RLCP: A Reinforcement Learning Method for Health Stage Division Using Change Points

Yijun Cheng, Jun Peng, Xin Gu, Xiaoyong Zhang, Weirong Liu, Yingze Yang and Zhiwu Huang*
School of Information Science and Engineering, Central South University, Changsha, China
Hunan Engineering Laboratory of Rail Vehicles Braking Technology, Changsha, China
Email: hzw@mail.csu.edu.cn

Abstract—**The health stage division problem has been attracting interest for its potential role in Prognostics and Health Management (PHM). In traditional health stage division, multiple health indicators (HIs) are extracted from original data, and then one suitable HI is constructed. According to the trend of the one HI, the whole degradation process is divided into different health stages based on the *change points*. The change points are the transition points between different health stages. But the one HI is difficult to construct and just suitable for special applications. Therefore how to consider multiple HIs simultaneously for health stage division is a big problem. In this paper, a reinforcement learning-based method by using change points (RLCP) is proposed to divide health state. The health stage division problem based on change points is modelled as a Markov Decision Process (MDP). Then reinforcement learning is introduced to solve the MDP effectively. The performance of the proposed RLCP is evaluated with the PRONOSTIA dataset. Experiment results show that the result of health stage division is optimal when compared with existing methods.**

I. Introduction

In recent years, PHM has attracted much attention from both academic and industrial community. The PHM is important for machinery to save significant costs because of its near-zero downtime [1]. The PHM consists of four step, i.e., data acquisition, HI construction, health stage division, and remaining useful life (RUL) prediction [2]. The health stage division is the important step to provide accurate inferences for better RUL prediction [3].

In the health stage division step, the whole degradation process is divided into different health stages. The trend of degradation process is hard to observe from original data directly. Therefore, various HIs are extracted from the original signal. According to the trends of HIs, different health stages of the whole degradation are obtained. But the trends of various HIs are different. How to divide the health stages when considering multiple HIs is a big matter.

Most works construct only one suitable HI, which is fused by multiple HIs [4–7]. According to the trend of one HI, the health state are divided into different health stages based on the *change points*. The change points are the transition points between different health stages [8].

Bechhoefer et al. [4] constructed a component HI from six appropriate condition indicators. The fused HI was used to divide the whole degradation process for gear fault prognosis. Liu et al. [5] presented a data-level fusion model for constructing a composite HI. The composite HI could characterize the different degradation states. Hong et al. [6] calculated a confidence value as an HI for identifying four degradation stages. Hu et al. [7] presented a probability evaluation method based on temperature data. The conditional probability was obtained by the proposed method to derive the degradation degree. These methods construct one HI effectively but just suitable for special applications.

To overcome the limitation above, some works consider multiple HIs simultaneously to divide the health stages. Soualhi et al. [9] detected the degradation states by a supervised classification algorithm called support vector machine (SVM). The classification algorithm depends on history data. Scanlon et al. [10] used a unsupervised clustering algorithm called k-means to determine the class label of degradation process. A smoothing algorithm was also employed to redefine the class label boundaries. However, these works did not divide the whole degradation using change points. The boundaries between different health stages are still not clear.

Consequently, how to consider multiple HIs and change points simultaneously for health stage division is a big challenge. In this paper, a reinforcement learning-based method by using change points (RLCP) is proposed to meet the challenge. The proposed RLCP is divided into three steps. First, due to the continuous temporal characteristics in the degradation process, six classical time domain features are extracted as multiple HIs. Second, the change points between different health stages are regarded as an agent. The health stage division problem based on change points is modeled as an MDP. Third, reinforcement learning is adopted to solve the MDP through the trial-and error learning with environment. The proposed RLCP is expected to obtain the health stage division when considering the multiple HIs simultaneously.

The remainder of the paper is organized as follows: In section 2, reinforcement learning is introduced briefly. In section 3, the health stage division is modeled as an MDP. In section 4, the proposed RLCP is described in detail. In section 5, our proposed RLCP is verified by the experiment in the PRONOSTIA.

