Data-driven Tracing Method of Overload of Transmission Network Equipment

Meng Li 1*, Wei Wu 1, Yi Lin 1, Tongyu Yan1, Pingfa Huang2 and Shengyi Lin2

1Economic and Technology Research Institute of State Grid Fujian Electric Power Company, Fuzhou, 350000, China
2State Grid Xintong Yili Technology Co.Ltd, Fuzhou, 350000, China
*Corresponding author’s e-mail: 995587581@qq.com

Abstract—It is of great significance to the tracing of heavy load or overload (HOD) of power transmission network equipment for the precise investment and planning in power grid. The measurement system of power transmission network is increasingly accurate and intelligent, and a large amount of operation data was generated. Therefore, it is theoretically feasible for the tracing method based on big data is proposed for the HOD transmission network equipment. Firstly, the messy operation data generated by the power grid should be preprocessed, including data coding, cleaning and resampling. Secondly, estimation of the importance of independent variables of mutual information was solved out four characteristic quantities which have a great influence on HOD by game theory. Then, with the characteristic quantity as the label, the Fuzzy C-means (FCM) was used to cluster the HOD data into four categories, corresponding to four causes of HOD. Finally, an example is given to analyze the operation data of a city’s transmission network in 2019, which verifies the effectiveness and accuracy of the proposed method. Compared with the manual tracing method, the algorithm can greatly improve the efficiency.

1. Introduction
In the power transmission network, high or low temperature, holidays and so on will lead to a large increase in load in a short term, which may cause short-term heavy-load or HOD (HOD) of equipment. On the other hand, the maintenance of nearby transmission lines or transformer, also resulting in a change in the distribution of power flow, may also give a rise to the HOD of grid equipment. HOD of transmission equipment causes huge hazard, if not eliminated in time, may bring about large-scale power failure and other irreversible losses. For power grid utility, different reasons of HOD have different ways to deal with: for HOD caused by line maintenance, power grid companies often adjust the unit output and control the power load; For HOD caused by load growth, grid companies may also consider adding transmission lines or transformers to expand transmission capacity. Therefore, it is very valuable to provide accurate cause analysis for grid planners, which can help grid planning and investment more accurate and reasonable.

| The Label | Load rate |
|-----------|-----------|
| Normal    | 0-0.8     |
| Heavy-load| 0.8-1     |
| HOD       | >1        |

Table 1. Load rate classification of Transformer.
However, with the increasing scale of modern power grid, the data can be used and need to be considered in the power grid planning process increases sharply, but the data such as power grid parameters and historical operation information have not yet formed intensive management. At present, manual processing of power grid operation data is not only inefficient, but also prone to human error, resulting in a relative shortage of planners and low degree of planning informatization and intelligence in different degrees at the present stage, which brings great obstacles to the work of power grid planning and analysis. Therefore, it is urgent to establish an intelligent diagnosis algorithm for HOD data of power grid operation, and realize intelligent processing and analysis of HOD data of power grid operation by means of informatization.

There have been related studies on the HOD analysis of the transmission network. Literature [1] analyzes the causes and control measures of the network line HOD, but focuses on the control measures, without studying how to specifically analyze the causes of HOD. In literature [2], through analyzing and mining the big data of power grid operation, the operation index, dispatching index and performance index was established to form the index system. However, the reasons for HOD of power transmission network equipment were also not analyzed. Literature [3] application of independent variable importance based on mutual information for big data technologies, such as the smart grid oriented transformer HOD phenomenon, analysis of various industries HOD probability, but in the end can only be given general industry, temperature and so on four important factors affecting the weight HOD, unable to concrete analysis of each transformer HOD causes. Literature [4] based on random forest predict distribution transformer HOD, improve sample size by heavy sampling, has obtained the good accuracy, but this article focuses on real-time operation of equipment and future short-term prediction, for there has been HOD device cannot be targeted to analyze its reasons, cannot help power grid planning personnel more clear huge amounts of data contains the HOD causes.

