Yingshuai Hao[1], Chengjun Huang[1], Wendong Zheng[1]

Shanghai Jiaotong University (1)

Partial Discharge Pattern Recognition of Medium-Voltage Switchgear Based on Association Rules Mining

***Abstract**. In the paper, a new partial discharge pattern recognition approach based on association rules mining is developed. Some statistical parameters are extracted from the sampled transient earth voltage data and classic Apriori algorithm is employed to mine the association rules between those parameters and the corresponding fault types. Moreover, using other experimental data obtained in the laboratory, the method is validated and made comparison with two conventional methods.*

***Streszczenie.** W artykule opisano metodę wykrywania wyładowań niezupełnych, poprzez wyznaczenie reguł powiązanych z wystąpieniem tych wyładowań i ich wyszukiwanie. Dzięki analizie statystycznej danych dotyczących stanów nieustalonych napięcia oraz zastosowaniu algorytmu a-priori, określono zasady powiązane z wystąpieniem zjawisk. Skuteczność działania rozwiązania została potwierdzona wynikami badań eksperymentalnych, które następnie porównano z dwoma konwencjonalnymi metodami. (**Wykrywanie schematu wyładowań niezupełnych w rozdzielni średniego napięcia w oparciu o wydobycie powiązanych reguł**).*

Keywords: Partial Discharge, Pattern Recognition, Association Rules Mining, Transient Earth Voltage
Słowa kluczowe: wyładowanie niezupełne, rozpoznawanie schematu, wydobywanie reguł powiązanych, przejściowe napięcie od ziemi.

Introduction

As a widely used equipment in substations, the medium-voltage switchgear plays a great role in ensuring the reliability of power supply. In order to prevent catastrophic power failure resulting from switchgear's insulation failure, people develop various methods to monitor its insulation condition. Among those methods employing different characteristics of insulation deterioration respectively, the online measurement of partial discharge (PD) and generated transient earth voltage (TEV) signal on metal enclosure proves to be a convenient and effective way to provide important information about the insulation condition [1].

Since the detrimental effect to insulation system varies with defect types, PD source classification is an essential stage in order to make a persuasive evaluation of switchgear's insulation condition and offer directions on later handling [2]. The most popular methods for classification are artificial neural network (ANN) [3] and decision tree [4-5]. The paper [3] presents several types of ANN available to PD source recognition. Since ANN has a parallel information-processing structure, its fault tolerance capability is at a high level. However, due to that it's like a 'black box' to the users, people have difficulty in understanding how those input parameters work and which plays a greater role in the process of classification. As to decision tree, the method only pays attention to those parameters that could distinguish different defects to the maximum extent, hence it can't make good use of all the input parameters. Besides, unlike ANN, the information-processing structure of decision tree is series, though it's much easier for users to understand how it works, the parameters' deviation can possibly lead to judgment errors.

Based on the analysis of ANN and decision tree, a novel two-step approach is presented in the paper, which combines ANN's high level of fault tolerance and decision tree's feature of easiness to understand. In the first step, the selected parameters get extracted from the original experimental data and then get discretized by fuzzy c-means clustering (FCM) algorithm. Apriori algorithm described in [6] is employed to perform association rules mining (ARM) between discretized parameters and corresponding defect type. In the second step, defect recognition process, which gets accomplished by means of fuzzy logic, is demonstrated in detail and the approach finally gets validated through the identification result of the testing records.

Experimental testing

1. TEV signal measurement

As demonstrated in Fig.1, the PD experiment is carried out with the testing circuit in accordance with IEEE standard [7]. The test object is subject to medium voltage as high as several kilovolts to stimulate the occurrence of PD phenomenon. When PD happens, the discharge current pulses excite a dynamic electromagnetic field in the surroundings and induce a TEV signal on the earthed metal enclosure of switchgear. The induced voltage signal gets detected and measured by a capacitive probe placed on the surface [8]. The measurement system has a bandpass of 3~100 MHz and the sampled signal can be amplified by 0dB, 20dB or 40 dB. The sampling time last 1024ms and hence 50 cycles of PD information can be obtained through one sampling.

