Article citation info:
Wang S, Zhang Y, Li Y, Gao S, Yang F. The mining of high risk equipment based on the algorithm of HR-Tree's decision. Diagnostyka. 2020;21(2):51-59. https://doi.org/10.29354/diag/122341

DIAGNOSTYKA, 2020, Vol. 21, No. 2

ISSN 1641-6414
e-ISSN 2449-5220
DOI: 10.29354/diag/122341

THE MINING OF HIGH RISK EQUIPMENT BASED ON THE ALGORITHM OF HR-TREE'S DECISION

Shanyuan WANG, Yujie ZHANG, Yao LI, Suisui GAO, Feiling YANG
Northeastern University, College of Information Science and Engineering,
Institute of Electrical Automation, 110004, wsyneu@163.com

Abstract

Due to the different construction of various subsystems in the power grid, the information of various systems are not closely connected. Nowadays, the network is complex and changeable where the automation is getting higher. This article takes high-risk equipment set of substation in Liaoyang as the research background. It constructs HR-Tree for the device set, and establishes a high-risk equipment evaluation system which based on the HR-Tree context. Then we calculate high-risk equipment sets in the structure of overall data set. By establishing the original data set and the prior knowledge system of equipment risk, the non-candidate high-risk equipment set is reduced in the local path of the high-risk equipment set. We refer to the process of reducing data as minus branch. After the threshold is established, the branches are reduced and the highest risk equipment set is obtained. Finally, we use the scoring system to find the probability of occurrence of associated devices, such information is more open. Example showed that such methods could effectively express high-risk device sets, and managers could get early warning information based on this. It helps people monitoring the power system, w hich could also provides new ideas for the monitoring project.

Keywords: high-risk equipment set, HR-Tree, minus branch, early warning

1. INTRODUCTION

State Grid Corporation of China includes many systems for collecting various types of information. These subsystems in the dispatch control system have great limitations in terms of data-exchanging and data-sharing, where a unified connection has not yet been established for the multi-source data of each system. The situation cannot meet the needs for development of automated application systems and the integration of smart grid information [1].

In the process of multi-dimensional information fusion, this article establishes an overall framework. The system will reasonably allocate information to categories and provide scientific theoretical basis according to the needs of on-site dispatch.

In this article, the establishment of an HR-Tree information network which based on the framework of fused information is helpful for information analysis. It is more efficient to explore high-risk equipment on the fuse platform of the power grid information.HR-Tree is one of the main methods for testing system's safety whose causality is clear. It can intuitively and comprehensively reflect the internal mechanism of faults. After analyzing the basic events, the contribution rate of basic events to faults can be obtained [2].

The research of industrial multi-dimensional data networks involves high-risk decision trees in many fields. HR-Tree generates high-risk equipment sets and reduces candidate branches to obtain the highest-risk equipment set. In addition, the method of this paper explores the association between high-risk equipment sets and then obtains the rules of failure occurrence, which are mainly reflected in the score of related equipment failures.

As an effective and basic method, system safety evaluation has been gradually applied to various engineering fields [3]. Traditional evaluation methods such as fuzzy comprehensive evaluation rely on a large amount of historical experience and expert opinions [4]. The literature [5] presents results on a methodology for high-power EM based risk assessment of large structures considering the example of smart grid substations. The methodology developed in this paper evaluates the threat, vulnerability, impact, and protective measures as indicators in various scenarios of both conducted and radiated intentional EM interference (IEMI) threats to these systems. In literature [6], Data-mining analysis was carried out using the C4.5 decision tree algorithm for the aforementioned three events using five different splitting criteria. The literature [7] introduce the APRICOIN algorithm, which combines frequent pattern mining and a fuzzy logic system, to assess the container's risk score. The frequent pattern growth algorithm is proposed to retrieve the key criteria for evaluating container risk. In literature [8], It presents a scientific literature information extraction architecture using text mining techniques to assess the human health risk of electromagnetic fields (EMFs) generated by wireless sensor devices in Internet of Things. To extract high-quality patterns in real-life applications, this literature [9] extends the occupancy measure to also assess the utility of

patterns in transaction databases. Research on power equipment and fault assessment methods at home and abroad has grown vigorously in the past 20 years and has formed a scientific theoretical basis.

Literature [10-11] obtained fault diagnosis classification by studying massive monitoring data of transformers, and proposed a collaborative variable prediction model based on Spark computing framework. Literature [12-13] used deep learning models for transformer fault diagnosis, while abandoning the shortcomings of traditional DBN-based deep learning models, and then proposed adaptive improvements. Finally, an adaptive deep learning model transformer fault diagnosis method was obtained. Literature [14-15] introduced the calculation of the critical importance of each component under each fault and the order of troubleshooting by using the T−S fuzzy gate algorithm, and combined with the component fault self-diagnostic program. The hybrid reasoning based on fuzzy fault tree analysis in literature [16-17] uses a targeted artificial intelligence search algorithm, which will more accurately search for the fault location of the system and provide a solution for fault analysis from multi-source data sources.

