Knowledge-Based Systems 000 (2017) 1–12

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Convolutional neural network-based hidden Markov models for rolling element bearing fault identification

Shuhui Wang, Jiawei Xiang*, Yongteng Zhong, Yuqing Zhou

College of Mechanical and Electrical Engineering, Wenzhou University, Wenzhou, PR China

ARTICLE INFO

Article history:
Received 22 June 2017
Revised 19 December 2017
Accepted 25 December 2017
Available online xxx

Keywords:
Convolutional neural network
HMM
Feature extraction
Rolling element bearing
Fault diagnosis

ABSTRACT

Vibration signals of faulty rolling element bearings usually exhibit non-linear and non-stationary characteristics caused by the complex working environment. It is difficult to develop a robust method to detect faults in bearings based on signal processing techniques. In this paper, convolutional neural network - based hidden Markov models (CNN–HMMs) are presented to classify multi-faults in mechanical systems. In CNN–HMMs, a CNN model is first employed to learn data features automatically from raw vibration signals. By utilizing the t-distributed stochastic neighbor embedding (t-SNE) technique, feature visualization is constructed to manifest the powerful learning ability of CNN. Then, HMMs are employed as a strong stability tool to classify faults. Both the benchmark data and experimental data are applied to the CNN–HMMs. Classification results confirm the superior performance of the present combination model by comparing with CNN model alone, support vector machine (SVM) and back propagation (BP) neural network. It is shown that the average classification accuracy ratios are 98.125% and 98% for two data series with agreeable error rate reductions.

1. Introduction

Failure of mechanical components may lead to machine breakdown, potentially causing economic loss, or even catastrophic results. Thus, it is essential to accurately detect incipient and imminent faults in the components of various mechanical structures [1,2]. Rolling element bearings are one such type of mechanical component.

Vibration analysis [3–10] is the most commonly used technique for rolling element bearing fault diagnosis. Randall et al. [3] applied three established techniques to benchmark vibration signals. Villa et al. [4] presented a vibration analysis-based method in wind turbines fault diagnosis. Feldman et al. [5] reported on the Hilbert transform applied to mechanical vibrations. Seshadrinath et al. [6] proposed a vibration analysis-based inter-turn fault diagnosis method in induction machines. Wavelet-based techniques [7,8] and intelligent fault diagnosis [9,10] using vibration signals were applied to rolling element bearings and gears respectively. However, as a result of the sophisticated working environment, the fault features in the vibration signals are usually obscured, which makes fault diagnosis difficult. Therefore, accurate extraction of fault features in mechanical fault diagnosis is a significant accomplishment. Time domain features and frequency domain features are sometimes not enough for fault diagnosis due to the complexity of the signals. Therefore, temporal-frequency methods and adaptive analysis methods have been developed, such as enhanced empirical mode decomposition [11], local mean decomposition [12], PCA method [13], and spectral kurtosis (SK) techniques [14–16]. Nevertheless, these signal processing techniques suffer from limited ability to represent features, and may even reduce original fault feature to some extent.

With the widespread development of intelligent fault diagnosis [17–23], much research is being conducted on mechanical fault diagnosis. Rafiee et al. exploited artificial neural networks to realize intelligent condition monitoring of a gearbox [17]. Zhang et al. utilized a support vector machine-based method for intelligent fault diagnosis of rotating machinery [18,19]. Wei et al. achieved intelligent fault diagnosis via adaptive feature selection [22]. Meanwhile, deep learning models have also been applied to intelligent fault diagnosis by many researchers. Shao et al. [24] applied the PSO method to optimize a deep belief network to achieve bearing fault diagnosis. Ince et al. [25] exploited 1D convolutional neural network for real-time motor condition monitoring. Sun et al. [26] demonstrated an automatic fault recognition model based on convolutional neural networks for multi-fault identification in trouble of running freight train detection systems (TFDSs). Lee et al. [27] performed process fault detection and classification based on a convolutional neural network model for semiconductor manufacturing. Wang et al. [28] presented an optimized convolutional

* Corresponding author.
E-mail address: jwxiang@wzu.edu.cn (J. Xiang).

https://doi.org/10.1016/j.knosys.2017.12.027
0950-7051/© 2017 Elsevier B.V. All rights reserved.

neural network for rolling element bearing fault diagnosis. In [29], Janssens et al. pointed out the superior feature-learning ability of the CNN model and gave the conclusion that CNN-based methods yield great classification results without extensive domain knowledge. Jing [30] proposed an adaptive multi-sensor data fusion method based on convolutional neural networks to detect faults in a planetary gearbox. In contrast to the traditional intelligent fault diagnosis method, deep learning-based methods that originate from artificial neural networks have a deep architecture which exhibits brilliant feature learning abilities. Deep learning emphasizes the learning of features layer by layer [31]. The input data to a deep model are somewhat prepared at a previous hidden layer and then fed into the next hidden layer via feature transformation. Each layer has a different expression of the input data, and with multiple layers, the last hidden layer can represent the input data with certain explicit features. Therefore, deep learning eventually achieves feature extraction with constant iteration and abstraction processes arranged hierarchically.

