K-nearest neighbor method for power transformers condition assessment

D V Tanfilyeva*, O V Tanfyev and Y V Kazantsev
Power Engineering Department, Novosibirsk State Technical University, 20, Prospekt K. Marksa, Novosibirsk, 630073, Russia
* E-mail: tanfileva@corp.nstu.ru

Abstract. Reliable power supply is the most crucial task of every power utility company making power transformers one of their chief assets. Thus, detection of power transformers abnormal condition is of high importance. Most general and approved tool of condition assessment is Dissolved Gas Analysis (DGA) of transformer insulation liquids. The identification task in fact implies conventional pattern recognition of measured parameters and their classification. The majority of prevalent classification and recognition methods require a priori knowledge of classes and symptoms, probability distribution laws and density functions etc. This paper describes application of two fault recognition methods – Bayesian classifier and k-nearest neighbours (KNN) algorithm. Studies have shown that KNN tool allows flexibility for concentration limits adjustment with operation conditions’ altering and has a high condition classifying accuracy.

1. Introduction
State-of-the-art monitoring and diagnostic systems help to determine the condition of utility companies equipment with some scheduled regularity and thus to extend its lifetime. All such expert systems have their own assessment and decision-making algorithms based on some standard or guideline [1-3] and therefore the task is narrowed to state parameter measuring and comparison with boundary conditions. Such approach can be adapted to the conventional test methods (short-circuit losses, dielectric loss tangent, capacity measuring) for new equipment or units without operation history and routine/major repairs records. Diagnostic methods like thermo-vision examination, Dissolved Gas Analysis (DGA)[4], partial discharge monitoring proved to be highly-informative and having regulated assessment criteria, provide easy statistic data collection, monitoring of various operational factors’ influence, and determining condition and health index[5]. Combination of diagnostic methods, lifespan and maintenance results enable power utilities to determine the number and extent of repairing activities with minimal risk of false incipient fault diagnosis.

The DGA diagnostic tools are included in national and international guides[1] as approved examination methods of faulty oil-filled equipment, allowing to detect incipient faults, such as discharge, overheating, discharge and overheating.

However, Key Gas Method is not yet perfect as the threshold gas concentrations might be revised for transformers with representative DGA sample data by plotting the distribution function[1,2].

According to IEC 60599 to classify fault type the key gases concentrations ratios have to be calculated. Ratios’ ranges corresponding to each type of abnormality are not updated, depending on life cycle, load and climatic conditions, which is the key gases major weakness.
Therefore, in diagnostic and monitoring systems the development of universal technique for equipment condition assessment is of high necessity.

2. Equipment condition classification methods

Nowadays the methods of pattern recognition have taken the shape of sole, solid mathematical theory. This section presents three groups of condition classification methods:

- probabilistic;
- statistical (without distribution parameters evaluation);
- machine learning.

Above-mentioned techniques can be used for limited number of tasks and particular database.

Thus, probabilistic methods (Bayesian and linear classifiers) are based on derivation of distribution laws of condition sets, distribution parameters and probabilities of belonging to i-th sets. The simplest form of a Bayesian decision boundary \[5,6,11,12\] may be derived if distribution law of i-th condition sets is normal. Multiparameter (log-normal, gamma-distribution, etc.) distribution laws complicate the mathematical expression of classifier. A further difficulty concerns the classifier derivation with large number of measured parameters. DGA tool implies 7 gases’ concentrations measuring and their further analysis after additional converting, filtering, centering and decorrelation in classifier [6, 7].

Generally, every classifier before being implemented to a diagnostic system, should be graphically interpreted, which enables presentation of equipment’ condition ranges and the present state. However, due to human perception visualization is limited to three dimensions. Still the Bayesian classifier can operate any number of classes and input parameters, but the task of condition parameter reduction arises, which can be solved by picking the most in-formative ones. The solution is factor analysis or non-linear transformation [8]. Despite of high accuracy of current condition assessment, the analytical de-termination of threshold is quite complicated. The use of probabilistic techniques requires sophisticated (time-consuming) preliminary statistical processing of input data.

