Research on the Key Feature Selection of Supercharger Based on Relief-F Arithmetic

Liao Guomin¹,.*, Hujing², b, Hu Guichuan³, c*

¹School of Digital Engineering, Chongqing College of Architecture and Technology, Chongqing, China
²School of Intelligent Manufacturing, Chongqing College of Architecture and Technology, Chongqing, China
³School of mechanical and power engineering, Chongqing University of science and technology, Chongqing, China
*email: 1643493318@qq.com, bemail: 278252640@qq.com, cemail: huguichuan@cqust.edu.cn

Abstract: Gas-jet supercharger is the core equipment of underground gas storage, whose working performance directly affects the working state of gas storage. Therefore, Relief-F algorithm was used to select the key features to determine the key factors causing the failure of the supercharger. Collecting the main state categories and related characteristic parameters of the supercharger, and calculating the feature scores for the normalized sample set by Relief-F algorithm. At the same time, ranking the feature scores and removing the feature whose importance score is less than the threshold value. Eleven key features, such as engine exhaust temperature I and engine exhaust temperature bar, were obtained, which provide a basis for the fault identification of supercharger.

1. Introduction

Chongqing and the surrounding northeast Sichuan area are rich in natural gas resources, which is one of the three major natural gas producing areas in China. Hence, the extraction, transportation and storage of natural gas has long been a concern of the industry. Gas storage in deep strata using closed strata structure is an effective way to store a large amount of natural gas, because of the large gas storage capacity, high pressure and low cost. This gas storage mode plays an important role in regulating the contradiction between the supply and demand of natural gas in winter and summer, ensuring the demand of users and optimizing the operation of the gathering and transportation system. Which is the most effective means to give full play to the capacity of long-distance transmission pipeline, solve the uneven gas consumption of users, and meet the requirements of urban peak regulation and emergency gas supply.

Moreover, gas-jet supercharger is the core equipment of underground gas storage, whose working performance directly affects the working state of gas storage. Its failure will cause the delay of the production of the gas storage, and in serious cases, it will cause production accidents, or even explosions, resulting in heavy casualties and serious economic losses.

In order to determine the key factors causing the fault of the supercharger, a small feature subset is obtained by feature selection of the supercharger, which makes the model better and the generalization ability stronger. Relief-F algorithm and Relief algorithm are common feature selection methods, on the
basis of feature selection, they also add feature weight calculation function, which is widely used in big data processing. But Relief algorithm can only be used to deal with two types of data in the training set, which has great limitations. In order to make it able to deal with multiple types of data, the Relief algorithm was extended on the basis of Relief algorithm, and Relief-F algorithm was obtained [1].

Through the feature selection process of Relief-F algorithm, the engine power cylinder exhaust temperature \( I \), engine power cylinder exhaust temperature bar and other key features related to the turbocharger fault are obtained, thus providing a basis for the discrimination of subsequent turbocharger faults.

2. Feature selection

2.1 Selection method

2.1.1 Relief algorithm

Relief algorithm is a famous multi-variable filtering feature selection algorithm proposed by Kira and Rendell in 1992 [2]. The basic idea of the Relief algorithm is to analyze the differences between different types of samples to determine the weight of the feature based on the degree of the differences. In other words, Relief algorithm obtains the correlation between features and categories by judging the ability of features to distinguish close samples, and then completes classification and extraction [3].

Relief algorithm measures the distinguishing ability of features by examining the differences between similar and heterogeneous neighbor samples of features. If the difference between similar samples is small, but the difference between different samples is large, the feature has a strong discriminative ability. The spacing in the Relief algorithm is defined according to the nearest neighbor method, which can solve the classification problem of nonlinear data more effectively. Moreover, the Relief algorithm avoids using any global search and heuristic search methods, so it shows more efficient performance than other encapsulated algorithms. These advantages make it widely used in the classification of large-scale data [4].

2.1.2 Relief-F algorithm

Relief algorithm is suitable for handling feature selection problems with large sample size, large dimension and less classification. But supercharger key feature selection is a multi-category feature selection problem, therefore, Relief-F algorithm was selected to select the key features of the supercharger.