The exploration methods for faults at home and abroad have gradually developed into intelligence in recent years. The rise and development of data mining analysis methods has opened up a new technical line for the evaluation and fault diagnosis of power equipment conditions. Data mining in industry proposed high requirements for more parameter's information of equipment condition[18]. In recent years, the algorithms about data mining analysis have been used in a variety of situations in industry.

Although the innovation of digital technology had brought convenience to system monitoring and equipment management, the integration of digital information in the power system has not yet matured. This is mainly because the framework of information fusion is not yet constructed, so the efficiency of information utilization has not yet meet the requirements. On dispatching department under the State Grid Corporation of China, multi-professional knowledge can be presented on several platform in a short period of time, but the manual release of commands is delayed in time. In summary, this paper reduces branches between various data sets, and the high-risk equipment set which obtained after branch reduction has guiding significance for the dispatch center. The method of this paper calculates the relationship between various data sets, and the estimated occurrence of the fault of device. Therefore, the research on electric power data mining still has great development prospects.

2. EVALUATION METHOD OF CHARACTERIZATION FROM EQUIPMENT RISK

Facts have proved that equipment failure often were the direct cause of power outages and the key factor in the expansion of accidents, so equipment risk is one of the cores of grid risk assessment. As shown in Figure 1, equipment risk impact is composed of equipment importance and equipment hidden dangers, and they are respectively composed of equipment cost, voltage level, level of power supply area, related scale, alarm level, impact of failure , and maintenance frequency.

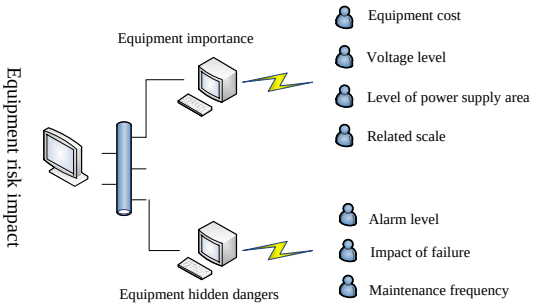

Fig. 1. The system of risk-assessment

As shown in Figure 1, equipment risk impact is mainly composed of equipment importance and equipment hidden dangers. The interaction between equipment importance and equipment hidden dangers finally determines the final risk value of equipment. Equipment importance and equipment hidden dangers can be divided into seven section in detail:

A, Equipment cost. From the economic point of view, the more expensive the equipment, the more important it is. Its level has a clear guiding effect on the evaluation equipment.

B, Voltage level. The basis of the establishment of the voltage level is the magnitude of the impact on the environment and residents after the failure of high-voltage equipment. The level is established by the magnitude of the loss caused by the environmental blackout.

C, Level of power supply area. The magnitude of the load is proportional to the risk factor. The higher the factor of risk, the greater the administrative level of the department in which the equipment is located. The high level of power supply area represents its high risk.

D, Related scale. In the process of equipment failure, the scale of the affected equipment is counted to determine the important impact of the equipment.

E, Alarm level. The alarm characterization form of the device is the weighted summation of each alarm level. Its formula is as follows:

$$WR = \sum_{i=1}^{t} w_i K_i \qquad (1)$$

In the formula, w_i is set as the alarm level, K_i is set as the frequency of occurrence of a certain alarm level, and t is the number of alarm levels. The larger the value of WR.The greater the alarm level, the greater the degree of risk damage.

F, The equipment's fault characterization is a weighted summation of various fault levels, and its formula is as follows:

$$GR = \sum_{i=1}^{s} g_i T_i \qquad (2)$$

In the formula, g_i is set as the fault level, T_i is set as the frequency of occurrence of a certain alarm level, and s is the number of alarm levels.

G, Maintenance frequency. Equipment maintenance is divided into planned maintenance and maintenance after failure. The number of maintenance times objectively represents the of equipment hidden dangers.

These non-quantifiable indicators are shown in Tables 1 to 3:

Table 1. Standard of equipment cost

Index	Level (Unit: 1,000$)	Level	Quantified value
Equipment cost	>100	Important	9
	100~50	More important	7
	50~20	medium	5
	20~10	Less important	3
	<10	minor	1

Table 2. Voltage level(unit: *kV*)

Voltage level	<66	110	220	330	>500
Level	Slight	Lighter	General	serious	Very
Quantified value	1	3	5	7	9

Table 3. Level of power-supply area

Index	Zone / ID	Level	Quantified value
Regional level	Factory /A	Important	9
	Mall/B	More important	7
	Shool/C	medium	5
	Park/D	Less important	3
	Suburbs /E	minor	1

The process is based on expert evaluation and manual weighting. The equipment importance is the same as the evaluation method of equipment hidden dangers. According to the actual situation, only these two indicators Equipment cost and Voltage level are static, while other indicators (including Level of power supply area, Related scale, Alarm level, Impact of failure, Maintenance frequency)are dynamically changing, which is also caused by the physical structure of the device itself.

