Health Assessment and Fault Classification for Centrifugal Pump Using Logistic Regression

Chen Lu and Jian Ma[1]
School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China
Science & Technology on Reliability & Environmental Engineering Laboratory, Beijing 100191, China

E-mail: majian3128@126.com

Abstract. Real-time health monitoring of industrial components and systems that can detect, classify and predict impending faults is critical to reducing operating and maintenance cost. This paper presents a logistic regression based prognostic method for on-line health assessment and failure modes classification. System condition is evaluated by processing the information gathered from access controllers or sensors mounted at different points in the system, and maintenance is performed only when the failure/ malfunction prognosis indicates instead of periodic maintenance inspections. The wavelet packet decomposition and fast Fourier transform (FFT) technique is used to extract features from non-stationary vibrations signals, wavelet package energies and fundamental frequency amplitude are used as features and Principal Component Analysis (PCA) is used to features reduction. Reduced features are input into logistic regression (LR) models to assess machine health condition and identify possible failure modes. The maximum likelihood method is used to determine parameters of LR models. The effectiveness and feasibility of this methodology have been illustrated by applying the method to a real centrifugal pump.

1. Introduction

Considerable efforts have been made to develop methods and tools to diagnose failures. However, limited results have been given on prognostics that can detect, analyze and correct equipment problems well before failures actually manifest, also provide system operators with a sufficient time window to schedule maintenance without disrupting the operations [1-4]. This paper presents a prognostic method for on-line centrifugal pump health assessment and root cause classification using multiple logistic regression (LR).

The paper is organized as follows. Section 2 provides state of the art of the prognostic methodology along with related mathematics. Section 3 illustrates test bed setup of centrifugal pump and assessment/classification results obtained from an application of the proposed schemes on real data. Section 4 concludes the paper with a summary and future research directions.

2. Methodology

The prognostic scheme is based on monitored data which contain centrifugal pump incipient failure signatures; intelligent mathematical techniques which can be incorporated to detect, evaluate the risk of failure over some protracted period of time and also classify which particular type of failure may occur.

2.1. Procedures of the Methodology

There are three major steps of the methodology. Step 1: feature extraction and selection that extracts and determines appropriate features for health assessment and root cause classification as well as to reduce the search space for fast computation. Step 2: health assessment and root cause classification by means of the logistic regression method.

[1] To whom any correspondence should be addressed.

2.2. Feature Extraction Using Wavelet Packet Decomposition

In practice, fault signal of centrifugal pump is usually distributed in both high and low frequency band where wavelet packet decomposition can reach a very delicate degree. As we all know, Wavelet packet analysis (WPA) constructs a more sophisticated method of orthogonal decomposition on the basis of multi-resolution analysis, which can divide the full frequency band of signal in multi-level, so that each band's signal contain more elaborate information about the original signal. Therefore wavelet packet decomposition is suitable to extract both low and high frequency features. Statistically analyzing all bands of signal decomposed by wavelet packet, energy index of all bands reflecting signal characteristics can be constructed. The determination of wavelet packet decomposition scale is an issue that can't be ignored. If wavelet packet decomposition scale is too little, the fault feature can't be extracted effectively, whereas too many scale will increase the dimension of feature vector, consequently the calculating rate can be affected [5]. Therefore in centrifugal pump health assessment and fault classification , according to the vibration signal characteristics, eight frequency band energy index E_{3j} can be calculated by three-layer decomposition.

$$E_{3j} = \int \left| S_{3j}(t) \right|^2 d_t = \sum_{k=1}^{n} \left| x_{jk} \right|^2 \tag{1}$$

where, $\mathrm{x}_{jk}\left(j = 0,1,\cdots,7; k = 1,2,\cdots,n\right)$ stands for the amplitude of reconstruction signal S_{3j}.

When centrifugal pump has a fault, the energy of each band signal will have a great impact, so the energy should be normalized into a feature vector *T*.

$$T = [E_{30}/E, E_{31}/E, E_{32}/E, E_{33}/E, E_{34}/E, E_{35}/E, E_{36}/E, E_{37}/E] \tag{2}$$

$$E = (\sum_{j=0}^{7} \left| E_{3j} \right|^2)^{1/2} \tag{3}$$

2.3. PCA based Feature reduction

For the sake of completeness, the PCA procedure employed in this paper will be briefly presented. Let us assume that at a given centrifugal pump operation stage *S* (in this paper, only the normal machine behavior and machine operation with a worn bearing/ Impeller are considered), the signal features *X* are characterized by the multivariate Gaussian distribution with mean $\overline{\mu}_s$ and the covariance matrix K_s. The symmetric matrix K_s can now be represented as

$$K_s = \sum_{i=1}^{r} \lambda_i \overline{v}_i \overline{v}_i^T = V \Lambda V^T \tag{4}$$

where *r* is the rank of the covariance matrix K_s, λ_i, $i = 1,2,...,r$ are the non-zero eigenvalues of K_s, $\overline{v}_i$ are the corresponding unit norm eigen-vectors and

$$V = \left[\overline{v}_1 \quad \overline{v}_2 \quad \cdots \quad \overline{v}_r \right]; \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_r \end{bmatrix} \tag{5}$$

