Electr Eng (2017) 99:1109–1119
DOI 10.1007/s00202-016-0481-3

ORIGINAL PAPER

A nearest neighbour clustering approach for incipient fault diagnosis of power transformers

Md Mominul Islam[1] · Gareth Lee[1] · Sujeewa Nilendra Hettiwatte[2]

Received: 14 September 2016 / Accepted: 4 November 2016 / Published online: 16 November 2016

Abstract Dissolved gas analysis (DGA) is one of the popular and widely accepted methods for fault diagnosis in power transformers. This paper presents a novel DGA technique to improve the diagnosis accuracy of transformers by analysing the concentrations of five key gases produced in transformers. The proposed approach uses a clustering and cumulative voting technique to resolve the conflicts and deal with the cases that cannot be classified using Duval Triangles, Rogers' Ratios and IEC Ratios Methods. Clustering techniques group the highly similar faults into a cluster providing a virtual boundary between dissimilar data. A cluster of data points may contain single or multiple types of faulty transformers' data with different distinguishable percentages. The *k*-nearest neighbour (KNN) algorithm is used for indexing the three closest clusters from an unknown transformer data point and allows them to vote for single or multiple faults categories. The cumulative votes have been used to identify a transformer's fault category. Performance of the proposed method has been compared with different conventional methods currently used such as Duval Triangles, Rogers' Ratios and IEC Ratios Method along with published results using computational and machine learning techniques such as rough sets analysis, neural networks (NNs), support vector machines (SVMs), extreme learning machines (ELM) and fuzzy logic. The experimental comparison with both published and utility provided data show that the proposed method can significantly improve the incipient fault diagnosis accuracy in power transformers.

Keywords Clustering techniques · Dissolved gas analysis (DGA) · Incipient faults · Machine learning techniques · *k*-Nearest neighbour (KNN) algorithm · Power transformer

1 Introduction

A power transformer is one of the most important and expensive components in power transmission and distribution systems [1]. Its precise operation is essential for ensuring the reliable and stable operation of a power system. Any fault in the power transformer may lead to the unscheduled outages resulting in interruption of power supply. Failure of transformers hampers the stability of operation and causes a great loss to the utilities. As sudden failure of a transformer can result in an explosion, it has significant implications both for quality of service and it poses a risk to both maintenance crew and the general public. Therefore, incipient fault diagnosis and condition monitoring of power transformers are both gaining attention [2], by the utilities to ensure continuous operation and minimizing the operational risks.

Due to the continuous operation and variable loading, transformers are always subjected to electrical, thermal, mechanical and chemical stresses. As a result, different types of combustible and non-combustible gases, such as hydrogen (H_2), oxygen (O_2), nitrogen (N_2), carbon dioxide (CO_2), carbon monoxide (CO), methane (CH_4), ethylene (C_2H_4), ethane (C_2H_6), acetylene (C_2H_2), propane (C_3H_8) and propylene (C_3H_6), are released and dissolved in transformer oil [3]. Moreover, the excessive thermal, electrical

✉ Md Mominul Islam
md.islam@murdoch.edu.au

Gareth Lee
Gareth.Lee@murdoch.edu.au

Sujeewa Nilendra Hettiwatte
sujeewa@nsbm.lk

[1] School of Engineering and Information Technology, Murdoch University, Murdoch, Australia

[2] School of Engineering, National School of Business Management, Colombo, Sri Lanka

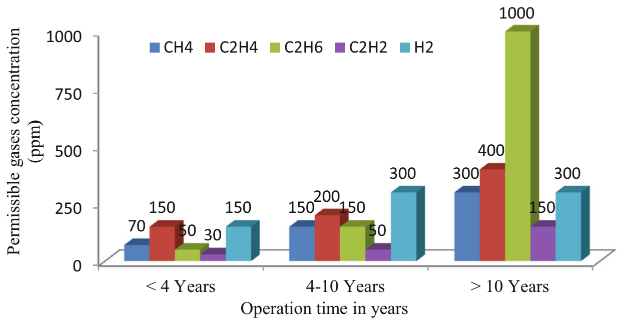

Fig. 1 Permissible concentration of dissolved gases in a healthy transformer [10]

and chemical stresses change the dielectric properties and mechanical strength of cellulose paper and produce furanoid compounds, namely 2-furfural (2-FAL), 5-Hydroxy methyl-2-furfural (5-HMF), 5-Methyl-2-furfural (5-MEF), 2-Furfurol (2-FOL) and 2-Acetylfuran (2-ACF), which are partially soluble in oil [4,5]. To monitor the insulation condition and detect faults in a transformer, different techniques such as dissolved gas analysis (DGA), Furan analysis, degree of polymerisation (DP) measurement, gas chromatography (GC), Mass chromatography (MC), high-performance liquid chromatography (HPLC) and moisture analysis are available [6]. Among these, DGA is a non-invasive, proven and widely accepted method to detect incipient faults in transformers. The DGA method can be used to continuously monitor the overall condition of transformers and generate advance warnings of newly developing faults. Therefore, operators can conveniently plan their remedial action following the total gas production or the production rate of individual gases that minimizes the risk of premature failure.