The theory of Relief-F algorithm is as follows: Assume that \( D \) is a training sample set and \( N \) is the number of features, \( \text{diff} (A, R_1, R_2) \) represents the distance between samples \( R_1 \) and \( R_2 \) under feature \( A \).

\[
\text{diff} (A, R_1, R_2) = \begin{cases} 
R_1[A] - R_2[A] & \text{continuous process} \\
|R_1[A] = R_2[A]| & \text{discrete process and } |R_1[A] = R_2[A]| \\
1 & \text{discrete process and } |R_1[A] \neq R_2[A]| 
\end{cases}
\]

So the total distance between samples \( R_1 \) and \( R_2 \) is defined as

\[
d(R_1, R_2) = \sum_{A=1}^{N} \text{diff} (A, R_1, R_2)
\]

The specific steps of implementation of Relief-F algorithm are as follows:

1. Giving the Moreover sample set \( K \) and the feature set \( F \), and initializing \( m \) (the iteration times) and \( k \) (the number of neighborhood samples selected).
2. Sample \( R \) is randomly selected from the sample set, and \( k \) nearest neighbors are selected from samples similar to \( R \) which denoted as \( H_j \), and also selecting the \( k \) nearest neighbors from samples different to \( R \) which denoted as \( M_j \), until the specified number of iterations is reached.
3. Calculating the feature importance score and feature importance ranking.
(4) Retaining the features whose importance scores exceed a given threshold.

2.2 Feature selection process
The feature selection process of supercharger related data based on Relief-F algorithm is as follows:

(1) Category set
According to the SCADA (Data Acquisition and Monitoring Control) system of an underground gas storage supercharger, there are six main states in the category set and shown in Table 1.

| No. | Main States Categories                      | Code |
|-----|--------------------------------------------|------|
| 1   | Intake pipe failure                        | $Z_1$|
| 2   | Valve failure                              | $Z_2$|
| 3   | Piston rod and stuffing box failure        | $Z_3$|
| 4   | Cylinder fault                             | $Z_4$|
| 5   | Frame fault                                | $Z_5$|
| 6   | Normal                                     | $Z_6$|

(2) Characteristic parameter set
By sorting out the operation data of the supercharger, a total of 15 characteristic parameters related to the operation of the supercharger are obtained, and the characteristic parameter set is $T$, which is shown in Table 2.

| No. | Characteristic Parameter                     | Code  | Unit |
|-----|---------------------------------------------|-------|------|
| 1   | Engine speed                                | $T_1$ | rad/min |
| 2   | Engine exhaust temperature I                | $T_2$ | ℃     |
| 3   | Engine exhaust temperature II               | $T_3$ | ℃     |
| 4   | Engine exhaust temperature III              | $T_4$ | ℃     |
| 5   | Engine cooling jacket temperature           | $T_5$ | ℃     |
| 6   | Engine coolant temperature                  | $T_6$ | ℃     |
| 7   | Compressor exhaust temperature I           | $T_7$ | ℃     |
| 8   | Compressor exhaust temperature II          | $T_8$ | ℃     |
| 9   | Compressor cooling jacket temperature      | $T_9$ | ℃     |
| 10  | Compressor inlet pressure I                | $T_{10}$ | MPa |
| 11  | Compressor outlet pressure I               | $T_{11}$ | MPa |
| 12  | Compressor coolant temperature             | $T_{12}$ | ℃     |
| 13  | Compressor outlet pressure II              | $T_{13}$ | MPa |
| 14  | Compressor body vibration                  | $T_{14}$ | mm/s |
| 15  | Maintenance time                            | $T_{15}$ | min |


(3) Sample set
There are 10,000 status data were randomly selected from the SCADA system, including the relevant data of equipment normal and equipment failure, to form a sample set. Which was used as the data source for the selection of key features of the supercharger. Partial sample set is shown in Table 3.

Table 3  Partial sample set

| No. | T₁   | T₂   | T₃   | T₄   | ... | T₁₅  | Z    |
|-----|------|------|------|------|-----|------|------|
| 1   | 284.07 | 238.04 | 235.35 | 47.52 | ... | 38.14 | Z₀   |
| 2   | 317.67 | 265.76 | 264.74 | 53.20 | ... | 42.62 | Z₀   |
| 3   | 288.81 | 240.63 | 238.74 | 47.91 | ... | 38.73 | Z₀   |
| 4   | 311.35 | 258.04 | 260.50 | 51.62 | ... | 41.78 | Z₀   |
| ... | ...   | ...   | ...   | ...   | ... | ...   | ...  |
| 10000 | 299.43 | 250.58 | 250.94 | 49.96 | ... | 40.14 | Z₀   |

(4) Data normalization
Different evaluation indexes often have different dimensions and dimensional units, which will affect the results of data analysis. In order to eliminate the dimensionless influence among the indicators, data normalization is needed to solve the comparability among the data indicators. After the normalization of the original data, each index is in the same order of magnitude, which is suitable for comprehensive comparative evaluation.

