International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 7970-7985

https://doi.org/10.70517/ijhsa464683

A Methodology for Multi-dimensional Mining and Analysis of Power Grid Operation and Maintenance Data through the Synergy of Computation and Information Theory

Songyao Feng[1], Zhengyan Huang[1,*], Junhao Song[1] and Xuexia Quan[1]
[1] The Information Center of Guangxi Power System Co., Ltd., Nanning, Guangxi, 530012, China
Corresponding authors: (e-mail: hzy900529@163.com).

Abstract With the rapid development of smart grid and the increasing growth of electric power equipment, operation and maintenance intelligence gradually turns into an important way for power grid enterprises to improve productivity. The research proposes a smart grid operation and maintenance system using ExtJs+Spring+iBatis architecture. It first improves the weighted fusion rule based on the D-S evidence theory of virtual union, proposes a grid diagnostic model with multi-source information fusion, and then establishes a grid state evaluation model using AR model and SOM neural network model. The results show that the fusion model based on the improved D-S evidence theory fuses the results of switching quantity analysis and electrical quantity analysis for diagnosis, and the diagnosis results are more accurate compared to a single source of information, and at the same time, the grid state evaluation method can quickly and effectively detect the state of power grid operation and maintenance. The combination of big data analysis technology and power equipment evaluation will be a useful attempt in the construction of smart grid, which improves the requirements for equipment testing parameters.

Index Terms D-S evidence theory, multi-source information fusion, AR model, SOM neural network, grid operation and maintenance

I. Introduction

Power grid is one of the basic systems operating in modern society, which connects the generation and consumption of energy and guarantees the production and living needs of urban and rural residents [1], [2]. However, due to the complexity, scale and distribution of the grid, the safe, stable and efficient operation of the grid has been one of the important research topics in the power industry [3], [4].

In the current era of informationization and intelligence, the use of data mining to analyze and predict the grid operation and maintenance data has become an indispensable part of the power industry [5], [6]. Grid O&M data is a variety of data generated during the operation and maintenance of the power system, which includes information such as power load management, operation status monitoring, energy consumption prediction, and electrical parameter analysis [7]-[9]. These data come from a variety of sources, fast transmission speed, large and trivial, if these data are not scientifically researched and analyzed, they may cause accidents in the power grid, and will greatly affect the production and life of the masses [10]-[12]. Therefore, grid operation and maintenance multidimensional data mining and analysis is of great significance for the safe operation of power grid [13], [14]. Grid operation and maintenance multidimensional data mining and analysis is the use of data mining technology to mine various data generated in the process of grid operation and maintenance, and to analyze and study these data to assist the operation and planning departments to timely find problems and solve problems to ensure the safe, stable and efficient operation of the grid [15]-[18].

Literature [19] introduced the grid dispatch operation analysis and data mining system, and used the data warehouse online analysis and processing technology to manage and analyze the massive data, which verified that the system can mine the laws of grid operation and help to improve the management of the power system. Literature [20] introduced the operation and maintenance management of distribution grids and examined the application of data mining in the operation and maintenance management of distribution grids for the situation of multi-timescale and multi-spatio-temporal data in distributed grids. Literature [21] proposed a research on the method of analyzing and predicting the cost of grid-connected production and technical improvement projects based on data mining, and analyzed the composition and influencing factors of the cost of the grid project from multiple perspectives of equipment and labor costs and overhead costs, and obtained accurate cost prediction results through the introduction of data mining technology. Literature [22] applied data mining techniques to the analysis of power quality, revealing the effectiveness of data mining techniques in identifying the problematic areas of power quality, and

providing support for the staff's power grid operation. Literature [23] proposed a novel load forecasting (LF) strategy for power grids based on data mining techniques, noting that, in addition to the novel load estimation, the strategy also used new outlier rejection and feature selection methods and verified their effectiveness. Literature [24] proposed a method for mining and analyzing secondary equipment defect data based on the FP-Growth algorithm, aiming at effective maintenance and operation of intelligent substations, based on the analysis of substation defect data, revealing the effectiveness of the method, which is effective in identifying the relationship between the nature of the defects, the causes of the defects, and so on. Literature [25] examined the data mining technology suitable for the operation and monitoring of power grid enterprises and constructed a data mining technology database, the research results can effectively reduce the duplication of resources, thus improving the lean operation and management of power grids.

The study proposes a grid operation and maintenance system based on big data mining, the overall framework of the system adopts ExtJs+Spring+iBatis architecture, and the hardware includes: perception integration chip, data transmission device, data fusion processor, central controller, and visualization device. The software part includes two modules, one of which investigates the smart grid fault diagnosis method based on multi-source information fusion. The D-S evidence theory fusion algorithm is analyzed, the improved weighted fusion rule based on virtual coalition of D-S evidence theory is proposed, and considering the normalization of the fusion result, a decision-making strategy conditioned on the nonpropositional fusion result M(¬) is proposed, and then the faulty components are subjected to the fault retrospective reasoning to analyze the accuracy of the protection, circuit breaker action, and alarm information. Secondly, the AR model and SOM neural network model are used to process a certain continuous power equipment state quantity, discretize the continuous and uninterrupted monitoring data into individual sequences in linear space, and calculate the transfer probability of the state quantity with time as the axis to establish the power grid state evaluation model. Finally, the performance of the software module is tested in combination with relevant arithmetic examples and experiments, which lays the foundation for ensuring the security, stability and integrity of power grid information.

II. Grid informationization operation and maintenance system design based on big data analysis

II. A.System design

In the face of grid information, “big network, big system, big centralized” and other trends, this paper based on big data analysis of grid information operation and maintenance system overall framework using ExtJs + Spring + iBatis architecture, grid operation and maintenance architecture shown in Figure 1. The architecture of the system using B / S architecture, hair at the same time through IE, FireFox and Chrome and other mainstream browser testing, so that the system UI has a strong compatibility, support for J2EE1.5 and Servlet3.0 specification, to achieve high performance, scalability and cross-platform features, the use of advanced five-layer system, support for a variety of operating environments, along the lines of the MVC idea, make the project level more clear. Make the project level more clear.

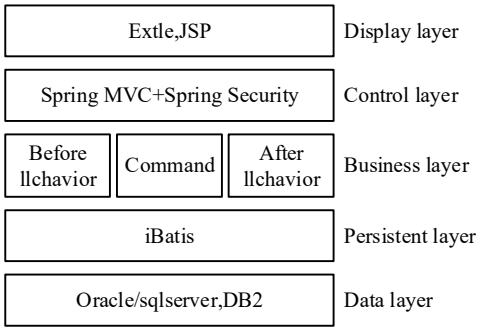


Figure 1: ExtJs + Spring + iBatis

II. B.System Hardware Design

Based on the system framework, several major hardware devices of the system are analyzed.

II. B. 1) Perception Integration Chip

Perception integration chip is able to extract information from the power consumption collection system, scheduling management, fault repair and other related basic systems are concentrated into a customized system of data integration chip. In the chip has a 12-bit A/D converter, integration rate of up to 1 MHz, 16 single-ended or 8

differential or combination of input methods, PCI bus data transmission, analog input channels of data integration trigger mode can be used pre-trigger, post-trigger, match trigger and delay trigger.

II. B. 2) Data transmission equipment

Data transmission equipment can be directly connected to the various microcontroller devices and industrial control intelligent communication equipment, PLC and instrumentation, such as the upper unit configuration software and wireless communication between the terminal equipment, but also with the touch screen or PLC and field terminal equipment data communication. Support: wireless 4G-RTU transmission application between PLC and PLC, multi-server center to manage multiple terminals at the same time, one master to multi-slave or multi-master to multi-slave communication, support for TCP/IP, UDP network transmission protocols, and transmission of device data to the website WEB server in the way of HTTP POST.

II. B. 3) Data fusion processor

The role of the data fusion processor is to clean and organize the data integrated together, including de-measurement, denoising, conversion, compression, fusion, and analysis. In this system, the data fusion processor is is a single chip CPU, i.e. microprocessor.

