Health Assessment for Piston Pump Based on Laplacian Eigenmaps-Random Forests Method

Haoren Wang¹, Yixiang Huang¹, Liang Gong¹, Pengfei Ye¹, Chengliang Liu¹ and ZhiYu Tao¹

¹State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University,
Shanghai, China

Abstract

A piston pump is one of the key components in the hydraulic system. Therefore, the failure of the pump may severely hurt the reliability of the hydraulic system and cause great loss. Currently, how to effectively evaluate the health condition of the piston pump remains an open problem. In this paper, a novel health assessment method based on an integration of Laplacian eigenmaps and random forests is proposed, which takes full advantage of both the fusion ability of correlated features enabled by the Laplacian eigenmaps and the optimized feature-selection ability provided by the random forest. The proposed method (LE-RF) is applied to piston pumps for validation. The results indicate that the Laplacian eigenmaps-random forest (LE-RF) technique can provide an effective tool for piston pump health condition assessment. Compare with other manifold methods, as well as some popular deep learning method, such as LLE, ISOMAP and deep DBN, the LE-RF method can achieve better and more accurate assessment results.

Keywords: Piston pumps, Feature extraction, Pattern recognition, Laplacian eigenmaps, Random forests.

1. Introduction

Piston pumps are widely used in various industrial applications, and unexpected component degradation in a hydraulic system may cause substantial losses in productivity. The degradation of a piston pump can be divided into several stages. In the first stage, shock and vibration of piston pump are within a controllable level. With increased abrasion, the pressure loss in the hydraulic system increases, frequencies of the vibration change and the hydraulic system works less smoothly. Finally, pressure drops below a certain threshold, and failure of hydraulic pump may occur. To avoid a possible safety injury and substantial productivity losses, certain preventive

health maintenance paradigms are currently applied, such as scheduled maintenance, which may operate in a generally inefficient and costly way. Thus, condition-based maintenance (CBM) [1,2], is considered to be one of the best solutions. The health condition of a piston pump is evaluated by monitoring its vibration signals, CBM improves the accuracy of estimating the remaining useful life (RUL) of the components, preventing unexpected breakdown and severe damage in productivity. Therefore, health assessment of hydraulic pump has been widely researched. For instance, Gao et al. [3,4], proposed a diagnosis method based on a discharge pressure pulsation model, the wavelet analysis was performed to separate the pressure signals when a fault occurs. Kong and Chen [5], built a combined method based on wavelet transformation, fuzzy logic and neural networks for fault diagnosis of a triplex pump. The test indicates failure characteristics of the fluid-end and dynamic-end are coincident with different scales of wavelet transformation, and the method has been proved by experimental results. Dong and He [6] presented a statistic modeling methodology of both diagnosis and prognosis in a unified framework, which is based on segmental hidden semi-Markov model (HSM). The model estimated parameters through the modified forward-backward algorithm, different abrasion conditions of the pump were successfully identified, and the remaining-useful-life of the pump is effectively predicted as well. Chen et al. [7], presented a Kolmogorov-Smirnov test method to analyze the impulsive energy of the vibration signal, and the fault degradation of the piston in the water hydraulic motor was detected. Zhao et al. [8], proposed a method on early fault diagnosis of hydraulic pump. Intermittent chaos and a symbol sequence statistic-based method were successfully applied for the early fault diagnosis on the hydraulic tube tester. Jun et al. [9], proposed a method based on the layered clustering algorithm, multiple faults occur simultaneously in an aircraft hydraulic pump can be detected by the proposed method.

