A data enlargement strategy for fault classification through a convolutional auto-encoder

Hao Cui[1], *Kesheng Wang*[1,*], *Yu Li*[1], *Binyuan Yang*[2], *Qiang miao*[2]

[1]Equipment Reliability, Prognostics and Health Management lab(ERPHM), School of Mechatronics Engineering, University of Electronic Science and Technology of China,611731 Chengdu, China

[2]Chengdu Yute Technology Co., Ltd., 610000 Chengdu, China

Abstract. The amount of data is of crucial to the accuracy of fault classification through machine learning techniques. In wind energy harvest industry, due to the shortage of faulty data obtained in real practice, together with ever changing operational conditions, fault detection and evaluation of wind turbine blade problems become intractable through conventional machine learning methods. In this paper, a modified unsupervised learning method, namely a convolutional auto-encoder based data enlargement strategy (ABE) is proposed for wind turbine blade fault classification. Limited simulation results for different levels of wind turbine icy blades are used for investigation. First, convolutional auto encoder is used to increase the amount of the data. Then, decision tree based xgboost tool, as an example, is used to demonstrate the effectiveness of data enlargement strategy for fault classification. The study shows that the proposed data enlargement strategy is an effective method to improve the fault classification accuracy through machine learning techniques.

1 Introduction

When the wind turbine operates at temperatures below zero degree Celsius, ice coating may occur, especially when it encounters humid air or rain [1]. Ice coating on wind turbine blades can be a catastrophic problem. It may lead to serious life safety accidents, shut down of the wind turbine and inevitably a decrease in wind farm power generation. Unfortunately, in real practice, to obtain the real icy blade data is problematic due to the ever-changing operational conditions and the difficulty to assess the severity of ice coating on the blade. Therefore, to make full use of data information for icy blade detection and evaluation become a critical issue.

There are several ways to monitor and identify the existence of ice coating on the wind turbine blade. W. Olsen conducted an ice coating experiment at the ice wind tunnel in NASA [2]. Their works detailed the growth process of ice coating and the thermodynamic model of ice coating is then improved. M.C. Homola et al. simulated the power loss of a 5MW wind turbine after icing coating [3]. The calculation results show that the drag coefficient of the blade becomes larger after the ice coating, and the lift coefficient decreases, resulting in a 27% reduction in the output power of the wind turbine. Huang proposed a wind turbine blade icing detection system [4]. The relationship between the output voltage and the ice thickness can be obtained from the proposed system. When the surface thickness of the sensor exceeds the threshold, it will issue a warning signal. However,the method is complicated,time consuming and costly.

Auto-encoder is a deep-learning method, used as a tool for extracting features from data. Autoencoder has many forms of variants, such as convolutional auto-encoders (CAE), denoising autoencoders (DAE), sparse autoencoders(SAE), etc. Several studies have used different auto-encoders for fault diagnosis in resent years. Siqin Tao et al. combined stacked auto-encoders and softmax regression methods for bearing faults diagnosis [5]. The method shows exceptional ability to exclude influences from noise. Shao and Jiang use the maximum correntropy as the loss function in a deep auto-encoder and the fish swarm algorithm is further used to optimize the parameters of the auto-encoder so that an improved structure of autoencoder is established for the fault diagnosis of the gearbox [6].

As far as literature survey, it is found that most of researches for fault diagnosis through auto-encoders is to take advantages of auto-encoders in unsupervised ability of data feature extraction, but few study has explored another unique feature of auto-encoders, namely data dimension enhancement. In the present work, we propose a data enlargement strategy for fault classification through a convolutional auto-encoder for wind turbine blade fault diagnosis. In the method, the convolutional auto-encoder is first applied to enhance the original data dimension in which an optimal selection of data enlargement strategy is proposed, and then Xgboost, as a typical tool for data classification, is used here for both data importance ranking and data classification.

The remainder of this paper is organized as follows: The main technical theory is introduced in section 2. After that,a case study of the proposed method applying in the

* Corresponding author: keshengwang@uestc.edu.cn

simulation data from Bladed software is presented in section 3. Conclusions together with possible future works are presented in Section 4.

