Proceedings of 2011 8th Asian Control Conference (ASCC)
Kaohsiung, Taiwan, May 15-18, 2011

MoA4.5

Control of Mechatronics Systems

Ball Bearing Fault Diagnosis Using Machine Learning Techniques

Hsuan-Wen Peng
Dept. of Mechanical Engineering
National Chung Cheng University, Taiwan
hsw@hotmail.com.tw

Pei-Ju Chiang
Dept. of Mechanical Engineering
National Chung Cheng University, Taiwan
pchiang@ccu.edu.tw

Abstract—**Ball bearing fault is one of the main causes of induction motor failure. This paper investigates in the fault diagnosis of ball bearing of three phase induction motor using random forest algorithm and C4.5 decision tree. The bearing conditions are classified to four categories: normal, bearing with inner race fault, bearing with ball fault and bearing with outer race fault. The statistical features used for classification are extracted from mechanical vibration signal in time domain and frequency domain. Principal component analysis (PCA) and linear discriminent analysis (LDA) are used to reduce the dimension and complexity of the feature set. The classification accuracy of random forest algorithm and C4.5 decision tree are analyzed and compared. The experimental results show that the random forest algorithm not only works better than the C4.5 decision tree but also can classify the ball bearing condition effectively.**

Keywords-Ball bearing fault diagnosis; Principal component anslysis; Linear discriminent analysis; C4.5 decision tree; Random forest algorithm

I. INTRODUCTION

The fault diagnosis of machine component is an important technique for industry because the machine component failure will cause the interruption of the production line and increase the cost. To diagnose the faulty condition of the machine components such as induction motor, bearing, hydraulic pump, chain box, etc., vibration signal has been widely used [1-14]. For example, in [1], the root-mean-square, crest factor, variance, skewness, etc., of the vibration signal analyzed in time and frequency domain are calculated as the features to diagnosis the fault condition of induction motor and predict the life of machine component. In [2], power spectrum density of the vibration signal of hydraulic pump was analyzed. In [3], the Hilbert transform and envelope analysis of the vibration signal in time-domain was used in the fault diagnose of roller bearings. The bispectrum of vibration signal of rolling element for fault diagnosis is demonstrated in [4]. In [5], wavelet transform was used to analyze the vibration of rolling element bearings.

Various features are then extracted from the analyzed signal and used to build classifiers. To reduce the computation complexity, feature dimensionality reduction techniques such as principal component analysis can be used [6]. To diagnose the faulty conditions, machine learning techniques such as artificial neural network [1,7,10,11,12], support vector machine [12], decision tree [2,6,8,9], random forest [13], fuzzy logic technique [1,2,8,10,11], and hidden Markov model [14] can be applied.

This paper focused on the fault diagnosis of ball bearing of induction motor using C4.5 algorithm, which is one type of decision tree, and random forest algorithm. The measured vibration signal is analyzed in time domain and frequency domain. Different features are extracted, processed and applied to the classifiers. The effectiveness of classifier, C4.5 and random forest algorithms, are compared. Experimental results show that the random forest algorithm performs better than the C4.5 algorithm. Different features process will also result in different diagnosis performance.

The structure of this paper is organized as follows. Feature extraction methods are presented in the next section. C4.5 and random forest algorithm, two types of machine learning techniques, are briefly introduced in section III. Experimental results are shown in section IV. The last section concludes this paper.

II. FEATURE EXTRACTION

Figure 1 outlines the procedures to develop a classifier to achieve fault diagnosis of ball bearing of three phase induction motor. First, the signals created by the machine with normal and faulty conditions are measured. Features extracted from the obtained data are then used to train the classifier. The developed classifier is then used to classify the operational conditions of the machine.

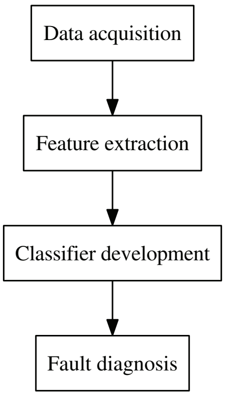

Figure 1. The general procedure for fault diagnosis.

