Shen *et al. Chin. J. Mech. Eng.* (2020) 33:16
https://doi.org/10.1186/s10033-020-00437-3

Chinese Journal of Mechanical Engineering

ORIGINAL ARTICLE Open Access

A Fast Multi-tasking Solution: NMF-Theoretic Co-clustering for Gear Fault Diagnosis under Variable Working Conditions

Fei Shen[1], Chao Chen[1], Jiawen Xu[1] and Ruqiang Yan[1,2*]

Abstract

Most gear fault diagnosis (GFD) approaches suffer from inefficiency when facing with multiple varying working conditions at the same time. In this paper, a non-negative matrix factorization (NMF)-theoretic co-clustering strategy is proposed specially to classify more than one task at the same time using the high dimension matrix, aiming to offer a fast multi-tasking solution. The short-time Fourier transform (STFT) is first used to obtain the time-frequency features from the gear vibration signal. Then, the optimal clustering numbers are estimated using the Bayesian information criterion (BIC) theory, which possesses the simultaneous assessment capability, compared with traditional validity indexes. Subsequently, the classical/modified NMF-based co-clustering methods are carried out to obtain the classification results in both row and column tasks. Finally, the parameters involved in BIC and NMF algorithms are determined using the gradient ascent (GA) strategy in order to achieve reliable diagnostic results. The Spectra Quest's Drivetrain Dynamics Simulator gear data sets were analyzed to verify the effectiveness of the proposed approach.

Keywords: Gear fault diagnosis, Non-negative matrix factorization, Co-clustering, Varying working conditions

1 Introduction

In those large-scale rotating machines, wear and tear always comes out in the teeth surface of driving gears if the pressure is not even or some extra impurities are mingled in the lubricating oil. Health monitoring technology of mechanical components has been proved to be effective at discovering early abrasion and reducing the failure rate [1–5]. As one of the major tasks in health monitoring, gear fault diagnosis (GFD) aims to assess the current gear state based on the obtained measurement data, then to inform users to take proper actions [6]. A GFD procedure generally consists of three main processes: (1) Data acquisition: data are collected from sensors to monitor the health status of gears; (2) Feature extraction: some feature extraction algorithms, such as wavelet transform (WT) [7] and least squares support vector machine (LSSVM) [8], are carried out based on the prior knowledge to provide recognizable features; (3) Fault recognition: classifiers are built to obtain gear faults with the analysis of the extracted features.

Clustering technology, one of unsupervised fault recognition approaches, has experienced long term development from partition based clustering to graph theory based clustering as listed in Table 1. Most of these algorithms were applied to fault diagnosis of rotating machinery. For instance, Yuwono et al. [9] combined particle clustering with a Hidden Markov Model (HMM) for bearing fault diagnosis; Pacheco et al. [10] classified gear fault severities using rough set theory. These researches have provided effective clustering applications related to machine fault diagnosis. However, they have a non-negligible limitation: each feature vector is treated as independent and uncorrelated unit in these clustering methods. In fact, strong correlation exists between machine working conditions and

*Correspondence: ruqiang@seu.edu.cn
[1] School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China
Full list of author information is available at the end of the article

Table 1 Clustering technology lists

Category	Algorithms
Partitional-based	Graph theories [15]; K-means [16]; C-means [17]; FCM [18], etc.
Hierarchical-based	Nearest neighbor [19]; Binary positive [20], etc.
Density and grid-based	AGRID [21]; GCHL [22], etc.
Graph theory-based	ACODF [23], etc.

targeted diagnostic tasks. There is no doubt that clustering analysis in one dimension will lose significant information hidden in another dimension. To overcome this limitation, a co-clustering strategy for variable working conditions GFD is proposed in this paper. Co-clustering was firstly utilized in the biology and medical domains since this concept was mentioned by Mirkin [11] in 1997. The joint clustering of genes shapes and locations promoted the discovery of genetic structure sequence [12]. Subsequently, co-clustering has been expanded to other fields such as text analysis [13] and search engines [14], etc.

In this study, co-clustering applications for gear fault diagnosis have been developed. Compared with previous GFD based on joint clustering methods, two highlights in this paper can be obtained: (1) when one varying working condition (such as rotating speed or load) and one diagnostic task (such as fault severity) are jointed in the same matrix, their correlations are extracted, and the classification accuracy of the latter can adjust with the range of the former; (2) when two diagnostic tasks (such as fault severity and fault type) are jointed in the same matrix, they can be classified at the same time, which improves the diagnosis speed compared with independent GFD strategy and offer a fast multi-tasking solution.

The remainder of this paper is organized as follows. Section 2 presents a brief summary about the applicability of co-clustering. It also describes the principle and basic framework of co-clustering to solve the GFD problem. Section 3 addresses the preparatory work of GFD, especially a short-time Fourier transform (STFT)-based feature extraction method. In Section 4, the co-clustering numbers are estimated based on the Bayesian information criterion (BIC). Then in Section 5, the traditional and modified NMF-theoretic co-clustering process is discussed in detail. To assess these algorithms, the gradient ascend algorithm is also implemented for parameters regulation in Section 6. Section 7 concentrates on the varying working condition GFD experiments using the Drivetrain Dynamics Simulator (DDS) system, which especially shows the superiority of co-clustering compared with classical clustering strategy such as X-means algorithm. Conclusions are drawn in Section 8 with discussions on the future GFD application based on joint clustering.

2 Co-clustering Framework of GFD

Traditional clustering can be defined as: dataset $\boldsymbol{X}$ exists in a limited data space, which can be represented with a $n \times d$ matrix, is composed of n elements: $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})^{\mathrm{T}}$. The purpose of general clustering is to segment dataset X into p categories: $C_k (k = 1, 2, \cdots, p)$.

$$\boldsymbol{X} = \left[\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_i \ \cdots \ \boldsymbol{x}_n \right]^{\mathrm{T}}, \tag{1}$$

where $i \in \{1, 2, \ldots, n\}$.

Different from general clustering, co-clustering could be defined as: dataset $\boldsymbol{X}$ exists in a limited data space, which can be represented with a $m \times n \times d$ matrix, is composed of $m \times n$ elements: $\boldsymbol{x}_{ij} = \left(x_{ij1}, x_{ij2}, \cdots, x_{ijd}\right)^{\mathrm{T}}$. The purpose of co-clustering is to segment dataset $\boldsymbol{X}$ into p and q categories in horizontal and vertical axis, respectively: $C_k^h (k = 1, 2, \cdots, p)$ and $C_k^v (k = 1, 2, \cdots, q)$.

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_{11} & \cdots & \boldsymbol{x}_{1i} & \cdots & \boldsymbol{x}_{1n} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ \boldsymbol{x}_{j1} & \cdots & \boldsymbol{x}_{ji} & \cdots & \boldsymbol{x}_{in} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{m1} & \cdots & \boldsymbol{x}_{mi} & \cdots & \boldsymbol{x}_{mn} \end{bmatrix}, \tag{2}$$

where $i \in \{1, 2, \ldots, n\}$, $j \in \{1, 2, \ldots, m\}$.

Currently, one challenge for fault diagnosis of gears is that they are often operated under the varying working conditions. To explain the influence of varying working conditions on the diagnosis performance, a typical example is given in Figure 1 to compare the classification results between classical and co-clustering strategies. In the extracted feature matrix, different colors represent different feature value. C1–C5 represent different fault categories; D1–D5 represent different working conditions. R1–R9 represent the 1st clustering result; r1–r5 represent the 2nd clustering result. It can be seen that the classical clustering results may be distorted due to the fact that different working conditions have direct influence on the extracted features, and this makes difference between final clustering categories (R1–R9) and real fault categories (C1–C5). On the contrast, the co-clustering can reflect the actual diagnosis results (R1–R5), which shows the robustness of co-clustering in GFD under interference environment.

Notice that, co-clustering is not limited in the two dimensions. Theoretically, it can be generalized to higher dimension ($n \geq 3$), thus giving an idea to solve

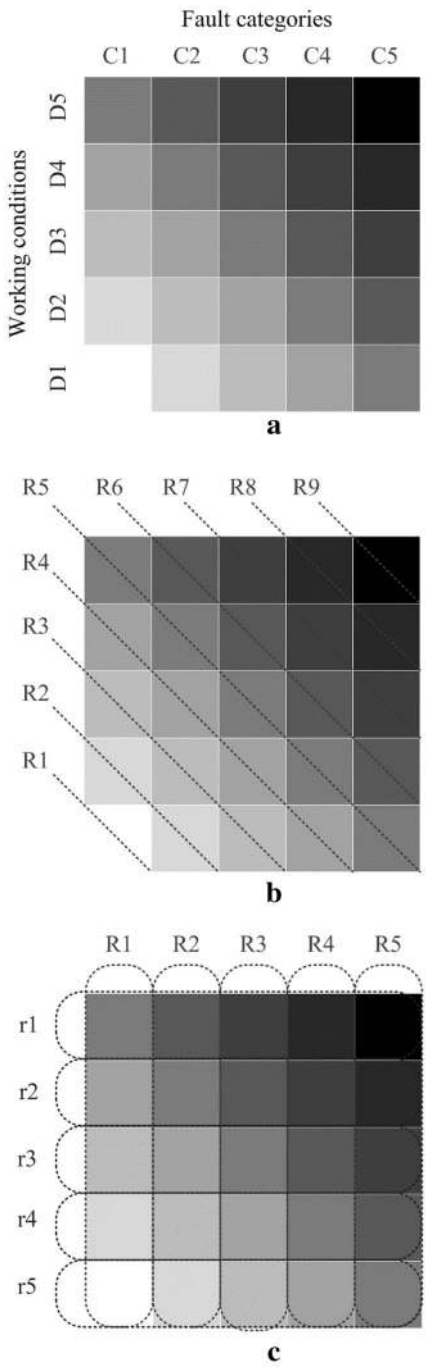

Figure 1 The comparison between classical clustering and co-clustering results: **a** extracted features, **b** classical clustering results, **c** co-clustering results