Fig.1 Heavy and HOD analysis flow of transmission network equipment based on big data

2. Data preprocessing of transmission network operation

2.1. Data coding
Based on big data analysis of HOD of transmission network equipment, the characteristic variables must be input first. The daily operation data, load properties and adjacent switching states of the transmission network equipment are discretized into 24 data points. For the weather, holidays and other character information and code.
TABLE 2 Data code table

| Name            | Description                                      | code  |
|-----------------|--------------------------------------------------|-------|
| Temperature     | -0°C， 0-10°C， 20-30°C， 30°C                   | 0,1,2,3,4 |
| Weather         | Sunny, rainy, cloudy, snowy                      | 1,2,3,4 |
| Holiday         | Yes, No                                          | 1,2   |
| Weekday or Weekend | Monday, Tuesday…Sunday.                     | 1-7   |
| Season          | Spring, Summer, autumn                          | 1,2,3,4 |
| Adjacent breaker| On, Off                                          | 0,1,2,3,4 |
| Adjacent Switch | On, Off                                          | 0,1   |
| Load category   | Industry, Commercial, Resident                   | 1,2,3 |
| Load rate       | Normal, high-load, HOD                          | 0,1,2 |

Therefore, the time sequence data of a specific substation can be recorded as: \( A = \{\alpha_1, \alpha_2, \ldots, \alpha_{24}\} \). Where, \( \alpha_i \) is the sequential state vector of the device, and there are 24 state vectors.

2.2. Data cleaning

At present, in the power system operation data storage file, the data is often missing or abnormal for a variety of reasons, so it is necessary to clean the data [8]. Data cleaning is mainly divided into: outlier analysis and missing value processing.

The analysis of outliers mainly includes screening out individual sample points in the sample data that deviate from other data points obviously. The boxplot can be used to detect outliers, the acceptance of real data is relatively high, and boxplot is the criterion for identifying abnormal data: we define value between \( Q_L - 1.5*\text{IQR} \) and \( Q_U + 1.5*\text{IQR} \) as normal value. Where, \( Q_L \) is lower quartile, It represents a quarter of all observed data with a value less than that; \( Q_U \) is upper quartile, it represents a quarter of all observed data with a value more than that; \( \text{IQR} \) is quartile interval, representing the difference of \( Q_U \) and \( Q_L \).

![Fig.2 Boxplot detects outliers](image)

The missing value processing methods mainly include fixed value, regression method and interpolation method.

For the time series data, there is always a continuous relation with the samples before and after, so this paper adopts Newton interpolation method to interpolate. The value of \( N \) is 5, which is related to the two samples before and after.

\[
f(x) = f(x_1) + (x - x_1)f[x_2, x_1] + (x - x_1)(x - x_2)f[x_2, x_1, x_3] + \ldots + (x - x_1)(x - x_2) \ldots (x - x_n)f[x_n, x_{n-1}, \ldots, x_1, x]\]

(1)
2.3. Bootstrap resampling

Efron, a professor of statistics at Stanford university, proposed the Bootstrap method in 1977. This method is widely used in research scenarios with limited samples or high sample imbalance [9]. At present, the transmission network operation has indeed produced a large number of data, but the outliers of these data are high, and the HOD rate of the current power grid is generally not high, so the actual HOD sample size is limited, and heavy sampling is needed to enrich the samples.

The basic principle of Bootstrap is shown in FIG. 3.

Fig.3 Chart of Bootstrap principle flow

3. Importance estimation of independent variables of mutual information

In the actual problems of power system, the data is complex. When analyzing the changes of dependent variables, the independent variables are often correlated, and the relative importance of independent variables cannot be directly obtained. The way to solve this problem is to discuss the relative importance of one independent variable in different combinations of all independent variables [10]. The correlation and importance degree of each influencing factor on HOD of equipment are analyzed, and the characteristics with high correlation and importance degree are extracted. The principle is as follows:

According to information theory, discrete random variables $X$, $X \in S_X$, The probability function is $p(x)$, we define its entropy as:

$$ H(X) = - \sum_{x \in S_x} p(x) \log_2 p(x) \tag{2} $$

The statistical dependency of two random variables $X$ and $Y$ is measured by mutual information $I(X; Y)$:

$$ I(X;Y) = \sum_{x \in S_x} \sum_{y \in S_y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{3} $$