II. B. 4) Centralized controller

The central controller, which is the device that controls the operation of all the hardware in the system by means of protocols, plays a central role on the motherboard. The central controller in this system integrates an integrated circuit chip with CPU and other circuits to form a complete microcomputer system. There are components such as RAM, ROM, a serial interface, a parallel interface, timers and interrupt scheduling circuits included in the microcontroller.

II. B. 5) Visualization equipment

The visualization device, i.e. the system display window, can be used for system operation data monitoring, data result display and adjustment of each parameter, and its composition structure is shown in Figure 2.

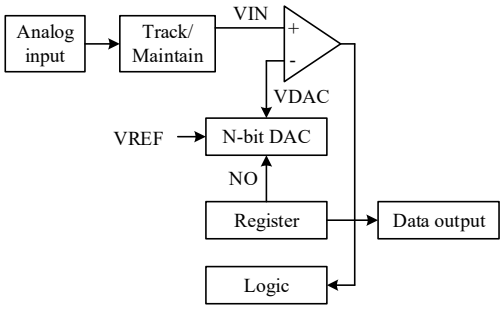


Figure 2: The basic structure of the visual device

III. Grid diagnostic model based on data fusion

III. A. Improving the D-S Evidence Theory Fusion Algorithm

III. A. 1) Conflict metrics for sources of evidence

In power grid fault diagnosis, through each evidence source processing analysis to get the suspicious component failure degree, according to statistical theory or expert experience decision-making preliminary judgment of the suspicious component whether there is a possibility of failure, therefore, remember h evidence source to judge the probability of component failure threshold for $\alpha_1, \alpha_2, ..., \alpha_h$, when the i th evidence source focal element "-" basic probability assignment $m_i(\cdot) \geq \alpha_i$, then let $FT_i(\cdot) = 1$; conversely, let $FT_i(\cdot) = 0$, you can get the fault table with binary attributes FT, you can intuitively determine whether the same proposition in the focal element of the possibility of conflict between different sources of evidence, and then calculate the conflict coefficients and the distance of conflict through the Jousselm function.

Taking two faulty evidence sources S1 and S2 as an example, their basic probability distribution functions are m_1 and m_2 respectively, and there are y suspected faulty elements and N faulty propositions in the same identification frame Θ with a probability threshold of α_1, α_2. The fault table of the two evidence sources can be obtained as FT. The n th faulty propositions have the joules denoted A_m and B_n in the fault table FT, and the

kernels consisting of joules A and B are denoted as $A=\{A_1,..,A_n,\ldots,A_N\}$ and $B=\{B_1,\ldots,B_n,\ldots,B_N\}$ respectively, and then the two Jousselm distances between the evidence sources can be expressed as:

$$J_{12}=\sqrt{\frac{1}{N}\sum_{n=1}^{N}(m_1(A_n)-m_2(B_n))^2\cdot d_n} \tag{1}$$

$$d_n=1-\frac{|A_n\cap B_n|}{|A_n\cup B_n|} \tag{2}$$

$$D_{12}=\begin{bmatrix}d_1 & \cdots & d_n & \cdots & d_N\end{bmatrix}^T \tag{3}$$

where, d_n is the focal element conflict coefficient, and D_{12} represents the matrix of conflict coefficients of each focal element of two evidence sources S1 and S2, and the coefficient is 0 term is modified i.e., if $d_n=0$, then make $d_n=1/2^y$, in order to avoid the problem that the conflict distance between different evidence sources is 0.

When facing h sources of evidence, the conflict coefficient matrix of each focal element of evidence source i and evidence source j is calculated respectively D_{ij}, and then the conflict distance matrix between the sources of evidence can be obtained J:

$$J=\begin{bmatrix}0 & J_{12} & ... & J_{1h}\\ J_{21} & 0 & ... & J_{2h}\\ \vdots & \vdots & & \vdots\\ J_{h1} & J_{h2} & ... & 0\end{bmatrix} \tag{4}$$

Element J_{ij} in matrix J represents the conflict distance between evidence source i and evidence source j.

In contrast to the conflict distance matrix, the similarity $s_{ij}=1-J_{ij}$ between any two evidence sources i and j generates a similarity distance matrix S. The sum of the similarity distances between evidence source i and the other evidence sources indicates the degree of support $Sup(i)$ for evidence source i by the multiple evidence source system:

$$Sup(i)=\sum_{j=1}^{n}s_{ij}-1 \tag{5}$$

Normalizing the level of support for each source of evidence to a probability space yields the credibility of source i in a multiple source system of evidence $Crd(i)$:

$$Crd(i)=\frac{Sup(i)}{\sum_{i=1}^{n}Sup(i)} \tag{6}$$

III. A. 2) Improved fusion rules based on virtual coalitions

The non-diagonal elements (i.e., $i\neq j$) in the conflict distance matrix J are judged to satisfy the classification conditions by setting the conflict distance threshold J_α, and then the multiple evidence sources are classified to form a virtual coalition. The process of categorizing the multiple evidence sources to form a virtual coalition of evidence may be described as follows:

Step 1: If $\max(J_{ij})\le J_\alpha$, it means that there is no strong conflict between the evidences and no need to classify and divide them; if $\max(J_{ij})>J_a$, further division is required.

Step 2: Find the minimum value $J_{ij},(i\neq j)$ of the non-diagonal elements of the distance matrix J, form a virtual union G from evidence source i with evidence source j, and replace i and j.

Step 3: Find the maximum value of the distances of Evidence Source i and Evidence Source j from the other evidence that constitutes the distance between Virtual Union G and the other evidence, i.e., $J_{Gf}=\max(J_{if},J_{if})$, and update the distance matrix J.

Step 4: If the minimum value of the non-diagonal element of the updated distance matrix J is not greater than the threshold J_α, then repeat steps 2 and 3 until $\min(J_{ij}) > J_\alpha$, which indicates that the distance between the alliances meets the classification requirements and the classification of evidence sources is completed. Thus, the h evidence sources that meet the conditions are divided into g virtual alliance $G_1, G_2, ..., G_g$, where any alliance G_k is composed of e evidence sources, namely:

$$G_k = \{m_1, m_2, \cdots, m_e\} \tag{7}$$

Evidence source i credibility $Crd(i)$ that represents the weight of evidence source i, and the weight of the virtual coalition depends on the credibility $Crd(i)$ of each evidence source within the coalition, it can be seen that the weight W_k of the virtual coalition G_k is:

$$W_k = Crd(1) + Crd(2) + \cdots + Crd(e) \tag{8}$$

Then the improved weighted fusion rule for D-S evidence theory based on virtual coalition can be obtained as:

$$m_k(A) = m_1 \oplus m_2 \oplus \cdots \oplus m_e \tag{9}$$

$$m'(A) = (W_1 \cdot m_1(A)) \oplus \cdots \oplus (W_k \cdot m_k(A)) \oplus \cdots \oplus (W_g \cdot m_g(A)) \tag{10}$$

where Eq. (9) represents the fusion of e sources of evidence in the formed coalition G_k separately to obtain the fusion result of the coalition, and Eq. (10) represents the weighted fusion of the g coalitions to obtain the final result.

III. A. 3) Improvements to the D-S Evidence Theory information fusion model

The main implementation steps of the improved D-S evidence theory weighted fusion method [26] based on virtual coalition are:

Step1: Threshold comparison of the basic probability assignments of the focal elements of each evidence source to obtain the fault table FT with binary attributes, which can intuitively determine the conflict of focal elements in the same proposition;

Step2: Calculate the focal element conflict coefficients of each proposition in the fault table respectively, and the conflict distance between evidence sources J, reflecting the degree of conflict between evidence sources;

Step3: categorize each evidence source according to a threshold comparison of the conflict distances to form a virtual coalition of evidence;

Step4: Calculate the trust degree Crd of the evidence sources in the virtual coalition G_k, and express the fusion weight W_k of this virtual coalition G_k, and carry out the D-S evidence theory improvement weighted fusion to get the fusion result.