Half-ladder model includes Half-lift ladder model and Half-step ladder model. In order to facilitate subsequent calculations, the quantized data is subjected to isotropic processing, that is, the quantized results are all located in the interval [0,1], in Half-lift ladder model, the smaller the value, the better the corresponding state and running status.

The expression of the Half-lift ladder scoring model is:

$$T(\mathrm{x}) = \begin{cases} 0, & 0 \le x \le a \\ (\mathrm{x}-\mathrm{a})/(\mathrm{b}-\mathrm{a}), & a \le x \le b \\ 1, & x \ge b \end{cases} \qquad (3)$$

The expression of the Half-step ladder scoring model is:

$$T(\mathrm{x}) = \begin{cases} 1, & 0 \le x \le a \\ (\mathrm{b}-x)/(\mathrm{b}-\mathrm{a}), & a \le x \le b \\ 0, & x \ge b \end{cases} \qquad (4)$$

We quantify each index using methods such as formula (3) and formula (4). Such algorithms are more accurate and fair which are widely used in engineering applications. Among them, a, b are thresholds; m is a quantized parameter value. According to the relevant regulations and maintenance experience settings, the revision of the threshold and parameter values of each state quantity be found above.

2.1. Assessment system in equipment risk impact

The above related equipment set is defined as: $T = \{t_1, t_2, ..., t_s\}, s \in N$, N is the number of devices, t_s is the device name, and under the index h_i, the data of the device can be quantified.

According to the establishment of each characterization of equipment risk in Section 2.1, the mathematical model for establishing the evaluation system is shown in formula (5):

$$FR_i = SI_i \times YI_i \qquad (5)$$

In the formula, FR_i represents equipment risk impact of device i, SI_i represents the equipment importance of device i, and YI_i represents equipment hidden dangers of device i. The indicator set of characterization from equipment importance is $W = \{h_1, h_2, ..., h_i\}$.In the indicator set, i is the number of member from equipment importance, which are the equipment cost, voltage level, level of power supply area, and Related scale.

Under the quantification of the index h_i, the equipment relative importance matrix is shown in formula (6):

$$G_i = \begin{matrix} t_1 \\ t_2 \\ \cdots \\ t_s \end{matrix} \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1N} \\ g_{21} & g_{22} & \cdots & g_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ g_{N1} & g_{N2} & \cdots & g_{NN} \end{bmatrix} \qquad (6)$$

As shown in formula (6), g_{ii} represents the relative importance value of the device t_i under the corresponding index h_i. The definition of relative importance in formula (6) is shown in formula (7):

$$g_{ij}=\begin{cases}2 & t_i > t_j;\\ 1 & t_i = t_j;\\ 0 & t_i < t_j;\end{cases} \quad (7)$$

After obtaining the sum G_i of the row vectors of the same index, the high-level relative importance under the index h_i is obtained: $g_i^{h_N}=\sum_{j=1}^{N} g_{ij}^{(h_N)}$. At the same time, the maximum eigenvalues: λ_i and corresponding eigenvectors: $X_i=(x_1,x_2,...,x_i)$ are obtained for the relative importance matrix between indicators: $X_i=(x_1,x_2,...,x_i)$. After that, the feature vector could be normalized to get the standard feature vector: $X_i^{'}=(x_1^{'},x_2^{'},...,x_i^{'})$, The relative importance of the equipment is obtained by formula (8), namely:

$$SI_i = X_i^{'}\times g_i^{h_N} \quad (8)$$

This article normalizes the relative importance calculated by each device: SI_i.

2.2. Definition of equipment high risk set

The risk value of device t_i is recorded as $R(t_i)$, which is the product of the support degree of the device: $\sup(t_i)$ and equipment risk impact: $FR(t_i)$, as shown in formula (9):

$$R(t_i)=\sup(t_i)\times FR(t_i) \quad (9)$$

Define the minimum equipment risk threshold as $\min-R$, and eliminate the risk equipment smaller than $\min-R$ in $R(t_i)$ of several devices. Let $J=\{j_1,j_2,...,j_s\}, s\in N$ be the collection of faulty things in the research carrier, and $j_i=\sum_{i=1}^{p}(t_i,c_i)$. c_i represents the frequency of occurrence of t_i in single transaction j_i, and p is the total number of transactions.

