Effective data pre-processing methods of anomaly detection for bridge bearings

Kawakin Core-Tech Co., Ltd. Regular Member ○Rongzhi Zuo
Saitama University Regular Member Ji Dang
Kawakin Core-Tech Co., Ltd. Kazuhiro Shimizu
Kawakin Core-Tech Co., Ltd. Akira Tsunoda
Kawakin Core-Tech Co., Ltd. Yasuhiro Suzuki

1. INTRODUCTION

Anomaly detection has been studied and applied in various industrial fields to ensure the safety of products [1]. For applying machine learning (ML) solutions to the interested product, bridge bearing, displacement measurement of a set of bridge bearings was conducted. Three data pre-processing methods were compared to demonstrate the necessary feature of a good-quality dataset of bridge bearings. A deep autoencoder model (DAE) was proposed and tested on three datasets respectively, the results show that the model trained on the short-time waveform samples has 100% case-level accuracy, in the meanwhile, the model performs better when the waveforms are more sampled from the time when the bearing is under live load. This paper also introduced a data mining method matrix profile (MP) [2] for sampling more representative short-time waveforms, which means the displacement under live load.

2. DATASET PREPARATION AND DAE MODEL

The measurement data is the horizontal displacement of the movable part of bridge bearings along the longitudinal direction of decks. The duration of the measurement is between 1 hour to 24 hours with a sampling rate of 100 Hz. To sample the short-time representation of the response of bridge bearings in service state, three sampling techniques were applied to prepare the training dataset respectively. The first one (a) is random sampling, randomly slicing 500 5.12-second waveforms per bearing case for setting the baseline of the dataset. The remaining two were proposed considering the performance representativeness of short-time waveforms, whether or not under live load might draw a considerable difference. If the waveform under no live load is not representative, inferring such data will be challenging. The second one (b), sliding window with descending order by the amplitude of short-time waveforms, increases the probability of data under live load. The third one (c) is similarity selection based on the motif result of MP analysis. MP can measure the similarity of the sub-sequence data at a given time window throughout the target time series data. The MP has two primary components; a distance profile and profile index. The distance profile is a vector of minimum Z-Normalized Euclidean Distances; Z-score normalization refers to the process of normalizing every value in a dataset such that the mean of all of the values is 0 and the standard deviation is 1, pairwise Euclidean Distances are compared and only the minimum values are kept per sub-sequence iteration. The profile index contains the location of its most similar sub-sequence, therefore multiple sub-sequences can be selected as a chain set. Based on the two components, the short-time waveform can be selected with high sub-sequence similarity, which means less pure white noise data. As a bonus, clustering analysis on top of MP analysis provides a good visualization of selected sub-sequence, cluster examples are shown in Figure 1, which can lead to a high-accuracy ML solution with human-in-the-loop design architecture. The selected waveforms were normalized by min-max normalization before feeding into the DAE model. As for the label of measurement cases, 20 cases are labeled as non-damage; of which 12 cases are after bridge bearing replacement, 2 cases are new bridge bearings, and the other 6 cases are human-labeled by inspection experts. There are 2 cases are human-labeled damage. The dataset splitting is shown in Table 1, automatically labeled ones are in the training group and human-labeled ones are in the test group. The DAE model was composed of convolution layers under an encoder-decoder architecture shown in Figure 2.

Table 1: Dataset splitting

Dataset	Cases	Size	Label
Training	12	(6000,512)	Non-damage
Validation	2	(1000,512)	Non-damage
Test-1	2	(1000,512)	Damage
Test-2	6	(3000,512)	Non-damage

3. RESULTS OF DAMAGE DETECTION

Considering the real application is on the case level, the case-level classification result was voted by the selected 500 examples with a threshold of 50%. The loss distributions for three datasets are demonstrated in Figure 3, Figure 4, and Figure 5; except for dataset a, loss clearly distributes towards opposite directions according to the label difference, non-damage ones have a smaller loss, and damage ones have a greater loss. The overall accuracy is 100% on both dataset b and dataset c at the case level, and the accuracy surpasses the threshold of 50% greatly at the sample level. The loss

Keywords: bridge bearing, anomaly detection, deep autoencoder, deep learning, matrix profile
Contact address: 2-2-7 Kawaguchi, Kawaguchi-shi, Saitama, Japan, 332-0015 Tel: +81-048-259-1113

distribution on dataset c is more concentrated, therefore the precision is increased as well.

4. CONCLUSIONS

This study demonstrated two techniques to decrease the occurrence of noise samples in the processed dataset, processing datasets through domain knowledge can boost model performance and provide an intuitive representation of sampled data. To align with the application side of anomaly detection, this research can be expanded as a regression problem to assess the urgency of countermeasures for functionally abnormal bridge bearings, and it will be studied with the database growth.