2 Related theory

2.1Data enlargement through CAE

Auto-encoders are popular unsupervised feature extraction deep-learning techniques which is suitable for dealing with highly nonlinear data [7]. Figure 1 shows an illustrative structure of a typical auto-encoder. A simple auto-encoder is, in fact, a feed-forward neural network with multiple hidden layers in which an encoder and a decoder is embedded in sequence. The role of the encoder is to represent the feature of the original data in terms of compressed data or code. The role of the decoder is to regenerate the compressed data to a reconstructed data with the same size of the input original data. An auto-encoder can be trained through minimizing the reconstruction errors between the orignal data and the reconstructed data. Specifically, for convolutional auto-encoder(CAE) which will be used in the present study, is to use convolutional layers at hidden layers to compress the data, as is shown the CAE part in Figure 3. The output of the k-th layer of the convolutional layer can be expressed as :

$$h_k = \sigma\left(x * W_k + b_k\right) \tag{1}$$

Where h_k is k-th layer output, σ is the activation function, W_k is k-th layer weight and b_k is the bias of the k-th layer. For highly non-linear data, it is reported that CAE can extract better features than simple auto-encoders.

Notice that the hidden layer or code shown in the Figure 1 has the capability to decrease or increase dimensions of the input data and capture the intrinsic feature of the data. Thus, in this study, the hidden or code layer is adopted as a data enlargement tool and further used as input for training the following machine learning technique.

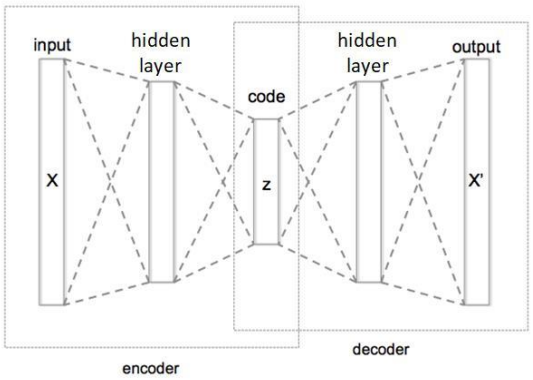

Fig.1. The structure of auto-encoders.

2.2Data for study from Bladed software

Bladed, a professional software for wind turbine design, is used to simulate the data from the wind turbine. The structural wind turbine parameters are not prescribed in here due to the confidentiality.The data generated by the simulation has five columns including wind speed, nominal pitch angle, measured power, and the two columns of the signal measured by the acceleometers in two perpendicular direction of the blade. The total amount of data generated is 144,000 points.

Figure 2 shows the scores of five columns of data obtained by Bladed importance calculated by the Xgboost built-in algorithm. Importance is calculated for a single decision tree by the amount that each attribute split point improves the performance measure, weighted by the number of observations the node is responsible for [8]. The performance measure may be the purity (Gini index) used to select the split points or another more specific error function. The result of importance score is depicted in Figure 2, the highest importance score of f3 means that the third column feature has the biggest discrimination ability for fault classification. Therefore, f3 is used for training the claassifer of Xgboost.

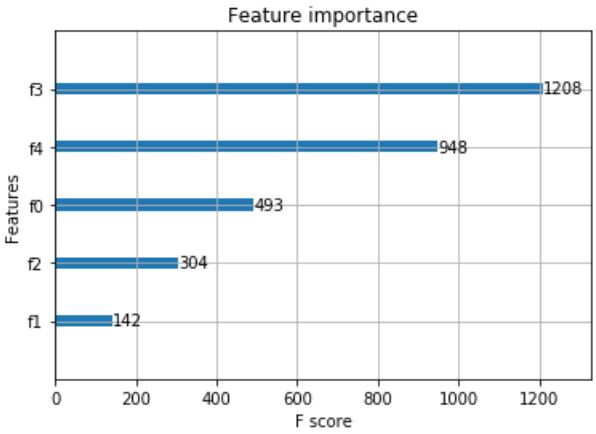

Fig.2. The importance of the original five-column features.

