Journal of Applied Mathematics and Physics, 2021, 9, 2927-2935
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Bearing Defect Detection Based on Acoustic Feature Extraction and Statistical Learning

Di Wu, Jinlong Sun, Zhifeng Liu, Ziqian Zhang, Min Huang, Zhongzhe Xiao*

School of Optoelectronic Science and Engineering, Soochow University, Suzhou, China
Email: *dwu42@stu.suda.edu.cn

How to cite this paper: Wu, D., Sun, J.L., Liu, Z.F., Zhang, Z.Q., Huang, M. and Xiao, Z.Z. (2021) Bearing Defect Detection Based on Acoustic Feature Extraction and Statistical Learning. *Journal of Applied Mathematics and Physics*, **9**, 2927-2935.
https://doi.org/10.4236/jamp.2021.911186

Received: October 14, 2021
Accepted: November 20, 2021
Published: November 23, 2021

Abstract

Bearings are widely utilized as key components in industrial scenarios. Therefore, the automatic and precise inspection of bearing defects is imperative for the manufacturing of the bearing. In this paper, a novel defect detection method based on acoustics is proposed to further improve both the accuracy and the efficiency of the defection process. We firstly constructed a labeled dataset composed of acoustic signals sampling from different bearings with a certain rotational speed. OpenSMILE is adopted to extract the acoustic features and the target acoustic feature dataset with 6373 features is formed. To further improve the efficiency of the proposed method, a feature selection strategy based on the chi-square test is adopted to eliminate the most inefficient features. Several statistical learning models are constructed and trained as the classifier. Eventually, the performance of classifiers is evaluated and achieves relatively high accuracy and efficiency with an extremely imbalanced dataset.

Keywords

Bearing Defect Detection, Feature Extraction, Machine Learning

1. Introduction

Bearings are important components widely used in production, especially in precision instrument production like optical, mechanical and electrical systems. Thus, the requirements of manufacturing precision of bearings are relatively high to ensure the performance and reliability of integrated systems. Under this circumstance, the classification of bearing defects is extremely meaningful in both industrial and economic aspects.

According to industrial experience and collected statistics, the common bearing conditions in production can be roughly categorized into three groups:

Qualified, Inner Ring Defect (IRD) and Ultra Precision Defect(UPD). Samples with Inner Ring Defect and Ultra Precision Defect should be recalled and destructed. The purpose of our work is to propose an algorithm that can precisely classify the samples and reduce defective rate. In this scenario, capability of recalling the defective products is the main descriptor of model performance.

In former research, the methods of detection of bearing defects were usually developed in perspective of time-domain signal, deep learning [1] and computer vision [2]. The deep learning model with high complexity may lead to massive amount of parameter, which will reduce the efficiency of training and classification. In our model, instead of using a complicated convolutional or recursive deep neural network, we constructed classifiers based on traditional machine learning methods. The model was trained with a dataset composed of filtered acoustic features. The acoustic feature dataset and statistical learning models brought many unique characteristics to our model.

We used openSMILE to extract acoustic features of original time domain signal, and built a new dataset based on the extracted features. With a certain vibration sequence, openSMILE can output the corresponding feature vector of the sample. Based on the extracted feature vectors, we constructed a new dataset. The acoustic feature dataset allowed to avoid using deep learning methods, such as one-dimensional Convolutional Neural Network to categorize.

Then, we performed feature selection to the acoustic feature dataset to further improve the performance of our model. Due to the restrained size of our dataset and the mathematical characteristics of most machine learning algorithms, an extremely high dimensional dataset will result in overfitting and unacceptable computational cost. To improve the accuracy and efficiency of the classifier, we use MATLAB and the built-in statistical algorithms to select features.

Finally, we applied statistical learning methods as the classifiers. In the industrial scenes, the prevalent deep learning models are unacceptably time-consuming in both training and predicting stages. To construct a model with balanced performance, we chose some of the statistical learning methods to improve the overall efficiency of the model. We also evaluated and compared the performance of the different classifiers.

The article is organized as follows: In Section 2 we introduced the basic structure and distribution of the original dataset. Section 3 presented the construction of ComParE 2016 Acoustic Feature Set and the procedure of feature extraction by openSMILE toolkit. In Section 4 we described feature selection with chi-square test. In Section 5, we trained the statistical learning models, listed related statistics, compared and analyzed their performances. Section 6 is the conclusion of the paper.