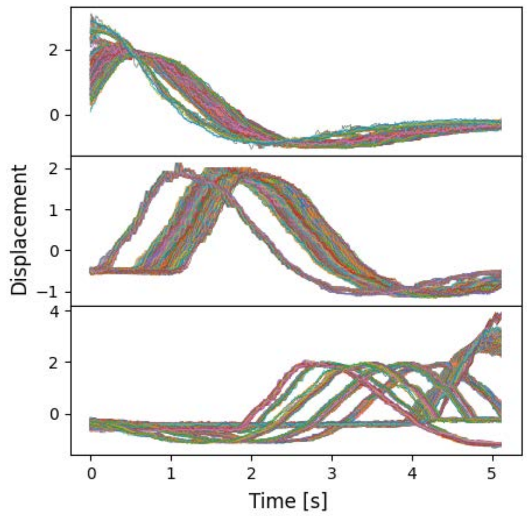

Fig. 1 Cluster examples of short–time waveforms

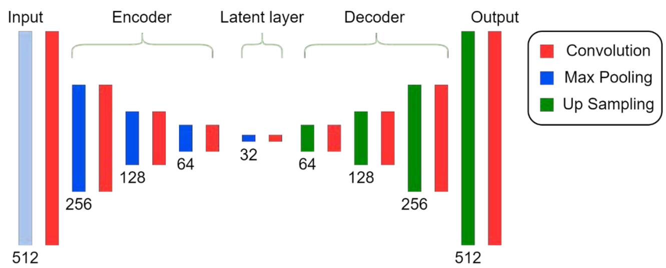

Fig. 2 The architecture of DAE model

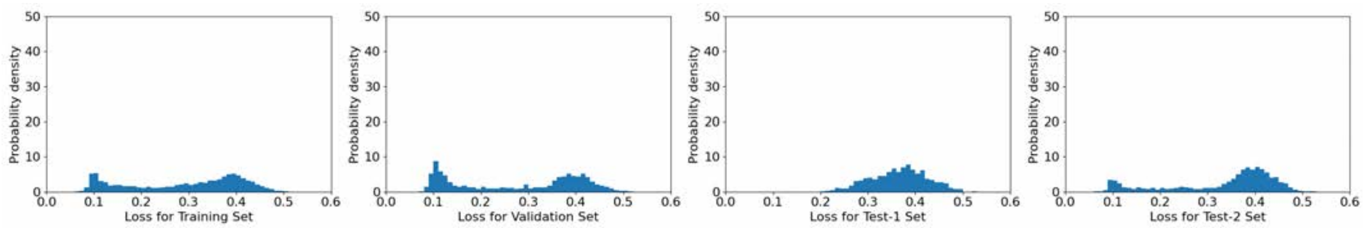

Fig. 3 Loss distribution for dataset a

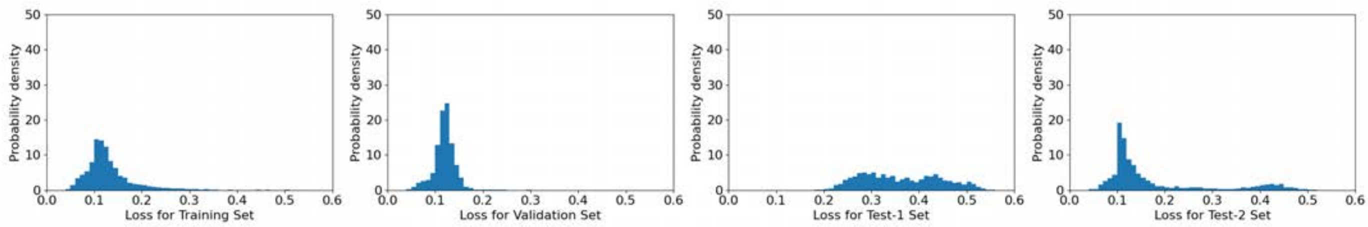

Fig. 4 Loss distribution for dataset b

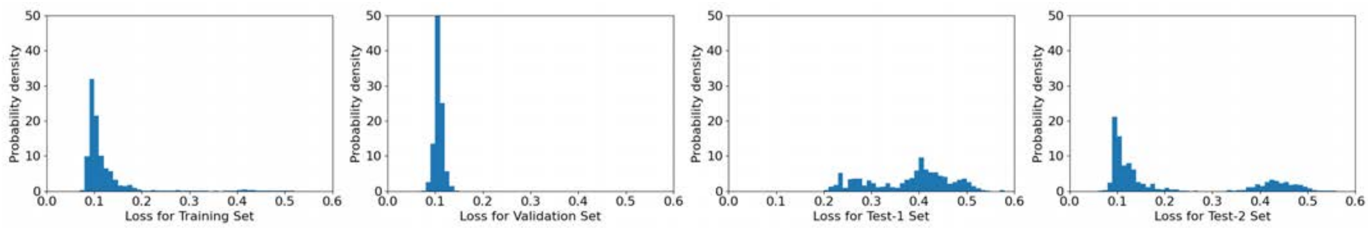

Fig. 5 Loss distribution for dataset c

REFERENCES

[1] Barry S.: Industrial Anomaly Detection: A Comparison of Unsupervised Neural Network Architectures, 2020, *IEEE Sensors Letters*

[2] Chin-Chia Y., Yan Z., Liudmila U., Nurjahan B., Yifei D., Hoang D., Diego S., Abdullah M., Eamonn K.: Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View that Includes Motifs, Discords and Shapelets, *2016, IEEE ICDM 2016*