International Journal of Science and Engineering Applications
Volume 14-Issue 11, 04 – 06, 2025, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1411.1002

Research on Data Analysis and Feature Extraction of Bearing Source Domain Based on Fault Mechanism

Shijie Teng
School of Electric Information
and Electrical Engineering
Yangtze University
Jingzhou, China

Abstract: To address the adaptability issue of bearing fault diagnosis models in cross-domain transfer tasks, this study focuses on source domain data selection, preprocessing, and mechanism-driven feature extraction. The Case Western Reserve University (CWRU) Bearing Data Center dataset is selected as the source domain data, covering four states: normal (N), outer race fault (OR), inner race fault (IR), and ball fault (B). Firstly, interference is eliminated through data preprocessing (unifying the sampling rate, applying band-pass filtering, slicing with sliding windows, and standardization). Secondly, based on the bearing fault mechanism, characteristic frequencies (BPFO, BPFI, BSF) are calculated using geometric parameters and rotational speed estimation. Then, multi-dimensional features are extracted from the time domain, frequency domain, envelope spectrum, and time-frequency domain to comprehensively capture fault information. Finally, the effectiveness of the features is verified using a radial basis function kernel support vector machine (RBF-SVM) with class weights, achieving an accuracy rate of 97.14% on the test set and a macro-average F1 score of 0.9705. The results indicate that the extracted features have good generalization and discrimination, laying a foundation for subsequent target domain transfer diagnosis.

Keywords: Bearings fault diagnosis; RBF-SVM; Transfer learning; CWRU dataset; Time-frequency analysis

1. Introduction

Bearings, as the core component of rotating machinery, their fault diagnosis is of vital importance for ensuring equipment safety. In fault diagnosis based on transfer learning, the quality of source domain data and the effectiveness of feature extraction directly determine the performance of the model in the target domain (such as different loads, rotational speeds, and equipment types). However, the original bearing vibration signals contain noise and sampling rate differences, and single-dimensional features cannot fully reflect the fault characteristics. This study addresses the above issues by carrying out the following work: (1) Selecting representative source domain data from the public CWRU dataset to ensure coverage of typical fault types; (2) Using mechanism-driven data preprocessing methods to improve signal quality; (3) Extracting multi-dimensional features from the time domain, frequency domain, envelope spectrum, and time-frequency domain; (4) Verifying the effectiveness of features through classification models. The ultimate goal is to provide a high-quality feature set for subsequent target domain transfer tasks.

2. Data Filtering and Preprocessing

2.1 Data Screening

Data filtering: The CWRU dataset was selected as the source domain data. This dataset is widely used in the field of bearing fault diagnosis. The selection criteria are as follows: Fault type: includes outer ring fault (OR), inner ring fault (IR), rolling element fault (B), and normal state (N); Sampling rate: covers 12kHz and 48kHz, consistent with the typical sampling rate in industrial scenarios; Load condition: within 0.3 horsepower range, ensuring diversity of working conditions; Sample quantity: after slicing, there are a total of 1925 samples (OR: 923, IR: 498, B: 504), avoiding extreme class imbalance.

2.2 Data Preprocessing

To eliminate the interfering factors in the original signal, signal preprocessing is necessary.The specific steps are as follows: 1. Sampling rate uniformity: Use the multi-phase resampling method to map the 48kHz signal to 12kHz, avoiding scale deviation in feature calculation; 2. Band-pass filtering and envelope extraction: Use an IIR filter to retain the "resonance-sensitive band" (the concentrated frequency band of fault impact signals) at 500 - 5000Hz, and extract the envelope signal

$$C(t) = |H\{x_b(t)\}| \tag{2.1}$$

through Hilbert transformation, amplifying the periodicity of the signal; 3. Sliding window slicing: Use a sliding window with a 1-second window length and a 50% overlap rate to divide the long signal into small samples, reducing the computational complexity; 4. Normalization: Normalize each window's signal according to

$$x[n] = \frac{x[n]\mu}{\sigma} \tag{2.2}$$

(where μ is the window mean and $\sigma$is the standard deviation), eliminating the influence of amplitude differences.

3. Mechanism-driven feature analysis

3.1 Speed estimation

The rotational speed f_r is the basis for calculating the fault characteristic frequency. To avoid errors caused by low-frequency disturbances in the original signal, the power spectrum $P_e(f)$ is calculated using the filtered envelope signal within the frequency range corresponding to the rotational speed

$$f_r \in [5,120]Hz$$

The peak of the spectrum is searched for within this range, and the frequency corresponding to this peak is the rotational frequency f_r, which is converted to rotational speed $RPM =$

International Journal of Science and Engineering Applications
Volume 14-Issue 11, 04 – 06, 2025, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1411.1002

$60 \times f_r$. The median of the rotational speeds in multiple windows of the same file is taken to further reduce random errors.

