Non-negative Matrix Factorization and Co-clustering: A Promising Tool for Multi-tasks Bearing Fault Diagnosis

Fei Shen¹, Chao Chen¹ and Ruqiang Yan¹,*

¹School of Instrument Science and Engineering, Southeast University
Nanjing, Jiangsu 210096, China

¹Email: ruqiang@seu.edu.cn

Abstract. Classical bearing fault diagnosis methods, being designed according to one specific task, always pay attention to the effectiveness of extracted features and the final diagnostic performance. However, most of these approaches suffer from inefficiency when multiple tasks exist, especially in a real-time diagnostic scenario. A fault diagnosis method based on Non-negative Matrix Factorization (NMF) and Co-clustering strategy is proposed to overcome this limitation. Firstly, some high-dimensional matrixes are constructed using the Short-Time Fourier Transform (STFT) features, where the dimension of each matrix equals to the number of target tasks. Then, the NMF algorithm is carried out to obtain different components in each dimension direction through optimized matching, such as Euclidean distance and divergence distance. Finally, a Co-clustering technique based on information entropy is utilized to realize classification of each component. To verify the effectiveness of the proposed approach, a series of bearing data sets were analysed in this research. The tests indicated that although the diagnostic performance of single task is comparable to traditional clustering methods such as K-mean algorithm and Gaussian Mixture Model, the accuracy and computational efficiency in multi-tasks fault diagnosis are improved.

1. Introduction

As one of the critical components in rotary machines, rolling element bearing occupies a significant position in modern mechanical equipment, which promotes a series of researches relating to the bearing fault mechanism, fault detection and fault precaution [1]. To highlight the weak fault information contaminated in the noise or reduce the influence of external factors, some time-frequency analysis methods have been studied to effectively extract features of bearing faults, including wavelet transform [2], matching demodulation transform [3], principal components analysis (PCA)[4], etc. Meanwhile, some modified classifiers were designed with the goal of improving the bearing fault diagnostic performance, such as the manifold learning classifier [5], support vector machine (SVM) [6] as well as neural network methods[7]. Other researchers even explored detection of multiple faults appearing in rotary machines. For example, Tang [8] demonstrated the high potential of kurtosis deconvolution to detect the compound faults of rolling bearings.

However, the approaches mentioned above have to suffer from the inefficiency when facing with those complicated or crossed diagnosis tasks. Aiming at the multi-task issue, a general strategy is to consider different tasks individually (e.g., crack size and crack location), thus producing increased computational loads, which, however, cannot meet the requirement for real-time diagnosis. Although some of multi-task strategies have been applied in other fields, like the diffusion least mean square
(LMS) [9] for network node recognition, the multi-task learning in context classification [10], their availability in bearing fault diagnosis is still unknown.

To introduce the concept of multi-task fault diagnosis, a novel idea which combines the non-negative matrix factorization (NMF) and Co-clustering is proposed in this paper. Different from the global characteristics of vector quantization (VQ) and PCA, NMF gives a good description about the local features, specializing in searching the small scale information from several tasks. Meanwhile, since the concept of Co-clustering is put forward by Higbee [11] in 1996, some outstanding algorithms have been developed, including the CTWC (Coupled Two-way Clustering), the Crossing Minimization, BCCA (Bi-Correlation Clustering Algorithm) [12], to realize the clustering in the row and line of a 2d matrix at the same time to meet some special requirements.

2. The basic principle of multi-task fault diagnosis

A challenge on multi-task bearing fault diagnosis is how to classify both fault locations and fault severity levels simultaneously. To tackle this challenge, two strategies were considered in traditional method: 1) Recognize these two tasks one by one, which does not consider the link between multi-purposes; 2) Subdivide the multi-task as a mixture-types classifying model (for example, 4 types for task 1 and 5 types for task 2 means 20 types for mixture task), which has the adverse effects on diagnostic accuracy. Therefore, a multi-task classifier is carried out in this paper to overcome the weaknesses of methods above.