The methods listed above can correctly diagnose faults and evaluate the degradation status under certain situations by a small dataset. However, various features are extracted from the original signals in actual applications. The number of features could be too large for the embedded computer to process. Thus, how to efficiently capture the sensitive features in the large feature set remains as an open research topic. Data reduction and fusion techniques are combined to eliminate noises and other irrelevant information. In the early 2000s, the manifold learning methods have emerged as a new dimension reduction technique, which can analyze the high-dimensional

geometric features of the nonlinear data, consequently gaining key information from a complex high-dimensional space. Since the first publication of manifold learning related algorithm in science, several typical algorithms have been developed, e.g., local linear embedding algorithm (LLE) [10], Laplacian eigenmaps (LE) [11], isometric mapping (ISOMAP) [12], locality preserving projections algorithm (LPP) [13]. Wherein the manifold learning is used for health assessment, Jiang et al. [14], proposed supervised Laplace mapping algorithm and successfully applied for gearbox fault diagnosis. Yu et al. [15], put forward the LPP improved algorithm and used it at fault detection and bearing health assessment. Li et al. [16], proposed a local tangent space alignment algorithm (LTSA) in a rotating machine, and successfully used it in bearing health condition assessments. Although the manifold learning method has been proposed in some common mechanical parts, the application for piston pumps and other hydraulic key components have not been demonstrated.

The manifold learning techniques proved to be effective for dimensional reduction, however, after dimensional reduction processing, not all features are effective for health states evaluation, so the evaluators need to be further refined to contain the degradation information, and it is necessary to find proper feature searching and selection methods. As a multiple classifier ensemble method [17], the random forests algorithm, proposed by Breiman [18], and applied in many fields [19], especially in machinery faults diagnosis [20-22], is taken into consideration in this study. Random forests can build a large number of decision trees [23]. Bagging [24], and boosting [25], are added to a unique depending subspace of origin datasets for training. Both variance and bias of classification have been considered to improve its generalization properties.

Considering the above reasons, the paper combined the manifold learning and random forest techniques and raised Laplacian eigenmaps-random forests (LE-RF) for the health assessment of piston pump. The proposed fusion method aims to find the intrinsic features. The method can better identify the piston pump health conditions and reduce the irrelevant information.

This paper is organized as follows. In Section II, the feature extraction method applied in this study is introduced. In Section III, the principal of the proposed LE-RF method is described theoretically. In Section IV, the experiment setup for the piston pump is given. In Section V, the

paper evaluates the performance of the LE-RF method for health assessment and compare results with other techniques. Finally, the conclusions are reached, and discussion regarding the future study is shown in Section VI.

2. Feature Extraction Method

2.1 Original high-dimensional features formation

The extraction of features is the key for piston pumps health status evaluation. However, the vibration signals contain noise and redundant information. By means of feature extraction, the raw signals are converted to high-dimensional feature space, and the characteristics of the vibration signals can be subsequently analyzed to evaluate the health status.

The wavelet packet decomposition [26,27], with a pair of high-pass filters and low pass filters is capable of extracting the high and the low frequency components of signals. Compared to the wavelet package decomposition, wavelet decomposition [28], is insufficient to make up for the high-frequency signal resolution. For an actual piston pump, the vibration signals have properties of the non-stationary characteristics. Therefore, the wavelet packet decomposition is better able to extract the characteristics of the signal, and obtain potential information, for analyzing all band of signal components.

Suppose that $\{\mu_n(x)|n \in Z_+\}$ is the orthogonal wavelet package with respect to the filter $\{h_n\}$, the coefficients in the subspace are as follows:

$$c_k^{j,n} = \int_{-\infty}^{\infty} f(x) 2^{\frac{j}{2}} \mu_n(2^j x - k) dx, \quad (1)$$

where j is a scale index, and l is the translation index. The coefficients U_{j-1}^{2n} and U_{j-1}^{2n+1} at coarser scales are

$$c_p^{j-1,2n} = \sum_l h(l-2p) c_l^{j,n} \quad (2)$$

$$c_p^{j-1,2n+1} = \sum_l g(l-2p) c_l^{j,n}, \quad (3)$$

where $\{h_n\}$ is low-pass filter and $\{g_n\}$ is high-pass filter. U_j^n is divided into the low frequency part U_{j-1}^{2n} and high frequency part U_{j-1}^{2n+1} , the decomposition standardized coefficients of each layer can be obtained.

