Improved information entropy weighted vague support vector machine method for transformer fault diagnosis

Hongru Zhang © | Jiaxiang Sun © | Kaining Hou | Qingquan Li | Hongshun Liu

Abstract
A combined model based on improved information entropy and vague support vector machine (IVSVM) is introduced into transformer fault diagnosis using dissolved gas analysis in oil (DGA). The improved information entropy method is used to obtain the weights of each gas and to weight the raw data, and the processed training data and the corresponding fault types are inputted into the vague support vector machine (VSVM) model to obtain classifiers. Firstly, the training data are weighted by the improved information entropy method to discretise the original data from the mixed state for subsequent classifier training. Then, the vague set divides the events into true, false and unknown factors, which can optimise the sub-interface of SVM and improve the accuracy of the boundary point classification. Finally, fault data from the literature and actual collections are selected for training and testing. By comparing with the widely used ratio method and artificial intelligence method, it can be concluded that the method described herein can effectively improve the accuracy of fault diagnosis. The result shows that this method has better applicability when facing actual fault type classification with higher data similarity.

1 | INTRODUCTION

Power transformers are very important power equipment and their stable operation is of great importance to the power supply. During operation, a transformer fault can occur due to environmental factors, human factors and other reasons. A small amount of gas is dissolved in the transformer insulating oil, and the composition of the gas in the transformer insulating oil changes more significantly when a fault occurs. The concentrations and relative proportions of by-product gases are closely related to the transformer insulation conditions [1].

Dissolved gas analysis DGA of oil can cover most transformer failure scenarios. The gases used for this analysis are hydrogen, methane, ethane, ethylene, acetylene, carbon monoxide, and carbon dioxide [2–5]. CO and CO₂ are produced due to the decomposition of cellulose in solid insulation and usually vary significantly in concentration at high temperatures and overheating, but not in other fault conditions. Therefore, conventional ratio methods largely ignore the use of CO and CO₂ [3]. The content and proportion of these gases vary depending on the fault conditions. For example, partial discharges in discharge failures will typically manifest as a surge in H₂ content. Low-energy discharges and high-energy discharges are accompanied by C₂H₂ production. The percentage of C₂H₄ and CH₄ is different for different degrees of overheating faults. Some scholars have proposed improved ratio methods to address the problems with the ratio method. For example, ratio methods involving more parameters have been proposed [6, 7], which divide the ratio conditions more carefully and can effectively improve its accuracy. Others have proposed new display methods, such as the pentagonal method [8, 9]. In addition, machine learning methods have been applied to fault diagnosis and achieved good results. For example, neural networks [10, 11], fuzzy set theory [12], Bayesian networks [13], set pair theory [14], support vector machines [15, 16], Parzen window [17] and other methods [18].

Herein, an improved vague support vector machine method (IVSVM) using the information entropy method is used. The information entropy method is used to process different classes of data to make them easier to distinguish, and then the VSVM method is used to classify transformer fault types by taking advantage of the higher accuracy of data classification at the boundary. Five gases, hydrogen, methane, ethane, ethylene, and acetylene, are chosen as the characteristic quantities for the study.

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.

© 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.
Open source fault data containing six fault types and actual collected operational data containing nine fault types are used for the training and testing data.

SVM as a shallow classifier has the advantage of high classification accuracy and fast classification speed. In addition, the number of selected features is moderate, all of them are gases and they have a high correlation with each other (all of them are decomposition products in transformer oil). The use of deep learning methods, such as deep neural networks, can lead to large time and space complexities, which can affect the effectiveness of learning classification. It is known from the analysis that the method described herein has better performances than traditional ratio and machine learning methods in fault classification. It also has better classification performance when facing poor data quality, and can effectively classify faults with similar data when the fault type is subdivided.

Section 2 introduces the method and flowchart presentation of the improved algorithm; Section 3 is a combination of data analysis results, selected open source data collected from the literature, and actual transformer operation data collected from a place in southwest China for algorithm performance proof; and Section 4 gives the conclusion.

2 | ALGORITHM AND IMPROVEMENTS

2.1 | Vague set

Vague sets theory [19] is a complement to fuzzy sets, which describes the object of study in terms of both true and false. Vague sets can better describe uncertainty and compensate for the lack of a single subordination function in fuzzy sets.

Definition:

Let \( U \) be a space of points, with a generic element of \( U \) denoted by \( x \).

A vague set \( V \) in \( U \) is characterised by a truth membership function \( t_v \) and a false membership function \( f_v \). \( t_v \) is a lower bound on the negation of \( x \) from the evidence against \( x \). \( t_v(x) \) and \( f_v(x) \) both associate a real number in the interval \([0,1]\) with each point in \( U \) and \( t_v(x) + f_v(x) \leq 1 \). Define \( m_v \) as the unknown part, where \( m_v(x) = 1 - t_v(x) - f_v(x) \).