This research method defines the equipment risk value of different attributes:

Definition 1: Calculate the equipment risk value of a single device t_i, and multiply the equipment risk impact of a single device by the frequency of the failure transaction c_i, as shown in formula (10):

$$R_0 = FR\times C_i \quad (10)$$

It calculates the equipment risk value of a single incident on the equipment set, s is the number of set including the incident, as shown in formula (11):

$$R_1=\sum_{i=1}^{s} R_{0i} \quad (11)$$

Calculate the equipment risk value of the total accident set of all equipment set, as shown in formula (12):

$$R_2=\sum_{i=1}^{p} R_{1i} \quad (12)$$

The sum of the risk values of all things for a single device t_s is defined as R_3:

$$R_3=\sum_{i=1}^{p} R_{0i} \quad (13)$$

Definition 2: The minimum risk threshold of the equipment is $\min-R$, which is a proportional value of R_2, and the proportional coefficient is η, as shown in formula (14):

$$\min-R = R_2\times\eta \quad (14)$$

the equipment risk value of a single incident on the equipment set in the article should meet the conditions: $R_1\geq\min-R$.

2.3. HR-Tree mining algorithm based on equipment risk impact data

The HR-Tree algorithm describes a tree structure for classifying specific parameters. The algorithm moves down recursively until it reaches the leaf node, and finally assigns the instance to the class of the leaf node [19].

The construct of the HR-Tree for the equipment risk impact was obtained in Section 2.2.The information of each equipment fault set name each node and connection point of the tree as H.name, H.count, H.link, H.parent, H. risk. H.name refers to the node name and device number. H.count refers to the number of branches passing through the node, and H.link refers to the connection with the same device number but not dependent on the existence of a branch. H.parent refers to the parent node common to a single node, and H.risk refers to the accumulation of risk value of a single node. The algorithm flow of HR-Tree is shown in Figure 2.

It can be seen from Figure 2 that the HR-Tree starts from the integration of the original data, which obtains the initial equipment risk set through the information screening of rule 1.It eliminates the low-threshold equipment set through rule 2 and the equipment risk threshold. This article defines the first rule, second rule and pruning of HR-Tree as follows:

Rule 1: If the risk value of a certain item of equipment does not meet the minimum risk threshold, the inclusion item of the device is eliminated. After a certain device set eliminates some branches, the branches need to be rebuilt.

Rule 2: If a certain device set is determined to be a high-risk set, any subset of the device set meets the minimum threshold of the minimum risk. This principle is also called downward closure.

pruning process:

Step 1: Perform risk value evaluation on the initial data of the equipment to obtain the initial risk equipment set.

Step 2: Calculate the minimum risk threshold as A and construct the global equipment risk set.

Step 3: Eliminate the set of equipment that does not meet the requirements, and reduce the number of branches to form a new HR-Tree.

Step 4: Find the equipment set with the highest risk from the local equipment set and determine the environmental safety index in the new situation.

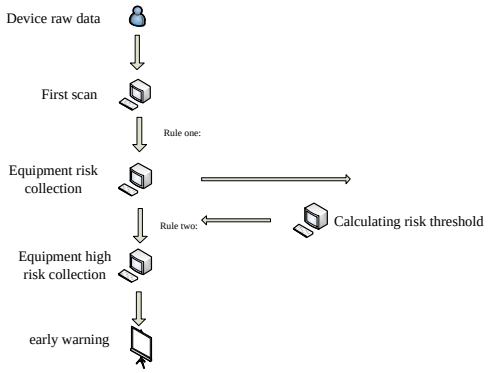

Fig. 2. HR-Tree's flow-chats

3. CASE ANALYSIS

3.1. The analysis of Equipment risk impact

This paper selects 9 equipments of a 220*kV* substation in Liaoyang to evaluate the equipment risk impact. This section categorizes the information of importance index from 9 devices, as shown in Table 4.

Table 4. Index value

Device	Number	Cost	Voltage	Area	Relative
Transformer	B1	9	5/220kV	9/A	9
Transformer	B2	7	3/110kV	9/A	9
Lines	S1	3	3/110kV	5/C	3
Bus bar	M1	5	5/220kV	7/B	5
Breaker	D1	9	3/110kV	9/A	9
Breaker	D2	9	5/220kV	7/B	9
Switchs	G1	1	3/110kV	9/A	1
Capacitors	H1	5	1/10kV	7/B	5
Capacitors	H2	7	3/110kV	5/C	7

A judgment matrix is constructed based on the information of importance equipment from the four indicators. The maximum eigenvalue is 4.2467, and its maximum eigenvector is:

$w = \{0.2246, 0.2301, 0.3361, 0.2092\}$.

The consistency check is: $CR = 0.0913 \le 0.1$.

The information in Table 4 is substituted into formulas (4) and (6). The calculation results of equipment importance are:

$SI_1 = \left[si_1, si_2, ..., si_9\right]$ = [1.9046, 0.9395, 1.4196, 1.1411, 0.9619, 0.6156, 11.343, 5.6603, 1.4888], and the results are normalized. The equipment hidden danger can be obtained. The equipment risk impact is further normalized the results are shown in Table 5.

