IOP Conf. Series: Earth and Environmental Science **133** (2018) 012001 doi:10.1088/1755-1315/133/1/012001

# Prediction of line failure fault based on weighted fuzzy dynamic clustering and improved relational analysis

**Xiaocheng Meng [1, 2], Renfei Che [1, 2], Shi Gao [1, 2, 3] and Juntao He[1, 2]**

[1]Key Laboratory of Power System Intelligent Dispatch and Control of Ministry of Education, Jinan, 250061, CHN

[2]School of Electrical Engineering, Shandong University, Jinan, 250061, CHN

[3]State Grid Hebei Maintenance Branch, Shijiazhuang, 050070, CHN

**Abstract.** With the advent of large data age, power system research has entered a new stage. At present, the main application of large data in the power system is the early warning analysis of the power equipment, that is, by collecting the relevant historical fault data information, the system security is improved by predicting the early warning and failure rate of different kinds of equipment under certain relational factors. In this paper, a method of line failure rate warning is proposed. Firstly, fuzzy dynamic clustering is carried out based on the collected historical information. Considering the imbalance between the attributes, the coefficient of variation is given to the corresponding weights. And then use the weighted fuzzy clustering to deal with the data more effectively. Then, by analyzing the basic idea and basic properties of the relational analysis model theory, the gray relational model is improved by combining the slope and the Deng model. And the incremental composition and composition of the two sequences are also considered to the gray relational model to obtain the gray relational degree between the various samples. The failure rate is predicted according to the principle of weighting. Finally, the concrete process is expounded by an example, and the validity and superiority of the proposed method are verified.

## 1. Introduction

As the information industry's rapid development, the world has been fully into the era of big data. At present, the power transmission and power distribution are the key points in the application of large data in power system[1]-[2]. The main aspect is the equipment fault diagnosis, that is, grid equipment status monitoring and early warning analysis.

In the data processing and feature mining of big data, the commonly used machine learning algorithms involve fuzzy clustering and gray relational analysis, etc. In the case of fuzzy clustering, the fuzzy clustering method is used in the literature [3-4] to cluster the original data, however, these fuzzy

clustering algorithms do not take into account the differences between the various factors, believe that the importance of the various factors between the indicators are equal, and this is clearly inconsistent with the actual. In literature [5-6], a method is proposed to solve the difference of factors. The article adopts the weighted way to distinguish the importance of the data, and discusses the problem where the weight should be added in the algorithm. In literature [7-8], some common algorithms of use of weight are discussed, and the application of each algorithm is put forward, and it is pointed out that the weighted weight should accord with the rationality analysis. In the case of gray relational analysis, the literature [9-10] points out the specific application of gray relational analysis in fault identification, index evaluation and equipment maintenance, but it does not improve the gray relational model itself.

Through the above related literature, and combined with the collection of relevant historical data and comprehensive analysis of electrical equipment early warning, the coefficient of variation is determined as the weight, and the classification of historical data is carried out by weighted fuzzy dynamic clustering; then, based on the basic idea and basic nature of gray relational analysis, the concept of Deng's slope is proposed, and the gray relational analysis is improved by using the incremental composition ratio and composition difference of two sequences, and the gray relational degree between the various sequences is obtained, and finally through the idea of the weight to predict the failure rate of the line.

## 2. Determination of the weight

In this paper, through the comparison of common weight assignment methods, combined with rationality analysis, the final introduction of coefficient of variation to determine the weight of each meteorological indicators.

The coefficient of variation uses the degree of variation of the evaluated object to carry out the assignment of the weight, and the dynamic assignment of the index can be realized. In the meteorological factors, the factors that have a high degree of influence on the line fault often have the characteristics of high degree of dispersion and high volatility, that is, larger variance. Therefore, this paper combines variance and coefficient of variation to determine the weight of each meteorological factors.

In this paper, a total of 10 meteorological sequences were collected to evaluate the meteorological indicators of a total of four, $c_{ij}$ indicates the $j$-th meteorological index of the $i$-th meteorological sequence, $i=1,2,3,\cdots,m$, $j=1,2,3,\cdots,n$, $m=10$, $n=4$。

The specific calculation steps are as follows:

- Calculate the mean square error of the $i$-th meteorological index

$$\begin{cases} S_j = \sqrt{\sum_{i}^{m} (c_{ij} - \bar{c}_j)^2 / m} \\ \bar{c}_j = \dfrac{\sum_{i=1}^{m} c_{ij}}{m} \end{cases} \tag{1}$$

Where $\bar{c}_j$ is the $j$-th meteorological index and $S_j$ is the mean square error of the $j$-th meteorological index.

