ISSN: 2710-0588

Research on Correlation Mining of Compressor Failure Factors based on Apriori Algorithm

Kai Zhang[1,*], Liqiong Chen[1], Sisi Zhang[2], Weihe Huang[1]

[1] College of Pelroleum and Natural Gas Engineering, Southwest Petroleum University, Chengdu 610500, China

[2] School of Law, Southwest Petroleum University, Chengdu 610500, China

[*]Corresponding Author

Abstract

In order to reduce the influence of uncertainty factors during the operation of centrifugal compressors, text data mining method is used to extract text data from 942 fault records of centrifugal compressors of a pipeline. The 942 data were classified by 5M1E analysis, and the data set was scanned several times based on Apriori algorithm, and the support, confidence and boost of the itemsets were calculated to discover all the frequent itemsets, so as to generate association rules. The results show that by setting the thresholds of support and confidence, 22 strong association rules are obtained, and "the degree of aging of compressor components" and "the degree of performance degradation of compressor components" are found to have higher support and more association rules, which provides a good solution for the evaluation and control of the factors affecting compressor failure. This provides an effective reference for evaluating and controlling the factors affecting compressor failure.

Keywords

Centrifugal Compressor; Apriori Algorithm; 5M1E Analysis; Influential Factors.

1. Introduction

In China's petrochemical field, natural gas compressor is one of the most important power protection equipment in the process of natural gas pipeline network management[1], centrifugal compressor as the provision of the main power energy equipment, if the failure occurs in the operation process, it will lead to the enterprise to suffer huge economic losses[2]. This paper adopts the data mining method, based on the Apriori algorithm for a pipeline centrifugal compressor (hereinafter referred to as the compressor) failure data to generate frequent itemsets, through the setting of the minimum support degree and confidence level to screen the key influencing factors and through the item set of the degree of enhancement of the judgement of whether to generate the correlation rules for the management of the compressor to provide an effective reference, and for the follow-up of the compressor failure assessment system to provide certain ideas.

2. Related Works

The research on fault diagnosis and technology of centrifugal compressors has become a common concern and research topic in related fields all over the world. Jianwen Xing et al[3] constructed a Bayesian network based on hazard and operability analysis to diagnose the faults of fuel-driven centrifugal compressor systems. Zhu Yongren et al[4] transferred the compressor operating parameters as input feature parameters of ant colony clustering algorithm for fault diagnosis of centrifugal compressors. Xu Ye et al[5] proposed a method of

ISSN: 2710-0588

centrifugal compressor surge diagnosis based on the fusion of multi-source information such as flow, pressure and vibration.Jun[6] applied the Bayesian network method to the reliability analysis of centrifugal compressor to analyse the compressor failure factors.Spuntrup[7] used the fault tree analysis method to model the reliability of the six main sub-systems, including the compressor body, the control and monitoring and the lubrication system, and determined the reliability of the system. Reliability modelling was carried out to identify the most common faults of the system.Golmoradi[8] obtained features from the compressor pre-processing signals as inputs to the support vector machine and subsequently optimised the support vector machine parameters using genetic algorithm.

Although there have been studies on compressor fault diagnosis, with the increase in the number of compressor production and years of service, the failure rate of compressors has gradually increased, and the problem of insufficient quantitative correlation analysis of the factors affecting compressor failures has also emerged.Apriori [9] is a classical correlation algorithm based on the set of frequent items. It can be used to analyse the influencing factors and explore whether there is an association between the influencing factors.

3. Methods

3.1. Overview of Association Rules

Association rule analysis is a commonly used technique in data mining[10] for uncovering the relationships between itemsets and itemsets in a dataset, i.e., to find out the correlations and dependencies between data items. In association rule mining, the most classical algorithm is Apriori algorithm[11], frequent itemsets can be effectively identified by this algorithm and construct association rules. Therefore, in this paper, Apriori algorithm is adopted for compressor failure influence factor analysis.

