Received September 24, 2020, accepted September 28, 2020, date of publication October 1, 2020,
date of current version November 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3028146

Investigating the Construction, Training, and Verification Methods of k-Means Clustering Fault Recognition Model for Rotating Machinery

QINGFENG WANG[1], JIAHE LIU[1,2], BINGKUN WEI[1], WENWU CHEN[3], AND SHUJIAN XU[3]

[1]Beijing Key Laboratory of Health Monitoring and Self-Recovery for High-End Machinery Equipment, Beijing University of Chemical Technology, Beijing 100029, China
[2]China Academy of Aerospace Standardization and Product Assurance, Beijing 100071, China
[3]State Key Laboratory of Safety and Control for Chemicals, China Petroleum and Chemical Corporation Research Institute of Safety Engineering, Qingdao 266071, China

Corresponding author: Qingfeng Wang (wangqf2422@buct.edu.cn)

This work was supported in part by the Chongqing Science and Technology Bureau under Grant cstc2018jszx-cyzdX0167, and in part by the China Petroleum and Chemical Corporation Ministry of Science and Technology under Grant 320059 and Grant 319022-1.

ABSTRACT Considerable studies have been carried out in recent years regarding fault diagnosis and prediction for the rotating machinery in industrial plants. However, few works present the use of clustering approaches applied to time series to diagnose machine faults. With the increasing practical requirement of safety, reliability, availability and maintainability of machinery running, predictive maintenance based on the technologies of fault diagnosis and prediction has also received significant attention in recent years. In the present study, under Cyber-physical systems (CPS) condition, k-means clustering analysis based on the fault case big data machine learning is applied to investigate the fault identification of the rotating machinery without external expert support. K-means cluster-based fault identification model, which includes the k-means cluster analysis module, fault mode – fault cluster centroid knowledge base module and fault identification module has been constructed. Moreover, the fault feature extraction and fault eigenvectors screening are studied in detail. The vibration data of surge, rubbing, misalignment and normal status of the centrifugal compressor in industrial plants are utilized to train and verify the effectiveness of the k-means cluster fault recognition model. The obtained result shows that recognition accuracy rates of the surge, rubbing and misalignment faults reach 94%, 100% and 80%, respectively. However, the effectiveness of the cluster analysis of vibration data for five or more operating states should be studied in the future.

INDEX TERMS K-means clustering, fault feature extraction, fault eigenvectors screening, fault cluster centroid, fault identification model.

I. INTRODUCTION

With the rapid development of Cyber-physical systems (CPS), artificial intelligence (AI) and big data, rotating machinery in the modern industry has become more large-scale, high-speed, automotive and intelligent. Primary fault detection and diagnosis of the rotating machinery has become the most important aspect in the system design and maintenance. Considering the increasing requirements of reliability, availability, maintainability and safety of the rotating machinery, the conventional maintenance strategies such as breakdown maintenance (BM), time-based maintenance (TBM), preventive maintenance (PM) and condition-based maintenance (CBM), which highly depend on the external expert are becoming less effective so that they gradually become obsolete [1]. Based on real-time monitoring of vibration, temperature, pressure and other parameters of a machine, fault diagnosis allows for detection and isolation of early developing faults, and thus predictive maintenance (PdM) has been proposed to serve as a countermeasure, which allows the maintenance to be performed only when it is required [2], [3]. PdM is expected to undertake the following five basic tasks, including predicting incipient faults, assessing the health status of the machine, identifying the failure mode of the machine, forecasting the remaining useful life or operation trend and formulating maintenance strategies

The associate editor coordinating the review of this manuscript and approving it for publication was Yan-Jun Liu.

and optimizing maintenance tasks. In diverse engineering applications, the demand for predictive maintenance is growing vigorously. However, achieving the abovementioned five basic tasks is still an enormous challenge.

Predictive maintenance is defined as the application of artificial intelligence technology to achieve the automatic fault diagnosis and identification of machines, rather than relying on a manual analysis through smarter external experts. Machine learning (ML) and data mining [4], [5] are two important branches of artificial intelligence technology. Using the machine learning model based on the ''black box'' principle, it can design the pattern recognition model indirectly by training the mapping relationship of the input and output data [6]. Global k-means clustering is the widely used partitional method, mainly adapted to machine learning and pattern recognition problems [7]. It is generally accepted that supervised, semi-supervised and unsupervised learning belong to the branch of the machine learning [8], [9]. Moreover, there are three methods for implementing predictive maintenance. These methods are called the mechanism model, data-driven model and hybrid model. Among them, the data-driven method does not need to know the failure mechanism of the machine in advance and does not rely on the experience and knowledge of external experts. Accordingly, the data-driven method has become one of the supporting technologies for predictive maintenance.

In recent years, many data-driven and deep learning methods have been effectively used for fault detection and pattern recognition, e.g. deep transfer learning method based convolutional neural network (DTLCNN) for bearing fault diagnosis under different working conditions [10], rolling bearing fault diagnosis model based on convolutional neural network (CNN) and long-short term memory (LSTM) neural network [11], Graph Convolution Broad Network (GCB-net) model by adding regular CNN and preserve more information for searching features in broad space through layer concatenation [12], SEAEN fault diagnosis approach featuring a sparse autoencoder (SAE) combined with an echo state network (ESN) [13], Broad Learning Adaptive Neural Control using for motor learning and generalization [14], adaptive Bayesian Algorithm using the failure dynamic under varying operating conditions [15], evolving Echo State Networks (ESNs) for dealing with fault diagnosis tasks [16], clustering algorithm using wavelet based probability density functions for identifying patterns [17] and clustering method based on density peak with symmetric neighborhood relationship [18].

It should be indicated that the supervised approaches of fault detection and pattern recognition such as the artificial neural networks (ANN) [19], [20], CNN [21], deep learning (DL) [22], [23], clustering [24] and support vector machine (SVM) [25] require a large amount of labeled normal and failure status data of the machine. In addition, it is generally assumed that these labelled failure status data include various failure types of machines [26]. However, obtaining the truth data of various failure types in industrial environments is a very challenging, expensive and time-consuming process [10], [27], [28]. Accordingly, it is a common method to simulate the component fault damage in the laboratory to obtain data of various fault types [29]. However, the failure damage type of the simulated machine parts is quite different from that in the real operating environment. Therefore, there are limitations in using the simulated failure type data for training the machine learning model. For tasks with little similarities, higher dimensions and few labeled data available, an effective algorithm called hierarchical lifelong learning algorithm to improve the lifelong machine learning system with shared representations [30] was proposed, but the algorithm only verified by Land Mine and Animals data set, the time series data set related to the equipment health status has not been verified.

K-means algorithm is an unsupervised machine learning algorithm, which does not require any previous knowledge to determine a set of clusters [31]. In this algorithm, distance is used as an index for measuring the similarity between data objects. In other words, the smaller the distance between data objects, the higher the similarity, and the more likely they are to be in the same class cluster [32]. Cluster centers and the objects assigned to them represent a cluster. For each sample assigned, the cluster center is recalculated based on the existing objects in the cluster. It is worth noting that an important feature of the k-means algorithm is that it tends to minimize the inter-class variance and increases the extra class distance [33]. The application of the k-means algorithm has successfully identified four operating states of machine tools to guide predictive maintenance [34]. The k-medoids clustering is employed to build the assessment model and achieve the degradation indicator for bearings [35]. Wavelet packet transform and K-means algorithm are employed to decompose, reconstruct, normalize and cluster the time series (off- line data) of multistate parameters under normal operation of wind turbine, to improve the generalization capability of long short-term memory (LSTM) prediction model [36]. Based on the collected exhaust fan vibration data, hierarchical clustering, k-means clustering, fuzzy c-means clustering and other clustering methods have been applied to compare the fault identification results, and investigations show that fault identifications obtained from different methods are almost consistent [24].

K-means algorithm has superior characteristics, including simple implementation, reasonable clustering effect and wide application in diverse problems. However, its model training has some technical challenges. Firstly, it is difficult to infer the number of clusters (K). Secondly, the selection of k initial cluster centers is difficult. Moreover, the inappropriate selection easily falls into a local minimum. Thirdly, off-group points and outliers affect the final results. Finally, the algorithm is only applicable to spherical clusters. It is worth noting that a spherical cluster data set means that each pair of points in the set and each point on a straight line segment connecting the two points are also in the set.

Considerable investigations have been performed in the past few years about the incipient fault detection and

diagnosis for machines [37], [38]. Due to the lack of sufficient label data and non-label data in the industrial environment, many fault diagnosis and prediction research results are based on experimental testing and simulation research. Studies have shown that data-driven methods often require run-to-failure data for complex systems and feedback data is not available [39]. Fortunately, the present study obtained raw time-series vibration monitoring data, which includes normal operation and fault status data of a centrifugal compressor of a company from PetroChina, and the raw vibration data generally has the characteristics of non-static, nonlinear and environmental interference. The labeled failure data collected include surge, misalignment, and rubbing, which are all typical failure modes of centrifugal compressors. Due to the lack of previous knowledge to distinguish fault data clustering, fault cluster analysis of centrifugal compressors can be regarded as an unsupervised learning process [40]. The contributions of this paper can be summarized as follows.

