Journal of Network Intelligence ©2024 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **9**, Number **1**, February **2024**

Data-Driven Fault Detection of Rotating Machinery Based on Density Peak Clustering

Yang Shen*

College of Information Engineering
Guangzhou Institute of Technology
Guangzhou 510075, P. R. China
sy9301@163.com

Hao Li

Department of Engineering
Krirk University
Bangkhen 10220, Thailand
007.lihao@163.com

Xie-Fei He

College of Information Engineering
Guangzhou Institute of Technology
Guangzhou 510075, P. R. China
hexiefei2017@163.com

*Corresponding author: Yang Shen

Received September 6, 2023, revised November 21, 2023, accepted December 24, 2023.

Abstract. : *Modern rotating machinery is becoming increasingly sophisticated and complex, making it difficult for manual Operation and Maintenance(O&M) methods to detect abnormal conditions and determine the causes of failures of rotating machinery and equipment in a timely manner. Data-driven intelligent O&M technology can establish the mapping relationship between equipment and operation status from large-scale historical monitoring data. Therefore, a rotating machinery fault detection method based on data mining technology is proposed. First, a sliding time window is used to divide the historical data sequences, and a subset of the sequences is represented as a fault feature vector according to the type of each sequence. The corresponding attribute values are calculated and the cluster centroids are discovered using local density. An improved density peak clustering method is designed using a new labeling delivery method. Then, the density peak clustering analysis is performed on the sample points in the healthy state in order to construct the baseline health state matrix. Finally, a nonlinear mapping function is used to calculate the difference between the real-time failure modes and the healthy baseline to quantitatively assess the operational status of the equipment. The experimental results show that the proposed fault detection method has high accuracy and stability for a variety of common rotor system faults.*

Keywords: Data mining; Density peak clustering; Rotating machinery; Fault detection; Data-driven

1. **Introduction.** Rotating machinery fault diagnosis has been one of the important research directions in the field of equipment condition monitoring and fault diagnosis. The operating condition of rotating machinery directly affects the safety and economy of the whole system. The development of rotating machinery fault monitoring and diagnosis

methods has gone through the process from traditional regular maintenance to condition monitoring and fault diagnosis [1,2]. The traditional method exists diagnosis lag, maintenance blind problem.

The emergence of condition monitoring and fault diagnosis technology has realized the change from regular maintenance to condition monitoring and condition based maintenance. Commonly used condition parameters are vibration, temperature, sound, current and so on [3,4]. Early failure warning can be realized through feature extraction and pattern recognition. Modern rotating machinery is becoming increasingly sophisticated and complex, which makes it difficult for manual operation and maintenance methods to discover the abnormal conditions of rotating mechanical equipment and determine the cause of failure in time [5]. Currently, intelligent diagnosis has become a development trend. Machine learning methods are used to establish fault diagnosis and prediction models [6, 7]. At the same time, remote online monitoring is carried out by combining the technology of the Internet of Things [8]. However, the challenges of rotating machinery fault diagnosis include the high-dimensional complexity of state information, the difficulty of identifying different fault mechanisms, and the improvement of the generalization performance of diagnostic models [9].

The traditional way of judging through manual experience inevitably suffers from large subjective differences and lagging diagnosis. Regular maintenance is also prone to over or under maintenance. These have prompted the shift from periodic maintenance to condition monitoring and fault diagnosis. Vibration, acoustic emission, and current signals are widely used for condition monitoring [10, 11]. Through signal processing and feature extraction, sensitive parameters can be selected to build condition assessment and fault diagnosis models. Commonly used features are time domain, frequency domain and time-frequency domain features. Wavelet transform, fuzzy and other methods can also be used to extract features. The high-dimensional complexity of the state, the difficulty of identifying different fault mechanisms, and the enhancement of model robustness and generalizability all require continuous research. Artificial Intelligence for IT Operations (AIOps) is a technology that applies artificial intelligence and machine learning to IT operations management [12]. AIOps collects, aggregates, and analyzes various types of operations data through automation to realize intelligent monitoring, fault prediction, abnormality diagnosis, problem location, and other functions of IT systems, networks, and applications, so as to enhance and optimize the IT operations management process. monitoring, fault prediction, abnormality diagnosis, problem location and other functions, so as to enhance and optimize the IT operation and maintenance management process. In the application of rotating machinery fault diagnosis, AIOps can collect and analyze the massive mechanical operating parameters, alarm information, operation and maintenance logs and other structured and unstructured data, and discover the hidden failure mode [13]. Through machine learning and other technologies to establish complex fault diagnostic models, to achieve accurate identification and localization of fault types and root causes. AIOps can predict the likelihood, time and severity of mechanical failures based on big data analysis, to achieve the prediction and early warning of failures. Through intelligent analysis, prediction, decision-making and other means, AIOps can significantly improve the automation and intelligence level of rotating machinery fault diagnosis and operation and maintenance, which has important research value.

The most commonly used machine learning techniques in AIOps are data mining algorithms such as k-mean clustering [14]. Clustering techniques are utilized to provide explanatory fault detection and enhance the interpretability of the detection process for engineers to parse and apply. As an advanced clustering algorithm, Density Peak Clustering (DPC) [15] can automatically cluster a large number of event logs and system faults in an unsupervised manner, discover the intrinsic connection between events and faults, and categorize them into different types of faults.DPC helps AIOps to extract knowledge and insights from complex O&M data automatically, so that O&M data can be automatically extracted and analyzed. and insights to make O&M decisions more intelligent, which deserves further research. Therefore, the research objective of this work is to utilize DPC to perform unsupervised clustering on the operation data of rotating machinery to automatically identify different types of fault states. The intrinsic connection between the monitored data is analyzed by clustering to realize AIOps for rotating machinery.

1.1. **Related Work.** Currently, research on fault detection in rotating machinery can be divided into the following categories.

