Mechanical Systems and Signal Processing 99 (2018) 459–477

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

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

Intelligent condition monitoring method for bearing faults from highly compressed measurements using sparse over-complete features

H.O.A. Ahmed [a], M.L.D. Wong [b], A.K. Nandi [a,c,*]

[a] Department of Electronic and Computer Engineering, Brunel University London, Uxbridge UB8 3PH, United Kingdom
[b] Heriot-Watt University Malaysia, Precinct 5, 62200 Putrajaya, Malaysia
[c] The Key Laboratory of Embedded Systems and Service Computing, College of Electronic and Information Engineering, Tongji University, Shanghai, China

ARTICLE INFO

Article history:
Received 30 July 2016
Received in revised form 26 April 2017
Accepted 20 June 2017
Available online xxxx

Keywords:
Compressed sensing
Sparse over-complete representations
Deep neural network
Sparse autoencoder
Bearing fault classification
Machine condition monitoring

ABSTRACT

Condition classification of rolling element bearings in rotating machines is important to prevent the breakdown of industrial machinery. A considerable amount of literature has been published on bearing faults classification. These studies aim to determine automatically the current status of a roller element bearing. Of these studies, methods based on compressed sensing (CS) have received some attention recently due to their ability to allow one to sample below the Nyquist sampling rate. This technology has many possible uses in machine condition monitoring and has been investigated as a possible approach for fault detection and classification in the compressed domain, i.e., without reconstructing the original signal. However, previous CS based methods have been found to be too weak for highly compressed data. The present paper explores computationally, for the first time, the effects of sparse autoencoder based over-complete sparse representations on the classification performance of highly compressed measurements of bearing vibration signals. For this study, the CS method was used to produce highly compressed measurements of the original bearing dataset. Then, an effective deep neural network (DNN) with unsupervised feature learning algorithm based on sparse autoencoder is used for learning over-complete sparse representations of these compressed datasets. Finally, the fault classification is achieved using two stages, namely, pre-training classification based on stacked autoencoder and softmax regression layer form the deep net stage (the first stage), and re-training classification based on backpropagation (BP) algorithm forms the fine-tuning stage (the second stage). The experimental results show that the proposed method is able to achieve high levels of accuracy even with extremely compressed measurements compared with the existing techniques.

1. Introduction

Rolling element bearings are among the most fundamental elements in rotating machinery and their failures are accountable for more substantial failures in the machine. Thus, roller bearings require an effective condition monitoring and machinery maintenance program to avoid machine breakdowns. In fact, machines' availability and functions can be monitored by

* Corresponding author at: Department of Electronic and Computer Engineering, Brunel University London, Uxbridge UB8 3PH, United Kingdom.
E-mail address: asoke.nandi@brunel.ac.uk (A.K. Nandi).

http://dx.doi.org/10.1016/j.ymssp.2017.06.027
0888-3270/

accurate, rapid and automatic fault detection techniques. Therefore, rolling element bearing condition monitoring (CM) [1] has attracted considerable attention from researchers in the past decades. In the course of their operation, rotating machines produce signals in different forms, e.g. noise, vibration, temperature, lubricating oil condition, etc. [2]. Various characteristic features can be observed from vibration signals that make it the best choice for machine condition monitoring. Vibration signal analysis can be performed in three main groups – time domain, frequency domain, and time-frequency domain analysis [3–6]. Time domain techniques extract features from the raw vibration signals using some statistical parameters, these include, peak-to-peak value, root mean square, crest Factor, skewness, kurtosis, Impulse factor, etc. [7]. The frequency domain analysis techniques have the ability to divulge some information based on the frequency characteristics that are not easily observed in time-domain. In practice, the time-domain signal is transformed into frequency-domain by using Fast Fourier Transform (FFT). The time-frequency domain has been used for non-stationary waveform signals which are very common when machinery fault occurs. Thus far, several time–frequency analysis techniques have been developed and applied to machinery fault diagnosis, e.g., Wavelet Transform (WT), Short Time Fourier Transform (STFT) adaptive parametric time-frequency analysis based on atomic decomposition, and non-parametric time-frequency analysis, including, Hilbert-Huang Transform (HHT), local mean decomposition, energy separation and empirical mode decomposition [8–15]. Spectral Kurtosis (SK) has been used effectively in the vibration-based condition monitoring of rotating machines. In this method, signal is first decomposed into the time-frequency domain where the kurtosis values are defined for each frequency group. In addition, Kurtogram concept has been proposed to increase the signal-to- noise ratio [16], and have been effectively used for both vibration and acoustic emission [17]. An alternative framework to strictly stationary vibration signal processing methods, cyclostationary analysis has been used for analysing vibration signals [18,19].