To analyse the measured gas concentrations, DGA techniques such as the family of Duval Triangles, Key Gas, Modified Rogers' Ratios, Doernenburg and IEC Ratios have been used over the last few decades [7,8]. The Duval Triangles are one of the preferred methods for many utility companies, as they do very well in classifying incipient faults, and can assess the state of insulation in transformers. However, there are cases where the Duval Triangles fail to produce any classification and there is a chance of misclassification near the boundary between adjacent regions [9]. To investigate these unusual cases, a link has been established with a large utility company in Western Australia having more than 350 power transformers in operation. The company primarily uses Duval Triangles to assess the overall condition of their transformers. They also consider IEC, Modified Rogers' Ratios and Key Gas methods before classifying their transformers into a fault category, especially for the cases when Duval Triangles fail or ambiguously classify a transformer. Additionally, over time, gases are produced by normal operation of transformers without indicating any fault. Consequently, there is a chance of misclassification of healthy transformers. According to [10], the permissible limit of dissolved gases in transformer oil corresponding to the operating time of a healthy transformer is shown in Fig. 1.

In this paper, a novel fault diagnosis technique has been proposed which can effectively classify the critical cases where there is a contradiction between various Duval Triangles and cannot be classified by the conventional ratio methods. The arrangement of the paper is as follows. Section 2 describes the motivation of the research. Section 3 presents the basic concept of k-means algorithm (KMA). Section 4 describes the proposed machine learning technique utilizing k-means clustering and k-nearest neighbour pattern classification. Section 5 presents the results that were achieved with the method, Sect. 6 presents a case study and Sect. 7 presents a summary of the results and conclusions.

2 Motivation of research

DGA is one of the proven methods that is used widely by many utilities for condition monitoring and fault diagnosis in power transformers. Due to their continuous operation, the normal ageing of transformer insulation is inevitable. Frequent overloading and short circuit incidents can create electrical, mechanical and thermal stresses in transformer insulation that degrade the dielectric properties of insulation and increase its ageing rate. The decomposition of insulating material increases the operational risk to transformers and produces several combustible and non-combustible gases that become partly dissolved in transformer oil. In DGA-based analysis, the concentration and production rate of gases are measured and continuously monitored to assess the insulation condition of a transformer and locate the sources of faults. To interpret the dissolved gases, a graphical technique like the Duval Triangles and different ratio methods such as Rogers' Ratios, Doernenburg Ratios, IEC Ratio, Single Gas Ratio and the Key Gas method are used over long periods of time. All of these conventional methods are very simple and easy to implement. However, different methods have different advantages and limitations. Therefore, comparison of the results from different methods on the same sample may lead to contradictions, and there is no clear way to prioritize one result over another, leading to ambiguity [11]. The accuracy of the IEC method is affected by the incomplete coding and absolute code boundary. It cannot identify the fault samples, if they fall outside the definite ratio limits. In addition, the interference problem between low-energy discharge (D1) and high-energy discharge (D2) of this method may lead to misleading classification [12]. The classification of Rogers' Ratios is not precise for detecting all faults [12]. It gives more accurate diagnosis for the low thermal (T1) fault. The Doernenburg method can only provide three types of diagnosis. It cannot distinguish the severity of thermal decomposition.

Table 1 Rogers' ratios [13]

Case	R2 = C_2H_2/C_2H_4	R1 = CH_4/H_2	R5 = C_2H_4/C_2H_6	Suggested fault diagnosis
0	<0.1	>0.1 to <1.0	<1.0	Unit normal
1	<0.1	<0.1	<1.0	Low-energy density arcing (LEDA)/PD
2	0.1 to 3.0	0.1 to 1.0	>3.0	Arcing-high-energy discharge (AHED)
3	<0.1	>0.1 to <1.0	1.0 to 3.0	Low-temperature thermal (LTP)
4	<0.1	>1.0	1.0 to 3.0	Thermal <700°C (T2)
5	<0.1	>1.0	>3.0	Thermal >700°C (T3)

Table 2 Ratio limits for respective faults based on IEC60599 (2007)

Case	Characteristic fault	C_2H_2/C_2H_4	CH_4/H_2	C_2H_4/C_2H_6
PD	Partial discharge	–	<0.1	<0.2
D1	Low-energy discharges	>1	0.1 to 0.5	>1.0
D2	High-energy discharges	0.6 to 2.5	0.1 to 1.0	>2.0
T1	Thermal fault <300°C	–	>1.0	<1.0
T2	Thermal fault 300 to 700°C	<0.1	>1.0	1.0 to 4.0
T3	Thermal fault >700°C	<0.2	>1.0	>4.0

All these ratio methods do not involve any mathematical formulation and their accuracy is dependent on the concentration and ratio of the key gases. Moreover, in some cases, the calculated ratios do not fall within any of the fault classes and remain unclassified. The Duval Triangles always gives a fault diagnosis even when a transformer is known to be healthy. The classical Duval Triangle cannot accurately detect the partial discharge (PD) and thermal fault. For mineral oil-filled transformers, if the fault classification is a thermal fault or a partial discharge by the classical triangular method, then Triangles 4 and 5 must be used for further clarification. In practice, there are cases where contradictory classifications are produced by Triangles 4 and 5. Moreover, all triangles have an unclassified region. Consequently, the accuracy of fault classification is dependent on the expert's experience supported by other ratio methods. The classification of transformers' incipient faults following the Rogers' Ratios, and IEC ratios has been tabulated in Tables 1 and 2, respectively.