The second group of statistic methods is the one without distribution parameters assessment: histogram method, k-nearest neighbors (KNN)[9,10], Parzen assessment, basis function decomposition, etc. The core strength of these is the fact that threshold derivation needs no validation of condition classes’ inclusion to certain distribution law. Besides, no numerical computation is required.

The machine learning group would comprise: neural networks[12-16], support vector machines[17], fuzzy-logic[18,19]. The benefits of this group are n-dimension space of conditions, classification of linearly separable and linearly non-separable condition classes. The apparent advantage is the opportunity to adjust the thresholds with operation conditions’ altering or with incipient fault detection in similar units during their maintenance. Such algorithms are software-friendly and have high computational speed. Nevertheless the limiting factors are input information quality, namely, its accuracy, sample ejections, homogeneity and preliminary class separability[8].

This paper reports on comparison of two methods: the Bayes Classifier and the KNN method (based on DGA) for faults classification in the group of 110 kV power transformers.

3. Bayesian classifier for DGA power transformer condition assessment

In a sense of false detection minimization the Bayesian classifier [6,7] offers the best solution and is easily applicable for low-dimensional indicator space (better analysis and visualization) in case of well-described classes and belonging probabilities. Bayesian decision rule, minimizing solution error for two classes \(C_1, C_2\), is:

\[
h(X) = -ln l(X) = -ln p(X / C_1) + ln p(X / C_2) \geq ln \frac{p(C_1)}{p(C_2)} \rightarrow X \in \{C_1, < C_2\}
\]
where $P(C_i)$ - is a priori belonging to class probabilities, $P(X / C_i)$ - conditional probability density, showing belonging of object to $i$- class.

In some cases classes could be expressed by normal law with probability density $p(X)$:

$$p(X) = N(X; M, \Sigma) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2} d^2(N, X; M, \Sigma)\right\},$$

where $N(X; M, \Sigma)$ - reduced notation of normal law with expected value vector $M$ and covariance matrix $\Sigma$.

To classify transformer abnormality types, two gas concentrations ratios are used as input parameters: $X \in \left[ x_1, x_2 \right] = \left[ \frac{C_2H_2}{C_2H_4}, \frac{CH_4}{H_2} \right]$. The research includes 311 DGA reports, 61, 130, 120 of which are discharge, overheat, overheat and discharge failure reports correspondingly.

Figure 1 summarizes areas of three fault classes, and it can be seen that they’re linear non-separable. Therefore, there is a considerable uncertainty in DGA results interpretation.

![Figure 1. Visualization of diagnostic indicators in condition space according to guide.](image)

To form separable classes, the non-linear transformation (logarithmic) of original gas concentration ratios is performed:

$$
\varphi(X) = \begin{bmatrix}
\ln\left(\frac{C_2H_2}{C_2H_4}\right) \\
\ln\left(\frac{CH_4}{H_2}\right)
\end{bmatrix}
$$

Taking logarithms allows expanding of points regions belonging to $i$-th classes thus increasing their separation, which enables Bayesian decision rule classification.

The expression of condition classes limits with 2-dimensional indicator distribution was derived in a form of canonical 2nd order curve equation:
\[ Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F \geq 0, \]  