However, features extracted from the data may correlate with each other, which may not only increase the computation complexity but also obstruct the classifier to diagnose the bearings effectively. To resolve this problem, the principal component analysis and linear discriminant analysis can be used to reduce the feature dimensions.

A. Principal Component Analysis (PCA).

In most of cases, the extracted features are not independent with each other. If some of features are closely related to each other, the PCA can be expressed these features as the linear combination of new features. That means that the principal component analysis can transform a number of correlated features into uncorrelated new features called principal components. The principal components are arranged in descending order according to their variance and can be done by singular value decomposition or covariance matrix. Neglecting the principal components with small variances will not lose significant information of the original data. The feature reduction can thus be done by projecting the original data onto the remaining principal components with high variances. For more details about PCA, please refer to [15].

B. Linear Discriminant Analysis

Similar to the principal component analysis, the goal of the linear discriminant analysis (LDA) is to perform dimensionality reduction. The difference between LDA and PCA is that the PCA is based on the sample covariance which characterizes the scatter of the entire data, and the LDA consider the scatter within-classes and the scatter between-classes to preserve the class discriminatory information as much as possible. The entire data are projected to the directions along which the classes are best separated. The detailed methodologies can be found in [15].

III. Mahine Learning

In this paper, we use C4.5 algorithm and random forest algorithm, which are widely used, to develop the classifiers.

A. C4.5 algorithm

The C4.5 algorithm proposed by Ross Quinlan [16] is one type of supervised machine learning algorithms used to generate a decision tree. The decision tree is built from a set of classified samples called training set. Fig. 2 shows an example of the decision tree. As shown in Fig. 2, a decision tree consists of numbers of nodes. Each node involves one feature used to classify the input data. These nodes are arranged in descending order based on the classification effectiveness. The top node, which is called “root”, is the node with best classification effectiveness; and the terminal nodes, which are called “leaf”, are the classification results. Each branch, which is a chain of nodes from root to leaf, represents one classification rule. The C4.5 algorithm uses the normalized information gain to select the most significant feature to classify the training set.

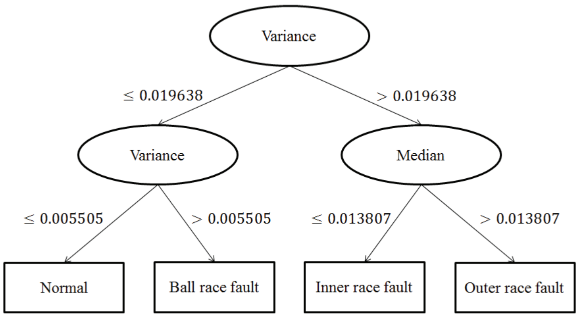

Figure 2. An example of C4.5 decision tree.

Assume that the training set T has K classes $\{C_1,C_2,\ldots,C_K\}$, the information entropy of node is calculated as follows:

$$\text{Info}(T) = -\sum_{i=1}^{K} \frac{\text{freq}(C_i,T)}{|T|} \log_2\left(\frac{\text{freq}(C_i,T)}{|T|}\right) \tag{1}$$

where $|T|$ is the number of cases in T and freq(C_i,T) is the number of cases C_i belonging to T. For the selected feature F_{test}, assume the training set is classified into N subsets $\{S_1,S_2,\ldots,S_N\}$, the expected information entropy can be calculated as:

$$\text{Info}_{\exp}(T,F_{test}) = \sum_{j=1}^{N} \frac{|S_j|}{|T|} \text{Info}(S_j) \tag{2}$$

where the $|S_j|$ is the number of cases S_j. The information gain of feature F_{test} can be calculated as

$$\text{Gain}(T,F_{test}) = \text{Info}(T) - \text{Info}_{\exp}(T,F_{test}) \tag{3}$$