(1) Fault detection based on physical model. This kind of method realizes fault detection by establishing a physical model of rotating machinery, analyzing its dynamic characteristics and designing sensitive parameters. Xu et al. [16] established a dynamic model of bearings and designed fault characteristic parameters based on differential rotational speed, which combined with the noise processing technology to realize the identification of various types of faults of bearings. Sun et al. [17] established a dynamic model of cracked rotor through theoretical analysis and simulation, and studied the natural frequency and stress distribution of the rotor. By establishing a physical model to analyze the influence of faults on dynamic parameters, it lays a theoretical foundation for the fault monitoring based on physical model, but the model simplification is still a difficult point in practical application. (2) Fault detection based on signal processing. This kind of method needs to analyze the signals of vibration, acoustic emission, current, etc., and extract the fault features through signal processing method for pattern recognition. Zhao and Zhang [18] proposed a new time-frequency decomposition method, which is applied to the condition monitoring of low-speed spiral reducer, and realized the effective detection of various problems, including tooth mating faults, tooth wear, etc. However, the signal processing process relies too much on feature selection and requires manual extraction of design fault-related features, which relies on the experience of experts, and there are difficulties in selecting features for different fault scenarios. For example, the selection of wavelet basis function, the number of modes of modal decomposition and other parameters have a great influence on the results, and need to be repeatedly tested.

(3) Data-driven fault detection. This type of method uses machine learning and other algorithms to learn fault patterns directly from operation and maintenance data to construct a detection model. Ye et al. [19] proposed a learning framework based on 1D convolutional neural network, which realizes the intelligent detection of bearing faults and improves the accuracy of fault diagnosis. Different from the neural network model, using the clustering results to classify the operational data is beneficial to discover the intrinsic patterns of the data and realize the fault detection. Typical clustering algorithms include k-means clustering, Gaussian mixture model, etc., which can realize unsupervised clustering of mechanical states. Dreher et al. [20] used k-means clustering to mine the state patterns of rotor machinery to detect bearing faults. As an advanced clustering algorithm, DPC can automatically cluster a large number of event logs and system faults in an unsupervised manner [21, 22], discovering the intrinsic connection between events and faults and categorizing them into different fault types.

1.2. **Motivation and contribution.** Due to the large uncertainty of the real-time state of rotating machinery under dynamic working conditions, the fault diagnosis method based on clustering algorithm needs to calculate the distance of all sample points in order to obtain the abnormality of the sample points.

In addition, the interpretability of fault diagnosis methods based on clustering algorithms still needs to be explored. Therefore, to address the problem of uncertainty in the real-time health state of rotating machinery under dynamic operating conditions, this work proposes a fault detection method based on DPC and Nonlinear Mapping Function (NMF).

The main innovations and contributions of this work include:

(1) Aiming at the problem of misclassification of clusters easily caused by manually selecting cluster centroids in the traditional DPCA algorithm, an Improved Density Peak Clustering (IDPC) method is designed by using a new labeling delivery method.

(2) Based on IDPC clustering to obtain representative sample points of large-scale fault sample points, it can effectively adapt to the efficient analysis of different types of rotating machinery AIOps systems.

(3) NMF is designed to calculate the real-time abnormality, thus realizing the quantitative assessment of equipment health status. Compared with existing similarity metrics, NMF has better interpretability and accuracy. In addition, the abnormality calculation of sample points avoids the inefficiency caused by the distance calculation of all sample points.

2. **Improved peak density clustering.**

2.1. **Density-based clustering algorithm.** Clustering algorithms are both classical and extremely important as an unsupervised learning method, especially in the current complex real-world network environment. Clustering algorithms are more suitable as a current technique for anomalous traffic detection since the network environment is not capable of obtaining class labeling of newly collected traffic on a large scale with a high degree of accuracy and efficiency. Compared with other clustering algorithms, Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [23, 24] starts from the key of data objects and clusters them according to their density correlation, which can find out any clusters in noisy spatial data. Find out arbitrary clusters, further study the connectability between different data objects and keep expanding, finally get the clustering result. As a classical algorithm, different from divisional clustering and hierarchical clustering algorithms, DBSCAN can find out the point with high density and then define the high density region around the point into clusters. The related definition concepts of DBSCAN algorithm are as follows:

1) Neighborhood. The circle represented by a data point as the center and e as the radius is defined as the neighborhood, that is, the set of data objects whose Euclidean distance from other points is less than e. The density value of the data point is the number of data points in the circle.

2) Core points. Given a neighborhood sample threshold Pmin, the center point of the circle which indicates that the number of data points in the circle is greater than Pmin is called a high density point or core point, otherwise it is called a low density point.

3) Density Direct. A peripheral data point m is said to be density-directed by a core data point n if the peripheral data point m is in the neighborhood of the core data point n. In other words, the high-density points in the neighborhood of the high-density point are connected, and so on, and then all such points are connected. If a low-density point is within the neighborhood of a high-density point, connecting it to the nearest high-density point is called a boundary point.

4) Density reachable. A data point k is said to be density reachable from data point n if there are the same number of sub-data points k in the peripheral data point m and core data point n.

5) Density connected. A peripheral data point m is said to be density connected to a core data point n if there exists a peripheral data point m that is reachable by the density of sub-data point k, and a core data point n that is also reachable by the density of sub-data point k.

The concepts related to the DBSCAN algorithm are shown in Figure 1. In this case, the dotted lines indicate the neighborhoods of the preset data points N_1, N_2, N_3, N_4, N_5. The Pmin indicates that there are 3 additional data points in the neighborhood, and it can be seen that there are 4 data points in each circle. Data point N_2 is directly accessible from data point N_1 density, data point N_3 is accessible from data point N_1 density, and data point N_6 is connected to data point N_1 density.

