Homework2
Shaft Health Assessment

Bing-Hong Hong
Department of Electrical Engineering
National Taipei University of Technology
Taipei, Taiwan
t106318072@ntut.edu.tw

Hsin-Ting Chou
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
justin0010523@gmail.com

Shao-Tse Chien
Department of Computer Science
National Taipei University of Technology
Taipei, Taiwan
t105820049@ntut.org.tw

Chih-Hao Chiang
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
sky41282007@gmail.com

***Abstract*—The data of shaft was transformed from time domain to frequency domain by Fourier transform. Then the Fisher criterion was utilized to find the important features which are 21.8 Hz and 84.6 Hz. Using these two features, we train the model with logistic regression. Finally, the shaft conditions were able to separate into three different conditions, which are health, unbalanced level 1 and level 2 shaft by logistic regression. Through setting up two different thresholds, we could separate different conditions of the shaft perfectly.**

Index Terms—unbalance shaft, Fisher criterion, Cross-validation, logistic regression, confusion matrix, ROC curve

INTRODUCTION

An unbalanced shaft was formed by distortion from stress, thermal distortion or deposits and oil buildup. Stress was added on the shaft and gradually making the shaft to be unbalanced with the ongoing manufacturing process. Thermal distortion happened if the shaft was operating in high temperatures, metal or other materials could become more easily to extend, causing shaft distortion. And deposits and oil buildup happened if the machine was involved in material handling. It would be possible that minerals, dust or dirt began to build up on the shaft, making distortion to occur. This could also happen if the shaft was exposed to oil. Oil gradually accumulated into parts, making the shaft to be unbalanced.

Due to the unbalanced force on the shaft, the shaft vibrated, causing the parts of the machine to lose and even damage other parts. The vibration also caused the machine to generate extra noise. Also with the vibration, additional force on rotation would pressure on the bearings holding up the shaft, making the life of a bearing to be shortened. Work conditions also became unsafe. With the extra vibration, parts of the machine were more easily to break. And with the increased possibilities of the machine to breakdown, maintenance time would also increase, causing loss of money and time [1].

This report is organized as follows. Section 2 mentioned the methodology and materials of this report. Section 3 demonstrated the performance by real case of unbalanced shaft and healthy shaft. Section 4 summarized the conclusions.

METHODOLOGY AND MATERIALS

2.1. Feature normalization

Feature normalization is a method which can scales independent variables or features of dataset into same range without changing differences in the ranges of features. Most of the times, features in dataset are highly varying in magnitudes and that is the reason why we need to do Feature Normalization. There are several methods to do Feature Normalization, such as Rescaling (Min-Max Normalization), Mean Normalization, Standardization (Z-score Normalization), Scaling to unit length, etc. In particular, we applied standardization to normalize our dataset, which was rescaled and had the attributes of normal distribution with μ =0 and σ=1 after doing standardization[2][3].

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

2.2. Fisher criterion

Fisher criterion is used to measure how suitable a single variable is for separating two classes. The formula of fisher criterion is shown in (2).

$$F = \frac{|\mu A - \mu B|^2}{\sigma A^2 + \sigma B^2} \tag{2}$$

The numerator of fisher criterion is the difference between the mean of two classes. Moreover, the denominator of fisher criterion is the sum of the two classes' variance. Therefore, the higher the fisher criterion is, the better the feature is. The diagram of fisher criterion is shown as follow.

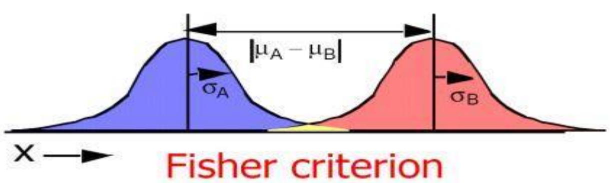


Fig 1. The diagram of Fisher criterion

2.3. Cross-validation

In order to train a machine learning model, the dataset is often separated to three subsets which are training set, validation set, and testing set. The training set and testing set are used to training and testing a machine learning model respectively. Moreover, the validation set is used to avoiding the overfitting during the training process. Therefore, we applied cross-validation to determine what the validation set is. Cross-validation divided the training set into some partitions; one of the partitions is regarded as the validation set, and the other partitions are regarded as the training set. Through selecting another partitions as validation set and choosing the other partitions as training set in each time, the final accuracy of training can be calculated by averaging the accuracy of each time. The diagram of cross validation is shown in Fig 2.