3.2 Calculation of Fault Feature Frequency

TBased on the bearing geometric parameters (number of rolling elements $n = 8$, diameter of rolling elements $d = 7.94$mm, pitch circle diameter $D = 39.04$ mm, contact angle $\phi = 0°$ and rotational frequency$f_r$from the CWRU dataset, the core fault characteristic frequencies are calculated as follows: Outer ring fault frequency (BPFO):

$$BPFO = \frac{n}{2} f_r \left(1 + \frac{d}{D} \cos \phi\right) \quad (3.1)$$

approximately 95-110 Hz;

Inner ring fault frequency (BPFI):

$$BPFI = \frac{n}{2} f_r \left(1 + \frac{d}{D} \cos \phi\right) \quad (3.2)$$

approximately 150 Hz;

Rolling element fault frequency (BSF):

$$BSF = \frac{D}{2d} f_r \left(1 - \left(\frac{d}{D} \cos \phi\right)^2\right) \quad (3.3)$$

approximately 120130 Hz.

3.3 Representative sample analysis

The average envelope spectra were plotted by category (Figure 1), and the results showed that the samples with outer ring faults (OR) exhibited obvious energy peaks at the BPFO and its harmonics, the samples with inner ring faults (IR) had significant peaks at the BPFI, and the samples with rolling element faults (B) showed concentrated energy at the BSF and its harmonics. This result was in perfect agreement with the theoretical characteristic frequencies, verifying the reliability of the mechanism-driven analysis.

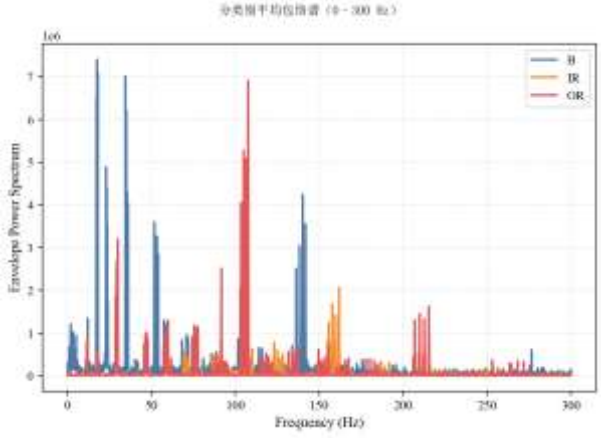

Figure.1 OR, IR, B sample average envelope spectrum

4. Dimension feature extraction

To comprehensively capture fault information, a feature set is constructed from four dimensions, and the core features and their physical meanings are shown in Table 1.

Table 1. Multidimensional feature set of data

Feature Dimension	Core feature	physical significance
time domain	Skewness, Peak Factor, Pulse Factor, Root Mean Square	Reflecting the impact intensity
frequency domain	Center of spectrum, spectral broadening, spectral entropy	Reflecting energy distribution
Envelope spectrum	Harmonic energy proportion, Fault index	Enlarge the periodicity
Time-frequency domain	Band energy, Time-frequency peak frequency	Incorporating time-frequency information

5. Verification of feature validity

5.1 RBFSVM classification model

To address the issue of class imbalance in the source domain data (where the number of OR samples is greater than that of other classes), class weights (proportional to the sample quantity) are set in the RBFSVM to balance the contribution of each class to the loss function. Grid search is used to optimize the hyperparameters C (penalty coefficient) and γ (kernel function width), and the data is divided into a training set and a test set in a 8:2 ratio.

5.2 Performance evaluation

The model test results are as follows: Accuracy: The overall accuracy of the test set reached 97.14%; Classification indicators: The macro average F1 score was 0.9705, among which the recall rate for the IR class was 100%, and the recall rates for the B class and OR class were both over 96% (Table 2); Stability: The macro average F1 score of the five-fold cross-validation was 0.9736 ± 0.0088, with small performance fluctuations and good generalization. Through tSNE dimensionality reduction, the high-dimensional features were mapped to a two-dimensional space (Figure 2), and the samples of the OR, IR, and B classes were clearly clustered, with distinct class boundaries. Only B and OR had a slight intersection due to the overlap of the multiplicative components under some loadings, verifying the good separability of the features. The random forest feature importance analysis showed that the energy proportion of the envelope spectrum harmonic (such as $E_R(BPFI)$, $E_R(BPFO)$) and the mechanism ratio feature (such as FI_IR, FI_OR) accounted for more than 70% in the top ten features, further confirming the rationality of feature construction.