The above method can constrain \( x_i \) to take values in subinterval \([t_v(x_i), 1 - f_v(x_i)]\) of \([0,1]\). Its vague value consists of three parts: truth membership function \( t_v(x) \), false membership function \( f_v(x) \), and unknown part \( m_v(x) \). The method of getting \( t, m, f \) of \( x \) is introduced in Subsection 2.2.

2.2 | Vague support vector machine

In this subsection, the VSVM model is first introduced and then the variable calculation method is introduced in detail.

The support vector machine [20, 21] is a sub-classification model that aims to find a hyperplane to partition the sample. The principle of segmentation is spacing maximisation, which ultimately translates into a convex quadratic planning problem to be solved. When the training sample is linearly divisible, a linearly divisible support vector machine is learnt by maximising the hard spacing; when the training sample is linearly indivisible, a non-linear support vector machine is learnt by maximising the kernel trick and the soft spacing. The model is as follows:

\[
\begin{align*}
\min & \left( \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \right) \\
\text{subject to} & \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \in \{1, -1 \} \\
& \quad \xi_i \geq 0, i = 1, 2, \ldots, n
\end{align*}
\]  

where \( w \) is the hyperplane normal vector, \( C \) is the penalty factor, \( n \) is the number of samples, \( \xi \) is the relaxation factor that represents the allowable error score in the linear indistinguishable case, \( y \) is the sample output, \( x \) is the sample input, and \( b \) is the threshold. Equation (2) is the constraint, and the Lagrange multiplier algorithm is introduced to solve the above problem to obtain the optimisation objective function.

The radial basis function is chosen as the kernel function of the support vector machine. The equation is as follows.

\[
K(x_i, x_j) = \exp(-g\|x_i - x_j\|^2)
\]  

where \( g \) is radial basis kernel function parameter, the value of which has a large impact on the prediction accuracy of the regression model. The larger \( g \) is, the stronger is the influence between the support vectors.

The SVM method relies heavily on penalty and relaxation factors for boundary points, and is prone to misclassify the points between \( y_i(w^T x_i + b) = 1 \) and the sub-interface. Herein, the VSVM method is proposed, which combines the vague set and SVM, to address this problem. The vague set can be used to measure the truth and false membership of things to improve the accuracy of the SVM classification of boundary points. The schematic diagram of the VSVM method is shown in Figure 1.

As shown in Figure 1, two types of faults are defined as \( F_1 \) and \( F_2 \) for X and dot, and the remaining fault is defined as Other Faults. The point in \( F_1 \) marked with green represents a case. \( t \) is the probability that the point belongs to fault \( F_1 \), \( f \) is the probability that the point belongs to \( F_2 \), and \( m \) is the probability that the point belongs to Other Faults. The VSVM model is shown below:
where \( w \) is the hyperplane normal vector, \( \xi_i \) is relaxation factor, \( t \), \( f \), \( m \) are the true dependence, false dependence, and unknown part of the vague set. In Figure 1, we get three sub-interfaces for these three sub-interfaces are \( t \) and \( f \), \( f \) and \( m \) and \( m \) and \( Q+ \) and \( Q- \) are the correction factor of \( \xi \). The smaller \( Q \) is, the smaller the role played by \( \xi \) and hence the less important \( x \) is. The points outside the support vector have less influence on the boundaries and it is easier to obtain more a reasonable boundary.

Set the Lagrange multiplier \( \alpha, \beta \) and the problem is transformed as follows:

\[
\begin{aligned}
&\min_{\omega, b, \xi, \alpha, \beta} \frac{1}{2}||\omega||^2 + C \left( \sum_{i=1}^{N} Q_i^+ \xi_i + \sum_{i=1}^{N} Q_i^- \xi_i \right) \\
&Q_i^+, Q_i^- \in \{t, f, m\} \\
&y_i (\omega x_i + b) \geq 1 - \xi_i, y_i \in \{1, -1\} \\
&\xi_i \geq 0, i = 1, 2, \ldots, n
\end{aligned}
\]

The solution of these parameters can be achieved using the SMO method [22]. The final solution obtains the support vector \( \omega \).

\[
w = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i)
\]

The classification function is

\[
y(x_i) = \text{sign}(\omega^T \phi(x) + b) = \text{sign} \left( \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \right)
\]

The parameters \( t, f, m \) of the V SVM in Equation (6) are obtained in combination with the KNN algorithm. KNN is the k-NearestNeighbor method, which is a common clustering method. The acquisition steps are as follows:

1. Define \( F_1 \) as fault type \( i \), \( X_{ij} \) as the \( j \)th data of fault type \( i \) in the sample. Select two types of faults from all the fault types to build a support vector machine model. For example, choose two faults in all fault types as \( F_1 \) and \( F_2 \). The data corresponding to fault \( F_1 \) and fault \( F_2 \) are \( D_1 = \{X_{11}, j = 1, 2, \ldots, X_{1m} \in F_1\} \) and \( D_2 = \{X_{21}, j = 1, 2, \ldots, X_{2m} \in F_2\} \). The other data in the sample are \( D_m = \{X_{mn}, j = 1, 2, \ldots, X_{mn} \in F_m\} \), \( F_m \) includes all fault types except \( F_1 \) and \( F_2 \).