Since the wavelet package decomposition forms the orthogonal bases, the decompositions will preserve energy. It shows that:

$$\sum_l (c_l^{j,n})^2 = \sum_p (c_p^{j-1,2n+1})^2 + \sum_p (c_p^{j-1,2n})^2. \quad (4)$$

After the raw signals are projected to the wavelet domain by the wavelet packet transform, the energy of the wavelet coefficients measures effects in each frequency span, therefore, if an energy measure can be defined as E_k^{j-1} , and the energy of decomposed coefficients is expressed as follows:

$$E_k^{j-1} = \sum_p (c_p^{j-1,k})^2, \quad (5)$$

where $c_p^{j-1,k}$ ($k \in Z_+$) stands for $c_p^{j-1,2n}$ or $c_p^{j-1,2n+1}$.

The usage of 8-level of Daubechies 6 wavelet decomposition gives full-band division. Figure 1 shows the vibration signals recorded by Channel 1 and all the high-frequency component of each level. After the wavelet package decomposition, totally 256 divided features are represented as energy measurements of level 8, and the energy of each sub-band is calculated by Equation (5). According to the characteristics of vibration signals, the six parameters (the absolute maximum, absolute minimum, absolute mean, variance, kurtosis and skewness) are selected as time-domain features. Both time and frequency related features have significant influence for the evaluation of the health state of the running status. Table 1 lists the 262 time and frequency related features form high dimensional space $X \in R^{N \times m}$, where N is the total number of samples and m ($m = 262$) is the number of the original features.

Figure 1. The original signals and decomposition components of channel 1

3. The LE-RF Algorithm

The flowchart of the health assessment process is shown in Figure 2. Extracting features of the original time-domain, frequency domain and time-frequency domain signals with mixed parameters have been used in the first stage. Since the nature of different mechanical components is corresponding to the acquired signals varies, the selection of a higher resolution and regularity of characteristic parameters is key to condition assessment. Secondly, the LE-RF method is selected as the data fusion method to refine the features. Thirdly, the confusion matrix and the recognition accuracy are used to evaluate the effects of health state assessment, the test set is also added to verify the recognition effects for other pumps in the same batch.

Figure 3 shows the architecture of the LE-RF algorithm. Approximated by the adjacency graph [29], the weights of the adjacency graph are chosen properly by the Laplacian Beltrami operator[30]. The LE method indicates the intrinsic relationship of each cluster and divides the original data space into separated structures by correlated features. The feature vector $\mathbf{x}$ ($\mathbf{x} \in R^m$) turns into $\mathbf{x}_1$ ($\mathbf{x}_1 \in R^{m'}$) by the LE method, and the local information of embedding [11]

preserves. Next, the random forests algorithm[18], creates a certain number of trees, and they vote for the most possible class, which reduces the number of different classes and benefits features reduction.

Figure 2. The flowchart of the health assessment for piston pump

Figure 3. The architecture of the LE-RF algorithm

3.1 The Laplacian eigenmaps

The Laplacian eigenmaps (LE) algorithm [11,31], is a geometry-based method to properly represent complicated data on high dimensional space. The algorithm focuses on the relationship

between the Laplace Beltrami operator, heat equation and graph Laplacian matrix, which provides computational efficiency benefits on non-linear data, and has local preserving properties and clusters naturally. The aim of the algorithm is to maintain the data points on average.

The paper defines that k points $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ in $\mathcal{R}^l$, and a weighted graph is set up with k nodes where each point connects adjacency points with a set of edges. An embedding map is established by computing the eigenvector of the graph Laplacian matrix. The steps of Laplacian eigenmaps are as follows:

Step 1 (Constructing an adjacency graph matrix): the k -nearest neighbor method [32,33], is used on the original dataset, and an edge between $\mathbf{x}_i$ and $\mathbf{x}_j$ is created in case i is among the n nearest neighbors of j .

Step 2 (Choose the weight of edges): There are two different approaches for weighting the edges: The heat kernel approach [11]: if node i is connected to j , W_{ij} is set as follows:

$$W_{ij} = e^{-(\|\mathbf{x}_i - \mathbf{x}_j\|^2)/\sigma^2}, \quad (7)$$

The simple approach: if the nodes i and j are connected by an edge, $W_{ij} = 1$; if the nodes i and j are not connected, $W_{ij} = 0$.