And the split information is given by

$$\text{SplitInfo}(T,F_{test}) = -\sum_{j=1}^{N} \frac{|S_j|}{|T|} \log_2\left(\frac{|S_j|}{|T|}\right) \tag{4}$$

The information gain (or called gain ratio) then is normalized as

$$\text{GainRatio}(T,F_{test}) = \frac{\text{Gain}(T,F_{test})}{\text{SplitInfo}(T,F_{test})} \tag{5}$$

The C4.5 algorithm selects the feature with maximum GainRatio(T) as the node according to

$$F_{node} = \max\left(\text{GainRatio}(T,F_{test})\right) \tag{6}$$

B. Random Froest Algorithmn

The random forest algorithm (RFA) proposed by Leo Breiman [17] is composed of numbers of individual decision trees. To avoid the dependence between each decision tree, features at each node to determine the split are selected randomly. To build the individual decision tree, the training set is separated into in-bag set (inBag) and out-of-bag set (ooBag) as shown in Fig. 3. The in-bag set is used to build a single decision tree; and the out-of-bag set is used to evaluate the classification accuracy of each decision tree. Each individual decision tree cast a vote for one class. The vote received by each class is tallied; and the class with the majority vote is declared to be the classified class as shown in Fig. 4.

IV. Experimental Results

A. Data Accuaction

In this paper, to investigate the fault diagnosis of ball bearing of three phase induction motor using C4.5 decision tree and random forest algorithm, sets of vibration data of induction motor with four different conditions (normal, inner race fault, ball fault and outer race fault) obtained from [18] are used.

The experimental apparatus in [18] is shown in Fig. 5 and the state of operation is listed on Table I. There are two test ball bearings installed at the drive end and the fan end of the induction motor. The experiments are performed using a bearing with normal condition, inner race fault, ball fault and outer race fault at the drive end, respectively. The fault is created by introducing a single point defect with diameter of 0.007 inch and depth of 0.011 inch. The motor shaft is rotated with speed 1794 rpm, and the motor horse power is 0 hp. Accelerometers with sampling rate of 12,000 are attached to the housing of both drive end and fan end.

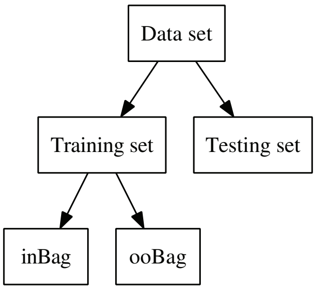

Figure 3. The processing of data set for random forest algorithm.

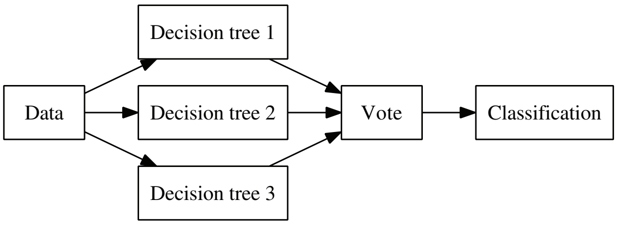

Figure 4. The procedures of classification for random forest algorithm.

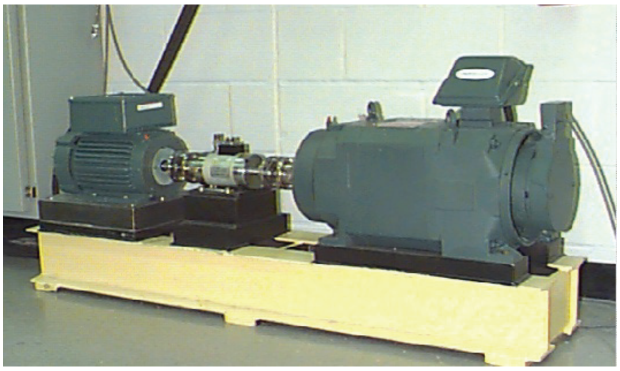

Figure 5. The experimental apparatus [18].