1) The construction, training and verification methods of k-means clustering fault recognition model for rotating machinery has been investigated and proposed.

2) An online k-means clustering fault recognition model by using time series raw vibration data of in-service rotating machinery is formulated.

3) Using time-series raw vibration data without noise reduction and dimension reduction signal processing as the input of the k-means clustering fault recognition model, the fault identification conclusion can be automatically given without depending the prior knowledge of external experts.

4) To the best of authors' knowledge, this is the first time that k-means cluster-based fault identification model is being applied for the fault diagnosis (e.g. surge, rubbing, misalignment) of in-service centrifugal compressor.

The present study reviews various methodologies and techniques in fault detection and identification research in industrial plants for predictive maintenance. Moreover, the k-means algorithm, which faces challenges for the engineering application has been described. The remaining sections of this study are organized as follows: In section 2, the k-means clustering algorithm is presented. Section 3 introduces construction, training and verification methods of k-means clustering fault recognition model for rotating machinery. Moreover, in section 4, the process of applying the typical fault vibration data of centrifugal compressors such as surge, rubbing, and misalignment and the vibration data under normal conditions to train and verify the k-means clustering model is comprehensively discussed. Finally, the article is summarized and the conclusion is presented in section 5.

II. K-MEANS CLUSTERING ALGORITHM

K-means algorithm is a clustering algorithm, which is established based on the partition. According to the principle of similarity, it divides data objects with respect to the similarity. More specifically, objects with a high degree of similarity are classified into the same clusters, while data objects with a high degree of heterogeneity are classified into different kinds

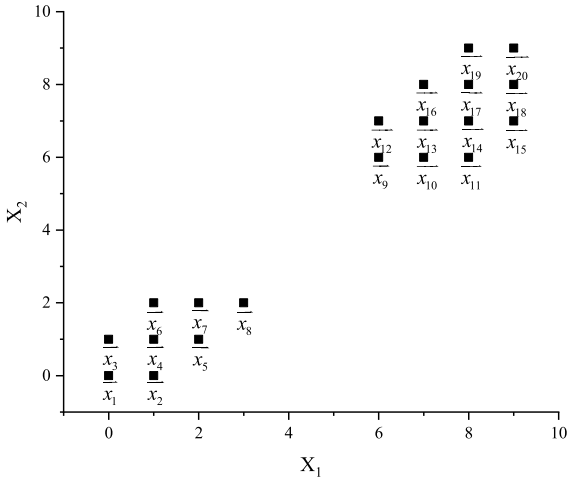

FIGURE 1. Initial cluster center.

of clusters [41]. The k-means clustering algorithm is an iterative clustering algorithm. Its steps are as the following: firstly, the data is pre-divide into K groups. Secondly, K objects are randomly selected as the initial clustering center (seed cluster center). Finally, the distance between each object and each seed cluster center is calculated, and each object is assigned to the cluster center closest to it. The k-means algorithm is a process of repeatedly moving the center point of a class, called centroids, to the average position of its containing members, and then re-dividing its internal members. This process is repeated until a certain termination condition is achieved, i.e. no or minimum number of objects is reassigned to different clusters, no or minimum number of cluster centers change, and the squared error reaches the local minimum.

As shown in Table 1, taking two-dimensional clustering as an example, the cluster analysis steps are as follows:

1) There are 20 samples $(\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_{20}})$ and each of them has 2 features. Moreover, the k-means clustering method is used for the sample classification.

TABLE 1. Sample value.

Sample	$\overrightarrow{x_1}$	$\overrightarrow{x_2}$	$\overrightarrow{x_3}$	$\overrightarrow{x_4}$	$\overrightarrow{x_5}$	$\overrightarrow{x_6}$	$\overrightarrow{x_7}$	$\overrightarrow{x_8}$	$\overrightarrow{x_9}$	$\overrightarrow{x_{10}}$
Feature1	0	1	0	1	2	1	2	3	6	7
Feature2	0	0	1	1	1	2	2	2	6	6
Sample	$\overrightarrow{x_{11}}$	$\overrightarrow{x_{12}}$	$\overrightarrow{x_{13}}$	$\overrightarrow{x_{14}}$	$\overrightarrow{x_{15}}$	$\overrightarrow{x_{16}}$	$\overrightarrow{x_{17}}$	$\overrightarrow{x_{18}}$	$\overrightarrow{x_{19}}$	$\overrightarrow{x_{20}}$
Feature1	8	6	7	8	9	7	8	9	8	9
Feature2	6	7	7	7	7	8	8	8	9	9

2) Suppose K = 2. Fig.1 shows that $\overrightarrow{Z_1}(1) = \overrightarrow{x_1} = (0, 0)^T$ and are selected as the initial clustering center.

3) Calculate the distance from $\overrightarrow{x_1}$ to the two cluster centers. $\left\|\overrightarrow{x_1} - \overrightarrow{Z_1}(1)\right\| = \left\|\begin{pmatrix}0\\0\end{pmatrix} - \begin{pmatrix}0\\0\end{pmatrix}\right\| = 0$, $\left\|\overrightarrow{x_1} - \overrightarrow{Z_2}(1)\right\| = \left\|\begin{pmatrix}0\\0\end{pmatrix} - \begin{pmatrix}1\\0\end{pmatrix}\right\| = 1$. Since

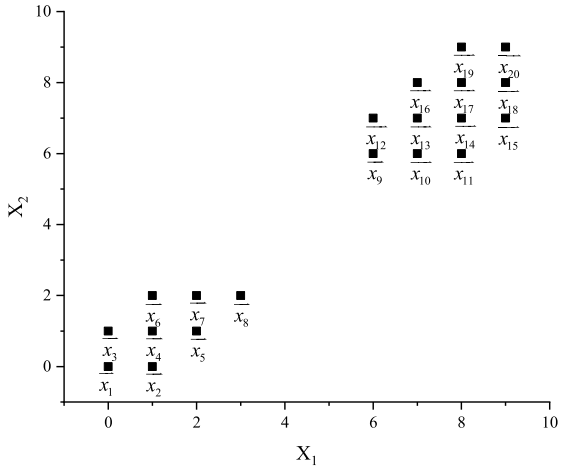

FIGURE 2. The establishment of a new cluster center.

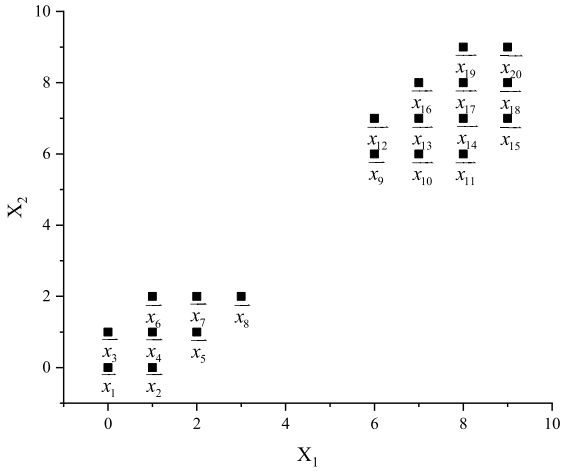

FIGURE 3. Final clusters and cluster centers.

$\left\|\overrightarrow{x_1} - \overrightarrow{Z_1}(1)\right\| < \left\|\overrightarrow{x_1} - \overrightarrow{Z_2}(1)\right\|$, then $\overrightarrow{x_1} \in \overrightarrow{Z_1}(1)$. Calculate the distance from $\overrightarrow{x_2}$ to the two cluster centers. $\left\|\overrightarrow{x_2} - \overrightarrow{Z_1}(1)\right\| = \left\|\begin{pmatrix}1\\0\end{pmatrix} - \begin{pmatrix}0\\0\end{pmatrix}\right\| = 1$, $\left\|\overrightarrow{x_2} - \overrightarrow{Z_2}(1)\right\| = \left\|\begin{pmatrix}1\\0\end{pmatrix} - \begin{pmatrix}1\\0\end{pmatrix}\right\| = 0$. Since $\left\|\overrightarrow{x_2} - \overrightarrow{Z_1}(1)\right\| > \left\|\overrightarrow{x_2} - \overrightarrow{Z_2}(1)\right\|$, then $\overrightarrow{x_2} \in \overrightarrow{Z_2}(1)$. By using the same method to calculate the distance of all samples, these samples can be divided into two new clusters $G_1(1)$ and $G_2(1)$ as follows:

$$G_1(1) = (\overrightarrow{x_1}, \overrightarrow{x_3}) \tag{1}$$

$$G_2(1 = (\overrightarrow{x_2}, \overrightarrow{x_4}, \overrightarrow{x_5} \ldots, \overrightarrow{x_{20}}) \tag{2}$$

4) The clustering center is recalculated based on the existing objects in the clusters $G_1(1)$ and $G_2(1)$. The new clustering center of the cluster can be calculated as follows:

$$\overrightarrow{Z_1}(2) = \frac{1}{N_1} \sum_{\overrightarrow{x_i} \in G_1(1)} \overrightarrow{x_i} = \frac{1}{2}(\overrightarrow{x_1} + \overrightarrow{x_3}) = (0, 0.5)^T \tag{3}$$

$$\overrightarrow{Z_2}(2) = \frac{1}{N_2} \sum_{\overrightarrow{x_i} \in G_2(1)} \overrightarrow{x_i} = \frac{1}{18}(\overrightarrow{x_2} + \overrightarrow{x_4} + \overrightarrow{x_5} + \ldots + \overrightarrow{x_{20}}) = (5.67, 5.33)^T \tag{4}$$

Fig.2 shows that the new clustering centers $Z_1(2)$ and $Z_2(2)$ can be marked separately in the clusters $G_1(1)$ and $G_2(1)$.

