Health state classification of a spherical tank using a hybrid bag of features and k-nearest neighbor.

HASAN, M.J., KIM, J. and KIM, J.-M.

2021

Health State Classification of a Spherical Tank Using a Hybrid Bag of Features and k-Nearest Neighbor

Md Junayed Hasan, Jaeyoung Kim, and Jong-Myon Kim

Department of Electrical and Computer Engineering, University of Ulsan, Ulsan, South Korea

Abstract Feature analysis plays an important role in determining the various health conditions of mechanical vessels. To achieve balance between traditional feature extraction and the automated feature selection process, a hybrid bag of features (HBoF) is designed for the health state classification of spherical tanks in this paper. The proposed HBoF is composed of (a) the acoustic emission (AE) features, and (b) the time and frequency based statistical features. A wrapper-based feature selector algorithm, Boruta, is applied to extract the most intrinsic feature set from HBoF. The selective feature matrix is passed to the k-nearest neighbor (k-NN) classifier to distinguish between normal condition (NC) and faulty condition (FC). Experimental results show that the proposed approach yields an average 100% accuracy for all working conditions. The proposed method outperforms the existing state-of-the-art approaches by achieving at least 19% higher classification accuracy.

Keywords Spherical tank; AE features; Boruta; Fault diagnosis

1 Introduction

Mechanical vessels play a very important role in day-to-day life with widespread application [1]. Specifically, in the oil and gas industry, the use of spherical tanks is required due to the cost effectiveness of building a sphere. With the increasing use of these types of spherical tanks for different industries, the number of accidents related to leakage from the bottoms of these tanks is also increasing [2]. As a result, improved safety precautions and maintenance are required [3, 4].

In this experiment, the main emphasis is on health state categorization through signals acquired from a spherical tank. Identifying the health condition (normal or faulty state) through signals at an early stage will make it easier to determine the necessary precautions at later stages. The acoustic emission (AE) velocity signals are considered for classification of the health state. Compared with old-style methods, AE is an economical and efficient detection process [5]. Additionally, AE signals can provide underlying information from low energy signals [6, 7] for a more substantial data-driven fault identification approach. AE-based diagnosis methods mostly rely on a procedure for analyzing the peak of the characteristic frequencies of the signals [8]. Pattern generation from acquired signal domains using several signal-imaging techniques can also differentiate between health conditions for further classification [9]. Several automated feature learning processes driven by deep learning-based algorithms have been studied to reduce the necessity of domain knowledge expertise [9–11]. Due to limitations in the amount of data, deep learning-based approaches are not capable of extracting meaningful features.

Herein, a data-driven hybrid feature extraction process is considered. The main contributions of this research can be summarized as follows. (1) An HBoF extraction method is designed by combining two types of analysis: analysis of the AE signal properties, and of the time-domain and frequency-domain based statistical properties; and (2) a wrapper-based non-redundant feature selection method, Boruta, is utilized to analyze all the key elements of the hybrid feature pool. Finally, the k-nearest neighborhood (k-NN) is applied for classification of the health state, using those selected features as input.

The rest of the paper is structured as follows. Section 2 provides details of the methodology, including the AE data acquisition system. The analysis of the experimental results and comparative discussions are provided in Sect. 3. The paper is concluded in Sect. 4.

2 Proposed Method

The proposed approach is divided into four sections: (1) data collection from a multisensory testbed, (2) feature extraction by HBoF, (3) feature selection by Boruta, and (4) k-NN-based classification.

2.1 Experimental Testbed and Dataset Acquisition

An experiment is performed on a self-designed test platform to collect AE signals. One AE sensor (WDI-AST) with four different channels is attached to collect the velocity AE signals. On the test rig, there are four different crack positions (825, 750, 1040 and 430 mm) to collect the velocity data with 1 MHz sampling frequency. The signal is measured through a trigger-based measurement technique for a specific amount of time.

2.2 Hybrid Bag of Features

It is difficult to obtain intrinsic information for different health types from a raw signal. To create the health condition-based feature matrix, two different sets of features are considered. For the AE features, the amplitude (F1), rise time (F2), and duration (F3) of the signals are computed. For the threshold value, the rms of the signal is considered. The specifics of the AE features are demonstrated in Fig. 1. For statistical analysis, from the time domain, the numerical features obtained are root mean square (F4), kurtosis (F5), skewness (F6), shape factor (F7), and impulse factor (F8). In the same manner, from the frequency domain, the features obtained are root mean square (F9), kurtosis (F10), and skewness (F11). Thus, in total, 11 features are extracted to create the designed HBoF. In Table 1, the numerical details of these statistical features are described.