2. Select a point in \( D_1 \), such as \( X_{11} \), and use the KNN method to obtain the \( k \)-neighbouring points it contains. The value of \( k \) is usually less than the square root of the data volume and herein there are three classes when building one VSVM classifier (\( k \) is at least 1), so \( k \) is an integer and \( k \in [1, \sqrt{n}] \). Subsection 2.4 introduces the selection of \( k \) value. An example is given to illustrate this, as shown in Figure 2.

The point marked with green is \( X_{11} \) and its \( k \)-neighbouring points are circled with a dotted line. The number of points belonging to fault \( F_1 \), fault \( F_2 \) and other faults are \( n_1, n_2 \) and

\[
\begin{aligned}
&\min_{\omega, b, \xi, \alpha, \beta} \frac{1}{2}||\omega||^2 + C \left( \sum_{i=1}^{N} Q_i^+ \xi_i + \sum_{i=1}^{N} Q_i^- \xi_i \right) \\
&Q_i^+, Q_i^- \in \{t, f, m\} \\
&y_i (\omega x_i + b) \geq 1 - \xi_i, y_i \in \{1, -1\} \\
&\xi_i \geq 0, i = 1, 2, \ldots, n
\end{aligned}
\]
In this example, set $k$ as 6, so $n_1 = 4$, $n_2 = 1$ and $n_m = 1$. The point $X_{12}$ parameters $t_{12}, f_{12}, m_{12}$ are as follows:

$$t_{12} = \frac{n_1}{k}$$

$$f_{12} = \frac{n_2}{k}$$

$$m_{12} = \frac{n_m}{k}$$

Therefore $t_{12} = 0.666, f_{12} = 0.167, m_{12} = 0.167$.

Besides, $k$ is a variable, so the value of the volume of changes when $k$ changes. That makes $m$ change from 0 to the maximum of $m$, and $t$ changes from 1 to the minimum of $t$. Therefore, this method of getting $t, f, m$ satisfies the Definition in Subsection 2.1.

(3) Get the $t, f, m$ parameters of each data item by the above method and build $\frac{n(n-1)}{2} \times 3$ VSVM classifiers by using Equations (4)–(10), $n$ is the number of the fault type.

### 2.3 Obtaining data weights by improved information entropy method

Before the data can be used, they need to be processed for more accurate results. The gases are normalised in the following way:

$$H_2: \frac{C(H_2)}{\sum C(All \ the \ 5 \ gases)}$$

$$Others: \frac{C(This \ gas)}{C(C_2H_6) + C(C_2H_4) + C(C_2H_2) + C(CH_4)}$$

It can be seen from the data shown later that the volume share of $H_2$ among the five gases may be high. When normalising the four gases other than hydrogen using the method of Equation (14), the addition of hydrogen may make the volume share of other gases too small and lead to missing information. When classifying actual faults with the same fault type, the normalisation method of the four gases in Equation (14) can highlight the differences.

The training data for fault type $F$ are $X^{(F)}$, $X^{(F)}$ is a matrix of $p$ rows and $q$ columns, $p$ is the number of the cases, $q$ is the number of features, and $q$ is the gas type used here, that is, five types. $X^{(F)}$ is shown below.

$$X^{(F)} = \begin{bmatrix}
    x_{11}^{(F)} & x_{12}^{(F)} & \cdots & x_{1i}^{(F)} & \cdots & x_{1q}^{(F)} \\
    x_{21}^{(F)} & x_{22}^{(F)} & \cdots & x_{2i}^{(F)} & \cdots & x_{2q}^{(F)} \\
    \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
    x_{i1}^{(F)} & x_{i2}^{(F)} & \cdots & x_{ij}^{(F)} & \cdots & x_{iq}^{(F)} \\
    \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
    x_{p1}^{(F)} & x_{p2}^{(F)} & \cdots & x_{pj}^{(F)} & \cdots & x_{pq}^{(F)}
\end{bmatrix}$$