3.2. Apriori Algorithm

The Apriori algorithm is generally based on three metrics: support, confidence, and lift. Support is a measure of how often an association rule appears in the dataset. It represents the ratio of the number of items of a particular rule to the total number of items. For example, the support of an item set A is the ratio of the number of items of A to the total number of items:

$$Support(A) = P(A) = \frac{N(A)}{N(ALL)} \tag{1}$$

Confidence is a measure of the probability that if one item set occurs, another item set also occurs. For example, the probability that B occurs at the same time as A occurs is the ratio of the number of items in the itemsets A and B to the number of all items containing A:

$$Confidence(A,B) = \frac{P(A \cup B)}{P(A)} = \frac{N(A,B)}{N(A)} \tag{2}$$

The degree of lift is a measure of how much the probability of an occurrence of one itemset is increased by the occurrence of another itemset. A lift greater than 1 indicates a positive association between the two itemsets, while less than or equal to 1 indicates no association between the two. For example, the probability of an item A appearing together with an item B, but also considering the probability of this item B appearing, is the ratio of the number of items in itemsets A and B to the number of items in A to the number of items in B:

ISSN: 2710-0588

$$Lift(A,B)=\frac{P(A\cup B)}{P(A)\times P(B)}=\frac{N(A,B)}{N(A)N(B)} \tag{3}$$

3.3. Basic Idea

Apriori algorithm is an association rule mining algorithm based on frequent itemsets.The basic idea of Apriori is as follows:

(1) Scan the dataset and calculate the support degree of each item, i.e., how often the item appears in the dataset. According to the set minimum support threshold, the items with support above the threshold are filtered out as candidate item set C1.

(2) Based on the candidate itemset C1, generate the frequent itemset L1, the itemset whose support is higher than the threshold. Generate candidate itemset CK by iterating layer by layer and use the dataset for support calculation and screening until no new frequent itemset can be generated.

(3) Generate association rules based on frequent item sets. For each frequent item set LK, generate all its non-empty subsets as the premise of the rule and the remaining items as the conclusion of the rule. Calculate the confidence level of the rules and filter the association rules whose confidence level is higher than a set threshold.

(4) Steps 2 and 3 are performed iteratively until no new frequent itemsets or association rules can be generated.

(5) Setting the lifting degree threshold to generate strong association rules between itemsets.