Table 5. Risk impact

Device	Number	Importance	(H)Dangers	(F)Risk
Transformer	B1	0.1679	1.0000	0.1679
Transformer	B2	0.0828	0.1209	0.0100
Line	S1	0.1251	0.1587	0.0199
Bus bar	M1	0.1006	0.1139	0.0115
Breaker	D1	0.0848	0.1589	0.0135
Breaker	D2	0.0543	0.1146	0.0622
Switch	G1	1.0000	0.0998	0.0998
Capacitor	H1	0.4990	0.1188	0.0593
Capacitor	H2	0.1313	0.1870	0.0246

Table 6 collects various types of alarm information for a 220*kV* substation in Liaoyang in 2019, and it sorts out the original data to get an example of the original set of equipment failures.

Table 6. Risk-set

Incident	Path	R1
L1	(B1,1),(S1,4),(M1,3),(G1,1)	0.3818
L2	(B2,5),(M1,1),(H2,3),(G1,1)	0.2351
L3	(D1,3),(B2,3),(H2,2)	0.1197
L4	(B1,1),(B2,4),(D1,2)	0.2349
L5	(H2,1),(B2,5),(D1,3),(M1,1)	0.1266
L6	(B2,4),(D1,1),(M1,1)	0.0650

According to the total value of the single transaction risk in Table 6, the total transaction impact of the equipment risk: R_3 can be calculated as: $\{B1, B2, S1, M1, D1, G1, H2\}$ = {0.6167, 0.7813, 0.3818, 0.8085, 0.5462, 0.6169, 0.4814}, and the equipment risk impact of the total accident set is calculated as 1.1631, the scale factor is 0.4, according to formula (15), the minimum risk threshold: $\min - R$ is 0.4652, and the equipment set below the risk threshold is removed: $\langle M1, B2, D1, H2 \rangle$ .This step is the basis for reducing branches. The S1 equipment is discarded and all branches were arranged according to the total transaction risk value.

3.2. The construction and reduction of HR-Tree

According to the basic data in section 3.1, HR-Tree is constructed. In this section, branch A is first established. The process is as follows:

(1) Establish the root node of the tree and name it top.

(2) Insert transaction set:

$L1$={(M1,3),(G1,1),(B1,1)},

H_{M1}.name = {M1},

H_{M1}.count =1,

H_{M1}.risk =R1(L1)-(R1({G1}L1)
+R1({B1}L1))
=0.3022-0.0998-0.1679=0.0345,

H_{G1}.risk =R1(L1)-R1({B1},L1)
=0.3022-0.1679=0.1343,

H_{B1}.risk =R1(L1)=0.3022,

The HR-Tree L1 path is constructed, as shown in Figure 3.

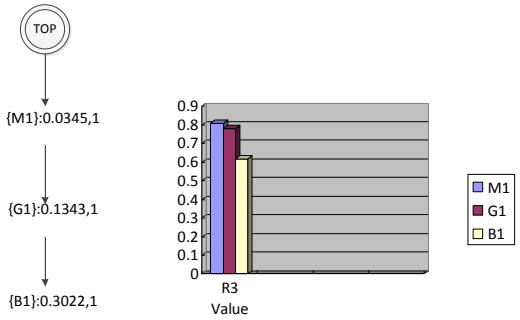

Fig. 3. L1path analysis

Insert the L2, L3, L4, L5, and L6 paths in the L1 path to form a partially complete HR-Tree, and connect the branches with the same numbered devices. As shown in Figure 4.

According to the total transaction risk value R_3, a non-candidate lower risk path is obtained, in which the path $\{H2\}$'s total transaction risk value is the smallest, it's thinning process is as follows:

Classify the three paths containing $\{H2\}$:

$\langle M1, B2, D1, H2\rangle$:0.1266,

$\langle M1, B2, G1, H2\rangle$:0.2351,

$\langle B2, D1, H2\rangle$:0.1197,

Then the each transaction risk(R_3) of them is:

$\langle B2\rangle$:0.4814, $\langle H2\rangle$:0.4814, $\langle M1\rangle$:0.3617, $\langle D1\rangle$: 0.2463, $\langle G1\rangle$:0.2351.

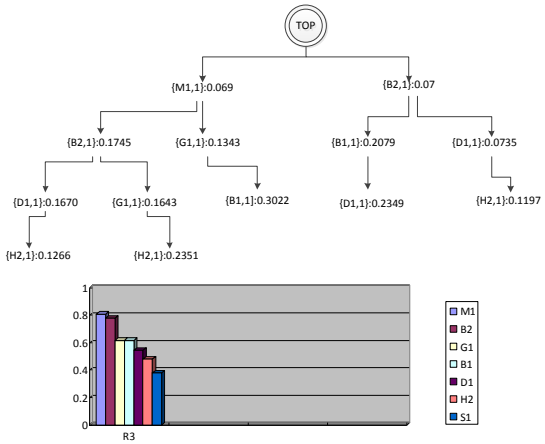

Fig. 4. Overall HR-Tree

Remove the non-candidate high risk set: $\langle G1\rangle$ and readjust the path. The process is shown in Table 7.