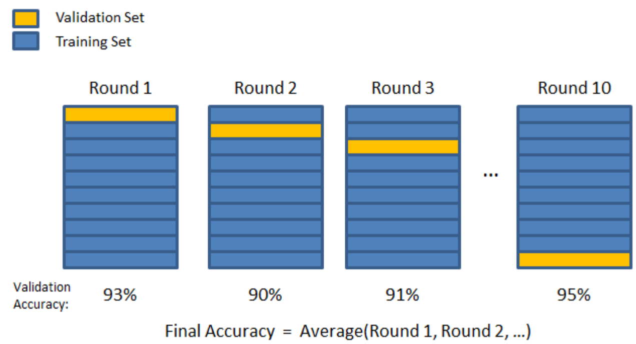


Fig 2. The illustration of cross-validation

2.4. Binary- and multi- class logistic regression

There are two types of methods in the classification field. One is binary classification. The other is multi-class classification. First of all, Binary classification can classify the elements of dataset into two groups [4]. There are lots of ways to do binary classification such as decision trees, random forests, support vector machines (SVM), and logistic regression. In contrast, the goal of Multi-class Classification is to classify the features of dataset into many different classes. The difference between Binary Classification and Multi-class Classification is that previous one only has two classes, but the latter one has more than two classes [5]. For example, binary classification just like a yes-no question, whereas multi-class classification resemble multiple choice question.

Logistic regression is a kind of binary classification while its result is a probability which is between 0 and 1. With the certain threshold, the data can be separate to two classes [6]. The logistic function is defined as:

$$\text{logistic}(\eta) = \frac{1}{1+exp(-\eta)} \tag{3}$$

The steps of logistic regression are similar to linear regression. In the linear regression model, the relationship between outcome and features is modeled as:

$$\eta = \hat{y}^{(i)} = \beta_0 + \beta_1 {x_1}^{(i)} + \cdots + \beta_p {x_p}^{(i)} \quad (4)$$

For logistic regression, the results are probabilities between 0 and 1, so we combine two functions together:

$$\mathrm{P}\left(y^{(i)} = 1\right) = \frac{1}{1+exp\left(-\left(\beta_0 + \beta_1 {x_1}^{(i)} + \cdots + \beta_p {x_p}^{(i)}\right)\right)} \quad (5)$$

2.5. Confusion matrix

Confusion matrix is a table we use to describe the performance of a classification model on test data which the true label has already known. In confusion matrix, there are four conditions which are true positive, true negative, false positive, and false negative [7]. Moreover, through these four conditions, we can calculate the further result, the true negative rate and the true positive rate, which can be used to calculate the ROC curve. The diagram of confusion matrix is shown in Fig 3.

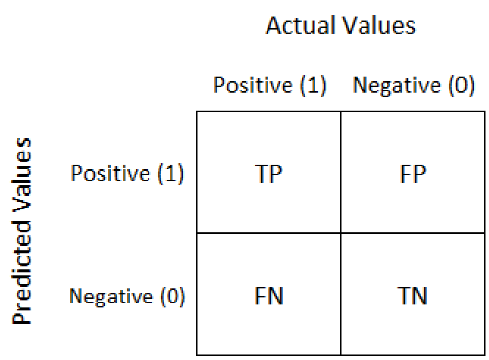


Fig 3. The illustration of confusion matrix

2.6. ROC curve

There are many machine learning algorithms in the world. How we compare each algorithm is an important issue. There is a common method to determine which algorithm is better called Receiver operating characteristic (ROC) curve, which sets different thresholds to draw the relation between the false positive rate and true positive rate. It should be noted that the higher the false positive rate we get, the higher the true positive rate we can reach. For example, if we have a Neyman-Pearson detector, which is Gaussian hypothesis shown in Fig 4, we can obtain the relation between the false positive rate and the true positive rate based on the certain threshold. Furthermore, there are two strategies to evaluate which algorithm is better in the ROC curve. The first strategy is that we can distinguish which the true positive rate of each algorithm is the highest based on the certain false positive rate, which means that algorithm is better than the others algorithm. The second strategy is more common used than the first strategy, which calculate the area under the curve (AUC) of ROC. Through comparing each algorithm of AUC, the best algorithm which has the biggest AUC can be distinguished [8]. The illustration of ROC curve is shown in Fig 5.