2. Dataset Construction

According to the experience of industrial production, defective bearings make a different sound from normal bearings when rotating. In our work, we built the bearing defect detection model based on sound signals.

The original dataset is composed of sound signals of different bearings rotating at a certain rotational speed of 1800 rpm. The sound signals are presented as numerical sequences; the value of elements in the sequences reflected the amplitude of sound signal when rotating.

The samples of sound signals were manually labelled and categorized into 3 groups: Qualified, Inner Ring Defection (IRD), and Ultra Precision Defection (UPD). The total number of the samples in the dataset is 708, which indicated that the dataset is relatively small scaled. In the original labeled dataset, the distribution of different samples is shown in **Table 1**. According to **Table 1**, the dataset is relatively imbalanced, this may lead to models low capacity of generalization, hence accuracy will not be the only evaluation standard of overall performance.

Table 1. Dataset distribution.

Defect Category	Number of Samples	Percentage
Qualified	627	88%
IRD	78	11%
UPD	9	1%

3. Feature Extraction

Traditionally, the classification problems about time domain signals are usually solved by deep learning methods. By using one-dimensional Convolutional Neural Networks (CNN), the model can automatically extract the features [3]. Recurrent Neural Networks (RNN) and their derivatives are another widely used category of Artificial Neural Networks in time-related problems. RNN can build the relationship between current output and previous hidden state, and one of the derivatives, Long Short-Term Memory model introduced gating system to solve the problem of gradient exploding and vanishing [4]. However, despite the outperforming results in accuracy of the deep learning models, their time cost of training and predicting is unacceptable in industrial bearing production.

To apply our model into the scene of industrial production, the balance of accuracy and efficiency is required. Thus, we considered using statistical learning methods. However, traditional statistical learning methods usually performs badly on sequential problems, so we performed feature extraction by using openSMILE to extract acoustic features of the time domain signals.

3.1. Feature Set

Effective feature extraction can output crucial acoustic features for bearing defect detection, which can directly improve the performance of prediction system. However, there are few studies about feature extraction of bearing defect detection. In our work, we adopted experience on emotion recognition to our model. In emotion recognition and natural language processing tasks, to process more complex sound signals, a more detailed feature set is need. We utilized the ComParE

2016 Acoustic Feature Set in feature extraction. The ComParE feature set contains 6373 features resulting from the computation of various functionals over low-level descriptor (LLD) contours [5]. The feature set has been proved effective in emotion recognition [6], thus it is very likely to be capable of defect detection.

3.2. Extraction with OpenSMILE

OpenSMILE is an open-source audio feature extractor. It can process the original vibration sequences and directly output the corresponding numerical values and name of the acoustic features [7]. With different configuration files, openSMILE can extract different features. The overall steps of feature extraction are described in **Figure 1**.

In our architecture, we used one of the default configuration files provided by the developers of openSMILE. We extracted 6373 acoustic features in total.The features can be divided into 65 categories by corresponding LLDs [8]. The LLDs are shown in **Table 2**.

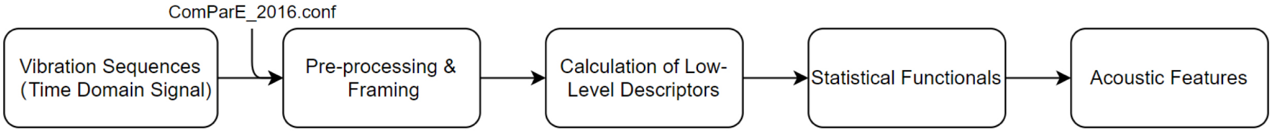

Figure 1. Process of feature extraction.

Table 2. 65 low-level descriptors (LLDs).

LLDs
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS energy ZCR (Zero-Crossing Rate)
RASTA-style auditory spectrum, bands 1 - 26 (0 - 8 kHz)
MFCC 1 - 14
Spectral energy 250 - 650 Hz, 1 kHz - 4 kHz
Spectral roll off point 0.25, 0.50, 0.75, 0.90
Spectral flux, centroid, entropy, slope
Psychoacoustic sharpness, harmonicity
Spectral variance, skewness, kurtosis
F_0
Prob. of voice
Log. HNR, Jitter (local, delta), Shimmer (local)

4. Feature Selection

Due to the unacceptable computational cost of deep learning model, we chose several classical machine learning models as the classifier. According to the mathematical theory, for most of the machine learning models, an extremely large number of dimensions will lead to overfitting, and result in perfect performance on training set and bad performance on testing set. An overfitted model does not have ideal performance in application for its lack of generalization capability, so the model cannot be used in industrial production under this circumstance. Besides, high dimensional dataset will greatly increase the computational cost of our model, which will have negative effect on the efficiency of our model.