To overcome the limitations of these conventional approaches, various computational and machine learning techniques such as support vector machines (SVMs) [2], neural networks (NNs) [14], extreme learning machines (ELM) [15], fuzzy logic [16,17] and rough sets (RS) detection [18] have been combined with DGA interpretation techniques to analyse the incipient faults in transformers. These new techniques have improved the accuracy of fault diagnosis and solved the interference problem between fault classes. The combined approach is helping researchers and utilities to explore the relationship between different fault patterns and their characteristic parameters. In this research, a modified clustering technique and k-Nearest Neighbour algorithm have been used, and a modified cumulative voting mechanism has been proposed to classify and predict the incipient faults in power transformers.

3 Basic concepts of K-means algorithm

Cluster analysis is prevalent in any discipline that aims to find the natural grouping, detect anomalies and identify salient features of data points in a given data set. The groups are called clusters and the region belonging to a cluster is a Voronoi cell [19], in which the density of similar data points is higher than in other regions. A good clustering technique generally uses the splitting, merging or randomized approaches for partitioning given data points into clusters so that the formal objective function is optimized [19]. The most common objective in clustering technique is to minimize the squared error between the empirical mean of a cluster and the points lying in its Voronoi cell. The k-means algorithm (KMA) is one of the simplest and widely used unsupervised learning algorithm that minimizes the clustering error and can be used to discover the natural grouping of data points [19,20]. For a set of n data points $X = \{x_1, x_2, \ldots, x_n\}$, in a real d-dimensional space R^d, KMA determines a set of K cluster $C = \{\mu_1, \mu_2, \ldots, \mu_k\}$ in R^d such that the mean squared Euclidean distance from each data point to the nearest centre is minimized. Each of the clusters is associated with a subset of X such that any x_i is a member of only one cluster. The subset of X that cluster around μ_k will be referred to as C_k. Each of the subsets is disjoint and, therefore, the union of all C_k provides the entire set of points and can be expressed as,

$$\bigcup_{k=1}^{K} C_k = X. \tag{1}$$

According to [21], the sum of the squared error for the set of clusters C can be defined as

$$\mathrm{J}\,(\mathrm{C}) = \sum_{k=1}^{K} \sum_{x_i \in C_k} x_i - \mu_k^2. \tag{2}$$

Although, the objective function J (C) decreases with the increasing number of clusters K,and become zero when $K = n$, the number of cluster centres and data points are equal. It needs to be minimized when $K < n$ using KMA. The steps of KMA are as follows [22–24]:

1. Place Kcentres into the d-dimensional space of the data points X. The locations are known as initial centres (μ_k) of the C_k clusters.
2. Assign data points to the Voronoi cell (group) which centre has the closest distance to form subsetsC_k.
3. Compute new cluster centres $\mu_k$from the mean of data points lying in their Voronoi cell.
4. Repeat steps 2 and 3 until the square error is reduced to a pre-determined value or the centroids are immobilized.

Although KMA is a simple and popular method in clustering applications, it is very sensitive to the initial positions of the cluster centres [21]. It is an NP-hard algorithm and, therefore, a globally optimal solution cannot be found, except for unrealistically small values of n and K. However, there are well-established heuristic algorithms, such as those employed here, for providing adequate albeit suboptimal clustering when n and K are larger. In this work, a number of heuristics like Lloyd's algorithm and Linde–Buzo–Gray (LBG) have been combined with conventional KMA [19]. These approaches offer a more efficient clustering algorithm that can minimize the clustering error, employed by the conventional KMA as a local search procedure. To solve the K clustering problem, the proposed hybrid approach proceeds in an incremental way. Initially, a single centre is calculated following the LBG algorithm and placed in the d-dimensional space. The centre is calculated from the geometric mean of the given data points. In each stage, the old centroids are split into two.Moreover, in each stage of the LBG algorithm, the nearest data points to each centre are computed and the centres are moved to the centroid of data points lying in their Voronoi cell. These steps are repeated until some convergence condition is met. The only difference between Lloyd's and LBG algorithm is that LBG specifies the initial placement of a centre which is absent in Lloyd's algorithm [19]. A detailed procedure of the LBG algorithm has been discussed in [25]. In the LBG algorithm, there is no guarantee that every cluster centre will have some data association. This limitation can be overcome by supervising the splitting and relocating any centres where cluster (C_k) becomes an empty set. The centres obtained from supervised LBG can be used as initial centres for KMA to take advantage of global minima. This hybrid method reduces the clustering error that results from the local convergence. Finally, the feature of clusters with the collaboration of k-nearest neighbour (KNN) algorithm can be used in a diagnostic decision table to classify the fault category of transformers.

4 Methodology

The methodology involved a development of a clustering process combined with a cumulative voting technique to determine the fault category of a transformer. This section includes data collection, pre-processing, model development, neighbour selection and training stage which have been discussed below.