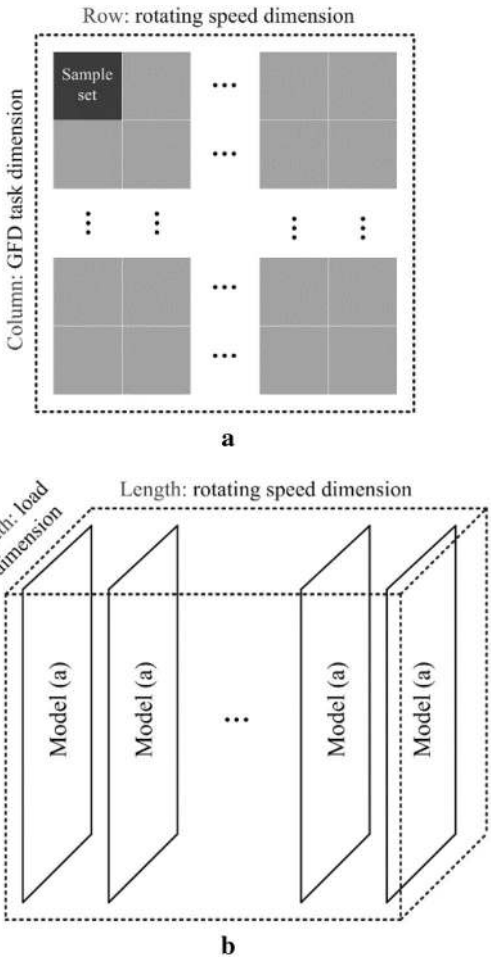

Figure 2 High dimension gear fault diagnosis models: **a** the 2D classification matrix; **b** the 3D classification matrix

the more complex problem such as the gear fault diagnosis problem under multiple working conditions. Figure 2 explains the mechanism of co-clustering GFD models under two working condition factors, such as varying rotating speed and varying load. In Model (a), the 2D classification matrix is structured in two dimensions: row for rotating speed and column for GFD task. In Model (b), the 3D classification matrix is structured in three dimensions: length for rotating speed, width for load and height for GFD task. Structured high dimension matrix is classified in each scale at the same time, which offers an idea for GFD under more than one working condition.

According to the description above, a co-clustering framework is constructed for gear fault diagnosis, which is shown in Figure 3. The main process can be divided into four sub-frames.

- Feature extraction sub-frame: as the input of this model, the gear vibration signals are collected using several tri-axial accelerometers installed in monitored mechanical equipment, which may be operated in varying working condition environment.

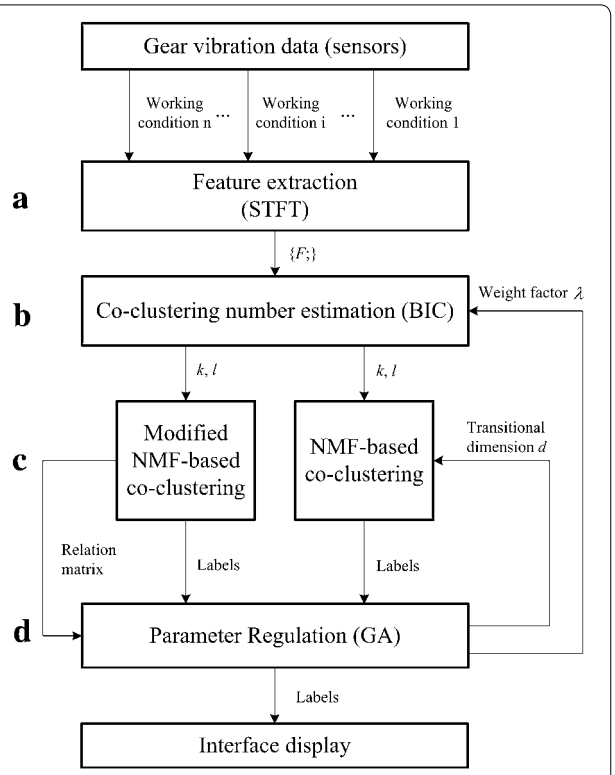

Figure 3 The co-clustering framework of GFD: **a** the feature extraction sub-frame; **b** the clustering number estimation sub-frame; **c** the co-clustering sub-frame; **d** the parameter regulation sub-frame

Then, the feature vectors $\{F\}$ are obtained using the short-time Fourier transform (STFT) approach, aiming to gain differentiable time-frequency features for effective co-classifier performance;

- Clustering number estimation sub-frame: the Bayesian information criterion (BIC) strategy is adopted to characterize the distribution character of all feature vectors $\{F\}$ and then estimate their co-clustering numbers k & l in row and column, respectively;
- Co-clustering sub-frame: given co-clustering numbers, the conventional as well as modified NMF-based co-clustering classifiers are put into practice to build the varying working condition GFD models and get the classification results in various tasks;
- Parameter regulation sub-frame: aiming to those adjustable parameters involved in BIC and the NMF algorithm, such as the weight factor λ and the transitional dimension d, the gradient ascent (GA) algorithm is implemented to find the optimal values, which reaches a reliable diagnostic accuracy.

More details of these four sub-frames will be given from Section 3 to Section 6.

3 STFT-based Feature Extraction

The performance of feature extraction algorithms depends critically on the characteristic of the input gear fault signals as well as the running environment of equipment. Because of the non-stationary property of the vibration signal, general algorithms, such as Fourier transform (FT), are not very useful in this regard [24]. As a kind of time-frequency analysis method, the main idea of short-time Fourier transform (STFT) is to execute Fourier transform in short sequential signals, cut by a sliding temporal window function $\gamma(t)$, such as a Hanning or Hamming window [25]. When analyzing the non-stationary vibration signal $s(t), (t = 1, 2, \ldots, t_0)$, we supposed that it can be approximated as smooth among the window function $\gamma(t)$. Therefore, the STFT of $s(t)$ is calculated by time-frequency units $STFT(t, w)$ and is given by

$$STFT(t,w) = \int_{-\infty}^{+\infty} s(\tau)\gamma^*(\tau - t)e^{-\mathrm{j}w\tau}\mathrm{d}\tau, \tag{3}$$

where τ means the position of window function $\gamma(s)$; w represents the frequency parameter of STFT. Figure 4 illustrates a STFT example for a chipped tooth signal with the Hamming window. It can be seen from Figure 4(d) that there are four main peaks (approximate 120 Hz, 500 Hz, 1160 Hz, and 1770 Hz) appearing in the frequency domain of the STFT spectrogram. The frequency at 1770 Hz possesses the most obvious peak, up to 0.5 V, compare to other peaks.

Since the short time Fourier spectrum reflects a distribution of energies among all frequencies and temporal intervals, we have to seek those 'meaningful' values from the STFT graph. Without any known knowledge of the running equipment, an effective method is to find each maximum in the partitions of the STFT spectrogram to represent the feature of sub-window. Figure 5 gives an example of extracted features in the chipped gear in Figure 4, including 10×10, 40×40 and 80×80 features. Generally, the row & column divided dimensions depends on the non-stationary degree of signal in time and frequency domain, respectively. This figure indicates that more details are emerged in high dimension features compared with low dimension features. However, the increase of dimension will bring the expansion of computational load as well as time consuming. Therefore, a middle dimension is suitable if the definition and dimension are both acceptable, such as the 40×40 features in Figure 5(c). Here $\boldsymbol{STFT}$ matrix is composed of N^2 sub-matrixes $\boldsymbol{st}_{ij} \in \boldsymbol{R}(n/N \times m/N)$, where $i \in \{1, 2, \ldots, N\}, j \in \{1, 2, \ldots, N\}$:

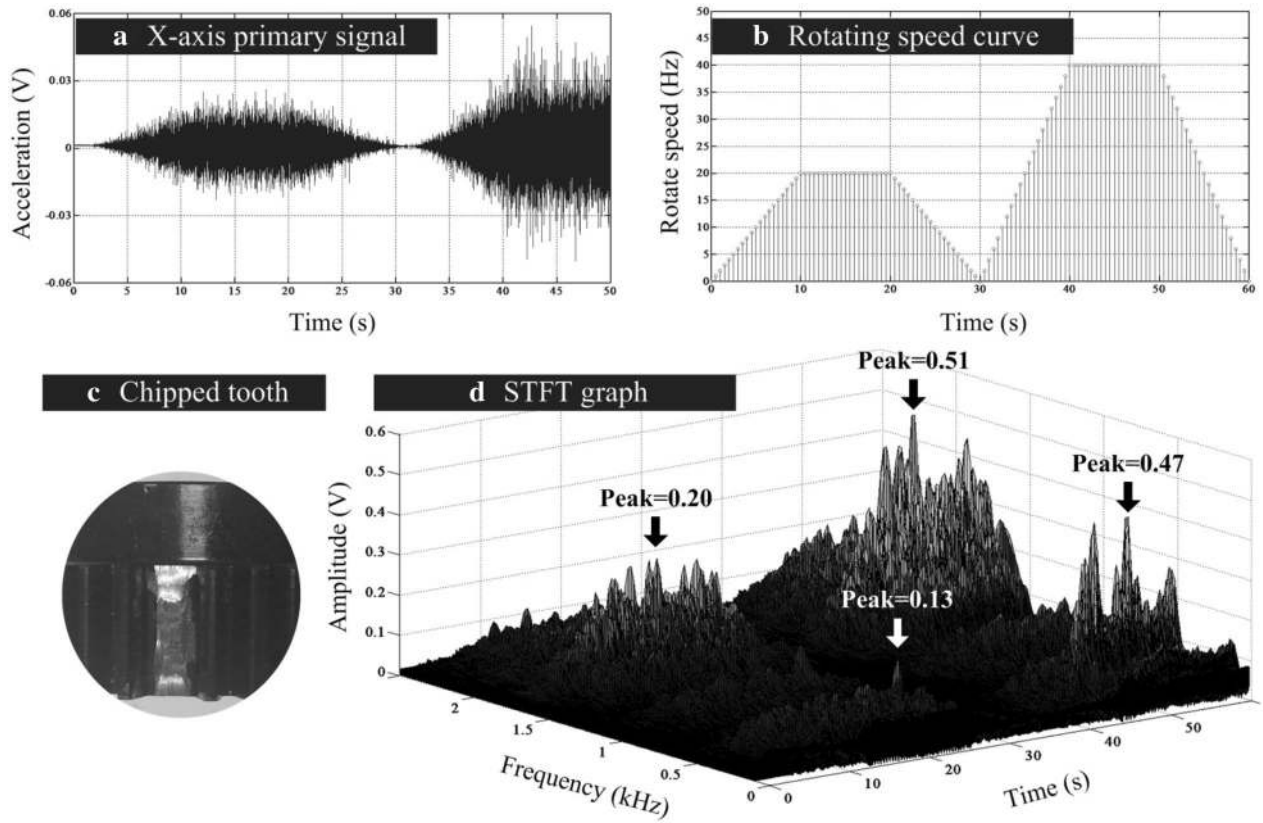

Figure 4 A STFT example for a chipped tooth signal with the Hamming window: **a** *X*-axis primary signal; **b** varying rotating speed curve; **c** fault gear with a chipped tooth; **d** the STFT graph

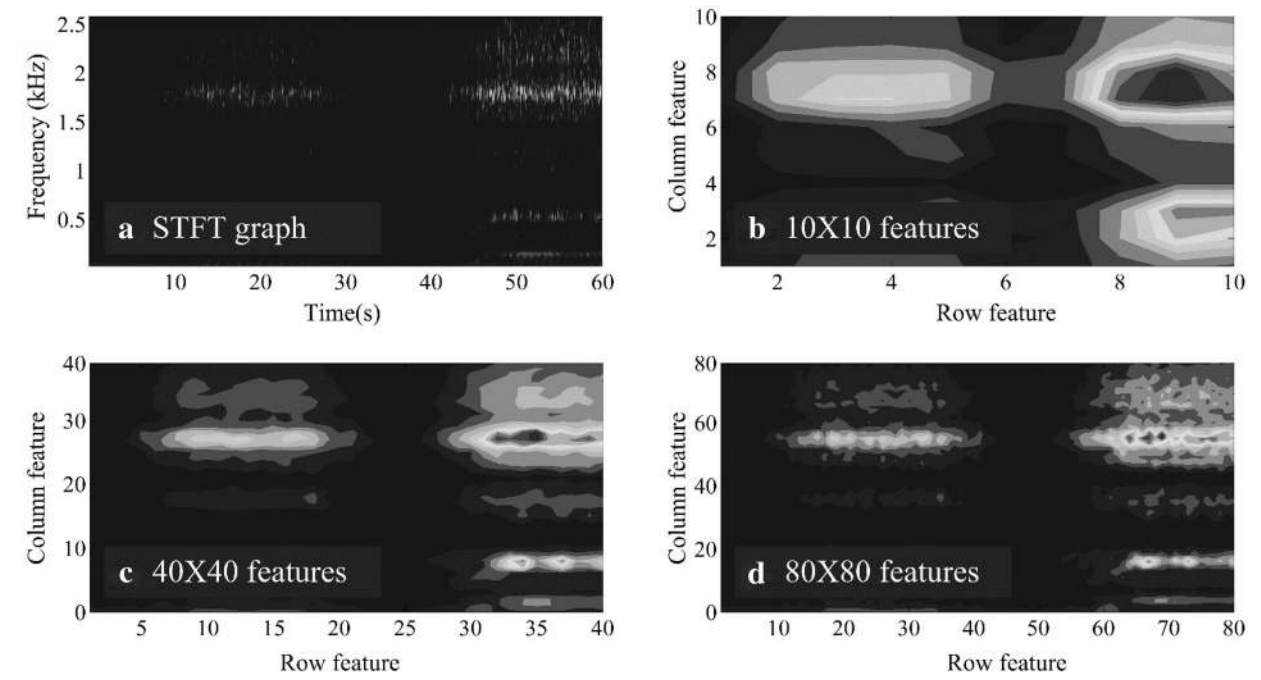

Figure 5 An example of extracted features in the chipped gear: **a** the STFT graph; **b** 10 × 10 features; **c** 40 × 40 features; **d** 80 × 80 features

$$\boldsymbol{STFT} = \begin{bmatrix} \boldsymbol{st}_{11} & \cdots & \boldsymbol{st}_{1j} & \cdots & \boldsymbol{st}_{1N} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ \boldsymbol{st}_{i1} & \cdots & \boldsymbol{st}_{ij} & \cdots & \boldsymbol{st}_{iN} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ \boldsymbol{st}_{N1} & \cdots & \boldsymbol{st}_{Nj} & \cdots & \boldsymbol{st}_{NN} \end{bmatrix}, \tag{4}$$

where $\boldsymbol{st}_{ij}$ represents coefficient at the location (i, j) int the sub-matrix of the STFT graph; The dimension N will be updated in the real GFD experiments in Section 7. Finally, the feature vectors from STFT are obtained and are given by

$$\boldsymbol{F} = \left\{\max(\boldsymbol{st}_{11}), \ldots, \max\left(\boldsymbol{st}_{ij}\right), \ldots, \max(\boldsymbol{st}_{NN})\right\}. \tag{5}$$

4 BIC-based Clustering Number Estimation

In most practical gear fault diagnosis applications, it is difficult to know the estimated number of clusters in advance, but the input clustering numbers always have direct influence in final clustering results. An optimal numbers estimation strategy based on the cluster validity indexes is found in historical literatures, including: the Calinski–Harabasz index [26], the Davies–Bouldin index [27], the weighted inter-intra index [28] and the in-group proportion index [29]. However, two drawbacks exist in the cluster validity indexes-based estimation strategy: (1) The estimation precision depends on selected clustering algorithm, dataset as well as the validity index. For instance, both using the IGP index, the affinity propagation (AP) method has more credible optimal clustering-number compared with the k-means strategy on account of the randomness of the latter [30]. (2) It is hard to apply the cluster validity indexes-based estimation strategy to 2D or higher dimension clustering like co-clustering because these validity indexes are mainly calculated according to the distance algorithms between different classes or within the same classes.

To find optimal numbers of co-clustering, an estimation algorithm based on Bayesian information criterion (BIC) is proposed. BIC is a statistical method which represents the descriptive power of a model to dataset [31], including: (1) the posterior likelihood of data estimation L; (2) The model complexity $|\Theta|$. The computational formula of BIC is given by

$$\mathrm{BIC} = \lambda L - \frac{1}{2}|\Theta|\log N, \tag{6}$$

where λ means the weight factor; N is the totality of samples. In clustering, the posterior likelihood of data estimation L is represented using the ratio of the mutual information entropy between after-clustering $\mathrm{I}(S^*; F^*)$ and before-clustering $\mathrm{I}(S; F)$. In Eq. (6), the entire meaning of L is that a good clustering must maintain original information entropy as possible as it can,

$$L = \mathrm{I}\left(S^*; F^*\right)/\mathrm{I}(S; F). \tag{7}$$

But in 2D co-clustering, the BIC model requires to take row and column clustering into consideration at the same time. So we redefine the parameters as follows:

Direction	Sample length	Sample size	Clustering number
Row	n	m	k
Column	m	n	l

According to these definitions, the BIC model complexity in co-clustering can be re-expressed as

$$|\Theta|_r \log m + |\Theta|_c \log n = nk\log m + ml\log n. \tag{8}$$

Substituting Eq. (7) and Eq. (8) into Eq. (6), we get Eq. (9) as follows:

$$\mathrm{BIC}(k, l) = \lambda \mathrm{I}\left(S^*; F^*\right)/\mathrm{I}(S; F) - \left(nk\log m + ml\log n\right)/2. \tag{9}$$

Further, this Bayesian information criterion can be extended to 3D field and is given by

$$\begin{aligned} \mathrm{BIC}(k, l, p) = {} & \lambda \mathrm{I}\left(S^*; F^*\right)/\mathrm{I}(S; F) \\ & - \left(nqk\log m + mql\log n + qp\log p\right)/3, \end{aligned} \tag{10}$$

where p is the clustering number of the 3rd classification; q is the size of the 3rd dimension. Based on the description of theory above, the details of BIC algorithm is given as:

Input:

a. For single variable working condition GFD, the sample matrix C is structured by

$$C_{40\times n} = \left[\begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{40} \end{bmatrix}_1 \cdots \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{40} \end{bmatrix}_i \cdots \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{40} \end{bmatrix}_n \right], \tag{11}$$

where $\begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{40} \end{bmatrix}^{\mathrm{T}}$ represents the STFT values in continuous approximate one minute, which depends on the change speed of working condition; $i \in \{1, 2, \ldots, n\}$ represents the random samples collected from different fault type and n is the number of sample.

b. For double variable working condition GFD, the sample matrix C is structured by

$$C_{n\times 40\times 40}=\begin{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{40} \end{bmatrix}_{11} & \cdots & \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{40} \end{bmatrix}_{i1} & \cdots & \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{40} \end{bmatrix}_{n1} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{40} \end{bmatrix}_{1j} & \cdots & \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{40} \end{bmatrix}_{ij} & \cdots & \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{40} \end{bmatrix}_{nj} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{40} \end{bmatrix}_{140} & \cdots & \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{40} \end{bmatrix}_{i40} & \cdots & \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{40} \end{bmatrix}_{n40} \end{bmatrix}. \tag{12}$$

Process:

a. Select an matrix construction method to build C matrix according to the number of running environments;
b. Initialize the row clustering number $k=1$, the column clustering number $l=1$, the weight factor $\lambda=0.5$;
c. For k from 1 to m, l from 1 to n:

$$\mathrm{BIC}(k,l)=\lambda \mathrm{I}\left(S^*;F^*\right)/\mathrm{I}(S;F)-\left(nk\log m+ml\log n\right)/2, \tag{13}$$

$$I(S;F)=\sum\nolimits_s\sum\nolimits_f p(s,f)\log_2\left[p(s,f)/p(s)p(f)\right], \tag{14}$$

where $p(s,f)$ means the joint probability distribution between row and column; $p(s)$ means the probability distribution in row vector; $p(f)$ means the probability distribution in column vector.

d. Search the max value of BIC during the whole k and l domain:

$$\arg\min_{k,l}\mathrm{BIC}(k,l). \tag{15}$$

Output:

a. For single variable working condition GFD, it outputs the number of final clustering numbers k and l;
b. For double variable working condition GFD, it outputs the number of final clustering numbers k, l, and p.