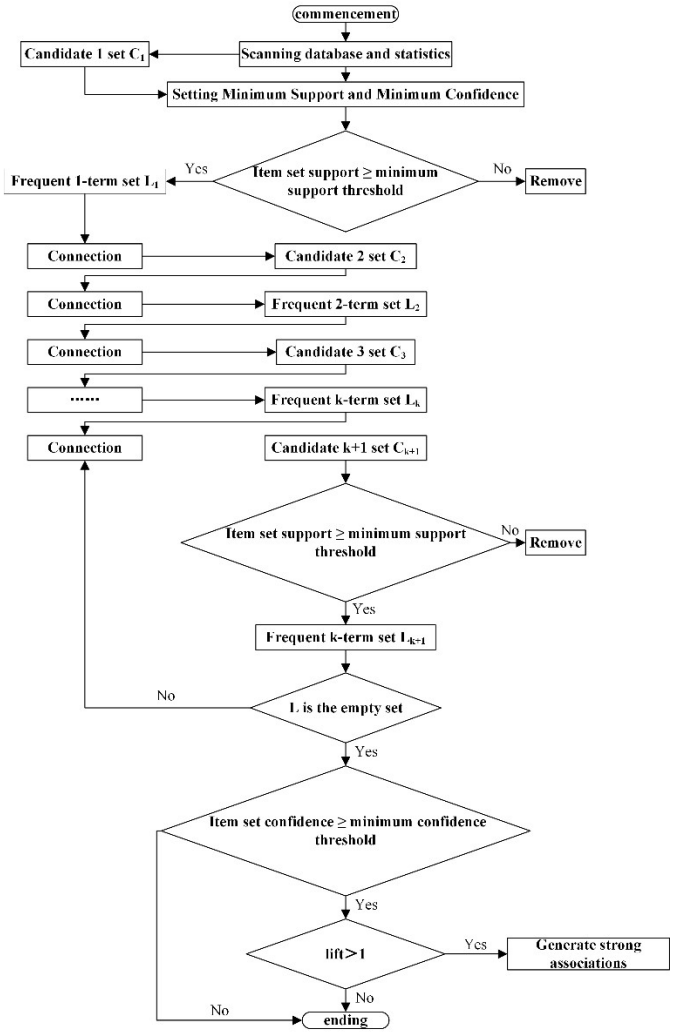

Fig.1 Basic flow chart of Apriori algorithm

ISSN: 2710-0588

4. Results and Discussion

4.1. Text Mining Overview

Text mining[12] technique is a technique for extracting and discovering meaningful information from large amounts of text data. The process of text mining is generally divided into six parts: text collection, text pre-processing, text linguistic processing, data analysis, result visualisation and structural output[13], as shown in Fig. 2.

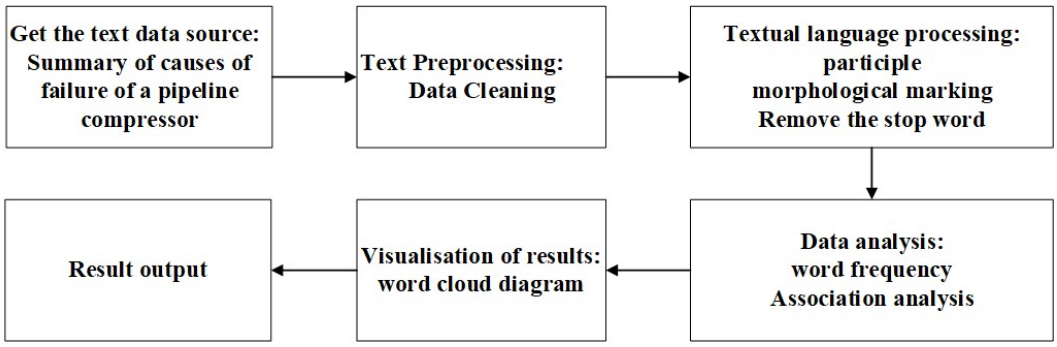

Fig.2 Text mining technology flow

4.2. Data Processing

Texts were collected from the summary table of failure data of a pipeline centrifugal compressor unit, and 942 failure data texts were obtained after deleting unimportant data. Python is used as the programming language of text mining technology, combined with Jieba library to group the text with split words, and Wordcloud library to visualise the results and draw the word cloud as shown in Fig.3. The font size of the word cloud indicates the frequency of occurrence of fault factors, and component damage, voltage fluctuation, signal abnormality, high temperature, phase loss, tripping and External lightning factor appear with high frequency counts, which are the main causal factors.