If the mutual information of two random variables is large, the correlation between the two random variables is large. The mutual information value can be solved by the entropy value:

$$ I(X; Y)=H(X)+H(Y)-H(XY) \tag{4} $$
The mutual information between two groups of random variables, considering the mutual information between \(X\) and \((Y, Z)\), has the following relationship:

\[ I(X; Y, Z) = H(X) + H(Y|Z) - H(X|Z) \]  

(5)

The relative importance of each eigenvalue is solved by using the game theory for \(P\) different characteristics, \(X = \{X_1, X_2, \ldots, X_p\}\). Function with the eigenvalue of mutual information \(I\), according to Shapley theorem, Calculate \(V_j\), which is the donate value of \(X_n\), the forum is as follow:

\[ V_j(X) = \sum_{S \subset X} \alpha_n(S)[I(S \cup X_j) - I(S)] \]  

(6)

\[ \alpha_n(S) = \frac{s!(p-s-1)!}{p!} \]  

(7)

\(S\) is subsets without \(X_n\), \(b\) is the number of variables in the \(S\), \(p\) is the number of variables.

4. Fuzzy c-means clustering

Fuzzy C-Means Clustering (FCM) algorithm hold the ability to divide data into two or more categories, was found in 1973 by Dunn, and improved by Bezdek, eventually was widely used in pattern recognition\[11\]. FCM cluster data by the follow objective function \[11\]:

\[ \min j = \sum_{c=1}^{c} \sum_{n=1}^{m} u_{ij}^m \|y_j - c_i\|^2 \]  

(8)

\(m\) is any real number greater than 1, \(u_{ij}\) represents degree of membership of \(x_i\) to cluster \(j\), \(x_i\) represents \(i\)th data, \(c_i\) represents \(d\) dimension center of the cluster, \([\| \cdot \|]\) is the canonical expression of any similarity measurement data and center.

In principle, FCM is to obtain the membership of each sample point to the center points of all categories by optimizing the objective function. According to the principle of maximum membership in fuzzy mathematics, determine the category of each sample \[12\]. Among it, Cluster evaluation index \(\text{CHI}(\text{Calinski-Harabasz})\) decides the number of clusters \(C\), the CHI indicator needs to consider the dispersion between different classes \(B\) and the compact \(W\) between the same class:

\[ CHI = \frac{\beta/(K-1)}{B/(N-K)} \]  

(9)

\[ B = \sum_{k=1}^{K} \sum_{l=1}^{N} w_{kl} \|c_k - \bar{x}\|^2 \]  

(10)

\[ w_{kl} = \begin{cases} 1, x_l \in X_k \\ 0, x_l \notin X_k \end{cases} \]  

(11)

Among them, \(\bar{x}\) represents the mean vector of all objects, \(c_k\) represents the center of \(k\)th cluster, \(w_{kl}\) represents the membership relationship between \(l\)th object and \(k\)th cluster, meanwhile, \(K\) is the number of clusters. \(N\) is the number of all the samples.

Based on characteristics set \(Y = \{y_1, y_2, L, y_1, y_2, y_3\}\), we divided samples into \(C\) clusters. Correspondingly, there are \(C\) cluster centers after that. Every sample \(j\) holds the membership \(u_{ij}\) with \(i\)th cluster. Thus, the classification problem is transformed to solve the optimization problem, whose objective function is as formula (8), and restrictions is as formula (12):

\[ \sum_{i=1}^{c} u_{ij} = 1, j = 1, 2, \ldots, n \]  

(12)

Using the Lagrange multiplier method to bring the constraints into the objective function, the above problem can be simplified to be as formula (13) and (14).

\[ u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|y_j - c_k\|^2}{\|y_j - c_k\|^2}ight)^{\frac{2}{m-1}}} \]  

(13)
The FCM algorithm can be attributed to the following pseudocode [13]:

FCM algorithm:

1: Initialize $U=[u_{ij}]$ as zero matrix $U^{(0)}$.
2: At step $k$, use $U^{(k)}$ calculate Center vector $C^{(k)}=[c_{j}]$, $c_{j} = \frac{\sum_{j=1}^{n}(y_{j}u_{ij}^{m})}{\sum_{j=1}^{n}u_{ij}^{m}}$ (14) The pseudo code of TSNE algorithm is as follows:

TSNE algorithm:

1: Initialize data set $X = \{x_{1}, x_{2}, \ldots, x_{n}\}$ Cost function parameters: $\text{perp}$, Optimization parameters: iteration times $T$, learning rate $\eta$. Momentum function $\alpha(t)$
2: Low-dimensional data $y^{T} = \{v_{1}, v_{2}, \ldots, v_{n}\}$
3: Begin formula(15), use $\text{perp}$ calculate $p_{ij} = \frac{p_{ij}^{0} + p_{ij}}{2n}$
4: From $t=1$ to $T$ execute formula(16) calculate $q_{ij}$, formula (17) calculate $\frac{\delta_{C}}{\delta y}$, $y^{T} = y^{T-1} + \eta \frac{\delta_{C}}{\delta y} + \alpha(t)(y^{T-1} - y^{T-2})$
5: End
6: End

5. TSNE data dimensionality reduction

Data visualization technology is an important technology for big data analysis. There are two main types of linear dimensionality reduction and nonlinear dimensionality reduction [14]. Linear dimensionality conclude PCA (Principal Components Analysis), LDA (Linear Discriminant Analysis). Actually, TSNE(stochastic neighbor embedding) is a mature nonlinear dimensionality reduction method, its essence is manifold learning. This paper uses it as a visualization of data dimensionality reduction after clustering, the basic principle is: In data points that are similar in high-dimensional space, the distance mapped to low-dimensional space is also similar. Euclidean distance is usually used to describe this similarity, and TSNE converts this distance relationship into a conditional probability to express similarity [15]. Two data points in high-dimensional space: $x_{i}$ and $x_{j}$, choose $x_{j}$ as its proximity in a conditional probability $P_{ji}$. Considering a GD (Gaussian Distribution) that $x_{i}$ is the center point, if $x_{j}$ is closer to $x_{i}$, then $P_{ji}$ is greater. Therefore, $P_{ji}$ can be define as:

$$p_{ji} = \frac{\exp(-\frac{1}{2\sigma^{2}})}{\sum_{k=1}^{n} \exp(-\frac{1}{2\sigma^{2}})}$$

(15)

TSNE adopts the symmetric SNE approach for the distribution in high dimensions, and the more general T distribution for the distribution in low dimensions, which is also symmetric.

$$q_{ij} = \left(1 + \frac{1}{\sigma^{2}}\right)^{-1}$$

(16)

$$\frac{\delta_{C}}{\delta y} = 4 \sum_{j}(p_{ij} - q_{ij})(x_{i} - x_{j})(1 + \frac{1}{\sigma^{2}})^{-1}$$

(17)

The pseudo code of TSNE algorithm is as follows:

5.1. TSNE algorithm:

1: Initialize data set $X = \{x_{1}, x_{2}, \ldots, x_{n}\}$ Cost function parameters: $\text{perp}$, Optimization parameters: iteration times $T$, learning rate $\eta$. Momentum function $\alpha(t)$
2: Low-dimensional data $y^{T} = \{v_{1}, v_{2}, \ldots, v_{n}\}$
3: Begin formula(15), use $\text{perp}$ calculate $p_{ij} = \frac{p_{ij}^{0} + p_{ij}}{2n}$
4: From $t=1$ to $T$ execute formula(16) calculate $q_{ij}$, formula (17) calculate $\frac{\delta_{C}}{\delta y}$, $y^{T} = y^{T-1} + \eta \frac{\delta_{C}}{\delta y} + \alpha(t)(y^{T-1} - y^{T-2})$
5: End
6: End
6. Case study
This article takes the heavy or HOD operation data of a provincial transmission network (220kV and above) from September to November 2019 as an example to analyze the causes of HOD of its equipment. From September to November, the province has a total of 3782179232 transmission grid operation data, including 4419 HOD data, involving 41 HODEd transformers and 26 lines. For the analysis of these 4,419 pieces of data, the analysis software is Anaconda3, and the python environment is 3.7.0. The input data feature is shown in Table 3 [16].