- Calculate the coefficient of variation for the $j$-th meteorological index.

$$\delta_j = \frac{S_j}{\bar{c}_j} \tag{2}$$

Where $\delta_j$ is the coefficient of variation of the $j$-th meteorological index.

- Normalize the coefficient of variation, get the weight of the first p meteorological indicators.

$$\begin{cases} A_j = \dfrac{\delta_j}{\sum \delta_j} \\ \sum\limits_{j=1}^{n} A_j = 1 \end{cases} \tag{3}$$

Where $A_i$ represents the weight of the $j$-th meteorological index.

## 3. Weighted fuzzy dynamic clustering

The traditional fuzzy dynamic clustering algorithm does not take into account the differences between the various factors, This paper uses a weighted approach to represent the different degrees of importance between the various factors.

The specific methods are as follows:

- Normalize the original meteorological data $c_{ij}$ according to the deviation standardization method.

$$d_{ij} = \frac{c_{ij} - \min\limits_{1\le j\le n}\{c_{ij}\}}{\max\limits_{1\le j\le n}\{c_{ij}\} - \min\limits_{1\le j\le n}\{c_{ij}\}} \tag{4}$$

- Calculate the weight of each meteorological factor according to the steps of the coefficient of variation

$$a = [a_1, a_2, a_3, \cdots, a_n] \tag{5}$$

- Do weighting after normaling data

$$f = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix} * \begin{bmatrix} a_1 & 0 & 0 & 0 \\ 0 & a_2 & 0 & 0 \\ \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & a_m \end{bmatrix} \tag{6}$$

- Solve Fuzzy Similarity Matrix.

$$R_{ij} = \frac{\sum\limits_{k=1}^{m}(f_{ik} \wedge f_{jk})}{\sum\limits_{k=1}^{m}(f_{ik} \vee f_{jk})} \tag{7}$$

Where $\vee$ for the big operation, $\wedge$ for the small operation.

- Solve $t(R)$ according to square method

$$\begin{cases} t(R) = R^{2k} \\ R \to R^2 \to R^4 \to R^8 \to \cdots R^{2k} = R^{2(k+1)} \\ R^2 = R \circ R = (h_{ij})_{m\times n},\ h_{ij} = \bigvee_{k=1}^{m}(R_{ik} \wedge R_{kj}) \end{cases} \tag{8}$$

- Determine the optimal threshold $\lambda$.

In this paper, the $F$ statistic is used to determine the optimal threshold.

The number of samples corresponding to $\lambda$ is $r$, the number of samples of the $j$-th class is $n_j$, the average of the $k$-th variable in the $j$-th class is $\bar{x}_k^{(j)}$, and the average of the $k$-th variable of the sample is $\bar{x}_k$:

$$\begin{cases} \bar{x}_k^{(j)} = \dfrac{1}{n_j}\sum_{i=1}^{n_j} x_{ik}^{(j)},\ (k = 1, 2, \cdots, m) \\ \bar{x}^{(j)} = (\bar{x}_1^{(j)}, \bar{x}_2^{(j)}, \cdots, \bar{x}_m^{(j)}) \end{cases} \tag{9}$$

$$\begin{cases} \bar{x}_k = \dfrac{1}{n}\sum_{i=1}^{n} x_{ik},\ (k = 1, 2, \cdots, m) \\ \bar{x} = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_m) \end{cases} \tag{10}$$

Define the F statistic as follows:

$$F = \frac{\sum_{j=1}^{r} n_j \left\| \bar{x}^{(j)} - \bar{x} \right\|^2 / (r-1)}{\sum_{j=1}^{r}\sum_{i=1}^{n_j} \left\| x_i^{(j)} - \bar{x}^{(j)} \right\|^2 / (n-r)} \tag{11}$$

$F$ obey $F$ distribution rule which the degree of freedom is $r-1$, $n-r$. For a given test level $\alpha$, check the $F_\alpha(r-1, n-r)$ distribution table, get the critical value of $F_\alpha$, if $F > F_\alpha$, then consider that there are significant differences between the various types. If the number of $F > F_\alpha$ is not just one, then compare the ratio of $(F - F_\alpha)/F_\alpha$ to choose the larger one.

## 4. Improvement of Gray Relational Analysis

Gray relational analysis is one of the important components of gray system theory, and it is a quantitative and comparative method for the development of gray dynamic process.

In the evaluation of the gray relational model, the four principles of the Deng's gray relation is the basic criterion of the test model, and it is also the basic idea to be followed in the gray relational analysis. But the defects of Deng's relational degree are obvious, including: the least absolute difference between the two levels and the maximum absolute difference between the two stages is serious; For a given

resolution factor, the dimensionless processing is not warranted; it cannot meet the normative requirements.