II. Reinforcement Learning

Reinforcement learning is a goal-directed leaning method, which could learn from the interaction with environment [11]. The learner is called *agent*, and everything it interacts with is called *environment* [12]. The agent incrementally updates its

978-1-5090-0382-2/16/$31.00 © 2018 IEEE

978-1-5386-1164-7/18/$31.00 ©2018 IEEE

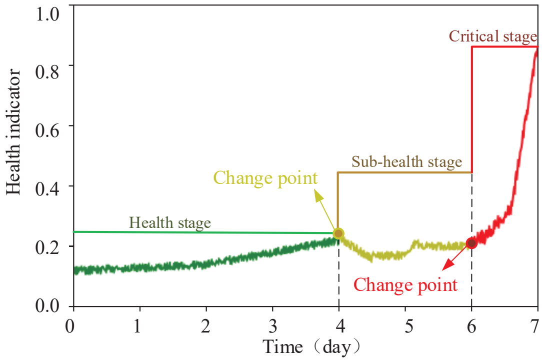


Fig. 1: The n health stages based on $n-1$ change points

knowledge as it interacts with environment by trial-and-error learning. At each time step t, the agent chooses the action A_t based on the observation O_t from the environment state S_t [13]. The state S_t generally exhibits the Markov property. The taken action A_t would change the environment state S_t. According to the change, the agent receives the immediate reward R_t from the environment.

The core of reinforcement learning is the mapping from the states to the actions, which is called *policy*. The goal of the agent is to maximize the long-term received reward in one episode. For each time step of one episode, the agent needs to choose an action from a set of possible actions [14] based on its learned policy.

To find the optimal policy, there are two basic problems of choosing actions [11]. One is the difficulties of evaluating the long-term rewards of different possible actions. The long-term rewards are not known until the end of the episode. And in the practical situation, we aren't aware of the complete knowledge of the environment. Only *experience* is known after the trial-and-error learning with the environment, which is the sample sequences of states and actions and rewards. So how to obtain the evaluations from the experience is important to reinforcement learning. The other is the dilemma between exploration and exploitation. If the agent only exploits the current maximum reward, the received reward may involve short-term reward. But if the agent only explores for the future reward, the maximum immediate reward may be missed. So the trade-off between exploration and exploitation also plays an important role in reinforcement learning.

III. System Model

The traditional degradation process is shown in Fig. 1 [2]. The whole degradation process is divided into n health stages based on one HI. The time is on the horizontal axis, and the HI is on vertical axis. Change points in the degradation process represent the transition points between different health stages. Hence the $n-1$ change points are the key to distinguish n health stages in degradation process.

For one HI, the change points are always set by alarm threshold. But considering multiple HIs, it is difficult to find the optimal change points. Therefore we regard the $n-1$ change points as an agent. And health stage division problem is built as an MDP, which is defined as a tuple $(\boldsymbol{S}, \boldsymbol{A}, T, R, \gamma)$. $\boldsymbol{S}$ represents the finite set of states, and every state s contains $n-1$ change points. $\boldsymbol{A}$ represents the finite set of actions, which contains all the action a from current state s moving into next state s'. $T(s,a,s')$ represents the state transition probability of the state s transforms into state s' after action a, which is deterministic in our problem. $R(s,a)$ represents the immediate reward received after taking action a in state s. $\gamma \in [0,1]$ represents the discount factor of future reward.

For each episode, the agent would move to the optimal state from the initial state as closely as possible. The policy $\pi(s)$ is the action sequence of every step in one episode, which is stated from the given initial state s. The return $G_t(\pi)$ based on $\pi(s)$ is the total discounted reward from time-step t in one episode as follow:

$$G_t(\pi) = R_{t+1} + \gamma R_{t+2} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}. \tag{1}$$

Theorem 1. *For any Markov Decision Process, there exists an optimal policy π^* that is better than or equal to all other policies,$\pi^* \geq \pi, \forall \pi$ [11].*

Based on the Theorem, the optimal policy $\pi^*(s)$ must exist. In the next section, we would introduce the proposed method to find the optimal policy $\pi^*(s)$ in given initial state s to maximize the expected return $\mathrm{E}_\pi(G_t|S_t = s)$.

IV. The Proposed RLCP Method for Health Stage Division

The proposed RLCP includes the following three steps. First, the feature extraction step extracts a list of feature from original signal. Second, the system model step models the health stage division as an MDP. Third, the reinforcement learning step solves the division problem eventually.

In the first step, the time domain features are used because the whole degradation process is the continuous time series. Six classical time domain features are extracted in this problem, which are the mean value (first moment), the root mean square, the crest factor, the average power, the skewness (third moment) and the kurtosis (fourth moment). The six features can measure the signal distribution in general [15]. The six feature and their equation are summarized in Table I, where x_i is the original data points, n is the number of points, and $\overline{x}$ is the mean value.