5) If the new and old cluster centers are not the same, go to step 3) and recalculate the sample distance. Then, update the cluster center. This process is repeated until no (or minimum number) cluster centers change again, and the squared error reaches local minimum. Fig.3 shows the final cluster and cluster centers.

III. CONSTRUCTION OF K-MEANS CLUSTER BASED FAULT IDENTIFICATION MODEL

The main purpose of this paper is proposing an online k-means clustering fault recognition model by using time series raw vibration data of in-service rotating machinery.

Fig.4 shows that the constructed k-means clustering fault recognition model is designed with three working mode: offline or online training, online fault recognition, and online fault mode-fault cluster centroid knowledge base enrichment.

The detailed principle of k-means cluster-based fault identification model is introduced in Part A to C of section III.

A. K-MEANS CLUSTER ANALYSIS METHOD

1) FAULT FEATURE EXTRACTION

The vibration signal contains information about the state of the machine and is often used for machine health evaluation. In order to characterize the performance degradation process of rotating machinery, root mean square (RMS), kurtosis factor, margin factor, crest factor and skew factor are extracted from the monitoring signals in the time domain.

The RMS value describes the strength and energy parameters of the vibration signal, and it can be mathematically expressed as follows:

$$X_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{5}$$

where x_i^2, N and X_{rms} represent the magnitude of the i-th point, the number of data points and the RMS value, respectively.

The kurtosis factor reflects the degree of deviation of the vibration signal from the normal distribution. More specifically, it is suitable for the diagnosis of surface damage faults and especially early faults. Moreover, it can reflect the shock characteristics of the vibration signal. The value of the kurtosis factor can be described as follows:

$$K_f = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^4}{X_{rms}^4} \tag{6}$$

where $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

The margin factor can effectively reflect the wear of the parts. The value of the margin factor can be represented as

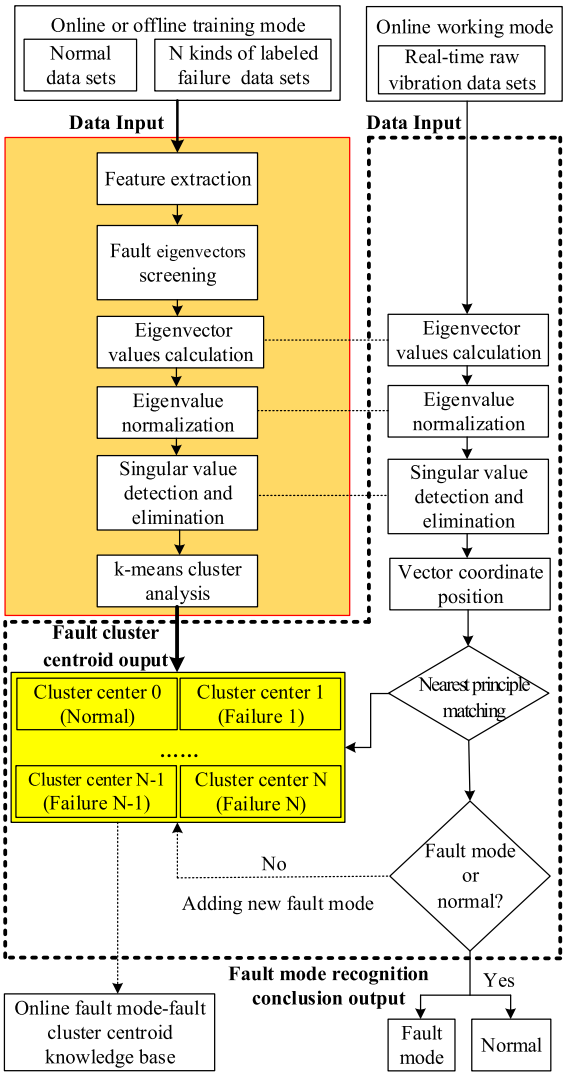

FIGURE 4. Framework schematic of k-means cluster-based fault identification model. The orange area in the solid red frame is the k-means cluster analysis module, which can be used for offline or online model training; the input data of this module is vibration monitoring normal state data and various labeled fault data, and the output is fault cluster centroid. The yellow area in the dashed frame represents the online fault mode – fault cluster centroid knowledge base enrichment working mode. The dashed frame is the fault identification module, which is used for online real-time fault mode recognition; the input data of the fault identification module is the raw data of real-time vibration monitoring, and the output is the fault mode recognition conclusion. The dashed frame is also called fault identification "black box".

follows:

$$L_f = \frac{\max\{x_i\}}{\frac{1}{N}\sum_{i=1}^{N}\sqrt{|x_i - \bar{x}|}} \tag{7}$$

The crest factor reflects the extension of the peak change, and excessive crest factors usually indicate local defects. The value of the crest factor can be calculated as follows:

$$C_f = \frac{\max\{x_i\}}{x_{rms}} \tag{8}$$

The skew factor reflects the asymmetry of the vibration signal and is more sensitive to wear faults. The value of the skew factor can be mathematically expressed as follows:

$$C_w = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{X_{rms}} \tag{9}$$

2) FAULT EIGENVECTORS SCREENING RULES

When the machine operates in the performance degradation state, the damage point of the component repeatedly hits the surface of other components in contact with it during its operation. Meanwhile, the vibration shock is generated, and the vibration time-domain signal can be directly collected. In this study, the time-domain fault characteristic parameters are directly selected as the mechanical fault characteristic parameters, and the fault identification method is investigated based on the cluster analysis. The time-domain characteristic parameters have different value ranges under different fault conditions. Moreover, it should be indicated that the sensitivity of time-domain characteristic parameters such as RMS value, kurtosis factor, margin factor, and peak factor to different faults is different.

The present study uses excellent, good, medium and poor values to measure the sensitivity of each characteristic parameter to different fault conditions. Moreover, it makes a semi-qualitative analysis of the difference between normal data and fault data characteristic parameters. When most of the fault data feature parameters and normal data feature ranges overlap, and the coincidence rate is greater than 80%, the sensitivity of this feature parameter is defined to be poor. on the other hand, when the characteristic parameters of the fault data and the characteristic parameters of the normal data overlap in a small part and the coincidence rate is less than 20%, the sensitivity of the characteristic parameter is defined as the medium. Furthermore, when the fault data feature parameter ranges do not coincide with the normal data feature parameter ranges, but most of the fault feature parameter value ranges overlap, and the coincidence rate is greater than 80%, the sensitivity of this feature parameter is defined as good. when the fault data feature parameter range does not overlap with the normal data feature parameter range, but the sensitivity value range between different fault feature values overlaps with a small part, and the coincidence rate is less than 20%, the sensitivity of this feature parameter is defined as excellent. Based on the life cycle "run to failure" data of the machinery, the above-mentioned characteristic parameter sensitivity discrimination rule is obtained by the statistical learning method, which can be used for the k-means cluster analysis.

3) FAULT EIGENVECTORS VALUE CALCULATION

For visualization point of view, the clustering of different failure modes in the 3-dimensional feature space will be shown obvious inter-class separation and intra-class aggregation without obvious aliasing. So, three fault-sensitive feature parameters are selected to form a three-dimensional fault eigenvector in this paper. The normal vibration data and N

labeled fault data of the machine are selected as training samples. Then, the fault eigenvector value of each vibration data is calculated.

4) NORMALIZATION OF THE FAULT EIGENVECTORS VALUE

In a multi-index evaluation system, due to the different nature of evaluation indices, they usually have different dimensions and orders of magnitude. When the level of each index differs greatly and the original index value is directly applied for the analysis, it will highlight the role of the index with higher value in the comprehensive analysis. Meanwhile, it relatively weakens the role of the index with lower value level. Therefore, in order to ensure the reliability of the results, it is necessary to standardize the original index data with the range from zero to one. From an empirical point of view, normalization is performed to make the features between different dimensions numerically comparable, which can greatly improve the accuracy of the classifier.