In this paper, the above-mentioned nature is used as the basic idea to improve the gray relational analysis, and take into account the indicators among measuring the degree of similarity of the sequence curve, only slope is the indicator with inheriting order property, so combine Deng's relational method with the conception of slop, named Deng's slop ,to solve the problem of inheriting order property of Deng's slop. And also, this paper uses two series of incremental composition ratio and composition difference to define the relational coefficient, which can make full use of the information contained in the data sequence, which reflects the more real degree of relation between data sequence.

The specific definition is as follows:

Define the Deng's slope of the sequence $X_i(i=0,1,2,\cdots,m)$ at $k$ is $Z_i(k)$, which is calculated as follows:

$$Z_i(k)=\frac{\min\limits_i\min\limits_k\left|\frac{x_i(k)-x_i(k-1)}{\frac{1}{m(n-1)}\sum\limits_{i=1}^{m}\sum\limits_{k=2}^{n}\left|x_i(k)-x_i(k-1)\right|}\right|+\rho\max\limits_i\max\limits_k\left|\frac{x_i(k)-x_i(k-1)}{\frac{1}{m(n-1)}\sum\limits_{i=1}^{m}\sum\limits_{k=2}^{n}\left|x_i(k)-x_i(k-1)\right|}\right|}{\left|\frac{x_i(k)-x_i(k-1)}{\frac{1}{m(n-1)}\sum\limits_{i=1}^{m}\sum\limits_{k=2}^{n}\left|x_i(k)-x_i(k-1)\right|}\right|+\rho\max\limits_i\max\limits_k\left|\frac{x_i(k)-x_i(k-1)}{\frac{1}{m(n-1)}\sum\limits_{i=1}^{m}\sum\limits_{k=2}^{n}\left|x_i(k)-x_i(k-1)\right|}\right|} \tag{12}$$

where $\rho\in[0,1]$,usually $\rho=0.5$.

The relational degree between $X_i$ and $X_0$ is defined as:

$$\gamma(X_0,X_i)=\begin{cases}\dfrac{1}{n-1}\sum\limits_{k=2}^{n}\dfrac{\operatorname{sgn}((x_0(k)-x_0(k-1))\cdot(x_i(k)-x_i(k-1)))}{1+\frac{1}{2}\left|Z_0(k)-Z_i(k)\right|+\frac{1}{2}\left(1-\frac{\min(Z_0(k),Z_i(k))}{\max(Z_0(k),Z_i(k))}\right)} \\ 1\end{cases} \tag{13}$$

When $(x_0(k)-x_0(k-1))$ and $(x_i(k)-x_i(k-1))$ are not 0 at the same time, $\gamma(X_0,X_i)$ equals the upper formula; otherwise $\gamma(X_0,X_i)$ equals 1.

And also,

$$\operatorname{sgn}((x_0(k)-x_0(k-1))\cdot(x_i(k)-x_i(k-1)))=\begin{cases}1 & when(x_0(k)-x_0(k-1))\cdot(x_i(k)-x_i(k-1))\geq 0\\ -1 & when(x_0(k)-x_0(k-1))\cdot(x_i(k)-x_i(k-1))<0\end{cases} \tag{14}$$

## 5. Early Warning of Failure Probability

When the new meteorological condition is obtained, first, determine which classification it belongs to according to the fuzzy clustering, and then calculate the relational degree between the reference sequence and the comparison sequence among the classification, final, calculate the failure rate $p$ under the meteorological condition according to weighted principle.

$$p = \frac{\sum_{i=1}^{n} p_i \gamma_{0i}}{\sum_{i=1}^{n} \gamma_{0i}} \tag{15}$$

Where $\gamma_{0i}$ is the gray relational degree of the reference sequence and the $i$-th comparison sequence in the classification, and $p_i$ is the failure rate of the $i$-th comparison sequence in the classification.

## 6. Example

Based on the historical data of a city power supply line, 10 sets of data sequences are selected for example verification. As shown in Table 1.