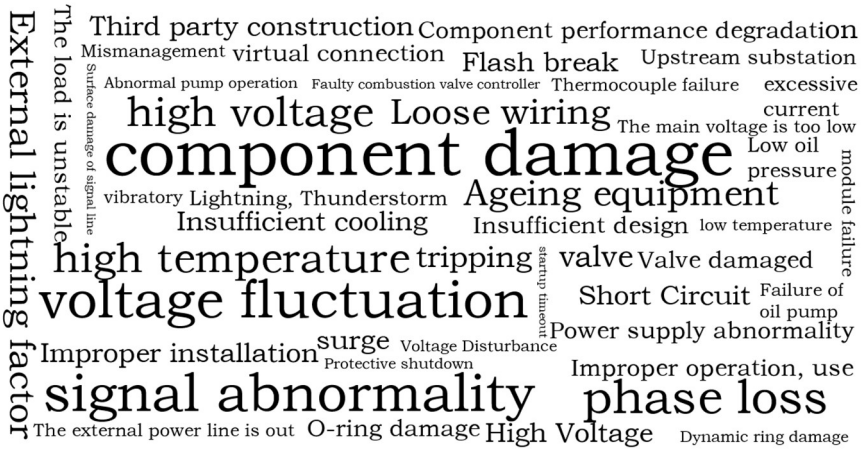

Fig.3 Text mining word cloud graph

The corresponding indicators need to be established after the classification of raw data. This paper adopts 5M1E analysis[14] as the scientific method to determine the classification index, man, machine, material, method, environment and measurement as the main factors affecting compressor failure. Based on the first-level index of "5M1E", combined with the word cloud diagram, the second-level index can be classified and labelled as shown in Table 1.

ISSN: 2710-0588

Table 1. Impact factor numbers

Level 1 indicators	Secondary indicators
A man	A1 Operator skill level; A2 Operator responsibility intensity; A3 Operator proficiency in performing maintenance procedures; A4 Operator training level; A5 Manager responsibility; A6 Supervisor responsibility
B machine	B1 Compressor Selection Reasonableness; B2 Compressor Design Reasonableness; B3 Compressor Component Ageing Degree; B4 Compressor Component Performance Degradation Degree; B4 Inspection Instrument Accuracy; B5 Auxiliary Tool Reasonableness
C material	C1 Degree of connection of compressor nodes; C2 Strength of component materials; C3 Durability of component materials; C4 Robustness of component materials; C5 Quality of lubricants
D method	D1 Degree of standardisation of operation; D2 Degree of reasonableness of maintenance plan; D3 Degree of reasonableness of maintenance method; D4 Degree of accuracy of fault diagnosis method; D5 Degree of standardisation of management; D6 Degree of reasonableness of training method for operators; D7 Degree of reasonableness of skill enhancement method.
E environment	E1 Compressor Operating Environment; E2 Compressor Parking Environment; E3 Severe Weather Environment; E4 Field Maintenance Environment
F measurement	F1 Compressor Signal Transmission Inspection Accuracy; F2 Compressor Component Quality Inspection Accuracy; F3 Compressor Construction Connection Inspection Accuracy; F4 Compressor Production Quality Inspection System

After dividing the indicators, the text-mined phrases were compared with the secondary indicators in Table 1 of the evaluation indicators, and the 942 failure data were compared one by one to form the text analysis data as shown in Table 2.

Table 2. Data related to the algorithm after preprocessing

Data number	Impact factor number
P1	C3, D5
P2	B3, A3
…	…
P562	C2, C3
…	…
P942	B1, B3, C2, C3

4.3. Parameter Setting

In order to make the association rules practical, the minimum support setting needs to be made according to the actual needs and the characteristics of the dataset, which helps to ensure the quality and accuracy of the generated rules and effectively avoid the appearance of duplicate rules. Similarly, the setting value of minimum confidence should balance the rigour of the rules and the number of generated rules to ensure the effectiveness of the mining results.

In this paper, we first set the minimum support to 0.02 and the minimum confidence to 0.4[15] to obtain a larger number of association rules, after which a more objective and regular analysis is carried out by changing the thresholds of minimum support and minimum confidence.

4.4. Analysis of Results

Then, by adjusting and setting the minimum support degree and the minimum confidence degree, the support degree of most influencing factors of the compressor is obtained, and generates a scatter plot as shown in Fig.4.

ISSN: 2710-0588

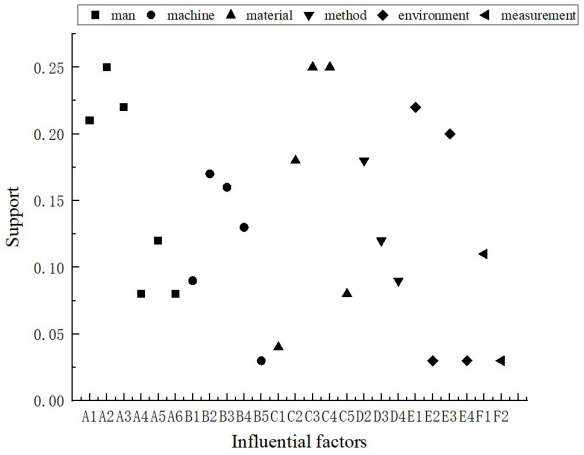

Fig.4 Calculation results of support degree

As can be seen in Fig. 4, only a small number of influencing factors are higher than 0.2, which is due to the selection of a large sample size and the low probability of each influencing factor appearing in the sample.