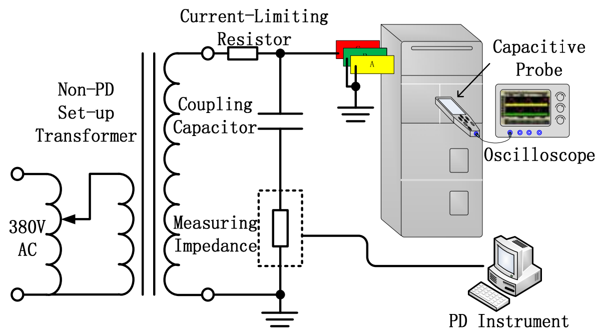

Fig.1. Testing circuit of PD

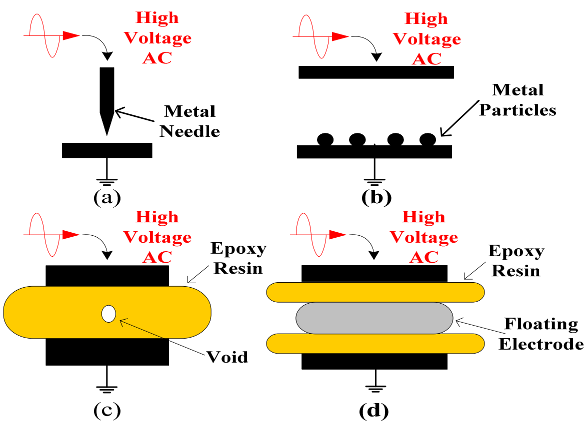

Fig.2. Artificial defect model

Fig.3. Factual experiment picture

2. Artificial defect model

In the experiment, four types of defect models are made and tested to simulate the typical PD sources in switchgear, including needle-to-plate corona PD, metal particle PD, internal void PD and floating electrode PD. The models are shown in Fig.2. Fig.3 describes the factual experiment circumstance.

Experimental data processing

1. Parameters selection

In order to denote the characteristics of PD pulse from different perspectives, various parameters, e.g. those concerning the discharge current waveform, have been proposed and examined by researchers. But in the paper, the phase features of PD pulse draw more attention and thus statistical parameters presented in [2] are extracted and three 2-dimension statistical graphs get obtained, i.e.

● Hvmax(θ) - the maximum measured voltage pulse-phase distribution; θ - phase angle.

● Hvmean(θ) - the mean voltage pulse-phase distribution; θ - phase angle.

● Hn(θ) - the number of PD pulses-phase distribution; θ - phase angle.

In view of the fact that to some kinds of PD, e.g. needle-to-plate corona PD, the pulse distribution in positive half cycle differs much from that in negative half [2], those parameters related to the shape of pulse distribution are extracted in positive and negative half cycles separately and hence totally 25 parameters are obtained, i.e.

● Skewness (Sk), kurtosis (Ku) and number of peaks (Pe) of positive and negative halves of Hvmax(θ), Hvmean(θ) and Hn(θ) respectively (a total of 18 parameters).

● Asymmetry (Asm) of Hvmax(θ), Hvmean(θ) and Hn(θ) (3 parameters).

●Cross correlation (cc) and modified cross correlation (mcc) factors of Hvmax(θ) and Hvmean(θ) (4 parameters).

2. ARM

When it comes to ARM, anyone can't avoid mentioning two popular algorithms: Apriori algorithm [6] and frequent-pattern growth (FP-growth) algorithm [9]. Since in the research there's no need to mine the maximum frequent item set in which FP-growth algorithm has obvious advantage over Apriori algorithm, classic Apriori algorithm is employed to perform ARM and the complexity of the task could decrease. But Apriori algorithm can only deal with discrete data, whereas parameters extracted from experimental data are continuous, therefore those parameters should firstly get discretized.

2.1 Parameters discretization

Discretization means that each parameter is divided into several intervals and then the original data is replaced by several membership values, each reflecting the extent to which that original data belongs to one particular interval. FCM algorithm presented in [10] is adopted in the research to perform discretization. Since the number of divided intervals is determined by users, given that in the research four types of PD defects are taken into account, the number is set to four. Meanwhile, since after discretization a value is replaced by several membership values, take the parameter of 'Sk+ max' for example, the membership values to 4 divided intervals is demonstrated in Fig.4.