**Table 1.** The statistics of line failure rate

| Meteorological sequence | Maximum wind speed （m/s） | average temperature (℃) | average relative humidity (%) | rainfall (mm) | Line failure rate (times·100/month/km) |
|---|---|---|---|---|---|
| x1 | 10.1 | 25.2 | 73 | 70.9 | 0.9414 |
| x2 | 5.2 | 24.3 | 79 | 32.3 | 0.3096 |
| x3 | 4.1 | 32.3 | 85 | 19.8 | 0.5616 |
| x4 | 4.3 | 30.7 | 84 | 119.5 | 0.4968 |
| x5 | 9.2 | 25 | 81 | 75.1 | 0.8316 |
| x6 | 3.5 | 30.5 | 89 | 25.7 | 0.5292 |
| x7 | 4.9 | 32.6 | 90 | 29.7 | 0.6156 |
| x8 | 8.2 | 24.9 | 80 | 82 | 0.8946 |
| x9 | 9 | 26.3 | 78 | 86.5 | 0.8496 |
| x10 | 3.8 | 31.6 | 88 | 23.2 | 0.5868 |

Using the coefficient of variation method, through the formula (1)~(3), the weight value of the maximum wind speed, the average temperature, the average relative humidity and the rainfall is $a = [0.3406, 0.1006, 0.0543, 0.5045]$.

Theoretically, the rainfall and the maximum wind speed are likely to cause the line fault, while the temperature and humidity have less influence. Therefore, the weight value obtained by the coefficient of variation method is consistent with the actual situation, which meets the requirements of the rationality analysis.

The original data are normalized and weighted according to (4)~(8), and the transitive closure matrix is obtained by the square method. The transitive closure matrix $t(R)$ is obtained as follows.

$$t(R)=\begin{bmatrix} 1 & 0.2899 & 0.2899 & 0.4650 & 0.8543 & 0.2899 & 0.2899 & 0.8543 & 0.8543 & 0.2899 \\ 0.2899 & 1 & 0.4626 & 0.2899 & 0.2899 & 0.4626 & 0.4626 & 0.2899 & 0.2899 & 0.4620 \\ 0.2899 & 0.4626 & 1 & 0.2899 & 0.2899 & 0.7371 & 0.6099 & 0.2899 & 0.2899 & 0.7371 \\ 0.4650 & 0.2899 & 0.2899 & 1 & 0.4650 & 0.2899 & 0.2899 & 0.4650 & 0.4650 & 0.2899 \\ 0.8543 & 0.2899 & 0.2899 & 0.4650 & 1 & 0.2899 & 0.2899 & 0.8630 & 0.8630 & 0.2899 \\ 0.2899 & 0.4626 & 0.7371 & 0.2899 & 0.2899 & 1 & 0.6099 & 0.2899 & 0.2899 & 0.7585 \\ 0.2899 & 0.4626 & 0.6099 & 0.2899 & 0.2899 & 0.6099 & 1 & 0.2899 & 0.2899 & 0.6099 \\ 0.8543 & 0.2899 & 0.2899 & 0.4650 & 0.8630 & 0.2899 & 0.2899 & 1 & 0.8691 & 0.2899 \\ 0.8543 & 0.2899 & 0.2899 & 0.4650 & 0.8630 & 0.2899 & 0.2899 & 0.8691 & 1 & 0.2899 \\ 0.2899 & 0.4626 & 0.7371 & 0.2899 & 0.2899 & 0.7585 & 0.6099 & 0.2899 & 0.2899 & 1 \end{bmatrix}$$

The cross matrix $\lambda$, obtained by passing the closure matrix, is $\lambda=[1,0.8691,0.8630,0.8543,0.7585,0.7371,0.6099,0.4650,0.4626,0.2899]$

According to the cross matrix $\lambda$, the dynamic clustering diagram is shown in Fig.1

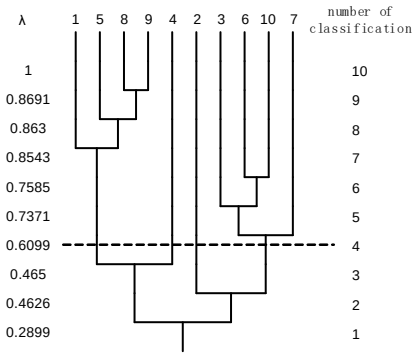


**Figure 1.** Dynamic clustering graph

This paper takes the significance level $\alpha=0.05$, by querying the F distribution table, according to the formula (9)~(11), calculates $F$ and $(F-F_\alpha)/F_\alpha$ in each classification. The results are shown in Table 2.