4.1 Data collection and processing

Data of combustible gases generated from the insulating oil in 376 power transformers have been collected by a large utility company in Western Australia. The gas concentrations are measured in parts per million (ppm) by analysing an oil sample drawn from each transformer under laboratory conditions. The measured gas concentrations have been analysed using the Rogers' Ratios, Duval Triangles, Doernenburg and IEC ratio methods and verified by the utility's experts before labelling them into a fault category. It is presumed that the final classification from the combined approach of different conventional methods, sophisticated software analysis and experts judgement is accurate and reliable. To verify the accuracy of suspected faulty transformers, these transformers have been removed from the services for investigation, and the findings have exactly matched with the expert's fault classification. The proposed method is based on a clustering technique that uses the percentage concentrations of the five combustible gases comprising Hydrogen (H_2), Methane (CH_4), Ethylene (C_2H_4), Ethane (C_2H_6) and Acetylene (C_2H_2). The sum of the five gas concentrations has been calculated as per (3). The summation is defined as the total combustible gases (TCG).

$$\mathrm{TCG} = H_2 + CH_4 + C_2H_4 + C_2H_6 + C_2H_2 \tag{3}$$

Therefore, the percentage of those individual gases has been calculated and used as an input for the proposed method to classify the testing data sets into seven targeted fault categories. The individual percentage calculation procedure and the targeted fault category with their fault code have been shown in Table 3.

Moreover, the collected data sets are divided into two subsets. The first subset (318 measurements) is used as a

Table 3 Input and targeted output of the proposed method

Input	Targeted fault category	Fault code
1. $\%H_2 = \frac{H_2}{TCG} \times 100$	1. Partial discharge	1. PD
2. $\%CH_4 = \frac{CH_4}{TCG} \times 100$	2. Discharge of low energy	2. D1
3. $\%C_2H_4 = \frac{C_2H_4}{TCG} \times 100$	3. Discharge of high energy	3. D2
4.$\%C_2H_6 = \frac{C_2H_6}{TCG} \times 100$	4. Thermal fault, $t < 150°C$	4. S
5. $\%C_2H_2 = \frac{C_2H_2}{TCG} \times 100$	5. Thermal fault, $150^{\circ}C < t < 300°C$	5. O
	6. Thermal fault, $300\ ^{\circ}C < t < 700°C$	6. C
	7. Thermal fault, $t > 700°C$	7. T3

Table 4 Probability of a transformer fault following their association with a cluster centre

Cluster centres	C (%)	D1 (%)	D2 (%)	O (%)	PD (%)	S (%)	T3 (%)
μ_1		100					
μ_2					80	10	10
μ_3	10						90
	……	……	…….	…….	……..	……	….
μ_{15}	100						
μ_{16}		25				75	
μ_{17}	67						33
…	……	……	…….	…….	……..	……	….
μ_{25}		8	92				
μ_{26}		86	14				
μ_{27}			100				
…	……	……	…….	…….	……..	……	….
μ_{32}				100			
μ_{33}				100			
μ_{34}						100	

Table 5 Modified voting metrics

Neighbours	C	D1	D2	O	PD	S	T3
C_3	10 × 0.65						90 × 0.65
C_2					80 × 0.25	10 × 0.25	10 × 0.25
C_{17}	67 × 0.10						33 × 0.10
Total vote	13.2				20	2.5	64.3

Table 6 Comparison of Rogers' ratios, IEC ratios and the proposed method

Test methods	Unresolved diagnosis	Wrong diagnosis	Accuracy (%)
Roger's ratios	2	9	75.67
IEC ratio	27	8	74.19
Proposed method	–	4	93.10

training data set and the second subset (58 measurements) which could not classified easily by the Duval Triangles or come with a conflicting classification due to the overlapping between different faults.

4.2 Proposed model

In this work, a hybrid clustering technique has been used because it has advantages over the Duval Triangles method. In Duval Triangles, five-dimensional key gases are mapped into a set of two-dimensional spaces to make a classification of transformer faults based on a set of linear boundaries. This dimension-reducing mapping throws away some valuable information. The proposed clustering technique preserves all of the five-dimensional gases information in the expectation that it can do better than the Duval Triangles and other ratio methods, in the cases when they fail or contradict each other. There are two stages in the proposed approach. First, a set of

Table 7 Fault diagnosis comparison between established methods and adapted methods

Case number	H_2	CH_4	C_2H_4	C_2H_6	C_2H_2	Diagnosis result							
						Expert diagnosis	Duval triangle	RS	RS-ANN	KMC-RS-ANN	Roger's ratios	IEC ratio	Proposed method
1	60	40	110	10	70	D2	D2	D2	D2	D2	D2	D2	D2
2	31	7	5	19	67	D2	D1	–	D2	D2	–	–	D1
3	293	50	15	13	120	D2	D2	D2	D2	D2	D2	D1	D2
4	57	7	4.5	19	71	D2	D1	–	D2	D2	–	–	D1
5	467	148	266	13	511	D2	D2	D2	DT	D2	D2	D1	D2
6	160	90	17	27	58	D1	D1	D1	D1	D1	–	–	D1
7	402	81	27	39	25	D1	D2	D1	D1	D1	–	–	D1
8	4	79	312	112	0	T2	T1	–	T2	DT	T2	T2	T2
9	180	180	4	74	3	DT	T3	DT	DT	DT	–	–	T2
10	1300	740	2000	260	71	T3	T3	T3	T3	T3	–	–	T3
11	42	97	600	157	0	T3	T3	Normal	Normal	Normal	T2	T2	T3
12	44	52	119	15	1	T3	T3	T3	T3	T3	T3	T3	T3
13	42	79	152	31	1	T3	T3	T3	T3	T3	T3	T3	T3
14	164	244	497	103	8	T3	T3	T3	T3	T3	T3	T3	T3
15	22	51	57	42	0	T2	T2	T2	T2	T2	T2	T2	T2
16	679	4992	3671	1823	0	T2	T1	T2	T2	T2	T2	T2	T2

clusters based around the global k-Means Algorithm (KMA) is generated. The clusters are representative of various fault categories. After clustering, the KNN algorithm has been used to decide which clusters are closest to the data set of an unclassified transformer. In this research, three closest clusters have been identified based on their Euclidean distances from the testing data. The specific procedure for clustering the data points and neighbours selection has been discussed in the following sections and a summary of the comparative performances can be seen in Tables 7, 8, 9. However, a workflow of the proposed model with the combination of Duval Triangles has been shown in Fig. 2.