In the second step, the health stage division problem is modelled as an MDP as shown in the section III. In the third step, through the analysis in section II, reinforcement learning is adopted to solve the MDP problem of health stage division based on change points. To find the optimal change points, the two problems mentioned in section II are solved as follows.

To solve the problem of evaluating the long-term rewards, the action-value function $Q_\pi(s,a)$ is designed for estimating

TABLE I: Time domain features

Mean	$\frac{1}{n}\sum_{i=1}^{n} x_i$
Root Mean Square (RMS)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n} {x_i}^2}$
Crest factor	$\frac{\max\left[x_i\right]}{RMS}$
The average power	$\frac{1}{n}\sum_{i=1}^{n} {x_i}^2$
Skewness	$\frac{\mathrm{E}\left[\left(x_i-\overline{x}\right)^3\right]}{RMS^3}$
Kurtosis	$\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^4}{RMS^4}$

how good the action a is in the given state s. $Q_\pi(s,a)$ is defined as the expected long-term return for taking action a in given state s following policy π,

$$Q_\pi(s,a) = E\left[G_t|S_t = s, A_t = a\right]. \tag{2}$$

For given state s, the degradation process division is evaluated by the Calinski-Harabaz index [16] as follow,

$$S = \frac{Tr(B)}{Tr(W)} \times \frac{N-k}{k-1} \tag{3}$$

where k represents the number of health stages, N represents the number of points in the whole degradation process, B is the dispersion matrix between different health stages, W is the dispersion matrix within one health stage.

Therefore, the immediate reward R is defined as the improvement of degradation process division between the next state s' and the current state s as follow,

$$R = S(s') - S(s). \tag{4}$$

According to *Bellman's principle of optimality prescribes* [17], the $Q_\pi(s,a)$ is decomposed into immediate reward R plus discounted $Q_\pi(s',a')$ of successor state s'. Therefore, Eq.2 could be written as follow,

$$Q_\pi(s_t,a_t) = \mathrm{E}\left[R + \gamma Q_\pi(s'_t,a'_t)\right]. \tag{5}$$

In practical situation, the information of environment is not known. So the expected long-term reward is obtained from experience. In each time step of one episode, the agent chooses the action a_t in current state s_t based on the known action-value function $Q(s_t,a_t)$. After the action a_t is taken, the received reward R_t and the next state s'_t are used to evaluate the estimated return $Q_e(s_t,a_t)$ as follow,

$$Q_e(s_t,a_t) = R_t + \gamma \max_a Q(s'_t,a). \tag{6}$$

Algorithm 1 The proposed RLCP method for health stage division

```
Input:
    All the state $s \in \boldsymbol{S}$ in the whole degradation process
    The time domain features of each state $s$ in $S$
    All the action $a \in \boldsymbol{A}$ in each state $s$
    Initial state $s_0$
Output:
    The optimal policy $\pi^*$ for optimal change points
Initialize:
    The action-value function $Q(s,a), \forall s \in \boldsymbol{S}, a \in \boldsymbol{A}$
Repeat for each episode:
    Initialize: $s = s_0$
    Repeat for each step $t$ of episode:
        Choose $a$ in corresponding $s$ by Eq.8
        Take action $a$, observe:
            the immediate reward $R$ (obtained by Eq.4)
            and the next state $s'$
        Update the $Q(s,a)$ by Eq.7
        Update the state $s \leftarrow s'$
    end for
end for
```

The action-value function Eq. 5 is updated towards to the estimated return Eq. 6 by the following Eq. 7,

$$\begin{aligned} Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[R_t + \gamma \max_a Q(s'_t,a) \\ -Q(s_t,a_t)] \end{aligned}, \tag{7}$$

where $\alpha \in (0,1)$ is the learning rate, which represents how quickly the agent learns from new ones. The action-value Q is updated constantly step by step. After a number of episodes, the Q derived from experience is approximate to the actual value.

To solve the problem of the trade-off between exploration and exploitation, the $\varepsilon - greedy$ algorithm is used in this paper. With the probability ε, the agent chooses an action at random. With the probability $1-\varepsilon$, the agent chooses an action based on the maximal action-value function. The mapping from state s to action a is as follow,

$$\pi\left(\mathrm{a}|s\right) = \begin{cases} \frac{\varepsilon}{m} + 1 - \varepsilon, & \text{if } a^* = \arg\max_{a \in A} Q(s,a) \\ \frac{\varepsilon}{m}, & \text{otherwise} \end{cases}, \tag{8}$$

where m is the number of all the actions. The $\varepsilon - greedy$ algorithm is a near-greedy algorithm. The greedy action is chosen most of the time, which is the process of exploitation. But every once in a while, the action is chosen randomly, which is the process of exploration. Therefore, the trade-off between exploration and exploitation is solved by $\varepsilon - greedy$.