**Table 2.** Distribution table of $F$ and $(F-F_\alpha)/F_\alpha$

| **Classification number** | 7 | 6 | 5 | 4 | 3 | 2 |
|---|---|---|---|---|---|---|
| **F** | 24.266 | 37.594 | 53.186 | 63.200 | 12.167 | 24.277 |
| **F distribution table** | 8.94 | 6.26 | 5.19 | 4.76 | 4.74 | 5.32 |
| **ratio** | 1.714 | 5.005 | 9.248 | 12.277 | 1.567 | 3.563 |

The results show that the samples divided into four classes have the best results, as $0.465 < \lambda \le 0.6099$. According to the dynamic clustering graph, the specific classification is: the first classification $L_1 = \{x_1, x_5, x_8, x_9\}$, the second classification $L_2 = \{x_2\}$, the third classification $L_3 = \{x_3, x_6, x_7, x_{10}\}$, the fourth classification $L_4 = \{x_4\}$. Taking the sample $\mathrm{x10}$ as an example, according to the formula (13)~(14), the not improved gray relational degree is $\gamma = [0.7752, 0.8116, 0.6435]$. According to the formula (15)~(17), the improved gray relational degree is $\gamma = [0.9189, 0.9102, 0.8076]$.

According to (18), predicted failure rate is: not improved gray relational degree is $p = 0.5654$; improved gray relational degree is $p = 0.5670$.

The predicted failure rate of all samples is shown in Table 3.

**Table 3.** Table of failure rate prediction

| Gray relation | | Not improved | | Improved | |
|---|---|---|---|---|---|
| Sample numble | Line failure rate | Predicted failure rate | Difference percentage | Predicted failure rate | Difference percentage |
| x1 | 0.9414 | 0.8572 | 0.0894 | 0.8583 | 0.0883 |
| x2 | 0.3096 | 0.3096 | 0.0000 | 0.3096 | 0.0000 |
| x3 | 0.5616 | 0.5785 | 0.0301 | 0.5766 | 0.0267 |
| x4 | 0.4968 | 0.4968 | 0.0000 | 0.4968 | 0.0000 |
| x5 | 0.8316 | 0.8948 | 0.0760 | 0.8956 | 0.0770 |
| x6 | 0.5292 | 0.5880 | 0.1111 | 0.5872 | 0.1096 |
| x7 | 0.6156 | 0.5593 | 0.0915 | 0.5595 | 0.0911 |
| x8 | 0.8946 | 0.8700 | 0.0275 | 0.8737 | 0.0234 |
| x9 | 0.8496 | 0.8856 | 0.0424 | 0.8891 | 0.0465 |
| x10 | 0.5868 | 0.5654 | 0.0365 | 0.5670 | 0.0337 |
| Average difference | | | 0.0504 | | 0.0496 |

It can be concluded from the table that the improved gray relational analysis improves the accuracy of the prediction and the predicted difference is within the allowable range, which verifies the effectiveness of the proposed method.

**References**

[1] Dewen Wang and Zhiwei Sun 2015 Big data analysis and parallel load forecasting of electric power user side *Proceedings of the CSEE* vol 35 pp 527-537.

[2] Teng Zhao, Yan Zhang and Dongxia Zhang 2014 Application technology of big data in smart distribution grid and its prospect analysis *Power System Technology* vol 38 pp 3305-3312.

[3] Jie Gai, Qunzhan Li and Jia Wang 2016 Modal parameter identification of low frequency oscillation through NExT-ERA based on fuzzy clustering *Power System Protection and Control* vol 44 pp 40-49.

[4] Yongjun Zhang, Chao Chen and Liang Xu 2011 Prediction of original reliability parameters of power system based on fuzzy clustering and similarity *Power System Protection and Control* vol 39 pp 1-5.

[5] Guifen Chen, Liying Cao and Guowei Wang 2009 Application of weighted spatially fuzzy dynamic clustering algorithm in evalution of soilfertility *Scientia Agricultura Sinica* vol 42 pp 3559-3563.

[6] Guowei Wang, Li Yan and Guifen Chen 2010 Weighted spatially fuzzy dynamic clustering algorithm *Computer Engineering and Applications* vol 46 pp 146-149.

[7] Lianju Ning and Meng Li 2011 Evaluation model for large and medium-sized industrial enterprises' technological innovation capability based on factor analysis method *Science Research Management* vol 32 pp 51-58.

[8] Wei Zhao, Jian Lin and Shufang Wang 2013 Influence of human activities on groundwater environment based on coefficient variation method *Environmental Science* vol 34 pp 1277-1283.

[9] Ganyun Lv, Haozhong Chen and Haibao Zhai 2004 Fault diagnosis of power transformer based on improved grey relationl analysis *Proceedings of the CSEE* vol. 24; pp. 121-126.

[10] Jian Shen, Shujuan Shi and Yang Zhou 2014 Surface water environmental quality assessment of Danjiankou valley based on improved grey relational analysis *Environmental Monitoring in China* vol 30 pp 41-46.