5 Modified NMF-based Co-clustering

5.1 NMF Theory

Non-negative factorization (NMF) is a kind of efficient data compression strategy, aiming to describe the high-dimensional data set using few base vectors with the help of the non-negative theory [32]. Different from the global characteristics of vector quantization (VQ) and principle component analysis (PCA) theory, NMF offers a good description about the local features, so specializing in searching the small scale information. Non-negative factorization has been investigated for feature extraction and recognition of rolling element bearing fault [33]. However, the unique advantage of co-clustering has not been explored for 2D or higher dimensions application. The basic NMF problem is stated as the following equation:

$$\boldsymbol{C}_{n\times m}\approx \boldsymbol{W}_{n\times d}\boldsymbol{H}_{d\times m}, \tag{16}$$

where $\boldsymbol{C}_{n\times m}$ is a $n\times m$ non-negative matrix; the basis matrix $\boldsymbol{W}$ and the gains matrix $\boldsymbol{H}$ are factorized from $\boldsymbol{C}_{n\times m}$; $d<(n\times m)/(n+m)$ represents the reduced rank. Therefore, $\boldsymbol{C}_{n\times m}$ can be linearly estimated by the sub-vectors $\boldsymbol{W}_{n\times d}$ and $\boldsymbol{H}_{d\times m}$. In order to obtain matrix $\boldsymbol{W}$ and $\boldsymbol{H}$, a large number of cost functions are defined to quantify the degree of approximation. In our strategy, the Euclidean distance is chosen as the cost function.

$$D(\boldsymbol{C}||\boldsymbol{WH})=\boldsymbol{C}-\boldsymbol{WH}^2. \tag{17}$$

The purpose of NMF is to find the $\boldsymbol{W}$ and $\boldsymbol{H}$, which possesses the smallest cost function: $\min_{\boldsymbol{W},\boldsymbol{H}} D(\boldsymbol{C}||\boldsymbol{WH})$ s.t., $\boldsymbol{W},\boldsymbol{H}\geq 0$. An iterative multiplicative algorithm is carried out based on the updated rule of $\boldsymbol{W}$ and $\boldsymbol{H}$, and is given by

$$\begin{aligned} \boldsymbol{H}_{r+1}&=\boldsymbol{H}_r\otimes\left[\left(\boldsymbol{C}\boldsymbol{H}_r^{\mathrm{T}}\right)\boldsymbol{\Theta}\left(\boldsymbol{W}_r\boldsymbol{H}_r\boldsymbol{H}_r^{\mathrm{T}}\right)\right],\\ \boldsymbol{W}_{r+1}&=\boldsymbol{W}_r\otimes\left[\left(\boldsymbol{W}_r^{\mathrm{T}}\boldsymbol{C}\right)\boldsymbol{\Theta}\left(\boldsymbol{W}_r^{\mathrm{T}}\boldsymbol{W}_r\boldsymbol{H}_r\right)\right], \end{aligned} \tag{18}$$

where $\otimes$ is the element-wise multiplication, $\boldsymbol{\Theta}$ is the element-wise division; r represents the iteration.

5.2 Classical NMF-based Co-clustering

Recently, the clustering application based on NMF has attracted much attention. Particularly, KIM, etc., explored the effective combination between cluster and NMF [34]. This paper extends its application from single cluster to co-clustering, aiming to solve the varying work condition or multi-tasks problem. Rely on the computational $\boldsymbol{W}\in\boldsymbol{R}^{n\times r}$ and $\boldsymbol{H}\in\boldsymbol{R}^{r\times m}$ above, two objective functions $\boldsymbol{J}_k$ & $\boldsymbol{J}_l$ are defined as follows:

$$\begin{aligned} \boldsymbol{J}_k&=\left\|\boldsymbol{W}-\boldsymbol{C}^r\boldsymbol{B}^r\right\|^2,\\ \boldsymbol{J}_l&=\left\|\boldsymbol{H}^{\mathrm{T}}-\boldsymbol{C}^c\boldsymbol{B}^c\right\|^2, \end{aligned} \tag{19}$$

where $\boldsymbol{C}^r=[\boldsymbol{c}_1,\boldsymbol{c}_2,\ldots,\boldsymbol{c}_k]^r\in\boldsymbol{R}^{m\times k}$ and $\boldsymbol{C}_2=[\boldsymbol{c}_1,\boldsymbol{c}_2,\ldots,\boldsymbol{c}_l]^c\in\boldsymbol{R}^{n\times l}$ represent the centroid matrix in row and column, respectively; The element $\boldsymbol{c}_j, j\in[1,2,\ldots,k]$ of $\boldsymbol{C}^r$ means the cluster centroid of the jth cluster in ***Task I***

and the element $\boldsymbol{c}_j, j \in [1, 2, \ldots, l]$ of $\boldsymbol{C}^c$ means the cluster centroid of the jth cluster in ***Task II***; $\boldsymbol{B}^r$ & $\boldsymbol{B}^c$ denote clustering assignment in ***Task I*** and ***Task II*** respectively. In ***Task I***, $\boldsymbol{B}^r_{ij} = 1$ means the ith sample belongs to the jth cluster, otherwise $\boldsymbol{B}^r_{ij} = 0$, and so is ***Task II***.

The purpose of co-clustering is to find sparse matrix $\boldsymbol{B}^r_{ij}$ and $\boldsymbol{B}^c$, which has only one in each row, with others being zero. Taking ***Task I*** and ***Task II*** as example, we redefine $\boldsymbol{C}^r$ and $\boldsymbol{C}^c$ as

$$\begin{aligned} \boldsymbol{C}^r &= \boldsymbol{W}\left(\boldsymbol{B}^r\right)^{\mathrm{T}}\left(\boldsymbol{D}^r\right)^{-1}, \\ \boldsymbol{C}^c &= \boldsymbol{H}^{\mathrm{T}}\left(\boldsymbol{B}^c\right)^{\mathrm{T}}\left(\boldsymbol{D}^c\right)^{-1}, \end{aligned} \tag{20}$$

where $\left(\boldsymbol{D}^r\right)^{-1} = \mathrm{diag}\left(|c_1|^{-r}, |c_2|^{-r}, \ldots, |c_k|^{-r}\right) \in \boldsymbol{R}^{k\times k}$, and $\left(\boldsymbol{D}^c\right)^{-1} = \mathrm{diag}\left(|c_1|^{-c}, |c_2|^{-c}, \ldots, |c_l|^{-c}\right) \in \boldsymbol{R}^{l\times l}$. Meanwhile, we set $\left(\boldsymbol{D}^r\right)^{-1} = \boldsymbol{D}^r_1\boldsymbol{D}^r_2$, $\left(\boldsymbol{D}^c\right)^{-1} = \boldsymbol{D}^c_1\boldsymbol{D}^c_2$, then Eq. (19) can be expressed as follows:

$$\begin{aligned} \boldsymbol{J}_k &= \left\| \boldsymbol{W} - \boldsymbol{W}\left(\boldsymbol{B}^r\right)^{\mathrm{T}}\left(\boldsymbol{D}^r\right)^{-1}\boldsymbol{B}^r \right\|^2 \\ &= \left\| \boldsymbol{W} - \boldsymbol{W}\left(\boldsymbol{B}^r\right)^{\mathrm{T}}\boldsymbol{D}^r_1\boldsymbol{D}^r_2\boldsymbol{B}^r \right\|^2 \\ &= \left\| \boldsymbol{W} - \boldsymbol{W}\left(\boldsymbol{M}^r\right)^{\mathrm{T}}\boldsymbol{N}^r \right\|^2, \end{aligned} \tag{21}$$

$$\begin{aligned} \boldsymbol{J}_l &= \left\| \boldsymbol{H}^{\mathrm{T}} - \boldsymbol{H}^{\mathrm{T}}\left(\boldsymbol{B}^c\right)^{\mathrm{T}}\left(\boldsymbol{D}^c\right)^{-1}\boldsymbol{B}^c \right\|^2 \\ &= \left\| \boldsymbol{H}^{\mathrm{T}} - \boldsymbol{H}^{\mathrm{T}}\left(\boldsymbol{B}^c\right)^{\mathrm{T}}\boldsymbol{D}^c_1\boldsymbol{D}^c_2\boldsymbol{B}^c \right\|^2 \\ &= \left\| \boldsymbol{H}^{\mathrm{T}} - \boldsymbol{H}^{\mathrm{T}}\left(\boldsymbol{M}^c\right)^{\mathrm{T}}\boldsymbol{N}^c \right\|^2, \end{aligned} \tag{22}$$

where $\boldsymbol{M}^r = \left(\boldsymbol{D}^r_1\right)^{\mathrm{T}}\boldsymbol{B}^r$, $\boldsymbol{N}^r = \left(\boldsymbol{D}^r_2\right)^{\mathrm{T}}\boldsymbol{B}^r$, $\boldsymbol{M}^c = \left(\boldsymbol{D}^c_1\right)^{\mathrm{T}}\boldsymbol{B}^c$, $\boldsymbol{N}^c = \left(\boldsymbol{D}^c_2\right)^{\mathrm{T}}\boldsymbol{B}^c$.