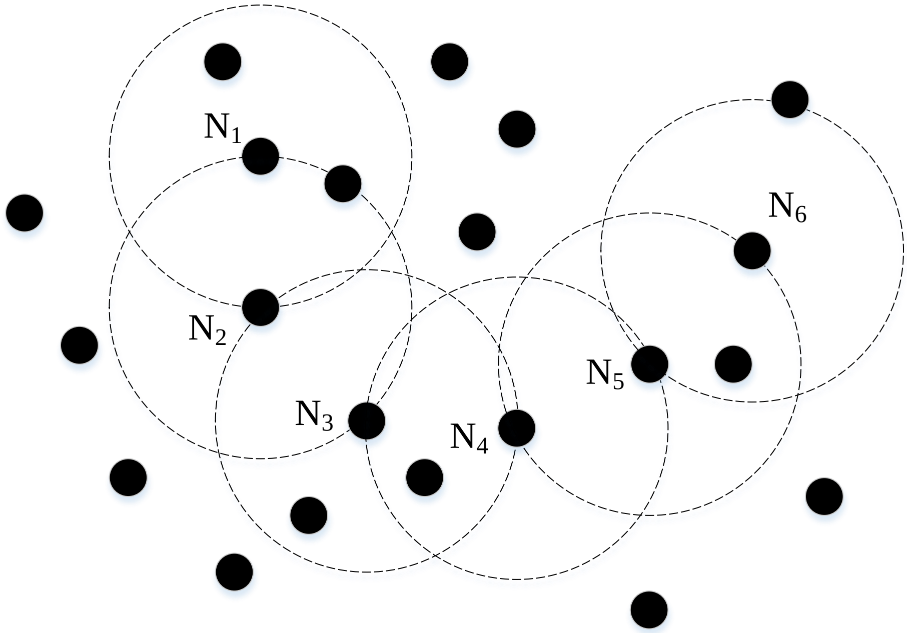


Figure 1. Schematic conceptualization of DBSCAN

2.2. **Peak density clustering algorithm.** Clustering algorithms mainly classify the points with closer Euclidean distance into one class cluster and the points with farther Euclidean distance into other class clusters. Although there are many types of clustering algorithms, the definition of clusters in clustering still has not been standardized.

The DPC algorithm is a novel density-based clustering algorithm [25], introduced in 2014. The algorithm first determines the centroid of the cluster, and then adds the rest of the data points to the corresponding class clusters. The DPCA algorithm selects the centroid of the cluster under 2 conditions: the first condition is that the selected centroid of the cluster is locally the peak density point within the neighborhood, i.e., the maximum density value; the second condition is that this centroid of the cluster is far away from the Euclidean distances of all other similarly localized data points with larger densities. The DPCA algorithm mainly The main purpose of DPCA algorithm is to calculate the relative distance δ and local density ρ of all sample data points, and then construct the corresponding decision diagram based on these two attribute values. Then, the data points with larger values of relative distance δ and local density ρ are selected as cluster centers. Other data points are added to the cluster where the data point with the smallest value δ of relative distance is located according to the decreasing value of local density ρ

from large to small. Where the relative distance δ_i of any data point i is defined as:

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{1}$$

where d_{ij} is the Euclidean distance of any data point i from another data point j.

If this data point i is the highest density in the whole globe, its distance is defined:

$$\delta_i = \max_j (d_{ij}) \tag{2}$$

The Euclidean distance of this data point i is equal to the maximum distance of other data points from this point. The local density ρ_i of any data point i is defined as:

$$\rho_i = \sum_{i \neq j} x(d_{ij} - d_c), \quad x(X) = \begin{cases} 1, X \geq 0 \\ 0, X < 0 \end{cases} \tag{3}$$

where d_c is the cutoff distance.

Equation (3) is used under the condition that the local density is more discriminative when the data volume is large. If the amount of data is small, it is necessary to use the Gaussian function to find the local density value ρ_i of data point i, ρ_i is the sum of the weighted values of the Euclidean distances of data points in the neighborhood of this data point i to solve the problem of low distinction of local density.

$$\rho_i = \sum_j \exp(-\frac{d_{ij}^2}{d_c^2}) \tag{4}$$

2.3. **IDPC.** This work proposes IDPC this algorithm consists of 2 main steps: first, identifying the 2 attribute values of each data point in the computed dataset; and second, constructing a decision map and starting clustering. Cluster centers are identified among the data points that are denser than the neighboring data points.

IDPC uses 2 different metrics to identify the cluster centers and then clusters the data points using the label propagation distance method. The steps of IDPC are as follows:

Step 1: Enter the data set D and truncate the distance d_c ;

Step 2: Discover the cluster centroids and calculate the point distance matrix using equations (1) and (2). Calculate the local density of points using Equation (3) and Equation (4), construct a decision diagram and select the cluster centroids;

Step 3: Form clusters, assign the labels of the cluster center points to the nearest neighbor points for clustering based on neighborhood distance matrix and density, and assign each remaining point to the nearest cluster center;

Step 4: Output the clustering result cluster C .

The corresponding data is computed in Step 2, which is consistent with the traditional DPC algorithm. The difference is that a new label passing method is proposed in Step 2, which finally forms clusters based on the processed cluster centroids, assigns a different label to each cluster centroid, and each cluster centroid passes its label to its nearest neighbor. For a data point i which does not have any label or processed, if its local density value ρ_i is less than ρ_j, then the data point gets the label of the fetched data point j. It can be shown that the time complexity of the IDPC algorithm is $O(N^2)$, where N is the number of sample data points in D.

Finally, the rules for judging the anomalous states in the dataset, in this work, the anomalous state samples are defined to satisfy the following conditions: local density $\rho_i < P_{\min}$, relative distance $\delta_i < \delta_r$, where $P_{\min}$ is the local density threshold.

$$P_{\min} = \frac{1}{N} \sum_{i=1}^{N} \rho_i - \gamma_\rho \tag{5}$$

The relative distance threshold δ_r is defined as follow:

$$\delta_\tau = \frac{1}{N}\sum_{i=1}^{N}\delta_i - \gamma_\delta \tag{6}$$

where γ_ρ and γ_δ are empirical parameters.

