69.3	82.1	63.0	77.4
	DT	70.7	82.9	69.7	82.3	75.0	85.2	69.4	82.0	72.2	88.5	70.8	82.7	65.5	79.8
6000	SVM	57.9	73.8	62.5	77.0	58.4	74.2	62.1	77.0	70.9	81.9	63.6	77.5	59.2	73.4
	KNN	75.4	85.9	76.7	86.2	75.3	85.5	74.0	85.1	75.6	85.4	75.6	85.7	66.5	79.5
	DT	73.3	84.0	74.2	84.5	76.1	85.5	74.1	84.9	76.1	84.6	77.2	86.5	65.6	79.2
9000	SVM	69.7	81.5	74.9	84.8	72	83.1	69.2	81.6	83.8	88.9	69.8	81.8	60.1	74.0
	KNN	78.9	87.9	80.8	88.5	79.8	88.1	75.7	86.0	86.5	90.6	78.5	87.1	71.6	82.2
	DT	78.4	86.6	82.5	88.6	80.1	87.5	78.0	86.5	88.6	91.8	83.9	89.6	75.4	85.1

Abbreviations: DT, decision trees; KNN, K-Nearest neighbours; SVM, Support vector machines.

for resampling algorithms, and Table 7 for ensemble algorithms. The value in bold is the best in the row. The comparison result is summarized as follows.

5.3.1 | Balance is not a panacea

By comparing the Balance ($N_N = 600$) and data set with different numbers of health samples (N_N), it shows that a balanced data set cannot guarantee an accurate classifier. The best classification for resampling and ensemble algorithms appears at $N_{NF} = 9000$, where the imbalance ratio is the largest. Therefore, a balanced data set is not necessary for transformer diagnosis. As indicated in Refs. [42, 43], imbalance is not the main source of classification difficulties. It will amplify the difficulties caused by the original data.

5.3.2 | More health data higher accuracy?

Previous studies about DGA mainly focus on fault classification. The health samples used in the previous research are limited and the influence of their appearance is seldom investigated. As the results in Tables 5–7 shown, the performance of classification algorithms increases with the growth of sample numbers, and reach the best performance

at $N_{NF} = 9000$. Although no additional fault data is added into the datasheet, the diagnosis accuracy is still improved. This is because more data means a wider coverage of the information. Therefore, the diagnosis efficiency of the transformer can be improved by adding exhaustive health samples that contribute more information to the problem.

5.3.3 | Inefficient resampling algorithms

The oversampling can be regarded as a preprocessing technique and classifiers are followed for prediction. By comparing Table 5 (CS + classifiers) and Table 6 (oversampling + classifiers), it reveals that cost sensitive is more effective than oversampling algorithms when dealing with DGA problem. This is because oversampling generates new samples based on the distance among existing samples. However, it is hard to define distance on the DGA data set, especially when data fusion is adopted. Due to the serious overlapping and the small quantity, the minority lacks a good representation or clear distribution structure. In this case, oversampling methods (e.g. SMOTE and its variants) that directly rely on analysing the neighbour relationship among a few samples usually fail, or even be counterproductive. The newly generated samples may be noisy to the original data set.

TABLE 7 Binary classification performance of ensemble algorithms and SPE (Unit: %)

Method	Bagging		AdaBoost		SMOBagging		SMOBoost		RUSBoost		EasyEns		BaCascade		SPE	
	Rec	Gm	Rec	Gm	Rec	Gm	Rec	Gm	Rec	Gm	Rec	Gm	Rec	Gm	Rec	Gm
600	59.0	75.6	63.3	78.9	53.7	77.3	75.4	88.3	56.0	81.8	63.8	82.5	-	-	-	-
1200	65.1	80.4	69.7	81.8	55.6	77.2	58.8	81.3	54.7	79.4	58.6	80.7	60.7	82.1	67.0	80.9
1800	68.7	81.4	70.2	81.4	56.2	82.6	52.2	80.2	54.9	80.4	55.2	86.2	64.7	84.5	75.2	85.2
2400	76.3	87.5	76.2	85.9	64.8	85.1	67.0	86.8	56.0	83.3	55.8	87.3	72.0	88.8	81.1	88.5
6000	84.7	90.3	83.5	89.6	79.0	86.7	78.6	89.7	53.1	80.3	54.3	87.7	70.0	91.0	88.6	90.2
9000	91.3	94.0	93.0	92.3	82.3	89.0	81.8	91.0	53.0	85.1	55.1	90.0	78.5	94.1	95.1	95.0