Step 3 (Set up the object function): to connect data points as close as possible, the weighed graph $\mathbf{G}$ is embedded to a lower dimensional space. $\mathbf{y} = (y_1, y_2, \dots, y_n)$ is defined to be such a map. A reasonable measurement method is used to choose an appropriate map, and the goal of optimization is to minimize the following objective function:

$$\sum_{ij} (y_i - y_j)^2 W_{ij}, \quad (8)$$

To avoid neighboring points, $\mathbf{x}_i$ and $\mathbf{x}_j$ are mapped far away from each other and the mechanism for minimizing objective function is introduced. The Laplacian matrix is postulated as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, $\mathbf{D}$ is a diagonal weight matrix, and the elements of matrix $\mathbf{D}$ are sums of $\mathbf{W}$, i.e., $D_{ii} = \sum_j W_{ij}$. It turns out that for any $\mathbf{y}$, the result can be got as follows:

$$\frac{1}{2} \sum_{ij} (y_i - y_j)^2 W_{ij} = \text{tr}(\mathbf{y}^T \mathbf{L} \mathbf{y}). \quad (9)$$

Therefore, the minimization problem is now simplified to find $\text{argmin}(\mathbf{y}^T \mathbf{L} \mathbf{y})$, the constraints are

$$\mathbf{y}^T \mathbf{D} \mathbf{y} = 1, \quad (10)$$

$$\mathbf{y}^T \mathbf{D} \mathbf{1} = 0. \quad (11)$$

The constraint $\mathbf{y}^T \mathbf{D} \mathbf{y} = 1$ avoids the arbitrary scaling factor; Matrix $\mathbf{D}$ is a measurement on vertices of the graph. The larger D_{ii} is, the more important y_i will be. Suppose $\mathbf{1}$ to be the constant function, once all vertices of $\mathbf{G}$ onto the real number 1, trivial solution occurs, the constraint $\mathbf{y}^T \mathbf{D} \mathbf{1} = 0$ is added to eliminate this situation.

Step 4: (Eigenmaps): compute the eigenvalues and eigenvectors for the generalized eigenvalue problem

$$\mathbf{L} \mathbf{f} = \lambda \mathbf{D} \mathbf{f} \quad (12)$$

$\mathbf{L} = \mathbf{D} - \mathbf{W}$ is a Laplacian matrix, which is always a symmetric, positive semidefinite matrix, is regarded as the operation function on vertices of graph $\mathbf{G}$.

The eigenvalues can be obtained through the eigenvalue decomposition algorithm, and 2 to $d+1$ items of the eigenvalues are corresponding to d -dimensional output the feature vector $\mathbf{y}$. The LE algorithm turns the problem into a calculation of eigenvalues and does not need to calculate iteration, so it requires little computational time.

3.2 Eigenvectors normalization

The resulting wavelet coefficients energy is normalized to limit the value of each data point within a certain range and avoid significant effects of extreme value. It is convenient for continuous processing. The formula z-score normalization method is:

$$e' = \frac{e - \mu_{train}}{\sigma_{train}}, \quad (6)$$

Where μ_{train} is the mean of a training set and σ_{train} is standard deviation from the same training dataset.

Table 1. Time domain and frequency related features

Features	Definition of Features	Dimensions
Absolute maximum	$x_{\max} = \max x_i $	1
Absolute	$x_{\min} = \min x_i $	1

minimum		
Absolute mean	$ \bar{x} = \sum_{i=1}^N x_i / N$	1
Variance	$D_x = \sum_{i=1}^N (x_i - \bar{x})^2 / (N - 1)$	1
Kurtosis	$x_k = \sum_{i=1}^N (x_i - \bar{x})^4 / N\sigma^4$	1
Skewness	$x_s = \sum_{i=1}^N (x_i - \bar{x})^3 / N\sigma^3$	1
Decomposition	All nodes energy of the eighth level	256
coefficient energy	WPD	

3.3 Random forests method

The random forests method [18], is a multiple classifier assembled by many tree-type classifiers. Once the paper imports a new input vector into the random subspace, the vector is examined by every tree in the forest, and they vote on the input one via classification. The forests select the most votes over all trees, and the most popular class is selected. The procedure is known as random forests.