Finally, a second order NMF is applied in $\boldsymbol{J}_k$ and $\boldsymbol{J}_l$, aiming to factorize $\boldsymbol{W}$ to $\boldsymbol{W}(\boldsymbol{M}^r)^{\mathrm{T}}$ and $\boldsymbol{N}^r$, to factorize $\boldsymbol{H}^{\mathrm{T}}$ to $\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{M}^c)^{\mathrm{T}}$ and $\boldsymbol{N}^c$. Therefore, the $\boldsymbol{B}^r$ and $\boldsymbol{B}^c$ matrix is obtained according to the second order NMF result. After that, the classifications in row and column are obtained from the $\boldsymbol{B}^r$ and $\boldsymbol{B}^c$ matrix.

$$\begin{cases} L^r_i = j, \boldsymbol{B}^r_{ij} = 1, \\ L^r_i \neq j, \boldsymbol{B}^r_{ij} = 0, \end{cases} \tag{23}$$

$$\begin{cases} L^c_i = j, \boldsymbol{B}^c_{ij} = 1, \\ L^c_i \neq j, \boldsymbol{B}^c_{ij} = 0. \end{cases}$$

5.3 Modified NMF-based Co-clustering

As described in Section 5.1, non-negative matrix $\boldsymbol{C}$ is factorized into two sub-matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ in conventional non-negative factorization. Although the physical meanings of $\boldsymbol{W}$ and $\boldsymbol{H}$ are clear: they represent the decomposition values in row and column respectively and promote the classification effect of co-clustering, the relation between two directions is still ill-defined. Hence, the typical NMF is improved and is given by

$$\boldsymbol{C}_{n\times m} \approx \boldsymbol{W}_{n\times k}\boldsymbol{L}_{k\times l}\boldsymbol{H}_{l\times m}, \tag{24}$$

where $\boldsymbol{L}_{k\times l}$ is named as 'the relation matrix', the value $\boldsymbol{L}_{ij}$ represents the link between the ith cluster in ***Task I*** and the jth cluster in ***Task II***; k is the clustering number in the row vector and l is the clustering number in the column vector. In modified NMF, the cost function and the update functions can be re-written as

$$D(\boldsymbol{C}||\boldsymbol{WLH}) = \boldsymbol{C} - \boldsymbol{WLH}^2, \tag{25}$$

$$\boldsymbol{H}_{r+1} = \boldsymbol{H}_r \otimes \left[\left(\boldsymbol{C}\boldsymbol{H}^{\mathrm{T}}_r\right)\boldsymbol{\Theta}\left(\boldsymbol{W}_r\boldsymbol{L}_r\boldsymbol{H}_r\boldsymbol{H}^{\mathrm{T}}_r\right)\right], \tag{26}$$

$$\boldsymbol{W}_{r+1} = \boldsymbol{W}_r \otimes \left[\left(\boldsymbol{W}^{\mathrm{T}}_r\boldsymbol{C}\right)\boldsymbol{\Theta}\left(\boldsymbol{W}^{\mathrm{T}}_r\boldsymbol{W}_r\boldsymbol{L}_r\boldsymbol{H}_r\right)\right], \tag{27}$$

$$\boldsymbol{L}_{r+1} = \boldsymbol{W}^{+}_{r+1}\boldsymbol{C}\boldsymbol{H}^{+}_{r+1}, \tag{28}$$

where $\otimes$ is the element-wise multiplication, $\boldsymbol{\Theta}$ is the element-wise division; r represents the iteration; $\boldsymbol{W}^+$ is the generalized inverse of $\boldsymbol{W}$: $\boldsymbol{W}\boldsymbol{W}^+\boldsymbol{W} = \boldsymbol{W}$; $\boldsymbol{H}^+$ is the generalized inverse of $\boldsymbol{H}$: $\boldsymbol{H}\boldsymbol{H}^+\boldsymbol{H} = \boldsymbol{H}$.

By introducing the matrix $\boldsymbol{L}$, the $\boldsymbol{W}$ and $\boldsymbol{H}$ are not required to be orthogonal in modified NMF strategy. Therefore, it expands the optional range of $\boldsymbol{W}$ and $\boldsymbol{H}$ and improves the factorization performance, which will be proved in Section 7.

6 GA-based Parameter Regulator

Three parameters need to be designed in the co-clustering-based GFD strategy, including: (1) the feature-dimension N in STFT; (2) the weight factor λ in BIC algorithm; (3) the transitional dimension d in traditional NMF-based co-clustering, which are listed in Table 2.

Among these three parameters, the feature-dimension N in STFT can be decided from the GFD experiments. λ and d will be adjusted using the gradient ascent (GA) regulatory mechanism [35, 36], whose fundamental is shown in Figure 6. The main idea of gradient ascent algorithm is to follow the fastest changing direction to find the

Table 2 Three main parameters

Symbol	Descriptions	Range
N	The feature-dimension in STFT	$N \leq 80$
λ	The weight factor in BIC	$0 \leq \lambda \leq 1$
d	The transitional dimension	$d \leq \min(m, n)$

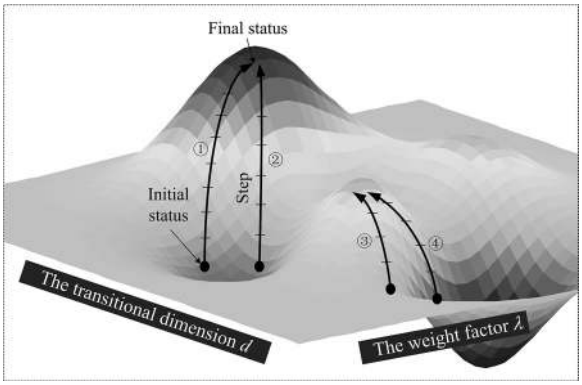

Figure 6 Fundamental of gradient ascent

maximum of diagnostic accuracy. In Figure 6, four different initial points of (λ, d) are listed as example, where the approximation curves ① and ② reach the global optimum while ③ and ④ are limited in the local optimum.

Meanwhile, three concepts are included in the gradient ascent model:

- The step length *sl*: it represents the speed along the gradient direction during the iteration. We initialize the step length as 0.02 in the parameter λ & $d/\min(m,n)$, where $0 \le \lambda \le 1$, $0 < d/\min(m,n) \le 1$;
- The learning function: it has been designed in NMF-based co-clustering classifier as:

$$\left\{L_i^r, L_i^c\right\} = NMF\left\{\boldsymbol{x}_i, \lambda, [d/\min(m,n)]\right\}, \tag{29}$$

where $\boldsymbol{x}_i$ represents extracted STFT feature vector from the ith sample; $NMF(\cdot)$ means the NMF-based co-clustering classifier, with three input parameters $\left\{\boldsymbol{x}_i, \lambda, [d/\min(m,n)]\right\}$; L_i^r and L_i^c means the clustering results of the ith sample in row and column.

- The validity function: it is calculated by the sum of correct classifications, and it assesses the effectiveness of classification.

$$\begin{aligned} &Ac(\lambda, [d/\min(m,n)]) \\ &= \sum_{i=1}^{m} zer\left(L_i^r - y_i^r\right) + \sum_{i=1}^{n} zer\left(L_i^c - y_i^c\right), \end{aligned} \tag{30}$$

$$zer(x) = \begin{cases} 0, x \neq 0, \\ 1, x = 0, \end{cases} \tag{31}$$

where y_i^r and y_i^c represents the label of the ith sample; $zer(\cdot)$ means the zero sign function. Notice that, the validity function can only be obtained in those training samples, whose classification labels are known. The optimal λ and $d/\min(m,n)$ is gained according to the training samples and is used in others, called testing samples.

It should be noted that in real GFD application, when performing GA algorithm: (1) If the step length is too large, the optimal parameter result might be skipped. But if the step length is too small, the iteration speed will be slow and cause too large computational load. (2) It is easy for the GA algorithm to be deep in the local optimum rather than the global optimum, which relies on the initial location of $\left\{\lambda, d/\min(m,n)\right\}$. Therefore, it is necessary to take these two factors into consideration to balance the computational accuracy and the time consumption.

7 Experiments and Performance Analysis

7.1 DDS Experimental System

The Spectra Quest's Drivetrain Dynamics Simulator (DDS) was used in this study for experimental verification, as shown in Figure 7. This system is composed of six units including: (1) speed regulator; (2) the driving motor; (3) the planetary gearbox; (4) the reduction gearbox; (5) brake device; (6) brake regulator. The faults occur in those gears in planetary & reduction gearboxes, under varying rotating speed and load conditions, which are adjusted using the speed regulator and the brake regulator, respectively. Four types of gear faults are studied: (1) root cracks; (2) missing teeth; (3) chipped teeth; (4) surface wear. The purpose of GFD is to classify these faults through 7 vibration sensors (3-axis for planetary gearbox; 3-axis for reduction gearbox; 1-axis for driving motor). In addition, in order to put the co-clustering methods into effect, we define different levels in four task sets listed in Table 3, including: (1) the fault type task (C1–C5); (2) the fault severity task (D1–D4); 3) the speed regulator task (E1–E5); (4) the load regulator task (F1–F5). Specially, the rotating speed curve and the load curve in Figure 8 was also conducted.