The performance of these oversampling algorithms can be improved by integrating the data information. MWMOTE and CGMOS show the best classification performance compared with other oversampling algorithms, because they identify the informative minority class samples before the generation of samples. They can effectively avoid the effect of noises during the data generation process. MAHAKIL obtains the worst result because it generates new samples based on evolutionary theory. New samples are generated simply by randomly among them and their bins. The algorithm provides more diversity compared with others. However, it also has a higher risk of generating noise.

5.3.4 | Ensemble algorithms are better?

As results in Table 7 show, Bagging and AdaBoost are superior to oversampling algorithms and other ensemble algorithms, except SPE. Bagging trains a model on resampled subsets, and takes the average. It cannot significantly reduce the bias, but can significantly reduce the variance. On the contrary, Boosting is to minimize the loss function sequentially, and its bias will gradually decrease. They all show a good performance on the data set in this work.

SMOTEBagging and SMOTEBoost are hybrid algorithms that integrating oversampling and ensemble. SMOTEBagging and SMOTEBoost perform worse than Bagging and AdaBoost, it indicates that SMOTE brings a negative effect on the ensemble classifiers, which is consistent with the previous discussion.

RUSBoost shows the worst performance of all. RUSBoost is an ensembles classifier integrating random undersampling and boosting. Undersampling will lose lots of information and shows a significant negative effect on the classification. Consequently, undersampling should be avoided in transformer diagnosis. EasyEnsemble is another type of undersampling based ensemble. It also shows a bad performance due to it will underfit the minority.

BalanceCascade iteratively discards majority samples that were well-classified by the current classifier. It may result in overfitting hard samples in late iterations and finally deteriorate the ensemble.

5.3.5 | SPE performs well

By comparing with other algorithms, SPE shows the highest recall and G-mean in most cases. It proves that SPE provides the best performance with different N_N . SPE considers characteristics of the data into the training, and fills the gap between data sampling strategy and the classifiers' capacity. The issues in other algorithms as discussed above are all considered in by classification hardness. It reaches the best at the $N_{NF} = 9000$, where has the most data and the highest imbalance ratio. It demonstrates the capacity of SPE in dealing with imbalance and data fusion.

TABLE 8 The comparison of computation times for all algorithms

Method	SVM	KNN	DT					
Time (s)	0.9	0.2	0.1					
Method	SMOTE	BorSMOTE	SafeSMOTE	ADASYN	MWMOTE	CGMOS	MAHAKIL	
Time(s)	3.6	3.9	4.3	0.5	263.5	8.8	0.1	
Method	Bagging	AdaBoost	SMOBagging	SMOBoost	RUSBoost	EasyEns	BaCascade	SPE
Time (s)	3.5	2.9	4.5	7.1	1.2	1.1	6.7	4.01

Abbreviations: DT, decision trees; KNN, K-Nearest neighbours; SVM, Support vector machines.

5.3.6 | Efficiency comparison

To demonstrate the efficiency, the computation times for all algorithms are evaluated when the number of health samples is 9000. All these algorithms are run on a workstation with Intel CPU E5-2699 2.30 GHz and 32 GB memory. The average computation times for five runs are listed in Table 8. SPE ranks 12 out of 18 algorithms, and 5 out of 8 ensemble methods. SPE provides the best classification performance without losing much efficiency and is the best choice in this work. It also reveals that oversampling will unavoidably increase the computation burden compared with individual classifiers. MWMOTE performs best among oversamplers, however, consumes too much computation time. Under-sampling algorithms such as RUSBoost and EasyEnsemble shows a fast training but a low accuracy.