3. IDPC-NMF based rotating machinery fault detection.

3.1. Subset partitioning and vectorized representation.

In order to meet the requirements of real-time anomaly detection, data sequences over a period of time need to be analyzed. The sliding window mechanism is a processing method for data sequences, which is able to divide the full amount of sequence data according to the generation time of the sequence or the number of elements. Therefore, this work is based on sliding time windows to segment historical data sequences. An example of the three sequence subsets is shown in Figure 2. The attributes of the sliding time window include the length of the window and the step size, the length of the window indicates the time span of each division, and the step size indicates the time interval of each slide. Generally speaking, the larger the length of the time window, the more information it covers, and the better the effect of anomaly detection, but for the fault detection task, a time window with a large span may lead to a reduction in the real-time and accuracy of detection, so the window length needs to be determined according to the actual situation. For the step length attribute, it is usually set to be smaller than the window length to obtain a larger subset of sequences and ensure that each sequence is classified into the corresponding fault subset.

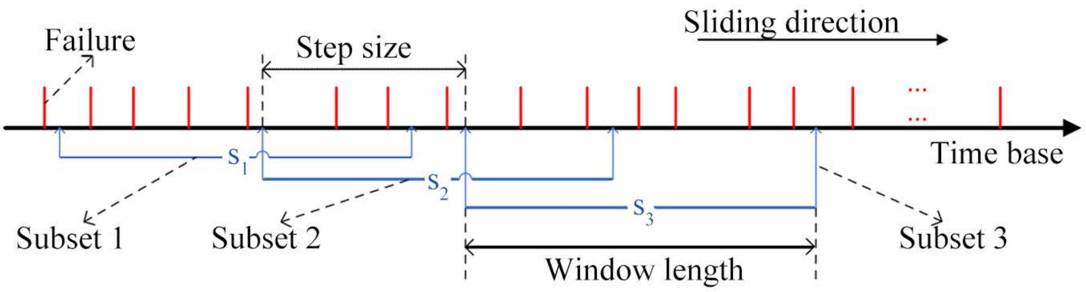


Figure 2. Examples of three subsets of sequences

After completing the division of the subset of sequences, a vectorized representation of them is performed. The number of different types of sequences is vectorized to represent them as eigenvalues.

3.2. IDPC-based benchmark matrix under different operating conditions.

The variation of fault characteristics of rotating machinery under different operating conditions is analyzed. Assuming that there exists an optimal value of the eigenvectors of the faults when the equipment operates normally under a certain working condition, at which time the equipment state is the healthiest, then the health state of the equipment corresponds to a set of optimal value eigenvectors under the condition of multiple working conditions. In this work, we refer to the concept of health baseline and define the baseline health state matrix to express the mode of the equipment state under multiple operating conditions.

It is assumed that there are m operating conditions characteristics of the rotating machinery during operation. Under a certain condition characteristic c, the health baseline of the equipment is $X^c = (x_1^c, x_2^c, \dots, x_k^c)$, $c = 1, 2, \dots, m$, where x_i^c denote the optimal value of the i-th characteristic under the condition characteristic c. The baseline health state matrix can be expressed as follows The baseline health state matrix $D_{m\times k}$ can be expressed as follows:

$$D_{m\times k} = \begin{bmatrix} X^1 \\ X^2 \\ \dots \\ X^m \end{bmatrix} = \begin{bmatrix} x^1 & x^1 & \dots & x_k^1 \\ x^2 & x^2 & \dots & x_k^2 \\ \dots & \dots & \dots & \dots \\ x^m & x_2^m & \dots & x_k^m \end{bmatrix} \tag{7}$$

This matrix stores a collection of health baseline vectors of the equipment under different operating conditions, so the baseline health state matrix is actually a collection of representative fault feature vectors. In this work, IDPC is used to cluster the sample points under health states.

IDPC determines the initial clustering center by calculating the sample attribute density and distance values [26], and then determines the category based on the distance between the sample points and the center point. A DPC decision map is generated from the density and distance two-dimensional coordinate graph to select the clustering center. Let the sample X contains $C = \{C_1, C_2, \dots, C_k\}$ categorys, that satisfies the conditions: $k \le N$, N is the total number of sample sets) and $X = C_1 \cup C_2 \cup \dots \cup C_k$. The distance r_{ij} between any two sample points x_i and x_j is shown as follow:

$$r_{ij} = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{in} - x_{jn}|^2} \tag{8}$$

where n denotes the dimension.

The density x_i of the point among the N points ρ_i is shown as follow:

$$\rho_i = \sum\nolimits_j \chi(r_{ij} - r_c) \tag{9}$$

where r_c is the distance threshold and the kernel function $\chi(x)$ is shown as follow:

$$\chi(x) = \begin{cases} 1, x < 0 \\ 0, x \ge 0 \end{cases} \tag{10}$$

Since $\chi(x)$ is not derivable, a Gaussian kernel function is often used as a replacement.

$$\rho_i = \sum\nolimits_j e^{-\frac{r_{ij}^2}{2r_c^2}} \tag{11}$$

The minimum distance δ_i of the point x_i is calculated as follow:

$$\delta_i = \begin{cases} \min_j(r_{ij}), if\ j\ s.t.\ \rho_j > \rho_i \\ \max_j(r_{ij}), \qquad otherwise \end{cases} \tag{12}$$

Calculate ρ_i and δ_i for N sample points and generate a decision map using both as horizontal and vertical axis coordinates, multiplying the density and distance values for all points.

$$\gamma_i = \rho_i \cdot \delta_i \tag{13}$$

Then all the sample points of ρ_i, δ_i and γ_i are arranged in descending order, and the point that is larger than all three points is taken as the center of clustering. Finally, for the non-center points, the distance value from the point to the center is used to identify the category.

The cluster centers obtained by the IDPC algorithm have the characteristics of maximum local density and maximum distance between cluster centers, so the IDPC clustering can learn the representative fault feature vectors of the health states from the positive samples. Each category represents different working conditions, the category cluster center represents the health baseline of the category, and the set of all cluster center vectors constitutes the baseline health state matrix.

First, the local density *rho* of each sample is calculated based on k-nearest neighbors, and the distance *delta* of each sample is calculated. Then, find the one with larger *rho* and *delta* at the same time as the clustering center. The IDPC-based mechanical failure benchmark matrix construction method is shown in Algorithm 1.