TABLE I. STATE OF MACHINE OPERATON

Parameter	*Value*
Speed of motor shaft	1797 rpm
Motor horsepower	0 hp
ball bearing defect location	Drive end
Size of defect	0.007 in (diameter) 0.011 in (depth)
Length of extracted signal	8192 samples

In this paper, two cases are considered:

Case I – *The source of the training set is same as the testing set:* To evaluate the effectiveness of the classifiers, as shown in Table II, 80 sets of vibration signal at the fan end and drive end under 4 different operation conditions (normal, inner race fault, ball fault and outer race fault) are measured. Among these 80 sets of data, 40 sets are used to train the classifiers (training set) and the other 40 sets (testing set) are used to test the classifiers.

Case II – *The source of the training set is different from the testing set:* To test the robustness of the classifier, as shown in Table II, 40 sets of vibration signal measured at the drive end under different operation conditions (normal, inner race fault, ball fault and outer race fault) are used to train the classifiers (training set); and another 40 sets of data (testing set) measured at the fan end are used to test the classifiers. In both cases, the length of the measured vibration signal is 8192.

B. Classifier Development and Performance Evaluation

The process block diagram of classifier development is shown in Fig. 6. The obtained training sets are analyzed in time domain and frequency domain. Thirteen features listed in Table III are then extracted from each domain. Features without reduction, with LDA, PCA, and both PCA and LDA are used to develop the classifiers, respectively. The testing sets are then used to evaluate the performance of each classifier. To evaluate the classifier performance, the classification accuracy is defined as follows:

$$\%Classification\_accuracy = \frac{number\_of\_right\_classification}{number\_of\_total\_test\_data} \times 100 \qquad (7)$$

Figures 7(a) and 7(b) show the classification accuracy of classifiers developed by C4.5 and RFA using the data of case I analyzed in time domain and frequency domain, respectively. Features are processed in 4 different ways: no feature reduction, reduced with LDA, PCA, and both LDA and PCA, respectively. In this work, the feature-dimensions were reduced to three and five dimensions by LDA and PCA, respectively. Figure 8(a) and 8(b) show the classification accuracy of classifiers developed by C4.5 and RFA using the data of case II analyzed in time domain and frequency domain, respectively. Same as Fig. 7, features are processed in 4 different ways: no feature reduction, reduced with LDA, PCA, and both LDA and PCA, respectively.

As shown in Figs. 7 and 8, different feature processes will result in different classification accuracy. From Figs. 7(a) and Fig. 8(b), when the feature dimensions are reduced by both LDA and PCA, the classification accuracy are reduced significantly. This may be caused by the over reduction of feature dimensions. From Figs. 7 and 8, in most cases, RFA performs better than C4.5. However, as seen from case I, when the features are properly processed, both the C4.5 and RFA can diagnose the fault effectively. Comparing the classification accuracy for case I and case II, case II does not work well as case I, since the signal of training set and testing set in case I are measured from different positions .

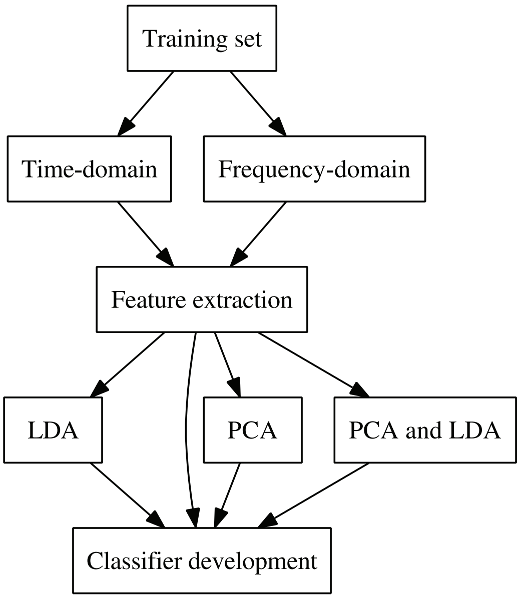

Figure 6. The processing steps of training set for calssifier development.