6 | CONCLUSION

This article presented a comprehensive investigation of the DGA with an imbalanced data set from the view of both data set and methodology. To overcome the unsatisfactory diagnosis performance, large numbers of health data are collected to improve the classification. Since the imbalance and data fusion are introduced, SPE is employed, where classification hardness is proposed to consider the data characteristic in the classification.

The performance of traditional interpretation methods and AI based methods including different classifiers and their interaction with data are thoroughly investigated. The experiment reveals that:

1. The diagnosis efficiency of the transformer can be improved by adding exhaustive health samples with using suitable imbalanced algorithms simultaneously. The quality of the data is the fundamental reason for the diagnosis performance. More data means a wider coverage of the information.
2. Cost sensitive is more effective than oversampling algorithms when dealing with data fusion problem because oversampling may generate wrong samples. These wrong samples may mislead classifiers and result in a negative effect on the diagnosis.
3. Though widely used to handle imbalance problems, undersampling algorithms have a negative effect on the transformer diagnosis. Therefore, undersampling based algorithms should be avoided when dealing with transformer diagnosis.
4. Ensemble algorithms (Bagging, Boosting, etc.) show a good performance on the transformer diagnosis especially SPE. SPE shows the best performance among all because it takes account the distribution of the data into the classification. This feature is important when data fusion is involved. Therefore, SPE is recommended and more data should be involved in the transformer diagnosis

ACKNOWLEDGEMENTS

The authors would like to thank State Grid Zhejiang Electric Power Co., Ltd., Research Institute for providing the valuable DGA data. This work is supported by Science and Technology Project of State Grid Corporation of China (No. 5500-202019090A-0-0-00).

ORCID

Yang Zhang  <https://orcid.org/0000-0003-0156-8952>
Hong Cai Chen  <https://orcid.org/0000-0001-5418-4122>
Yaping Du  <https://orcid.org/0000-0003-0649-4649>
Xin Yao  <https://orcid.org/0000-0001-8837-4442>

REFERENCES

1. Wang, M., Vandermaar, A.J., Srivastava, A.J.: Review of condition assessment of power transformers in service. *IEEE Electr. Insul. Mag.* 18(6), 12–25 (2002)
2. Aj, C., Salam, M.A., Rahman, Q.M., Wen, F., Ang, S.P., Voon, W.: Causes of transformer failures and diagnostic methods—a review. *Renew. Sustain. Energy Rev.* 82, 1442–1456 (2018)
3. Li S., Li J. Condition monitoring and diagnosis of power equipment: Review and Prospective. *High Voltage* 2(2), 82–91 (2017)
4. Duan, L., Hu, J., Zhao, G., Chen, K., Wang, S.X., He, J.: Method of inter-turn fault detection for next-generation smart transformers based on deep learning algorithm. *High Volt.* 4(4), 282–291 (2019)
5. Cheng, Y., Bi, J., Chang, W., Xu, Y., Pan, X., Ma, X.: Proposed methodology for online frequency response analysis based on magnetic coupling to detect winding deformations in transformers. *High Volt.* 5(3), 343–349 (2020)
6. Dutta, S., Mishra, D., Baral, A., Chakravorti, S.: Estimation of de-trapped charge for diagnosis of transformer insulation using short-duration polarisation current employing detrended fluctuation analysis. *High Volt.* 5(5), 636–641 (2020)