4.3 Clustering procedure

The clustering of the training data points has been completed using modified KMA. To perform the clustering of training data points $X = \{x_1, x_2, \ldots, x_n\}$, an initial cluster centre $\mu_1 \in C$ for KMA is computed following the LBG algorithm (mean of all data points) and placed in the five-dimensional space $\left(R^5\right)$ formed by the percentages of five gas concentrations. The number of cluster centres gradually increased to K (where $K = 2^n$ for $n = 1, 2, \ldots, 6$) through successive iterations. In each stage of iteration, the old centre(s) is split into two and Euclidean distances from the centres to all data points are calculated. Let the C_j represent the data points (neighbourhood) in a Voronoi cell for which μ_j is the nearest centre. The set of data points lying to the cluster C_j can be expressed as follows

$$C_j = \left\{x_i : \left|\left|x_i - \mu_j\right|\right|^2 \leq \left|\left|x_i - \mu_k\right|\right|^2 \forall\, k = 1, 2, 3, \ldots, K\right\} \tag{4}$$

where $||x_i - \mu_j||^2$ is the Euclidean distance between a training data points x_i $(i = 1, 2, 3, \ldots, n)$ and the cluster centre μ_j, and K is the number cluster centres. After finishing the allocation of all data points to the Voronoi cells of cluster centres, for the next iteration, the position of cluster centres moves to the centroid of data in subset C_j. The new position of the cluster centres can be calculated by the following equation.

$$\mu_j = \frac{\sum_{x_i \in C_j} x_i}{\sum_{x_i \in C_j} 1} \tag{5}$$

In the next stage, the distances of all points from the new position of centres are again calculated and associated to centres having smallest Euclidean distances. These steps are

Table 8 Comparison between decision tree, ratio and proposed methods

Case number	H_2	CH_4	C_2H_4	C_2H_6	C_2H_2	Diagnosis result					
						Actual diagnosis	New approach DGA	Modified new approach DGA	Roger's ratios	IEC ratio	Proposed method
1	117	17	3	1	1	PD	PD,D1	PD	–	–	PD
2	32930	2397	0	157	0	PD	PD,D1	PD	PD	PD	PD
3	78	20	13	11	28	D1	D1,D2	D1	–	D1	D2
4	1230	163	233	27	962	D1	D1,D2	D1	D2	D1	D1
5	8200	3790	4620	250	277	D2	D2,T1	D2	–	–	D2
6	130	140	120	2	0	T1	D2,T1	T1	T3	T3	T2
7	78	66	2.6	283	0	T1	D2,T1	T1	Normal	PD	T1
8	30.4	117	138	44.2	0.1	T2	T2,T3	T2	T3	T2	T2
9	27	90	63	42	0.2	T2	T1,T2	T1	T2	T2	T2
10	1100	1600	2010	221	26	T3	T2,T3	T3	T3	T3	T2

Table 9 Comparison of different adapted fault diagnosis methods

Case number	H_2	CH_4	C_2H_4	C_2H_6	C_2H_2	Diagnosis result							
						ANN	SVM	ELM	SaE-ELM	Duval triangle	Roger's ratios	IEC ratio	Proposed method
1	103	5.8	7.3	5	0.7	T1	T3	T1	T1	S	–	–	T1
2	416	21	43.1	10.5	1	T1	T3	T1	T3	T3	–	–	T3
3	59	53	60.3	17.7	0.8	T2	T2	T2	T2	C	–	–	T2
4	10.5	4.8	4.8	5	2.2	D1	D1	D1	D1	DT	–	–	D1
5	137	97	29	12	1.5	T2	T2	T2	T2	C	LEDA/PD	–	T2
6	89	73	6.8	6	5	D2	D2	D2	D2	DT	–	–	T2
7	240	157	127	98	0.8	T2	T2	T2	T2	C	LEDA/PD	–	T2
8	116	104	51	36	0	T2	T2	T2	T2	C	LEDA/PD	–	T2

repeated until μ_j becomes immobilized or the square error J (C) is reduced to a pre-determined value [22].

4.4 Neighbor selection and voting

In this section, the distances to three closest cluster centres μ_i, $\mu_j$andμ_k from any unknown transformer data point x is measured and sorted into ascending order of distance. In a later stage, the clusters have been used in a voting mechanism to classify a faulty transformer. In the case of conventional voting, a decision is taken based on the majority opinion. For instance, if one neighbouring cluster is associated with a T1 fault, the second one might be associated with T2 and third neighbour might be labelled as a T1 fault, then the majority of votes for T1 will classify the transformer as having a T1 type fault. As the clusters are sometimes straddling the boundaries of the Duval Triangles method, a cumulative voting system has been introduced where each of the three clusters can vote for multiple fault categories. In most cases, they vote for one fault category; sometimes they can also vote for two or three categories because they might be closer to a corner where three faulty regions join together. In the final step, a distance matrix has been used for cumulative voting. Hence, the cluster that is closest has a stronger bearing on the overall result; the next farthest has a weaker bearing and the next most distant one has a still weaker bearing on the final result. Mathematically, the voting weight of any cluster can be expressed by the following equation.

$$W_p = \frac{S - ||x - \mu_p||^2}{2 * S} \tag{6}$$

where $S = \sum \left|\left|x - \mu_i\right|\right|^2 + \left|\left|x - \mu_j\right|\right|^2 + ||x - \mu_k||^2$ and $p \in \{i, j, k\}$ are the three nearest clusters centres from any data point x. Finally, the cumulative votes are added up in a weighted fashion to determine the classification of an unknown transformer most effectively.

