Research on Condition Evaluation Algorithm of Oil-immersed Transformer Based on Naive Bayes

Mingchao Yong\textsuperscript{1a,*}, Jianmin Xue\textsuperscript{1b}, Weijie Wang\textsuperscript{1c}, Bangtian Wang\textsuperscript{2a}, Lidong Lv\textsuperscript{1a}, Qingshan Wang\textsuperscript{1d}, Hongguang Shi\textsuperscript{1e}, Bingxin Wu\textsuperscript{1f}, Bogen Chen\textsuperscript{2b}

\textsuperscript{1}Xuchang XuJi Software Technology Co., Ltd., Xuchang 461000, China
\textsuperscript{2}Fuzhou Tianyu Electric Co., Ltd., Fuzhou 350012, China
\textsuperscript{3}School of Electrical & Information Engineering, Anhui University of Technology, Ma’anshan 230009, China

\textsuperscript{1a,*}ymc0112@163.com, \textsuperscript{1b}xjc_xuejianmin@126.com, \textsuperscript{1c}xjc_wangweijie@126.com, \textsuperscript{2a}13703741338@139.com, \textsuperscript{3a}13813936243@139.com, \textsuperscript{1d}xjc_wangqingshan@126.com, \textsuperscript{1e}xjc_shihongguang@126.com, \textsuperscript{1f}wubingxincc@163.com, \textsuperscript{2b}13559148186@139.com

Abstract—The improved three ratio method is a general algorithm for condition evaluation of oil immersed transformer. It has high accuracy in pre-test and laboratory analysis. However, the accuracy of DGA online monitoring data is not high, resulting in the decline of the positive judgment rate of transformer condition evaluation using the improved three ratio method, which is difficult to support the development requirements of transformer intelligence. To solve this problem, a DGA state evaluation method based on Naive Bayesian algorithm is proposed. The algorithm first performs preprocessing such as median filtering on the DGA online monitoring data to remove invalid data, then uses a triple composed of three conditional attributes to describe the characteristics of the DGA data, and finally calculates the a priori probability of training samples and the a posteriori probability of test samples by naive Bayesian algorithm for state evaluation. The verification of the algorithm on the measured data set shows that the accuracy of the algorithm is better than the improved three ratio method, and the algorithm is feasible and effective.

1. Introduction

Oil-immersed power transformer is the main component of the power system operation, and its safe and stable operation is an important condition to ensure the stability of the power system. Under the influence of electricity, heat and other factors, the oil-paper insulation material in transformer operation will undergo aging and cracking, resulting in low-molecular hydrocarbons and CO\textsubscript{2}, CO and other gases dissolved in the oil. Therefore, the monitoring and analysis of the dissolved gas in the oil can indirectly reflect the insulation fault of the transformer. Since the construction of online transformer monitoring in China, Dissolved Gas Analysis (DGA) has been the main content of online transformer monitoring and has actively carried out research based on faulty gases\textsuperscript{[1-3]}. According to statistics, among the transformers that failed, overheating faults accounted for 63\%, high-energy discharge faults accounted for 18.1\%, overheating and high-energy discharge accounted for 10\%, spark discharge accounted for 7\%, and damp faults accounted for 1.9\%\textsuperscript{[1]}. The fault analysis based on DGA data can identify the above-mentioned faults relatively accurately. Based on this series of guidelines and specifications issued by the State Grid Corporation of China, DGA is determined as a...
necessary configuration for a new 110 kV substation. DGA monitoring data and on-site diagnosis results are important reference opinions for maintenance, which puts forward higher requirements for the accuracy of DGA-based transformer fault diagnosis algorithms.

The DGA-based transformer condition assessment algorithm generally uses rule-based methods such as the cube diagram method, the electrical research method, the four-ratio method, and the David triangle method [4-5]. In particular, the "improved three-ratio" method proposed by Chinese scholars such as Cao Dunkui after a large number of experimental statistics has become the industry standard for transformer condition assessment based on DGA data in China. However, the rule-based diagnosis method has high requirements for the integrity and accuracy of the monitoring data. The boundary data processing is relatively difficult and the lack of monitoring data is not allowed.