III. B. Smart grid diagnostic model based on multi-source information fusion

III. B. 1) Diagnostic Decision Strategies and Evaluation of Alarm Information

In grid fault diagnosis, when there are multiple faulty elements, the fault propositions $A_1, A_2, \ldots, A_y$ representing y suspected faulty element form the identification framework Θ of the whole fault event, and the basic probability distribution function of each of the h evidence sources is $m_1, m_2, \cdots, m_h$.

In order to represent the non-faulty case of the suspected element, the nonproposition "$\neg$" is added as the concatenation set of each fault proposition, i.e:

$$\neg = \{A_1 \cup A_2 \cup \cdots \cup A_y\} \tag{11}$$

Then a new identification framework Θ is formed for:

$$\Theta' = \{A_1, A_2, ..., A_y, \neg\} \tag{12}$$

At this point, the total number of faulty propositions in the recognition framework Θ' is $N = y + 1$.

Set the fusion result $M(\neg)$ of the non-propositions as the diagnostic decision threshold, set $1/N$ as the confidence difference threshold, and N as the total number of faulty propositions in the recognition framework Θ', when the fusion result of proposition A_n satisfies equation (13):

$$\begin{cases} M(A_n) > M(\neg) \\ M(A_n) - M(\neg) > \dfrac{1}{N} \end{cases} \tag{13}$$

After the final fault element is obtained from the diagnostic decision, fault retracement is carried out according to the grid topology, protection configuration, protection principle and circuit breaker action rules to determine the accuracy of alarm information, protection and circuit breaker action.

III. B. 2) Smart grid fault diagnosis based on multi-source information fusion

This paper proposes a smart grid fault diagnosis method based on multi-source information fusion, and the diagnosis model is shown in Figure 3. In case of grid fault, the control center receives various protection and circuit breaker alarm information as well as faulty electrical quantity data, and through the fusion diagnosis and analysis of the alarm information and electrical quantity data, the decision-making results are derived in order to ensure the rapid processing of faulty components and normal power supply of non-faulty components.

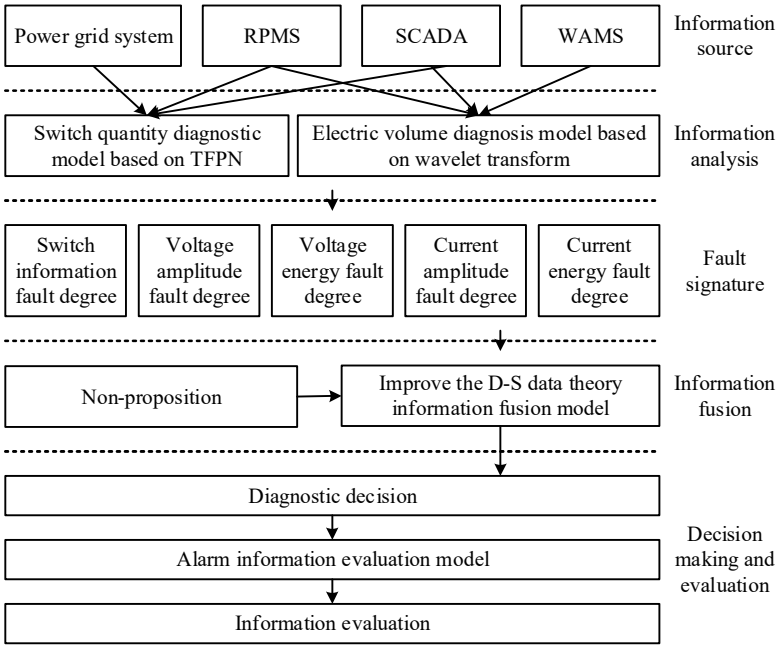


Figure 3: Smart Grid fault diagnosis model based on multi-source information fusion

III. C. Example analysis

Figure 4 shows the IEEE39 simulation system.

III. C. 1) Calculation example 1

A short-circuit fault occurs on line L_{16-17}, main protections $L_{(16)-17m}(18ms)$ and $L_{(17)-16m}(17ms)$ operate, circuit breaker $_{(17)-16}(66ms)$ trips, but $CB_{(17)-16}$ refuses to operate, line remote backup protections $L_{(21)-16s}(836ms)$, $L_{(24)-16s}(844ms)$, $L_{(19)-16s}(838ms)$, and $L_{(15)-16s}(837ms)$ operate, tripping circuit breakers $CB_{(21)-16}(888ms)$, $CB_{(24)-16}(890ms)$, $CB_{(19)-16}(880ms)$, and $CB_{(15)-16}(865ms)$, and backup protection $L_{(16)-21p}(422ms)$ malfunctions resulting in circuit breaker $CB_{(16)-21}(440ms)$, where all the time-scaled information is in parentheses.

According to the received alarm information, the set of possible faulty components is obtained after fault area search. $\{L_{16-17}, L_{16-21}, L_{16-24}, L_{15-16}, L_{16-19}\}$ The TWFCPN fault probability of each component in the set of faulty components is calculated according to the time-weighted fuzzy colored Petri net diagnostic model, and the corresponding TWFCPN fault probability characterization is obtained. The related line current waveform data are extracted, L_{16-17} faulty phase currents, improved EMD results and marginal spectral analysis are shown in Figs. 5

to 7, and the HHT frequency distortion and HHT amplitude distortion are calculated.The fusion results of the three fault probability characterizations and the error squared distances between each component and the cluster center after the clustering is stabilized are shown in Table 1, and the decision model calculates the center value of the clustering C_1 as 0.9188 and the center value of cluster C_2 is 0.0188, then C_1 is a faulty cluster. From the table, it can be seen that component L_{16-17}, and C_1 of the center also error square distance compared to C_2 is much smaller, then component L_{16-17}, belongs to cluster C_1, diagnosis for faulty components, diagnosis results are correct.