Preparing for the multi-task classifier, short-time Fourier transform (STFT) is adopted for feature extraction because of the non-stationary characteristics of vibration signals.

\[
STFT(t,w) = \int_{-\infty}^{+\infty} s(\tau) \gamma^*(\tau-t)e^{-jwt} d\tau
\]

where \(\gamma(t)\) is the window function; \(s(\tau)\) is the collecting vibration signals. The results of STFT reflect the energy distribution both in time and frequency domain.

In the STFT-based feature extraction strategy, the time-frequency power spectrum graph is segmented by \(M \times N\) windows, and the maximum value in each window is chosen as a feature. So there will be \(M \times N\) features, which are defined as: \(F = \{f_{i,j}\}, i \in [1, 2, ..., M], j \in [1, 2, ..., N]\). A chirp signal as well as its feature graph is shown in Figure 1 as an example, where selection of \(M\) and \(N\) values depends on the non-stationarity in the time-frequency domain. More fluctuation in curves requires more segments.

![Figure 1. The power spectrum graph of chirp signal.](image-url)
Then, we assume $CmCnx^2$ samples that are chosen from 2-task sets: 1–$Cm$ category for task 1; 1–$Cn$ category for task 2. $x^2$ is the number of each sub-category. The process of multitask diagnosis is shown in Figure 2 and described as follows:

1) A group of $M \times N$ feature matrixes is constructed using the STFT features, where each matrix is composed of the identical dimension sub-feature $f_{(i,j)}$, $i \in [1,2, ..., M], j \in [1,2, ..., N]$ of all $CmCnx^2$ samples:

$$V_{(i,j)} = \begin{bmatrix}
    f_{(i,j)}^{1,1} & f_{(i,j)}^{1,2} & \cdots & f_{(i,j)}^{1, Cn}
    
    f_{(i,j)}^{2,1} & f_{(i,j)}^{2,2} & \cdots & f_{(i,j)}^{2, Cn}
    
    f_{(i,j)}^{xM,1} & f_{(i,j)}^{xM,2} & \cdots & f_{(i,j)}^{xM, Cn}
\end{bmatrix}$$ (2)

2) A Co-clustering classifier based on information entropy is carried out in each 2D feature matrix after NMF algorithm, and then the clustering matrixes are obtained, where two different types are divided into $n$ and $m$ groups in horizontal and vertical grid respectively. But it is possible that there are a bits of missing categories in each matrix, such as that shown in Figure 2(c):

$$R_{(i,j)} = \begin{bmatrix}
    C_{(i,j)}^{1,1} & C_{(i,j)}^{1,2} & \cdots & C_{(i,j)}^{1,Cn}
    
    C_{(i,j)}^{2,1} & C_{(i,j)}^{2,2} & \cdots & C_{(i,j)}^{2,Cn}
    
    \cdots & \cdots & \cdots & \cdots
\end{bmatrix} , \ C_{(i,j)}^{m,n} = \{f_{(i,j)}\}, T_1[f_{(i,j)}] = m \cap T_2[f_{(i,j)}] = n \tag{3}$$

3) The final categories are confirmed by a fuse approach, such as weight fusion, in $M \times N$ result matrixes.

---

**Figure 2.** The basic principle of multi-task fault diagnosis
3. The Non-negative Matrix Factorization

Since the non-negative characteristic of energy features, we considered the NMF algorithms for separating two tasks in the same non-negative matrix $V$ [13]. Two sub-matrixes are represented as $W$ and $H$, respectively.

$$V \approx WH$$  \hspace{1cm} (4)

where the dimension of $V$ is $x_{C_m} \times x_{C_n}$, which is then approximately factorized into an $x_{C_m} \times r$ matrix $W$ and an $r \times x_{C_n}$ matrix $H$. Usually $r$ is chosen to be smaller than $x_{C_m}$ or $x_{C_n}$, so that $W$ and $H$ are smaller than the original matrix. Factorization results mean two compressed versions of the original data matrix.