To avoid overfitting, we performed feature selection to decrease the number of features. By using chi-square tests, we computed the corresponding p-values of features. According to basic knowledge of statistics, a small p-value indicates that the corresponding feature is dependent on label, thus, it is an important feature.

We performed feature selection with MATLAB and its built-in function fscchi2. The fscchi2 function automatically compute the p-values, then output the vector of predictor scores.The predictor score is $-\log(p)$, therefore, a larger predictor score indicates the corresponding predictor is more label-dependent and contains more information. We selected 600 features with largest predictor scores and built the final dataset. The features with the largest predictor scores are shown in **Table 3**.

The most label-dependent attributes are mainly related to spectral LLDs, such as spectral variance and spectral slope. This is reasonable for the sound signals of bearing rotation are generally periodic and simple signals when compared to human emotional signals.

Table 3. The most label-dependent features and corresponding predictor sores.

Feature	LLD	Predictor Score
pcm_fftMag_spectralVariance_sma_quartile1	Spectral Variance	86.78
pcm_fftMag_spectralSlope_sma_quartile3	Spectral Slope	80.58
pcm_fftMag_spectralVariance_sma_amean	Spectral Variance	79.13
pcm_fftMag_spectralSlope_sma_quartile2	Spectral Slope	78.07
pcm_fftMag_spectralSlope_sma_amean	Spectral Slope	76.44
pcm_fftMag_spectralVariance_sma_rqmean	Spectral Variance	76.20
pcm_fftMag_spectralVariance_sma_quartile2	Spectral Variance	75.49
pcm_fftMag_spectralSlope_sma_peakMeanAbs	Spectral Slope	74.88
logHNR_sma_de_minPos	Log.HNR	73.70
pcm_fftMag_spectralSlope_sma_quartile1	Spectral Slope	71.05

5. Classifiers and Performance Analysis

5.1. Classification Algorithms

The training of machine learning models is implemented on WEKA platform, which integrated various statistical learning methods [9]. We selected several prevalent statistical learning methods as classifiers: Simple Logistic Regression, C4.5 Decision Tree, Discrete AdaBoost and LogitBoost.

Simple Logistic Regression is a classical linear classification algorithm. It is widely used in various fields. It is easy to train and use, however, due to the relatively simple model structure, Simple Logistic Regression may have unsatisfying performance on complex nonlinear problems. Thus, we applied LogitBoost Algorithm. LogitBoost Algorithm is an application of boosting procedure. Boosting means combining the performance of several weak classifiers and produce a strong learnable classifier. The most famous boosting machine learning algorithm is AdaBoost. LogitBoost is a unique derivative of AdaBoost based on logistic regression and maximum log-likelihood cost function [10]. To compare the classification capabilities of different AdaBoost algorithms, we also implemented Discrete AdaBoost, also known as AdaBoost M1, it is another widely used boosting algorithm [11]. To investigate the effect of Decision Tree on bearing defect detection, the C4.5 Decision Tree Algorithm is also included to comparison [12].

5.2. Performance Analysis

Due to the limited scale of the dataset, we implemented the evaluation of models with 10-flod cross validation. Since the dataset is extremely imbalanced, and the final purpose is to detect the defective samples, to evaluate the overall performance, in addition to accuracy, we also made use of other information such as AUC (area under ROC curve), recall and confusion matrices.

The accuracies, AUC and recalls are presented in **Table 4**. The simplest classifier, logistic regression, performs well in both Qualified and IRD categories, with an overall accuracy of 89.22%. However, even though its recall of the UP category is 0.364, which is relatively good, the corresponding AUC is as low as 0.545. According to knowledge about machine learning, a low AUC value, which is close to 0.5, indicates that the classifier tends to classify the samples randomly, so the correct classification of UP samples is the result of random classification.

The boosting methods, LogitBoost and Discrete AdaBoost perform differently under the imbalanced circumstance. LogitBoost has a high accuracy of 91.46%, the highest value in selected classifiers, and performs well in IRD samples, with a recall of 0.592. LogitBoost also have a high AUC value in all three categories. By optimizing the distribution of dataset, the performance of LogitBoost may be improved. However, Discrete AdaBoost tends to categories all of the samples as Qualified samples. This is a common problem in imbalanced machine learning. C4.5 Decision Tree has an accuracy of 89.36%, and its recalls in IRD and UPD categories are close to Logistic Regression, while the AUC is relatively low. This may prove that the C4.5 Decision Tree is not reliable enough in defect detection.