Entropy is a physical concept in thermodynamics and is a measure of how chaotic a system is. The higher the entropy, the more chaotic the system is, and the less information it carries. The lower the entropy, the more ordered the system is and the more information it carries. Information entropy [23] draws on the concept of entropy in thermodynamics to describe the average amount of information in a source. Defining $e_j^{(F)}$ as the information entropy when the fault is $F$, $e_j^{(F)}$ is calculated as in Equation (16).

$$e_j^{(F)} = - \sum_{i=1}^{p} x_{ij} \ln x_{ij}$$

The classification herein is based on the different percentages of various gas contents in different fault conditions. For example, the proportion of ethylene content in low-energy and high-energy discharges far exceeds that in thermal and partial discharge faults; the proportion of hydrogen content in partial discharge faults far exceeds that in other types of faults. Define $\omega_j^{(F)}$ as the weight of the $j$-th gas of fault type $F$, defined as in Equations (17)–(19):

$$\delta_j^{(F)} = \frac{1 - e_j}{\sum_j (1 - e_j)}$$

$$b_j^{(F)} = \frac{1}{\delta_j^{(F)}}$$

$$\omega_j^{(F)} = \frac{b_j^{(F)}}{\sum_j b_j^{(F)}}$$

Equation (17) is the coefficient of variation in the entropy method. The larger $g$, the more scattered the data distribution. Herein the weight of a gas is considered in a fault type. Therefore, the more concentrated the volume fraction of the gas is distributed, the more representative it is of that fault type, that is the greater the weight of the gas. In summary, the definition of $g$ here is opposite to that of the entropy value method, and therefore the inverse is taken here, as shown in Equation (18). The weighted vector $P^{(F)}$ of fault $F$ is defined as Equation (20):

$$P^{(F)} = \left[ \omega_1^{(F)} \omega_2^{(F)} \cdots \omega_q^{(F)} \right]^T$$

### 2.4 Algorithm flowchart

According to the description above, the flow chart of the IVSVM method is shown in Figure 3. In conjunction with
Figure 3, the steps of the IVSVM approach described here are as follows:

1. Normalising the gas content of the raw data. The normalisation method can be referred to Equation (14).
2. Weighting coefficients for each gas content in the training data for different fault types were calculated using the improved information entropy method. The training data were then weighted.
3. Setting the value of $k (k \in [1, \sqrt{n}])$ and using weighted training data to obtain $t, f$ and $m$ of each data item.
4. Building $\frac{n(n-1)}{2} \times 3$ VSVM classifiers, the number of transformer fault is $n$.
5. The testing data are weighted separately and then entered into the classifier for classification. The classification rules are as follows.

When using the category 1 and 2 classifiers, one vote is counted if the testing data are classified as fault 1 or fault 2, and half a vote is counted if the testing data are classified as other than faults 1 and 2. All classifiers are used and votes are counted, and the final classification is the one with the most votes.

6. Changing $k$ value and carry out steps (3)–(5) until the accuracy of the training data is the best possible.

3 | CASE STUDIES AND ANALYSIS

3.1 | Open source data analysis in international standards and literature

A total of 342 cases from the literature [24–30] were used here for testing. PD is partial discharge, LED is low energy discharge, HED is high energy discharge, T1 is low temperature overheat, T2 is medium temperature overheat, and T3 is temperature overheat. The types and numbers of fault data are shown in Table 1.

The IEC three-ratio method, Rogers method, Duval triangle method, SVM method and BPNN (BP neural network) method were used for comparison with the algorithms described here. The number of cases correctly predicted by each method for each fault type is shown in Figure 4 and the accuracy of each algorithm is shown in Figure 5.

Figure 4 shows the number of cases correctly predicted by each method for each fault type. The horizontal axis refers to the type of fault and the vertical axis refers to the fault types. The dark blue bar labelled “testing data” refers to all the data used for testing (Table 1, testing data). The other bars represent the number of cases diagnosed correctly by each method for each fault type. From Figure 4, the total number of testing data is 117 and the correct prediction number of the IEC
three-ratio method, Rogers method, Duval triangle method, SVM method, BPNN and IVSVM are 83, 91, 99, 100, 101 and 106.

From Figure 5, the accuracy of the IEC three-ratio method, Rogers method, Duval triangle method, SVM method, BPNN and IVSVM are 70.9%, 77.8%, 84.6%, 85.5% and 90.6%. When the data quality is low, that is, there are more boundary points, the accuracy of the IEC three-ratio method and Rogers method is lower. The Duval triangle method shows similar accuracy to the SVM method, because the classification boundaries are more detailed and more realistic. SVM, BPNN and IVSVM have higher accuracy rates than the traditional ratio methods. The main reason for this is that the classification of boundary points is more realistic and less likely to be misclassified.