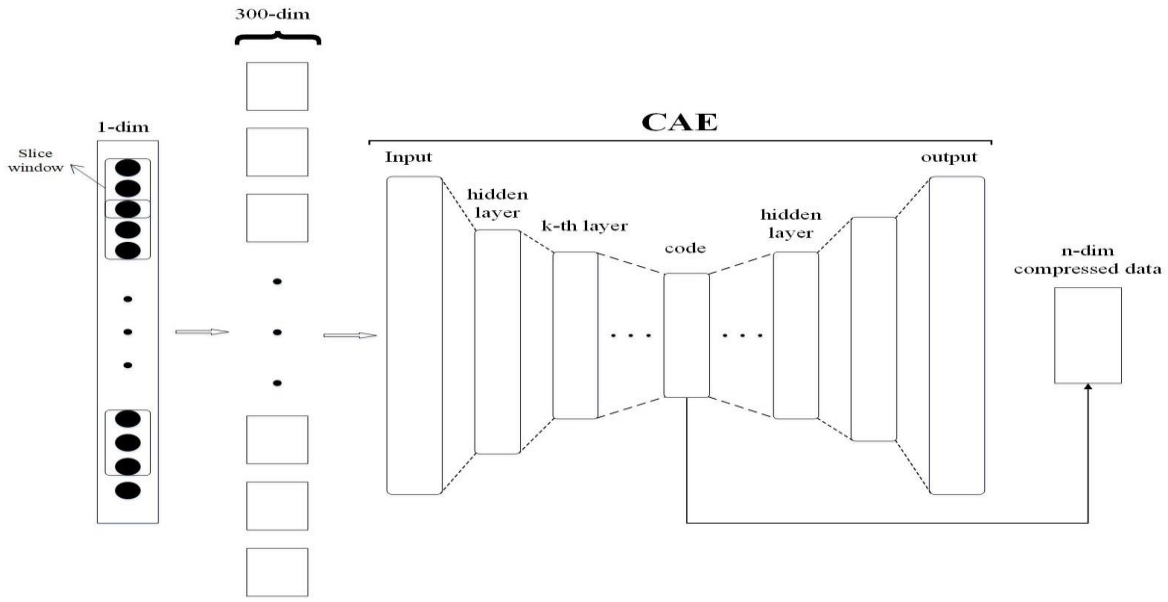

Fig.3. The process of getting compressed data.

However, if only one feature is being used for classification under the condition of limited data amount, it may be difficult to distinguish different kinds of ice coating states. Therefore, data slicing is applied here to increase number of data samples, but in each sample of the data will, in any case, contain less useful information compared with whole data. Once the training data are obtained, the CAE can be implemeted and convolutional auto-encoder is performed for feature extraction. The dimensional formula that generates the data based on the convolutional layer calculation:

$$x' = \frac{x - 2p - f}{s} + 1 \tag{2}$$

Where x' is the dimension of the input data, x is the dimension of the compressed data, p is the number of padding of the convolutional layer, f is the filter size and s is the strides of filter sliding. Figure 3 shows the process of obtaining the compressed data.

3 Results

Figure 4 shows the classification accuracy by using xgboost classifier. The training data is the third column feature. When the number of decision trees is 23, the classification accuracy is up to 71% in Figure 4. Figure 5 is the highest classification accuracy (90%) with the generated data dimension of 20. It is found that with the variation of the generation data dimension, the classification accuracies are different. Based upon the data structure, six different dimensions choices are applied and six different classification accuracy are obtained. Further, for the compressed data from convolution layer, we applied Xgboost to calculate the values of importance to each column. Then we calculate the standard deviation of the importance values to indicate the difference of importance values. An example of the importance values for the compressed data with dimension of 20 is shown in Figure 6. The relationship between accuracy and standard deviation under different compressed data dimensions are listed in Table 1.

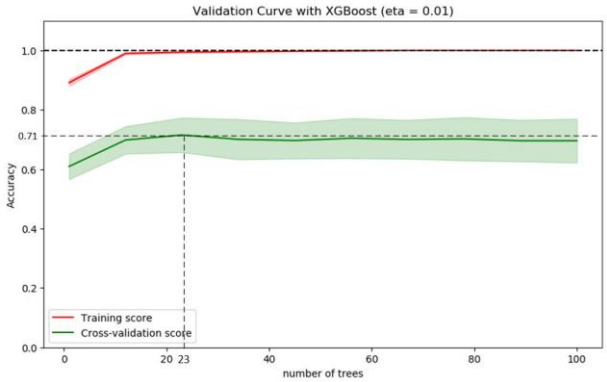

Fig.4. Classification accuracy of raw data.

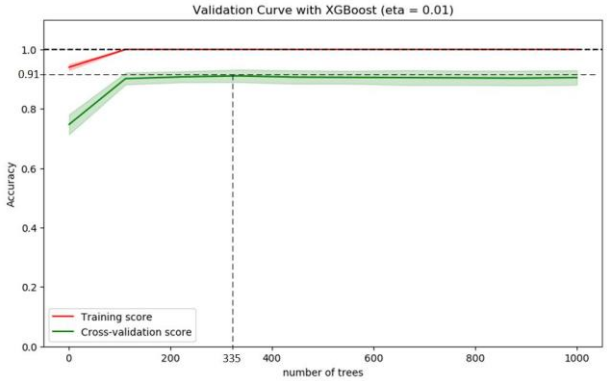

Fig.5. When the data dimension is 20, the accuracy of the five classifications is performed.

By comparing Figure 2 and Figure 6, it can be seen that the distribution of the importance valuess is significantly different.