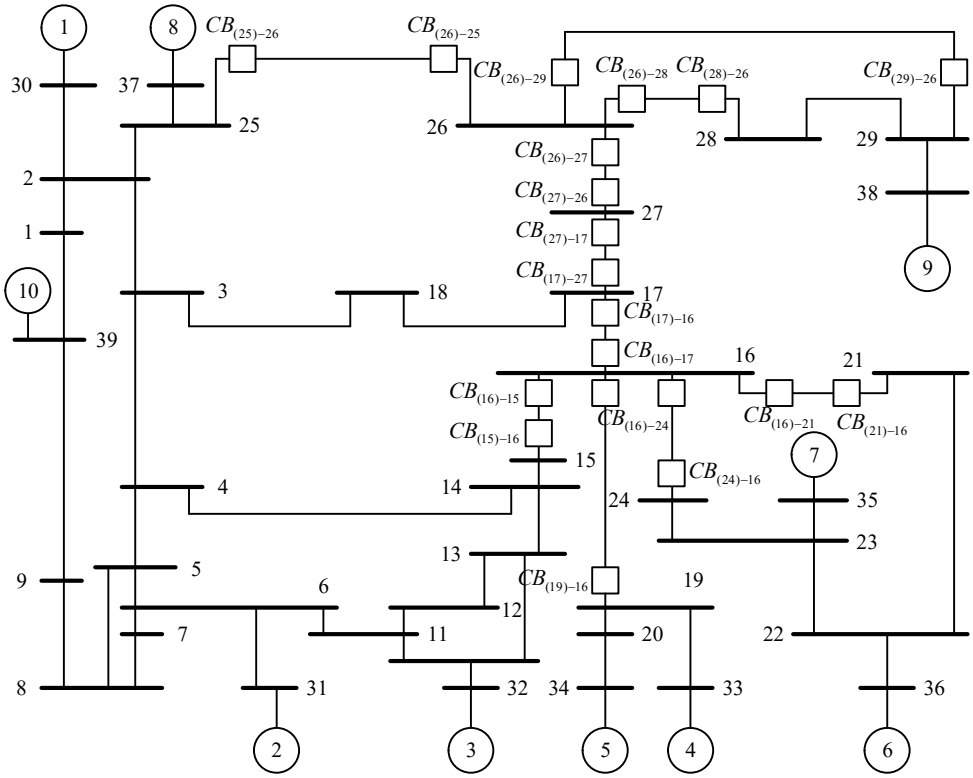


Figure 4: IEEE39 System

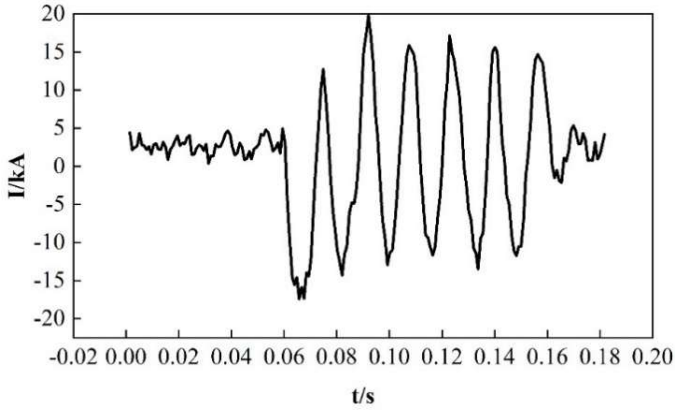


Figure 5: Fault phase current

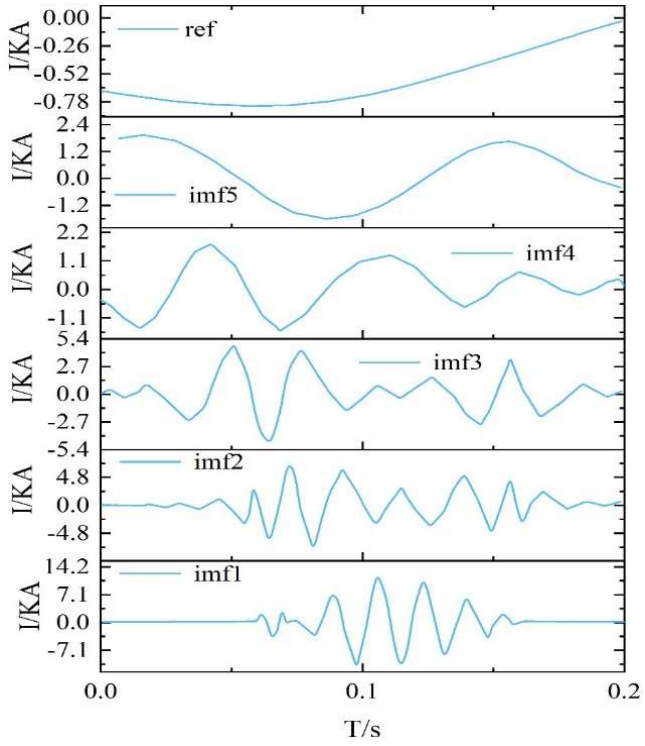


Figure 6: Improved EMD results

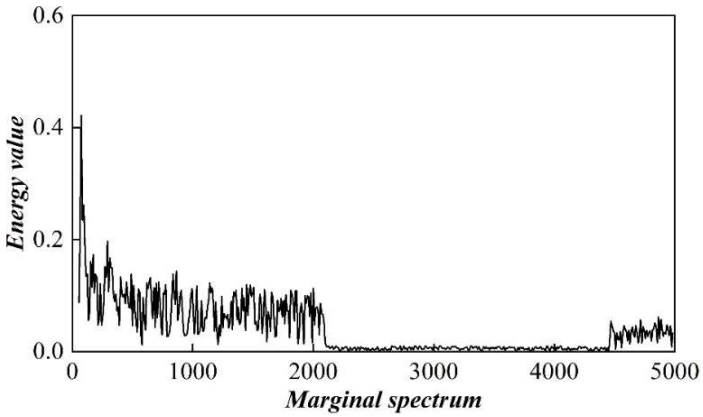


Figure 7: Marginal spectral analysis

Table 1: Example 1 diagnosis fusion results

Line	L_{16-17}	L_{16-21}	L_{16-24}	L_{15-16}	L_{16-19}	μ
ADD	0.5772	0.1055	0.0952	0.0741	0.0482	0.12
FDD	0.6011	0.0833	0.0722	0.0688	0.0688	0.13
FFD	0.4949	0.3349	0.0000	0.0000	0.0000	0.17
Melding	0.9292	0.0501	0.0111	0.0009	0.0069	0.098
C_1	1.0×10^{-8}	0.7452	0.8058	0.8158	0.8225	——
C_2	0.8088	6.88×10^{-4}	4.455×10^{-5}	7.355×10^{-5}	1.12×10^{-4}	——

Based on the switching fault diagnosis method, during the accident, there are 7 protection actions, 6 circuit breaker trips between, 5 lines in the outage area, where the protection and circuit breaker mis-activation. As the timing information of the misoperation $L_{(16)-21p}(422ms)$ and $CB_{(16)-21}(440ms)$ is close to the normal timing, the timing weighted fuzzy colored Petri net model cannot give low confidence to the misoperation information through the timing information, which leads to misdiagnosis, as shown by the relative fault characterization probability of

TWFCPN in Table 1, and the non-faulty component L_{16-21} has a high fault probability characterization. After the fusion of electrical and switching quantities, the fault probability characterization of faulty element L_{16-17} is increased compared to that before fusion, while the fault probability characterization of non-faulty element L_{16-21} is decreased, which improves the accuracy of fault diagnosis.

III. C. 2) Example 2

A short-circuit fault occurs on line L_{26-27} and line L_{17-27}, main protection $L_{(26)-27m}(21ms)$ and $L_{(27)-26m}(18ms)$ operate, circuit breaker $CB_{(27)-26}(71ms)$ trips, but $CB_{(26)-27}$ refuses to operate, line remote backup protection $L_{(25)-26s}(844ms)$, $L_{(27)-17}(438ms)$, and $L_{(29)-26s}(851ms)$ operate, tripping circuit breakers $CB_{(25)-26}(908ms)$, $CB_{(28)-26}(892ms)$, and $CB_{(29)-26}(899ms)$, main protection $L_{(17)-27m}(20ms)$ operates, but main protection $L_{(27)-17m}$ refuses to operate, backup protection $L_{(27)-17p}$ loses its information, and circuit breakers $CB_{(17)-27}(71ms)$, and $L_{(27)-17}(438ms)$ trip.

According to the received alarm information, after the fault region search to get the possible fault components set of $\{L_{26-27}, L_{26-29}, L_{26-28}, L_{25-26}, L_{17-27}\}$. Based on the TWFCPN model and improve the HHT to get the failure probability characterization of each component, and then according to the improvement of the D-S evidence theory for fusion, and after the C-mean method of decision-making diagnostics to get the clustering stabilization of the components of each component after the cluster distance from the center of the clusters of the error squared distance, the algorithm of the 2 diagnostic fusion results are shown in Table 2. The center value of cluster C_1 calculated by the decision model is 0.4833, and the center value of cluster C_2 is 0.0088, then C_1 is a faulty cluster. From the table it can be seen that components L_{26-27} and L_{17-27} distance from the center of cluster C_1 error squared distance is much smaller compared to cluster C_2, then L_{26-27} and L_{17-27} belong to cluster C_1, diagnosis for faulty components, diagnostic results are correct.