Algorithm 1 IDPC-based Mechanical Failure Benchmark Matrix

```
# Import the required libraries
import numpy as np
from sklearn.neighbors import NearestNeighbors
# Input
X = mechanical sensor data
# Peak density clustering
# 1. Calculate the local density $\rho$ for each sample
knn = NearestNeighbors(n_neighbors=5)
knn.fit(X)
distances, _ = knn.kneighbors(X)
$\rho = np.sum(np.exp(-distances), axis = 1)$
# 2. Calculate the delta for each sample
$\delta = np.min(distances, axis = 1)$
#3. find the peak density ($\rho$ and $\delta$ both larger) as the center of clustering
cluster_centers = []
for i in range(len(X)) do
    if $\rho[i] == np.max(\rho)$ and $\delta[i] == np.max(\delta)$ then
        cluster_centers.append(i)
    end if
end for
# 4. Assign other samples to clusters where the nearest clustering center is located
clusters = [[] for _ in range(len(cluster_centers))]
for i in range(len(X)) do
    if i not in cluster_centers then
        closest = np.argmin(np.sqrt((X[i] - X[cluster_centers])**2))
        clusters[closest].append(i)
    end if
end for
# Perform troubleshooting
faulty_clusters = []
for c in clusters do
    if determines that c represents a fault state then
        faulty_clusters.append(c)
    end if
end for
print("Faulty benchmark matrix:", faulty_clusters)
```

3.3. **Nonlinear mapping function.** The abnormality of a device can be obtained by calculating the similarity between the sample points and the healthy baseline. However, most of the traditional similarity measures use various distance measures between vectors

[27], which do not take into account the weighting factors of different features or the effect of different variations of the features on the anomaly metric value.

Thus it does not accurately reflect the degree of abnormality of the current state of the device compared to the healthy state. To address the above problems, this work proposes a weighted nonlinear mapping function to quantitatively assess the degree of abnormality of a device. The feature weights of the sample points are designed. Among the multidimensional features of the sample points, the changes of different features have different impacts on the health state of the equipment, so it is necessary to weight the original sample points, i.e., to give more weight to the fault features that have a greater impact on the health state of the equipment. The formula for the designed combined weight w_i is as follows:

$$w_i = \frac{\theta \cdot \zeta_i}{\sum\limits_{i=1}^{k} \zeta_i} + \frac{(1-\theta)\,\eta_i}{\sum\limits_{i=1}^{k} \eta_i} \tag{14}$$

$$\zeta = \frac{\bar{x}_i}{\sigma_i} \tag{15}$$

$$\eta_i = \frac{N_c}{x_i'} \tag{16}$$

where σ_i is the standard deviation of the feature, $\bar{x}_i$ is the mean value of the feature, k is the feature dimension, N_c is the number of positive sample points under the working condition characteristic c, and θ is the scale parameter.

The feature weights are composed of two parts, including the distribution weights and importance weights of the features. The distribution weight is determined by the inverse coefficient of variation of the feature values, which is defined here as the inverse of the coefficient of variation. In historical data, if the distribution of feature values in a dimension is more concentrated, it has higher weight in anomaly detection. The importance weight of the feature is determined by the mean value of the fault importance of the sample points to be tested. In practice, the proportion parameter can be used to adjust the ratio of the two to obtain better model performance.

Let x_i denote the i-th eigenvalue of the sample to be tested, x_i^c denote the i-th eigenvalue of the health baseline, and w_i denote the combination weight of the i-th eigenvalue, then the abnormality δ_n^c of the sample point to be tested under the condition of working condition characteristic c is calculated as follows:

$$\delta_n^c = \sum_{i=1}^{k} w_i \cdot \mathrm{ReLU}(x_{i'} - x_i^c) \tag{17}$$

where ReLU denotes the linear correction function.

The greater the abnormality of the sample point to be tested, the worse the health status of the equipment.

4. Experimental results and analysis.

4.1. Experimental environment and experimental dataset.

The experimental hardware environment is: Intel Core i5 2.2GHz processor, 6G RAM, 400G hard disk, GTX1060 discrete graphics card. The experimental software environment is: Windows 7 operating system, Matlab 2012 (R2012a) simulation software.

The Society for Machinery Failure Prevention Technology (MFPT) dataset was selected for the failure detection experiments of rotating machinery. The MFPT dataset, released in 2022, contains vibration signals of many types of bearings under different operating conditions, with a total data volume of more than 150GB, and more than 20 types of bearings. The main parameters of the MFPT dataset are shown in Table 1.

4.2. **Clustering performance validation.** The performance of IDPC algorithm is compared with k-means algorithm [28] and DBSCAN algorithm [29], which are evaluated in terms of four performance metrics, namely, running time, completeness, homogeneity, and accuracy, respectively, as shown in Table 2. The IDPC algorithm and k-means algorithm are set to set the number of clustering centers to 25, and the default parameters are selected for DBSCAN algorithm.

Table 1. Main parameters of the MFPT dataset.

Parameters	**Clarification**
Name	MFPT Bearing Data Set
Release time	2022
Data content	150GB vibration signals of different types of bearings
Bearing type	More than 20 types including ball bearings, roller bearings, etc.
Failure mode	Normal, inner ring failure outer ring failure, etc.
Transducers	Vibration acceleration sensors
Sampling frequency	96 kHz
Speed range	600-6000 RPM
Loading force	0-8000 pound.
Data set format	MAT files

Table 2. Performance Comparison of Different Clustering Algorithms.

Performance indicators	***k*-means**	**DBSCAN**	**IDPC**
Running time/s	3.44	5.79	3.84
Completeness	0.6	0.631	0.604
Homogeneity	0.779	0.775	0.783
Accuracy	0.81198	0.79016	0.82013

It can be seen that in terms of running time, the k-means algorithm takes the shortest time of 3.44 s, the IDPC algorithm is second only to it, while DBSCAN has a longer running time. In terms of completeness higher values are better, DBSCAN algorithm has the highest value, IDPC algorithm remains second and k-means algorithm has the lowest value. In terms of homogeneity, the IDPCA algorithm has the highest value, the k-means algorithm is second and the DBSCAN algorithm has the lowest value. In terms of clustering accuracy, the IDPC algorithm has the highest accuracy, the k-means algorithm has the second highest and the DBSCAN algorithm has the lowest.