4.5 Training stage

During the training stage, 318 transformers' gas concentrations out of 376 collected from the utility company have been used to develop the proposed clustering technique. These training data points were excluded from the test set. The fault categories of the training data points are labelled by the utility experts. The number of individual fault category of the training samples is shown in Fig. 3.

To create a cluster of points (rows) in a five-dimensional Euclidean space, the relative percentages of the five gases have been calculated and put into the individual columns of a matrix X. The matrix X can be expressed as

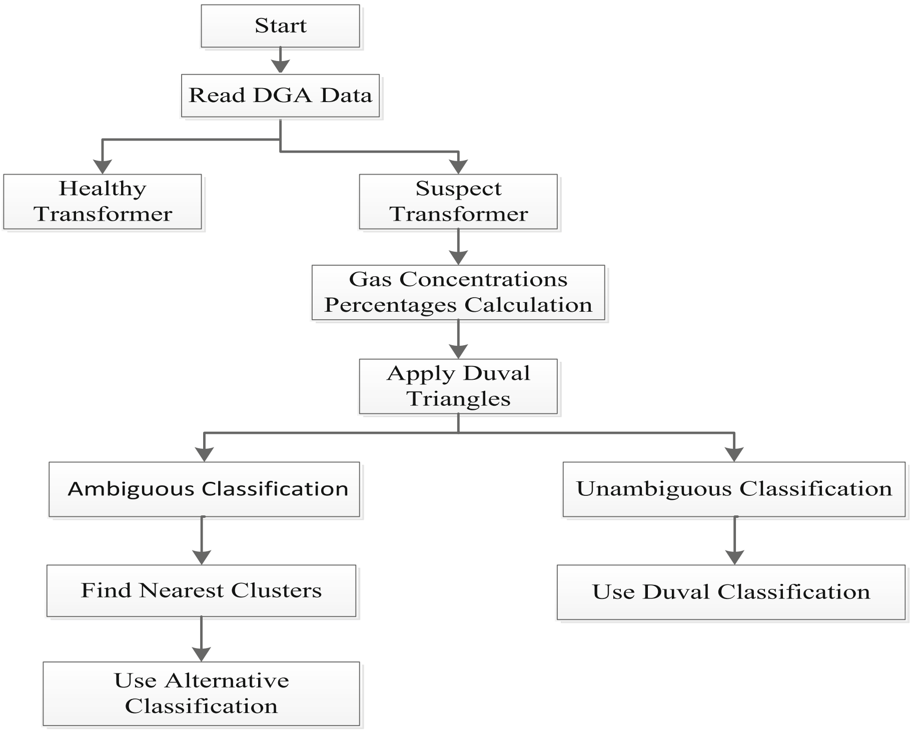

Fig. 2 Workflow of the proposed model for practical application

$$\boldsymbol{X} = \begin{bmatrix} \%H_{2_{1,1}}, & \%CH_{4_{1,2}}, & \%C_2H_{4_{1,3}}, & \%C_2H_{6_{1,4}}, & \%C_2H_{2_{1,5}} \\ \%H_{2_{2,1}}, & \%CH_{4_{2,2}}, & \%C_2H_{4_{2,3}}, & \%C_2H_{6_{2,4}}, & \%C_2H_{2_{2,5}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \%H_{2_{318,1}}, & \%CH_{4_{320,2}}, & \%C_2H_{4_{320,3}}, & \%C_2H_{6_{320,4}}, & \%C_2H_{2_{318,5}} \end{bmatrix}$$

After creating the Euclidean space, 64 five-dimensional cluster centres were created following the LBG algorithm so that the training data can be partitioned around the nearest cluster centres. After each LBG iteration, each cluster centroid (initial centre is the mean of X) was split into two until 64 clusters are formed. The centres having zero association with the training data points (where C_j was an empty set) were removed, and the remaining were used as initial centres in KMA. To ensure optimum performance, the 34 best cluster centres have been chosen using KMA. The selection process includes splitting the centres having large number of data associations, and removing or combining the centres having lower numbers of data associations. The association is evaluated by indexing the data points to the centres based on their distances to them. The iterative process was continued until each of the clusters become representative of a particular fault class or multiple classes with different distinguishable percentages. Moreover, there was no cluster where C_j was exceptionally large set. A summary of the probabilities of each fault category based on its association with each cluster centre is shown in Table 4.