Table 2: Example 2 diagnosis fusion results

Line	L_{26-27}	L_{26-29}	L_{26-28}	L_{25-26}	L_{17-27}	μ
ADD	0.3245	0.1088	0.0778	0.0811	0.3105	0.12
FDD	0.4418	0.0288	0.0206	0.0051	0.4011	0.13
FFD	0.5348	0	0	0	0.3233	0.17
Melding	0.5419	0.0106	0.0068	0.0079	0.4223	0.098
C_1	3.588×10-3	0.2253	0.2271	0.2268	3.578×10-3	——
C_2	0.2888	3.22×10-6	1.1×10-6	6.2×10-7	0.1518	——

Based on the switching fault diagnosis method, during the accident, there are six protection actions, six circuit breaker trips between them, and five lines are in the blackout area, in which the protection information is lost. In the case of information loss, based on the time-weighted fuzzy colored Petri net model can get diagnostic results, but the corresponding failure probability of fault components is not obvious characterization, and can not completely determine whether the components are faulty or not, such as the FFD fault characterization probability shown in Table 2, the probability of failure of fault component L_{17-27} fault characterization of small. After the fusion of electrical and switching quantities, the L_{17-27} fault probability characterization of the backup protection information loss improves very significantly compared to the fault probability characterization obtained by FDD, which makes the accuracy of the fault diagnosis results improve.

IV. Multi-source information fusion for grid condition assessment

IV. A. Data pre-processing

Grid in the operation process is often vulnerable to their own and external factors, the most typical of their own factors is the aging of the equipment, and the most typical external factors are the environment in which the equipment is located and the weather conditions, which leads to the measurement of the data have more interference data and thus bring about a greater error. At the same time, another important reason for the incomplete acquisition of data is the communication in harsh environments. Therefore, in order to ensure an accurate assessment of the state of the equipment and to ensure the quantity and quality of the acquired data, it is necessary to carry out appropriate pre-processing of the data.

In the pre-processing of the collected data, it is necessary to do the following aspects of work:

(1) Data organization, first of all, to achieve data cleanup, including the smoothing of noisy data, the addition of vacant items and the correction of data inconsistency, and secondly, the need for data integration, i.e., the integration of the collected data and storage.

(2) Data transformation.

(3) Data statute, the purpose of which is to take some effective methods to realize the compression of data, such as clustering methods.

IV. B. Grid O&M state assessment of state quantities

IV. B. 1) Time series autoregressive modeling

Due to the low dynamics of electrical equipment during its operation, autoregressive models (AR) with strong memory of time series are widely used in the assessment. The state quantities of electrical equipment in normal operation can be divided into two categories, one category can be realized to be directly fitted by AR(1) because it already belongs to a smooth sequence, such as ground current; the other category of state quantities, although they will undergo periodic changes, are small in magnitude so that they can be fitted by AR with adjustments, such as oil temperature. Thus, the first-order AR model is used to fit the equipment operating state data with the following formula:

$$x_t = \alpha x_{t-1} + e_t = \alpha^t x_0 + \sum_{i=0}^{t-1} \alpha^i e_{t-i} (\alpha < 1) \tag{14}$$

In Eq. x_t denotes the time series of online monitoring data, e_t is the white noise of this state quantity obeying a normal distribution, and $e_t \sim N(\mu_e, \lambda^2)$, and therefore x_t obeys a normal distribution of $N(\mu, \sigma^2)$, where parameters μ and σ satisfy the following equation:

$$\mu = \mu_e / (1-\alpha) \tag{15}$$

$$\sigma^2 = (\alpha^2 \mu^2 + \lambda^2 + \mu_e^2) / (1-\alpha^2) \tag{16}$$

When a device is in normal operation, all of its state quantities should be within the corresponding thresholds, so that for all independent variables t, it is assumed that x_t is within interval $[a,b]$, i.e., $a \le x_t \le b$.

For all $a \le x_{t+k} \le b$, it can be derived:

$$a - \alpha^k x_t \le e_{t+k} + \alpha e_{t+k-1} + \cdots + \alpha^{k-1} e_{t+1} \le b - \alpha^k x_t \tag{17}$$

Since $e_t \sim N(\mu_e, \lambda^2)$, it follows from Eq. (17) that the whole sequence can be satisfied to belong to the interval $[a,b]$ only if α is less than a restriction α_0.

However, the AR model has its limitations, which are manifested in the inability to accurately detect anomalous conditions of the monitoring data outside the state volume threshold. This is mainly due to the inconsistency between the data and the state of the equipment. The gradual degradation of equipment performance during operation and the existence of potential failures make the data parameters of equipment in abnormal states often within the thresholds of the guidelines, which leads to incorrect state evaluation.

IV. B. 2) SOM quantification of time series

Self-organizing neural network (SOM) belongs to a kind of neural network, which is mainly used in the fields of data clustering and data dimensionality reduction [27]. The model is trained by inputting sample data in the input layer, and the weights are modified by continuous input data during the training process to make the model stable. Then real-time data is input, so that the neurons in the output layer and the neurons in the input layer are compared, and finally only one neuron in the output layer becomes the winner of the competition, and the output of this acquired neuron represents the classification of the input pattern, and it can be seen that SOM neural network belongs to the unsupervised learning.

Since SOM enables unsupervised classification, the entire sequence x_t is used as the input to SOM and sequence $c = \{C_1, C_2, \ldots, C_n\}$ is used as the output of the network, then each x_t is trained to belong to node C_j by the formula:

$$j = i(x_t) = \arg\min d(x_t, C_i(t)) \tag{18}$$

After continuous input of data, the weights are modified to minimize the distance of x_t from its final output node as shown in equation (19):

$$\begin{cases} C_i(t+1) = C_i(t) + \gamma(t)[x_t - C_i(t)], i \in N_j(t) \\ C_i(t+1) = C_i(t), i \notin N_j(t) \end{cases} \tag{19}$$

where $\gamma(t)$ is the learning rate of the SOM neural network, which takes a value between 0 and 1, and it decreases as t continues to increase.

The sequence of state quantities to be evaluated, x_t, is fed into the SOM neural network model after the classification process, and x_t becomes a time series of discrete points in a linear space, $C_t \in \{C_1, C_2, ..., C_N\}$. i.e:

$$C_t = C_{i(x_t)} \tag{20}$$

C_t denotes the node closest to x_t at moment t, so C_t completes the discretization of time series x_t.

IV. B. 3) Big data analysis during time series changes

Another feature of SOM neural network is that the nodes in the competitive layer are correlated with each other, and the relationship between the input data can be reflected by the network topology. In the network topology, each neuron node has a strong correlation with nodes within its neighborhood and a weak correlation with those outside the domain due to the competition during SOM training. It is due to this feature of SOM neural network that the quantized time series C_t can be regarded as a transfer from one neuron to another in the topology, thus tapping into the change rule of data over time.

1) Probability density function of neurons

If the correlation between neurons is expressed in terms of the first-order transfer probability P, then the first-order transfer probability between neurons in the AR(N) model is $P[c_{t+1} | c_t, c_{t-1}, \ldots\ldots, c_1]$, so the first-order transfer probability between neurons in the AR(1) model can be simplified to $P[c_t + 1 | c_t]$.

Let the value of $\{C_1, C_2, ..., C_N\}$ be $\{1, 2, ..., N\}$, then the probability that the value of c_t is C_I at the moment of t can be introduced by Eq:

$$P[c_t = C_l] = P[i(x_t) = I] \tag{21}$$

The probability density function of $i(x_t)$ can be derived as:

$$\begin{cases} P[i(x_t) = I] = P[\arg\min \| x_t - C_i \| = I] \\ P[\| x_t - C_I \| \leq \| x_t - C_j \|, \forall j \neq I \end{cases} \tag{22}$$

The $\| \cdot \|$ in Equation (22) denotes the Euclidean distance.