In recent years, some experts have used intelligent algorithms to conduct research on transformer condition assessment based on DGA data. Literature [6-7] uses fuzzy theory and rough sets to evaluate the state of transformers. This method has the advantages of simple structure and fast diagnosis, but its learning ability is poor. Literature [8-10] describes the application of neural network algorithm in transformer state assessment. The advantage of this algorithm is strong self-learning ability and parallel processing ability, but the convergence speed is slow. Bayesian algorithm has good applications in distributed power grid fault diagnosis, power quality monitoring, distribution network reliability diagnosis, voltage transformer reliability analysis, comprehensive transformer diagnosis, and mechanical system reliability analysis [11-13]. The Bayesian algorithm obtains the posterior probability of the event according to the prior probability of the event, and then classifies the event according to the magnitude of the posterior probability. Bayesian algorithm has the characteristics of simple structure, self-learning and fast calculation speed. Based on this, a naive Bayesian transformer condition assessment algorithm based on DGA online monitoring data is proposed. The algorithm preprocesses the data set based on the monitoring data of DGA, uses triples composed of three conditional attributes to describe the characteristics of DGA data, and forms training samples and test samples. The naive Bayes algorithm is used to calculate the prior probability of the training sample and the posterior probability of the test sample to perform state evaluation.

2. Transformer diagnosis algorithm based on DGA

2.1. Relationship between the internal fault of the transformer and the gas content in the oil

At present, DGA online monitoring usually adopts the technical principle of gas chromatography or photoacoustic spectroscopy to finally analyze the fault gases such as hydrogen (H₂), methane (CH₄), acetylene (C₂H₂), ethylene (C₂H₄), ethane (C₂H₆), carbon monoxide (CO), carbon dioxide (CO₂) and other malfunctioning gases. The relationship between the type and content of gas is closely related to the fault type of transformer, and the fault type can be reflected indirectly according to the fault gas. The improved three-ratio method is the most widely used algorithm in the ratio method. According to the literature [1], the statistical analysis of 671 transformers shows that the positive judgment rate can reach 97.1% under laboratory conditions. China's domestic industry standards and national standards also use the improved three-ratio method as the main method for evaluating the internal state of transformers. However, the accuracy of online monitoring data is quite different from that of laboratory data. For example, literature [14] stipulates that the accuracy error of monitoring is 30%, which meets the accuracy requirements. If the online monitoring data with this accuracy is used, the positive judgment rate of diagnosis will be reduced, and the evaluation results are difficult to support the development requirements of transformer intelligence. Therefore, it is of great significance to study more effective algorithms in line with the development direction of transformer intelligence [15].

2.2. Naive Bayes Classifier

The Bayesian classification method is based on statistics in mathematics. According to the existing sample data examples, the Bayesian formula is used to determine the classification problem of the data to be tested. The main idea of Bayesian method is to link the prior probability and posterior probability...
of the event, and use the prior information to predict the posterior probability of the event. The Bayesian classifier has the ability of self-learning, which can add new data records to the known prior information, thereby further affecting the posterior probability of the event and improving the accuracy of event prediction. The Bayesian formula is shown in (1):

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)}$$ (1)

Where $P(C|X)$ is the posterior probability of $C$ under condition $X$, $P(C)$ is the prior probability of $C$, $P(X|C)$ is the posterior probability under condition $C$, and $P(X)$ is the prior probability of $X$.

Let $A$ represent the condition attribute variable set and $C$ represent the fault type variable. Suppose there are $n$ conditional attribute variables, $Val(A_i)$ ($1<i<m$) represents the value range of conditional attribute variables, and $Val(C)$ represents the value range of fault type variables, then $A=<A_1,A_2,...,A_n>$, the value of $A_i$ is $a_i \in Val(A_i)$, and the value of $C \in Val(C)$. Let $t$ represent training samples, $t=<a_1,a_2,...,a_n,c_t>$, $1<i<m$ represent training sample instances, $x$ represent test samples, $x=<a_1,a_2,...,a_j,...,a_n>$ indicates the test sample instance.

The posterior probability formula of sample $x$ belonging to class $c_j$ is shown in (2):

$$P(c_j|a_1,a_2,...,a_n) = \frac{P(a_1,a_2,...,a_n|c_j) \times P(c_j)}{P(a_1,a_2,...,a_n)} = \alpha \times P(c_j) \times P(a_1,a_2,...,a_n|c_j)$$ (2)

Among them, $P(c_j|a_1,a_2,...,a_n)$ is the posterior probability of the test sample classified into the fault type $c_j$; $P(a_1,a_2,...,a_n|c_j)$ is the conditional probability of the fault type $c_j$; $P(c_j)$ is the prior probability of the fault type $c_j$; $\alpha = 1/P(a_1,a_2,...,a_n)$, $\alpha$ is a constant.