To find an approximate factorization $V \approx WH$, some cost functions are defined to quantify the degree of approximation. Such a cost function can be constructed using some distance measures between two non-negative matrixes $A$ and $B$:

1) Euclidean distance:

$$D_1(A||B) = \|A - B\|^2 = \sum_{ij} (A_{ij} - B_{ij})^2$$  \hspace{1cm} (5)

2) Divergence distance:

$$D_2(A||B) = \sum_{ij} \left( A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \right)$$  \hspace{1cm} (6)

Equations (5) and (6) are both lower bounded by zero if and only if $A = B$. In addition, the divergence distance is asymmetric between these two matrixes ($D_2(A||B) \neq D_2(B||A)$). Then, the formulation of NMF is considered to be an optimization problem according to the definition of two distances:

$$\min_{W,H} D_1(V||WH) \quad \text{s.t.} \quad W, H \geq 0$$ 

or 

$$\min_{W,H} D_2(V||WH) \quad \text{s.t.} \quad W, H \geq 0$$  \hspace{1cm} (7)

To solve the optimization problem, traditional numerical optimization approaches like gradient descent and conjugate gradient are applied to find local minima. But the convergence of the former is slow and the computation of the latter is complex. So a multiplicative update rule is designed in our paper to provide a good compromise between speed and ease of implementation.

1) For $\min_{W,H} D_1(V||WH)$, the update rule of $W$ & $H$ can be designed as [14]:

$$H_{\alpha \mu} \leftarrow H_{\alpha \mu} \frac{(W^T H)_{\alpha \mu}}{(W^T W H)_{\alpha \mu}}$$  \hspace{1cm} (8)

$$W_{i \alpha} \leftarrow W_{i \alpha} \frac{(V H^T)_{i \alpha}}{(W H^T)_{i \alpha}}$$

2) For $\min_{W,H} D_2(V||WH)$, the update rule of $W$ & $H$ can be designed as:

$$H_{\alpha \mu} \leftarrow H_{\alpha \mu} \frac{\sum_i W_{i \alpha} V_{i \mu} / (W H)_{i \mu}}{\sum_k W_{i k}}$$  \hspace{1cm} (9)

$$W_{i \alpha} \leftarrow W_{i \alpha} \frac{\sum_{\mu} H_{\alpha \mu} V_{i \mu} / (W H)_{i \mu}}{\sum_{\nu} H_{\alpha \nu}}$$
Note that, the divergence is invariant under these updates if and only if $W$ and $H$ are at a stationary point of the divergence. Also, each update consists of multiplication by a factor. In particular, it is straightforward to see that this multiplicative factor is unity when $V = WH$.

As an example of NMF results for two-task matrix, here the full matrix $V$ was defined as a $300 \times 200$ task matrix and each $100 \times 100$ represent different categories, whatever in horizontal or vertical direction. As shown in Figure 3, the dimensions of sub-matrix $W$ and $H$ are $300 \times r$ and $r \times 200$, respectively. When the dimension $r$ is set less (e.g. $r=10$), the classification result is pure and the computational load is also small. However, if the types of full matrix $V$ are not clear, we must increase the dimension of sub-matrix for higher classification accuracy in next classifier.

![Figure 3](image_url)

**Figure 3.** An example of NMF results for two-task matrix.

4. The Co-clustering Based on Information Entropy

After the completion of NMF, the matrix $V$ have been divided into two sub-matrixes $W$ and $H$ in the horizontal and vertical direction respectively. To maximize the gap between different categories and minimize it between same categories, we design an information entropy-based clustering model for co-clustering.