(3)

with the following coefficients:

\[
A = \frac{e_{22} - q_{22}}{2\Delta_2}, \quad B = \frac{e_{12} - q_{12}}{\Delta_1}, \quad C = \frac{e_{11} - q_{11}}{2\Delta_1},
\]

\[
D = \frac{m_{12}q_{22} + m_{22}q_{12} - m_{11}e_{11} + m_{12}e_{12}}{\Delta_2},
\]

\[
E = \frac{m_{12}q_{11} + m_{22}q_{12} - m_{11}e_{11} + m_{12}e_{12}}{\Delta_1},
\]

\[
F = \frac{e_{12}m_{11}^2 + e_{22}m_{21}^2 + m_{11}m_{21}(e_{12} + e_{22}) - q_{11}m_{11}^2 + q_{22}m_{21}^2 + m_{11}m_{21}(q_{12} + q_{22}) + \frac{1}{2} \ln \frac{\Delta_1}{\Delta_2} - \ln \frac{P_1}{P_2}}{2\Delta_2}
\]

where \( M = \begin{bmatrix} m_{11} & m_{12} \\ m_{11} & m_{21} \end{bmatrix} \) expected value matrix \( C_i \) classes, \( \Sigma_1 = \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix} \), \( \Sigma_2 = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \) covariance matrix for first and second classes.

This 2nd order curve (3) may be ellipse, hyperbole or parabola, or pair of lines (parallel, intersecting or coincidental).

Figure 2. Visualization of diagnostic indicators in condition space according to guide. Derived Bayesian reasoning rule (Figure 2) is applicable for two classes’ separation, whereas defect type recognition includes 3 condition classes. Thus, this task should be solved successively: pick out 1 class, combine 2nd and 3rd, plot first classes threshold (repeat for 2nd and 3rd). Picking and combining of training classes is random, depending on conformity of resulting reasoning rule. Total recognition error with rule (3) is around 12.86%.
4. KNN approach implementation in DGA assessment of power transformers condition

Transformer fault type identification algorithm includes: development of measured gases concentrations database, their ratios computing \( \frac{C_2H_2}{C_2H_4} ; \frac{CH_4}{H_2} \); classification to three groups according to [1]. There are no requirements to the sample set size in first iteration for decision-making threshold and the major limitation is the verification of diagnosis while repair or with the help of additional diagnostic tools.

During database extension with new measured parameters vector (belonging of which to \( i \)-th class is unknown), the evaluation of its relative position to other “historical” classes is conducted. Position of new vector is assessed by identifying the distance to \( k \)-nearest neighbors. It’s recommended to choose an odd number of elements \( (k= 3, 5, 7, \ldots 19) \), less than 20 [8]. The next step is the selection of metric (q.e. analytical expression of distance from the currently classified indicator vector to the \( k \)-nearest ones), tabulated in table 1. The condition assessment follows the next rule: the current vector is related to the class, most of \( k \)-nearest neighbor vectors are.

The operating experience shows that one or two gases concentration excess might not always be the evidence of incipient fault and later samples can show no indication of defect. According to standard methods, such sample cannot be neglected and the transformer is supervised, which leads to extra costs of lab oil analysis. KNN method should reduce the number of samples with one or two key gases limits excess.

| Table 1. Major metrics of KNN method. |
|--------------------------------------|
| Metric                              | Expression                                           |
| 1 Euclidean distance is the metric  | \( D_E = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \),     |
| calculated with the help of         | where \( x_i, y_i \) – vector coordinates             |
| Pythagorean theorem                 |                                                       |
| 2 C- Euclidean metric is standard    | \( D_{SE}(\bar{x}, \bar{y}) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{\sigma_i^2}} \), |
| Euclidean distance, where the input  | where \( \sigma_i \) - standard deviation of \( x_i, y_i \) |
| data are normalized with their       |                                                       |
| standard deviation                  |                                                       |
| 3 Citiblok                          | \( D_S = \sum_{i=1}^{n} |p_i - q_i| \),                                      |
| where \( n \) - dimension of space, | where \( p_i, q_i \) – compared vectors.            |
| \( p_i, q_i \) – compared vectors. |                                                       |
| 4 Mahalanobis distance – generalized | \( D_M(x) = \sqrt{(x - \bar{x})^T S^{-1} (x - \bar{x})} \) |
| Euclidean distance.                  | where \( x \) – multidimensional vector \( [x_1, x_2, \ldots, x_n]^T \) |
| \( \bar{x} \) – set with mean value \( [\mu_1, \mu_2, \ldots, \mu_n]^T \) |
| \( S \) – covariance matrix         |                                                       |

Consider condition assessment study of two 31.5 MVA transformers 1T and 2T of “ZAO RES” power utility company by means of KNN tool.

