An Improved Decision Tree Algorithm for Condition Monitoring on Storage Power Station of Internet Things

Gengze Li[1], Shuaixuan Li[2] , Jun Yan[1]

[1]State Grid Xin Yuan Company Limited, Beijing 100052, China

[2]State Grid Xinyuan Bailianhe Pumped Storage Company Limited, Huanggang, Hubei 438600, China

***Abstract*—power station is an important basic power generation organization, and its operation status is related to the continuous power generation capacity. At present, a large number of physical network equipment and intelligent equipment are used in pumped storage power station, which makes its data mass growth and its operation state become a difficult problem. Accurate operation monitoring results can provide decision support that power generation planners and government, but also reasonably dispatch corresponding resources. In the past, decision tree algorithm was used in operation condition monitoring, which has the problem of data distortion and affects the accuracy of monitoring results. Based on the above reasons, this paper combines the wavelet function and decision tree algorithm, proposes an improved decision tree algorithm to eliminate redundant data in order, and uses wavelet function to cluster distorted data, so as to improve the accuracy and computational efficiency of the algorithm. Matlab simulation results show that: decision tree algorithm can eliminate 90% of redundant data, reduce the impact of feature data extraction on decision tree. At the same time, the improved accuracy is 98%, the calculation time is less than 25s is better than that, the decision tree algorithm. Therefore, the improved algorithm can optimize the condition monitoring of pumped storage power station.**

***Keywords*—condition monitoring, Decision tree, Internet of things, Wavelet algorithm**

I. INTRODUCTION

Pumped storage power generation belongs to the category of primary industry activities, which is not only the guarantee of China's economic transformation in 2020, but also the basis of the upgrading of the primary industry [1]. The integration and Internet of things not only expands the scope, but realizes its own structure optimization. Continuous monitoring and analysis can help enterprise managers and government departments to make decisions and improve the safety production capacity and comprehensive competitiveness of pumped storage industry [2].

In the past, decision tree algorithm is the main method to monitor the state. the algorithm can monitor and analyze the complex unstable time data in the operation state, it still can not avoid the shortage of data redundancy in matlab algorithm, and can not meet the requirements of massive data calculation in the condition monitoring of power station. Some scholars believe that [3], the state monitoring of pumped storage power station presents the trend of big data, and this trend is increasingly obvious. The accuracy and effectiveness of traditional algorithms such as neural network, vector regression, chaotic time and decision tree in operation state monitoring are decreasing day by day. It is suggested to integrate the above algorithms with redundant data elimination function. Some scholars believe that [4], although the decision tree method reduces the length of the coefficient sequence by 1/2, it does not have time-shift invariance and can not avoid data distortion. Therefore, it proposes to use steady-state and discrete functions to modify, so as to achieve the purpose of reconstructing the operation state monitoring results.

Based on the above reasons, this paper combines the discrete wavelet function and decision tree algorithm, and improves the accuracy of the results by eliminating the redundant data, revising the data and reconstructing the overall data, so as to better monitor the operation status.

II. MATHEMATICAL DESCRIPTION OF THE DEVELOPMENT MATURITY OF POWER STATION CONDITION MONITORING IN THE INTERNET OF THINGS

The key of operation state analysis is to quantify the relevant indicators, and describe the practice link, production content and development direction of mathematically, can pave the way for later monitoring results judgment and analysis.

A. Judgment Process of Condition Monitoring Results

The operation state judgment includes three aspects: the power generation form x_i, the influence of Internet of things on the power generation x_j, and the power generation

sustainability x_k. The power generation forms include the infrastructure x_{i1}, the proportion of different power generation forms x_{i2}, the degree of cooperation among wind power, energy storage power station and solar energy x_{i3}; The promotion degree of Internet of things generation x_{j1}, the integration degree of Internet of things and power generation x_{j2}, the promotion level of Internet of things to power generation x_{j3}. The cooperation degree among thermal power, hydropower, wind energy is x_{k1}, the cooperation between different power generation equipment is x_{k2}, the cooperation between different power generation departments is x_{k3}. The above analysis shows that there are many aspects involved in the operation state, and the collected data is massive (cloud data, a large number of applications of intelligent devices), complex (there are a large number of unstructured data), which greatly reduces the "micro" and "Overview" effect of the calculation results, resulting in the "distortion" of the monitoring. Since mass and complexity are the inevitable trend of the development [5] is the focus on the problem of "data distortion".