During the model, we assume the classification of task 1 is listed as $\{w_1, w_2, ..., w_n\}$, where $n$ is the total number of task 1, and the classification of task 2 is listed as $\{h_1, h_2, ..., h_m\}$, where $m$ is the total number of task 2. The mutual entropy $I(W; H)$ between task 1 and task 2 can be calculated as:

$$I(W; H) = \sum_w \sum_h p(w, h) \log_2 \frac{p(w, h)}{p(w)p(h)}$$

where discrete random variable $\alpha \in \{w_1, w_2, ..., w_n\}$, $\beta \in \{h_1, h_2, ..., h_m\}$; $p(w, h)$ means the joint probability distribution between $\alpha$ and $\beta$; $p(w)$ means the probability distribution of $W$; $p(h)$ means the probability distribution of $H$.

Reference [15] proved the optimal formula when the Co-clustering reaches the optimum values:

$$\arg \min_{W^*, H^*} \{I(\alpha; \beta) - I(\alpha^*; \beta^*)\}$$

And in this paper, we choose the Kullback-Leibler (KL) to represent the $I(\alpha; \beta) - I(\alpha^*; \beta^*)$:

$$I(\alpha; \beta) - I(\alpha^*; \beta^*) = KL[p(w, h), q(w, h)]$$

$$q(w, h) = p(w^*, h^*)p(\alpha|w^*)p(\beta|h^*)$$

where $KL[g, h]$ is the Kullback-Leibler distance between probability distribution function $g(x)$ and $h(x)$, which represents the relative entropy between them and is calculated in equation (13) or (14):
\[
KL[p(\alpha, \beta), q(\alpha, \beta)] = \sum_{S} \sum_{S_i} p(w) KL[p(\beta|w), q(\beta|w^*)]
\]

Therefore, the smallest mutual information entropy loss can be acquired by minimizing the \(KL\) distance between \(p(\beta|w)\) and \(q(\beta|w^*)\) or the \(KL\) distance between \(p(\alpha|h)\) and \(q(\alpha|h^*)\), and we can find the best mapping function. The detailed procedures can be described as following steps:

1) Initialize the probability density function: the sub-matrix \(W & H\) are divided into \(k & l\) groups according to the farthest segmentation criteria \([16]\), then calculate the initial probability distribution \(p(w, h)\) and \(q(w, h)\);
2) Update the row clustering: search a new category label \(i\) for each row using the constraint condition (15), to reduce the \(KL\) distance of equation (13) as far as possible.
\[
i = \arg \min_{S, j} KL[p(\beta|w), q(\beta|w^*)]
\]

Meanwhile, update the probability distribution \(q(w, h)\) with (12);
3) Update the column clustering: search a new category label \(j\) for each column using the constraint condition (16), to reduce the \(KL\) distance of equation (14) as far as possible.
\[
j = \arg \min_{S, j} KL[p(\alpha|h), q(\alpha|h^*)]
\]

Meanwhile, update the probability distribution \(p(w, h)\) with (12);
4) Compare the mutual entropy \(I(\alpha; \beta)\) with a threshold value \(I_o(\alpha; \beta)\), if \(I(\alpha; \beta) < I_o(\alpha; \beta)\), the program outputs the result of Co-clustering, otherwise, return to step (2) for next update. Particularly, a disturbance search algorithm \([16]\) is adopted during the update process in row clustering or column clustering to avoid the local optimum.

Finally, a result fusion based on the weight of each feature is carried out to obtain the final diagnosis result:
\[
i = \frac{1}{M \times N} \sum_{f=1}^{M \times N} \delta_f \times i_f
\]
\[
j = \frac{1}{M \times N} \sum_{f=1}^{M \times N} \gamma_f \times j_f
\]

where \(M\) means the number of frequency domain segmentation of STFT power spectrum graphs; \(N\) means the number of time domain segmentation of STFT power spectrum graphs; \(\delta_f\) and \(\gamma_f\) denote the weight of the \(f^{th}\) feature in task 1 and task 2, respectively.