7.2 GFD Experiments and Performance Analysis

7.2.1 Experimental Setup

The varying rotating speed and load are designed using the regulator curves in Figure 8 for the experiment. In order to enlarge the data analytical ability of algorithms, 10 repeat collections were implemented to increase the sample points to 10 times ($5120 \times 50 \times 10$) in each group, which are segmented by the 2.5 s sub-signals. Therefore, the sample number for fault type recognition can be gained in row clustering task (fault type) and column clustering tasks (rotating speed and load), and are listed in Table 4A and 4B.

In the fault type recognition experiments, several tests are compared using the models as follows: (1) X-means clustering; (2) Guassian Mixture Model (GMM)

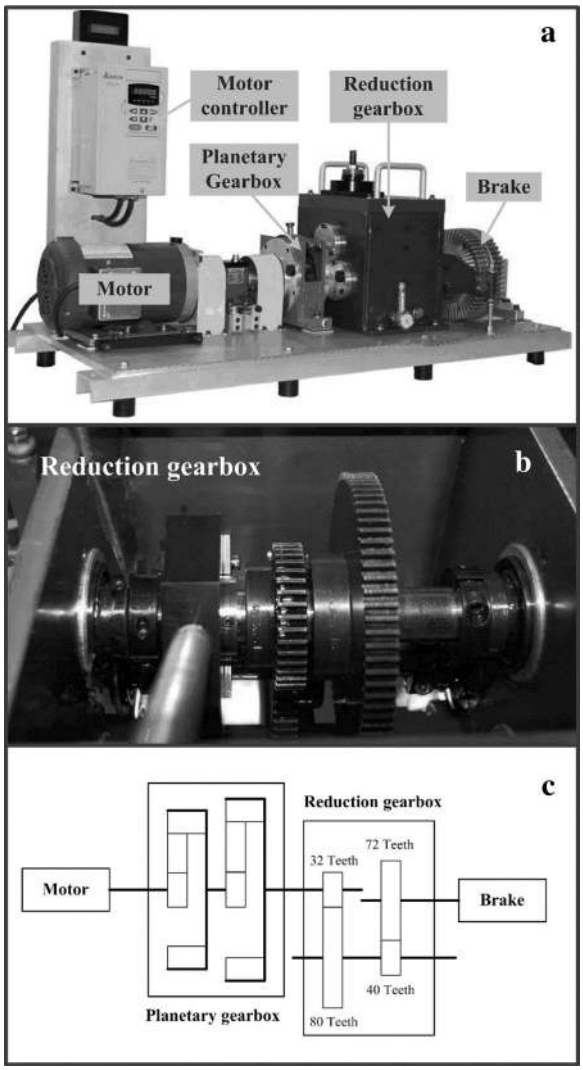

Figure 7 The Drivetrain Dynamics Simulator (DDS) system: **a** the whole physical map of DDS system; **b** the internal structure image of DDS system; **c** transmission ratio of reduction gearbox

Table 3 Define levels in different tasks

Task I	C1	C2	C3	C4	C5
Fault type	Health	Root	Missing	Chipped	Surface
Task II	D1	D2	D3	D4	
Fault severity	Health	Slight	Medium	Heavy	
Task III	E1	E2	E3	E4	E5
Speed regulator (Hz)	< 5	5–15	15–25	25–35	> 35
Task IV	F1	F2	F3	F4	F5
Torque regulator (N·m)	< 1.83	1.83–5.49	5.49–9.14	9.14–12.80	> 12.80

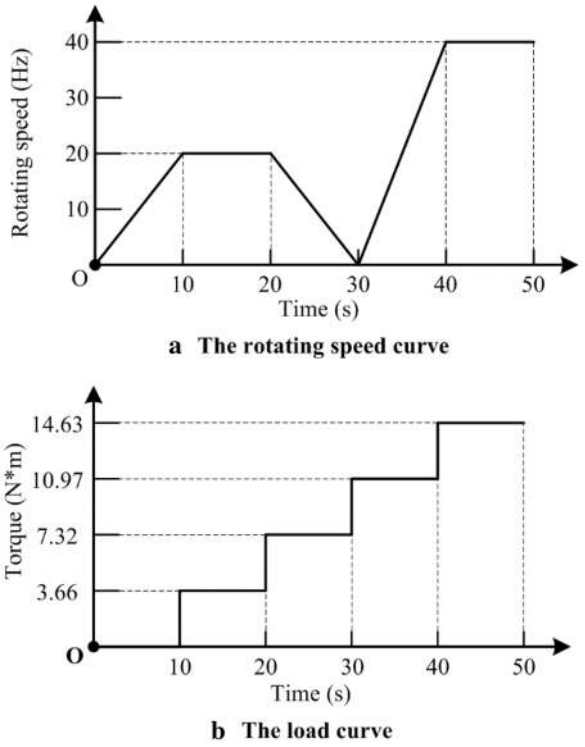

Figure 8 The varying working condition curves

clustering; (3) NMF-based co-clustering; (4) Modified NMF-based co-clustering. For co-clustering, the related parameters were set as: $\lambda = 1, d = 30$.

7.2.2 Experimental Results and Discussion

Experiments are carried out, and the column task in Table 4 is considered to improve the effectiveness of row task in co-clustering method. In order to assess the quality of models, the Pr (C_i) and Re (C_i) indexes of clustering results are calculated and given by

$$\Pr(C_i) = \frac{\text{Corrected classfied } C_i \text{ samples}}{\text{All } C_i \text{ samples}}, \tag{32}$$

$$\text{Re}(C_i) = \frac{\text{Corrected classfied } C_i \text{ samples}}{\text{All classfied } C_i \text{ samples}}, \tag{33}$$

where the Pr (C_i) index reflects the ability that a model identifies correct samples while the Re (C_i) index reflects the ability that a model finds all correct samples.

At first, we observed the fault type recognition results using the NMF-based co-clustering strategy under varying rotating speed and load environment, which are listed in Table 5. This table indicates that the diagnostic accuracy of co-clustering strategy in various load conditions (97.8%) is better than that in various rotating speeds

Table 4 Sample number for (A) rotating speed fault recognition and (B) load fault recognition

A	Column task (rotating speed)					
		E1	**E2**	**E3**	**E4**	**E5**
Row task (fault type)	C1	25	50	70	10	45
	C2	25	50	70	10	45
	C3	25	50	70	10	45
	C4	25	50	70	10	45
	C5	25	50	70	10	45
B	**Column task (load)**					
		F1	**F2**	**F3**	**F4**	**F5**
Row task (fault type)	C1	40	40	40	40	40
	C2	40	40	40	40	40
	C3	40	40	40	40	40
	C4	40	40	40	40	40
	C5	40	40	40	40	40

Table 5 NMF-based co-clustering results for fault type recognition ($k = 6$)

Rotating speed	Pr (%)	Re (%)	Load	Pr (%)	Re (%)
C1	92.0	99.5	C1	97.0	100
C2	95.0	99.3	C2	94.5	100
C3	96.5	93.7	C3	98.0	94.2
C4	98.5	97.5	C4	99.5	98.0
C5	99.5	93.0	C5	100	97.1
Total	96.3	96.5	Total	97.8	97.2

(96.3%). This can be explained by two possible reasons: (1) The classify boundary of the latter is more clear than the former since that the rotating speed presents a gradual change characteristic from 0 Hz to 40 Hz while the brake load jumps from 0 to 14.63 N·m; (2) Changing the rotating speed has a stronger interference effect than just changing the load to collected vibration signal, which have been verified in our previous study. Secondly, to illustrate clustering performance in the varying rotating speed model, the NMF-based co-clustering results are listed in Table 6 ($k=6$; $l=6$) and Figure 9, where 200 samples are tested in each category. Some details can be seen here: (1) the misclassification cases always appear between 'Health' and 'Surface' or between 'Root', 'Missing' and 'Chipped' because the time domain features of the former are more similar but the frequency domain features of the latter are alike. For example, the 'Chipped' type is easy to be classified as the 'Missing' type if the crack of chipped tooth is large enough; (2) The 6th category (R2 and R3) occurs in the C2 type when the number of clustering is set as 6, which means the discrete ability and the inconsistency exists inside the 'Root' samples. Interestingly, although the differences exist in local precision and recall index of different categories, the total precision and recall are very nearly the same in these two tables.