B. Description of Condition Monitoring Data Flow

Stable wavelet function can extract redundant discrete data for comprehensive analysis, and keep the order of data coefficients to reduce the "data distortion" rate [6]. At the same time, the stable wavelet function uses the discrete extraction method to ensure the time shift invariance of the data and complete the single-phase feature extraction of the data.

(1)Assuming that the result of operation monitoring is A and $Al=\{a_1,a_2,\ldots\ldots,a_n\}$, the relationship between A and the input data is as follows.

$$\sum_{l}^{n} A_l \xleftarrow[f(\bullet)]{TS(\bullet)} \oint_{n\ \text{isnature}} \prod_{\text{one by one}} \overleftarrow{k\text{-means deal}} \sum_{i}^{n} x_i \underset{-g(\bullet)}{\overset{g(\bullet)}{\longleftrightarrow}} \sum_{j}^{n} x_j \underset{-g(\bullet)}{\overset{g(\bullet)}{\longleftrightarrow}} \sum_{k}^{n} x_k \tag{1}$$

Among them, i, j, k are natural numbers, $TS(\cdot)$ is decision tree function, $f(\cdot)$ is stable wavelet function, *k-means* $(\cdot)$ is prior data clustering function, $g(\cdot)$ is forward function among different input indexes, and $-g(\cdot)$ is reverse function.

(2) Suppose that the arbitrary result a_l in the operation state and the input z in the decision tree algorithm (the generation form x_i, the influence of Internet of things on the generation x_j, and the generation duration x_k), p is the data proportion (structured data >70%, semi-structured data >70%, unstructured data >70%), q is the data distortion processing method (reconstruction=1, coefficient order = 2, discrete elimination=3, feature extraction =4, clustering=4), then a_l is described as $\overline{\overset{o}{a}\log A_l^{o,p,q}}$, o, p, q are natural numbers. Among them, in order to reduce the influence of large change of value on c_l, $log(\cdot)$ function is used to deal with it, and the average value of the function is taken, and it is noted that the function is single-phase ordered calculation.

(3) The power generation form x_i, the influence of Internet of things on power generation x_j, and the power generation sustainability x_k are fused by $g(\cdot)$, and the data are analyzed by standardization. The calculation formula is as follows.

$$g(x)=\underbrace{\left\langle \overbrace{\left(\sum_{i=1}^{n}\varepsilon/2x_i+\sum_{j=1}^{n}\varphi/2x_j+\sum_{k=1}^{n}\iota/2x_k\right)\bullet\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}+\Gamma/\sum_{i,j,k=1}^{n}(x_i+x_j+x_k)}^{\left\|\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}\right\rangle}}_{x_i,x_j,x_k} \tag{2}$$

Among them, ε, φ, ι and are the power generation form x_i, the influence of Internet of things on power generation x_j, and the weight coefficient of power generation sustainability x_k.

$\Gamma/\overset{o}{a}_{i,j,k=1}^{n}(x_i+x_j+x_k)$ is the error adjustment coefficient of the above data, $\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}$ is the judgment matrix of distorted data in decision tree algorithm and $\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}$ is the reconstruction judgment matrix in decision tree algorithm.

(4) The data in operation are mainly published data, and relevant websites and yearbooks. The calculation formula of steady-state wavelet function is as follows:

$$\sum_{1}^{T+1} A_{i+1} = \sum_{1}^{T}\sum_{1}^{i} A_i + \lim_{i\to\infty} f\left(\overbrace{\sum_{1}^{T}\sum\int x_i \underset{and}{\longleftrightarrow}\int x_j \underset{and}{\longleftrightarrow}\int x_k}\right) \tag{3}$$

Among them, $f\left(\overbrace{\sum\int x_i \underset{and}{\longleftrightarrow}\int x_j \underset{and}{\longleftrightarrow}\int x_k}\right)$ is the wavelet function, $\sum_{1}^{T}\sum_{1}^{i} A_i$ is the running state ordered data link at t time, $\sum_{1}^{T}\overbrace{\sum\int x_i \underset{and}{\longleftrightarrow}\int x_j \underset{and}{\longleftrightarrow}\int x_k}$ is the included data at t time, $\overset{T+1}{\underset{1}{a}} A_{i+1}$ is the running state data link at T + 1 time