5. Experiments and Performance Analysis
The bearing data set from Western Reserve University Bearing Data Center Website were used for the experimental study. As shown in Figure 4, the bearing apparatus consists of a drive motor, a torque transducer, and a dynamometer. Here 16 channel vibration data were collected using accelerometers, which were attached to the drive end and fan end of the motor housing with magnetic bases. Otherwise, speed and horsepower data were collected using the torque transducer/encoder, both with 12 kHz sampling frequency.

The bearing data set from Western Reserve University Bearing Data Center Website were used for the experimental study. As shown in Figure 4, the bearing apparatus consists of a drive motor, a torque transducer, and a dynamometer. Here 16 channel vibration data were collected using accelerometers,
which were attached to the drive end and fan end of the motor housing with magnetic bases. Otherwise, speed and horsepower data were collected using the torque transducer/encoder, both with 12 kHz sampling frequency.

![Drive motor, Torque transducer, Dynamometer](image)

**Figure 4.** The bearing test system.

**Table 1.** Two fault diagnostic tasks.

| Classification | C1 | C2    | C3    | C4    | C5    | C6   |
|----------------|----|-------|-------|-------|-------|------|
| Fault Severity | 0  | 0.178 | 0.356 | 0.533 | 0.712 | —    |
| Fault Location | Health | Inner | Ball  | Outer 3.00 | Outer 6.00 | Outer 9.00 |

Firstly, we extracted and compared the *STFT* features in different diagnostic tasks. Figure 5 illustrated the power spectrum graphs of fault severity task (Fault location: inner; Load: 0 horsepower; Rotate speed: 1797Hz). Figure 6 gave the power spectrum graphs of fault location task (Fault severity: 0.533; Load: 0 horsepower; Rotate speed: 1797Hz). From Figures 5 and 6, we can find the power spectrum graphs between different types have a difference in the frequency-axis, but are stable in the time-axis. Obviously, the influence of fault severity and fault location on the frequency domain is larger than that on the time domain. Based on the observed power spectrum graphs, we segmented these graphs using a (100×2) window, where 100 means there being 100 segmentations in the range of 0~500Hz, and 2 is the number of windows from 0s to 120s. Therefore, the dimension of feature vector of each sample is 200, and we chose 100 (10×10) samples in each fault category for NMF algorithm.

![Power spectrum graphs](image)

**Figure 5.** The power spectrum graphs of fault severity task.
Secondly, the NMF and the Co-clustering strategy were carried out for 100×(5+6) samples. We compared the performance of bearing fault diagnosis as well as their computational loads when the $r$ value in $W$ and $H$ increases from 1 to 100, the fault diagnosis accuracy curve along with the relative time cost is shown in Figure 7. It can be seen that the diagnosis accuracy grows from 74.16% (①) to 97.08% (③) when the dimension of NMF sub-matrix rises from 1 to 100, meanwhile, the computational load also appear an exponential increase from approximately 0 to 100. According to Figure 7, a balance point is found in $r=39$ (②) where the diagnosis accuracy is satisfactory enough (96.04%), also the diagnosis keeps a low level (15.21%). So, the dimension of NMF is designed as 39 when proposed method is applied to the Western Reserve University bearing data set.

![Figure 6. The power spectrum graphs of location task.](image)

![Figure 7. The fault diagnosis accuracy curve as well as the relative time cost curve.](image)
To verify the effectiveness of the proposed approach, we compared its performance with some classical clustering algorithms, such as K-means algorithm and GMM (Gaussian Mixture Model) algorithm. The fault diagnosis results are listed in Table 2. Some conclusions can be obtained from this table:

1) The bearing diagnostic accuracy in fault severity (91.16%) is weak smaller than that in fault location (94.08%) in all five strategies, which means that task 1 mainly lies on the time domain features, while task 2 depends on the frequency domain features. So increasing the time domain segmentation number (more than 2) of NMF is a good measure to improve the classification performance of task 1;

2) During classical clustering algorithms, although the accuracy of “Task 1+Task 2” strategy is higher than the “Task 1” strategy, the time cost of former is also larger than the later (increase by about 30%). Meanwhile, the GMM algorithm requires more time, even though possessing higher performance than K-mean approach;

3) The NMF-based Co-clustering offers good fault diagnosis performance compared with classical clustering algorithms. In addition, task 1 and task 2 are classified at the same time, which guarantees a low computational load, only 70.7% as compared with the K-means.