TABLE II. DATASET INFORMATION

Type of set	*Class*	*Case I*	*Case II*
Training set	Normal	Drive end (5) Fan end (5)	Drive end (10)[a]
	Inner race fault	Drive end (5) Fan end (5)	Drive end (10)
	Ball fault	Drive end (5) Fan end (5)	Drive end (10)
	Outer race fault	Drive end (5) Fan end (5)	Drive end (10)
	Total	40 sets	40 sets
Testing set	Normal	Drive end (5) Fan end (5)	Fan end (10)
	Inner race fault	Drive end (5) Fan end (5)	Fan end (10)
	Ball fault	Drive end (5) Fan end (5)	Fan end (10)
	Outer race fault	Drive end (5) Fan end (5)	Fan end (10)
	Total	40 sets	40 sets

[a] The number in the parentheses is the number of data sets.

TABLE III. THIRTEEN FEATURES EXTRACTED FROM THE DATA

Feature	Equation
Standard deviation	$x_{std} = \sigma_X = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}$
Variance	$x_{var} = \sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$
Median	$x_{median} = \text{median}\left(x\right)$
Kurtosis	$x_{kurtosis} = \frac{1}{\left(N-1\right)\sigma_X{}^4}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^4$
Skewness	$x_{skewness} = \frac{1}{\left(N-1\right)\sigma_X{}^3}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^3$
Minimum	$x_{min} = \min\left(x\right)$
Maximum	$x_{max} = \max\left(x\right)$
Range	$x_{range} = x_{max} - x_{min}$
Sum	$x_{sum} = \sum_{i=1}^{N} x_i$
Mean	$x_{mean} = \bar{x} = \mu_X = \frac{1}{N}\sum_{i=1}^{N} x_i$
Norm	$x_{norm} = \lVert x \rVert_2 = \sqrt{\sum_{i=1}^{N}\lvert x_i \rvert}$
Root-mean-square	$x_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$
Crest factor	$x_{CF} = \frac{x_{\max}}{x_{rms}}$

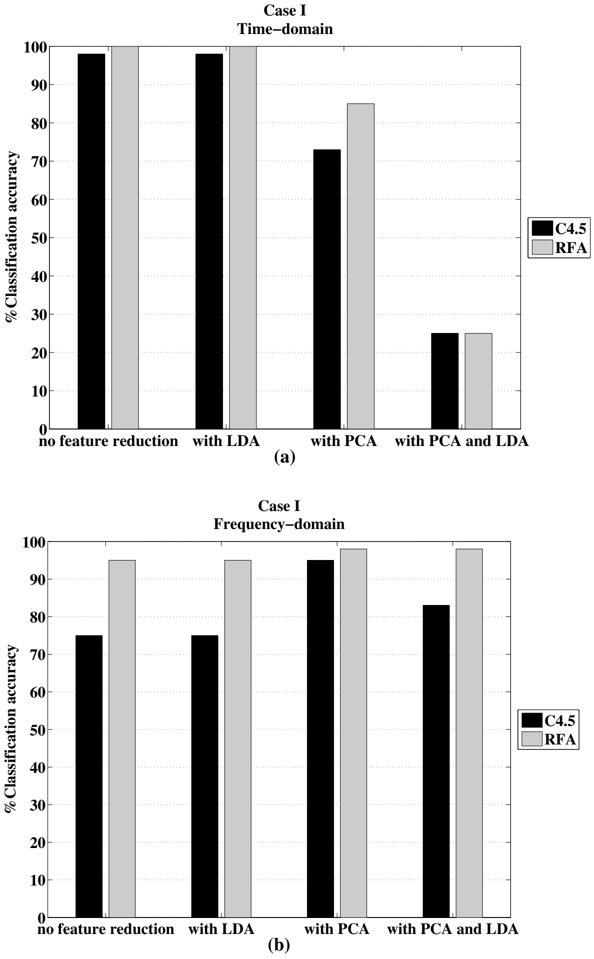

Figure 7. (a) Classification accuracy of classifiers developed by C4.5 and RFA using the data of case I analyzed in time domain (b) Classification accuracy of classifiers developed by C4.5 and RFA using the data of case I analyzed in frequency domain. The star symbol represents imperfect classification.