Table 6 Fault type classification details under varying rotating speed ($k = 6$)

Rotating speed	C1	C2	C3	C4	C5	Pr (%)	Re (%)
R1	184				1	92.0	99.5
R2	1	138				95.0	99.3
R3		52	2				
R4		10	193	3		96.5	93.7
R5			5	197		98.5	97.5
R6	15				199	99.5	93.0
Total						96.3	96.5

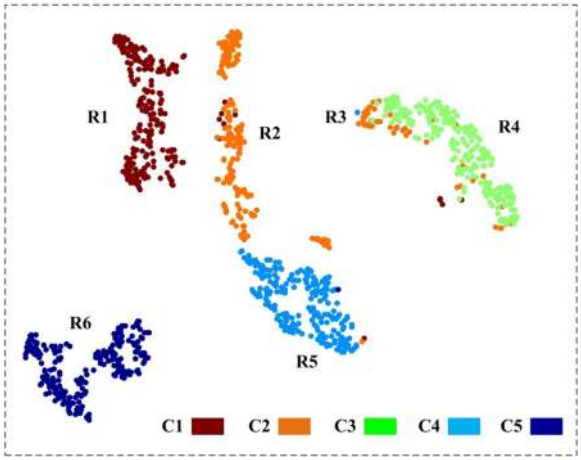

Figure 9 The scatter plot of the classification results ($k=6$)

Especially, we reclassify the fault types of testing samples in Table 4 using the modified NMF-based co-clustering strategy. Here the normalized relation matrix $\boldsymbol{L}_{k\times l}$ ($k=l=5$) can be gained and given by

Table 4A:	0.0801	0.0918	0.2563	0.0965	0.3236
	0.3285	3.7987	3.8255	3.2661	3.5314
	0.2155	**6.9619**	**6.5298**	**6.6931**	**7.0714**
	0.7031	**6.8263**	**7.2954**	**6.7465**	**7.2176**
	0.2250	0.1191	0.0545	0.2875	0.4563
Table 4B:	3.7887	3.5302	4.1173	2.1937	2.4488
	3.7157	0.1592	3.4741	1.9078	2.2279
	1.9611	1.3846	1.5855	3.8276	3.2316
	3.2774	0.2309	4.7511	3.9760	3.5468
	0.8559	0.4857	0.1722	0.9344	3.7734

Notice that, the element l_{ij} represents the relevance between the ith category in row task and the jth category in column task. In the $\boldsymbol{L}_{k\times l}$ matrix, the l_{ij} rises up to approximate 7 in 'Missing' & 'Chipped' fault type with the increase of rotating speed. The effect of speed regulator on different fault types presents the following rank: Missing≈Chipped>Root>Health≈Surface. However, there is a less link between load and fault types, seeing from the $\boldsymbol{L}_{k\times l}$ matrix under changing load environment. In order to further verify the necessity for modified NMF, Table 7 gives the evaluation indexes of this approach. It shows that an improved precision occurs in the varying rotating speed dataset but makes no difference in the varying load dataset. That is because the $\boldsymbol{L}_{k\times l}$ matrix in modified strategy cuts off the link between $\boldsymbol{W}$ and $\boldsymbol{H}$ in the former, which promotes the separation ability between row and column task. As can be seen in Table 7, the middle dimension of $\boldsymbol{W}$ and $\boldsymbol{H}$ is not limited in a single value, like traditional NMF method does, thus improving both the flexibility of selected dimension and the GFD precision (97.0% and 97.8%). Also, it can be known that the recognition performance of various loads (100%) is superior to rotating speed (95.3%) on account of the continuity of speed regulator.

Table 7 Evaluation indexes of modified co-clustering ($k=6$)

Experiments	Pr (%)	Re (%)
Varying rotating speed		
Row task (C1–C5)	97.0	97.2
Column task (E1–E5)	95.3	93.5
Varying load		
Row task (C1–C5)	97.8	97.5
Column task (F1–F5)	100	100

Finally, concentrating on the varying working condition GFD, the traditional clustering strategies, including the K-means and the GMM methods, and co-clustering approaches were compared using two selected indexes: the precision and the time consumption. The performance comparison results are listed in Table 8. According to this table, although the time complexity of single algorithm of 1D clustering is smaller than joint strategy (4.923 s < 7.991 s), the accumulation of computational load for two tasks is larger than co-clustering proposed (4.923 s + 4.856 s > 7.991 s). Meanwhile, the co-clustering have an apparent precision increase in varying working condition GFD, about 12.51% in varying rotating speed and 7.00% in varying load. Therefore, this experiment proves the superiority of co-clustering in gear fault diagnosis under variable working conditions.

7.3 Parameter Regulation Experiments

7.3.1 STFT Dimension Adjustment Experiments

During the STFT dimension adjustment experiments, we adjusted the feature dimension N from 10 to 80 one by one, and then the diagnostic precisions of Task I as well as the time consumptions of co-clustering model were observed and were drawn in Figure 10. It can be seen that the diagnostic accuracy increases from 78.76% to 97.54% when the feature dimension N increases from 10 to 80, meanwhile, the computational load indicates an exponential increase from approximately 16.78 s to 48.45 s. It can be seen from Figure 10 that $N=42$ is considered as an appropriate dimension, in which the diagnosis accuracy is satisfactory enough (97.02%), while the time consumption keeps at a low level (25.12 s). Although the precision will continue to improve up to 97.54% if we

Table 8 Performance comparison of different GFD models

Model	Varying rotating speed		Varying load	
	Row clustering precision (%)	Time consumption (s)	Row clustering precision (%)	Time consumption (s)
X-means	84.1	4.923 + 4.856	90.5	4.919 + 4.774
GMM methods	87.7	6.845 + 6.018	92.3	6.274 + 5.909
NMF	96.3	7.991	97.8	6.845
Modified NMF	97.0	9.362	97.8	8.647

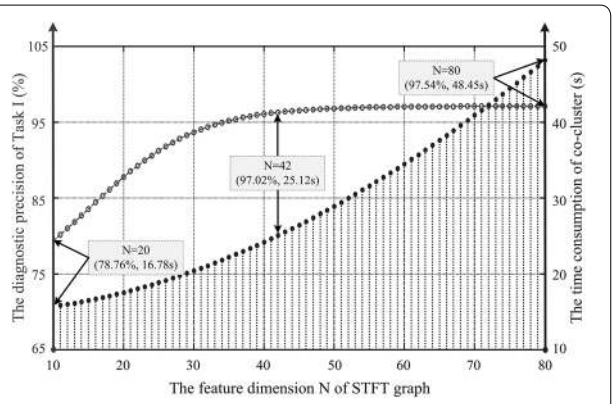

Figure 10 Diagnostic precision and time consumptions of co-clustering model

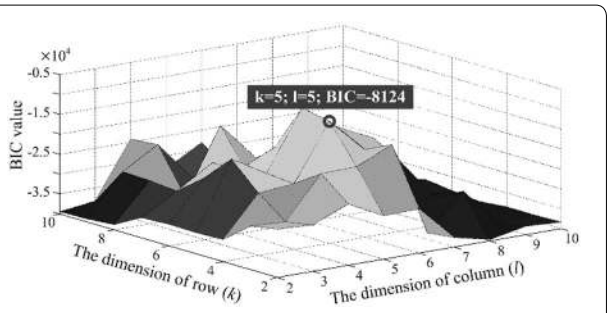

Figure 11 Varying load BIC results

continuously increase the feature dimension, but computational cost will also increase quickly.

7.3.2 BIC Algorithm Experiments

Based on the dataset from Drivetrain Dynamics Simulator (DDS) system, we tried to adjust the number of cluster in row and column, respectively. Generally, the search range of clustering number is from 2 to 10: $2 \leq k \leq 10$; $2 \leq l \leq 10$. Figure 11 illustrates an example of the BIC results in varying load GFD experiments when the clustering number changes from 2 to 10. It can be seen that the peak of BIC value (−8124) exists at the point (5,5), which means that the optimal co-clustering results happens when the row and the column clustering numbers both equal to five. According to the real dataset and the standard labels in Table 3, the BIC-based estimation algorithm satisfies the requirements of practical gear fault diagnosis applications.

For further study on the BIC estimation algorithm, the BIC method was compared with a kind of traditional self-adapting classification estimation algorithm: X-means, which is an improved strategy of K-means. The estimations of clustering number in BIC as well as X-means algorithm are listed in Table 9. On one hand, for the BIC algorithm, with the increase of clustering number k, the diagnostic precision presents an increasing trend from 60.0% to 96.9% first, then it begins to slightly decrease when the clustering number k is larger than 5. On the other hand, by comparing the estimation results between X-means and BIC algorithm, we find the estimation of clustering number using the X-means is much larger than the normal ($k = 12$) and also its diagnostic precision is only 84.1%, which proves the superiority of BIC-based clustering number estimation.

7.3.3 GA Parameter Regulator Experiments

As shown in Figure 4, the gradient ascent algorithm was used in the varying working condition datasets to obtain the weight factor λ and the transitional dimension d. The regulator results of the gradient ascent algorithm are listed in Table 10. Here the final (λ, d), the number of iterations as well as the precision of row task are observed using 4 initial (λ, d) values: (0.5, 500), (0.5, 200), (0.7, 500), (0.7, 200). After gradient ascent, it could be seen that four initial values all reached the global optimum in varying load experiments. But in varying rotating speed experiments, only two points reach the global optimum, which achieves 99.3% diagnostic accuracy. In addition, it can be found that the less iterations occur when the distance between initial and final (λ, d) point is short. Therefore, these tests prove that the performance of GA algorithm depends on the selected initial point to a great extent.

Table 9 Estimation results of X-means and BIC algorithm

Algorithm	Clustering number	Sub-clustering number of row task					Precision of row task (%)
		Health	Root	Missing	Chipped	Surface	
BIC & co-clustering	$k=3$			1	1	1	60.0
	$k=4$	1		1	1	1	79.8
	$k=5$	1	1	1	1	1	96.9
	$k=6$	1	2	1	1	1	96.3
	$k=7$	1	2	1	1	2	95.1
	$k=8$	1	2	2	1	2	94.0
	$k=9$	1	3	2	1	2	90.4
X-means	$k=12$	2	3	3	2	2	84.1

Table 10 Regulator results of gradient ascent algorithm

Dataset	Initial (λ, d)	Number of iterations	Final (λ, d)	Precision of row task (%)
Varying rotating speed	(0.5, 500)	21	(0.46, 156)	99.3
	(0.5, 200)	6	(0.46, 158)	99.3
	(0.7, 500)	8	(0.80, 406)	90.5
	(0.7, 200)	14	(0.80, 411)	91.4
Varying load	(0.5, 500)	14	(0.60, 296)	97.0
	(0.5, 200)	9	(0.60, 285)	96.9
	(0.7, 500)	13	(0.64, 296)	97.0
	(0.7, 200)	8	(0.64, 279)	96.3

8 Conclusions

A NMF-theoretic co-clustering strategy is presented in this paper to offer a fast multi-tasking solution to solve the gear fault diagnosis problem under variable working conditions. Here the time-frequency features are extracted from the STFT spectrogram, and are utilized to structure the 2D matrix for joint clustering. Experiments indicate that 97.02% diagnostic precision can be achieved when the STFT dimension is set as 42. Meanwhile, seeing from the results of the BIC-based optimal clustering number estimation, they are close to the practical categories, no matter in varying rotating speed or varying load dataset. After NMF, row and column clustering task can be identified at the same time, with approximately 10% improved accuracy and less time cost compared with those single task clustering algorithms, such as X-means and GMM algorithm. There is an internal connection in most of gear failure signals. The proposed co-clustering strategy has better performance than independent clustering strategy because the modified NMF helps to provide a relation matrix, which shows a strong correlation between different rotating speeds and fault types. Therefore, the NMF-based co-clustering has a good potential to apply in the gear fault diagnosis of large-scale rotating machines under varying working conditions. In the future, co-clustering with higher dimension will probably apply in the more complex working conditions or more diagnostic tasks to improve the gear fault diagnosis performance.