| Serial | Feature variables          | Serial | Feature variables         |
|--------|----------------------------|--------|---------------------------|
| 1      | Daily maximum temperature  | 11     | Transformer cooling       |
| 2      | Daily minimum temperature  | 12     | Production date           |
| 3      | Weather                    | 13     | Nearby equipment failure  |
| 4      | Weekday or Weekend         | 14     | Nearby equipment          |
| 5      | Holiday                    | 15     | Line length               |
| 6      | Season                     | 16     | Line split number         |
| 7      | Transformer startup time   | 17     | Allowable Current         |
| 8      | Transformer capacity       | 18     | Voltage level             |
| 9      | Main and standby           | 19     | Outage sign               |
| 10     | Transformer protection     | 20     | Main connection           |

6.1. Characteristic variables analysis
In order to extract feature quantities with high correlation and high importance, in combination with the foregoing, in the calculation example, the feature quantities and probability distributions are counted based on historical data. Taking the transformer as an example, Figure 3 shows the temperature and load rate statistics from August to October 2019. Through the calculation formula of mutual information and the theory of countermeasures, the importance of each feature is solved. The four more important features are as follows: Figure 4. It can be seen from Figure 4 that the maintenance of equipment near the transmission equipment has a high degree of impact on the HOD of the equipment, with an importance of 0.82; the daily maximum temperature has a significant impact on the load, which may also cause HOD of the transmission grid equipment, the daily maximum temperature The importance is 0.65; the daily minimum temperature has little effect on the equipment, the importance is 0.43; the importance of the holiday is 0.3, this is because the impact of the holiday on the load is relatively unstable, sometimes industrial users adjust the holiday, the holiday continues to produce, industrial load Stable, while the commercial load and residential load increase sharply during the holiday period, so the total load is increased, which may cause HOD of transmission equipment, but sometimes industrial users take a normal holiday, which will cause the load to drop and will not cause heavy or HOD.

Fig.4 Graph of temperature and load rate of a transformer from August to October, 2019
6.2. Analysis of clustering results

After encoding and cleaning the original data and re-sampling, the four categories are obtained by fuzzy C clustering. The load rate curves of the four categories are shown in Figure 5. The time series of category one and two refer to the time range of the data used, unit: days; the load rate is the load rate of 10 days per day. It can be seen from category 1 and category 2 that the load rate suddenly increases at a certain period of time, which may be caused by the maintenance of nearby equipment, which causes the power flow to change. The similarity of the curve is high, and the cause of HOD of the device cannot be derived from the load rate curve only. Therefore, the fuzzy C mean is the four more important feature quantities such as comprehensive equipment maintenance and so on. See the difference between the four categories.

The normal load growth will result in long-term heavy-duty operation of the transmission equipment. The load rate curve is shown in (d). This type of equipment runs for a long time near the heavy-load line, and it should be expanded in time. New transmission equipment should be built to cope with the normal increase in load. Effective and accurate HOD analysis can provide decision support for precise investment in transmission grids [17].

In Figure 7, the clustering results are reduced by TSNE to a two-dimensional plane, and the data is clearly divided into four categories, corresponding to the four types of HOD causes: red is the
maintenance factor of nearby equipment; yellow is the holiday factor; green is the temperature influence factor; Blue is the normal load growth factor.

![TSNE graph of clustering result](image)

**Fig.7 TSNE graph of clustering result**

6.3. **Algorithm accuracy analysis**

Sampling verification of the clustering results of the calculation examples and comparing with the results of manual calculation. The calculation time is only 23.44 seconds. The time for manual analysis is not easy to quantify. The short is one week and the long is January. The big data method is much faster than manual analysis. In terms of accuracy, the correct rate of manual analysis is only 76.2%.

**TABLE 4 Accuracy and calculating time of big data analysis.**

| Computer       | Win10 64bits; ROM 12G; |
|----------------|------------------------|
| Software       | Anaconda3 python3.7.0  |
| Accuracy       | 93.1%                  |
| Time of calculate| 23.44s                |

![Accuracy comparison](image)

**Fig.8 Comparison of big data analysis and human analysis in accuracy**

7. **Conclusion**

This paper proposes a data-driven method for tracing the HOD of transmission grid equipment. It analyzes the cause of HOD of transmission grid equipment through big data technology, reduces the manual calculation of the cause of HOD of transformers by grid planning designers, and analyzes the amount of labor and time to improve the analysis efficiency and correct rate. The analysis of calculation examples shows that the proposed method can effectively screen out the main HOD impact feature quantities, and the fuzzy C clustering can quickly and accurately determine the cause of HOD of different devices, with an accuracy of up to 93.1%, and visualize the clustering results, Can effectively improve the level of grid planning informatization.
It is worth noting that fuzzy C clustering requires sufficient data. In practice, if the time window is short, the insufficient data will affect the accuracy of the HOD analysis. How to analyze the cause of HOD more accurately in the case of insufficient data is the direction of the next research.