Since both X and C are one-dimensional arrays, Equation (22) can be expressed as:

$$P[i(x_t) = I] = P[| x_t - C_I | < | x_t - C_{I-1} |] \wedge P[| x_t - C_I | < | x_t - C_{I+1} |] \tag{23}$$

Let $a = (C_l + C_{l+1})/2$ and $b = (C_l + C_{l-1})/2$, since x_t obeys a normal distribution, the probability distribution function of x_t can be expressed as a standard normal distribution function:

$$\begin{aligned} P[i(x_t) = I] &= P[b < x_t < a] \\ &= F_x(a) - F_x(b) \\ &= \phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right) \\ &= f(I, \alpha, \mu_e, \lambda) \end{aligned} \tag{24}$$

When $I = 1$:

$$P[i(x_t) = I] = \phi\left(\frac{a-\mu}{\sigma}\right) \tag{25}$$

When $I = N$:

$$P[i(x_t)=I]=1-\phi(\frac{b-\mu}{\sigma}) \tag{26}$$

2) Transfer probability between neurons

The AR model has low dynamics, so the transfer probability between neurons that are close to each other is large, while the transfer probability between neurons that are far apart is small. The second order probability distribution function can be expressed as:

$$P[c_t=C_{I1},c_{t+k}=C_{I2}]=P[x_t\in(b_1,a_1),x_{t+k}\in(b_2,a_2)] \tag{27}$$

Eq. $C_{I1},C_{I2}\in\{C_1,C_2,...,C_N\},I_1=(a_1,b_1),I_2=(a_2,b_2)$. Since x_t belongs to the normal distribution, the secondary normal distribution function of x_t:

$$\int_{b_1}^{a_1}\int_{b_2}^{a_2}\frac{1}{2\pi\sigma^2\sqrt{1-\rho^2(k)}}e^{\left[\frac{(x-\mu)^2+(y-\mu)^2-2\rho(k)(x-\mu)(y-\mu)}{2\sigma^2}\right]}dxdy \tag{28}$$

where $\rho(k)=\alpha^k$, denotes the autocorrelation function of the first-order AR process.

In this model there is only one step transfer of neurons, and Eq. (27) can be reduced to:

$$P[c_{t+1}=C_{I2}\mid c_t=C_{I1}]=\frac{P[c_{t+1}=C_{I2},c_t=C_{I1}]}{P[c_t=C_{I1}]} \tag{29}$$

IV. C. Experimental results and analysis

IV. C. 1) Oil temperature and load test

Online monitoring data of 220kv main transformer during normal operation is taken from the local monitoring center to assess the condition of the equipment in terms of transformer oil temperature, equipment load, ambient temperature, etc., and the data for the year 2022 is taken as a sample from September 15 to September 22, and data from 14:00-24:00 of September 23 is taken as a sample to be tested. These state variables were collected at a rate of one per minute, totaling 500 sets. The parameters are λ=0.02, α=0.88, μ_e=0 after fitting by AR (1) model, and the number of neurons of SOM is set to 15, and the probability that the data of two neighboring moments belong to the same neuron is maximum. The transformer detection signal is shown in Fig. 8, and the equipment load and oil temperature have a smooth rising trend after 400 points.

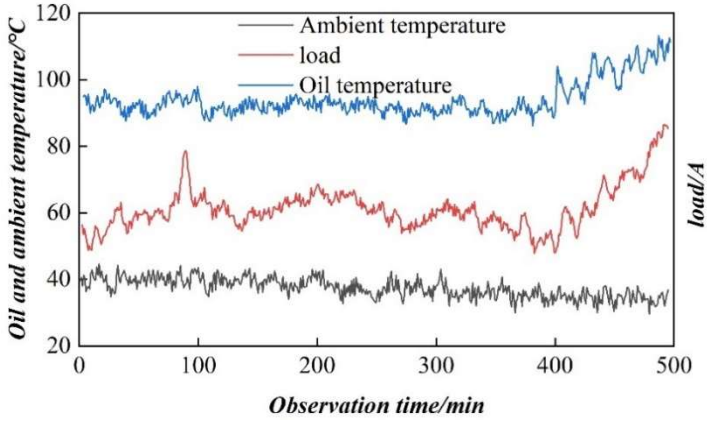


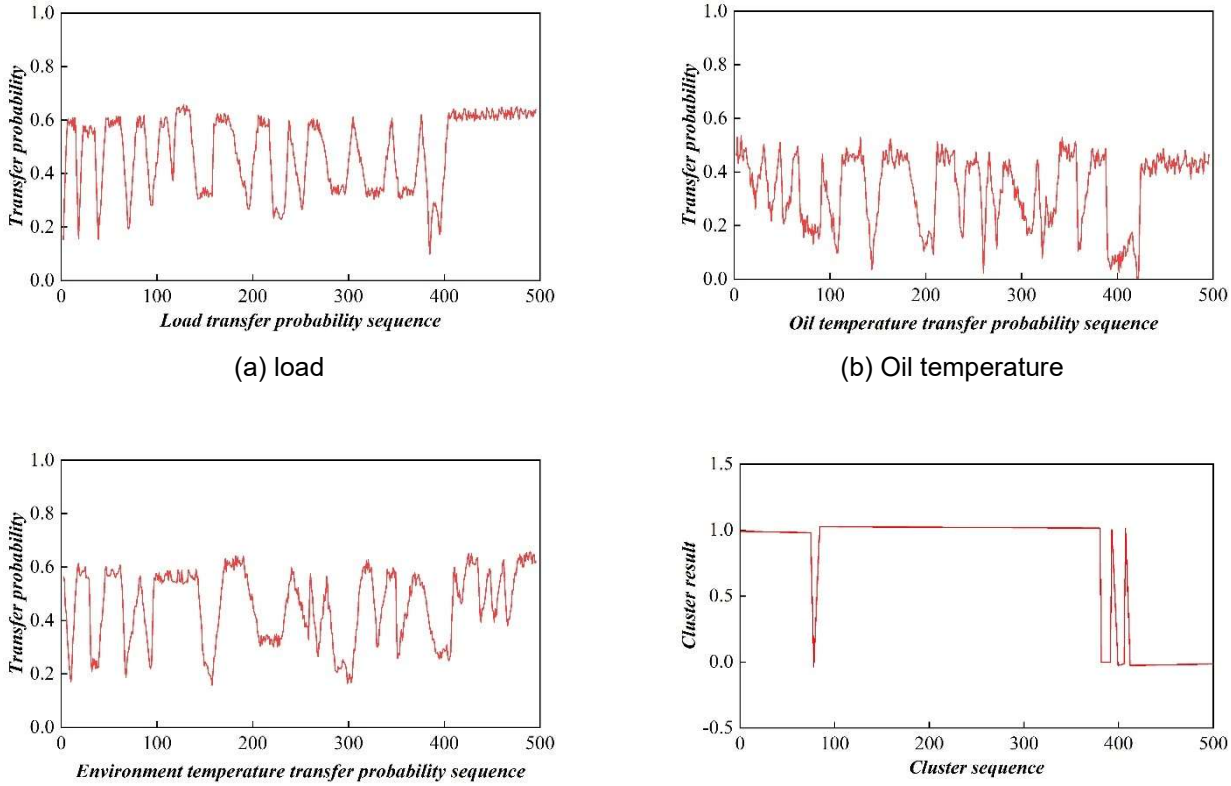
Figure 8: Detection signals of transformer

The results of anomaly detection by the method in the paper are shown in Fig. 9, (a) to (d) are the load, oil temperature, ambient temperature and clustering results, respectively. The following conclusions can be obtained:

(1) There is a 0 point of oil temperature at the observation moment 80min, the clustering results in this case also show abnormal values, but due to the high time transfer probability, the clustering results are normal, and it is judged to be caused by the sensor abnormality, which can be ignored.

(2) The transfer probability of oil temperature and load is 0 at the time of observation. The value of the transfer probability sequence fluctuates up and down at the observation moment (400-430)min, and there are many zero points. The clustering results also show that most of the observations at observation moments (400 to 500)min do

not belong to the conventional clustering. As a result, the operation is abnormal from the observation time at 385min, and there is a rapid increase in oil temperature and load. And deviation from normal values, at this time should be issued early warning of abnormal operating conditions.