Confusion matrices also provided direct visual information about the performance of different models. We presented the confusion matrices of selected models in **Figure 2**. The diagonal blocks are painted blue, a deeper blue color

Table 4. Performance statistics.

Classifier		Recall	AUC	Accuracy
Logistic Regression	Qualified	0.936	0.862	89.22%
	IRD	0.605	0.857	
	UPD	0.364	0.545	
LogitBoost	Qualified	0.968	0.932	91.46%
	IRD	0.592	0.916	
	UPD	0.091	0.919	
Discrete AdaBoost (AdaBoost M1)	Qualified	0.992	0.884	88.66%
	IRD	0.145	0.870	
	UPD	0.000	0.883	
Decision Tree (J48)	Qualified	0.941	0.699	89.36%
	IRD	0.579	0.701	
	UPD	0.364	0.653	

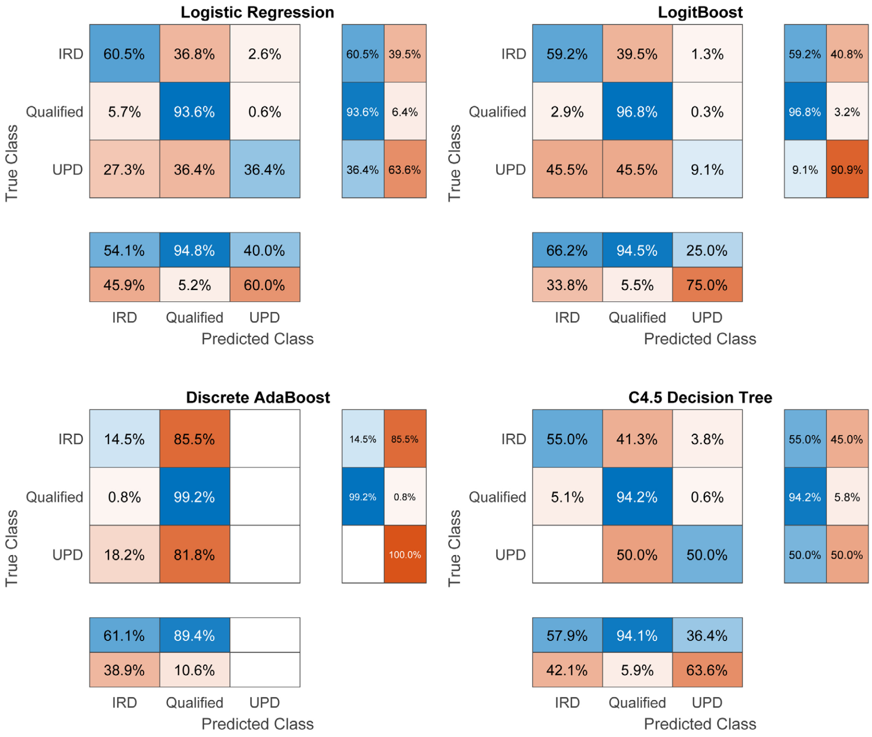

Figure 2. Confusion matrices.

represents better performance on the corresponding category. According to the figure, we can conclude that Logistic Regression and C4.5 Decision Tree have relatively satisfying performance. This may indicate that defect detection is not a extremely complex nonlinear question, and methods with simple structures are capable of classification. However, as we mentioned in previous analysis, Logistic Regression and C4.5 Decision Tree have low values of AUC, which proved that the classification procedures of the two algorithms are performed randomly. This may explain their good performances on rare categories in an extremely imbalanced scenario.

To compare the efficiency of machine learning model and evaluate the improvement brought by feature selection, we trained the model with original acoustic feature dataset and outputted relative statistics. The time cost data of model training with dataset before and after feature selection is shown in **Table 5**.

Table 5. Average time cost of model training.

Classifier	Average Time Cost (Seconds)	
	6373 Features	600 Features
Logistic Regression	10.35	0.68
LogitBoost	18.01	0.98
Discrete AdaBoost (AdaBoost M1)	3.41	0.58
C4.5 Decision Tree (J48)	4.04	0.25

According to the statistics, LogitBoost Algorithm have highest time cost while obtaining relatively ideal accuracy and recall. When traininig with 6373 features, the average time cost of 18.01 seconds is unacceptable in industrial application. High training time cost may result in difficulty in updating and maintaining the model, which will greatly reduce the efficiency. However, after feature selection, the time costs were greatly lowered, made the models more practical in production.

