ISSN 2706-655X Vol.3, Issue 4: 20-23, DOI: 10.25236/IJFET.2021.030403

Comprehensive evaluation model of abnormal production data based on Entropy Weight Method and TOPSIS Method

Linfeng Jiang, Fujie Sun, Jingxi Lian*

South China Agriculture University, Guangzhou, Guangdong, 510642, China
**Corresponding author: jlfmaths@163.com*

Abstract: *The paper is to build a mathematical model of the time series data recorded by the equipment in the production area from 00:00:00-22:59:59 on a certain day and the data used in the experiment is from the May Day Mathematical Contest in Modeling. According to the two characteristics of risky data sustainability and linkage, the Risky Data Detection Model based on partition and statistical analysis is established, and then dividing the abnormal group into non-risky data and risky data. After normalization, establishing a Comprehensive Evaluation Model based on Entropy Weight Method and TOPSIS Method, scoring the degree of abnormal on the data among each moment.*

Keywords: *Risk Anomaly Data, Entropy Weight Method, TOPSIS Method*

1. Introduction

Ensuring safety and preventing risks are the most fundamental base line to promote the high-quality development of production enterprises, and the data generated in the production process can reflect the potential risks immediately. According to the time series data recorded by the equipment from 00:00:00-22:59:59, we establish a mathematical model to judge whether risk anomaly data is or not. The quantitative evaluation method for the degree of risk anomaly data is also given, and evaluate the risky degree of data at each time by using the hundred-mark system, and then find the five moments with the highest abnormal score in the data and the number of the corresponding anomaly sensor at the five times, and evaluate the obtained results by mathematical model.

2. Establishment and solution of the model

2.1 Establishing Risk Anomaly Detection Model

2.1.1 Descriptive statistics of the data

First, this paper conducts exploratory analysis of the data, conducting dimensionality reduction processing according to the descriptive statistics such as standard difference and extreme difference. Then, drawing the quantile-quantile plot to roughly judge whether the variable conforms to the normal distribution. Finally, selecting the small fluctuating variables by standard difference, reducing the characteristic dimension according to the machine-learning fliter method to reduce the number of independent variables.

2.1.2 Filter the abnormal data by Box-type Diagram Method

The Tukey Box-plot Method adopts the denominator and the mean to avoid the outlier interference to the value, so this paper is used to filter the abnormal data. The following are the box chart screening steps:

1. Calculate the first and third quartile: (Q1,Q3)
2. Calculate distance: IQR = Q3-Q1
3. Output the interval of normal data: (Q1-1.5IQR,Q3+1.5IQR)

The data out of the normal-interval can be considered outliers according to empirical criteria, which are used to filter all sensor data, identify the outliers and mark them.

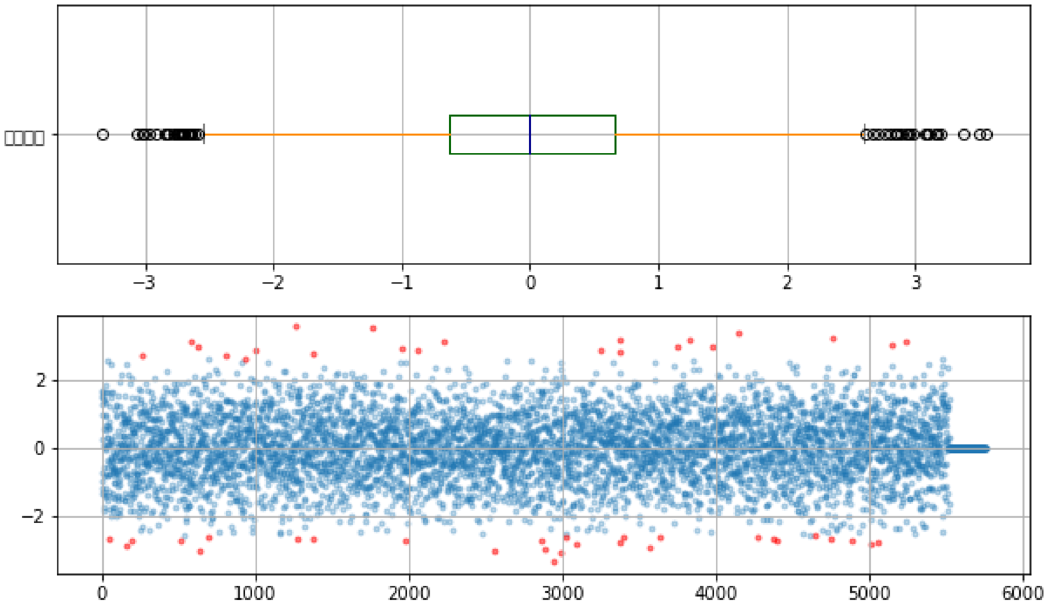

Figure 1: Box-plot and outlier scatter graph

2.1.3 Distinguish risk / non-risk anomaly data by Divide and Conquer Algorithm

This paper gives the following rules for the screening of anomalous data with linkage and sustainability:

Linkage: When a set of abnormal data appears more than N times at the same time, the linkage can be considered to exist among the row data. According to the significance principle of $\alpha = 5\%$ and the number of sensors, N=5 is identified as the linkage determination threshold. The data appears no less than 5 outliers at the same time can be considered to be of linkage.

Sustainability: When the data of the same index occur multiple ouliers over time and the number of abnormal data occurred continuously is more than K=3, that is, the partial correlation function truncates in the second order, which means y_t is affected by y_{t-1} and y_{t-2}. The interaction influence between two consecutive time series, sustainability of the data can be determined in the time period.