(5) In order to eliminate redundant data, feature data is proposed to reduce the impact of the attributes of structured, semi-structured and unstructured data on the results of running state [7], and *K-means* clustering is needed in the early stage. The purpose of *K-means* clustering is to select the clustering center S on the data entry point of the improved decision tree, and analyze different data through iterative calculation. The specific formula is as follows.

$$\left|S\right| = \lim_{x\to\infty} \overrightarrow{A_l/(\sum_{l=0}^{n} A_l + s_i^T)^2} + \varsigma \tag{4}$$

Among them, $\left|S\right|$ is the coefficient sequence after data reconstruction, A_l is the result of running state, s_i^T is the

clustering point i and the clustering center at time T, ς is cluster the allowable error, the error precision is set by the specific enterprise of pumped storage power station.

III. THE CONSTRUCTION OF OPERATION CONDITION MONITORING MODEL

A. Description Construction of "Data Distortion" Setting Operator in Operation Condition Monitoring

The improved decision tree algorithm takes into account the problem of poor data continuation, which increases the probability of "data distortion" in operation condition monitoring, so it is necessary to build a setting operator to preprocess the "data". Assuming that the setting operator function is $Q(x)$, the calculation formula is as follows.

$$Q(c)\begin{cases} x_{i\cap j\cap k}\in s_i^T,\ \overrightarrow{0<Q(x)<(\max(s_i^T)}, \text{and } Q(x_{i\cap j\cap k})\neq 0, \text{into } L\{\cdots\}, \text{or delet} \\ x_{i\subset\cup j\subset\cup k}\in s_i^T,\ \overrightarrow{\min(s_i^T)<Q(x)}, Q(x_{i\subset\cup j\subset\cup k})\neq 0, \text{into } L\{\cdots\}, \text{or delet} \\ \text{IF}\left|\overrightarrow{x_{i\cap j\cap k}}\right|\preceq\left|x_{i+l\cap j+l\cap k+1}\right|\cap\left|x_{i\subset\cup j\subset\cup k}\right|\preceq\left|\overrightarrow{x_{i+l\subset\cup j+l\subset\cup k+1}}\right|, \text{and } Q\left|x_{i\cap j\cap k}\right|\subset\cup\left|x_{i\subset\cup j\subset\cup k}\right|\neq 0, \\ \text{set } \left|x_{i+l\cap j+l\cap k+1}=x_{i\cap j\cap k}\right|\subset\cup\left|x_{i+l\subset\cup j+l\subset\cup k+1}=x_{i\subset\cup j\subset\cup k}\right|, \text{or delet} \end{cases} \quad (5)$$

Through the analysis of the above functions, the data of different centers can be compared s_i^T. Among them, $x_{i\cap j\cap k}$ is independent of different inputs, $x_{i\subset\cup j\subset\cup k}$ is Collaboration data between different inputs; For the comparison of the new operation state data $\overrightarrow{0<Q(x)<(\max(s_i^T)}$ and $L\{\cdots\}$ the original sequence data set; If the direction of the included data $L\{\cdots\}$ is different from that of the original middle end mantissa, it is redundant data and eliminated; If it is greater than the end data, the end data will be exchanged with the newly included data [9].

B. The Influence of the Relationship

The influence of Internet of things on the power generation, and the integration operator of the power generation sustainability are built. The influence of the relationship between different input indexes x_i, x_j and x_k on the results should be fully considered, so the fusion operator should be constructed. Assuming that the overall fusion degree of the three is $Z=\{Z_1, Z_2,\ldots\ldots,Z_n\}$, and the local fusion degree is $z=\{z_1, z_2, \ldots\ldots, z_n\}$, then the fusion formula is as follows.

$$\sum_{i=1}^{n}(Z_i\rightleftharpoons z_i)\begin{cases} z_i Z_i=0 \text{ or } 1 \\ g\{x\}[(Z_{x_{i,j,k}}\xleftarrow{\text{data is L}}x_{i,j,k})/\{Z_{x_{i,j,k}}-\max(z_{x_{i,j,k}})\}\{z_{x_{i,j,k}}-\min(Z_{x_{i,j,k}})\}/n\begin{bmatrix}1&0&0\\0&\ddots&0\\0&0&1\end{bmatrix}\approx 0 \text{ or } \approx 1 \quad z_i\neq Z_i \end{cases} \quad (6)$$

Among them, $g(x)$ is the fusion function, $\{Z_{x_{i,j,k}}-\max(z_{x_{i,j,k}})\}\{z_{x_{i,j,k}}-\min(Z_{x_{i,j,k}})\}/n$ is the wavelet decomposition of the fusion degree value, $\begin{bmatrix}1&0&0\\0&\ddots&0\\0&0&1\end{bmatrix}$ is the decision matrix of the decomposed data, and $\overset{\circ}{\underset{i=1}{\overset{n}{a}}}(Z_i\rightleftharpoons z_i)$ is the final result of the fusion operator.