Selecting part of the data with different fault types and using different algorithms for analysis, the diagnostic results are shown in Table 2. Data with an asterisk (*) are fault data with poor quality. The results are varied and do not match the actual faults when analysed by different methods. Data with two asterisks (**) are of average quality. This data are usually on the boundary between different types of faults and different methods may give different classification results but will be like the actual type of fault. The rest of the data in Table 2 are fault data with good quality, the results of the analysis of these data using different methods are basically the same and consistent with the actual failure.

It can be seen from Table 2 that the IEC three-ratio method and Rogers method have lower accuracy when the quantity quality is poor, and the Rogers method also has some data that cannot be judged because of the boundary conditions. The Duval triangle method has higher accuracy due to more detailed classification and more complex classification conditions. SVM, BPNN and IVSVM methods have higher accuracy than the traditional ratio methods and have better results in dealing with data that are prone to misclassification. IVSVM works better when the data quality is poor.

### 3.2 Analysis of actual operational data

Considering the practical applicability of the method, the algorithm shown here was used to train and test 630 of the collected fault data in southwest China. The data were collected from the data management system of China Southern Power Grid. These are real faults in actual operations and they were verified by experts and the inspection companies. For convenience, the fault codes A–I are used to indicate the fault conditions. The types and numbers of these faults are shown in Table 3.

Nine types of faults are included within the 630 cases. Each type of fault data has 70 cases. Fifty cases of each fault were used as training data and the remainder were used as testing data. A portion of the data is selected and optimised by weighting using the improved information entropy method. The data corresponding to various types of faults are shown in Table A1. The gas weighting factor for each fault is shown in Table 4.

The effect of information entropy weighting is shown for this selected part of the data. Since the data have five characteristic quantities, three of them, CH₄, C₂H₆ and C₂H₄, were selected for display as shown in Figures 6(a) (d). C₂H₆, C₂H₃, and H₂ were selected for display as shown in Figures 6(e) and (f).

It can be seen from Figure 6(a) that the data of the same kind as the raw data are scattered before processing and that there is mixing of different classes of data. Classes A, C, and F are mixed, class I has a point at the top that is very far from the other points, and class E also has a point that has this edge point situation. Therefore, when using the data in this case to classify or cluster directly, misclassification will occur. A different perspective can also reveal that these problems do exist, rather than the angle of observation, as shown in Figure 6(b).

Figure 6(c) shows the data after the weighting process. The data become concentrated and the boundaries between the different classes of data are clear. In Figure 6(a), the mixing of classes A, C and F is effectively improved, and the problem of classes I and E is also solved. There is also
a mixing of the observation classes G and I at this angle, but a different angle reveals that this is not the case, as shown in Figure 6(d).

After observing the weighted optimisation results of content CH\textsubscript{4}, C\textsubscript{2}H\textsubscript{6} and C\textsubscript{2}H\textsubscript{4}, C\textsubscript{2}H\textsubscript{4} was selected as the reference standard to analyse the weighted optimisation effect of the two remaining gases (C\textsubscript{2}H\textsubscript{2}, H\textsubscript{2}). It can be observed from Figure 6(e) that there is a mixture of classes C and F, and classes G and I. After optimisation, it can be observed that the data shown in Figure 6(f) are more compact within classes and the class spacing is obvious. The improved information entropy method used here has an obvious effect on the discretisation of data, which is of obvious help in the subsequent classification work.

Two data from each class from the testing data are selected for analysis, which contain the edge points of classes E and I as
TABLE 4 Gas weighting factor for each fault

| Fault type | CH₄ (ppm) | C₂H₆ (ppm) | C₂H₄ (ppm) | C₂H₂ (ppm) | H₂ (ppm) |
|------------|-----------|------------|------------|------------|----------|
| A          | 0.637072  | 0.010799   | 0.329875   | 0.002577   | 0.019677 |
| B          | 0.510076  | 0.124917   | 0.345981   | 0          | 0.019026 |
| C          | 0.28875   | 0.031886   | 0.659487   | 0.005209   | 0.014667 |
| D          | 0.161221  | 0.011171   | 0.000676   | 0          | 0.826931 |
| E          | 0.029385  | 0.004131   | 0.002949   | 0          | 0.963535 |
| F          | 0.460542  | 0.03074    | 0.502721   | 0.002239   | 0.003759 |
| G          | 0.182171  | 0.188593   | 0.305947   | 0.181804   | 0.141485 |
| H          | 0.222171  | 0.33631    | 0.280622   | 0.098305   | 0.062592 |
| I          | 0.257935  | 0.134717   | 0.161985   | 0.25889    | 0.186473 |

FIGURE 6 Improved entropy information method for data weighting
in Figure 6(a). The Duval triangle method, which has better results, as described above, is used for comparison with SVM, BPNN and IVSVM. The classification result is presented in Table 5.