The probability matrix of Table 4 shows that the single clustering technique could wrongly classify an unlabelled measurement, if it is close to a cluster on the boundary between multiple Voronoi regions (fault categories). The misclassification rate could increase greatly for a measurement closest to a cluster centre like μ_{17} which Voronoi region comprises 67 percent C type faulty data points and 33 percent $T3$ type faulty data points. To deal with this interference problem, the KNN algorithm has been used for the 1st, 2nd and 3rd nearest neighbouring cluster centres detection based on their distances. Moreover, the cumulative weighted voting, based on the distances from the centre has been proposed to identify a fault class of an unknown transformer. For instance, in case of three neighbours clustering approach, if the cluster centres μ_3, μ_2, and μ_{17} are the 1st, 2nd and 3rd nearest neighbours of a training data point, and their distance weight factors are 65%, 25% and 10%, respectively, of total weight then the voting process can be expressed as follows:

From Table 5, the maximum cumulative vote belongs to T3 since it has the largest column total. Consequently, the transformer will be classified as having a T3 type fault.

5 Results and discussion

This research targeted the examples that cannot be classified easily by the Duval Triangles or come with a conflicting classification due to the overlapping between different faults.

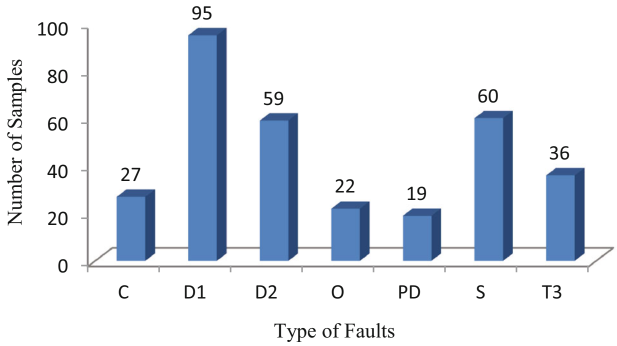

Fig. 3 Training samples following the fault categories

Therefore, a subset of 58 transformer measurement samples out of 376 have been selected as a test set and classified according to the proposed method. All 58 were excluded from the training examples. A comparison of fault classification based on modified Roger's Ratios, the IEC ratios and proposed method on 58 targeted transformer measurements has been summarized in Table 6.

The accuracy of each method shown in Table 6 has been calculated by comparing them with expert classifications. The overall accuracy of the proposed method is 93%. The method makes occasional errors but it does provide a useful decision support mechanism for engineers who are trying to deal with the critical cases. A similar type of experiment with slightly different method has been conducted by other researchers [12,15,18]. They used their own training sets (not disclosed in the literature) to train up their classifier to classify the test sets labelled by the experts in the power industry. To verify the performance of the proposed approach, it has been applied on their published test sets. In [18], researchers applied the rough sets (RS) analysis technique and artificial neural networks (ANNs) combined with RS and k-means clustering (KMC) algorithms to determine transformer fault categories. A comparison of their methods, different established methods and the proposed method is shown in Table 7.

The proposed method was also applied on the test samples available in [12] where each fault has been found using a decision tree (formed from if–else conditions) based on the gas contrition limits for particular types of fault. They considered the decision tree method as a new approach to DGA. According to this method, they found overlapping between different faults with the same gas limit used in the decision tree. To solve the overlapping problem, they included additional gas ratios in the decision tree and named it as Modified New Approach DGA. The proposed method has been compared with the decision tree method and different ratio methods to evaluate its performance. The comparison has been summarized in Table 8.

In [15], the researchers have used different artificial intelligence and machine learning techniques such as ANN, SVM, ELM, and SaE-ELM to classify the fault category of a transformer. They compared the classification of all the four methods and the decision taken from the majority votes. A comparison of their methods and the proposed method is shown in Table 9.

6 Case study and analysis

To provide a deeper understanding, all 34 literature cases have been used in a case study. The proposed method differs from the experts' diagnoses in three cases (Case number 2, 4 and 9), out of 16 shown in Table 8. According to the experts' judgements, cases 2, 4 and 9 are D2, D2 and DT faults but are classified as D1, D1 and T2 respectively, by the proposed method. Duval Triangle classified the faults as D1, D1 and DT correspondingly, but the other two conventional methods, Roger's ratio and IEC ratio, failed to diagnose these cases. As the DT fault category has been omitted from the proposed method due to the insufficient training data points, it has been classified as a T2 fault. Moreover, gas concentrations collected from the utility company predominantly used Duval Triangles to classify their transformers' fault categories which could be a reason for misclassification in cases 2 and 4, respectively. The proposed method effectively solved the overlapping problem in 7 cases out of 10 shown in Table 9. The IEC method could not detect any faults in the 8 cases shown in Table 9. Even the performance of Roger's ratio is not satisfactory, but the proposed method accurately classified all cases except for case number 6. That case is classified by the ANN, SVM, ELM, and SaE-ELM methods as a D2 fault. According to the Duval Triangle, the fault category of the sample is DT fault, which is a combination of thermal and dielectric. The proposed method has classified the case as a T2 type fault. As the data point is close to the three neighbouring cluster centres, which Voronoi regions comprise multiple types of faults with closed percentage of probabilities, it is clear that case 6 has been misclassified. This problem could be overcome by increasing the number of clusters with a much larger number of training data points.

In this research, the number of cluster centres and their positions are carefully chosen through a continuous iteration process and their performances have been tested before being applied to an unlabelled measurement. Deliberately preserved extra-dimensional information has helped to accurately classify 93 percent of the cases, which other conventional established methods could not cope with. Moreover, the proposed method performed well on the published data as shown in Tables 8, 9, respectively.