C. The Accuracy Operator of Operation Condition Monitoring

The accuracy of monitoring is an important judgment index of the improved algorithm [10], so to construct the accuracy operator, the specific calculation formula is as follows.

$$\log A_l^{o,p,q}=\begin{cases} 0<\overline{(\sum\log A_l^{o,p,q})/n<1} \\ \cap\overline{\log A_l^{o,p,q}<\|\max[\log A]\|\cap\log A_l^{o,p,q}>\|\min[\log A]\|} \\ \sum_{i=1}^{n}(Z_i\rightleftharpoons z_i) \\ \overline{\log A_l^{o,p,q}\propto\log A_{l-1}^{o-1,p-1,q-1}}\cap\prod_{x(x)}[\{\lim_{x\to\infty}\{\sum_{l=1}^{n}\max(\log A_l^{o,p,q})\underset{\prec}{\overset{\propto}{\rightleftharpoons}}\min(\log A_l^{o,p,q})\}] \end{cases} \quad (7)$$

Because of the output result of running state $\log A_l^{o,p,q}$, the precision range is between 0 ~ 100%. In the process of continuous calculation, $\lim_{x\to\infty}\{\sum_{l=1}^{n}\max(\log A_l^{o,p,q})\underset{\prec}{\overset{\propto}{\rightleftharpoons}}\min(\log A_l^{o,p,q})\}$ is the difference between the maximum value and $\overset{0\sim 1}{\overline{\log A_l^{o,p,q}\ \mu\ \log A_{l-1}^{o-1,p-1,q-1}}}$ is the minimum value, which tends to 0, indicating that the accuracy is improved. The smaller the value is, the lower the distortion rate is. In the above calculation process, the data integration requirements should be met, namely $\overset{\circ}{\underset{i=1}{\overset{n}{a}}}(Z_i\rightleftharpoons z_i)$.

D. The Steps of Improving the Operation Condition Monitoring in Decision Tree Algorithm

Step 1. Firstly, the input data is preprocessed by *K-means* to eliminate redundant data and illegal data, and the complexity of data is reduced by clustering;

Step 2.The processed data $\overrightarrow{\overset{\circ}{a}\log A_l^{o,p,q}}$, $A_l=\{a_1, a_2, \ldots\ldots, a_n\}$, are fused $\overset{\circ}{\underset{i=1}{\overset{n}{a}}}(Z_i\rightleftharpoons z_i)$ and decomposed by stable wavelet $\begin{bmatrix}1&0&0\\0&\ddots&0\\0&0&1\end{bmatrix}$.

Step 3. Calculate the decomposed data $\overset{0\sim 1}{\overline{\log A_l^{o,p,q}\ \mu\ \log A_{l-1}^{o-1,p-1,q-1}}}$, and repeat 1-2 steps until the iteration times meet 100 times, or the data value is illegal, and output the monitoring results,

Step 4. By comparing the improved decision tree algorithm with the decision tree algorithm, the accuracy and computing

time of the two algorithms are obtained.

IV. THE ACTUAL CASE OF IMPROVED DECISION TREE IN THE GENERATION TIME

A. The Case Introduction

In order to simplify the calculation process, electric energy is taken as the representative case. The introduction of case related parameters ara follows. The location case is plain area, mountain area, and other area[11]; The monitoring time is from 2019 to 2020; The data sources come from China Power Grid [12], National Bureau of statistics, website, statistical yearbook and domestic capital data in various regions; The data source time is from 2015 to 2018; The initial fusion value is Z_1=0.7, z_1=0.02; Input index are the power generation form x_i, the influence of Internet of things on the power generation x_j, and the power generation sustainability x_k; Iterations is 100time [13] and allowable precision of clustering is ς =0.001.

B. The "Data Distortion" of Improved Decision Tree Algorithm

The stable wavelet function, *K-mean* clustering and g (·) fusion function are used to process and reconstruct the "data" in the decision tree algorithm. The "1/2" method is used to reduce the redundant data, and the eigenvalues are sorted to extract the distortion of the data. The results are shown in Figure 1.