7. Yang, D., Chen, W., Shi, H., Wan, F., Zhou, Y.: Raman spectrum feature extraction and diagnosis of oil-paper insulation ageing based on kernel principal component analysis. *High Volt.* (2020)
8. Bakar, N., Abu-Siada, A., Islam, S.: A review of dissolved gas analysis measurement and interpretation techniques. *IEEE Electr. Insul. Mag.* 30(3), 39–49 (2014)
9. Singh, S., Bandyopadhyay, M.N.: Dissolved gas analysis technique for incipient fault diagnosis in power transformers: a bibliographic survey. *IEEE Electr. Insul. Mag.* 26(6), 41–46 (2010)
10. Ghoneim, S.S.M., Taha, I.B.M.: A new approach of DGA interpretation technique for transformer fault diagnosis. *Int. J. Electr. Power Energy Syst.* 81, 265–274 (2016)
11. Mirowski, P., LeCun, Y.: Statistical machine learning and dissolved gas analysis: a review. *IEEE Trans. Power Deliv.* 27(4), 1791–1799 (2012)
12. IEC 60599: mineral oil-filled electrical equipment in service – guidance on the interpretation of dissolved and free gases analysis, pp. 1–78. IEC Standards (2015)
13. IEEE guide for the interpretation of gases generated in mineral oil-Immersed transformers, pp. 1–96. IEEE Std C57.104-2008 (Revision of IEEE Std C57.104-1991) (2019)
14. Duval, M., dePablo, A.: Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases. *IEEE Electr. Insul. Mag.* 17(2), 31–41 (2001)
15. Dornenburg, E., Strittmatter, W.: Monitoring oil-cooled transformers by gas-analysis. *Brown Boveri Rev.* 61(5), 238–247 (1974)
16. Rogers, R.: IEEE and IEC codes to interpret incipient faults in transformers, using gas in oil analysis. *IEEE Trans. Electr. Insul.* 5, 349–354 (1978)
17. Hao, X., Cai-xin, S.: Artificial immune network classification algorithm for fault diagnosis of power transformer. *IEEE Trans. Power Deliv.* 22(2), 930–935 (2007)
18. Guardado, J., Naredo, L., Moreno, P., Fuente, C.: A comparative study of neural network efficiency in power transformers diagnosis using dissolved gas analysis. *IEEE Power Eng. Rev.* 21(7), 71 (2001)
19. Abu-Siada, A., Hmood, S.: A new fuzzy logic approach to identify power transformer criticality using dissolved gas-in-oil analysis. *Int. J. Electr. Power Energy Syst.* 67, 401–408 (2015)
20. Al-Janabi, S., Rawat, S., Patel, A., Al-Shourbaji, I.: Design and evaluation of a hybrid system for detection and prediction of faults in electrical transformers. *Int. J. Electr. Power Energy Syst.* 67, 324–335 (2015)
21. Bacha, K., Souahlia, S., Gossa, M.: Power transformer fault diagnosis based on dissolved gas analysis by support vector machine. *Electr. Power Syst. Res.* 83(1), 73–79 (2012)
22. da Silva, A.C.M., Garcez Castro, A.R., Miranda, V.: Transformer failure diagnosis by means of fuzzy rules extracted from Kohonen self-organizing map. *Int. J. Electr. Power Energy Syst.* 43(1), 1034–1042 (2012)
23. Weigen, C., Chong, P., Yuxin, Y., Yilu, L.: Wavelet networks in power transformers diagnosis using dissolved gas analysis. *IEEE Trans. Power Deliv.* 24(1), 187–194 (2009)
24. Dai, J., Song, H., Sheng, G., Jiang, X.: Dissolved gas analysis of insulating oil for power transformer fault diagnosis with deep belief network. *IEEE Trans. Dielectr. Electr. Insul.* 24(5), 2828–2835 (2017)
25. Islam, M.M., Lee, G., Hettiwatte, S.N.: Application of Parzen Window estimation for incipient fault diagnosis in power transformers. *High Volt.* 3(4), 303–309 (2018)
26. Ma, H., Ekanayake, C., Saha, T.K.: Power transformer fault diagnosis under measurement originated uncertainties. *IEEE Trans. Dielectr. Electr. Insul.* 19(6), 1982–1990 (2012)
27. Cui, Y., Ma, H., Tapan, S.: Improvement of power transformer insulation diagnosis using oil characteristics data preprocessed by SMOTEBoost technique. *IEEE Trans. Dielectr. Electr. Insul.* 21(5), 2363–2373 (2014)
28. Senoussaoui, M.E.A., Brahmi, M., Fofana, I.: Combining and comparing various machine-learning algorithms to improve dissolved gas analysis interpretation. *IET Gener. Transm. Distrib.* 12(15), 3673–3679 (2018)