4.3. **Comparison of fault detection results.** The sequences in the healthy state are first divided according to a window of fixed length of time.

In order to simulate a multi-operating condition environment, each sample point was collected from two operating conditions of the equipment, idle or busy, and the length of the time window was set to be 2 hours with a step size of 1.5 hours. After data preprocessing, a total of 144 idle-time working condition sample points and 96 busy-time working condition sample points were obtained. Then the maximum-minimum normalization was applied to each dimension of each sample point to eliminate the influence of different orders of magnitude of features in each dimension on the clustering results [30].

$$x_i' = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}, i = 1, 2, \cdots, k \tag{18}$$

IDPC clustering is performed on the normalized sample points. The Gaussian kernel function is used in this experiment to define the local density of a point. The truncation distance was set to be the top 2% after sorting the distances between all sample points in ascending order. The results of the clustering analysis are shown in Figure 3. The horizontal coordinates in the IDPC clustering decision diagram represent the local density of each sample point, and the vertical coordinates represent the cluster center distance of each sample point.

From the decision diagram, it can be intuitively found that the IDPC clustering can accurately divide all sample points into two classes, and the two points in the upper right corner have higher local density and larger cluster center distance, thus representing the healthy baseline of the alarm features under the two kinds of working conditions. Then the weights of the feature combinations under the two working conditions are calculated separately, and here the scale parameter is set to $\theta = 0.3$, and the results are shown in Table 3.

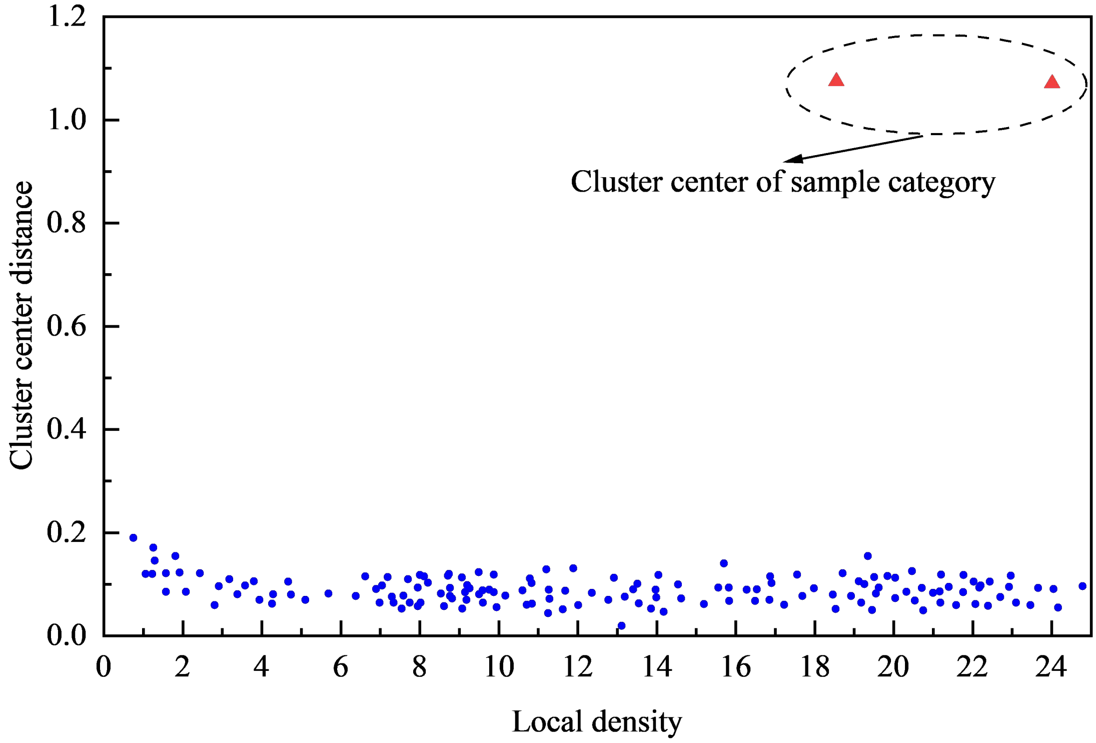
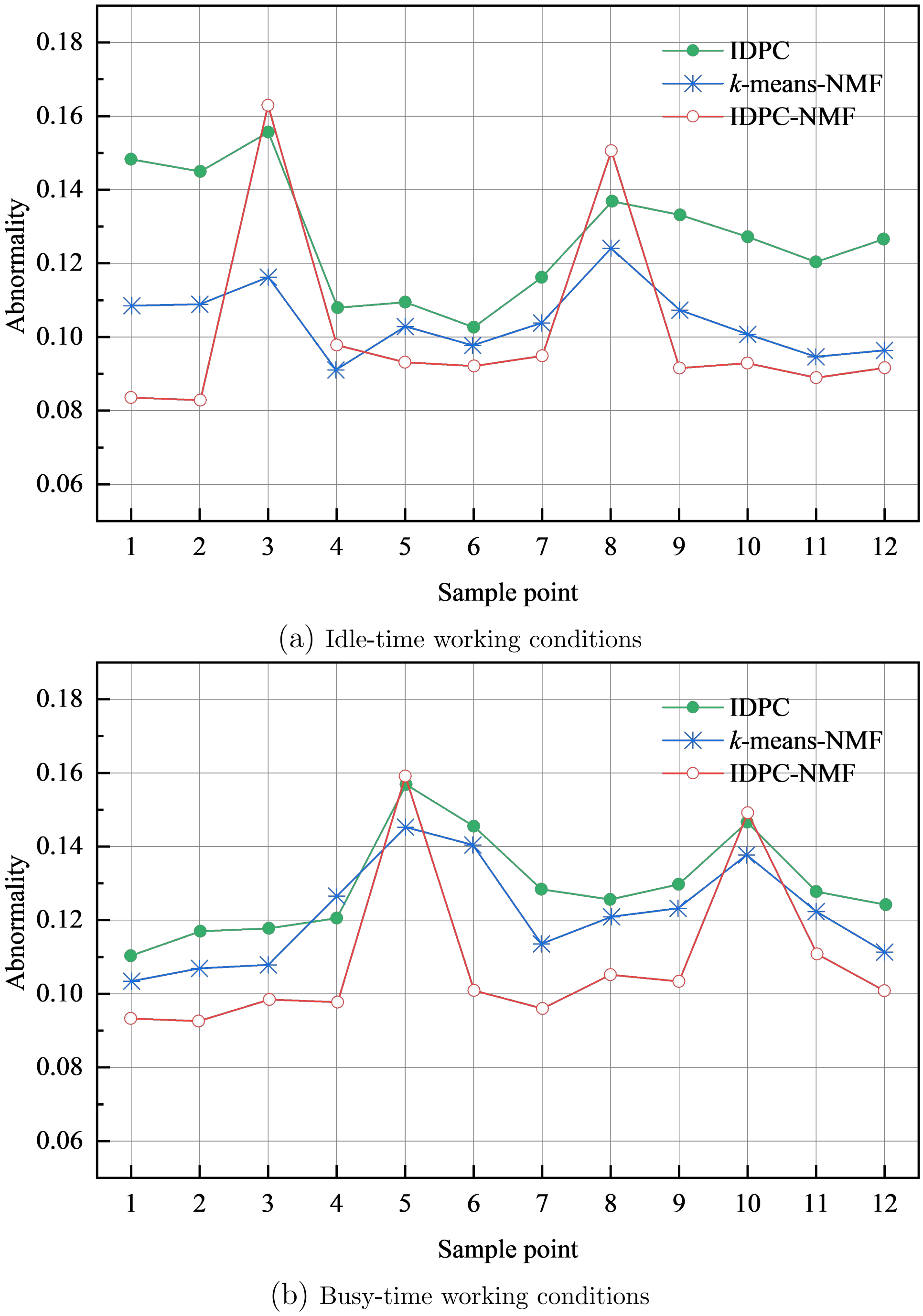