(a) load

(b) Oil temperature

(c) Ambient temperature

(d) Cluster result

Figure 9: Abnormal detection of transformer

After reviewing the logs and records, at 20:20 on September 23, the transformer received a dispatch order to operate at a high load factor of 1.1~1.25 for 5 minutes before resuming normal operation at 9:30pm. The detection is the same as the actual, and the method in the paper is able to accomplish the transformer anomaly detection.

There is no threshold standard for oil temperature, load, and ambient temperature in transformer condition assessment. Using the threshold determination method, 1.25 times is selected based on experience. That is, if the measured data exceeds 1.25 times the normal value, it is judged as abnormal with the following results.

(1) Abnormal oil temperature at observation moment 85min.

(2) Abnormal oil temperature and load at the same time after 470min at the observation moment. The abnormality detection conclusion of the threshold judgment method lags behind the actual situation and cannot eliminate the misjudgment caused by the sensor abnormality.

IV. C. 2) Ice cover experiments

Ice cover data for 500kv transmission lines were selected at the local monitoring center. The ice cover is highest from November to February each year, and the data collected by the ice attachment device from September to November, when the year 2022 is under heavy ice cover, is used as a training sample. The transmission line ice cover monitoring data is shown in Figure 10. The data from December 2022 to January 2023 is the sample to be tested, and the sampling period of the device is one hour.

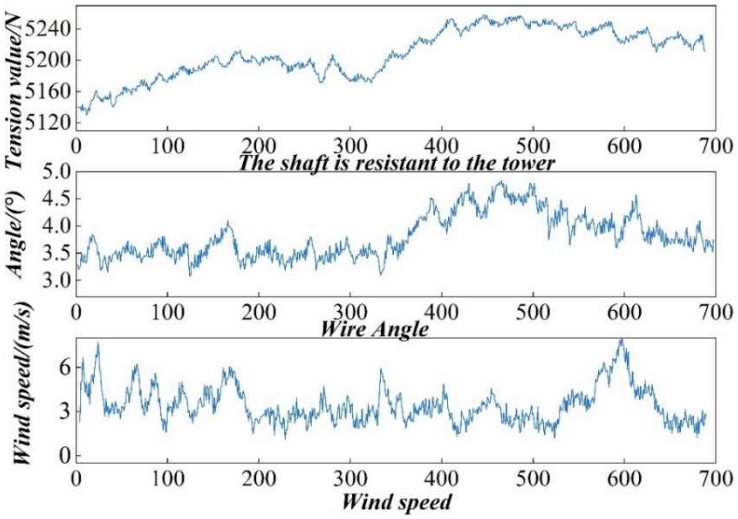
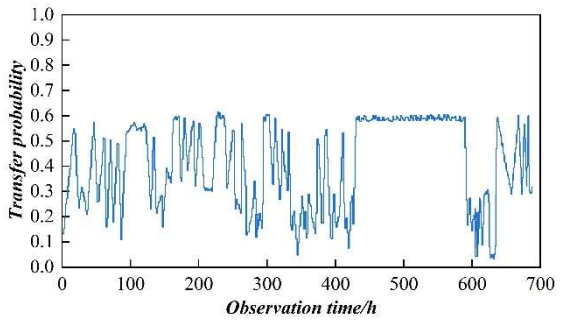


Figure 10: Monitoring data of transmission line icing

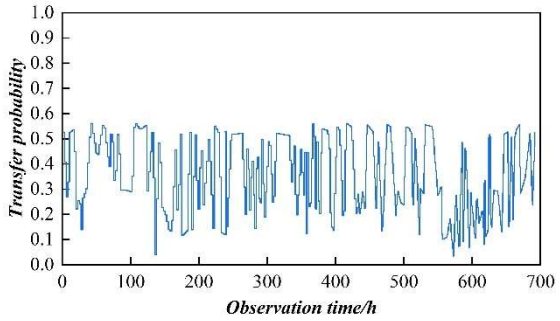
After training the SOM network and clusters, the data in Fig. 10 are used as inputs, and through the detection process in the paper, the transfer probability sequence and clustering results are obtained as shown in Fig. 11, with (a) to (d) being the axial direction of the tension-resistant tower, the tilt angle of the conductor, the wind speed, and the clustering results, respectively. The following conclusions can be obtained:

(1) At the observation moment 355h, there are a large number of data close to 0 or 0. In the observation moment (370~420) h section, the wire tension and inclination angle are almost all 0 points, from the clustering results, the data section does not belong to the conventional clustering, and there is a clustering abnormality in the observation moment (420~600) h section. Therefore, there is an ice-covering abnormality at the observation moment 370h, and it can be judged that the transmission line status is abnormal on that day.

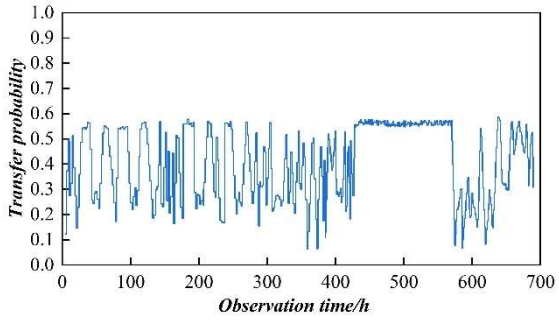
(2) The observation time has many zeros in the transformed probability series of conductor tension, inclination and wind speed in the (610~630)h section, and the clustering result is between normal and abnormal, which determines that the thickness of the ice cover is abnormally increased, the line is detected abnormally, and the daily ice cover is recorded. The ice thickness on the line has been about 2.5 mm since last December. Due to the weather (cold, light snow, fog) the ice thickness on the line has increased rapidly. During a special inspection of the coating on the ice on the 20th, the ice thickness near the monitoring equipment was found to be more than 11 mm. after the line inspector reported this to the authorities, the competent authority sent out an anomaly detection message, which was consistent with the results of the anomaly detection. The anomaly was detected after December 28th but the thickness of the overlying ice was only 9.1 mm, which could be caused by anomalies of tension due to the increase in the wind speed and not due to the overlaying of the ice. The line inspector was able to detect an anomaly in the ice thickness of about 2.5 mm.



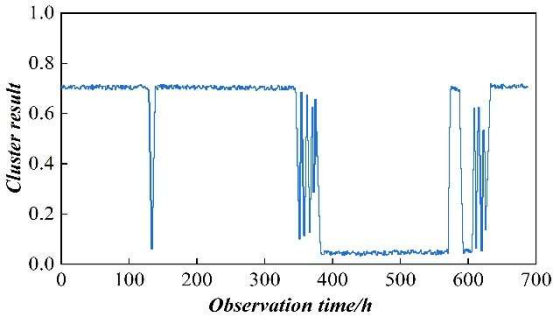
(a) The shaft is resistant to the tower



(b) Wire Angle



(c) Wind speed



(d) Cluster result

Figure 11: Abnormal detection of transformer

When assessing the status of transmission lines, the early warning value of the early ice cover thickness is set to 11 mm. based on this warning value, the threshold judgment method is used for anomaly detection. Figure 12 shows the time-varying curve of the ice cover thickness. The threshold anomaly appears only at the observation time 450h (corresponding to December 21), which is later than the method in the paper, and the threshold anomaly also appears in the observation time (585~620) h section, which is different from the actual one. It can be concluded that the method in the text is superior to the threshold judgment method.

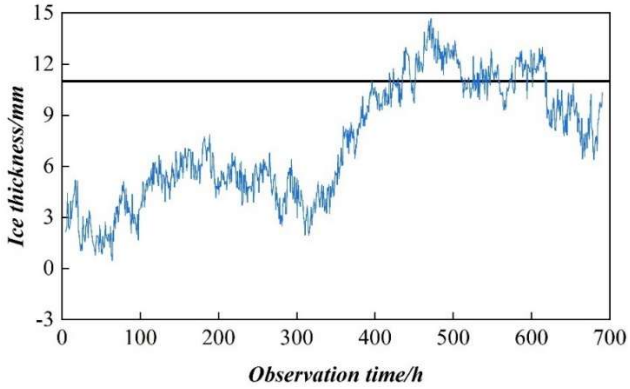


Figure 12: The time curve for the thickness of the ice

V. Conclusion

The study of grid informationized operation and maintenance system provides basic methods and technical support for constructing informationized operation and maintenance operation library, informationized operation and maintenance operation workload standard, and difference coefficients of operation volume of different types of systems, etc., which has important extended application value.