Due to the positive semi definiteness of K_s, all its eigenvalues are real and greater than, or equal to zero. Each eigenvalue λ_i, $i = 1,2,...,r$ depicts the amount of the covariance matrix energy projected in the direction of the corresponding eigenvector $\overline{v}_i$. When there exists a high degree of correlation among the components of *X*, only a few of the eigenvalues in Λ ccount for most of the energy3 in the

covariance matrix K_s. Thus, assuming that eigenvalues λ_i, $i=1,2,...,r$ are arranged in descending order, (3) can be represented as

$$K_s = \sum_{i=1}^{p} \lambda_i \overline{V}_i \overline{V}_i^T = V_p \Lambda_p V_P^T \tag{6}$$

where

$$V = \left[\overline{V}_1 \quad \overline{V}_2 \quad \cdots \quad \overline{V}_p \right]; \ \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_p \end{bmatrix} \tag{7}$$

p is the number of the principal components of K_s, λ_i, $i=1,2,...,r$ are the largest p eigenvalues of K_s, and $\overline{V}_i$ are the corresponding unit norm eigenvectors.

A query item $\tilde{X}$ can now be transformed into a p component random variable $\tilde{Y}$ given as

$$\tilde{Y} = T\left(\tilde{X} - \overline{\mu}_s\right), \ T = \Lambda_P^{-1/2} \Lambda_p \tag{8}$$

If $\tilde{X}$ belongs to the class of signals from centrifugal pump state S, then $\tilde{Y}$ should be normally distributed with zero mean and variance I_p, where I_p is the unity matrix of order p. Thus, for each query item $\tilde{X}$, its adherence to the class S can be assessed through the Euclidean norm of the vector $\tilde{Y}$, which in turn corresponds to assessment and classification based on the Logistic Regression of the query item from the training classes [6].

2.4. Logistic Regression Method

The machine condition description from daily maintenance records/logs is a dichotomous problem (either normal or failed) which can be represented using a logistic regression function [1]. The goal of logistic regression is to find the best fitting model to describe the relationship between the categorical characteristic of dependent variable (the probability of an event, constrained between 0 and 1) and a set of independent variables. The logistic function is

$$prob(event) = p(x) = \left(1 + e^{-g(x)}\right)^{-1} \tag{9}$$

The logistic or logit model is

$$Logit = g(x) = \log\left(p(x)\left(1 - p(x)\right)^{-1}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{10}$$

where $g(x)$ is a linear combination of the independent variables $\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_k$.

The pre-condition for figuring out $P(x)$ is determining parameters α and $\beta_1, \cdots, \beta_k$ in advance. Due to the fact that the dichotomous dependent variable makes estimation using ordinar y least squares inappropriate, rather than choosing parameters that minimize the sum of squared errors, estimation in logistic regression chooses parameters of α and $\beta_1, \cdots, \beta_k$ using the maximum likelihood method [1]. Then, the probability of failure for each input vector x can be calculated according to Eq. (9).

3. Experimental Result

The methodology has been implemented in a centrifugal pump (see, Fig. 1) to evaluate health

condition dynamically. Also fault mode analysis was performed in order to find out the possible root cause.