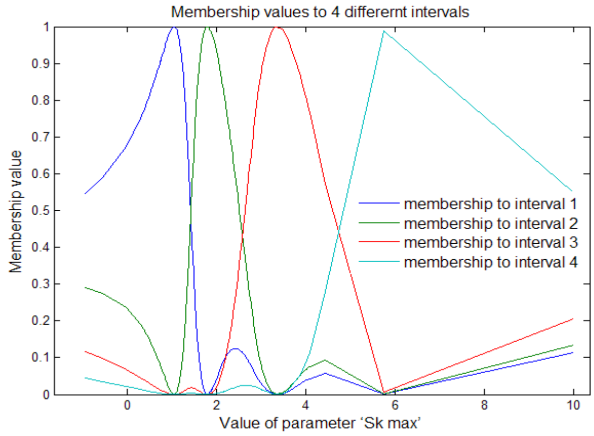

Fig.4. Membership values to 4 divided intervals

In Fig.4, each colored line denotes the membership value of certain parameter value to one of the 4 intervals, and it can be seen that the membership value ranges between 0 and 1. Since the membership value varies according to different intervals, the interval which a particular value belongs to is determined by the maximum membership value and hence the parameters can get discretized.

2.2 Experimental data resampling with Bootstrap

When implemented to mine association rules, though for Apriori algorithm there's no lower limit of data number, given that the algorithm is initially proposed to apply in a large database [11], the available experimental data records are insufficient to obtain reliable results. Therefore a resampling technique named Bootstrap [12] is utilized to generate enough data from the original records.

The initial data of each type of PD has 100 records, in the research, 80% are used as training data and the rest are employed to test the accuracy of the approach. After resampling with Bootstrap, totally the training set consists of 3200 records and the test data increase to 800 records.

2.3 Apriori algorithm implementation

Conventionally Apriori algorithm is used to mine the association rules among user-interested attributes, since now the research's focus is to perform PD pattern recognition, the key point is to mine the association rules between the discretized parameters and PD patterns.

In the process of ARM, the minimum support and confidence values play a significant role in determining the number of mined rules. If both parameters are set too small, many useless rules get obtained and hence the complexity of the task increases. But if those parameters are set too large, some original records differing from the majority may be skipped for their small percentage in the training data. In order to make a compromise between the complexity of the task and the variety of the original records, the minimum support value is set to 0.015 and that of confidence is 0.75. After implementing Apriori algorithm, the number of rules is shown in Table 1.

Table 1. Obtained Rules (only those whose length≤5 presented)

Length of obtained rules	Number of obtained rules
2	5
3	602
4	14442
5	132399

It should be noted that since the mined rules contain both discretized parameters and defect types, the rule with a length of *n* indeed contains *n-1* discretized parameters.

2.4 Rule selection

From Table 1, it can be seen that the number of rules with an upper length limit of 5 is over 150 thousand. If these rules are applied to PD pattern recognition directly, the process of recognition becomes extremely complicated. Therefore it is necessary to carry out selection from the obtained rules.

In the study, each rule is evaluated by the following two parameters: confidence and completeness [11]. Since each rule contains two parts: the former is the rule condition composed of parameter intervals, denoted by A, and the latter is the corresponding PD type, denoted by B, the confidence and completeness are defined as follows.

(1) $$confidence=\frac{num(A\cap B)}{num(A)}$$

(2) $$completeness=\frac{num(A\cap B)}{num(B)}$$

where: $num(A\cap B)$ denotes number of records in the training data satisfying A and B, while $num(A)$ and $num(B)$ denotes number of records satisfying A or B respectively.

After calculating the respective completeness and confidence of these rules, they are ranked by product of the two evaluation indexes in descending order. Each rule then compares with those in its front. If there is an inclusion relationship between the rule and any of the rules in front of it, then the rule gets eliminated. An instance is that the first rule concerning corona PD is Sk- num_1&Asm max_4 ('-' means the negative half cycle, 'max' means Hvmax(θ), 'num' means Hn(θ) and the last figure denotes the interval number), and the sixth one is Sk- num_1&Asm max_4&Ku+ max_1, then the latter gets deleted because the former contains all the instances of the latter. Table 2 demonstrates 10 of the selected 20 rules of corona PD.