The input data are obtained from the eigenvector value of normal and N labeled failure training data. In order to prevent the phenomenon that the large characteristic index is prominent and the small characteristic index is excluded in the cluster analysis, the dimensionless processing of the fault eigenvectors value must be performed. The present study utilizes the extreme value normalization method to compress the data between [0,1] and form a normalized eigenvector matrix. The specific method is mathematically expressed as follows:

$$e' = \frac{e - e_{\min}}{e_{\max} - e_{\min}} \tag{10}$$

where $e'$and e represent the normalized value and the current fault eigenvectors value, respectively. Moreover, $e_{\max}$and$e_{\min}$denote the maximum and minimum values of fault eigenvectors value, respectively.

5) SINGULAR VALUE DETECTION AND ELIMINATION

When performing the cluster analysis based on fault eigenvector values, the singular value has a great impact on the cluster analysis results. In this study, a method of self-learning early warning control limit based on Beta distribution to detect and eliminate singular value has been proposed.

A Beta distribution is a type of multi-parameter statistical distribution. It is the most basic bounded distribution and can be approximated to any form of distribution by adjusting its parameters. Its density function is defined as:

$$f(e; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} e^{\alpha-1} (1 - e)^{\beta-1} \tag{11}$$

The random variable$e$obeys the Beta distribution with parameters αandβ. $e$is usually described as:

$$e \sim Be(\alpha, \beta) \tag{12}$$

The shape parameters α and β are important parameters that determine the nature of the Beta distribution. The self-learning control limit is established by defining the four types of model training data sets and estimating the shape parameters based on prior knowledge or expertise and then calculating the control limit. During actual condition monitoring activities, different control limits can be automatically learned based on the model training data such as normal data, rubbing data, surge data and misalignment data, which makes the singular value detection more flexible. The process of self-learning can be summarized as follows:

- Normalize the model training data sets;
- Use the maximum likelihood estimation to calculate the beta distribution shape parameters of the statistical data under the four types of model training situation;
- Determine the normalized control limit by determining the two-sided quantile corresponding threshold value;
- Obtain the self-learning control limit;
- Singular value detection and elimination by anti-normalization.

6) K-MEANS CLUSTERING PROCESS

After the data sample extracts the fault feature, fault eigenvectors should be screened out and fault eigenvectors value should be calculated. Moreover, the fault eigenvectors value should be normalized and fault eigenvectors singular value should be eliminated. This algorithm requires a specified number of clusters to separate the data.

- Determine the Number of Clusters

K-means randomly selects the cluster centroids and then assigns the sample fault feature vectors to the closest cluster. Through multiple iterations, the cluster centroids are moved to the average position of all feature vectors. Since the exact number of clusters is not known in advance, the optimal number of clusters and clustering effect is required to determine the k value in k-means. In the present study, the below method or silhouette coefficient is applied to estimate the number of clusters [34].

- Determine the fault cluster centroid

The main purpose of the k-means clustering is to assign each fault eigenvector to a specific cluster [42]. To this end, centroids of k initial fault eigenvector points are randomly determined. Secondly, each fault eigenvector point in the data set is assigned to a cluster, while the nearest centroid for each point is searched to assign it to the cluster corresponding to the centroid. Finally, the centroid of each cluster is updated to the average of all fault eigenvector points in the cluster.

B. FAULT MODE-FAULT CLUSTER CENTROID KNOWLEDGE BASE CONSTRUCTION

If there is 1 set of labeled normal sample data and N sets of labeled fault sample data, k-means cluster analysis is carried out according to the abovementioned steps. Then, the corresponding N + 1cluster centroids are obtained accordingly. The N-type fault cluster centroids correspond to N kinds of machine faults, and the normal cluster centroids correspond to the normal operating state of the machine. Moreover, the N

+ 1 cluster centroids correspond to the N + 1 machine operating state. A knowledge base about the fault mode corresponding to cluster centroids is constructed based on k-means clustering analysis.

C. FAULT IDENTIFICATION PROCESS BASED ON K-MEANS CLUSTERING FAULT IDENTIFICATION MODEL

Rotating machinery can be considered as a complex system composed of multiple parts, and its performance degradation process is often determined by the damage of key and vulnerable parts. Due to the vibration coupling mechanism and mutual influence of multiple parts, the failure mode recognition of the rotating machinery has certain difficulty and uncertainty. Based on the historical condition monitoring data of the rotating machinery, the fault data that has occurred is labeled, the fault eigenvectors are extracted, and the centroids of the fault eigenvector clusters of different faults on the same space are calculated. If the fault eigenvector cluster centroid of the real-time condition monitoring data is calculated on the same space, the distance between all the centroids of the historically labeled fault eigenvector clusters and the current data fault eigenvector cluster centroid can be used to determine the fault type [43].

Fig.4 shows that based on the incoming real-time monitoring data sets, the steps of the k-means cluster analysis to identify the fault mode can be described as follows:

1) Calculate the fault eigenvectors value of each data point of each monitoring data set to form the fault eigenvectors data sets;

2) Normalize the fault eigenvectors values of the monitoring data and the training data together;

3) Perform singular value detection on the fault eigenvectors value, and remove the singular value;

4) The normalized fault eigenvectors value corresponds to the fault eigenvectors point in the three-dimensional space coordinate system;

5) According to the nearest neighbor matching rule, the fault eigenvectors coordinates match the nearest cluster centroid points in the cluster knowledge base:

- If 80% of the fault eigenvectors coordinates of the fault eigenvectors data sets can match the same cluster centroid in the cluster knowledge base, the current fault mode or normal health state of the machine can be determined.
- If 80% of the fault eigenvectors coordinates of the fault eigenvectors data sets cannot match with any cluster centroid in the cluster knowledge base, then the monitoring data set corresponding to the feature vector set should analyze the fault mechanism to determine whether the machine belongs to a new fault mode or not. If it belongs to a new fault mode, it is necessary to add a new fault cluster centroid through the cluster analysis algorithm and update the fault cluster knowledge base.

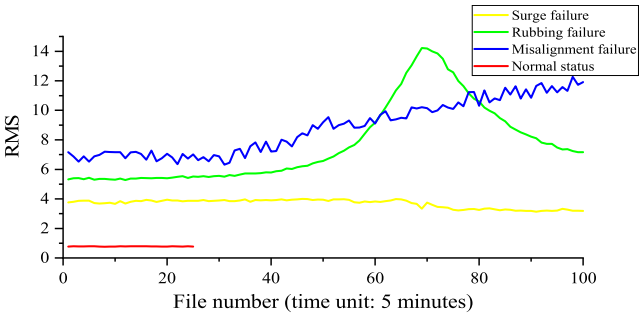

FIGURE 5. RMS curves of surge, rubbing, misalignment and normal status.

IV. K-MEANS CLUSTERING FAULT RECOGNITION MODEL TRAINING AND ENGINEERING APPLICATION VERIFICATION

Obtaining high-quality condition monitoring data of the entire performance degradation process of the rotating machinery from normal operation to the fault shutdown is of significant importance to establish a valuable k-means cluster fault recognition model. The present study obtained labeled data in the normal state and failure status of centrifugal compressors, and the failure mode of centrifugal compressor includes surge, rotor rubbing and shaft misalignment.

TABLE 2. Illustration of centrifugal compressor health status labeled data collected in industrial environment.

Compressor	RPM	Sensor location	Sensor type	Sampling points
C-301	6408	3H	Displacement	1024
120-K-20012	8500	2V	Displacement	1024
M-101	6000	1H	Displacement	1024
210-K-1650	6050	5H	Displacement	1024
210-K-1650	6050	6V	Displacement	1024
120-K-20012	8500	2V	Displacement	1024
M-101	6000	1V	Displacement	1024

Compressor	FS	Data set	Health status	Data usage
C-301	5120	312 sets	Surge	Training
120-K-20012	5120	155 sets	Rubbing (I)	Training
M-101	5120	335 sets	misalignment	Training
210-K-1650	5120	25 sets	Normal	Training
210-K-1650	5120	144 sets	Surge	Verification
120-K-20012	5120	109 sets	Rubbing (II)	Verification
M-101	5120	335 sets	Misalignment	Verification

A. DATASET DESCRIPTION

As described in Table 2, labeled data from four types of centrifugal has been gathered, the raw data belongs to vibration data which comes from displacement sensors mounted on the bearing block, each set of data contains 1024 points, and the vibration data sampling frequency and sampling points are 5120 and 1024 respectively. FS in the Table 2 stands for sampling frequency.

1) TRAINING SAMPLE DATA DESCRIPTION

In cooperation with the condition monitoring data center of a company in PetroChina, 100 sets of C-301 centrifugal compressor surge failure data, 100 sets of 120-K-20012

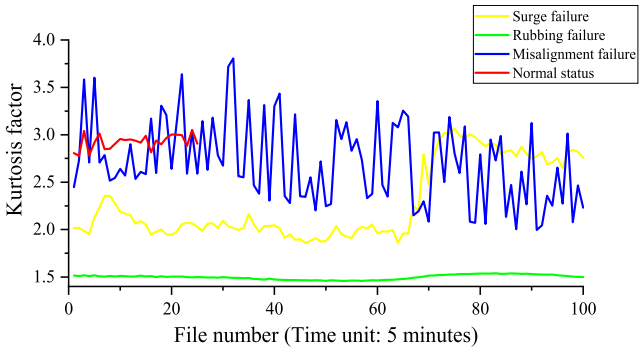

FIGURE 6. Kurtosis factor curves of surge, rubbing, misalignment and normal status.