All the above techniques use data that have been recorded satisfying the Shannon/Nyquist sampling theorem, in which the sampling rate must be at least double the maximum frequency present in the signal. One aspect of much of the literature on using Nyquist sampling rate is that it may result in measuring a large amount of data. It is clear that acquiring a large amount of data requires large storage and time for signal processing and this also may limit the number of machines that can be monitored remotely across wireless sensor networks (WSNs) due to bandwidth and power constraints.

Compressive sensing (CS) also called compressed sensing or compressed sampling [20,21] is a new technique that supports sampling below the Nyquist rate. CS is being considered in a large diversity of applications including medical imaging, seismic imaging, and radio detection and ranging, communications and networks [22–27]. The basic idea of CS is that original signals can be reconstructed from fewer measurements far below than the Shannon sampling rate using sparse representation and a well-designed measurement matrix. The literature on compressive sensing shows a variety of approaches to reconstruct signals from few measurements [28–31]. In an analysis of effects of compressive sensing on the classification of bearing faults after reconstructing the original signal, Wong et al. [32] found slight performance degradation with a large reduction of bandwidth constraint caused by using CS. Similarly, Li et al. in [33] shown the possibility to detect the fault of train's rolling bearing from the reconstructed signal based on compressive sensing. However, signal reconstruction techniques may not be practical in all applications and make no attempt to address the question of whether or not it is possible to learn in the compressed domain rather than recovering original signals. For instance, bearing vibration signal is always acquired for faults detection and estimation, and as long as it is possible to detect faulty signals in the measurement domain, then it is not necessary to recover the original signal to identify faults. A significant analysis and discussion on the subject of how to solve a range of signal detection and estimation problems given compressed measurements without reconstructing the original signal was presented by Davenport et al. in [34].

Over the past years, most research in compressive sensing based methods has emphasized the use of compressed measurements, sparse representations and incomplete signal reconstruction for the bearing fault diagnosis. For example, Tang et al. [35] developed a sparse classification strategy based on compressive sensing by extracting and classifying fault features through sparse representation combined with random dimensionality reduction. Zhang et al. [36] suggest a bearing fault diagnosis method based on the low-dimensional compressed vibration signal by training several over-complete dictionaries that can be effective in signal sparse decomposition for each vibration signal state. Another learning dictionary basis for extracting impulse components is described by Chen et al. in [37]. Tang et al. [38] proposed an interesting approach in which authors attempted to observe the characteristic harmonics from sparse measurements through a compressive matching pursuit strategy during the process of incomplete reconstruction. The outcomes of these studies corroborate the efficiency of compressive sensing in machinery fault diagnosis.

Even though the efficiency of CS in machine fault diagnosis has been validated in these studies, most of the fault classification improvements in the literature were achieved by increasing the sampling rate, otherwise highly compressed measurements attained poor classification accuracy. The aim of this work is to improve the efficiency of using highly compressed measurements for signal classification purpose. With this goal, this work explores the possibility of learning sparse overcomplete representations from highly compressed measurements in an unsupervised manner based on a deep learning approach.

Various deep learning architectures, e.g., Convolution Deep Neural Networks (CNNs), Deep Belief Networks (DBNs), Recurrent Neural Networks (RNNs), and stacked Autoencoders (AEs), have been used for reducing dimensionality or extracting features from signals. Unlike the standard Neural Network (NN), the architecture of CNN is usually composed of a convolutional layer and a sub-sampling layer, also called a pooling layer. CNN learns abstract features by alternating and stacking convolutional layers, and pooling operation. The convolution layers convolve multiple local filters with raw input data and generate invariant local features and the pooling layers extract most significant features. DBNs are generative