Acknowledgement
This work is supported by State grid science and technology project (52130N18000S)

References
[1] Li Shihui, Shi Min, Dong Liang, et al. Causes Analysis Control Measures on Line HOD of Power Network. Hebei Power Technology. 2008(04): 39-42.
[2] Liu Kaicheng, He Guangyu, Chen Rui, et al. Implementation of Transmission Network Operation Evaluation System in Fujian Power Grid. Electric Power Construction. 2014, 35(08):18-23.
[3] He Jianzhang, Wang Haibo, Ji Zhixiang, et al. Analysis of Factors Affecting Distribution Transformer HOD in Smart Grid. Power System Technology. 2017, 41(01): 279-284.
[4] He Jianzhang, Wang Haibo, Ji Zhixiang, et al. HOD Forecasting of Distribution Transformers Based on Random Forest Theory. Power System Technology. 2017, 41(08): 2593-2597.
[5] Wang Dong, Zhu Yongli, et al. Fault diagnosis of transformer based on fuzzy C means clustering and improved relevance vector machine. Electrical Measurement & Instrumentation. 2019, 56(12):14-19.
[6] Xie Wei, Zhao Qi, Guo Naiwang, et al. Application of the improved parallel fuzzy kernel C-means clustering algorithm in power load forecasting. Electrical Measurement & Instrumentation. 2019,56(11):49-54+60.
[7] Wu Lei, Fang Bin, Diao Liping, et al. Imbalanced data resampling based on oversampling and under-sampling[J]. Computer Engineering Applications, 2013 (21) : 172-176
[8] Liu Huankun, Luan Wenpeng, Xu, Yan, et al. Data Cleaning Method for Distribution Transformer. Power System Technology, 2017, 41(03): 1008-1014.
[9] Gao Yanping, Wang Jing. Modeling Method of Repeated Sampling Under Finite Sample Condition. Statistics & Decision, 2019, 35(08): 22-25.
[10] Xu Wenjin, Xu Yao, Xie Qin. et al. Improved K-Means Traffic Data Clustering Based on Mutual Information and Divergence[J]. Computer Systems & Application, 2020, 29(01):171-175.
[11] Sun Lijing, Sheng Wanxing, Wu Ming, et al. Estimation Method of DG Clusters Grid-connected Inverter’s Output Power Confidence Interval Based on Fuzzy C-means Algorithm[J]. Power System Technology, 2019,43(05):1495-1503.
[12] Dong Zhihui, Lin Lingxue. Dynamic Reconfiguration of Distribution Network Based on Improved Fuzzy C-means Clustering of Time Division[J]. Power System Technology, 2019, 43(07): 2299-2305.
[13] Sheng Wanxing, Ji Yu, Wu Ming, et al. Dynamic Clustering Modeling of Regional Centralized Photovoltaic Power Plant Based on Improved Fuzzy C-Means Clustering Algorithm[J]. Power System Technology, 2017, 41(10): 3284-3291.
[14] Dong Anguo, Zhang Qian, Liu Hongchao, et al. Hyperspectral Image Classification Based on TSNE and Multiscale Sparse Auto-Encoder[J]. Computer Engineering and Applications, 2019, 55(21): 177-182+219.
[15] Maaten L V D.Accelerating t-SNE using tree-based algorithms[J]. Journal of Machine Learning Research, 2014, 15 (1) : 3221-3245.
[16] Li Junqin, Li Qiqua, Shi Tianyu, et al. Feature selection method for wind power prediction based on data mining[J]. Electrical Measurement & Instrumentation, 2019, 56(10): 87-92.
[17] Liu Xuna, Wei Jun, Zhang Wentao. Investment benefits evaluation and decision for distribution network based on information entropy and fuzzy analysis method[J]. Power System Protection and Control, 2019, 47(12): 48 - 56.