Table 2. The performance comparison between K-means, GMM, and NMF-based Co-clustering.

| Methods                      | Accuracy of task 1 (mean±std) | Accuracy of task 2 (mean±std) | Total accuracy (mean±std) | Relative time cost with K-means (%) |
|------------------------------|-------------------------------|-------------------------------|---------------------------|-----------------------------------|
| Task 1⊙Task 2: K-means      | 87.73±4.54                   | 91.46±3.68                   | 90.76±1.40                | 76.2                              |
| Task 1+Task 2: K-means      | 91.12±4.01                   | 94.55±3.84                   | 95.77±1.56                | 100.0                             |
| Task 1⊙Task 2: GMM          | 89.44±5.91                   | 92.13±3.50                   | 93.25±2.02                | 84.1                              |
| Task 1+Task 2: GMM          | 94.26±5.39                   | 95.46±3.77                   | 95.05±2.39                | 112.3                             |
| NMF-based Co-clustering     | 93.26±4.05                   | 96.82±3.93                   | 95.43±2.13                | 70.7                              |

Task 1⊙Task 2: an algorithm for the 5×6=30 mixture categories;
Task 1+Task 2: an algorithm for 5 categories of task 1 firstly, then for 6 categories of task 2;

6. Conclusion
A fault diagnosis method based on NMF and Co-clustering strategy is proposed for multi-tasks bearing fault diagnosis. Here the high-dimensional matrixes are constructed using the STFT features, where the dimension of each matrix equals to the number of target tasks. With the NMF method, both fault severity and fault location of bearings can be identified at the same time. The time cost of the proposed method improves about 30% compared with classical clustering algorithms such as K-means and GMM, with acceptable diagnostic performance. Meanwhile, the suitable dimension of sub matrixes by NMF is 39. In that case, both classification accuracy and time cost are satisfactory thus being appealing for the multi-task application in real time bearing fault diagnosis system.

Acknowledgments
This project is supported by National Natural Science Foundation of China (Grant No. 51575102).

References
[1] Yu J 2011 J. Mechanical Systems and Signal Processing 25 2573-88.
[2] Li P, Kong F R and He Q B 2013 J. Measurement 46 497-505.
[3] Wang S, Chen X, Li G, Li X and Liu Z 2014 J. Instrumentation and Measurement 63 1372-83.
[4] Widodo A and Yang B S 2011 J. Expert Systems with Applications 38 8430-37.
[5] Tang B P, Song T and Li F 2012 J. Renewable Energy 62 1-9.

9
[6] Widodo A and Yang B S 2008 J. Mechanical Systems and Signal Processing 21 2560-74.
[7] Gebrael N Z and Lawley M A 2008 J. Automation Science and Engineering 5 154-163.
[8] Tang G, Wang X and He Y 2016 J. Journal of Mechanical Science and Technology 30 43-54.
[9] Chen J, Richard C and Sayed A H 2015 J. Signal Processing 63 2733-48.
[10] Xue Y, Liao X and Carin L 2007 J. Machine Learning Research 8 35-63.
[11] Higbee K 1996 J. Technometrics 40 852.
[12] Hochreiter S, Bodenhofer U and Heusel M 2010 J. Bioinformatics 26 1520-27.
[13] Zhongliang F U, Liu J and Jintao L I 2016 J. Geomatics.
[14] Lee D D and Seung H S 1999 J. Nature 401 788-791.
[15] Dhillon I S, Mallela S and Modha D S 2003 J. Kdd 89-98.
[16] Ke W 2015 J. International Journal on Digital Libraries 16 1-15.