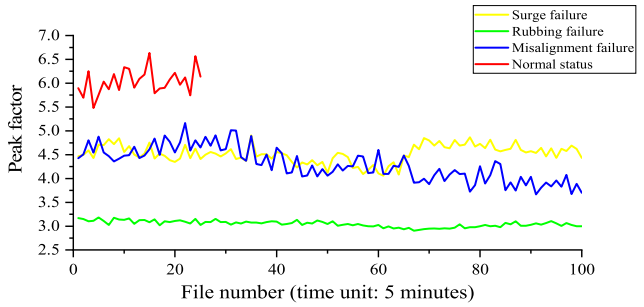

FIGURE 7. Peak factor curves of surge, rubbing, misalignment and normal status.

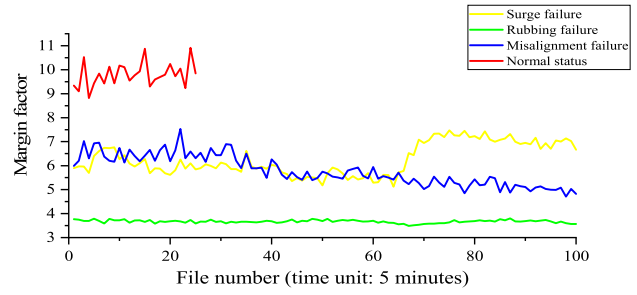

FIGURE 8. Margin factor curves of surge, rubbing, misalignment and normal status.

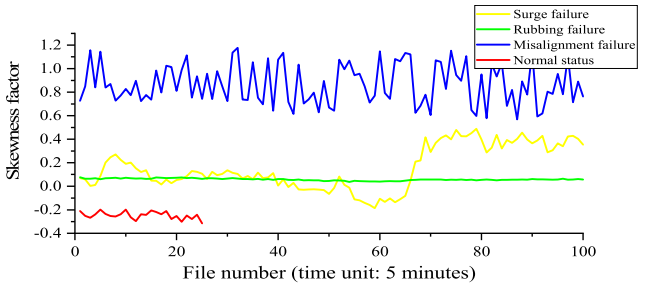

FIGURE 9. Skewness factor curves of surge, rubbing, misalignment and normal status.

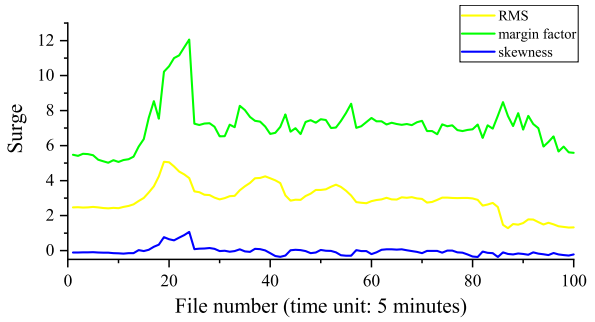

FIGURE 10. The RMS, margin factor and skewness factor curves of 100 sets of 210-K-1650 centrifugal compressor surge failure data.

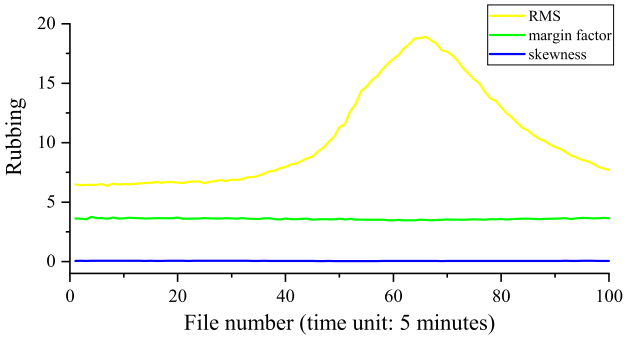

FIGURE 11. The RMS, margin factor and skewness factor curves of 100 sets of 120-K-20012 centrifugal compressor rubbing fault data.

centrifugal compressor impact failure data, 100 sets of M-1010 centrifugal compressor misalignment failure data and 25 sets M-1010 centrifugal compressor normal state data are taken. Then, these data are mixed. The mixed data is used as the training sample data of the k-means clustering fault recognition model.

The characteristic parameter curves (including RMS, kurtosis factor, crest factor, margin factor and skewness factor) of training sample data (100 sets of three kinds of fault data and 25 sets of normal data) can be seen in Fig.5∼Fig.9.

2) THE MODEL VERIFICATION DATA DESCRIPTION

The characteristic parameter curves (including RMS, margin factor and skewness factor) of 100 sets of 210-K-1650 centrifugal compressor surge failure data can be seen in Fig.10.

The characteristic parameter curves (including RMS, margin factor and skewness factor) of 100 sets of 120-K-20012 centrifugal compressor rubbing failure data can be seen in Fig.11.

The characteristic parameter curves (including RMS, margin factor and skewness factor) of 100 sets of M-101 centrifugal compressor misalignment failure data can be seen in Fig.12.

B. TRAINING OF K-MEANS CLUSTER FAULT RECOGNITION MODEL FOR THE CENTRIFUGAL COMPRESSOR

Time-domain characteristic parameters that are more sensitive to component damage and wear, such as RMS, kurtosis factor, peak factor, skewness factor and margin factor are selected as centrifugal compressor fault sensitive character-

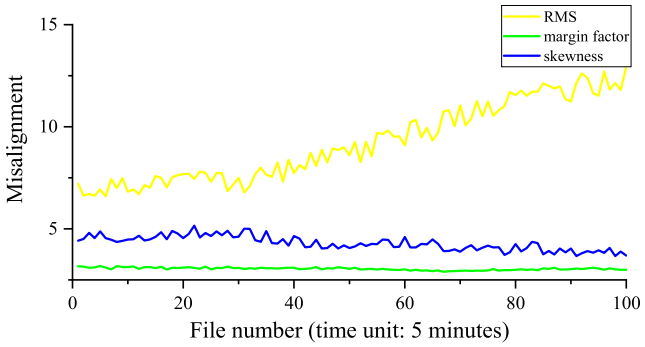

FIGURE 12. The RMS, margin factor and skewness factor curves of 100 sets of M-101 centrifugal compressor misalignment fault data.

istic parameters. Based on the fault case and normal state data gathered, the fault sensitive characteristic parameter values of the centrifugal compressor are calculated. Moreover, Table 3 shows that the range of values of each sensitive characteristic parameter is obtained by using statistical analysis methods.

TABLE 3. Value range of different characteristic parameters under each working condition.

Fault characteristic parameters	Normal status		Surge failure		Rubbing failure		Misalignment failure	
	Min	Max	Min	Max	Min	Max	Min	Max
RMS	0.76	0.80	4.01	5.28	5.28	14.2	6.33	12.3
Kurtosis factor	2.78	3.05	1.46	3.06	1.46	1.54	1.99	3.80
Peak factor	5.48	6.63	2.91	4.91	2.91	3.18	3.66	5.16
Skewness factor	-0.32	-0.20	0.04	0.49	0.04	0.08	0.57	1.18
Margin factor	8.82	10.92	3.48	7.74	3.48	3.80	4.71	7.50

According to the fault eigenvectors screening algorithm, the fault sensitive feature parameters are screened. Table 4 shows the screening results.

TABLE 4. Sensitivity of fault characteristic parameters to various types of faults.

Fault characteristic parameters	Surge failure	Rubbing failure	Misalignment failure
RMS	Excellent	Good	Excellent
Kurtosis factor	Poor	Excellent	Poor
Peak factor	Good	Excellent	Good
Skewness factor	Good	Excellent	Excellent
Margin factor	Good	Excellent	Good

Table 4 shows that the RMS and the skewness factor have a higher sensitivity to distinguish normal status and fault status. Moreover, it is observed that the coincidence rate between faults is very low. The kurtosis factor has poor sensitivity. The peak factor coincidence rate of surge fault and axial misalignment fault is 100%. Moreover, the margin factor coincidence rate of surge fault and axial misalignment fault is 85%. Therefore, the margin factor with a low coincidence rate is selected. In the present study, root mean square, skewness factors and margin factors are selected to form three-dimensional fault eigenvectors points. Moreover, k-means clustering analysis is carried out based on these points. In the present study, root mean square, skewness factors and margin factors are selected to form three-dimensional fault eigenvectors points.

TABLE 5. Clustering analysis results of labeled training sample data using the fault eigenvectors combination of RMS, skewness factor and margin factor.

Health status	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Surge	100groups	0	0	0
Rub	0	100 groups	0	0
Misalignment	0	0	100 groups	0
Normal	0	0	0	25 groups

As explained in section "*TRAINING SAMPLE DATA DESCRIPTION*", there are 4 kinds of machine running states in the training sample, and the k-means cluster analysis k value is 4. According to the aforementioned k-means cluster analysis method, the training sample data can be divided into 4 clusters, and the centroid coordinates of these 4 clusters can be obtained.