Figure 3. Clustering decision diagram for IDPC

Table 3. Health baseline and feature weights for two operating conditions.

Working condition	Health baseline	Combination weights (distribution weights + importance weights)
Idle-time	0.24	0.33+0.67
Busy-time	0.59	0.24+0.76

This experiment validates the effectiveness of IDPC-NMF. Firstly, 12 sample points of the device under two working conditions are collected for the experiment, and each working condition contains two anomaly sample points. IDPC-NMF, k-means-NMF and IDPC are used to calculate the abnormality, respectively. The experimental results are shown in Figure 4 From the experimental results, it can be seen that the IDPC-NMF proposed in this work is able to detect the abnormal sample points more obviously under the two working conditions, i.e., sample point 3 and sample point 8 under the idle-time working condition and sample point 5 and sample point 10 under the busy time working condition, which show a more obvious increase in the degree of abnormality. However, non-detection and misidentification occurred when using k-means-NMF and IDPC, which



(a) Idle-time working conditions

(b) Busy-time working conditions

Figure 4. Anomalies of different methods under two operating conditions

is due to the fact that IDPC-NMF is able to attenuate unimportant feature variations and increase the effect of variations in important features on the abnormality.

5. **Conclusion.** Aiming at the uncertainty problem of real-time health state of rotating machinery under dynamic working conditions, this work proposes a fault detection method based on DPC-NMF. A new label passing method is used to design the IDPC, which solves the problem of manually selecting the cluster centroids in the traditional DPCA algorithm that is prone to misclassification of clusters. Based on IDPC clustering to obtain representative sample points of large-scale fault sample points, it can be effectively adapted to analyze different types of rotating machinery AIOps systems with high efficiency. NMF is designed to calculate the real-time abnormality so as to realize the quantitative assessment of equipment health status. The abnormality calculation of sample points avoids the inefficiency caused by the distance calculation for all sample points. Experimental results show that the proposed fault detection method has high accuracy and stability for a variety of common rotor system faults. However, NMF is usually more complex than linear functions, and the computational cost is usually higher than linear functions, requiring more computational resources and time. Further research will be conducted on how to reduce the complexity of NMF.

Funding Statement. This work was supported by Tertiary Education Scientific research project of Guangzhou Municipal Education Bureau (No. 202235310); Guangdong Provincial Special Fund for Science and Technology Innovation Strategy (Cultivation of Science and Technology Innovation for College Students) (pdjh2022b0951).

REFERENCES

[1] Z. Zhu, Y. Lei, G. Qi, Y. Chai, N. Mazur, Y. An, and X. Huang, “A review of the application of deep learning in intelligent fault diagnosis of rotating machinery,” *Measurement*, 112346, 2022.

[2] Y. Ma, Y. Peng, and T.-Y. Wu, “Transfer learning model for false positive reduction in lymph node detection via sparse coding and deep learning,” *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 2, pp. 2121-2133, 2022.

[3] T.-Y. Wu, J. C.-W. Lin, U. Yun, C.-H. Chen, G. Srivastava, and X. Lv, “An efficient algorithm for fuzzy frequent itemset mining,” *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 5787-5797, 2020.

[4] T.-Y. Wu, J. Lin, Y. Zhang, and C.-H. Chen, “A Grid-Based Swarm Intelligence Algorithm for Privacy-Preserving Data Mining,” *Applied Sciences*, vol. 9, no. 4, 774, 2019.

[5] Y. Zhang, T. Zhou, X. Huang, L. Cao, and Q. Zhou, “Fault diagnosis of rotating machinery based on recurrent neural networks,” *Measurement*, vol. 171, 108774, 2021.

[6] X. Li, H. Shao, S. Lu, J. Xiang, and B. Cai, “Highly efficient fault diagnosis of rotating machinery under time-varying speeds using LSISMM and small infrared thermal images,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 12, pp. 7328-7340, 2022.