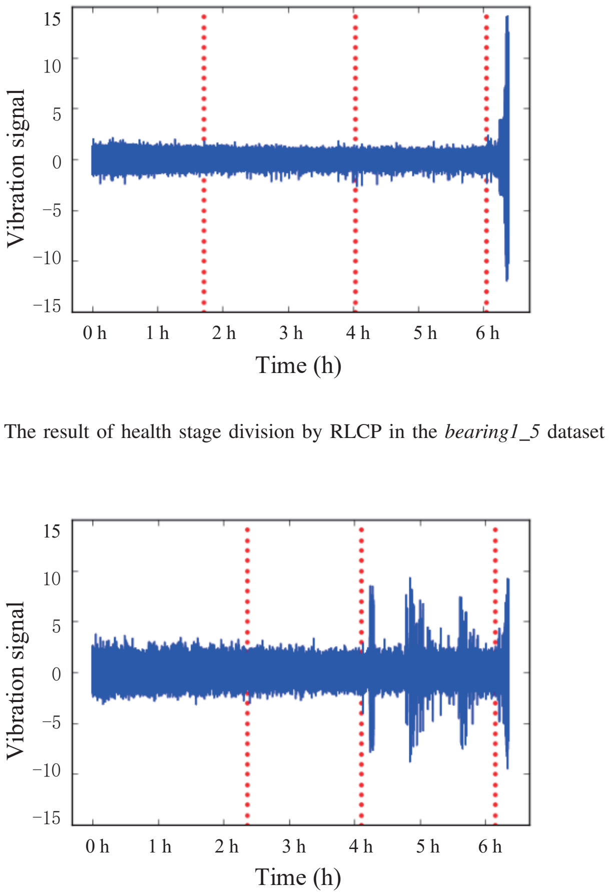
After overcoming the two problems, reinforcement learning solves the MDP eventually. The agent has found the optimal policy $\pi^*(A|s)$ that maximizes the action-value over all policies. The pseudocode of the proposed RLCP method is described in Algorithm 1.

V. Experiment demonstration

In this section, the proposed RLCP is verified in the PRONOSTIA. The PRONOSTIA comes from IEEE PHM 2012 prognostic challenge [18]. The experimental platform is able to conduct the bearing degradation in a few hours. The vibration signals are collected at a sampling frequency of 25.6 kHz every 10 s for a period of 0.1 s. Different datasets in different conditions are obtained in this experimental platform. The *Bearing1_5* and *Bearing1_6* dataset of bearing 1 are used in this experiment, which have almost the same recording duration.

As explained in section IV, six features on the horizontal accelerometers are extracted firstly. We extract the six features in a period of 0.1 s after 10 s interval. Second, the MDP process is built. Three change points for four health stage division are regarded as an agent. State $\boldsymbol{S}$ contains different combination of three change points. Each change points can move to left, right and not move. Action $\boldsymbol{A}$ contains 27 actions for three different actions of three different change points. The Initial state s_0 is (1.66 h, 3.31 h, 5.24 h). Finally, the proposed RLCP is adopted for health stage division. 500 episodes are ran in this experiment.

After episode by episode, the results of health stages division by the proposed RLCP of the *bearing1_5* and *bearing1_6* dataset are obtained in Fig. 2. In order to represent the results intuitively, the original vibration signals of two bearing dataset are set as the vertical axis. The three red dotted lines represent three change points to divided the whole degradation into four health stages.



(a) The result of health stage division by RLCP in the *bearing1_5* dataset

(b) The result of health stage division by RLCP in the *bearing1_6* dataset

Fig. 2: The result of health stage division by the proposed RLCP

Fig. 2(a) shows the optimal change points are (1.69 h, 4.00 h, 6.00 h) in *bearing1_5* dataset. Fig. 2(b) shows the optimal change points are (2.36 h, 4.09 h, 6.14 h) in *bearing1_6* dataset. The initial stage is the health stage. The second stage is the sbu-health stage, the third stage is the critical stage, and the final stage is the fault stage. The results demonstrate the proposed RLCP can divide the whole degradation based on change points when considering multiple HIs.