Table 7. {H2}Path of information

Original path	Adjusted path	N.count
L5<M1 B2 D1 H2>:0.1266	<M1 B2 D1 H2>:0.1266	1
L2<M1 B2 G1 H2>:0.2351	<M1 B2 H2>:0.1353	1
L3<B2 D1 H2>:0.1197	<B2 D1 H2>:0.1197	1

The minimum global risk value of local devices is shown in Table 8.

Table 8. Minimum risk value

Device ID:	*M1*	*B2*	*D1*	*H2*
Value:	0.0115	0.0300	0.0135	0.0246

The risk value of the new path is defined as $H.\text{risk}'$, the minimum risk value of the device is defined as $H.\min$, and the support degree of the new path about device is $H.\text{count}'$, then the new definition formula is:

$$\text{H.risk}' = \text{H.risk} - \sum H.\min \times \text{H.count}' \quad (15)$$

Reconstruct the local path risk value and describe the first path in Table 7:

$\text{H}_{M1}.\text{risk}' = \text{H}_{L5}.\text{risk} - \sum H_{M1}.\min \times \text{H}_{L5}.\text{count}' = 0.1266 - 0.03 - 0.0135 - 0.0246 = 0.0585.$

$\text{H}_{B2}.\text{risk}' = \text{H}_{L5}.\text{risk} - \sum H_{B2}.\min \times \text{H}_{L5}.\text{count}' = 0.1266 - 0.0135 - 0.0246 = 0.0885.$

$\text{H}_{D1}.\text{risk}' = \text{H}_{L5}.\text{risk} - \sum H_{D1}.\min \times \text{H}_{L5}.\text{count}' = 0.1266 - 0.0246 = 0.102.$

$\text{H}_{H2}.\text{risk}' = \text{H}_{L5}.\text{risk} - \sum H_{H2}.\min \times \text{H}_{L5}.\text{count}' = 0.1266.$

The partial branch reduction is shown in Figure 5.

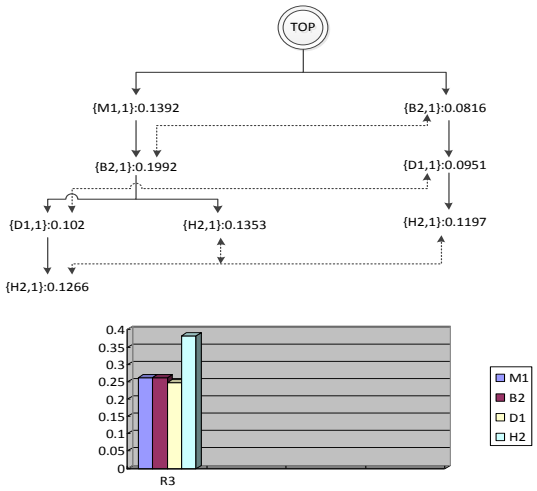

Fig. 5. Branch reduction

The transaction risk set of the local HR-Tree in Figure 5 is collectively recorded as:

$\{D1, B2\}$:0.1971,

$\{B2, M1\}$:0.1992,

$\{D1, B2, M1\}$:0.102,

$\{H2, B2, M1\}$:0.1353,

$\{H2, D1, B2\}$:0.1197,

$\{H2, D1, B2, M1\}$:0.1266。

The same method is used to analyze the risk value of local path set from other devices $\{B1, B2, S1, M1, D1, G1\}$, and the minimum risk threshold $\min - R$ is also used as the threshold. The highest proportional risk is η. By calculation,

the highest risk of equipment set can be obtained: $\{D1, B1, B2\}$:0.2349, $\{H2, D1, B2\}$: 0.2463, $\{B1, G1, M1\}$:0.3022.

According to on-site dispatching, high-risk collections have a great impact on the environment, and this information plays an important role in early warning of dispatchers and maintenance teams. After investigation, the on-site 110*kV* transformer and 10*kV* capacitor had a long investment time and aging phenomenon, which were highly consistent with the early warning of high-risk equipment.

After that, this section attempts to find the connection between transactions, and then explores the law of transaction occurrence, and provides a reference for the dispatchers on site. The simulation results are shown in Figure 6.