Table 2. Ten association rules of corona PD

Number	Association Rules
1	Sk- num_1 & Asm max_4
2	Sk+ num_1 & Asm num_4
3	Pe+ max_4 & mcc max_3
4	Sk- num_1 & mcc mean_3
5	Ku+ max_1 & Pe- max_4 & Asm max_4
6	Sk- num_1 & Ku- num_1 & mcc max_3
7	Ku+ max_1&Pe- mean_4&Asm max_4&mcc max_3
8	Sk+ num_1&Sk- num_1&Ku+ max_1&mcc max_3
9	Sk- num_1&Ku+ max_1&Ku+ num_1 &mcc max_3
10	Sk- num_1&Ku+ mean_1&Ku+ num_1&mcc max_3

Pattern recognition

1. Selection of membership function

Since the obtained rules reflect the association relationship between certain parameters' intervals and PD patterns, before these rules get applied, it's necessary to calculate the membership values in particular intervals. Fig.4 demonstrates that the obtained membership after discretization is similar to triangle membership function, which is depicted in Fig.5. In addition, compared to other forms of membership functions, e.g. trapezoidal membership function, the triangle one is relatively simple because its parameters are easy to get determined from the result of discretization. For these reasons, in the study the triangle membership function is selected and based on it the calculation of membership is finished as the first step of pattern recognition.

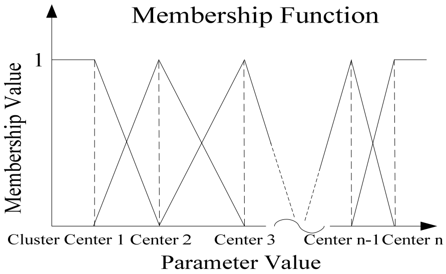

Fig.5. Triangle Membership Function

2. The principle of pattern recognition

From the calculated membership values in certain intervals of a particular rule, one needs to measure the degree of satisfying rule condition and the possibility of belonging to corresponding PD type. Besides, since each PD pattern has 20 rules, it's entirely possible that several rules get satisfied at the same time. If so, how to measure the overall possibility is a key issue as well. In the study, a method called 'mean-product' principle is developed and illustrated in Fig.6.

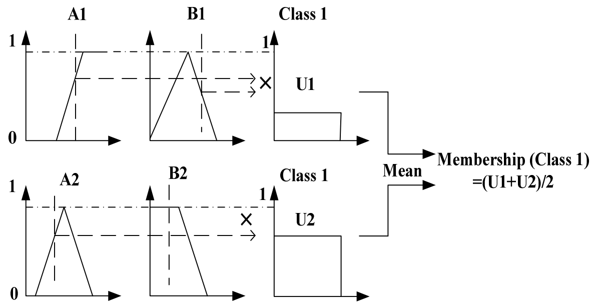

Fig.6. Mean-product principle

After working out the mean possibilities with which the sampled data belongs to each PD type, one can judge the fault through comparing those values.

PD pattern recognition result and analysis

Through employing triangle membership function, mean-product principle and the selected 20 rules of each PD type presented above, the PD patterns of the test records get identified and the result is shown as follows:

Table 3. PD recognition result

PD type	Correct recognition rate
Floating PD	190/200
Void PD	192/200
Particle PD	185/200
Corona PD	193/200

Table 4. PD recognition result of BPNN and decision tree

Method	PD type	Correct recognition rate
BPNN	Floating PD	192/200
	Void PD	194/200
	Particle PD	190/200
	Corona PD	200/200
Decision tree	Floating PD	179/200
	Void PD	182/200
	Particle PD	171/200
	Corona PD	190/200

From Table 3, it can be concluded that the correct recognition rate varies with PD types. Generally speaking, the correct recognition rates all surpass 92%, which can be seen as a satisfying result. The recognition results utilizing back-propagation neural network (BPNN) and decision tree respectively are shown as follows.

From Table 4, it can be concluded that the method has almost same recognition accuracy as BPNN and better recognition rate compared to DT.

As to the role that certain statistical parameter plays in recognition, it can be evaluated by a new parameter defined in the following formula.

$$imp_j = \sum_{i=1}^{20} completeness_i \bullet confidence_i \bullet \sigma_{ij} \tag{3}$$

where:

$$\sigma_{ij} = \begin{cases} 0 & if\ \ parameter\ j\ doesn't\ appear\ in\ rule\ i \\ 1 & if\ \ parameter\ j\ appears\ in\ rule\ i \end{cases} \tag{4}$$

$completeness_i$ means the completeness of rule i and $confidence_i$ means the confidence of rule i. Hence the relative extent of importance can get approximately determined through the calculation of imp. Table 6 presents the parameter intervals involved in the recognition of corona PD.