DGA results were used as the original vector of indicator space, namely, relative and absolute values of gas concentrations in transformers of Novosibirsk region electrical grid of (1999 - 2016) time period. This data set contains 1340 samples of 7 gases each. There’s no use of such amount of training information and only 180 samples (90 reports for each class) are used. Report data
arrangement is random (q.e. DGA results with preliminary classification according to guide [1,2] are submitted chaotically) to enhance the quality of recognition [8]. Recognition results are verified with DGA data of two transformers. DGA results of one substation are tabulated in tab.2. It is important to note that the first sample of table 2 (of 31.5 MVA transformer 1T) presents the case, when the transformer is supervised because of one gas limit excess, but the further inspection showed no fault.

![Figure 3](image_url)

**Figure 3.** Classification of measured concentrations vectors as the thermal fault with Euclidian distance metric and \( k=13 \).

The results of the above calculations show that the most precise metric is citiblok \((k=13)\), similar with Euclidean distance \((k=11)\). As indicated in table 2 the sole threshold concentration of hydrogen puts the whole sample vector to “no-fault” class. For 2T the fault state is proved by 2 methods. Graphical interpretation of KNN fault classification method is shown in figure 3.

The ratio of two gas concentrations is adopted as an input space of indicators. Since, this space of indicators is non-separable, the nonlinear transformation of input data is required and the natural logarithm is the optimal one.

Test report of 2T gas concentrations on 03.02.2006, with thermal fault predicted was chosen as the study case consistent with guide [1].

| Date       | 1T | 2T | CH₄/H₂ | CH₄ | C₂H₄ | C₂H₆ | CO₂ | CO | Result       |
|------------|----|----|--------|-----|------|------|-----|---|--------------|
| 18.01.06   |    | 1985 | 0.020  | 0.187| 0.040| 0.10 | 0.156| 0.230| Fault 76% / 23% |
| 01.02.06   | 0.178| 2983 | 0.947  | 9.916| 0.98 | 0.363| 0.100| Fault 0% / 100% |
| 01.02.06   | 0.629| 0.056| 0.233  | 0.124| 0.06 | 0.187| 0.250| Norm 100% / 0%   |
| 03.02.06   | 0.154| 1591 | 8.014  | 6.790| 0.43 | 0.351| 0.150| Fault 100% / 0%  |
| 07.07.06   | 0.009| 0.009| 0.010  | 0.002| 0.00 | 0.023| 0.016| Norm 100% / 0%   |
| 10.07.06   | 0.008| 0.009| 0.010  | 0.002| 0.00 | 0.022| 0.016| Norm 100% / 0%   |
| 14.07.06   | 0.004| 0.014| 0.010  | 0.002| 0.00 | 0.026| 0.016| Norm 100% / 0%   |
| 20.07.06   | 0.001| 0.005| 0.010  | 0.000| 0.00 | 0.025| 0.016| Norm 100% / 0%   |
| 25.07.06   | 0.207| 0.642| 6.497  | 6.488| 0.55 | 0.573| 0.183| Fault 100% / 0%  |
Figure 3 shows that the measured gas concentrations vector is correctly classified as “Discharge and overheat” with the help of KNN tool. The study revealed, that after k=13, recognition accuracy and quality keeps the same level.

5. Conclusions
The paper presents the KNN approach to transformer condition identification, which could be implemented in monitoring and diagnostic systems of any oil-filled equipment. The major advantages of this method are the opportunity to adjust the thresholds with operation conditions altering, high accuracy and the ease of software implementation.

