Multi-SoftMax Convolutional Neural Networks and Their Application to Complex Fault Diagnosis of Planetary Gearboxes

Authors: Jianhua Shan, Lü Qin, Zhang Shenlin, Meng Rui, Wang Xiaoyi, Jianhua Shan

Date: 2017-12-26T00:00:00+00:00

Abstract

Existing planetary gearbox fault diagnosis methods exhibit limitations: first, conventional methods suffer from complexity and are ineffective at diagnosing planetary gear fault types; second, convolutional neural network-based approaches are primarily utilized for gearbox fault diagnosis and are rarely employed for planetary gearbox diagnosis. To effectively diagnose complex fault types and variable operating conditions, this paper proposes for the first time a fault tree structure, an operating condition parallel structure, and a multi-SoftMax convolutional neural network. The fault tree structure enables unified processing of various complex fault types and allows for examination of diagnostic performance at each node. The operating condition parallel structure can handle variable operating conditions and predict rotational speed and load. Utilizing vibration data from a laboratory planetary gearbox, a series of tests were conducted, demonstrating that the proposed method can accurately diagnose complex faults and variable operating conditions in planetary gearboxes with an accuracy of 97%, thereby verifying the strong generalization capability of the multi-SoftMax convolutional neural network and the advantages of the fault tree structure.

Full Text

Preamble

DOI: 10.3901/JME.20..***

Multi-SoftMax Convolutional Neural Network and Its Application in the Diagnosis of Planetary Gearbox Complex Faults

SHAN Jianhua, LV Qin, ZHANG Shenlin, MENG Rui, WANG Xiaoyi

(School of Mechanical Engineering, Anhui University of Technology, Maanshan 243002)

Abstract

Existing methods for planetary gearbox fault diagnosis suffer from two primary limitations: First, traditional approaches are computationally complex and cannot effectively identify specific fault types in planetary gears. Second, convolutional neural network (CNN)-based methods have been predominantly applied to conventional gearbox fault diagnosis, with minimal application to planetary gearboxes. To address complex fault types and variable operating conditions, this paper proposes, for the first time, a fault tree structure, a working condition parallel structure, and a multi-SoftMax convolutional neural network. The fault tree structure enables unified processing of various complex fault types while allowing examination of diagnostic performance at each node. The working condition parallel structure handles variable operating conditions and predicts rotational speed and load. Experimental validation using vibration data from a laboratory planetary gearbox demonstrates that the proposed method accurately diagnoses complex faults and variable operating conditions in planetary gearboxes, achieving 97% accuracy. These results verify the strong generalization capability of the multi-SoftMax convolutional neural network and the advantages of the fault tree structure.

Keywords: Fault diagnosis; Planetary gearbox; Deep learning; Convolutional neural network

0 Introduction

Planetary gearboxes offer significant advantages including high transmission ratios, compact size, and substantial load-bearing capacity, making them widely employed in wind power generation, metallurgy, shipbuilding, and lifting equipment. However, gearbox failures can cause severe safety accidents, rendering fault diagnosis critically important. The gear motion in planetary gearboxes represents typical compound motion, producing more complex vibration responses than those from bearings or fixed-axis gearboxes, which presents unique challenges for fault diagnosis.

Numerous studies have addressed these challenges. YU from the University of Toronto proposed a planetary gearbox fault diagnosis method based on wavelet transform and time-synchronous averaging. In 2017, CHENG Junsheng from Hunan University introduced a method based on ASTFA and SDEO demodulation. In 2016, HUANG Wentao from Harbin Institute of Technology developed a resonance-based sparse signal decomposition method with adaptive quality factor optimization for planetary gearbox fault diagnosis. However, these methods are only suitable for steady-state conditions, involve complex mathematics, and exhibit suboptimal diagnostic performance. To overcome these limitations, an innovative approach is required.

In recent years, deep learning has emerged as the most active field within machine learning, achieving remarkable success in natural language processing and image recognition through powerful feature learning capabilities. As a key component of deep learning, convolutional neural networks have been primarily applied to two-dimensional image recognition, offering several advantages: (1) Automatic feature learning from large datasets eliminates the complex and tedious manual feature extraction process, reducing dependence on expert knowledge. (2) The architecture employs local connections and weight sharing, dramatically reducing network parameters, lowering training difficulty, improving training speed, and enhancing generalization capability.