Hidden Markov models (HMMs) have strong ability in dynamic time series modeling and pattern recognition [32–34]. Edmondo et al. [33] applied artificial neural networks to emission probabilities estimation within HMMs as a combination model to recognize speech signals. Volkmar et al. [34] investigated various methods of combining classic HMMs with maximum margin HMMs and neural networks for offline handwritten test recognition. In recent decades, many researchers have introduced HMMs into rotating machinery fault diagnosis [35–37]. Ocak et al. [35] modeled the normal bearing state using HMMs and tracked the rolling element bearings states by the HMM probabilities attenuation. Tobon-Mejia et al. [36] performed HMM to continuously assess the state and estimate the remaining useful life of mechanical components. Boutros and Liang [37] applied HMM to detect mechanical faults of cutting tools, as well as bearings, which obtained an approximately 95% fault severity classification accuracy and an approximately 96% fault location classification accuracy, respectively. Ocak and Loparo [38] proved the high accuracy of an HMM-based fault diagnosis scheme for representing various bearing states.

Based on the strong feature learning ability of CNN models and the excellent pattern recognition capacity of HMMs, a convolutional neural network-based hidden Markov model is proposed to identify faults in rolling element bearings. The performance of the combination model is investigated with benchmark data from Case Western Reserve University (CWRU) and experimental data from the Machinery Fault Simulator Magnum. This model takes advantage of both the CNN and HMMs for their strong ability in data feature learning and pattern recognition, respectively.

The remainder of this paper is set out as follows: descriptions of CNNs and HMMs are given in Section 2. Section 3 gives the description for the proposed model. Experimental investigations are conducted in Section 4 using two case studies with the proposed model. Finally, the conclusion is given in Section 5.

2. Theoretical background

2.1. Brief description of the CNN

A CNN is a kind of deep learning model with a distinctive architecture, namely a convolutional layer and a subsampling layer. There are three main traits of a CNN model, which are, local field, subsampling and weight sharing. Lecun and his colleagues [39] first proposed a commonly used CNN architecture for handwriting recognition, which now has become a benchmark for deep learning models. As shown in Fig. 1, there exist three main layers in the architecture, the input layer, the hidden layers and the output layer. The hidden layers consist of the convolutional layers and subsampling layers. Each convolutional layer is directly followed by a subsampling layer. The CNN model was originally designed to process 2D maps [40]. For this purpose, we generally call the inputs to the input layer "input maps". Fig. 1 shows only one input map in order to simplify the illustration of the CNN model.

The architecture shown in Fig. 1 is Input (S_0)-C_1-S_1…-S_{l-1}-C_l-S_l…C_L-S_L-Output. When L convolution and subsampling operations are performed, the input data will transferred to output data through $2L$ operations.

Considering the q-th feature map $x_q^{S_{l-1}}$($q=1,2,\ldots,Q$, where Q is the number of feature maps in the hidden layer S_{l-1}), the generalized feature map $x_k^{C_l}$($k=1,2,\ldots,K$, where K is the number of feature maps in the hidden layer C_l) can be represented by

$$x_k^{C_l} = f\left(\sum_{q\in M_k} x_q^{S_{l-1}} * w_{qk}^{C_l} + b_k^{C_l}\right) \tag{1}$$

where $f(\bullet)$ is the output activation function, M_k represents a selection of feature maps in the hidden layer C_l, $w_{qk}^{C_l}$ is the weight matrix of kernel in the hidden layer C_l, $b_k^{C_l}$ is the bias vector in the hidden layer C_l.

A subsampling layer produces downsampled versions of the feature maps in the hidden layer C_l as

$$x_k^{S_l} = f\left(\beta_k^{S_l} down\left(x_k^{C_l}\right) + b_k^{S_l}\right) \tag{2}$$

where $x_k^{S_l}$ ($k=1,2,\ldots,K$, where K is the same number as shown in Eq. (1)) is the k-th feature map in the hidden layer S_l, $\beta_k^{S_l}$ is the k-th scaling factor in the hidden layer S_l, $b_k^{S_l}$ is the k-th bias in the hidden layer S_l, and $down(\bullet)$ represents a subsampling function. It points out that Q and K will be changed along with the Lth convolution and subsampling operations.

In general, CNN accomplishes its training process by a feedforward pass and a backpropagation pass. In the feedforward pass, the outputs of a previous layer are transmitted to the next layer. In the backpropagation pass, the training error is propagated backward hierarchically updating the weights and biases of each layer [39]. The training procedures are presented in detail in [41].

2.2. Conventional GMM-based HMM model

An HMM is a statistical model that learns from input data. As a generative model, it learns the joint probability $p(input,label)$ from inputs and corresponding labels. Therefore, classification can be achieved by calculating conditional probability $p(label|input)$ using Bayes' theorem via maximum likelihood.

An HMM is a dual stochastic process that contains invisible hidden states $S=\{S_1,S_2,\cdots,S_n,\cdots,S_N\}$ and observation sequences. The observation sequence $O=\{O_1,O_2,\cdots,O_t,\cdots,O_T\}$ indicates the existence of a same length hidden states sequence of N states. Thus, we define an HMM h by

$$h=\{\pi, A, B\} \tag{3}$$

where π represents the prior probabilities, an N-length vector that denotes the probability of S_n being the first states in a hidden states sequence. A is an $N\times N$ matrix containing the transition probabilities among the hidden states. B represents the emission probabilities transferring hidden state S_n to observation value O_t. For a GMM-HMM, B is described by the Gaussian mixture model

$$gmm(O)=\sum_{m=1}^{M} w_m \cdot g_{(\mu_m,\sum_m)}(O) \tag{4}$$

where $gmm(\bullet)$ is the mixture of M Gaussians, $g_{(\mu_m,\sum_m)}(\bullet)$ is the Gaussian probability density function (G-pdf), μ_m, Σ_m denote the

Please cite this article as: S. Wang et al., Convolutional neural network-based hidden Markov models for rolling element bearing fault identification, Knowledge-Based Systems (2017), https://doi.org/10.1016/j.knosys.2017.12.027