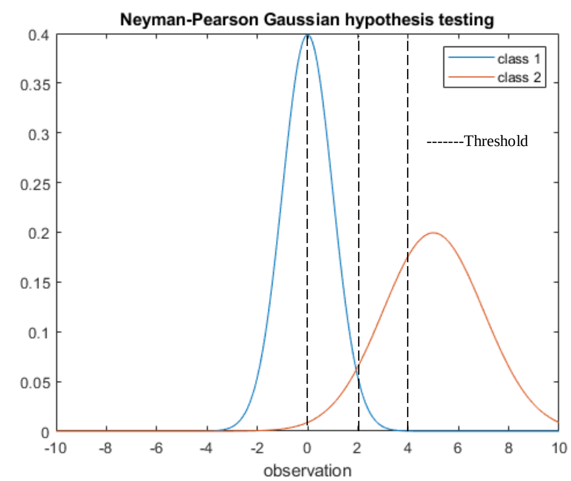


Fig 4. Neyman-Pearson Gaussian hypothesis testing

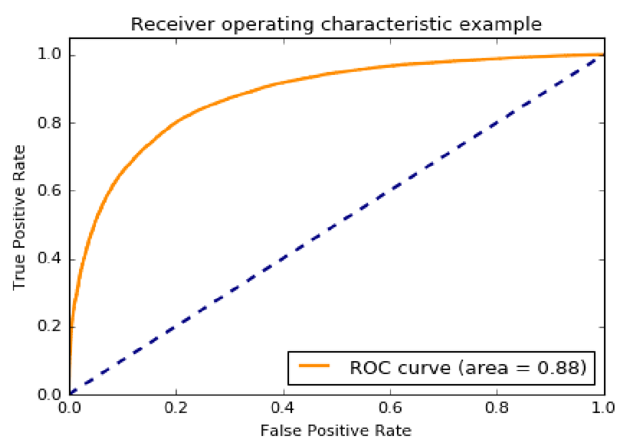


Fig 5. ROC curve

EXPERIMENTAL RESULTS

In the experimental results, we applied standardization to normalize our date in the preprocessing part. Then the Fourier transform was utilized to convert the data from time domain to frequency domain. After we convert the data to frequency domain, the feature extraction technique, Fisher criterion, was applied to search which feature is important. Finally, we applied logistic regression and cross-validation to train our machine learning model, and we use confusion matrix and ROC curve to evaluate how better the algorithm is. The flow chart of experiment is shown in the Fig 6.

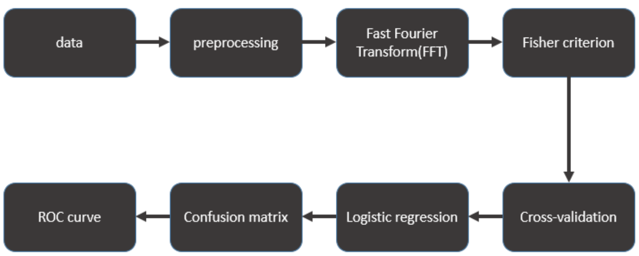


Fig. 6 The flow chart of experiment

3.1. Unbalance shaft and balance shaft data

In the data we obtained, the shaft is rotating at 20 Hz. The sampling rate of the data is at 2560 Hz. We have two sets of data, training and testing data. Training data consists of two parts, faulty and healthy. Each having 20 observations of vibration data. Testing data consists of 30 observations of healthy, level 1 and level 2 unbalanced condition with no labels. Each observation of the shaft vibration contains 38400 records.

3.2. Data preprocessing

In data preprocessing, we applied the standardization to normalize the signal, which mean is zero and standard deviation is one. Then data was transformed from the time domain to the frequency domain by Fourier transform. The result of Fourier transform is shown in Fig 7. It should be noted that we can observe that there has some significant different between the healthy shaft and the unbalanced shaft in Fig 7.

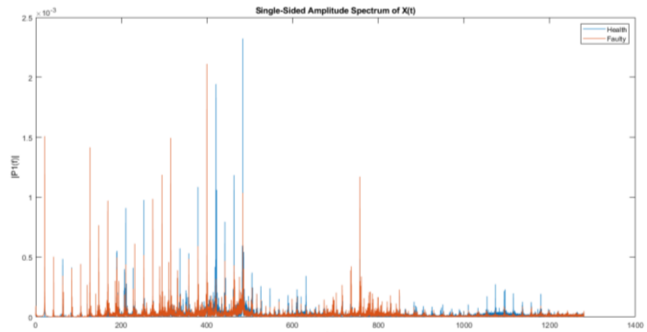


Fig 7. The result of Fourier transform

3.3. Feature extraction

In feature extraction, the fisher criterion was applied to find which frequency can distinguish between healthy shaft and unbalanced shaft better. The fisher criterion of each feature is shown in Fig 8. In our case, the frequency 21.8 Hz and 84.6 Hz were selected by the fisher criterion, which has the higher score than the other features.