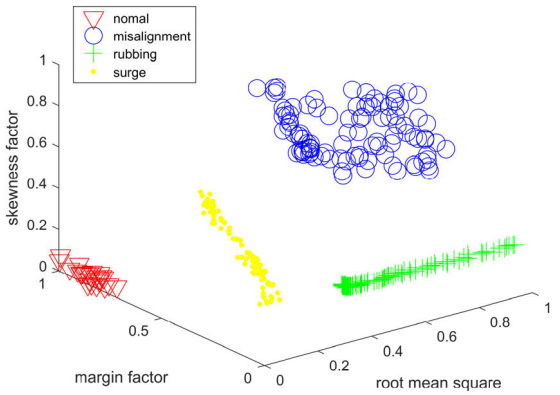

FIGURE 13. Clustering analysis graph of training samples for normal, surge, rubbing and misalignment.

Table 5 shows that 100 groups of surge fault signal samples are divided into cluster 1. 100 groups of rubbing fault signal samples are divided into cluster 2, 100 groups of misalignment fault signal samples are divided into cluster 3 and 25 groups of normal signals are divided into cluster 4. The dot "•" represents cluster 1, which is classified as a surge fault, and the clustering centroid coordinate is A (0.23, 0.34, 0.28). "+" represents cluster 2, which is classified as a rubbing fault, and the clustering centroid coordinate is B (0.44, 0.03, 0.25). "◯" represents cluster 3, which is classified as the misalignment fault, the clustering centroid coordinate is C (0.60, 0.28, 0.77). "∇" represents cluster 4, which is classified as the normal signal, the clustering centroid is D (0.002, 0.84, 0.05). Fig.13 illustrates that by using the k-means clustering analysis fault recognition model, the

normal, surge, rubbing and misalignment of the four types of machine operating status can be completely distinguished.

If the RMS, skewness factor and margin factor are not selected, the kurtosis factor, peak factor and margin factor are selected as the input parameters of the k-means cluster fault recognition model. Table 6 presents the results of the cluster analysis. Among the 100 sets of surge failure sample data, 68 groups are divided into cluster 1, 32 groups are divided into cluster 3, 100 groups of the misalignment failure sample data have 41 groups divided into clusters in 3 and 59 groups are divided into cluster 1.

TABLE 6. Clustering analysis results of labeled training sample data using the fault eigenvectors combination of kurtosis factor, peak factor and margin factor.

Health status	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Surge	68 groups	0	32 groups	0
Rub	0	100 groups	0	0
Misalignment	59 groups	0	41 groups	0
Normal	0	0	0	25 groups

Fig.14 shows the results of the cluster analysis. When the kurtosis factor, the peak factor and the margin factor are selected as the input parameters of the k-means cluster fault recognition model for the cluster analysis, surge faults and misalignment faults cannot be distinguished. Meanwhile, the fault identification effect is poor.

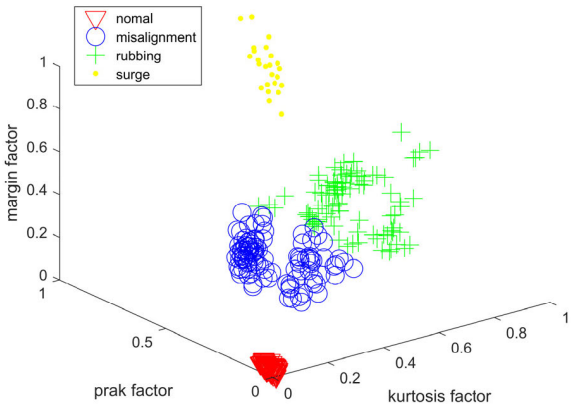

FIGURE 14. Clustering analysis graph of training samples for the kurtosis factor, peak factor and margin factor.

C. ENGINEERING APPLICATION VERIFICATION OF K-MEANS CLUSTER FAULT RECOGNITION MODEL FOR THE CENTRIFUGAL COMPRESSOR

100 sets of 210-K-1650 centrifugal compressor surge failure data are used as the model verification data and each set of data has 1024 points. Moreover, the RMS, margin factor and skewness factor are selected as the fault sensitive feature parameters. According to the aforementioned fault recognition process based on the k-means clustering fault recognition model, the fault recognition process can be described as follows:

- Calculate the value of the fault eigenvectors for each point of each group of data to form a fault eigenvector sets;
- Normalize the fault eigenvectors value after the singular value is eliminated;
- Perform the singular value detection on the fault eigenvectors value and remove the existing singular value;
- Plot the fault eigenvectors points, including RMS, margin factor and skewness factor in the three-dimensional space coordinate system corresponding to the normalized fault eigenvectors value;
- Calculate the centroid coordinates of the corresponding fault eigenvectors point of each group of data;
- According to the principle of the nearest neighbor matching, the clustering attribute closest to the four cluster centroids in the clustering knowledge base is used as the health status of the machine corresponding to the fault eigenvectors array, thereby achieving the fault recognition.

Table 7 shows the verification results. Among the 100 sets of the verification data, 94 sets of centroids are close to the cluster A centroid, 4 sets of centroids are close to the cluster C centroid, and 2 sets of centroids are close to the cluster D centroid. It is concluded that the machine health status corresponding to the verification data can be judged as a surge failure with an accuracy rate of 94%.

TABLE 7. The distance between the centroid corresponding to surge fault datasets and the four-cluster centroid.

Verification data group	Distance from cluster centroid A	Distance from cluster centroid B	Distance from cluster centroid C	Distance from cluster centroid D
Data group 1	0.19	0.38	0.77	0.63
Data group 2	0.19	0.38	0.77	0.64
Data group 3	0.18	0.38	0.76	0.63
…	…	…	…	…
Data group 100	0.28	0.48	0.88	0.60

Another case of 120-K-20012 centrifugal compressor rubbing fault is selected for the engineering verification. Select 100 sets of data as the model verification data where each set of data obtains 1024 points. Moreover, RMS, margin factor and skewness factor are select as fault sensitive characteristic parameters. According to the aforementioned fault recognition method based on the k-means cluster fault recognition model, the distance between the centroid of the fault-sensitive feature vector group corresponding to each group of data and the four cluster centroids of the cluster knowledge base is calculated. Table 8 shows that the centroid of the fault-sensitive feature vector group corresponding to 100 sets of the verification data is closest to the cluster centroid B. Therefore, it is concluded that the machine health status corresponding to the verification data can be judged as a rubbing failure with an accuracy rate of 100%.

Another case of M-101 centrifugal compressor misalignment fault is selected for the engineering verification. 100 sets

TABLE 8. The distance between the centroid corresponding to rubbing fault datasets and the four-cluster centroid.

Verification data group	Distance from cluster centroid A	Distance from cluster centroid B	Distance from cluster centroid C	Distance from cluster centroid D
Data group 1	0.33	0.13	0.65	0.90
Data group 2	0.33	0.19	0.65	0.90
Data group 3	0.34	0.13	0.65	0.91
…		…	…	…
Data group 100	0.35	0.06	0.62	0.92

of data are selected as the model verification data where each set of data obtains 1024 points. Moreover, RMS, margin factor and skewness factor are selected as fault sensitive characteristic parameters. According to the aforementioned fault recognition method based on the k-means cluster fault recognition model, the distance between the centroid of the fault-sensitive feature vector group corresponding to each group of data and the four cluster centroids of the cluster knowledge base is calculated. Table 9 shows that among 100 sets of the verification data, 80 sets of centroids are close to the cluster C centroid and 20 sets of centroids are close to the cluster A centroid. Therefore, it is concluded that the machine health status corresponding to the verification data can be judged as a misalignment failure with an accuracy rate of 80%.

TABLE 9. The distance between the centroid corresponding to misalignment fault datasets and the four-cluster centroid.

Verification data group	Distance from cluster centroid A	Distance from cluster centroid B	Distance from cluster centroid C	Distance from cluster centroid D
Data group 1	0.46	0.57	0.22	0.87
Data group 2	0.53	0.70	0.31	0.85
Data group 3	0.44	0.62	0.35	0.77
…	…	…	…	…
Data group 100	0.69	0.54	0.49	1.12

D. FAULT RECOGNITION ACCURACY RATE COMPARISION OF K-MEANS CLUSTERING MODEL AND SUPPORT VECTOR MACHINE MODEL

Taking the RMS, margin factor and skewness factor as the characteristic parameters of the support vector machine (SVM) fault recognition model, raw vibration data serve as the input of SVM fault recognition model. The SVM fault recognition model uses the same training sample data as k-means clustering model, composition of training sample data has been explained in section ''training sample data description''. As shown in Fig.15, labeled failure data are divided into four fault modes, surge is the first category, rubbing is the second category, misalignment is the third category, and normal state are the fourth category.