1 Fault Tree Structure and Multi-SoftMax Convolutional Neural Network

1.1 Development of Convolutional Neural Networks

In recent years, scholars have begun applying CNNs to rotating machinery fault diagnosis, primarily for bearing faults. In 2016, WANG Jinjiang from China University of Petroleum used CNNs for rotating machinery fault diagnosis with one-dimensional vibration data as input. In 2017, CHEN Lu from Beihang University, ZHANG Wei from Harbin Institute of Technology, and LIU Ruonan from Xi' an Jiaotong University all converted one-dimensional vibration signals into two-dimensional representations for CNN-based fault diagnosis. In 2017, Min Xia from the University of British Columbia proposed a multi-sensor fusion CNN approach for rotating machinery fault diagnosis. However, these studies exhibit several shortcomings: (1) Input sample lengths are arbitrarily fixed, making them difficult to generalize across different sampling frequencies and rotational speeds. (2) They primarily diagnose bearing faults with distinct vibration patterns, whereas planetary gearbox faults show minimal waveform differences, presenting significantly greater diagnostic challenges. (3) Operating conditions are simplified with fixed or narrowly varying speeds, without addressing variable condition diagnosis. (4) For compound faults, composite failures are treated as a single independent class, preventing examination of individual fault diagnostic performance.

To address these limitations, this paper proposes an improved solution for planetary gearbox fault diagnosis. The innovations include: (1) A formula for calculating sample data length that adapts to different speeds and sampling frequencies, combined with random sampling to enhance network generalization and simplify testing. (2) A fault tree structure that uniformly handles both simple and compound faults while enabling performance evaluation at each node. (3) A multi-SoftMax network architecture to implement the fault tree structure diagnosis. (4) The adoption of large convolutional windows, large pooling strides, and the latest DenseNet structure to simplify the network and solve variable-condition planetary gearbox fault diagnosis.

1.2 Convolutional Neural Network Architecture

Traditional CNNs consist of input layers, convolutional layers, nonlinear activation layers, pooling layers, fully connected layers, and SoftMax output layers, with convolutional, activation, and pooling layers forming the core. As shown in [Figure 1: see original paper], the input layer is typically an image represented as $0X$. The convolutional layer performs feature extraction from the previous layer using convolution kernels, where iX represents the feature map of layer i, iW represents the weight matrix of layer i' s convolution kernel, $\otimes$ denotes convolution operation, and ib represents the bias vector of layer i.

The nonlinear activation layer maps features from the previous layer to the next through nonlinear functions, expressed as $F(iX)$, where F is a nonlinear function. Common activation functions include ReLU and its variants: Leaky ReLU, Random ReLU, Parametric ReLU, Shifted ReLU, and Exponential LU.

The pooling layer downsamples feature maps, reducing both dimensionality and network parameters. Maximum pooling is most common, expressed as $sbp(saX)$, where $s \times s$ is the pooling window size, and a and b represent pixel coordinates.

The CNN objective is to minimize loss, comprising empirical risk and structural risk. Cross-entropy loss is commonly used for empirical risk, while weight L2 norm serves as structural risk.

1.3 Fault Tree Structure

In practice, fault types can be extremely complex, including simple faults, compound faults, and varying fault severities. To uniformly handle these scenarios, a fault tree structure was designed, as shown in [Figure 2: see original paper]. The tree has "Normal (no fault)" and "Faulty" as root nodes, branching downward to leaf nodes. "Normal" is a leaf node, while "Faulty" branches into simple and compound faults. Simple faults are categorized by component (sun gear, bearing, planet gear, carrier), with each component further divided by fault type (broken tooth, crack, wear, missing tooth). For example, sun gear cracks are classified by severity as slight, medium, or severe. Compound faults involve multiple components failing simultaneously. This structure can be adapted for other machinery fault types.

1.4 Working Condition Parallel Structure

Beyond fault types, planetary gearboxes operate under variable conditions where rotational speed and load change. A working condition parallel structure was designed, as shown in [Figure 3: see original paper], incorporating several typical speeds and loads based on actual operating conditions. Additional condition attributes can be added to this structure as needed.

1.5 Multi-SoftMax Convolutional Neural Network Structure

The multi-SoftMax CNN architecture includes input layers, convolutional layers, max pooling layers, average pooling layers, and a multi-SoftMax output layer. The output layer contains multiple SoftMax score vectors, denoted as $\{SoftMax_1, ..., SoftMax_n\}$. As shown in [Figure 4: see original paper], each $SoftMax_i$ is fully connected to the final feature layer. Traditional CNNs are special cases of this architecture with a single SoftMax output vector.

The empirical risk loss differs significantly from traditional CNNs. For example, in [Figure 2: see original paper], if a planetary gearbox has a sun gear fault but the specific location is unknown, the corresponding SoftMax index set is $\{1, 2, 3\}$. If the fault is a medium-severity crack, the index set becomes $\{1, 2, 3, 5, 9\}$. Thus, samples correspond to multiple SoftMax vectors with variable counts.