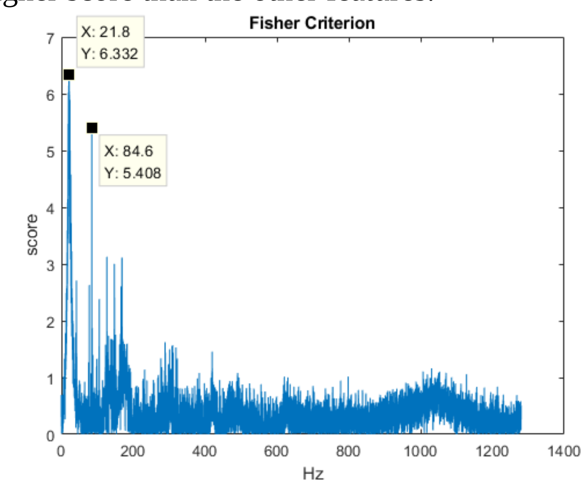
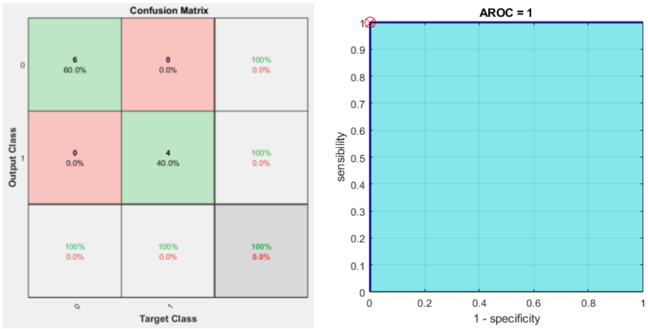


Fig 8. The result of Fisher criterion

3.4. Classification result

The training set was separated to four partitions by the cross-validation so that there have 10 data in the validation set and 30 data in the training set. The logistic regression algorithm was utilized to determine whether the shaft is health or not. Moreover, the confusion matrix and ROC curve were applied to evaluate the performance of logistic regression algorithm. The result of confusion matrix and ROC curve were showed once in Fig 9(a) and Fig 9(b) respectively since those were the same in cross-validation.



(a) (b)

Fig 9. The result of (a) confusion matrix (b) ROC curve

Finally, we utilized this logistic regression model to test the testing set. Although the testing set did not have the label, we can observe that there has three different level which might have been the healthy, the unbalanced level 1, and the unbalanced 2 shaft in the Fig 10.

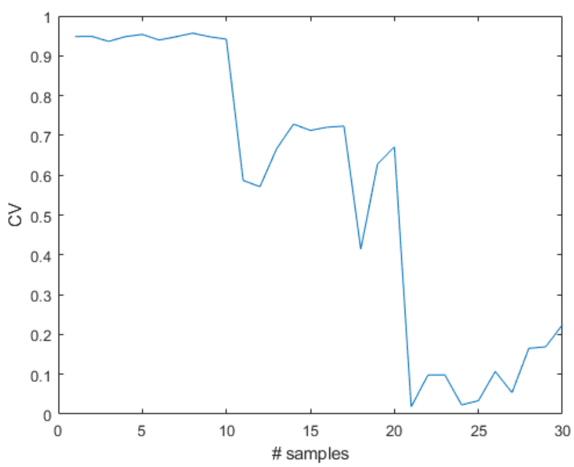


Fig 10. The testing result of logistic regression

CONCLUSION

In conclusion, after we finished the preprocessing of the data, we applied Fourier transform to convert the data from time domain to frequency domain. Then the two frequencies of the data, 21.8 Hz and 84.6 Hz, were selected by the scores of Fischer criterion. Using these two features, we train the model with logistic regression and are able to separate different shaft conditions (healthy, level 1 and level 2) with a very significant difference. Finally, through setting up two different thresholds, we could separate different conditions of the shaft perfectly.

REFERENCES

[1] https://www.testdevices.com/three-major-ways-manufactured-rotating-component-can-become-unbalanced/

[2] https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029

[3] https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e

[4] https://en.m.wikipedia.org/wiki/Binary_classification

[5] https://en.m.wikipedia.org/wiki/Multiclass_classification

[6] https://en.m.wikipedia.org/wiki/Logistic_regression

[7] Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006

[8] Poor, H. Vincent. An introduction to signal detection and estimation. Springer Science & Business Media, 2013.