While data in training dataset, the current trees are replaced. However, roughly one-third of the test samples are left out by the algorithm. These left-out samples are called out-of-bag samples, which are used for a test dataset to avoid unbiased estimate classification errors [19]. Due to the existence of these unbiased samples, there is no need for cross-validation or searching another test dataset to obtain an unbiased estimate. Both low bias and low correlation are necessary for accuracy. The low bias is determined by the out-of-bag estimate, and randomization is applied to achieve the aim of low correlation.

The steps for tree growing are as follows:

We randomly select N training bootstrap samples from the original dataset, and each tree is grown on them. About $(2/3)/N$ of the original data is used for training the tree, and the rest $(1/3)/N$ is used to validate the classifiers. At each node, the paper input M variables, then randomly choose m predictors from them ($m \ll M$). While the tree grows, the value of m remains steady. The initial reference value of m is $m = \lfloor \log_2(M) + 1 \rfloor$ or $m = \sqrt{M}$, and the optimized m can be found

by adjusting m around the initial value until the out-of-bag estimate achieves minimum error. These m predictors are used to split the node; only one out of m predictor is selected for the best split results in each node.

All trees can be grown as large as they can extend, there is no limit for their growth. In the current study, the importance of the variable estimated by classification accuracy has become a trend in many studies [34]. The algorithm computes the classification accuracy of out-of-bag data and the mean descent in accuracy. Bootstrap samples are $b = 1, \dots, B$, $\bar{D}_j$ stands for the importance measure of variable x_j . The calculation algorithm is as follows:

Step 1: Select the initial value of b : $b = 1$, then choose the out-of-bag data points $\mathcal{L}_b^{oob}$.

Step 2: Use tree T_b to classify $\mathcal{L}_b^{oob}$, then calculate the number of correctly classified items, R_b^{oob} .

Step 3: In order to calculate the importance measure of variables $x_j, j = 1, \dots, N$: the original values of x_j is replaced in $\mathcal{L}_b^{oob}$, and add the replaced results into $\mathcal{L}_b^{oob}$; then use tree T_b to classify $\mathcal{L}_b^{oob}$ and calculate the number of correctly classified items, $R_{b,j}^{oob}$.

Step 4: Repeat the algorithm starting at Step 1 and make the iterations while $b = 2, \dots, B$.

Step 5: For x_j , the importance measure $\bar{D}_j$ can be calculate as follows:

$$\bar{D}_j = \frac{1}{B} \sum_{b=1}^B (R_b^{oob} - R_{b,j}^{oob}). \quad (13)$$

Step 6: The importance measure $\bar{D}_j$ is normalized by the algorithm of z-score, and s_j is the descending standard deviation of correct classifications:

$$z_j = \bar{D}_j / (s_j / \sqrt{B}). \quad (14)$$

Step 7: the statistical distribution is assumed as a Gaussian distribution, and z_j is converted to a meaningful value.

4. Experimental Results and discussion

4.1 Experimental System

The test bed is shown in Figure 4. The type of the piston pumps is Kawasaki K3V. Five piston pumps of typical performances of different health states were carefully selected. These pumps used for this test were previously working in the same condition, but with different total running

time and corresponding degradation degrees. The status of their performances was preliminary labelled by the human expert from hearing the operation sounds and feeling the vibration degrees from touching the vibration of the pumps. Thus, the main purpose is to find an automated way to map the sensor signals to the labels of different performance of the pumps. The five piston pumps show differences in many aspects, such as pressure, efficiency and so on. Table 2 shows running time of actual piston pumps and their abbreviated labels. The table was designed according to actual pump maintenance plan in the factory.

The speed of piston pumps was set at 2200 r/min and the test lasts approximately 800 s of the stable working condition. Three vibration sensors were installed at the outlet of piston pump, on the upper-left, the upper-right and the bottom. The angle between any pair of two sensors was 120 degrees. The sampling frequency of the sensor was 50 kHz, and original signals were shown in Figure 5 by three channels in the first 0.08 s.