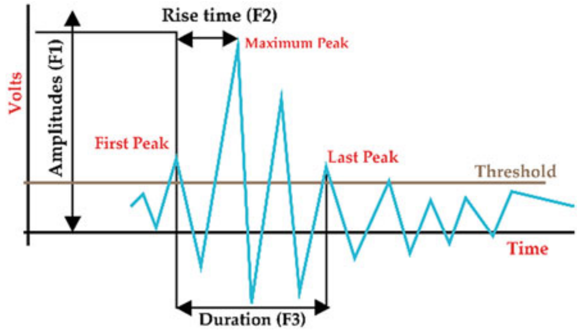

Fig. 1 Illustration of acoustic emission (AE) signal feature

Table 1 Numerical explanation of statistical features

Feature	Equation	Feature	Equation	Feature	Equation
F4	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} X_i^2}$	**F5**	$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i-\overline{X}}{\sigma}\right)^4$	**F6**	$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i-\overline{X}}{\sigma}\right)^3$
F7	$\frac{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i-\overline{X}}{\sigma}\right)^3}{\frac{1}{N}\sum_{i=1}^{N}\lvert X_i\rvert}$	**F8**	$\frac{\max(\lvert X\rvert)}{\frac{1}{N}\sum_{i=1}^{N}\lvert X_i\rvert}$		
F9	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} F_i^2}$	**F10**	$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{F_i-\overline{F}}{\sigma}\right)^4$	**F11**	$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{F_i-\overline{F}}{\sigma}\right)^3$

Here, x is the time-domain raw signal, and F is the frequency domain signal. N is the total number of samples

2.3 Feature Selection by Boruta

Boruta finds the most relevant and intrinsic feature information from data. As a first step, it duplicates the original feature set and then rearranges the feature values, which are called shadow features. Each of those shadow feature sub-sets is then trained by the random forest classifier to validate the significance of the important feature set by the mean decrease impurity (MDI) matrix. If the MDI value is higher, then the set is important. In the second step, it runs a similar test for the original feature set. For this test, Z score is calculated. Z score is the number of standard deviations a measure is from the mean. The algorithm considers whether the original set of features has a higher Z score than most of its shadow features. If the score is high, it is logged as a vector named hits. Thus, the iteration is continued till reaching the predefined set of iteration numbers and, at the end, a hit table is generated.

2.4 Fault Classification Using k-Nearest Neighbor (k-NN)

To validate the considered optimal feature sets in terms of classification performance, a k-NN classifier is used. k-NN has a simple architecture with less computational complexity [6]. k-NN categorizes the trials relying on the votes of the k-nearest neighbors, which are identified by certain distance parameters [12].

3 Experimental Result Analysis and Discussion

3.1 Dataset

The standard AE dataset of spherical tanks is used to conduct a test. A 0.1 s velocity signal with 1 MHz sampling frequency is used for consideration of each health state (NC and FC). The particulars of the dataset are provided in Table 2.

Table 2 Details of the considered dataset

Health condition	Crack type	Crack size (mm)	Channels
Normal condition (NC)	No crack	No crack	4
Faulty condition (FC)	Pinhole crack	3	4

3.2 Result Analysis

Raw AE signals have no intrinsic information to reveal different health conditions. Therefore, the HBoF is designed and Boruta is applied to get the most intrinsic feature information. From Boruta, the five most important features are calculated (i.e., F1, F3, F4, F5, and F10). These five features are collected from AE analysis, time domain, and frequency domain. The robustness of the HBoF is shown by considering all the important information from the signals.

The selected features from Boruta are each provided to the k-NN. The dataset is divided into training and testing sets at the respective proportion of 60/40. Sensitivity is considered for calculating the classwise accuracy. The final classification accuracy is obtained after 6-fold cross validation. The proposed approach achieves 100%classification accuracy when the optimal value of k is 8 in k-NN algorithm (illustrated in Fig. 2b). Along with the proposed approach, several comparisons are made to establish the robustness. From the HBoF, for feature selection, instead of Boruta, non-dimensional feature reduction techniques such as PCA and t-SNE are applied to get the intrinsic feature information for final classification. In Table 3, the classification