The SoftMax index set includes vectors from both fault tree and working condition parallel structures. In [Figure 3: see original paper], for speed 2 and load 3, the working condition index set is $\{10, 11\}$. When both structures are present, samples include SoftMax vectors from both.

The empirical risk loss for multi-SoftMax CNN is the average cross-entropy loss across all included SoftMax vectors. For a single $SoftMax_i$ vector, the loss is:

$$\text{Loss}_i = -\sum_{j=1}^{n} y_{ij} \log(sm_{ij})$$

where sm is the $SoftMax_i$ score vector, n is its dimension, and y_{ij} is the corresponding label value.

The sample loss is:

$$\text{Loss} = \frac{1}{|S_i|} \sum_{SoftMax_i \in S_i} \text{Loss}_i$$

where S_i is the sample' s SoftMax vector set and $|S_i|$ is its cardinality. During backpropagation, gradients are only passed through SoftMax vectors in S_i.

Sample labels are determined such that each $SoftMax_i$ corresponds to one label element, with values representing attribute indices (starting from 1). For example, if $SoftMax_3$ represents bearing faults, the label element is 2. A medium-severity sun gear crack corresponds to label $[2, 1, 1, 2, 2]$.

2 Planetary Gearbox Fault Diagnosis Methodology

The multi-SoftMax CNN fault diagnosis process comprises four steps: training database creation, network architecture design, network training, and fault diagnosis.

(1) Training Database Creation: This involves generating sample data and labels. For each operating condition, samples are created randomly by extracting continuous data points longer than one maximum data period (the number of sampling points per revolution at the lowest speed) from random positions in the vibration data. The preferred sample length is $k \times 2^n$ points, where k and n are positive integers. This random approach enhances model generalization and simplifies testing. Sufficient samples are generated to form the training database, with labels created according to the fault tree and working condition parallel structures.

(2) Multi-SoftMax CNN Architecture Design: The network is designed based on DenseNet. The input layer is $H \times 1 \times K$, where H is sample length, 1 is data width, and K is data dimension (number of sensors). Multi-sensor signal acquisition improves diagnostic accuracy, equivalent to multi-sensor fusion at the data level.

(3) Network Training: The multi-SoftMax CNN is trained using the training database. Mini-batch training is employed, with each mini-batch containing different samples, ensuring each sample participates in training only once. Hyperparameters are tuned to optimize network parameters and generalization performance.

(4) Fault Diagnosis: Test samples are input into the trained network to obtain multi-SoftMax score vectors. For fault tree vectors, diagnosis proceeds from the root node downward based on maximum values until reaching a leaf node. Using [Figure 2: see original paper] as an example, $SoftMax_1$ (root node) is evaluated first: if the maximum is at position 1, the diagnosis is normal and terminates; if at position 2, a fault is detected and evaluation continues to $SoftMax_2$. Working condition vectors are evaluated directly. For improved reliability, multiple samples can be collected, with the most frequent label selected as the final diagnosis.

3 Planetary Gearbox Fault Diagnosis Case Study

This case study uses self-measured fault data from a laboratory planetary gearbox test rig, as shown in [Figure 5: see original paper], with vibration data collected from two sensors. The gearbox parameters are listed in .

** Experimental Planetary Gearbox Parameters**

Parameter	Sun Gear	Planet Gear	Ring Gear	Carrier
Module (mm)				
Pressure Angle (deg)				
Tooth Width (mm)				
Mass (kg)				
Pitch Diameter (mm)				

The experiment includes four sun gear rotational frequencies: 10 Hz, 15 Hz, 20 Hz, and 25 Hz. Due to experimental constraints, the fault types are simplified compared to [Figure 2: see original paper]. Simple sun gear faults include broken tooth, crack, wear, and missing tooth at 10 Hz, 20 Hz, and 25 Hz. Compound faults combine sun gear faults (broken tooth, crack, wear, missing tooth) with bearing outer ring faults at 15 Hz and 20 Hz. Load conditions include loaded and unloaded states.

Vibration waveforms under different conditions are shown in [Figure 6: see original paper]. Under identical conditions, normal, broken tooth, and crack waveforms are very similar, demonstrating the diagnostic difficulty. Under the same condition, wear patterns differ significantly from other fault types but resemble patterns from other conditions (e.g., wear at one condition resembles broken tooth at another). Missing tooth patterns are distinct and easily identifiable.