From Table 5, Duval triangulation has a high accuracy in fault diagnosis when facing normal data, but there are more serious misclassification cases when facing edge points and data with low gas concentration. SVM has misclassification when facing edge points, and it is easy to be misclassified into similar classes when most of the data are similar. IVSVM is more accurate in classifying boundary points because of its restriction on boundary points.

Duval triangulation is ineffective when classifying multiple fault types, so SVM, BPNN and IVSVM were chosen for comparison. The learning rate of BPNN is 0.01 and the number of iterations is 500. The parameters of SVM, BPNN and IVSVM were optimised using the particle swarm algorithm. The maximum number of evolutions is 300, the maximum number of populations is 20, update rates of parameters C and g and BPNN's w and b are all set to 1.1. The best C and g for SVM are (37.68,105.78) and for IVSVM (14.69,73.59). Figure 7 shows the best accuracy changes and the average adaptation rates of three methods. Table 6 shows the accuracy of the testing data and training data for the three methods, SVM, BPNN and IVSVM.

Combining Figure 7 and Table 6, the training accuracy of SVM is 87.3% and the test accuracy is 84.4%; the training accuracy of BPNN is 89.1% and the test accuracy is 86.7%; and the training accuracy of IVSVM is 94.2% and the test accuracy is 91.7%. IVSVM has more improvement than SVM and BPNN due to the pre-processing of the data for better quality and the ability of VSVM to better classify the boundary points. From the diagnostic results of all three methods, the accuracy of the SVM and BPNN methods is similar. When facing the classification of different types of faults, sometimes SVM is better than BPNN, and sometimes BPNN is better than SVM, which proves that both methods have their advantages. IVSVM outperforms the other two methods in classifying all types of faults. The accuracies of both training and testing are higher than the other two methods for misclassified classes (e.g., fault type F), proving that IVSVM is more effective in the case of mixed data in multiple fault classifications.

4 CONCLUSION

Transformer fault monitoring is an important part of power system operation. An improved vague support vector machine method is proposed herein. The method uses an improved information entropy method to weight the dissolved gases in transformer oil and then builds multiple vague support vector machine classifiers for transformer fault diagnosis. The important findings are as follows:

| No. | H₂ (ppm) | CH₄ (ppm) | C₂H₆ (ppm) | C₂H₄ (ppm) | C₂H₂ (ppm) | Actual fault | Duval triangle | SVM | BPNN | IVSVM |
|-----|----------|-----------|------------|------------|------------|--------------|---------------|-----|------|-------|
| 1   | 13       | 20.5      | 5.7        | 46.1       | 7          | A            | T3            | I   | A    | A     |
| 2   | 22       | 32.3      | 6.3        | 81.4       | 13.5       | A            | T3            | A   | A    | A     |
| 3   | 11       | 77.4      | 0.5        | 87.1       | 40.4       | B            | T3            | B   | B    | B     |
| 4   | 5        | 80.3      | 0.7        | 83.3       | 50.6       | B            | T3            | B   | B    | B     |
| 5   | 6.66     | 47.14     | 1.51       | 86.56      | 17.35      | C            | T3            | C   | C    | C     |
| 6   | 11.45    | 60.45     | 2.35       | 110.66     | 18.76      | C            | T3            | B   | B    | C     |
| 7   | 1035.9   | 49.79     | 0.3        | 1.79       | 9.1        | D            | T₁            | D   | D    | D     |
| 8   | 1400     | 52.35     | 0.6        | 1.75       | 9.68       | D            | T₁            | D   | D    | D     |
| 9   | 368.05   | 47.64     | 0.1        | 2.5        | 3.17       | E            | T₁            | D   | D    | D     |
| 10  | 211.9    | 7.54      | 0.06       | 1.52       | 9.51       | E            | T₁            | D   | D    | D     |
| 11  | 78       | 227.7     | 1.6        | 365.3      | 48.7       | F            | T₃            | F   | F    | F     |
| 12  | 58.95    | 111.4     | 1.14       | 199.38     | 24.59      | F            | T₃            | H   | H    | H     |
| 13  | 144      | 69.9      | 88.8       | 104.6      | 13.8       | G            | HED           | G   | G    | G     |
| 14  | 107.39   | 47.8      | 177.31     | 63.68      | 6.32       | G            | LED           | G   | G    | G     |
| 15  | 77.82    | 72.41     | 1.59       | 67.01      | 24.76      | H            | T₂            | F   | F    | H     |
| 16  | 63.23    | 69.35     | 1.67       | 68.28      | 25.2       | H            | T₂            | H   | H    | H     |
| 17  | 14       | 3.8       | 6.9        | 20.3       | 2.9        | I            | D + T         | A   | I    | I     |
| 18  | 19       | 13.7      | 7.8        | 9.1        | 4.8        | I            | HED           | I   | I    | I     |