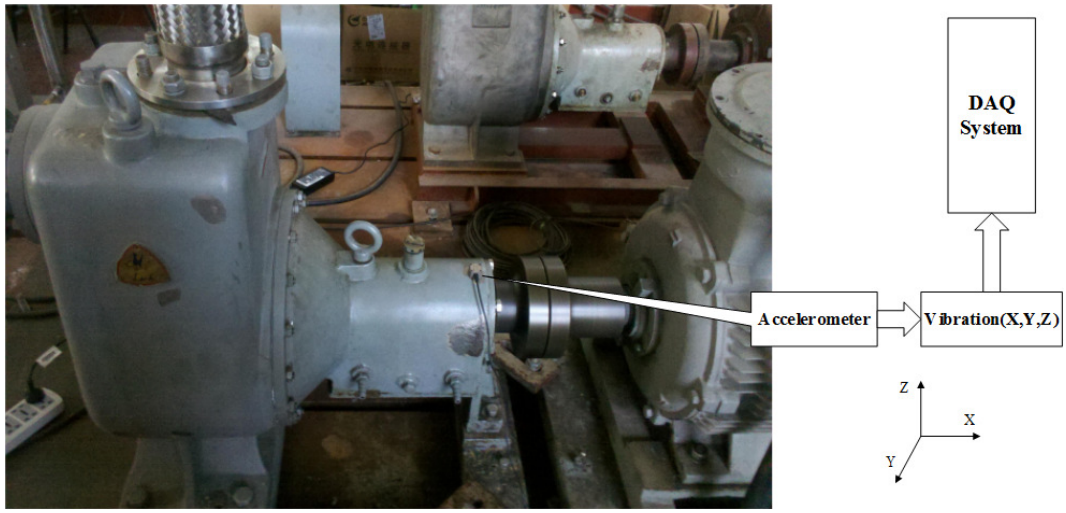

Figure 1. Centrifugal pump data acquisition system.

3.1. Data Acquisition System Description

Three vibration signals are acquired from an installed accelerometer, with a sampling rate as 10.24kHz.

3.2. Feature Extraction and Reduction

First of all, FFT was used for each vibration signal to obtain fundamental frequency amplitude. Second, three-level wavelet packet decomposition using Daubechies wavelet (DB10) was adopted for each vibration signal, then fundamental frequency amplitude and package energies were used as features. A subset of feature components was determined using PCA. In this case, after feature reduction, a 4-dimension feature vector is finally selected as feature vector for health assessment and fault mode classification.

3.3. LR Models Training

- LR model trained for health assessment.
  120 sets of data were used as training data, including 40 sets of data sampled under normal conditions (P(x)=0) versus 80 sets of fault data (P(x)=1). The parametersα and $\beta_1,\cdots,\beta_4$ were estimated using the maximum likelihood method to eventually obtain the model for performance assessment as LR1.
- LR models trained for root cause classification.
  Fault mode 1 (bearing roller wearing): 40 sets of bearing roller wearing data (P(x)=1) versus 40 sets of non-wearing data (P(x)=1) were used to train the classification model (LR2) for fault mode 1 using the maximum likelihood method;
  Fault mode 2 (bearing inner race wearing): 40 sets of bearing inner race wearing data (P(x)=1) versus 40 sets of non-wearing data (P(x)=0) were used to train the classification model (LR3) for fault mode 2 using the maximum likelihood method;
  Fault mode 3 (bearing outer race wearing): 40 sets of bearing outer race wearing data (P(x)=1) versus 40 sets of non-wearing data (P(x)=0) were used to train the classification model (LR4) for fault mode 3 using the maximum likelihood method;

Fault mode 4 (Centrifugal Pump impeller wearing): 40 sets of Impeller wearing data (P(x)=1) versus 40 sets of non-wearing data (P(x)=0) were used to train the classification model (LR5) for fault mode 4 using the maximum likelihood method.

3.4. Validation

Four sets of centrifugal pump vibration data (one set for each fault mode) were used for validation.

- Set 1 (Fault mode 1): 10 data under normal condition versus 40 bearing roller wearing data;
- Set 2 (Fault mode 2): 10 data under normal condition versus 32 bearing inner race wearing data;
- Set 3 (Fault mode 3): 10 data under normal condition versus 40 bearing outer race wearing data;
- Set 4 (Fault mode 4): 10 data under normal condition versus 30 bearing Impeller wearing data.

Note that the confidence value (CV) is calculated based on the probability of failure. Define CV=1−P(x) [1]. When the centrifugal pump is running normally, CV is close to 1, if the centrifugal pump is going to fail, CV is approaching 0 correspondingly. If the confidence value is less than a prefixed threshold, for example, 0.8, the root cause classification module will be triggered, and features are input into fault modes classification models to calculate the probability of each fault.

Figure 2 (a), (b), (c) and (d) shows the overall health assessment of the four sets of centrifugal pump vibration data using model LR1, the probability of different fault modes conducted from LR2, LR3, LR4 and LR5 is shown in Fig. 3(a), (b), (c) and (d).

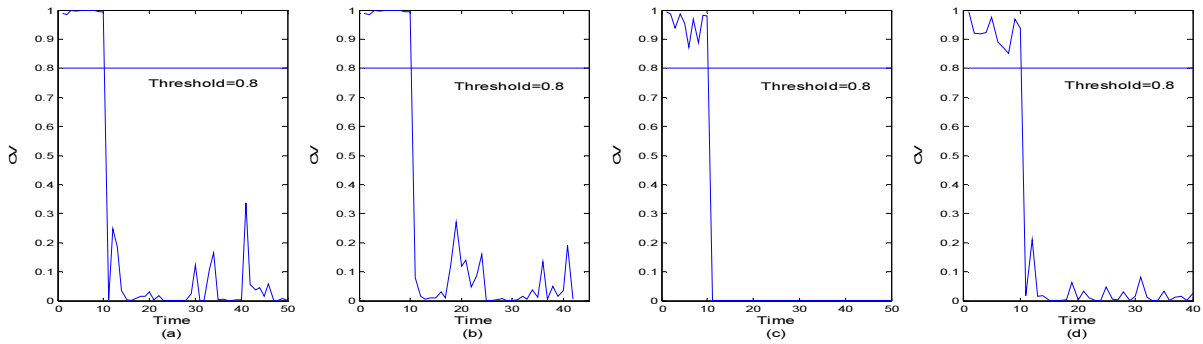

Figure 2. Health assessment result of four fault modes.