Table 5. Parameter intervals and the respective value of imp

Parameter interval	Value of imp
Asm num_4	0.5000
Ku- num_1	0.5984
Sk+ num_1	0.9797
Ku+ num_1	1.9492
mcc mean_3	2.6924
Pe- max_4	3.1260
Sk- num_1	3.7972
Pe- mean_4	3.9387
Ku+ max_1	4.0131
Ku+ mean_1	4.0158
Asm max_4	4.4921
mcc max_3	4.6903

From Table 5, totally there are 12 different parameters in the selected 20 rules of corona PD, so it can be seen that not all parameter intervals, even not all extracted parameters, participate in the recognition, and since the value of parameter 'imp' varies according to different parameter intervals, people can judge the extent of importance of particular parameters, which can't be obtained from ANN.

Conclusion

A new method employing ARM and fuzzy logic to perform PD pattern recognition is proposed in the paper. The main procedures, including the generation and selection of association rules, the calculation of membership values, and the principle of recognition, are described in detail. Then, the accuracy of applying the approach to recognize test data is presented and for comparison purpose, the results employing two conventional approaches are presented as well and thus the performance of the approach gets verified. Finally, how to evaluate the role that certain parameter plays in the process of recognition is demonstrated.

Acknowledgments: The authors would like to thank the Key Laboratory of Control of Power Transmission and Conversion, Ministry of Education, for research support.

REFERENCES

[1] Y.Ming, The application of temporary earth voltage (TEV) measurement in the online monitoring of the partial discharge of HV switch cabinet, *2010 China International Conference on Electricity Distribution(CICED)*, 2010, 1-6.

[2] F.H.Kreuger, E.Gulski, Classsification of partial discharge, *IEEE Transactions on Electrical Insulation*, 28(1993), Issue 6, 917-931.

[3] K.Tan, F.Li, Some types of artifical neural network used for recognition of partial discharge (in Chinese), *High Voltage Engineering*, (1996), Issue 4, 3-7.

[4] T.K.Abdel, R.M.Sharkawy, Partial discharge pulse pattern recognition using an inductive inference algorithm, *IEEE Transactions on Dielectrics and Electrical Insulation*, 12(2005), Issue 2, 320-327.

[5] T.K.Abdel, R.M.Sharkawy, Partial discharge pattern classification using fuzzy decision tree approach, *IEEE Transaction on Instrumentation and Measurement*, 54(2005), Issue 6, 2258-2263.

[6] R.Agrawal, Mining association rules between sets of items in large database, *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, Washington D.C, United States, 1993, 207-216.

[7] IEEE Standard for High-voltage Switchgear (Above 1000V) Test Techniques-Partial Discharge Management, IEEE STD C37.301-2009, 1-63.

[8] P.Brown, Nonintrusive partial discharge measurements on high voltage switchgear, *IEEE Colloquium on Monitors and Condition Assessment Equipment (Digest No.1996/186)*, 1996, 10/1-10/5.

[9] J.Han, J.Pei, Mining frequent patterns without candidate generation, *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 29(2000), Issue 2, 1-12.

[10] J.C.Bezdek, R.Ehrlich, FCM: The fuzzy c-means clustering algorithm, *Computers& Geosciences*, 10(1984), Issue 2-3,191-203.

[11] Z.Yang, W.H.Tang, Association rule mining based dissolved gas analysis for fault diagnosis of power transformers, *IEEE Transactions on Systems, Man and Cybernetics, Part C: Application and Reviews*, 39(2009), Issue 6, 597-610.

[12] A.Zoubir, B.Boashash, The bootstrap and its application in signal processing, *IEEE Signal Process.Mag.*, 15(1998), No.1, 56-77.

***Authors**: Yingshuai Hao, graduate student in high voltage and online monitoring of electric equipments at the key Laboratory of Control of Power Transmission and Conversion, Ministry of Education, Dept. of Electrical Engineering, Shanghai Jiao Tong University, China. E-mail: ys_hao@sjtu.edu.cn.*

Chengjun Huang, associate professor in high voltage at Dept. of Electrical Engineering, Shanghai Jiao Tong University, China.

E-mail: chijhuang@sjtu.edu.cn.