Abbreviations: BPNN, BP neural network; HED, high energy discharge; IVSVM, information entropy and vague support vector machine; LED, low energy discharge; SVM, support vector machine.
### TABLE 6  Accuracy of SVM, BPNN, IVSVM (%)

| Fault Type | SVM Training accuracy | SVM Testing accuracy | BPNN Training accuracy | BPNN Testing accuracy | IVSVM Training accuracy | IVSVM Testing accuracy |
|------------|-----------------------|----------------------|------------------------|-----------------------|-------------------------|------------------------|
| A          | 84                    | 80                   | 88                     | 85                    | 94                      | 95                     |
| B          | 92                    | 85                   | 90                     | 90                    | 96                      | 90                     |
| C          | 90                    | 80                   | 86                     | 85                    | 92                      | 85                     |
| D          | 94                    | 95                   | 92                     | 95                    | 98                      | 100                    |
| E          | 78                    | 75                   | 84                     | 85                    | 90                      | 90                     |
| F          | 88                    | 85                   | 86                     | 80                    | 92                      | 85                     |
| G          | 90                    | 90                   | 94                     | 90                    | 96                      | 95                     |
| H          | 86                    | 85                   | 92                     | 80                    | 94                      | 90                     |
| I          | 84                    | 85                   | 90                     | 90                    | 96                      | 95                     |
| **Total**  | **87.3**              | **84.4**             | **89.1**               | **86.7**              | **94.2**                | **91.7**               |

Abbreviations: BPNN, BP neural network; IVSM, information entropy and vague support vector machine; SVM, support vector machine.
Five types of dissolved gases in transformer oil are selected as the characteristic quantities for transformer fault diagnosis. The improved information entropy method is used to obtain the weights of the training data and to weight the training data and the test data.

VSVM is used to divide the weighted data into three parts and build multiple one-to-one classifier models. The parameters of VSVM are obtained by the KNN method and the particle swarm method for optimisation. VSVM can ensure the accuracy of one-to-one classification while coping with the problem of easy misclassification of interclass points.

A total of 342 data from collected open source international data sets and 630 actual transformer fault data from a location in southwest China were used to test the method proposed herein. By comparing the results with various ratio methods as well as SVM and BPNN methods, it was found that the proposed method paper has a higher accuracy. In the face of more types of faults in actual operation and the characteristics between classes not being obvious, this method can better classify the faults. It plays a certain role in transformer fault monitoring in a region in southwest China, and has a more important guiding significance for the actual operation and maintenance work.

ACKNOWLEDGEMENTS
This project was supported by National Natural Science Foundation of China (51,777,117), Natural Science Foundation of China State Grid Corporation Smart Grid Joint Fund (U1966209).

ORCID
Hongru Zhang https://orcid.org/0000-0002-9882-0175
Jiaxiang Sun https://orcid.org/0000-0002-4070-5502

REFERENCES
1. Duval, M.: Dissolved gas analysis: it can save your transformer [J]. IEEE Elect. Insul. Mag. 5(6), 22–27 (1989)
2. Duval, M.: New techniques for dissolved gas-in-oil analysis [J]. IEEE Electr Insul Mag. 19(2), 6–15 (2003)
3. CIGRE working group 09 of study committee 12: Life time evaluation of immersed transformers, IEEE. New York (2019)
4. Rogers, R.R.: IEEE and IEC codes to interpret incipient faults in transformers using gas in oil analysis [J]. IEEE Trans. Dielectr. Electr. Insul. 13(5), 348–354 (1978)
5. CIGRE, M.: Dissolved gas analysis in power transformers based on the five combustible gases [J]. IEEE Trans. Dielectr. Electr. Insul. 22(5), 2507–2512 (2015)

How to cite this article: Zhang H, Sun J, Hou K, Li Q, Liu H. Improved information entropy weighted vague support vector machine method for transformer fault diagnosis. High Voltage. 2021;1–13. https://doi.org/10.1049/hve.2.12095
## APPENDICES