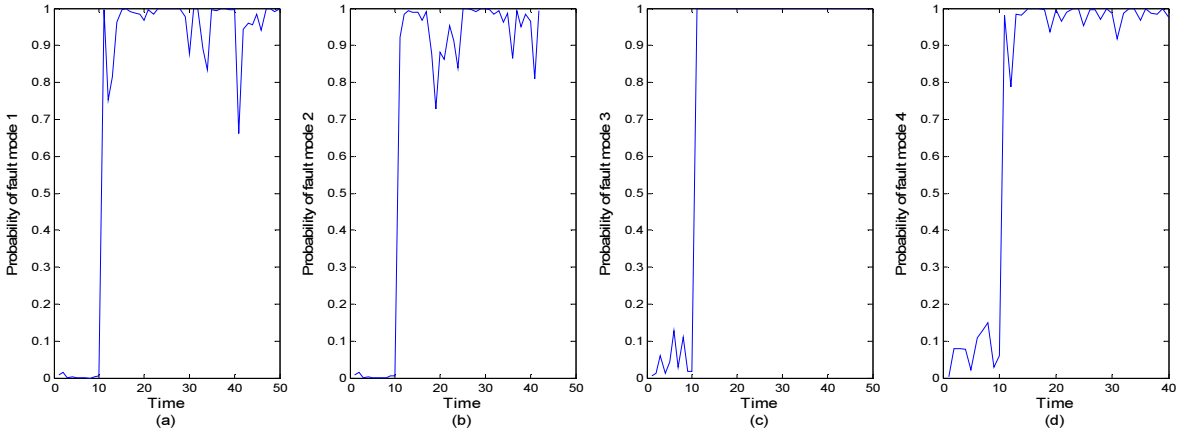

Figure 3. Probability of fault modes 1, 2, 3 and 4.

In Fig. 2, both bearing and impeller problems can be detected from the CV drops, but it is hard to clarify what the difference is of the four drops and what caused the drops. In this methodology, the root cause classification module is triggered as long as the confidence value is below a predetermined

threshold (0.8) by inputting the corresponding features into the trained models (LR2, LR3, LR4 and LR5) to calculate the probability of fault modes. From time 10, the probability of fault 1 (bearing roller wearing), fault 2 (bearing inner race wearing), fault 3 (bearing outer race wearing) and fault 4 (impeller wearing) is very high, see the solid line in Fig. 3(a), (b), (c) and (d) respectively. Consequently minor probability of failure of these points is indicated in Fig. 3.

4. Conclusions

A logistic regression based approach for centrifugal pump health assessment and root cause classification has been performed in this paper. LR combined with maximum likelihood technique is an effective and efficient tool for health assessment and root cause classification dynamically. WPT combined with PCA is a suitable feature extraction step where appropriate features can be obtained from non-stationary signals. The method is generic and shows promising results for the analysis of both stationary and non-stationary signals, which could be applied to other centrifugal pump.

However, four types of centrifugal pump fault modes are considered in this paper, more fault modes should be taken into consideration for better performance on centrifugal pump health assessment and root cause classification. Also when the process is not time shifted, the coefficients of WPT can be used as features directly instead of using packet energy, which will be investigated further in future research and application.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant Nos. 61074083, 50705005 and 51105019), as well as the Technology Foundation Program of National Defense (Grant No. Z132010B004).

References

[1] Yan J and Lee J 2005 *J. ASME Journal of Manufacturing Science and Engineering* **127** 912–4
[2] Yan J, Lee J and Koc M 2002 *Fifth International Conference on Managing Innovations in Manufacturing (MIM)* Milwaukee WI 172–8
[3] Kacprzynski G J and Roemer M J 2000 *International COMADEM Congress* Houston TX
[4] Liao L X and Lee J 2010 *J. Expert Systems with Applications* **37** 240–52
[5] Yen G G and Lin K C 2000 *IEEE Trans. Ind. Electron.* **47** 650–67
[6] Djurdjanovic D, Ni J and Lee J 2002 *ASME International Mechanical Engineering Congress and Exposition IMECE* 32032