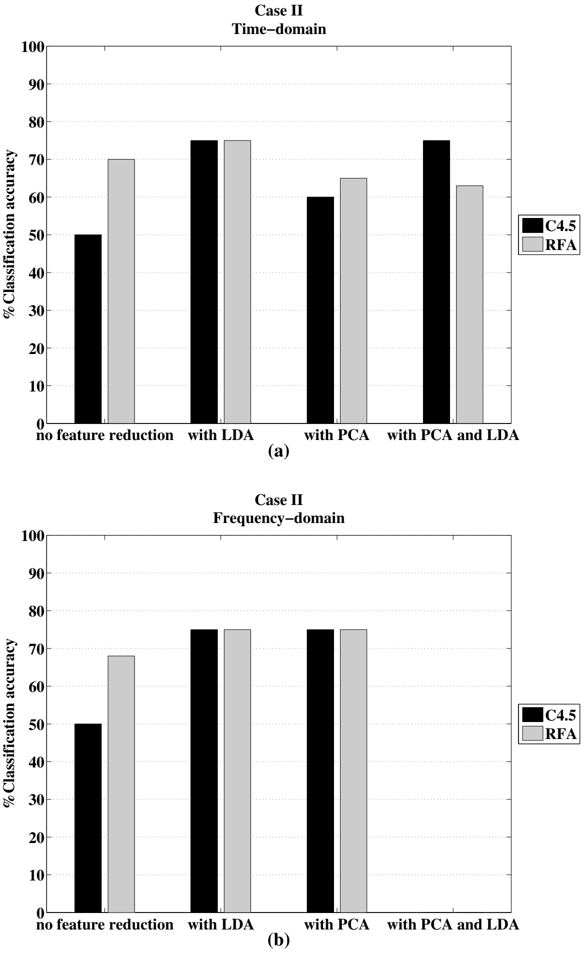

Figure 8. (a) Classification accuracy of classifiers developed by C4.5 and RFA using the data of case II analyzed in time domain (b) Classification accuracy of classifiers developed by C4.5 and RFA using the data of case II analyzed in frequency domain. The star symbol represents imperfect classification.

V. CONCLUSION

In this paper, C4.5 decision tree and random forest algorithm are used to diagnose the fault of ball bearing of three phase induction motor. The features for classifier development are extracted from the vibration signal analyzed in time domain and frequency domain, and processed by PCA and LDA to reduce the feature dimensions. The results show that different feature processes will result in different classification accuracy. In most cases, the random forest algorithm performs better than the C4.5. However, with proper feature process, both C4.5 and random forest can diagnose the fault effectively. In the future, fault diagnosis using other machine learning algorithms will be investigated and compared.

ACKNOWLEDGMENT

The data used in this paper is obtained from Case Western Reserve University Bearing Data Center, which provides access to ball bearing test data for normal and faulty bearings. And the classifier development are performed by a free data mining software WEKA.

REFERENCES

[1] Zengbing Xu, Jianping Xuan, Tielin Shi, Bo Wu and Youmin Hu, “A novel fault diagnosis method of bearing based on improved fuzzy ARTMAP and modified distance discriminant technique,” Expert Systems with Applications, vol. 36 (9), pp. 11801-11807, November 2009.

[2] Kaveh Mollazade, Hojat Ahmadi, Mahmoud Omid and Reza Alimardani, “An Intelligent Combined Method Based on Power Spectral Density, Decision Trees and Fuzzy Logic for Hydraulic Pumps Fault Diagnosis,” International Journal of Computer Systems Science and Engineering, vol. 3, Number 2008.