### TABLE A1  Part of the testing data

| Fault Type | $\text{CH}_4$ (ppm) | $\text{C}_2\text{H}_6$ (ppm) | $\text{C}_2\text{H}_4$ (ppm) | $\text{C}_2\text{H}_2$ (ppm) | $\text{H}_2$ (ppm) |
|------------|----------------------|-----------------------------|-----------------------------|-----------------------------|-------------------|
| A          | 20.5                 | 7                           | 46.1                        | 5.7                         | 79.3              |
| A          | 29.03                | 10.85                       | 69.14                       | 5.19                        | 114.21            |
| A          | 32.3                 | 13.5                        | 81.4                        | 6.3                         | 133.5             |
| A          | 11.12                | 2.76                        | 24.43                       | 5.45                        | 43.76             |
| A          | 14.51                | 3.45                        | 31.22                       | 6.34                        | 55.52             |
| B          | 65.76                | 29.82                       | 78.47                       | 0                           | 174.05            |
| B          | 75.2                 | 53.7                        | 95.1                        | 0                           | 224               |
| B          | 77.4                 | 40.4                        | 87.1                        | 0                           | 204.9             |
| B          | 106.5                | 59.6                        | 97.4                        | 0                           | 263.5             |
| B          | 80.3                 | 50.6                        | 83.3                        | 0                           | 214.2             |
| B          | 59.43                | 33.69                       | 74.05                       | 0                           | 167.17            |
| C          | 45.37                | 15.47                       | 75.56                       | 1                           | 137.4             |
| C          | 47.14                | 17.35                       | 86.56                       | 1.51                        | 152.56            |
| C          | 41.97                | 15.1                        | 76.53                       | 1.65                        | 135.25            |
| C          | 60.45                | 18.76                       | 110.66                      | 2.35                        | 192.22            |
| C          | 62.43                | 18.87                       | 112.13                      | 2.15                        | 195.58            |
| D          | 49.79                | 9.1                         | 1.79                        | 0                           | 60.68             |
| D          | 35.36                | 7.59                        | 1.68                        | 0                           | 44.63             |
| D          | 61.45                | 12.91                       | 2.04                        | 0                           | 76.4              |
| D          | 52.35                | 9.68                        | 1.75                        | 0                           | 63.78             |
| D          | 75.4                 | 15.03                       | 1.75                        | 0                           | 92.18             |
| E          | 34.2                 | 1.5                         | 1                           | 0                           | 36.7              |
| E          | 47.64                | 3.17                        | 2.5                         | 0                           | 53.31             |
| E          | 20.4                 | 12.5                        | 20.5                        | 0                           | 53.4              |
| E          | 19.62                | 2.02                        | 0.86                        | 0                           | 22.5              |
| E          | 7.54                 | 9.51                        | 1.52                        | 0                           | 18.57             |
| E          | 39.6                 | 2.6                         | 1                           | 0.4                         | 43.6              |
| F          | 94.9                 | 30.6                        | 171.4                       | 0.4                         | 297.3             |
| F          | 55.76                | 21.02                       | 80.68                       | 0                           | 157.46            |
| F          | 227.7                | 48.7                        | 365.3                       | 1.6                         | 643.3             |
| F          | 178.4                | 48.2                        | 308                         | 3.4                         | 538               |
| F          | 111.45               | 24.59                       | 199.38                      | 1.14                        | 336.56            |
| F          | 151.9                | 43.5                        | 304.4                       | 1.3                         | 501.1             |
| G          | 42.7                 | 5.1                         | 50.3                        | 5.54                        | 153.5             |
| G          | 57.5                 | 6                           | 59.1                        | 6.68                        | 189.4             |
| G          | 69.9                 | 13.8                        | 104.6                       | 8.88                        | 277.1             |
| G          | 83.1                 | 16.4                        | 176.6                       | 308.9                       | 585               |
| Fault Type | $\text{CH}_4$ (ppm) | $\text{C}_2\text{H}_6$ (ppm) | $\text{C}_3\text{H}_8$ (ppm) | $\text{C}_4\text{H}_8$ (ppm) | $\text{H}_2$ (ppm) |
|-----------|---------------------|---------------------|---------------------|---------------------|---------------------|
| G         | 54.76               | 9.17                | 103.27              | 74.99               | 242.19              |
| G         | 47.8                | 6.32                | 63.68               | 177.31              | 295.11              |
| H         | 72.41               | 24.76               | 67.01               | 1.59                | 165.77              |
| H         | 68.58               | 25.01               | 71.6                | 1.78                | 166.97              |
| H         | 69.35               | 25.2                | 68.28               | 1.67                | 164.5               |
| H         | 76.6                | 23.69               | 68.87               | 1.97                | 171.13              |
| H         | 60.43               | 22.24               | 66.82               | 1.62                | 151.11              |
| H         | 69.87               | 24.79               | 72.26               | 1.85                | 168.77              |
| I         | 4.38                | 1.31                | 7.63                | 15.65               | 28.97               |
| I         | 3.8                 | 2.9                 | 20.3                | 6.9                 | 33.9                |
| I         | 8.6                 | 1.3                 | 5.2                 | 20.7                | 35.8                |
| I         | 13.7                | 4.8                 | 9.1                 | 7.8                 | 35.4                |
| I         | 5.58                | 1.57                | 9.75                | 15.05               | 31.95               |
| I         | 6.76                | 0.9                 | 4.58                | 18.9                | 31.14               |