Normalization results of sample set is shown in Table 4.

Table 4  Normalization results of sample set

| No. | T₁   | T₂   | T₃   | T₄   | ... | T₁₅  | Z    |
|-----|------|------|------|------|-----|------|------|
| 1   | 0.23 | 0.27 | 0.27 | 0.26 | ... | 0.29 | Z₀   |
| 2   | 0.78 | 0.8  | 0.76 | 0.8  | ... | 0.81 | Z₀   |
| 3   | 0.29 | 0.3  | 0.35 | 0.34 | ... | 0.35 | Z₀   |
| 4   | 0.7  | 0.65 | 0.69 | 0.65 | ... | 0.71 | Z₀   |
| ... | ...  | ...  | ...  | ...  | ... | ...  | ...  |
| 10000 | 0.52 | 0.5  | 0.47 | 0.51 | ... | 0.52 | Z₀   |

3. Results and analysis

3.1 Feature importance score
Relief-F algorithm was used to score and order the features of the normalized sample set. The importance score of each features is shown in Table 5.

Table 5  The importance score of each features

| Features | Score | Features | Score |
|----------|-------|----------|-------|
| T₁       | 1674  | T₉       | 3354  |
| T₂       | 3490  | T₁₀      | 3462  |
| T₃       | 3396  | T₁₁      | 3468  |
| T₄       | 3352  | T₁₂      | 3448  |
| T₅       | 3458  | T₁₃      | 3220  |
| T₆       | 3282  | T₁₄      | 3314  |
| T₇       | 3486  | T₁₅      | 3166  |
| T₈       | 3526  |          |       |
3.2 Select Key Features

The features with importance score less than 3300 were removed, and keep 11 key features, which is shown in Table 6.

| No. | Characteristic Parameter                      | Code | Unit |
|-----|---------------------------------------------|------|------|
| 1   | Engine exhaust temperature I                | $T_2$ | °C   |
| 2   | Engine exhaust temperature II               | $T_3$ | °C   |
| 3   | Engine exhaust temperature III              | $T_4$ | °C   |
| 4   | Engine cooling jacket temperature           | $T_5$ | °C   |
| 5   | Compressor exhaust temperature I            | $T_7$ | °C   |
| 6   | Compressor exhaust temperature II           | $T_8$ | °C   |
| 7   | Compressor cooling jacket temperature       | $T_9$ | °C   |
| 8   | Compressor inlet pressure I                 | $T_{10}$ | MPa |
| 9   | Compressor outlet pressure I                | $T_{11}$ | MPa |
| 10  | Compressor coolant temperature              | $T_{12}$ | °C   |
| 11  | Compressor body vibration                   | $T_{14}$ | mm/s |

4. Conclusion

This paper takes a real large gas storage supercharger as an example to extract the operating characteristic parameters of the supercharger. Relief-F algorithm is used to calculate and sort the feature importance score of the normalized sample set, and remove the features with an importance score less than 3300. Finally, 11 key features were extracted, such as engine exhaust temperature I and engine cooling jacket temperature, and providing theoretical basis for subsequent fault prediction.

Acknowledgments

This paper is from the science and technology research project of Chongqing Education Commission, Research on cable health monitoring system of cable-stayed bridge based on intelligent inspection robot (KJQN201905203).

Reference

[1] Xiao Yan, Wang Qigang, Wang Bin. Object based land-use classification based on hybrid feature selection method of combining Relief F and PSO[J]. Transactions of the Chinese Society of Agricultural Engineering. 2016, 32(4): 211-216.
[2] Kira K, Rendell L A. The feature selection problem: Traditional methods and a new algorithm[C]. AAAI, 1992, 2: 129-134.
[3] Qu Pengcheng. Research and application of data driven fault prediction algorithm for reciprocating compressor[D]. Qingdao University of Science & Technology. 2020, 6.
[4] Huang Xiaojuan. Relief-based Feature Selection Algorithms[D]. Suzhou University. 2018, 6.
[5] TAO Yang, WANG Fan. Principal component selection algorithm based on Relief-F and its application in process monitoring[J]. CIESC Journal. 2017, 68(4): 1525-1532.