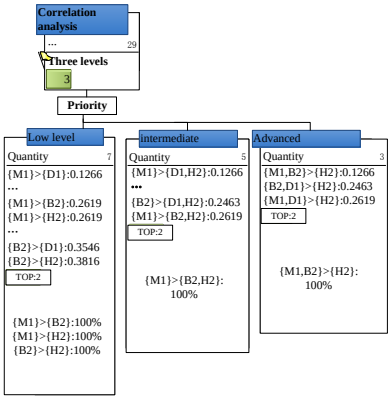

Fig. 6. The correlation of transaction

As shown in Figure 6, The transaction set is divided into three levels, namely low level, intermediate level, and high level. The set of each level is arranged according to the score of the occurrence of transactions. Each set of levels can get two sets of transactions with a high score system, which also shows that the probability of these high-scoring transactions is high. The 100% correlation transaction are: $\{M1\} > \{B2\}$, $\{M1\} > \{H2\}$, $\{B2\} > \{H2\}$, $\{M1\} > \{B2, H2\}$, $\{M1, B2\} > \{H2\}$. The above symbol indicates the occurrence of the previous item, and the occurrence probability of the latter item is 100%.

From the above analysis, the method research in this article believes that Transformer and Capacitor will also be damaged after Bus bar was damaged, and the loss must be stopped in time: turn off the relevant switches. Transformer 's damage also needs to pay attention to Capacitor at the same time. The correlation between device sets is also significant. When Bus bar and Transformer were damaged at the same time, we must know that Capacitor has also been damaged.

These chain of equipment damage occurred, we will simulate the on-site early warning measures in the process of knowing the chain reaction, which is very meaningful for the protection of the project. The basis for early warning is based on the scores generated after the occurrence of each transaction set. From Figure 6, we can see that the score of the occurrence of transactions. The selection of thresholds in this section varies according to the level of the transaction occurrence chain.

In the low-level environment, the threshold is selected as: 0.35. In the intermediate-level environment, the threshold is selected as: 0.24. In the advanced environment, the threshold is selected as: 0.24. From this we can see that each level of transaction generation chain has two transactions that get the highest score, and we could give early warning of these transactions. The content of the early warning is the highest scoring transaction set. This article hopes that such early warning can be brought to the field operation staff for maintenance guidance.

3.3. The Influence of the number of branches on HR-Tree data mining

This article extracts equipment data from the three accidents in the past five years, and selects equipment data of three different units. Each number means a group of data about a transformers and accessories, and the size of the control group ranges from few to many. At the same time, we set the scale factor in the experimental method to 0.4, and Table 9 shows the data mining results of the high-risk equipment set during the three accidents.

Table 9. The influence of the number of branches

Inspection scope	$1^{\#}$	$1^{\#}2^{\#}$	$1^{\#}2^{\#}3^{\#}$
Number of branches:	20	32	45
High risk set:	3	5	6
Abnormal devices:	10	11	14

Table 9 shows the relationship between the number of branches and the set of high-risk equipment. Abnormal devices represent the number of damaged device on site. As can be seen from Table 9, the number of branches represents the number of high-risk equipment sets in other means. The number of branches increases, and the number of high-risk equipment sets also increases. The increase of high-risk equipment sets has a warning effect on site supervisor. From the contrast of the abnormal events on the site, this is in line with reality.

4. CONCLUSION

With the increase of the automation degree of power equipment, the operation efficiency of personnel in handling accidents is also efficient. Nowadays, the power network framework has gradually changed. Digitalization has brought a huge operational revolution to dispatchers and substation employees, but the imformation is still necessary to be standardized and summarized. Without analysis of multi-dimensional information sources, it will bring the result of digital

accumulation which cause the dispatcher's inability to issue orders in a timely manner.

On the basis of the fault set obtained after minusing branch, the correlation between the fault sets is calculated by the algorithm of article, so that the early warning of each fault occurrence was precise. The innovation of this article can be summarized from two parts:

(1) Based on the multi-dimensional information foundation, this article quantifies the information into a digital network. This method establishes an HR-Tree for the digital network, then calculates the candidate high-risk set from the overall context. The process of branch reduction mainly strives for the redundancy of information. This method refers to the simplest calculation when the same result is obtained, which saves a lot of time and space for information processing.

(2) This article selects the highest risk value of candidate set from each candidate set, it accumulates historical data of each set and the operating status of the environment. The method could predicts the operating status of the environment through these prediction results: this paper can obtain the relationship about these equipments, which displayed on a 100-point scale. Managers could choose the cause of failure from high-scoring devices, which can achieve innovation that reduces workload.

SOURCE OF FUNDING

This work is supported by National Natural Science Foundation of China (Grant, No.61673093).