3.1 Database Creation

The sampling frequency is 8192 Hz. At the minimum 10 Hz rotation frequency, the sun gear completes one revolution with 820 sampling points, defining the data period. 1280 is selected as the sample length—exceeding 820 and satisfying $k \times 2^n$ (with $k = 5$, $n = 8$). For each condition, 1280-point segments are extracted from random positions, as shown in [Figure 7: see original paper]. This relatively short length reduces network complexity and training difficulty while accelerating training. Random sampling eliminates the need for specific extraction points during testing and theoretically enables unlimited sample generation.

The corresponding fault tree and working condition parallel structures are shown in [Figure 8: see original paper], with SoftMax index set $[1, 2, 3, 4, 5, 7, 8, 10]$. Since indices $[3, 4, 8]$ in [Figure 2: see original paper] contain only single attributes, these SoftMax vectors can be omitted, resulting in final index set $[1, 2, 5, 7, 10]$.

For a sample at 20 Hz with wear fault under simple fault conditions, the label is $[2, 1, 3, 3]$.

3.2 Multi-SoftMax CNN Architecture

The network employs large convolutional kernels and strides, with pooling layers using large windows and strides to reduce network depth. The activation function is $\mathrm{Max}(-1, X)$. To address the trade-off between excessive depth (training difficulty) and insufficient depth (poor feature extraction), the DenseNet dual-path architecture is adopted, as detailed in [Figure 9: see original paper] and .

** Multi-SoftMax CNN Layer Parameters**

Layer	Operator Size	Stride	Feature Map Size	Multi-SoftMax Output
Input	-	-	$1280 \times 1 \times 2$	-
Conv1	-	-	$1280 \times 1 \times 48$	-
Pool1	-	-	$321 \times 1 \times 48$	-
Conv2	-	-	$321 \times 1 \times 96$	-
Pool2	-	-	$80 \times 1 \times 96$	-
Conv3	-	-	$80 \times 1 \times 192$	-
Pool3	-	-	$40 \times 1 \times 192$	-
Feature	-	-	$40 \times 1 \times 192$	-
SoftMax1	-	-	$1 \times 1 \times 192$	$1 \times 1 \times 16$
SoftMax2	-	-	$1 \times 1 \times 192$	$1 \times 1 \times 5$

3.3 Test 1

Hyperparameters: For iterations $i < 7000$, learning rate = 0.001, regularization = 0.00001, mini-batch = 32, momentum = 0.9; for $i \geq 7000$, learning rate = 0.0005. Weights are initialized with Gaussian random numbers (mean = 0, variance = 0.1) and biases to zero. Input data is vibration waveforms multiplied by 10. Training employs early stopping at 21,400 iterations.

Each condition provides 100 test samples. Diagnostic performance uses two metrics: (1) Overall accuracy requiring complete label matching achieves 97.55%, demonstrating accurate diagnosis of complex planetary gearbox faults. (2) Recall rate (proportion of correctly diagnosed samples per class) for each node, as shown in [Figure 10: see original paper]. Simple fault recall is 99.96%, indicating nearly all simple fault samples are correctly classified as simple faults.

3.4 Test 2

In practice, planetary gearboxes experience complex variable loads not fully represented in training data. To evaluate load generalization capability, Test 2 uses unloaded data for training and loaded data for testing, with identical labels and network architecture.

Hyperparameters are similar to Test 1, but training stops at 6,000 iterations (earlier than Test 1), indicating reduced training difficulty. Each loaded condition provides 100 test samples, achieving 88.83% accuracy. Node recall rates are shown in [Figure 11: see original paper].

Although accuracy is lower than Test 1, collecting multiple samples per condition (e.g., 100) and selecting the most frequent diagnosis significantly reduces misclassification probability. Compound fault recall substantially exceeds its sub-node recalls, indicating most compound fault samples are correctly identified as compound faults, though specific fault type classification within compound faults shows errors (e.g., many broken tooth samples misclassified as wear). This demonstrates the fault tree structure' s advantage: when fault type details are ignored, recall rates for fault presence and simple/compound classification remain high (minimum 96.92%).

3.5 Test 3

Planetary gearboxes often operate under variable speeds. To evaluate speed generalization, Test 3 uses 10 Hz, 15 Hz, and 25 Hz data for training and 20 Hz data for testing, excluding speed score vectors. The pre-SoftMax network structure matches Test 1.

Training stops at 5,500 iterations (earlier than Test 1), indicating reduced training difficulty. Each condition provides 100 test samples, achieving 76.37% accuracy. Node recall rates are shown in [Figure 12: see original paper].