Further, each of the 6 major factors of 5M1E has a different degree of influence on compressor failure. As can be seen from the figure, the probability of occurrence of "machine", "method" and "measurement" indicator factors is low. What is more special is that the four indicators contained in the "ring" polarisation is more obvious, the higher two indicators are related to the environment of the natural climate and the operating environment, need to focus on it. In addition, the support for the impact indicators contained in "people" and "materials" is mostly distributed in the middle and high ranges, indicating that the main impact factors of the compressor are related to "people" and "materials". This indicates that the main influence factors of the compressor are strongly related to "people" and "materials", and therefore the main influence factors of the compressor should be improved from "people" and "materials" to improve the reliability of the compressor.

According to the results shown in Fig. 4, the minimum support degree is 0.1 and the minimum confidence degree is 0.5, and 22 association rules are obtained, and Table 3 shows the output results of association rules.

Taking the influence factor A1 in Table 3 as an example, the confidence level of the correlation factor between the A1 factor and the A2 factor is 0.63, and the result shows that when the A1 factor causes the compressor failure, there is a 63% probability that the factor A2 occurs at the same time. Further, the degree of enhancement of factor A1 and factor A2 is 2.37, the results show that when factor A1 occurs, the probability of A2 occurring simultaneously with the probability of only factor A2 is 2.37, the results are greater than 1, indicating that A1 and A2 have a strong correlation, and the rest of the correlation factors, A3, has a degree of enhancement of 3.16, B3 has a degree of enhancement of 2.13, and E3 has a degree of enhancement of 1.99, which suggests that the factor A1 causing Compressor failure is not the cause of a single factor, but the result of multi-factors.

The confidence and enhancement output results can be analysed to reduce the potential failure that may be brought by the association factors by analysing the former factors, effectively reducing the failure rate of the compressor. In addition, correlation rules can be used to indirectly control some of the less controllable factors. For example, in Table 3, A1 is the E3 correlation factor with a confidence level of 53% and an enhancement of 1.99. Due to the large degree of variation of the E3 factor in the severe weather environment, it is difficult to accurately assess it, so it is possible to monitor A1 and thus reduce the uncertainty brought by the E3 factor.

ISSN: 2710-0588

It can also be found in the results of Table 3 and Fig. 5 that the B3 and B4 influencing factors have a high confidence level and more association rules, and appear with the highest frequency, which corresponds to the information of component damage, equipment aging, and component performance degradation in the word cloud diagram, and the influencing factors of the association factors of B3 and B4 as the main influencing factors of compressor failures are: the skill level of the operator (A1), the intensity of the operator's sense of responsibility (A2), operator proficiency in performing maintenance procedures (A3), durability of component materials (C3), and degree of rationality of maintenance schedules (D2). The five influencing factors are closely related to compressor aging and component performance degradation, thus creating a correlation with B3 and B4. This shows that preventing compressor aging and maintenance as much as possible can greatly reduce the probability of compressor failure and improve compressor reliability.