7 Conclusions

A new DGA diagnosis technique has been developed that is based on a clustering approach combined with a modified KNN cumulative voting approach that considers interneighbour distances. The experimental results show that it correctly classifies 93% of the difficult cases where Duval's triangle is unable to make a classification. This result is compared with other methods like the Rogers' Ratios and IEC method, all of which fail in a significant fraction of cases (See Tables 7, 8, 9). The focus of this work has been to develop a method that compliments Duval's widely adopted triangles method rather than replacing it. The interaction can be seen from the workflow diagram shown in Fig. 2. It can be seen that the proposed method is very suitable for incipient fault diagnosis in power transformers.

One of the weaknesses of the proposed technique is that it does not deal well with the mixture of dielectric and thermal (DT) faults. Most other systems considered also have difficulties differentiating DT from other fault categories. Further work needs to be done in this area, particularly building classifiers with large number of training examples that can deal with very uncommon faults more effectively.

References

1. Das N, Abu-Siada A, Islam S (2013) Impact of conducting materials on furan-spectral correlation of transformer oil. In: Power engineering conference (AUPEC), 2013 Australasian Universities, pp 1–4
2. Khmais Bacha, Seifeddine Souahlia, Gossa M (2012) Power transformer fault diagnosis based on dissolved gas analysis by support vector machine. Electr Power Syst Res 83:73–79
3. Abu-Siada A, Hmood S, Islam S (2013) A new fuzzy logic approach for consistent interpretation of dissolved gas-in-oil analysis. IEEE Trans Dielectr Electr Insul 20:2343–2349
4. Malik H, Azeem A, Jarial RK (2012) Application research based on modern-technology for transformer Health Index estimation. In: Systems, signals and devices (SSD), 2012 9th international multi-conference on, pp 1–7
5. Abu-Siada A, Lai SP, Islam SM (2012) A Novel fuzzy-logic approach for furan estimation in transformer oil. IEEE Trans Power deliv 27:469–474
6. Saha TK (2003) Review of modern diagnostic techniques for assessing insulation condition in aged transformers. IEEE Trans Dielectr Electr Insul 10:903–917
7. Ashkezari AD, Saha TK, Ekanayake C, Ma H (2011) Evaluating the accuracy of different DGA techniques for improving the transformer oil quality interpretation. In: Universities power engineering conference (AUPEC), 2011 21st Australasian, pp 1–6
8. Bakar NA, Abu-Siada A, Islam S (2014) A review of dissolved gas analysis measurement and interpretation techniques. IEEE Electr Insul Mag 30:39–49
9. Hettiwatte SN, Fonseka HA (2012) Analysis and interpretation of dissolved gases in transformer oil: a case study. In: Condition monitoring and diagnosis (CMD), 2012 international conference on, pp 35–38
10. Singh S, Bandyopadhyay MN (2010) Dissolved gas analysis technique for incipient fault diagnosis in power transformers: A bibliographic survey. IEEE Electr Insul Mag 26:41–46
11. IEC, 60599 Ed. 2.1 (1999) Mineral oil-impregnated electrical equipment in service-guide to the interpretation of dissolved and free gases analysis
12. Ghoneim SSM, Taha IBM (2016) A new approach of DGA interpretation technique for transformer fault diagnosis. Int J Elect Energy Syst 81:265–274
13. (2008) IEEE guide for the interpretation of gases generated in oil-immersed transformers. IEEE STD. C57.104
14. Zakaria, F, Johari D, Musirin I (2012) Artificial neural network (ANN) application in dissolved gas analysis (DGA) methods for the detection of incipient faults in oil-filled power transformer. In: Control system, computing and engineering (ICCSCE), 2012 IEEE international conference on, pp 328–332
15. Li S, Wu G, Gao B, Hao C, Xin D, Yin X (2016) Interpretation of DGA for transformer fault diagnosis with complementary SaE-ELM and arctangent transform. IEEE Trans Dielectr Electr Insul 23:586–595
16. Huang YC, Sun HC (2013) Dissolved gas analysis of mineral oil for power transformer fault diagnosis using fuzzy logic. IEEE Trans Dielectr Electr Insul 20:974–981
17. Abu-Siada, A, Arshad, M, Islam S (2010) Fuzzy logic approach to identify transformer criticality using dissolved gas analysis. In: IEEE PES general meeting, pp 1–5
18. Sheng-wei Fei YH (2012) A multi-layer KMC-RS-SVM classifier and DGA for fault diagnosis of power transformer. in: R.P.o.C. Science (ed), pp 238–243
19. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY (2002) An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Anal Mach Intell 24:881–892
20. Yin M, Hu Y, Yang F, Li X, Gu W (2011) A novel hybrid K-harmonic means and gravitational search algorithm approach for clustering. Expert Syst Appl 38:9319–9324
21. Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recognit Lett 31:651–666
22. Niknam T, Taherian Fard E, Pourjafarian N (2011) An efficient hybrid algorithm based on modified imperialist competitive algorithm and K-means for data clustering. Eng Appl Artif Intell 24:306–317
23. Kwedlo W (2011) A clustering method combining differential evolution with the K-means algorithm. Pattern Recognit Lett 32:1613–1621
24. Li Y, Wu H (2012) A clustering method based on k-means algorithm. Phys Procedia 25:1104–1109
25. Kekre HB, Sarode MTK (2009) Vector quantized codebook optimization using k-means. Int J Comput Sci Eng 1:283–290