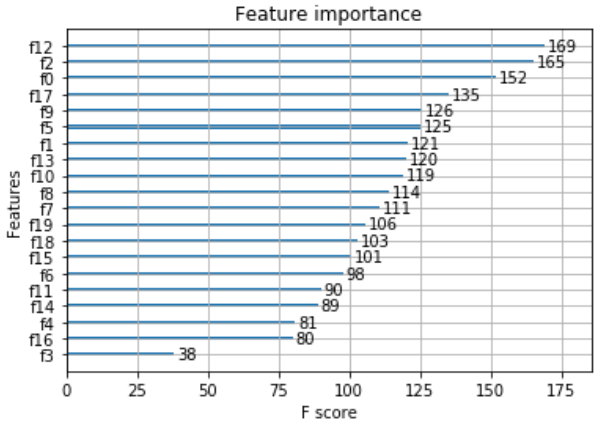

Fig.6. Feature score of 20-dimensional data.

By plotting table 1 in Figure 7, a trend can be seen from the figure that, the smallest standard deviation gives , the highest the classification accuracy. Therefore, a data enlargement strategy to improve fault classification can be proposed and the logic is shown in Figure 8.

Table 1. Relationship between predictions and stand deviation.

Dimentions	Accuarcy	Standard deviation
3-dim	41%	171.35
5-dim	71%	64.52
10-dim	73%	76.07
15-dim	77%	47.24
20-dim	91%	30.49
30-dim	75%	42.33

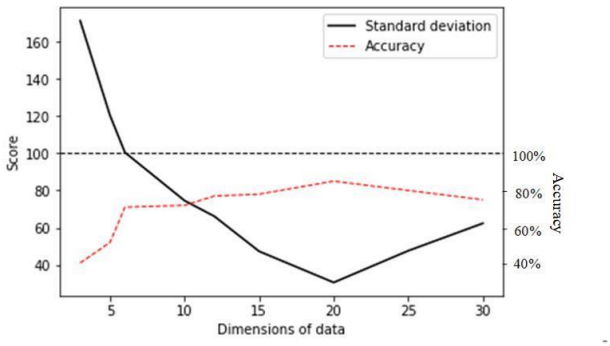

Fig.7. Relationship between classification accuracy and standard deviation in different dimensions.

The application steps of the proposed strategy is explained as follow:

Step1.Slice raw data to get multiple columns of data.

Step2.Obtain intermediate layer (code) output by convolutional auto-encoder.

Step3.Calculate the importance scores of different features extracted by the convolutional auto-encoder by the importance algorithm and calculate the standard deviation of each importance score.

Step4.Find the data with the smallest standard deviation of importance scores.

Step5. Input the selected data from step4 into Xgbosst for classification.

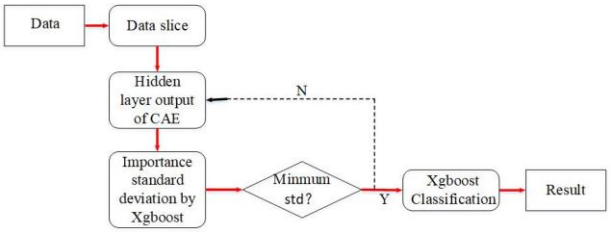

Fig.8. Data increasement via CAE.

4 Conclusions

In this paper, we propose a data enlargement strategy to improve the fault classification accuracy under limited data samples. The proposed method features advantageous for improving the accuracy of fault classification.

Further theoretical studies of the proposed method as well as more data test should be researched. Feasibility investigation of the proposed method with the real data from wind turbine is on the way.

5 References

1. E Muljadi, CP Butterfield, Pitch-controlled variable-speed wind turbine generation, IEEE Trans Ind Appl, 37 (2001) 240-24.

2. W. Olsen, E. Walker, Experimental evidence for modifying the current physical model for ice accretion on aircraft surfaces, Third International Workshop on Atmospheric Icing of Structures, Vancouver, Canada, 1986.

3. M.C. Homola, M.S. Virk, P.J. Nicklasson, Performance losses due to ice accretion for a 5 MW wind turbine, Wind Energy, 15 (2012) 379-389.

4. CY Huang, YE Lin, Experimental Study of Wind Turbine Blade Icing Detection System, Instrument Technique& Sensor, 6 (2014) 86-92.

5. Siqin Tao, Tao Zhang, Jun Yang, Bearing fault diagnosis method based on stacked autoencoder and softmax regression, IEEE, Hangzhou, China, September 2015.

6. Shao Hai dong , Jiang Hongkai , Zhao Huiwei, A novel deep auto-encoder feature learning method for rotating machinery fault diagnosis, Mechanical System and Signal Processinbg, 95(2017) 187-204.

7. O. Chen, D.Simig, G. Weisz, Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction, ICANN 2011: Artificial Neural Networks and Machine Learning – ICANN, (2011) 52-59.

8. Jason Brownlee, Feature Importance and Feature Selection With XGBoost in Python. 23 September. 2018 <https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python>.