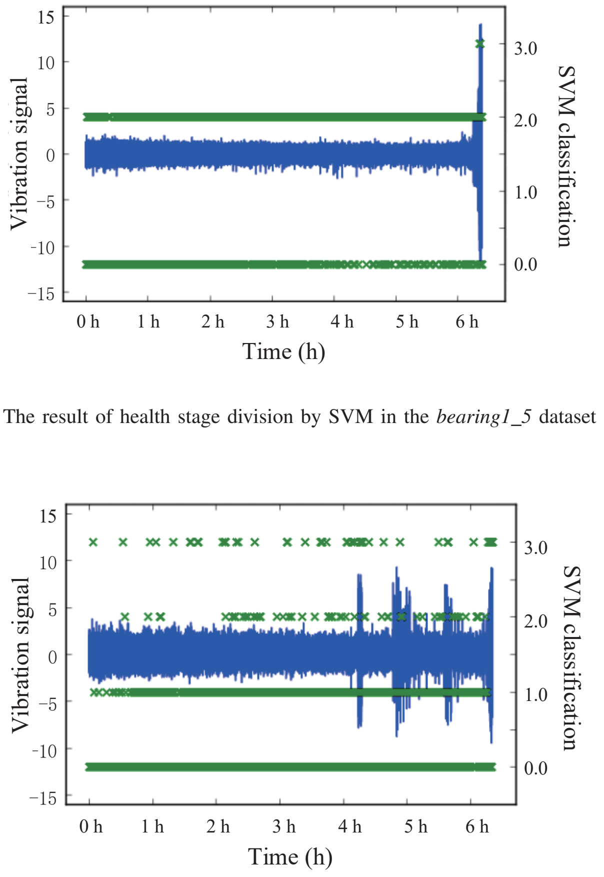
SVM [9] and k-means [10] are adopted to divide health stage for comparison. After the six feature extraction, the whole degradation of *Bearing1_6* is used to train the SVM classifier. Then the SVM classifier is used to predict the degradation of *Bearing1_5*. Same as the health stage division in *Bearing1_6*. The results of SVM are shown in Fig. 3. K-means is adopted to cluster the degradation of *Bearing1_5* and *Bearing1_6*. After clustering, the smoothing algorithm is employed to re-cluster the boundary points. The results of k-means are shown in Fig. 4.

As shown in Fig. 3, 4, SVM and k-means are used to divide the whole health state into discrete cluster. The green x represents the cluster results of SVM. The black point represents the cluster results of Kmeans. The whole health state are divided into 4 clusters, which are (0, 1, 2, 3) on the right vertical axis. Because the two methods did not divide the whole degradation based on change points, the division results are not obtained in the continuous time series. The comparison results demonstrate the importance of considering multiple HIs and change points simultaneously.

VI. Conclusion

To overcome the difficulty in the fusion of multiple HIs, the RLCP is proposed for health stage division. The proposed RLCP is divided into three steps. First, six classical time domain features are extracted as multiple HIs. Second, the health stage division problem based on change points is modeled as an MDP. Third, reinforcement learning is adopted to solve the MDP eventually. The experiment demonstrates that the proposed RLCP finds the optimal change points compared with the classification algorithm called SVM and the clustering algorithm called k-means.

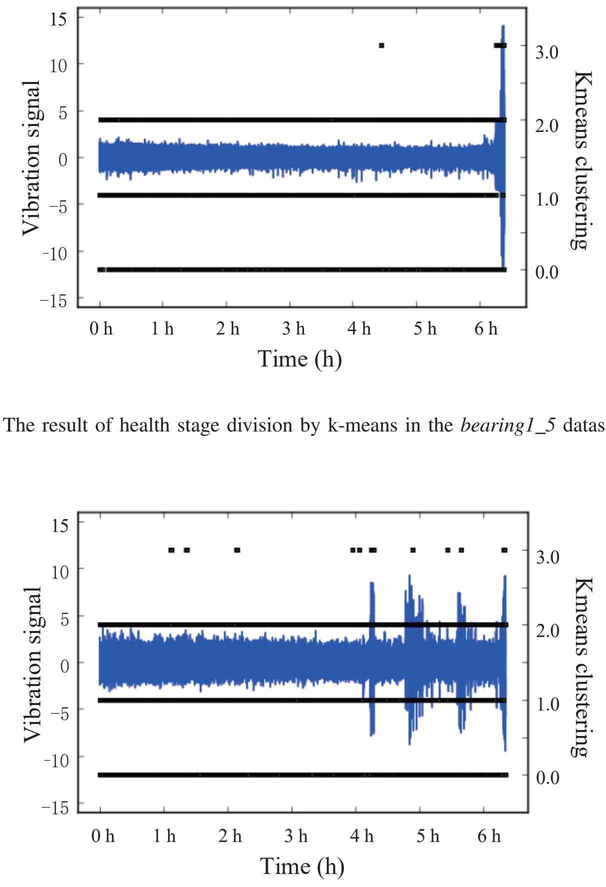
After the health stage division, the future work would focus on the RUL prediction.