References
[1] IEEE Std C57.104-2008 “IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers” 2008.
[2] Mineral oil-impregnated electrical equipment in service - Guide to the interpretation of dissolved and free gases analysis 1999 IEC 60599 Ed. 2.1, 1999
[3] Duval M and de Pabla A 2001 Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases IEEE Electr. Insul. Mag. 17(2) 31-41.
[4] Duval M 1989 Dissolved gas analysis. It can save your transformer IEEE Electr. Insul. Mag. 5(6) 22-7.
[5] Jahromi A, Piercy R, Cress S, Service J and Wang Fan 2009 An approach to power transformer asset management using health index IEEE Electr. Insul. Mag. 25(2) 20-34.
[6] Fukunaga K 1990 Statistical Pattern Recognition (New York: Academic Press, 2nd edition).
[7] Theodoridis S, Koutroubas K 2003 Pattern Recognition (Academic Press Elsevier).
[8] Haykin S 1999 Neural Networks A Comprehensive Foundation (Prentice Hall, Upper Saddle River, New Jersey 07458 2nd edition).
[9] Benmahamed Y, Teguar M and Boubakeur A 2017 Application of SVM and KNN to Duval Pentagon I Transformer Oil Diagnosis IEEE Trans. Elec. Insul. 24(6) 3443-51.
[10] Md Mominul Islam, Lee G and Hettiwatte S N 2016 Incipient fault diagnosis in power transformers by clustering and adapted KNN Australasian Universities Power Engineering Conf. (AUPEC) 25-28 Sept. 2016 Brisbane, QLD, Australia
[11] Fang-yu Ren, Shu-bin Si, Zhi-qiang Cai and Shuai Zhang 2015 Transformer fault analysis based on Bayesian networks and importance measures Journal of Shanghai Jiaotong University (Science) 20 353.
[12] Jiejie Dai, Hui Song, Gehao Sheng and Xiuchen Jiang 2017 Dissolved gas analysis of insulating oil for power transformer fault diagnosis with deep belief network IEEE Trans. Dielectr. Electr. Insul 24(5) 2828-35.
[13] Barbosa F R, Almeida O M, Braga A P S, Amora M A B and Cartaxo S J M 2012 Application of an artificial neural network in the use of physicochemical properties as a low cost proxy of power transformers DGA data IEEE Trans. Dielectr. Electr. Insul 19(1) 239-46.
[14] Miranda V, Castro A R G and Lima S 2012 Diagnosing faults in power transformers with autoassociative neural networks and mean shift IEEE Trans. Power Del. 27(3) 1350-7.
[15] Guardado J L, Naredo J L, Moreno P and Fuerte C R 2001 A Comparative Study of Neural Network Efficiency in Power Transformers Diagnosis Using Dissolved Gas Analysis IEEE Trans. Power Del. 16(4) 643-7.
[16] Carita A J Q, Leita L C, Junior A P P M, Godoy R B and Sauer L 2013 Bayesian networks applied to failure diagnosis in power transformer IEEE Latin America Trans. 11(4) 1705-82.
[17] Li J Z, Zhang Q G, Wang K, Wang J Y, Zhou T C and Zhang Y Y 2016 Optimal dissolved gas ratios selected by genetic algorithm for power transformer fault diagnosis based on support vector machine IEEE Trans. Dielectr. Electr. Insul. 23(2) 1198-1206.
[18] Khan S A, Equbal M D and Islam T 2015 A comprehensive comparative study of DGA based transformer fault diagnosis using fuzzy logic and ANFIS models IEEE Trans. Dielectr. Electr. Insul. 22(1) 590-6.

[19] Mofizul Islam S, Wu T, Ledwich G 2000 A Novel Fuzzy Logic Approach to Transformer Fault Diagnosis, IEEE Trans. Electr. Insul. 2 177.