[7] Y. Cheng, M. Lin, J. Wu, H. Zhu, and X. Shao, “Intelligent fault diagnosis of rotating machinery based on continuous wavelet transform-local binary convolutional neural network,” *Knowledge-Based Systems*, vol. 216, 106796, 2021.

[8] L. Chen, W. Gan, Q. Lin, S. Huang, and C.-M. Chen, “OHUQI: Mining on-shelf high-utility quantitative itemsets,” *The Journal of Supercomputing*, vol. 78, no. 6, pp. 8321-8345, 2022.

[9] C.-M. Chen, L. Chen, W. Gan, L. Qiu, and W. Ding, “Discovering high utility-occupancy patterns from uncertain data,” *Information Sciences*, vol. 546, pp. 1208-1229, 2021.

[10] W. Gan, L. Chen, S. Wan, J. Chen, and C.-M. Chen, “Anomaly Rule Detection in Sequence Data,” *IEEE Transactions on Knowledge and Data Engineering*, pp. 1-1, 2022.

[11] J. Lu, W. Qian, S. Li, and R. Cui, “Enhanced K-nearest neighbor for intelligent fault diagnosis of rotating machinery,” *Applied Sciences*, vol. 11, no. 3, 919, 2021.

[12] A. Dibaj, M. M. Ettefagh, R. Hassannejad, and M. B. Ehghaghi, “A hybrid fine-tuned VMD and CNN scheme for untrained compound fault diagnosis of rotating machinery with unequal-severity faults,” *Expert Systems with Applications*, vol. 167, 114094, 2021.

[13] G. Yu, T. Lin, Z. Wang, and Y. Li, "Time-reassigned multisynchrosqueezing transform for bearing fault diagnosis of rotating machinery," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 2, pp. 1486-1496, 2020.

[14] M. Shafapourtehrany, P. Yariyan, H. Özener, B. Pradhan, and F. Shabani, "Evaluating the application of K-mean clustering in Earthquake vulnerability mapping of Istanbul, Turkey," *International Journal of Disaster Risk Reduction*, vol. 79, 103154, 2022.

[15] Y. Chen, X. Hu, W. Fan, L. Shen, Z. Zhang, X. Liu, J. Du, H. Li, Y. Chen, and H. Li, "Fast density peak clustering for large scale data based on kNN," *Knowledge-Based Systems*, vol. 187, 104824, 2020.

[16] G. Xu, D. Hou, H. Qi, and L. Bo, "High-speed train wheel set bearing fault diagnosis and prognostics: A new prognostic model based on extendable useful life," *Mechanical Systems and Signal Processing*, vol. 146, 107050, 2021.

[17] W. Sun, T. Li, D. Yang, Q. Sun, and J. Huo, "Dynamic investigation of aeroengine high pressure rotor system considering assembly characteristics of bolted joints," *Engineering Failure Analysis*, vol. 112, 104510, 2020.

[18] Y. Zhao, and H. Zhang, "Recent Patents on Third Generation Bearing Testing Machine," *Recent Patents on Engineering*, vol. 16, no. 4, pp. 63-80, 2022.

[19] M. Ye, X. Yan, N. Chen, and M. Jia, "Intelligent fault diagnosis of rolling bearing using variational mode extraction and improved one-dimensional convolutional neural network," *Applied Acoustics*, vol. 202, 109143, 2023.

[20] N. R. Dreher, I. O. de Almeida, G. C. Storti, G. B. Daniel, and T. H. Machado, "Feature analysis by k-means clustering for damage assessment in rotating machinery with rolling bearings," *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 44, no. 8, 330, 2022.

[21] W. Guo, W. Wang, S. Zhao, Y. Niu, Z. Zhang, and X. Liu, "Density peak clustering with connectivity estimation," *Knowledge-Based Systems*, vol. 243, 108501, 2022.

[22] Z. Shu, S. Yang, H. Wu, S. Xin, C. Pang, L. Kavan, and L. Liu, "3D shape segmentation using soft density peak clustering and semi-supervised learning," *Computer-Aided Design*, vol. 145, 103181, 2022.

[23] N. Hanafi, and H. Saadatfar, "A fast DBSCAN algorithm for big data based on efficient density calculation," *Expert Systems with Applications*, vol. 203, 117501, 2022.

[24] Y. Yang, C. Qian, H. Li, Y. Gao, J. Wu, C.-J. Liu, and S. Zhao, "An efficient DBSCAN optimized by arithmetic optimization algorithm with opposition-based learning," *The Journal of Supercomputing*, vol. 78, no. 18, pp. 19566-19604, 2022.

[25] Y. Han, K. Li, F. Ge, and W. Xu, "Online fault diagnosis for sucker rod pumping well by optimized density peak clustering," *ISA transactions*, vol. 120, pp. 222-234, 2022.

[26] Y. Yang, J. Cai, H. Yang, and X. Zhao, "Density clustering with divergence distance and automatic center selection," *Information Sciences*, vol. 596, pp. 414-438, 2022.

[27] L. Pineda, T. Fan, M. Monge, S. Venkataraman, P. Sodhi, R. T. Chen, J. Ortiz, D. DeTone, A. Wang, and S. Anderson, "Theseus: A library for differentiable nonlinear optimization," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3801-3818, 2022.

[28] E. Zhang, H. Li, Y. Huang, S. Hong, L. Zhao, and C. Ji, "Practical multi-party private collaborative k-means clustering," *Neurocomputing*, vol. 467, pp. 256-265, 2022.

[29] H. Chen, M. Liang, W. Liu, W. Wang, and P. X. Liu, "An approach to boundary detection for 3D point clouds based on DBSCAN clustering," *Pattern Recognition*, vol. 124, 108431, 2022.

[30] X. Zhang, Y. Chen, J. Jia, K. Kuang, Y. Lan, and C. Wu, "Multi-view density-based field-road classification for agricultural machinery: DBSCAN and object detection," *Computers and Electronics in Agriculture*, vol. 200, 107263, 2022.