Table 3. Association rule output

Influential factors	relevant factors	confidence	lift
A1	A2	0.63	2.37
	A3	0.72	3.16
	B3	0.51	2.13
	E3	0.53	1.99
A2	A1	0.50	2.37
	B3	0.53	1.89
	B4	0.51	2.13
A3	A1	0.67	3.17
	B3	0.51	1.67
	B4	0.52	1.62
	E3	0.52	1.21
B2	B1	0.67	3.75
C2	C3	0.94	3.73
	C4	0.71	3.52
C3	C2	0.71	3.73
	B4	0.63	2.13
	C4	0.93	3.79
D2	B3	0.56	2.12
	C3	0.61	2.63
	C4	0.67	1.21
E1	E3	0.81	4.40
F1	D4	0.53	1.12

5. Conclusion

This paper adopts Apriori algorithm for data mining, extracts the relevant factors for 942 compressor failure records, and obtains the following main conclusions by setting the minimum support degree and confidence degree, and combining the enhancement degree:

(1) In the compressor failure, "human" and "material" have high support, which can reduce the probability of compressor failure by reducing the influence of human factors and using better compressor materials.

(2) for the uncontrollable compressor failure factors, can control with its strong correlation rule factors for indirect control, so as to achieve the uncontrollable factors into indirectly controllable.

(3) The use of strong correlation rules of correlation analysis, found that the compressor aging and component performance degradation has a high degree of confidence and more correlation

ISSN: 2710-0588

rules, which can be the focus of the research object, in order to improve the reliability of the compressor.

References

[1] Zhao Jingyan, Ge Kai, Chu Chenkeng, et al. Current application and prospects of China-made natural gas compressors[J]. Natural gas industry, 2015, 35(10): 151-156.

[2] Yang Lan, Feng Siyang. Current status and development trend of centrifugal compressor fault diagnosis research[J]. China Science and Technology Journal Database Industry A, 2022, (4): 58-60.

[3] Xing Jianwen, Liang Xi, Li Wei. Research on compressor fault location method based on HAZOP analysis and Bayesian network[J]. Automation Instrumentation, 2022, 43 (3): 11-14,19.

[4] Zhu Yongren, Cai Jie, Shan Yingying. A fault diagnosis method for centrifugal compressors based on ant colony clustering algorithm[J]. Oil and Gas Storage and Transportation, 2019, 38(4): 424-428.

[5] Xu Ye, Huang Wenjun, Mi Junpeng, et al. Surge diagnosis method of centrifugal compressor based on multi-source data fusion[J]. CIESC Journal, 2023, 74(7): 2979-2987.

[6] Jun HB, Kim D. A Bayesian network-based approach for fault analysis[J]. Expert Systems with Applications, 2017, 81: 332-348.

[7] Spuntrup F.S, Londono J.G, Skourup C, et al. Reliability improvement of compressors based on asset fleet reliability data[J]. IFAC-Papers Online, 2018, 51(8): 217-224.

[8] Golmoradi M, Ebrahimi E, Javidan M. Compressor fault diagnosis based on SVM and GA[J]. Vibroengineering Procedia, 2017, 12: 49–53.

[9] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules[C]//Proceedings of the 20th International Conference on Very Large Data Bases. Santiago: Morgan Kaufmann, 1994: 487-499.

[10] Liu Wenya, Xu Yongneng. Association Rule Mining of Metro Failures Based on lmproved Apriori Algorithm[J]. Journal of Weapons and Equipment Engineering, 2021, 42(12): 210-215.

[11] Arshiya. Trends on Data Mining Concepts and its Techniques[J]. Test Engineering and Management, 2020,Vol.83.

[12] Tobback, E.;Naudts, H. Belgian economic policy uncertainty index: Improvement through text mining[J]. International Journal of Forecasting, 2018, Vol.34(2): 355-365.

[13] Zheng Binbin, Feng Tingting, Wang Jiahe, et al. Causes and correlation analysis of urban gas accidents based on text mining[J]. Chinese Journal of Safety Science, 2023, 33(7): 190-195.

[14] Li Tangzhenhao, You Xiaoyue. Analysis of quality influencing factors of assembly building based on Apriori[J]. Journal of Tongji University (Natural Science Edition), 2022, 50 (2): 147-152.

[15] Cheng Congcong, Zhao Yicheng, Jiag Linjing, et al. Correlation Analysis on the Hidden Risk Factors of Tailings Pond Based on Text Mining[J]. Mining Research and Development, 2021, 41(11): 26-33.