(a) The result of health stage division by SVM in the *bearing1_5* dataset

(b) The result of health stage division by SVM in the *bearing1_6* dataset

Fig. 3: The result of health stage division by SVM



(a) The result of health stage division by k-means in the *bearing1_5* dataset

(b) The result of health stage division by k-means in the *bearing1_6* dataset

Fig. 4: The result of health stage division by k-means

Acknowledgment

The authors would like to thank the organizing committee of IEEE PHM 2012 prognostic challenge for sharing the degradation data of rolling element bearings. And this work was supported by National Natural Science Foundation of China (61672537, 61672539, 61772558, 61502055).

References

[1] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systemsreviews, methodology and applications," *Mechanical systems and signal processing*, vol. 42, no. 1, pp. 314–334, Jan, 2014.

[2] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to rul prediction," *Mechanical Systems and Signal Processing*, vol. 104, pp. 799–834, May, 2018.

[3] K. Liu, N. Gebraeel, and J. Shi, "A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis," *IEEE Transactions on Automation Science and Engineering*, vol. 10, no. 3, pp. 652–664, Apr, 2013.

[4] E. Bechhoefer and D. He, "A process for data driven prognostics," *IEEE International Conference on Prognostics and Health Management*, Dec, 2012, pp. 24–26.

[5] K. Liu, N. Z. Gebraeel, and J. Shi, "A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis," *IEEE Transactions on Automation Science and Engineering*, vol. 10, no. 3, pp. 652–664, Apr, 2013.

[6] S. Hong, Z. Zhou, E. Zio, and W. Wang, "An adaptive method for health trend prediction of rotating bearings," *Digital Signal Processing*, vol. 35, pp. 117–123, Dec, 2014.

[7] Y. Hu, H. Li, X. Liao, E. Song, H. Liu, and Z. Chen, "A probability evaluation method of early deterioration condition for the critical components of wind turbine generator systems," *Mechanical Systems and Signal Processing*, vol. 76, pp. 729–741, Aug, 2016.

[8] S. J. Bae, T. Yuan, S. Ning, and W. Kuo, "A bayesian ap-

proach to modeling two-phase degradation using change-point regression," *Reliability Engineering & System Safety*, vol. 134, pp. 66–74, Feb, 2015.

[9] A. Soualhi, K. Medjaher, and N. Zerhouni, "Bearing health monitoring based on hilberthuang transform, support vector machine, and regression," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 1, pp. 52–62, Jul, 2015.

[10] P. Scanlon, D. F. Kavanagh, and F. M. Boland, "Residual life prediction of rotating machines using acoustic noise signals," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 1, pp. 95–108, Nov, 2013.

[11] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[12] B. Abdulhai and L. Kattan, "Reinforcement learning: Introduction to theory and potential for transport applications," *Canadian Journal of Civil Engineering*, vol. 30, no. 6, pp. 981–991, 2003.

[13] O. Abul, F. Polat, and R. Alhajj, "Multiagent reinforcement learning using function approximation," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 30, no. 4, pp. 485–497, Nov, 2000.

[14] B. Abdulhai, R. Pringle, and G. J. Karakoulas, "Reinforcement learning for true adaptive traffic signal control," *Journal of Transportation Engineering*, vol. 129, no. 3, pp. 278–285, May, 2003.

[15] A. Soualhi, H. Razik, G. Clerc, and D. D. Doan, "Prognosis of bearing failures using hidden markov models and the adaptive neuro-fuzzy inference system," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 6, pp. 2864–2874, Jul, 2014.

[16] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, Sep, 1974.

[17] R. Bellman, "The theory of dynamic programming," *Bulletin of the American Mathematical Society*, vol. 60, no. 6, pp. 503–515, 1954.

[18] P. Nectoux, R. Gouriveau, K. Medjaher, E. Ramasso, B. Chebel-Morello, N. Zerhouni, and C. Varnier, "Pronostia: An experimental platform for bearings accelerated degradation tests." *IEEE International Conference on Prognostics and Health Management*, Denver, CO, USA, Dec, 2012.