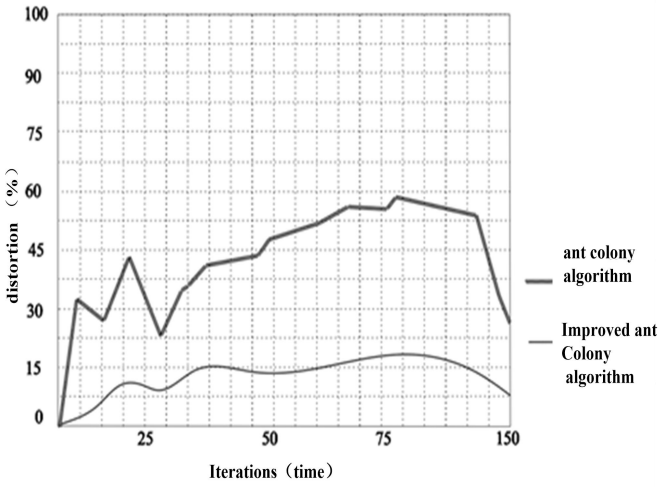

Fig.1 Comparison of distortion between improved decision tree and decision tree algorithm

It can be seen from Figure 1 that the abnormal variable of the improved decision tree algorithm is below 15%. Although the change trend of the improved decision tree algorithm is the same as that of the decision tree algorithm in the early stage, the overall abnormal variable is significantly lower than that of the decision tree algorithm. The above results show that the pre-processing of operation condition monitoring data, as well as the later fusion and reconstruction calculation, can make up for the "defect" of data disorder extraction in decision tree algorithm, which is consistent with the domestic results [14].

C. The Accuracy of Decision Tree Algorithm for Operation Condition Monitoring

The monitoring accuracy of input indicators (power generation form x_i, impact of Internet of things on power generation x_j, and power generation sustainability x_k is analyzed and compared with the actual situation from 2019 to 2020.

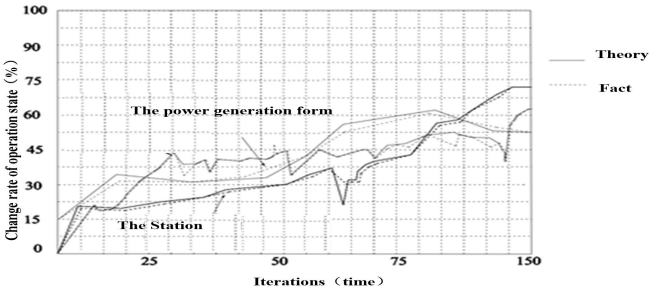

Fig. 2 monitoring results of improved decision tree

It can be seen from Figure 2 that the improved decision tree algorithm can better monitor the running state. Although the power generation form is 28 iterations, the impact of Internet of things on the power generation is 49 iterations, and the power generation duration is 62 iterations, the overall monitoring results are consistent with the actual situation.

D. The Time of Operation Condition Monitoring

Compare the time of old algorithm and improved algorithm, and the results are shown in Figure 3.

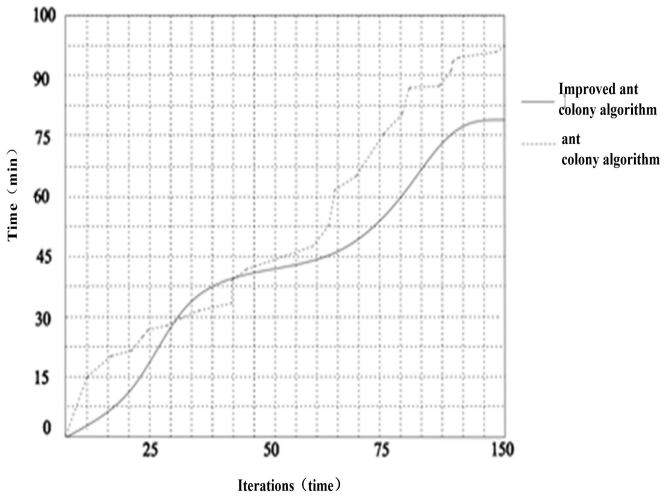

Fig. 3. The running condition monitoring time of improved decision tree algorithm

It can be seen from Figure 3 that the processing time of the improved algorithm is less than 75s, which is significantly lower than the monitoring time of the old, and the curve of the improved is smoother, which indirectly proves that the "data distortion" rate is low and the accuracy is high.