[3] Dejie Yu, Junsheng Cheng and Yu Yang, "Application of EMD method and Hilbert spectrum to the fault diagnosis of roller bearings," Mechanical Systems and Signal Processing, vol. 19 (2), pp. 259-270, March 2005.

[4] Fidel Ernesto Hernandez Montero and Oscar Caveda Medina, "The application of bispectrum on diagnosis of rolling element bearings: A theoretical approach," Mechanical Systems and Signal Processing, vol. 22 (3), pp. 588-596, April 2008.

[5] Chebil, J., Noel, G., Mesbah, M. and Deriche, M., "Wavelet Decomposition for the Detection and Diagnosis of Faults in Rolling Element Bearings," Jordan Journal of Mechanical and Industrial Engineering, vol. 3, pp. 260-267, Number 4.

[6] Weixiang Sun, Jin Chen and Jiaqing Li, "Decision tree and PCA-based fault diagnosis of rotating machinery," Mechanical Systems and Signal Processing, vol. 21 (3), pp. 1300-1317, April 2007.

[7] B. SAMANTA and K. R. AL-BALUSHI, "ARTIFICIAL NEURAL NETWORK BASED FAULT DIAGNOSTICS OF ROLLING ELEMENT BEARINGS USING TIME-DOMAIN FEATURES," Mechanical Systems and Signal Processing, vol. 17 (2), pp. 317-328, March 2003.

[8] V. Sugumaran and K.I. Ramachandran, "Automatic rule learning using decision tree for fuzzy classifier in fault diagnosis of roller bearing," Mechanical Systems and Signal Processing, vol. 21 (5), pp. 2237-2247, July 2007.

[9] Ngoc-Tu Nguyen, Jeong-Min Kwon and Hong-Hee Lee, "Fault diagnosis of induction motor using decision tree with an optimal feature selection," Power Electronics, 2007. ICPE '07. 7th Internatonal Conference, pp.729-732, 22-26 Oct. 2007.

[10] Zengbing Xu, Jianping Xuan, Tielin Shi, Bo Wu and Youmin Hu, "Application of a modified fuzzy ARTMAP with feature-weight learning for the fault diagnosis of bearing," Expert Systems with Applications, vol. 36 (6), pp. 9961-9968, August 2009.

[11] Yaguo Lei, Zhengjia He and Yanyang Zi, "Application of a Novel Hybrid Intelligent Method to Compound Fault Diagnosis of Locomotive Roller Bearings," Journal of Vibration and Acoustics, vol. 130, pp. 0345011-03450116, June 2008.

[12] P.K. Kankar, Satish C. Sharma and S.P. Harsha, "Fault diagnosis of ball bearings using machine learning methods, Expert Systems with Applications," vol. 38 (3), pp. 1876-1886, March 2011, in press.

[13] Zhiyuan Yang and Qinming Tan, "The Application of Random Forest and Morphology Analysis to Fault Diagnosis on the Chain Box of Ships," Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium, pp.315-319, 2-4 April 2010.

[14] Zhinong Li, Zhaotong Wu, Yongyong He and Chu Fulei, "Hidden Markov model-based fault diagnostics method in speed-up and speed-down process for rotating machinery," Mechanical Systems and Signal Processing, vol. 19 (2), pp. 329-339, March 2005.

[15] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski and Lukasz A. Kurgan, Data Mining: A Knowledge Discovery Approach, 1st ed, Springer, 2007, pp. 133-152

[16] J. Ross Quinlan, C4.5:programs for machine learning, Morgan Kaufmann Publishers, Inc., 1993.

[17] Leo Breiman, "Random Forests", Machine Learning, vol. 45 (1), pp. 5-32, 2001.

[18] Loparo, K. A., Bearings vibration data set. Case Western Reserve University, 2003.
<http://www.eecs.cwru.edu/laboratory/bearing/download.htm>