REFERENCES

1. Li G, Tang J. A new HR-tree index based on hash address[C]. 2010 2nd International Conference on Signal Processing Systems.2010;pp. V3-35-V3-38. https://doi.org/10.1109/ICSPS.2010.5555818
2. Xing Yu. research on power transformer fault diagnosis technology based on fault tree and signal analysis. Harbin Institute of Technology. 2017.
3. Ren Dongmei, Zhang Yuyang, DONG Xinling. Application of improved principal component analysis-Bayes discriminant analysis method to petroleum drilling safety evaluation. Journal of Computer Applications.2017,37(06):1820-1824. https://doi.org/10.11772/j.issn.1001-9081.2017.06.1820
4. Yu H, Wu ZR, Bao T F. Multivariate analysis in dam monitoring data with PCA. Science China Technological Sciences. 2010;53(4):1088 — 1097. https://doi.org/10.1007/s11431-010-0060-1
5. Lanzrath M, Suhrke M Hirsch H. HPEM-based risk assessment of substations enabled for the smart grid. IEEE Transactions on Electromagnetic Compatibility. 2020;62(8):173-185. https://doi.org/10.1109/TEMC.2019.2893937.
6. Karaolis MA, Moutiris JA, Hadjipanayi D, Pattichis CS. Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Transactions on Information Technology in Biomedicine.2010;14(3):559-566. https://doi.org/10.1109/TITB.2009.2038906
7. Samiri MY, Najib M, Elfazziki A, Abourraja MN, Boudebous D, Bouain A. APRICOIN: An adaptive approach for prioritizing high-risk containers inspections. IEEE Access.2017;9(5):18238-18249. https://doi.org/10.1109/ACCESS.2017.2746838
8. Lee, Sang-Woo, Kwon, Jung-Hyok, Lee, Ben, Kim, Eui-Jik. Scientific literature information extraction using text mining techniques for human health risk assessment of electromagnetic fields. Sensors and Materials.2020;32(1): 149-157. https://doi.org/10.18494/SAM.2020.2572
9. W. Gan, JC Lin, P. Fournier-Viger, H. Chao and P. S. Yu. HUOPM: High-Utility Occupancy Pattern Mining. IEEE Transactions on Cybernetics. 2020;50(3):1195-1208. https://doi.org/10.1109/TCYB.2019.2896267
10. Ma Lijie, Zhlj Yongli, Zheng Yanyan. Research on transformer fault diagnosis and optimization based on parallel variable prediction model. Power System Protection and Control. 2019,47(6):82-89. https://doi.org/10.7667/PSPC180399.
11. Zhang Xuewei, Li Hanshan. Research on transformer fault diagnosis method and calculation model by using fuzzy data fusion in multi-sensor detection system. Optik. 2019;176(9):716-723. https://doi.org/10.1016/j.ijleo.2018.09.017
12. Mou Shanzhong, Xu Tianci, Fu Ao, Wang Men G, Bai Ru. Fault diagnosis method of transform er based on adaptive deep learning model. Sout hern Power System Technology. 2018;12(10):14-19. https://doi.org/10.13648/j.cnki.issn1674-0629.2018.10.003
13. LiLi Mo. Transformer fault diagnosis method ba sed on support vector machine and ant colony. Advanced Materials Research. 2013;659(1):54-58. https://doi.org/10.4028/www.scientific.net/AMR.659.54
14. Li Wenfeng, You Qinghe, Liao Qiang. Research on remote fault diagnosis method based on T-S fuzzy FTA. Control Engineering of China.2018, 25(9):1703-1708. https://doi.org/10.14107/j.cnki.kzgc.160089
15. Chaolong Zhang, Yigang He, Bolun Du, Lifen Yuan, Bing Li, Shanhe Jiang. Transformer fault diagnosis method using IoT based monitoring system and ensemble machine learning. Future Generation Computer Systems.2020,108(3): 533-545. https://doi.org/10.1016/j.future.2020.03.008
16. Arturo Mejia-Barron, Martin Valtierra-Rodriguez, David Granados-Lieberman, Juan C. Olivares-Galvan, Rafael Escarela-Perez. The application of EMD-based methods for diagnosis of winding faults in a transformer using transient and steady state currents. Measurement. 2018; 117(12): 371-379. https://doi.org/10.1016/j.measurement.2017.12.003
17. Xu Jinyong, Luo Shijun, Zhang Zida. Fault diagnosis system of wheel loader hydraulic system based on fuzzy fault tree analysis. Journal of Jilin University (Engineering and Technology Edition). 2007;37(3):569-574. https://doi.org/10.13229/j.cnki.jdxbgxb2007.03.017
18. Hu Jun, Yin Liqun, Li Zhen, Guo Lijuan, Duan Lian, Zhang Yubo. Fault diagnosis method of transmission and transformation equipment based on

big data mining technology. High Voltage Engineering.2017;43(11):3690-3697. https://doi.org/10.13336/j.1003-6520.hve.20171031026

19. Wang Tao, Sun Zhipeng, Cui Qing, Zhang Zhilei, Zhang Tianwei. Research on fault diagnosis of power transformer based on classification decision tree algorithm. Electrical Engineering. 2019;20(11):16-19.

Received 2020-03-13
Accepted 2020-05-12
Available online 2020-05-13

Shanyuan WANG (1994-), man, Communication author, Postgraduate, Research on power grid operation and power dispatch control technology and application, E-mail：wsyneu@163.com