Abbreviations

GFD: gear fault diagnosis; NMF: non-negative matrix factorization; STFT: short-time Fourier transform; BIC: Bayesian information criterion; GA: gradient ascent; DDS: Drivetrain Dynamics Simulator.

Authors' Contributions

RY designed the experiment, FS and CC analyzed the data. All authors read and approved the final manuscript.

Authors' Information

Fei Shen received his B.Sc. and M.Sc. degree from *Southeast University*, *China*, in 2014 and 2016 respectively. Now he is pursuing his PhD degree at *School of Instrument Science and Engineering, Southeast University, China*. His main research interest is machine fault diagnosis.

Chao Chen received his B.Sc. and M.Sc. degree from *Jiangsu University*, *China*, in 2011 and 2014 respectively. Now he is pursuing his PhD degree at *School of Instrument Science and Engineering, Southeast University*, *China*. His main research interest is machine fault diagnosis.

Jiawen Xu is currently an associate researcher at *School of Instrument Science and Engineering, Southeast University*, *China*.

Ruqiang Yan received his B.Sc. and M.E. degree from *University of Science and Technology of China* in 1997 and 2002 respectively, and received his Ph.D. degree in 2007 from *University of Massachusetts, USA*. Now he is a professor and a Ph.D. supervisor at *Xi'an Jiaotong University*, *China*. His main research

interests include machine condition monitoring and fault diagnosis, signal processing, and wireless sensor networks.

Funding
Supported by National Natural Science Foundation of China (Grant No. 51575102) and Jiangsu Postgraduate Research Innovation Program (Grant No. KYCX18_0075).

Competing Interests
The authors declare no competing financial interests.

Author Details
[1] School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China. [2] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China.

Received: 6 March 2019 Revised: 18 January 2020 Accepted: 10 February 2020
Published online: 26 February 2020

References

[1] G B Wang, Z J He, X F Chen. Basic research on machinery fault diagnosis-what is the prescription. *Journal of Mechanical Engineering*, 2013, 49(1): 63–72. (in Chinese)
[2] S P Yang, Z H Zhao. Improved wavelet denoising using neighboring coefficients and its application to machinery fault diagnosis. *Journal of Mechanical Engineering*, 2013, 49(17): 137–141. (in Chinese)
[3] X H Jin, Y Sun, J H Shan, et al. Fault diagnosis and prognosis for wind turbines: An overview. *Chinese Journal of Scientific Instrument*, 2017, 38(5): 1041–1053.
[4] MMM Islam, J Kim, SA Khan, et al. Reliable bearing fault diagnosis using Bayesian inference-based multi-class support vector machines. *The Journal of the Acoustical Society of America*, 2017, 141(2): EL89–EL95.
[5] A Singh, A Parey. Gearbox fault diagnosis under fluctuating load conditions with independent angular re-sampling technique, continuous wavelet transform and multilayer perceptron neural network. *IET Science, Measurement & Technology*, 2017, 11(2): 220–225.
[6] S H Kia, H Henao, G A Capolino. Fault index statistical study for gear fault detection using stator current space vector analysis. *IEEE Transactions on Industry Applications*, 2016, 52(6): 4781–4788.
[7] R Yan, RX Gao, X Chen. Wavelets for fault diagnosis of rotary machines: A review with applications. *Signal Processing*, 2014, 96(PART A): 1–15.
[8] C Chen, F Shen, R Q Yan. Enhanced least squares support vector machine-based transfer learning strategy for bearing fault diagnosis. *Chinese Journal of Scientific Instrument*, 2017, 38(1): 33–40.
[9] M Yuwono, Y Qin, J Zhou, et al. Automatic bearing fault diagnosis using particle swarm clustering and Hidden Markov Model. *Engineering Applications of Artificial Intelligence*, 2016, 47: 88–100.
[10] F Pacheco, M Cerrada, R V Sánchez, et al. Attribute clustering using rough set theory for feature selection in fault severity classification of rotating machinery. *Expert Systems with Applications*, 2017, 71: 69–86.
[11] B Mirkin. Mathematical classification and clustering. *Journal of the Operational Research Society*, 1997, 48(8): 852.
[12] A Tanay, R Sharan, M Kupiec, et al. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences*, 2004, 101(9): 2981–2986.
[13] M Ailem, F Role, M Nadif. Graph modularity maximization as an effective method for co-clustering text data. *Knowledge-Based Systems*, 2016, 109: 160–173.
[14] S Schmidt, S Schnitzer, C Rensing. Text classification based filters for a domain-specific search engine. *Computers in Industry*, 2016, 78: 70–79.
[15] Y Xu, V Olman, D Xu. Clustering gene expression data using a graph-theoretic approach: An application of minimum spanning trees. *Bioinformatics*, 2002, 18(4): 536–545.
[16] S Javadi, S M Hashemy, K Mohammadi, et al. Classification of aquifer vulnerability using K-means cluster analysis. *Journal of Hydrology*, 2017, 549: 27–37.
[17] C Xu, P L Zhang, G Q Ren, et al. Engine wear fault diagnosis based on improved semi-supervised fuzzy C-means clustering. *Journal of Mechanical Engineering*, 2011, 47(17): 55–60.
[18] J Q Zhang, G J Sun, L Li, et al. Study on mechanical fault diagnosis method based on LMD approximate entropy and fuzzy C-means clustering. *Chinese Journal of Scientific Instrument*, 2013, 34(3): 714–720.
[19] A J Gallego, J Calvo-Zaragoza, J J Valero-Mas, et al. Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation. *Pattern Recognition*, 2018, 74: 531–543.
[20] T T Van, T M Le. Content-based image retrieval based on binary signatures cluster graph. *Expert Systems*, 2017: e12220.
[21] D Fitrianah, A N Hidayanto, H Fahmi, et al. ST-AGRID: A spatio temporal grid density based clustering and its application for determining the potential fishing zones. *International Journal of Software Engineering and its Applications*, 2015, 9(1): 13–26.
[22] A H Pilevar, M Sukumar. GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases. *Pattern Recognition Letters*, 2005, 26(7): 999–1010.
[23] C F Tsai, C W Tsai, H C Wu, et al. ACODF: A novel data clustering approach for data mining in large databases. *Journal of Systems and Software*, 2004, 73(1 SPEC. ISS.): 133–145.
[24] K Y Chen, L S Chen, M C Chen, et al. Using SVM based method for equipment fault detection in a thermal power plant. *Computers in Industry*, 2011, 62(1): 42–50.
[25] R E Precup, P Angelov, B S J Costa, et al. An overview on fault diagnosis and nature-inspired optimal control of industrial process applications. *Computers in Industry*, 2015, 74: 1–16.
[26] M Arumugam, J Raes, E Pelletier, et al. Enterotypes of the human gut microbiome. *Nature*, 2011, 473(7346): 174–180.
[27] R Bhola, N H Krishna, K N Ramesh, et al. Detection of the power lines in UAV remote sensed images using spectral-spatial methods. *Journal of Environmental Management*, 2018, 206: 1233–1242.
[28] F Gasperini, J M Forbes, E N Doornbos, et al. Wave coupling between the lower and middle thermosphere as viewed from TIMED and GOCE. *Journal of Geophysical Research A: Space Physics*, 2015, 120(7): 5788–5804.
[29] L Wang, Y Zhang, S Zhong. Typical process discovery based on affinity propagation. *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, 2016, 10(1): JAMDSM0001–JAMDSM0001.
[30] W Zhang, X Wu, W P Zhu, et al. Unsupervized image clustering with SIFT-based soft-matching affinity propagation. *IEEE Signal Processing Letters*, 2017, 24(4): 461–464.
[31] A Barcaru, H G J Mol, M Tienstra, et al. Bayesian approach to peak deconvolution and library search for high resolution gas chromatography – Mass spectrometry. *Analytica Chimica Acta*, 2017, 983: 76–90.
[32] J K Liu, H M Schreyer, A Onken, et al. Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization. *Nature Communications*, 2017, 8:149, https://doi.org/10.1038/s41467-017-00156-9.
[33] Yuguang Niu, Shilin Wang, Ming Du. A combined Markov chain model and generalized projection nonnegative matrix factorization approach for fault diagnosis. *Mathematical Problems in Engineering*, 2017, 2017: 1–7.
[34] H Kim, H Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 2007, 23(12): 1495–1502.
[35] A V Gorshkov, T Calarco, M D Lukin, et al. Photon storage in -type optically dense atomic media. IV. Optimal control using gradient ascent. *Physical Review A - Atomic, Molecular, and Optical Physics*, 2008, 77(4).
[36] G Rancan, T T Nguyen, S J Glaser. Gradient ascent pulse engineering for rapid exchange saturation transfer. *Journal of Magnetic Resonance*, 2015, 252: 1–9.