Test 3 accuracy is significantly lower than Tests 1 and 2, indicating speed variation impacts vibration waveforms more than load variation. This is confirmed by normal gearbox waveforms at different speeds in [Figure 13: see original paper]. Average absolute amplitude values are: 0.0032, 0.0056, 0.0075, and 0.012 m/s^2 for 10–25 Hz respectively—a fourfold maximum difference.

Despite 76.37% accuracy, recall rates for fault presence and simple/compound classification remain near 100%. Lower accuracy stems primarily from misclassifying specific fault types within compound faults (e.g., most broken tooth and crack errors).

Conclusions

This paper makes three key contributions: (1) First proposal of fault tree and working condition parallel structures for unified handling of complex fault types and variable conditions, with node-level diagnostic performance visibility. (2) First proposal of multi-SoftMax CNN architecture for these structures, incorporating DenseNet principles and Max(-1, X) activation for easier training. (3) Successful diagnosis of laboratory planetary gearbox faults with 97.55% accuracy in Test 1, validating the approach' s effectiveness.

References

[1] LEI Y G, HE Z J, et al. Research advances of fault diagnosis technique for planetary gearboxes[J]. Journal of Mechanical Engineering, 2011, 47(19): 59-67.

[2] ZHOU Z K. Summary of research on fault diagnosis technology of wind turbine gearbox[J]. Technology and Market, 2016, 23(4): 25-28.

[3] ZOU J CH, SHEN Y D. Review of gearbox fault diagnosis under variable working conditions[J]. Mechanical Drive, 2012, 36(08): 124-127(132).

[4] YU Jing, YIP L, MAKIS V. Wavelet analysis with time-synchronous averaging vibration data for fault detection, diagnostics, and condition-based main-

tenance[C]//2nd International Conference on Mechanical and Electronics Engineering, August 1-3, 2010, Kyoto, Japan. IEEE, 2010: 132-136.

[5] CHENG J S, YANG X K, et al. A method of planetary gearboxes fault diagnosis based on ASTFA and SDEO demodulation[J]. Noise and Vibration Control, 2017, 37(2): 137-142.

[6] HUANG W T, FU Q, DOU H Y. Resonance-based sparse signal decomposition based on the quality factors optimization and its application of composite fault diagnosis to planetary gearbox[J]. Journal of Mechanical Engineering, 2016, 52(15): 44-51.

[7] LI Y D, HAO Z B, LEI H. Summary of the convolutional neural network[J]. Journal of Computer Applications, 2016, 36(9): 2508-2515, 2565.

[8] WANG J J, ZHUANG J H, DUAN L X, et al. A multi-scale convolutional neural network for featureless fault diagnosis[C]. 2016 International Symposium on Flexible Automation, Cleveland, Ohio, U.S.A., 1-3 August, 2016.

[9] LU C, WANG Z Y, ZHOU B. Intelligent fault diagnosis of rolling bearing using hierarchical convolutional network based health state classification[J]. Advanced Engineering Informatics, 2017, 32: 139-151.

[10] ZHANG W, PENG G L, LI C H. Bearings fault diagnosis based on convolutional neural networks with 2-D representation of vibration signals as input[C]. MATEC Web of Conferences 95, 13001(2017).

[11] LIU R N, MENG G T, et al. Dislocated time series convolutional neural architecture: an intelligent fault diagnosis approach for electric machine[J]. IEEE Transactions on industrial informatics, 2017, 13(3): 1083-4435.

[12] XIA M, LI T, et al. Fault diagnosis for rotating machinery using multiple sensors convolutional neural networks[C]. IEEE/ASME Transactions on Mechatronics, 2017, 1-9.

[13] HUANG G, LIU ZHUANG, et al. Densely connected convolutional networks[EB/OL]. arXiv:1608.06993v3 [cs.CV] 3 Dec 2016.

[14] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[15] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]. Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2012: 1106-1114.

[16] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2009: 248-255.

[17] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. arXiv:1409.1556v6 [cs.CV] 10 Apr 2015.

[18] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[EB/OL]. arXiv:1409.4842v1 [cs.CV] 17 Sep 2014.

[19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[EB/OL]. arXiv:1512.03385v1 [cs.CV] 10 Dec 2015.

[20] CLEVERT D, UNTERTHINER T, HOCHREITER S. Fast and accurate deep network learning by exponential linear units (ELUs)[EB/OL]. arXiv:1511.07289v5 [cs.LG] 22 Feb 2016.

Author Biography:

SHAN Jianhua (Corresponding Author), male, born 1979, Ph.D., Professor. Research interests: deep learning, fault diagnosis.
E-mail: 379751793@qq.com

Note: Figure translations are in progress. See original paper for figures.