In this paper, a new grid information diagnostic model is designed based on the improved D-S evidence theory, and after the fusion of each body of evidence using information fusion technology, the fault probability characterization of faulty components increases, and the fault probability characterization of non-faulty components decreases. Due to its integrated consideration of switching and electrical information, it reduces the influence of uncertainty facets such as refusal to act, false action and loss of information on the fault diagnosis results, and makes the results more accurate.

In addition, it also combines big data analysis technology with equipment evaluation methods. After the pre-processing of multi-source data, AR time series model and SOM neural network are processed for a continuous state quantity to discretize the continuous data into individual sequences and calculate the transfer probability of the state quantity based on the time axis, and the state assessment model is tested by taking transformers and transmission lines as an example, which verifies that the designed state assessment general components can meet the needs of actual production.

References

[1] Moreno Escobar, J. J., Morales Matamoros, O., Tejeida Padilla, R., Lina Reyes, I., & Quintana Espinosa, H. (2021). A comprehensive review on smart grids: Challenges and opportunities. Sensors, 21(21), 6978.

[2] Soltan, S., Mazauric, D., & Zussman, G. (2015). Analysis of failures in power grids. IEEE Transactions on Control of Network Systems, 4(2), 288-300.

[3] Protalinskiy, O., Savchenko, N., & Khanova, A. (2019). Data mining integration of power grid companies enterprise asset management. In Cyber-Physical Systems: Industry 4.0 Challenges (pp. 39-49). Cham: Springer International Publishing.

[4] Suo, N., & Zhou, Z. (2021). Computer assistance analysis of power grid relay protection based on data mining. Computer-Aided Design and Applications, 18(S4), 61-71.

[5] Lu, R., Liu, N., Li, D., Luo, X., & Fan, Y. (2021, August). Intelligent monitoring analysis of power grid monitoring information based on big data mining. In Journal of Physics: Conference Series (Vol. 1992, No. 3, p. 032132). IOP Publishing.

[6] Liu, D. N., Jiang, X. F., & Zhang, S. Y. (2013). Operating Analysis and Data Mining System for Power Grid Dispatching. Advanced Materials Research, 787, 611-617.

[7] Dehghan Nezhad, M. T., & Sarbishegi, M. M. (2023). Data Mining Applications in Smart Grid System (SGS). In Handbook of Smart Energy Systems (pp. 1557-1573). Cham: Springer International Publishing.

[8] Shiomi, R., Shimasaki, H., Takano, H., & Taoka, H. (2019). A study on operating lifetime estimation for electrical components in power grids on the basis of analysis of maintenance records. Journal of International Council on Electrical Engineering, 9(1), 45-52.

[9] Karagiorgos, N., & Siozios, K. (2018). Data analytic for improving operations and maintenance in smart-grid environment. In IoT for Smart Grids: Design Challenges and Paradigms (pp. 147-161). Cham: Springer International Publishing.

[10] Xing, Y., Meng, C., Li, C., Xi, G., Jia, X., Bai, Y., ... & Zhang, Z. (2020, October). Lean operation and maintenance evaluation technology of power grid equipment based on improved big data cleaning method. In 2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2) (pp. 2749-2752). IEEE.

[11] Gu, Z., Zhu, M., Zhu, W., Zhou, J., Zhu, Q., & Wang, J. (2020, June). Optimal design of intelligent analysis and control system for power grid operation and maintenance. In 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC) (pp. 1445-1449). IEEE.

[12] Iftikhar, H., Sarquis, E., & Branco, P. C. (2021). Why can simple operation and maintenance (O&M) practices in large-scale grid-connected PV power plants play a key role in improving its energy output?. Energies, 14(13), 3798.

[13] Wang, J. (2021, October). Research and Application of Operation and Maintenance Simulation Technology of Provincial Intelligent Distribution Network Based on Data Mining. In 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (Vol. 5, pp. 1247-1251). IEEE.

[14] Yang, X., Yu, M., & Liu, F. (2022, January). Construction of power network operation and maintenance cost prediction model based on data information mining. In 2022 International Conference on Big Data, Information and Computer Network (BDICN) (pp. 124-127). IEEE.

[15] Lydia, E. L., Kumar, B. P., & Ramya, D. (2018). Generation of dynamic energy management using data mining techniques basing on big data analytics isssues in smart grids. International Journal of Engineering & Technology, 7(2.26), 85-89.

[16] Wang, H., Liu, Z., Xu, Y., Wei, X., & Wang, L. (2020). Short text mining framework with specific design for operation and maintenance of power equipment. CSEE Journal of Power and Energy Systems, 7(6), 1267-1277.

[17] Guan, W., Chen, H., & Zhang, J. (2024, April). Data Mining Methods and Intelligent Analysis Application for Foundation Treatment in Power Grid Engineering. In 2024 Photonics & Electromagnetics Research Symposium (PIERS) (pp. 1-7). IEEE.

[18] Zhao, L., Feng, Z., Wang, N., Zhu, J., & Li, L. (2023, April). Big Data Mining Analysis of Power Grid Based on Apriori Optimization. In Journal of Physics: Conference Series (Vol. 2476, No. 1, p. 012088). IOP Publishing.

[19] Zhou, H., Liu, D., Li, D., Shao, G., & Li, Q. (2013). Operating Analysis and Data Mining System for Power Grid Dispatching. Energy and Power Engineering, 5(4), 616-620.

[20] Zhao, X., Luo, L., Ma, G., Cai, Z., Gu, Z., & Wang, Q. (2018, November). Operation and Maintenance Management and Decision Analysis in Distribution Network Based on Big Data Mining. In 2018 International Conference on Power System Technology (POWERCON) (pp. 4855-4861). IEEE.

[21] Liu, F., Ying, Q., & Zhang, C. (2024, May). A Data Mining-based Method for Analyzing and Predicting the Cost of Power Grid Production and Technical Improvement Projects. In 2024 3rd International Conference on Energy, Power and Electrical Technology (ICEPET) (pp. 829-832). IEEE.

[22] Grigoras, G., Neagu, B. C., & Scarlatache, F. (2023). Data Mining-Based Approaches in the Power Quality Analysis. In Smart Grid 3.0: Computational and Communication Technologies (pp. 93-119). Cham: Springer International Publishing.

[23] Saleh, A. I., Rabie, A. H., & Abo-Al-Ez, K. M. (2016). A data mining based load forecasting strategy for smart electrical grids. Advanced Engineering Informatics, 30(3), 422-448.

[24] Zhao, M., Wang, B., Zu, Y., Wang, D., & Liu, Z. (2023, July). Data mining and analysis for defects of secondary equipment in power grid. In Third International Conference on Mechanical, Electronics, and Electrical and Automation Control (METMS 2023) (Vol. 12722, pp. 455-462). SPIE.

[25] Chen, G., Qiu, J., Pan, T., Niu, D., Pu, D., Zhao, Z., ... & Qin, L. (2018, May). Research on Data Mining Technology of Company Operation Monitoring. In 8th International Conference on Social Network, Communication and Education (SNCE 2018) (pp. 72-80). Atlantis Press.

[26] Liang Xing,Luo Yuanxing,Deng Fei & Li Yan. (2022). Application of Improved MFDFA and D-S Evidence Theory in Fault Diagnosis. Applied Sciences,12(10),4976-4976.

[27] Jiyang Zhu & Xue Han. (2024). Big data clustering algorithm of power system user load characteristics based on K-means and SOM neural network. Multimedia Tools and Applications,84(10),1-15.