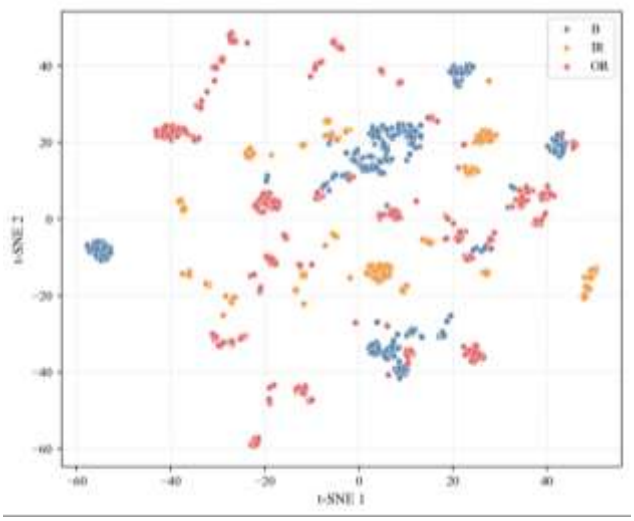

Figure.1 OR, IR, B sample average envelope spectrum

Table 2. Test set classification performance indicators

Fault Type	Precision	Recall	F1 Score	Sample Count
B (Rolling Element Failure)	0.9706	0.9802	0.9754	101
IR (Inner Ring Failure)	0.9901	1.0000	0.9950	100
OR (Outer Ring Failure)	0.9945	0.9837	0.9891	184

6. Conclusion

This study uses the CWRU dataset as the source domain to complete data screening, mechanism-driven preprocessing, and multi-dimensional feature extraction. The effectiveness of the features is verified through RBFSVM, and the following conclusions are drawn: 1. The preprocessing process effectively eliminates sampling rate differences and noise interference, laying a good foundation for feature extraction; 2. Features based on fault mechanisms (such as BPFO, E_R) are crucial for fault classification and are highly consistent with physical laws; 3. The multi-dimensional feature set has good generalization and separability, and is suitable for subsequent target domain transfer diagnosis tasks. Future research will focus on using this feature set to reduce the distribution differences between the source domain and the target domain, further improving the performance of transfer diagnosis.

7. References

[1] Yao, C., Wang, S. B., Chen, B. Y., Mei, G. M., Zhang, W. H., Peng, H., & Tian, G. R. (2022). An Improved Envelope Spectrum via Candidate Fault Frequency Optimization-gram for Bearing Fault Diagnosis. Journal of Sound and Vibration, Elsevier.

[2] Xu, X. F., Xu, D. Y., Xu, S., Guo, N. X., & Xu, H. Y. A Bearing Fault Diagnosis Method Combining Spectral Clustering and Association Rules. Computer Measurement & Control. Moorthy, V. M., Muthukumaran, B. M., Britto Manoj, R. C. S., & Arul Elango. 2024. A Secure and Resilient Smart Energy Meter. IEEE Access, 12.

[3] Liu, Y. F., Liu, H. B., Wei, M., & Li, M. F. (2025). A Fault Diagnosis Method for Dry Vacuum Pump Bearing Based on Finite Element Simulation with Deep Transfer Learning. IEEE Access (Early Access), 1-1.

[4] Zhou, P., Wu, D. S., Xu, J. C., Wang, Z. N., & Ma, D. Z. (2024). Fault Diagnosis Method for Rolling Bearing Based on Probabilistic Diffusion Models Under Imbalanced Data. IEEE Sensors Journal, 24(23), 40059-40068.

[5] Jia, Z. H., Wang, C. S., Zhu, H. B., Li, J., & Yang, G. (2019). A Rolling Bearing Fault Diagnosis Method Based on the Fusion of Decision Tree and Neural Network. Highlights of Sciencepaper Online, 12(2).

[6] Cheng, Y., et al. (2025). Research on Candidate Fault Frequency Optimization and Improved Envelope Spectrum for Bearing Fault Diagnosis (Matlab Code Implementation). 51CTO Blog.

[7] Anonymous. (2025). A Rolling Bearing Fault Diagnosis Method Based on Multimodal Knowledge Graph. IEEE Transactions on Industrial Informatics

[8] Samanta, B. (2004). Bearing Fault Diagnosis Using SVM and ANN with Genetic Algorithm for Feature Selection. Journal of Vibration and Control, 10(5), 713-732.

[9] Lei, Y. G., Lin, J., & He, Z. J. (2018). A Review on Data-Driven Fault Diagnosis for Rotating Machinery. IEEE Transactions on Industrial Informatics, 14(3), 1181-1196.