V. CONCLUSION

The Internet of things is an important national basic industry [15,16], which is related to the strategic security of national development [17], and it is required to optimize the power

generation and accelerate the development of power station condition monitoring [18]. Based on this background, this paper improves the previous decision tree algorithm, and constructs the improved decision tree algorithm by combining the stable wavelet function and *K-mean* clustering. Matlab simulation results show that the improved decision tree algorithm can effectively reduce the data distortion (<15%), and the accuracy of operation condition monitoring is more than 98%, which is consistent with the actual test value. At the same time, the running condition monitoring time of the decision tree algorithm is less than 75s, which is significantly better than the decision tree algorithm.

REFERENCES

[1]Ji Tonghui. Research on evaluation method of power generation structure of low carbon economy pumped storage power station based on generalized distance minimum and rough set]. *Ecological Economy*. 2018, 34 (04): 40-44.

[2] Maryam E. Multi-method approach for the comparative analysis of solar and wind energy industry structures in Germany and Iran. *Int. J. of Energy Technology and Policy*. 2018, 14(2/3).

[3] Chang Fengrui, Zeng Zihao. Research on power generation structure transformation path of Pumped Storage Power Station under supply side Reform. *Coal Economy Research*. 2018, 38 (12): 12-16.

[4] Askari S. a critical note on inverse fuzzy time series algorithms. *Fuzzy Decision Trees and Systems*. 2020.

[5] Yang Liu. Simulation of abnormal data mining algorithm in semi-structured decision tree. *Computer Simulation*. 2020, 37 (10): 230-234.

[6] Abhijit K, Bharat N, Chetan P, et al. Weather Prediction for Tourism Application using Time Series Algorithms. *International Journal of Innovative Technology and Exploring Engineering*, 2020, 9(11).

[7] Bishan W. optimized model of energy industry chain considering low carbon development mechanism. *Energy Sources*, Part A: recovery, utilization, and environmental effec decision tree. 2020, 42 (21).

[8] Wei Wen, Zhao Zhan. Research on Key Technologies of abnormal subject mode monitoring of decision tree P-control chart. *Electrical Measurement and Instrumentation*. 2021, 58 (02): 47-52.

[9] Zheng Zhixue. Research on fault monitoring of cloud computing cluster based on improved decision tree algorithm. *Information Recording Materials*. 2021, 22 (05): 171-172.

[10] Hou D. Determine the impact on fisheries by predicting the migration of fish near Scotland. IOP Conference Series: Earth and Environmental Science. 2021, 631(1).

[11] Guo s, Feng h, Feng a W, et al. Automatic quantification of subsurface defec decision tree by analyzing laser ultrasonic signals using revolutionary neural networks and wavelet transform. 2021.

[12] Dick O E, Glazov A L. Estimation of the synchronization between intermittent photic stimulation and brain response in hypertension disease by the recurrence and synchrosqueezed wavelet transform. *Neurocomputing*. 2021, 455.

[13] Bandyopadhyay K S, Pramanik S, Ghosh R, et al. A New Combinational Technique in Image Steganography. *International Journal of Information Security and Privacy*, 2021, 15(3).

[14] Hsien-Chu W, Wen-Li F, Chwei-Shyong T, et al. An image authentication and recovery system based on discrete wavelet transform and convolutional neural networks. *Multimedia Tools and Applications*. 2021.

[15] Hutchison Z L, Gill A B, Sigray P, et al. A modelling evaluation of electromagnetic fields emitted by buried subsea power cables and encountered by marine animals: Considerations for marine renewable energy development. *Renewable Energy*. 2021, 177.

[16] Chen Shaofei, Liu Xiaojie, Li Hongyang, et al. Outlier screening method for blast furnace ironmaking data. *Journal of Iron and Steel Research*. 2021, 33 (06): 467-475.

[17] Su Linping, Dong Zixian, Li Wei, et al. Personalization supporting multi-attribute generalization (α, l. K) anonymous model. *Computer Technology and Development*. 2021, 31 (06): 88-93.

[18] Zhu Heng Dong, Ma Ying Cang. Semi supervised sparse subspace clustering based on marker discrimination and local linear reinforcement. *Computer Application Research*. 2021: 1-6.