Using the 100 sets surge data collected from 210-K-1650 centrifugal compressor as the SVM model verification data, and the verification result shows that 57 sets of data

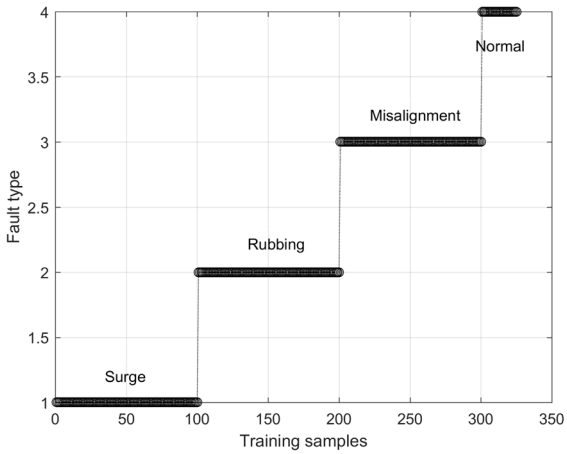

FIGURE 15. SVM fault training sample classification.

have been partitioned as surge fault, 27 sets of data have been partitioned as rubbing fault, and 16 sets of data have been partitioned as misalignment fault (Fig.16). It can be calculated that recognition accuracy rate of the surge fault of SVM model is 57%.

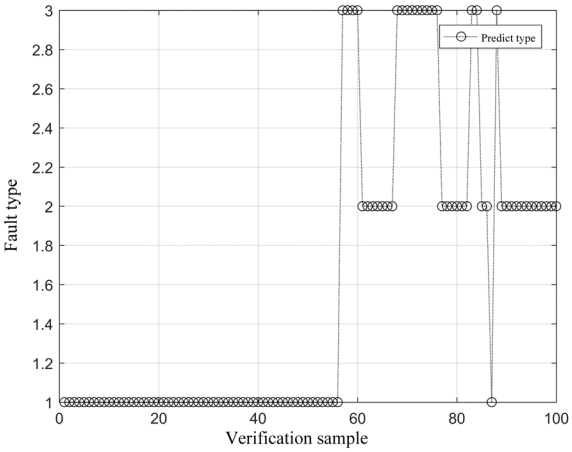

FIGURE 16. SVM surge fault sample classification.

Using the second 100 sets rubbing fault data collected from 120-K-20012 centrifugal compressor as the SVM model verification data, and the verification result shows that all 100 sets of data have been partitioned as rubbing fault (Fig.17). It can be calculated that recognition accuracy rate of the rubbing fault of SVM model is 100%.

Using the second 100 sets misalignment fault data collected from M-101 centrifugal compressor as the SVM model verification data, and the verification result shows that 3 sets of data have been partitioned as surge fault, 15 sets of data have been partitioned as rubbing fault, and 72 of data have been partitioned as misalignment fault (Fig.18). It can be calculated that recognition accuracy rate of the misalignment fault of SVM model is 72%.

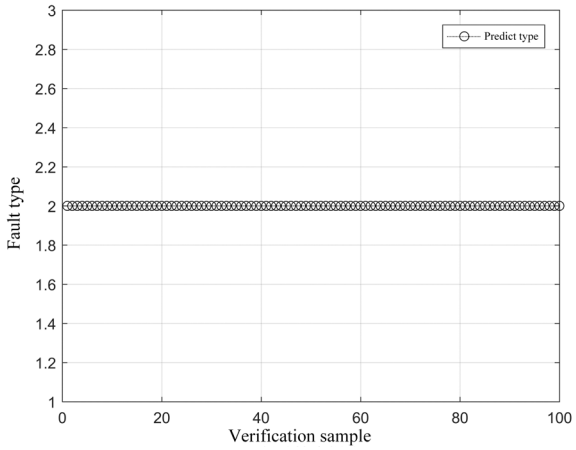

FIGURE 17. SVM rubbing fault sample classification.

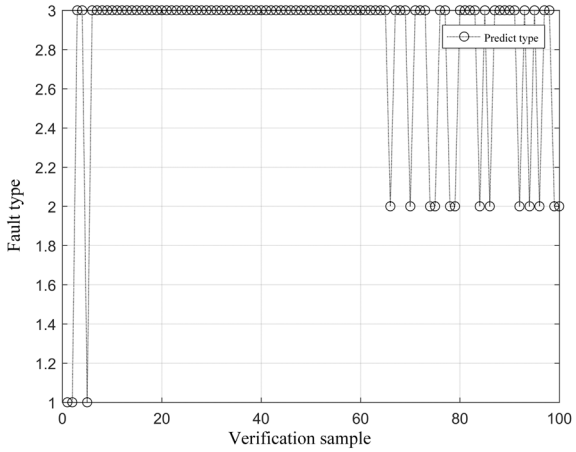

FIGURE 18. SVM misalignment fault sample classification.

TABLE 10. The accuracy rate comparison of fault recognition between K-means model and SVM model.

Fault recognition model	Surge	Rubbing	Misalignment
k-means	94%	100%	80%
SVM	57%	100%	72%

Support vector machine as a supervised learning method, fault recognition needs to be combined with training samples to obtain the optimal model for classification, while unsupervised learning k-means cluster analysis can be trained to obtain cluster centers, according to the distance criterion can automatically identify the failure mode, it is more suitable for engineering applications. As shown in Table 10, for surge and misalignment faults identification, the comparison results between k-means model and SVM model shows the former has higher fault recognition accuracy rate than the latter.

V. CONCLUSION

The main purpose of this paper is proposing an online k-means clustering fault recognition model by using time series raw vibration data of in-service rotating machinery. The k-means clustering fault recognition model has been constructed based on the ''black box'' principle, which is designed with three working modes such as offline or online model training, online fault recognition, online fault mode-fault cluster centroid knowledge base enrichment and so on. Using time-series raw vibration data as the input of the k-means clustering fault recognition model, the fault identification conclusion can be automatically given without depending on the prior knowledge of external experts.

The raw vibration data of surge, rubbing, misalignment and normal status of the in-service centrifugal compressor are used to train and verify the effectiveness of the k-means cluster fault recognition model. The result shows that surge fault, rubbing fault and misalignment fault recognition accuracy rate reach 94%, 100% and 80%, respectively. In addition to surge, rubbing, and misalignment fault modes, centrifugal compressors also have fault modes such as rotor unbalance, oil film oscillation and so on, the effectiveness of the k-means cluster analysis of vibration data for five or more operating states needs to be further studied, and the generalization of the constructed model needs to be further verified and perfected too.

ACKNOWLEDGMENT

The authors would like to thank all the reviewers and editors for their valuable comments and work.

REFERENCES

[1] K. T. P. Nguyen and K. Medjaher, ''A new dynamic predictive maintenance framework using deep learning for failure prognostics,'' *Rel. Eng. Syst. Saf.*, vol. 188, pp. 251–262, Aug. 2019.
[2] T. P. Carvalho, F. A. A. M. N. Soares, R. Vita, R. D. P. Francisco, J. P. Basto, and S. G. S. Alcalá, ''A systematic literature review of machine learning methods applied to predictive maintenance,'' *Comput. Ind. Eng.*, vol. 137, Nov. 2019, Art. no. 106024.
[3] M. Kordestani, M. Saif, M. E. Orchard, R. Razavi-Far, and K. Khorasani, ''Failure prognosis and applications—A survey of recent literature,'' *IEEE Trans. Rel.*, early access, Sep. 17, 2019, doi: 10.1109/TR.2019.2930195.
[4] R. Liu, B. Yang, E. Zio, and X. Chen, ''Artificial intelligence for fault diagnosis of rotating machinery: A review,'' *Mech. Syst. Signal Process.*, vol. 108, pp. 33–47, Aug. 2018.
[5] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, ''Deep learning and its applications to machine health monitoring,'' *Mech. Syst. Signal Process.*, vol. 115, pp. 213–237, Jan. 2019.
[6] W. Zhang, D. Yang, and H. Wang, ''Data-driven methods for predictive maintenance of industrial equipment: A survey,'' *IEEE Syst. J.*, vol. 13, no. 3, pp. 2213–2227, Sep. 2019.
[7] S. Manochandar, M. Punniyamoorthy, and R. K. Jeyachitra, ''Development of new seed with modified validity measures for k-means clustering,'' *Comput. Ind. Eng.*, vol. 141, Mar. 2020, Art. no. 106290.
[8] A. Purarjomandlangrudi, A. H. Ghapanchi, and M. Esmalifalak, ''A data mining approach for fault diagnosis: An application of anomaly detection algorithm,'' *Measurement*, vol. 55, pp. 343–352, Sep. 2014.
[9] V. Vercruyssen, W. Meert, G. Verbruggen, K. Maes, R. Baumer, and J. Davis, ''Semi-supervised anomaly detection with an application to water analytics,'' in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 527–536.
[10] J. Zhu, N. Chen, and C. Shen, ''A new deep transfer learning method for bearing fault diagnosis under different working conditions,'' *IEEE Sensors J.*, vol. 20, no. 15, pp. 8394–8402, Aug. 2020.
[11] M. Qiao, S. Yan, X. Tang, and C. Xu, ''Deep convolutional and LSTM recurrent neural networks for rolling bearing fault diagnosis under strong noises and variable loads,'' *IEEE Access*, vol. 8, pp. 66257–66269, 2020.