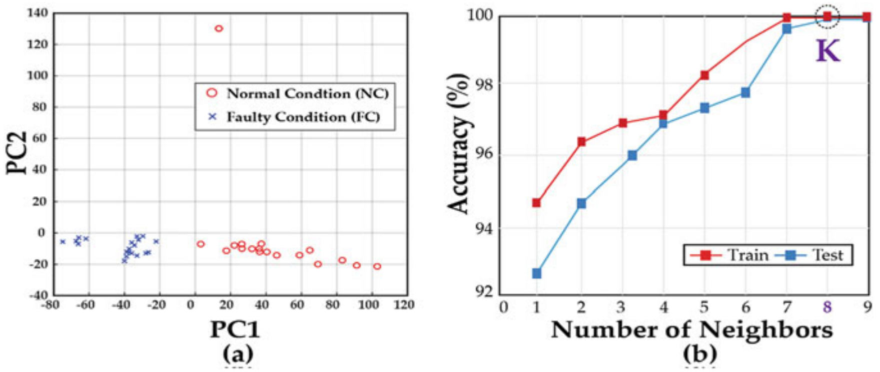

Fig. 2 **a** Boruta feature space. The five features selected from Boruta are embedded into 2D space by PCA for visualization purposes only. **b** Various categorization accuracies as a function of the number of neighbors (k). The optimal value is k = 8

Table 3 Classification accuracy of different methods

Approach	Classification accuracy (%)		Average classification accuracy (%)	Decrement from the proposed method (%)
	NC	FC		
Proposed	100	100	100	–
HBof + t-SNE + k-NN	75.5	35	55.25	44.75
HBoF + PCA + k-NN	79.5	82.5	81	19

accuracies from different approaches are described in a very detailed way. From the information shown there, the necessity of finding out the ranked features is demonstrated, as opposed to keeping all the information.

4 Conclusion

This paper presented a hybrid feature selection method called HBoF, which is composed of AE feature analysis and statistical information from time and frequency analysis. To select the most intrinsic features from the proposed HBoF, feature wrapper Boruta is applied. Thereafter, k-NN is used for final classification, which leads to a 100% average accuracy for both normal and faulty conditions (NC and FC). Comparative analysis with different non-linear feature dimensionality reduction techniques (i.e.; PCA, and t-SNE) was performed to validate the performance. The proposed approach outperformed the PCA and t-SNE based methods by respective 19% and 44.75% classification accuracies.

Acknowledgements This work was supported by the Technology Infrastructure Program funded by the Ministry of SMEs and Startups (MSS, Korea).

References

1. Saidur R (2010) A review on electrical motors energy use and energy savings. Renew Sustain Energy Rev 14:877–898
2. Morofuji K, Tsui N, Yamada M, Maie A, Yuyama S, Li ZW (2003) Quantitative study of acoustic emission due to leaks from water tanks. Group. 21:213–222
3. Luo T, Wu C, Duan L (2018) Fishbone diagram and risk matrix analysis method and its application in safety assessment of natural gas spherical tank. J Clean Prod 174:296–304
4. Korkmaz KA, Sari A, Carhoglu AI (2011) Seismic risk assessment of storage tanks in Turkish industrial facilities. J Loss Prev Process Ind 24:314–320
5. Li W, Dai G, Wang Y, Long F (2011) Study of tank acoustic emission testing signals analysis method based on wavelet neural network. In: ASME pressure vessels and piping conference
6. Pandya DH, Upadhyay SH, Harsha SP (2013) Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN. Expert Syst Appl 40:4137–4145
7. Niknam SA, Songmene V, Au YHJ (2013) The use of acoustic emission information to distinguish between dry and lubricated rolling element bearings in low-speed rotating machines. Int J Adv Manuf Technol 69:2679–2689
8. Kang M, Kim J, Kim J (2015) High-performance and energy-efficient fault diagnosis using effective envelope analysis processing unit. IEEE Trans Power Electron 30:2763–2776
9. Amar M, Gondal I, Wilson C (2015) Vibration spectrum imaging: a novel bearing fault classification approach. IEEE Trans Ind Electron 62:494–502
10. Sohaib M, Kim C-H, Kim J-M (2017) A hybrid feature model and deep-learning-based bearing fault diagnosis. Sensors 17:2876
11. Hasan MJ, Sohaib M, Kim J-M (2019) 1D CNN-based transfer learning model for bearing fault diagnosis under variable working conditions
12. Chen X, Xu J-B, Guo W-Q (2013) The research about video surveillance platform based on cloud computing. In: 2013 international conference on machine learning and cybernetics. IEEE, pp 979–983