Vibration Signal Generation Using Conditional Variational Autoencoder for Class Imbalance Problem

J. U. Ko[1], M. Kim[1], H. B. Kong[1], J. Lee[1] and B. D. Youn[1, 2*]

[1]Department of Mechanical and Aerospace Engineering, Seoul National University, Seoul 08826,
Republic of Korea
[2]Onepredict Inc., Seoul 08826, Republic of Korea

[*]Corresponding author: bdyoun@snu.ac.kr

1. Introduction

Nowadays, a large amount of data is obtained from the industrial fields such as power plants and automated factories owing to the development of sensor techniques. In many cases, this industrial data is class-imbalanced which is composed of large normal and a little of fault data since most engineering system is designed not to fail. There are three types of solutions for class imbalance problem: algorithm level, cost level, and data level approach. Algorithm level method mandates a classifier to be biased toward the minor classes [1] and cost level approach also biases a classifier by implementing a misclassification loss [1]. Data level method reduces the amount of data of major classes (down-sampling) or increases the amount of data of minor classes (over-sampling). Among those three approaches, prior knowledge of both the classifier and the application domain is needed to adopt the algorithm level approach, and the misclassification loss of cost level method is hard to be defined [1]. Therefore, we adopted over-sampling by generating new fault data instead of algorithm and cost level approaches.

To generate fault data, we employed conditional variational autoencoder (CVAE) [2], which is a powerful generative model when label information is given. Multi-layer perceptron (MLP) based CVAE model is developed and validated by the testbed data obtained from Bentley Nevada RK4 testbed. The generated data is quite similar to the raw vibration signals.

2. Conditional Variational Autoencoder (CVAE)

Conditional variational autoencoder (CVAE), a variation of vanilla variational autoencoder (VAE), generates a data using label information. The basic architecture of it is shown in Fig. 1.

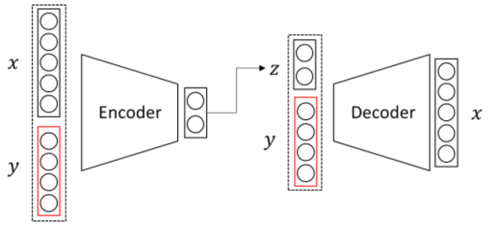

Fig. 1 Architecture of CVAE

Here, *x* is the input data, *y* is the corresponding label vector, *z* is the vector of latent variables which are encoded by encoder.

Eq. (1) is the loss function of CVAE. The first term is the reconstruction loss and the second one is a regularization term which makes the distribution *q*(·) closer to prior distribution *p*(·). By minimizing Eq. (1), CVAE learns the data distribution and it is able to generate new data of given label. We can sample z from *p*(·) since *q*(·) becomes similar to *p*(·) after training.

$$\tilde{L}_{CVAE}\left(\mathbf{x},\mathbf{y};\theta,\phi\right)=\frac{1}{m}\sum_{i=1}^{m}\log p_{\theta}\left(\mathbf{y}\,|\,\mathbf{x},z^{(i)}\right)-KL\left(q_{\phi}\left(\mathbf{z}\,|\,\mathbf{x},\mathbf{y}\right)\|\,p_{\theta}\left(\mathbf{z}\,|\,\mathbf{x}\right)\right) \quad (1)$$

3. Proposed Vibration Signal Generative Model

We developed a vibration signal generative model using concept of CVAE. The proposed model is illustrated in Fig. 2.

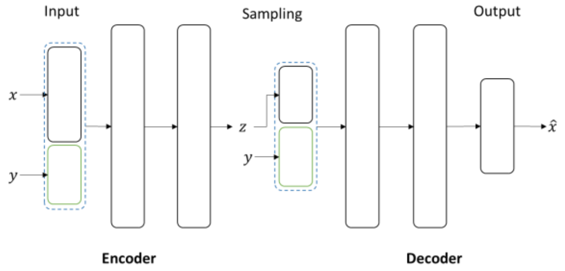

Fig. 2 The architecture of proposed model

x is vibration signal and *y* is the one-hot label vector. Thus, the number of nodes at input layer is the sum of the length of signal and the label vector. Encoder and decoder are made up of three hidden layers, whose activation functions are exponential linear unit, hyperbolic tangent function, and linear function at last layer. After training the model, we only used decoder part to generate new fault data.

4. Case Study

We tested the performance of proposed generative model with RK4 testbed data. The fault classes are unbalance, misalignment, rubbing and oil whirl. The input sequence length is 128, which contains around 4 cycles of sample points.

After training, we produced data with respect to various latent vector *z*. The generated data is shown in Fig. 3. In the figure, red line is the true vibration signal of each fault and blue line is the generated data by the proposed model.

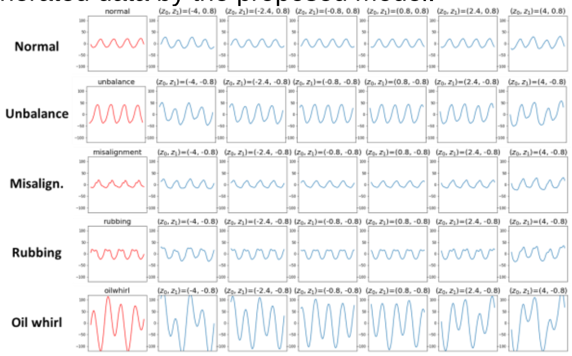

Fig. 3 Generated fault data by proposed model

As you can see in Fig. 3, all data shows very similar time-series trend to the true data. For example, the generated signals of unbalance have larger amplitude than normal data, and the rubbing signals appear to be cut off near the peak points.

5. Conclusion

We proposed a generative model using CVAE to mitigate the class imbalance problem by producing scarce fault data. The generated data of each fault class is very close to the true fault data. In the future, we will validate the proposed model by more various data to check whether the model simply memorizes the training data or truly learns the data distribution.

Acknowledgment

This work was supported by the R&D project (R17GA08) of Korea Electric Power Corporation (KEPCO).

References

[1] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid approaches, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and REVEIWS)*, 42 (4), (2011), 463-484.

[2] K. Sohn, X. Yan, and H. Lee, Learning structured output representation using deep conditional generative models, *Advances in neural information processing systems*, (2015), 3483-3491.