[12] T. Zhang, X. H. Wang, X. M. Xu, and C. P. Chen, "GCB-Net: Graph convolutional broad network and its application in emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Aug. 27, 2019, doi: 10.1109/TAFFC.2019.2937768.

[13] J. Long, Z. Sun, C. Li, Y. Hong, Y. Bai, and S. Zhang, "A novel sparse echo autoencoder network for data-driven fault diagnosis of delta 3-D printers," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 3, pp. 683–692, Mar. 2020.

[14] H. Huang, T. Zhang, C. Yang, and C. L. P. Chen, "Motor learning and generalization using broad learning adaptive neural control," *IEEE Trans. Ind. Electron.*, vol. 67, no. 10, pp. 8608–8617, Oct. 2020.

[15] M. Rezamand, M. Kordestani, M. E. Orchard, R. Carriveau, D. S-K. Ting, and and M. Saif, "Improved remaining useful life estimation of wind turbine drivetrain bearings under varying operating conditions (VOC)," *IEEE Trans. Ind. Informat.*, early access, May 7, 2020, doi: 10.1109/TII.2020.2993074.

[16] J. Long, S. Zhang, and C. Li, "Evolving deep echo state networks for intelligent fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4928–4937, Jul. 2020.

[17] M. Kordestani, A. Alkhateeb, I. Rezaeian, L. Rueda, and M. Saif, "A new clustering method using wavelet based probability density functions for identifying patterns in time-series data," in *Proc. IEEE EMBS Int. Student Conf. (ISC)*, May 2016, pp. 1–4.

[18] C. Wu, J. Lee, T. Isokawa, J. Yao, and Y. Xia, "Efficient clustering method based on density peaks with symmetric neighborhood relationship," *IEEE Access*, vol. 7, pp. 60684–60696, 2019.

[19] M. Unal, M. Onat, M. Demetgul, and H. Kucuk, "Fault diagnosis of rolling bearings using a genetic algorithm optimized neural network," *Measurement*, vol. 58, pp. 187–196, Dec. 2014.

[20] M. Schlechtingen and I. Ferreira Santos, "Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection," *Mech. Syst. Signal Process.*, vol. 25, no. 5, pp. 1849–1875, Jul. 2011.

[21] J. Zhu, N. Chen, and W. Peng, "Estimation of bearing remaining useful life based on multiscale convolutional neural network," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 3208–3216, Apr. 2019.

[22] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, "Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," *Mech. Syst. Signal Process.*, vols. 72–73, pp. 303–315, May 2016.

[23] S. Khan and T. Yairi, "A review on the application of deep learning in system health management," *Mech. Syst. Signal Process.*, vol. 107, pp. 241–265, Jul. 2018.

[24] N. Amruthnath and T. Gupta, "A research study on unsupervised machine learning algorithms for early fault detection in predictive maintenance," in *Proc. 5th Int. Conf. Ind. Eng. Appl. (ICIEA)*, Apr. 2018, pp. 355–361.

[25] Z. Chen, K. Gryllias, and W. Li, "Intelligent fault diagnosis for rotary machinery using transferable convolutional neural network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 339–349, Jan. 2020.

[26] K. C. Gryllias and I. A. Antoniadis, "A support vector machine approach based on physical model training for rolling element bearing fault detection in industrial environments," *Eng. Appl. Artif. Intell.*, vol. 25, no. 2, pp. 326–344, Mar. 2012.

[27] M. Alexandre, F. Carina, C. Jaques, G. Konstantinos, C. Bram, J. Karl, and H. Kilian, "Condition monitoring of gears under medium rotational speed," in *Proc. 24th Internation Congr. Sound Vibrat.*, London, U.K., 2017, p. 8.

[28] L. Bull, K. Worden, G. Manson, and N. Dervilis, "Active learning for semi-supervised structural health monitoring," *J. Sound Vibrat.*, vol. 437, pp. 373–388, Dec. 2018.

[29] T. J. Rogers, K. Worden, R. Fuentes, N. Dervilis, U. T. Tygesen, and E. J. Cross, "A Bayesian non-parametric clustering approach for semi-supervised structural health monitoring," *Mech. Syst. Signal Process.*, vol. 119, pp. 100–119, Mar. 2019.

[30] T. Zhang, G. Su, C. Qing, X. Xu, B. Cai, and X. Xing, "Hierarchical lifelong learning by sharing representations and integrating hypothesis," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Feb. 27, 2019, doi: 10.1109/TSMC.2018.2884996.

[31] K. Dhalmahapatra, R. Shingade, H. Mahajan, A. Verma, and J. Maiti, "Decision support system for safety improvement: An approach using multiple correspondence analysis, t-SNE algorithm and K-means clustering," *Comput. Ind. Eng.*, vol. 128, pp. 277–289, Feb. 2019.

[32] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas, "Randomized dimensionality reduction for k-means clustering," *IEEE Trans. Inf. Theory*, vol. 61, no. 2, pp. 1045–1062, Feb. 2015.

[33] G. Hamerly and J. Drake, *Accelerating Lloyd's Algorithm for K-Means Clustering, Partitional Clustering Algorithms*. New York, NY, USA: Springer, 2015, pp. 41–67.

[34] E. Uhlmann, R. P. Pontes, C. Geisert, and E. Hohwieler, "Cluster identification of sensor data for predictive maintenance in a selective laser melting machine tool," *Procedia Manuf.*, vol. 24, pp. 60–65, Jan. 2018.

[35] A. Rai and S. H. Upadhyay, "Bearing performance degradation assessment based on a combination of empirical mode decomposition and k-medoids clustering," *Mech. Syst. Signal Process.*, vol. 93, pp. 16–29, Sep. 2017.

[36] J. Zhang, H. Sun, Z. Sun, W. Dong, Y. Dong, and S. Gong, "Reliability assessment of wind power converter considering SCADA multistate parameters prediction using FP-growth, WPT, K-Means and LSTM network," *IEEE Access*, vol. 8, pp. 84455–84466, 2020.

[37] P. Baraldi, F. Di Maio, M. Rigamonti, E. Zio, and R. Seraoui, "Clustering for unsupervised fault diagnosis in nuclear turbine shut-down transients," *Mech. Syst. Signal Process.*, vols. 58–59, pp. 160–178, Jun. 2015.

[38] N. A. A. Shashoa, G. Kvaščev, A. Marjanović, and Ž. Djurović, "Sensor fault detection and isolation in a thermal power plant steam separator," *Control Eng. Pract.*, vol. 21, no. 7, pp. 908–916, Jul. 2013.

[39] E. Zio, "Prognostics and health management of industrial equipment," in *Diagnostics and Prognostics of Engineering Systems: Methods and Techniques*. Hershey, PA, USA: IGI Global, 2012, pp. 333–356.

[40] V. Kavitha and M. Punithavalli, "Clustering time series data stream—A literature survey," *Int. J. Comput. Sci. Inf. Secur.*, vol. 8, no. 1, pp. 289–294, 2010.

[41] T. Mitsa, *Temporal Data Mining*. Boca Raton, FL, USA: CRC Press, 2010.

[42] M. Reder, N. Y. Yürüşen, and J. J. Melero, "Data-driven learning framework for associating weather conditions and wind turbine failures," *Rel. Eng. Syst. Saf.*, vol. 169, pp. 554–569, Jan. 2018.

[43] S. Al-Dahidi, F. Di Maio, P. Baraldi, E. Zio, and R. Seraoui, "A framework for reconciliating data clusters from a fleet of nuclear power plants turbines for fault diagnosis," *Appl. Soft Comput.*, vol. 69, pp. 213–231, Aug. 2018.

QINGFENG WANG received the Ph.D. degree from the College of Mechanical and Electrical Engineering, Beijing University of Chemical Technology, China, in 2011. He is currently an Associate Researcher with the College of Mechanical and Electrical Engineering, Beijing University of Chemical Technology. He is also a Security Expert of the State Administration of Work Safety of China. His research interests include electromechanical equipment monitoring, diagnosis, and maintenance.

JIAHE LIU received the B.S. degree from the College of Electrical Engineering, Beijing University of Chemical Technology, China, in 2017, where he is currently pursuing the M.S. degree. His current research interests include fault diagnosis and health evaluation.

BINGKUN WEI received the B.S. degree from the College of Electrical Engineering, Beijing University of Chemical Technology, China, in 2018, where he is currently pursuing the M.S. degree. His current research interests include fault diagnosis and health evaluation.

WENWU CHEN is currently a Professorate Senior Engineer with the SINOPEC Qingdao Research Institute of Safety Engineering, China. He is mainly engaged in the research of integrity management, fault diagnosis, and prediction technology application in oil refining and chemical equipment.

SHUJIAN XU is currently a Professorate Senior Engineer with SINOPEC Qingdao Research Institute of Safety Engineering, China. His research interests include anti-corrosion technology, material failure analysis technology, and asset integrity management technology.