The Bayesian algorithm calculates the posterior probabilities of all classes $c_j$ ($j=1, 2, ..., m$) according to formula (2), and then classifies the group of test sample instances into the class with the largest posterior probability. The Bayesian classifier combines causality and probability knowledge, conforms to people's normal thinking, has good results, and has a wide range of applications in practice. The naive Bayes classifier is the earliest Bayes classifier, based on the conditional independence of each attribute variable, that is, it is assumed that the probability of a certain attribute value is completely unaffected by other attributes. Due to the mutual independence between attributes, the calculation formula of posterior probability in formula (1) can be transformed into formula (3):

$$P(c_j|a_1,a_2,...,a_n) = \alpha \times P(c_j) \times \prod_{i=1}^{n} P(a_i|c_j)$$ (3)

The Bayesian classifier has a good performance in describing the polymorphism of system events and the non-determinism of the logical relationship between events, and can perform forward and backward probabilistic inferences, that is, it can derive the result from the cause, and the cause can be derived from the result. At the same time, the Bayesian classifier can directly obtain the failure rate of the system, the posterior probability of the root node, etc. Its superior state description and reasoning mechanism makes the Bayesian classifier widely used in the field of reliability analysis of complex systems [6-9].

3. DGA diagnosis algorithm based on Naive Bayes

3.1. Feature set extraction

Naive Bayesian operations need to construct feature data sets. The three ratio rules of the improved three-ratio method are representative in the ratio method, and the specific values of each attribute variable of the Naive Bayes feature set are also taken to these three ratios.

The algorithm feature set is represented by a triple $V$:

$$V = \{C_1H_2 / C_2H_4, CH_3 / H_2 , C_1H_4 / C_2H_8 \}$$ (4)

3.2. Principles of State Assessment Algorithms

The algorithm first collects the relevant information of the dissolved gas, and performs data processing
to obtain the feature set required by the algorithm, and form a training sample set and a test sample set. The training sample set is \( T = \langle A, C \rangle \), where \( A \) represents the conditional attribute set, that is, \( A = \{A_1, A_2, A_3\}, Val(A_1) = Val(A_2) = Val(A_3) = \{0, 1, 2\} \), \( C \) represents the type attribute, \( Val(C) = \{\text{low energy discharge, high energy discharge, low energy discharge and overheating, low temperature overheating, medium temperature overheating, high temperature overheating, high energy discharge and high temperature overheating}\} \). The type attribute format in the test sample set is consistent with the training sample set, but the fault type \( C \) in the test sample set is empty. The algorithm flow chart is shown in Figure 1:

4. Test Results and Discussions

4.1. Example data set
The calculation example data is collected from the DGA online monitoring data of the oil-immersed power transformer in actual operation, and the transformer covers the voltage level of 110kV to 500kV. The sample data set contains 525 complete samples. The distribution of fault types of the data samples is shown in Table 1.

| Fault type                                      | Data set sample size |
|------------------------------------------------|----------------------|
| low energy discharge                           | 137                  |
| low energy discharge and overheating            | 29                   |
| high energy discharge                           | 91                   |
| low temperature overheating                     | 4                    |
| medium temperature overheating                  | 33                   |
| high temperature overheating                    | 192                  |
| high energy discharge and high temperature overheating | 39                   |
4.2. Calculation example results and analysis
Case based reasoning the characteristic indexes of the classifier in the case retrieval process select
seven characteristic gases: H₂, CH₄, C₂H₆, C₂H₄, C₂H₂, CO₂ and CO. The search algorithm adopts the
case retrieval algorithm based on Pearson product moment correlation coefficient. The Pearson
product-moment correlation coefficient between the source case X and the target case Y can be
expressed as formula (5):

\[
 r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}
\]

Where \(X_i\) and \(Y_i\) represent the ith characteristic index of examples \(x\) and \(y\) respectively. The \(\bar{X}\) and \(\bar{Y}\) respectively represent the mean values of the examples \(x\) and \(y\).