29. Prati, R.C., Batista, G.E.A.P.A., Monard, M.C.: Learning with class skews and small disjuncts. In: Bazzan, A. L. C., Labidi, S., *Advances in artificial intelligence—SBIA 2004* 3171, pp. 296–306. Springer Berlin Heidelberg, Berlin, Heidelberg (2004). https://link.springer.com/chapter/10.1007/978-3-540-28645-5_30
30. He, H., Garcia, E.A.: Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 21(9), 1263–1284 (2009)
31. He, H., Ma, Y.: *Imbalanced learning: Foundations, algorithms, and applications*. Wiley-IEEE Press, Hoboken, NJ (2013)
32. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), IEEE, pp. 1322–1328
33. Barua, S., Islam, M.M., Yao, X., Murase, K.: MWMOTE--Majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Trans. Knowl. Data Eng.* 26(2), 405–425 (2014)
34. Akbani, R., Kwek, S., Japkowicz, N.: Applying support vector machines to imbalanced datasets. In *European conference on machine learning*, 2004: Springer, pp. 39–50
35. Sun, Y., Kamel, M.S., Wong, A.K.C., Wang, Y.: Cost-sensitive boosting for classification of imbalanced data. *Pattern Recogn.* 40(12), 3358–3378 (2007)
36. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* 42(4), 463–484 (2012)
37. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.* 37(3), 297–336 (1999)
38. Liu, Y., Yao, X.: Ensemble learning via negative correlation. *Neural Netw.* 12(10), 1399–1404 (1999)
39. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: SMOTEBoost: improving prediction of the minority class in boosting. Paper presented at the European conference on principles of data mining and knowledge discovery, 2003
40. Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: RUS-Boost: a hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man Cybern. Syst. Hum.* 40(1), 185–197 (2010)
41. Wang, S., Minku, L.L., Yao, X.: Resampling-based ensemble methods for online class imbalance learning. *IEEE Trans. Knowl. Data Eng.* 27(5), 1356–1368 (2015)
42. Napierala, K., Stefanowski, J.: Types of minority class examples and their influence on learning classifiers from imbalanced data. *J. Intell. Inf. Syst.* 46(3), 563–597 (2015)
43. Liu, Z. et al.: Self-paced ensemble for highly imbalanced massive data classification. Paper presented at the 36th IEEE International Conference on Data Engineering, Dallas (2020)
44. Kumar, M.P., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: Lafferty, J., Williams, C., Shawe-Taylor, J., Culotta, A. (eds.) *Advances in neural information processing systems*, vol.3, pp. 1189–1197. Curran Associates, Inc. (2010)
45. Taha, I.B.M., Ghoneim, S.S.M., Duaywah, A.S.A.: Refining DGA methods of IEC Code and Rogers four ratios for transformer fault diagnosis. Paper presented at the 2016 IEEE Power and Energy Society General Meeting (PESGM), 2016
46. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357 (2002)
47. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In: *International Conference on Intelligent Computing*, 2005: Springer, pp. 878–887
48. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Safe-level-smote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2009: Springer, pp. 475–482
49. Zhang, X., Ma, D., Gan, L., Jiang, S., Agam, G.: CGMOS: certainty guided minority OverSampling. Paper presented at the Proceedings of the 25th ACM International on Conference on Information and Knowledge Management—CIKM '16, 2016
50. Bennin, K.E., Keung, J., Phannachitta, P., Monden, A., Mensah, S.: MAHAKIL: diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction. *IEEE Trans. Software Eng.* 44(6), 534–550 (2018)

51. Lin, C.F., Wang, S.D.: Fuzzy support vector machines. *IEEE Trans. Neural Netw.* 13(2), 464–471 (2002)
52. Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. In: 2009 IEEE Symposium on Computational Intelligence and Data Mining, 30 March–2 April 2009, pp. 324–331
53. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. B Cybern.* 39(2), 539–550 (2009)

How to cite this article: Zhang Y, Chen HC, Du Y, et al. Power transformer fault diagnosis considering data imbalance and data set fusion. *High Voltage*. 2021;6:543–554. <https://doi.org/10.1049/hve2.12059>