6. Conclusions

The classification of three categories of bearing conditions: Qualified, Inner Ring Defect, Ultra Precision Defect, was investigated with integrated statistical learning methods. We adopted ComParE 2016 Acoustic Feature Set in our model, which has been proved to be effective in various fields including emotion and language recognition. The ComParE 2016 Acoustic Feature Set configuration file was integrated into openSMILE, by using the file and openSMILE toolkit, we extracted 6373 acoustic features related to 65 LLDs. Then we performed feature selection with a chi-square test and selected 600 of most label-dependent attributes. The important attributes are mainly related to spectral LLDs. In the WEKA platform, we trained several statistical learning classifiers. The best accuracy of 91.46% was obtained by LogitBoost model. The model also performed relatively well in minor categories. In IRD category, the model acquired a recall of 0.592. The AUC of the model is also relatively high, with a weighted average value of 0.930. We also compared the time cost of models trained with 6373 and 600 features and concluded that feature selection greatly improved the efficiency.

Since the dataset is small-scaled and unevenly distributed, the overall performance of the model can still be improved. A larger and evenly distributed dataset will be constructed in near future. With a more reasonable dataset, the aim of future work will focus on improving performance in defective categories, compare and investigate the time cost of the defect detection model.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Kumar, A., Zhou, Y.Q., Gandhi, C.P., Kumar, R. and Xiang, J.W. (2020) Bearing Defect Size Assessment Using Wavelet Transform Based Deep Convolutional Neural Network (DCNN). *Alexandria Engineering Journal*, **59**, 999-1012. https://doi.org/10.1016/j.aej.2020.03.034

[2] Deng, S., Cai, W.W., Xu, Q.Y. and Liang, B. (2010) Defect Detection of Bearing Surfaces Based on Machine Vision Technique. 2010 *International Conference on Computer Application and System Modeling*, Taiyuan, 22-24 October 2010, V4-548-V4-554. https://doi.org/10.1109/ICCASM.2010.5620311

[3] Peng, D.D., Liu, Z.L., Wang, H., Qin, Y. and Jia, L.M. (2018) A Novel Deeper one-Dimensional CNN with Residual Learning for Fault Diagnosis of Wheelset Bearings in High-Speed Trains. *IEEE Access*, **7**, 10278-10293. https://doi.org/10.1109/ACCESS.2018.2888842

[4] Graves, A., Mohamed, A.-R. and Hinton, G. (2013) Speech Recognition with Deep Recurrent Neural Networks. 2013 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, 26-31 May 2013, 6645-6649. https://doi.org/10.1109/ICASSP.2013.6638947

[5] Schuller, B., Steidl, S., Batliner, A., *et al.* (2016) The Interspeech 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language. 17*th Annual Conference of the International Speech Communication Association*, 2001-2005. https://doi.org/10.21437/Interspeech.2016-129

[6] Deb, S., Dandapat, S. and Krajewski, J. (2017) Analysis and Classification of Cold Speech Using Variational Mode Decomposition. *IEEE Transactions on Affective Computing*, **11**, 296-307. https://doi.org/10.1109/TAFFC.2017.2761750

[7] Eyben, F., Wllmer, M. and Schuller, B. (2010) Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. *Proceedings of the* 18*th ACM International Conference on Multimedia*, 1459-1462. https://doi.org/10.1145/1873951.1874246

[8] Weninger, F., Eyben, F., Schuller, B., *et al.* (2013) On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common. *Frontiers in Psychology*, **4**, 292. https://doi.org/10.3389/fpsyg.2013.00292

[9] Hall, M., Frank, E., Holmes, G., *et al.* (2009) The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, **11**, 10-18. https://doi.org/10.1145/1656274.1656278

[10] Friedman, J., Hastie, T. and Tibshirani, R. (2000) Additive Logistic Regression: A Statistical View of Boosting (with Discussion and A Rejoinder by the Authors). *The Annals of Statistics*, **28**, 353-360. https://doi.org/10.1214/aos/1016218223

[11] Freund, Y., Schapire, R. and Abe, N. (1999) A Short Introduction to Boosting. *Journal-Japanese Society for Artificial Intelligence*, **14**, 771-780.

[12] Quinlan. J.R. (2014) C4. 5: Programs for Machine Learning. Elsevier.