The correlation coefficient \(r\) is used to describe the correlation degree between examples. It is a
number between -1 ~ 1. The larger the value of \(r\), the closer the target example is to the source
example. Taking 1 indicates that there is a linear relationship between variables \(X\) and \(Y\), that is, \(Y\)
increases with the increase of \(X\), and all points fall on a straight line. When -1 is taken, all points fall
on a straight line, but the variable \(Y\) decreases with the increase of \(X\). A correlation coefficient of 0
indicates that there is no linear correlation between variables.

Figure 2 shows the test results of this algorithm and the three-ratio method on the data set.

![Figure 2](image)

(a) The first ten-fold cross-validation result      (b) The second ten-fold cross-validation result
(c) The third ten-fold cross-validation result              (d) The average result

Fig 2 Comparison results of naive Bayes algorithm and improved three-ratio method
The above verification results show that:
1) The positive judgment rate of Naive Bayes is significantly better than the improved three-ratio
   method in the 6 groups, and the positive judgment rate of the two is equivalent only in the low
temperature and overheating fault types.
2) The DGA diagnosis algorithm based on Naive Bayes is feasible and effective.
5. Conclusion

In view of the high error in the current DGA online monitoring, the commonly used improved three-ratio method has a low positive judgment rate. This paper proposes a DGA diagnosis algorithm based on Naive Bayes based on a statistical method. The algorithm first uses triples to represent data features, and then uses Bayesian formula to classify the sample to be tested. Numerical example experiments show that the overall positive judgment rate of the algorithm is better than the improved three-ratio method, which can be applied to the requirements of DGA online monitoring on-site early warning and graded diagnosis, and has great engineering practical value.

Acknowledgments
This work was financially supported by the R&D project of XuJi Group "Development of 220/110kV Transformer Intelligent Terminal " (B492CH213013).

References

[1] CAO, D. (2005) Analysis and diagnosis of gases dissolved in transformer oil and fault check. Chinese Power Publishing House, Beijing.
[2] CAO, D. (2008) Transformer operation and maintenance and fault analysis. China Electric Power Press, Beijing.
[3] ZHAO F, LI S. (2015) Fault Diagnosis for Traction Transformer Based on DGA and Improved Grey Correlation Analysis Model. J. High Voltage Apparatus., 51(01):41-45.
[4] Yan C. (2006) Matters needing attention in judging transformer faults by using dissolved gas analysis method in oil. J. Transformer., 43(9):38-41.
[5] YANG T, LI J, ZENG X, et al. (2009) Large Power Transformer Fault Diagnosis Based on a Diagnosis Model Combined with Several Methods. J. Automation of Electric Power Systems., 33(20):92-95.
[6] WU Z, YANG J, ZHU Y, et al. (2010) Power transformer fault diagnosis based on rough set theory and support vector machines. J. Power System Protection and Control., 38(18):80-83.
[7] GONG M, ZHANG Y, LIU Y, et al. (2015) Fault diagnosis of power transformers based on back propagation algorithm evolving fuzzy Petri nets. J. Power System Protection and Control., 43(3):113-117.
[8] CHENG S, CHENG X, YANG L. (2014) Application of wavelet neural network with improved particle swarm optimization algorithm in power transformer fault diagnosis. J. Power System Protection and Control., 42(19):37-42.
[9] ZHANG R, HAN C, LI X. (2013) Transformer fault diagnosis method based on IPSO-BP neural network. J. Computer Systems & Applications., 22(4):125-128.
[10] CHENG J, XIONG W, DUAN Z, et al. (2010) Application of Gray Neural Network to Fault Diagnosis of Transformer. J. High Voltage Apparatus., 46(08):0-56.
[11] WANG Y, LYU F, LI H. (2006) Synthetic fault diagnosis method of power transformer based on rough set theory and bayesian network. J. Proceedings of the CSEE., 26(8):137-141.
[12] ZHU Y, WU L, LI X. (2005) Synthesized diagnosis on transformer faults based on bayesian classifier and rough set. J. Proceedings of the CSEE., 25(10):159-165.
[13] Li J, Hu N, Qin G, et al. (2003) Bayesian network and its application for device fault diagnosis. J. China Mechanical Engineering., 14(10):896-900.
[14] GB 7252 — 2001 Guide for the analysis and diagnosis of gases dissolved in transformer oil. S. 2001.
[15] ZHANG W, YUAN J, WANG S, et al. (2015) A calculation method for transformer fault basic probability assignment based on improved three-ratio method. J. Power System Protection and Control., 43(7):115-121.