After the linkage and sustainability discrimination of abnormal data and unioned processing,the abnormal data is divided into non-risk abnormal data and risk abnormal data.

2.2 Establishing a Comprehensive Evaluation Model based on Entropy Weight Method and TOPSIS Method

2.2.1 Index Weight is obtained by the Entropy Weight Method

Due to the fact that the indicators may be temperature, concentration, pressure, etc., and the unit of measurement is not unified, the first step will be standardized and the different nature index values are homogenized. Then, the normal and non-risk abnormal data can be marked, ahead of dimensionality reduction processing which drops the all normal and non-risk abnormal data sensors and selects the appropriate indicators. The risk abnormal data can be transformed by the following processing, aiming to describe the abnormal degree of risk abnormal data.

$$x_{ij} = | x_{ij} - u | \tag{1}$$

Among them, u is the mean value of all the standard normal data.

Then, normalizing the data:

$$x_{ij}^{'} = \frac{\max\left\{x_{1j},\cdots,x_{nj}\right\} - x_{ij}}{\max\left\{x_{1j},\cdots,x_{nj}\right\} - \min\left\{x_{1j},\cdots,x_{nj}\right\}} \tag{2}$$

According to the following formula, the matrix of the proportion of the i-th time in the j-th index is obtained.

$$p_{ij} = x_{ij} \Big/ \sum_{i=1}^{n} x_{ij} \quad , \quad i = 1,2,\cdots,n, j = 1,2,\cdots,m \tag{3}$$

The calculation formula of information entropy is the

$$E_j = -\frac{1}{\ln n}\sum_{i=1}^{n} p_{ij}\ln(p_{ij}), \quad j = 1,2,\cdots,m \tag{4}$$

According to the information entropy redundancy formula and the weight calculation formula

$$g_j = 1 - E_j \tag{5}$$

$$W_j = g_j \Big/ \sum_{j=1}^{m} g_j \quad , \quad j = 1,2,\cdots,m \tag{6}$$

According to the above steps, we can calculate the weight of each index.

2.2.2 Rating is obtained by the TOPSIS Method

After the normalized treatment, the standardized matrix is

$$A = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}_{n\times m} \tag{7}$$

The optimal scheme consists of the maximum value A^+ for each column element in A:

$$A^+ = (\max\{x_{11}, x_{21}, \cdots, x_{n1}\}, \max\{x_{12}, x_{22}, \cdots, x_{n2}\}, \cdots, \max\{x_{1m}, x_{2m}, \cdots, x_{nm}\})$$

The worst scheme consists of the minimum value A^- for each element of the column in A:

$$A^- = (\min\{x_{11}, x_{21}, \cdots, x_{n1}\}, \min\{x_{12}, x_{22}, \cdots, x_{n2}\}, \cdots, \min\{x_{1m}, x_{2m}, \cdots, x_{nm}\})$$

The proximity of each evaluation object to the optimal scheme and the worst scheme:

$$D_i^+ = \sqrt{\sum_{j}^{m} W_j (A_j^+ - x_{ij})^2}, \quad i = 1,2,\cdots,n \tag{8}$$

$$D_i^- = \sqrt{\sum_{j}^{m} W_j (A_j^- - x_{ij})^2}, \quad i = 1,2,\cdots,n \tag{9}$$

Among them, W_j represents the weight of the j-th index.

The relative proximity of the object to the worst scheme:

$$C_i = \frac{D_i^+}{D_i^+ + D_i^-}, \quad i = 1,2,\cdots,n \tag{10}$$

Among them, $C_i \in [0,1]$, $C_i \to 0$ indicates that the better the evaluation object is.

Finally, converting it into a hundred-mark system:

$$s_i = \frac{C_i}{\max\left\{C_1, C_2, \cdots, C_n\right\}} \times 100, \quad i = 1, 2, \cdots, n \tag{11}$$

Through sorting and filtering, the five moments with the highest abnormal degree score are obtained. Moreover, the five abnormal sensors with the largest weight, which is corresponding to the risk abnormal data at each time, are obtained. The results are as follows:

Table 1: Results

	First high score	Second high score	Third high score	Fourth high score	Fifth high score
Abnormal degree score	100.00	99.32	97.35	97.05	96.14
Exception time number	2465	2531	2383	2394	2466
Abnormal sensor number	45	45	45	45	45
Abnormal sensor number	90	90	90	90	90
Abnormal sensor number	44	44	44	44	44
Abnormal sensor number	62	40	62	62	40
Abnormal sensor number	40	27	40	40	27

3. Evaluation and promotion of the model

This paper uses the Euclidean Distance to define the degree of anomaly and determines the weight of each index by the Entropy Weight Method. Furthermore, using the TOPSIS Method to make the score more objective and reliable. However, this paper does not consider the endogeneity between the indicators. When calculating abnormal degree score, it may be imprecise. The model can be applied to the safety assessment of production lines in the industry, it can also be used in real-time monitoring and risk warning of auto parts.

References

[1] Jiang Qiyuan, Xie Jinxing, Ye Jun, Mathematical Model (4th Edition), Beijing: Higher Education Publishing House, 2011.

[2] Eric Mathers, Python Programming: Getting Started to Practice (2nd Edition), Beijing: People's Post and Telecommunications Press, 2020.

[3] Machine Learning Analysis of Ali Liyun Tianchi Competition, Beijing: Electronic Industry Publishing House, 2020.