Figure 4. The testbed setup

Table 2. The condition of different pumps

Abbr. Codes	Runtime	Actual Runtime
P1	0 ~ 500h	400h
P2	500 ~ 1000 h	800h
P3	1000 ~ 2000 h	1400h
P4	2000 ~ 3000 h	2200h
P5	More than 3000 h	4300h

Figure 5. The original signal of the piston pump from three channels

4.2 Dimensional reduction results

The total number of samples is 500, and the number of each condition is 100. We selected 100 pieces of the vibration signal from each piston pump, when the piston pump run steadily. We built a 262-dimension matrix $X^{500 \times 262}$ for the sample points. LE, LPP, LLE, ISOMAP, PCA and KPCA are used for the dimensionality reduction process, and the methods transform high-dimensional space into low-dimensional space for feature extraction and pattern recognition analysis. We set neighboring factor $k = 5$, and used a maximum likelihood estimator method [35] to calculate operator embedding dimension, and the embedding dimension d' is estimated as 23.

(a) LE

(b) LPP

(c) LLE

(d) ISOMAP

(e) PCA

(f) KPCA

Figure 6. Different manifold method for dimensional reduction

As shown in Figure 6(a), the states are divided into five categories, and within-class distance is small, the clustering is obvious, and it is easy to distinguish with different class. P4 and P5 have slight overlap, but there is no impact on the health status evaluation. P1, P2 and P3 have a large among-class distance and a relatively small within-class distance; their health statuses are clearly different. In Figure 6(b), LPP gathers the samples into five groups, but P1 and P2 merge together, and the boundary between P4 and P5 is not clear. In Figure 6(c), LLE divides the samples into five groups, overlaps between P3 and P5 are obvious and the boundary of P1 and P2 is blurred. In Figure 6(d), although the boundary between P1 and P2 is clear, P3, P4 and P5 have merged. The distances between the feature vectors of the different classes are also larger than those calculated by LE method.

In Figure 6(e), PCA fails to differentiate P3, P4 and P5, and P1 and P2 are overlapped. In Figure 6(f), the usage of the kernel function results in a large within-class distance, P3, P4 and P5 are collinear, and P1 and other clusters are surrounded by P2, it appears that the clustering effects are not as good as in the LE method.

To evaluate the classification accuracy, we use the random forests classification method to validate the model classification accuracy. We set the number of trees to $k = 800$ and 70% of the data (about 350 samples) was used for training and the rest (about 150 samples) for validation. The training set is used to train the classification model, and the validation set is used to verify the accuracy of the trained model. The dimension reduction method (LE, LPP, LLE, ISOMAP, PCA, and KPCA) is applied before classification method, then the accuracy is obtained through random forests and calculated accuracy for 10 times, and Figure 7 shows the detail of recognition results with different manifold methods. After processing, numerical average is calculated. Table 3 shows numerical average of different manifold method. The LE method has higher classification accuracy than the other methods, indicating that the method using the LE to evaluate the piston pumps operating state has the highest discrimination.

Figure 7. Reduced data recognition results of the different manifold methods

Table 3. The results of the different manifold methods by average

	LE	LPP	LLE	ISOMAP	PCA	KPCA
Averages	98.2432	95.1097	95.1097	92.5976	94.5270	61.4865

Figure 8. Classification accuracy of different embedding dimensions

Considering the influence of different embedding dimensions, we calculate the classification accuracy using different embedding dimensions. The LE method is selected as the dimensional reduction method with different embedding dimension, and accuracy is calculated by random forests with the same configuration as above. The results shown in figure 8 suggest the relationship between classification accuracy and the embedding dimension. When the embedding dimension d' belongs to the integer set (23,262), the feature dimension is relatively high, and the accuracy become higher with a decrease in the embedding dimension. When the embedding dimension d' is

23, the accuracy achieves its highest value; when the embedding dimension d' belongs to the integer set $[1,23]$, the accuracy descends with the decrease of the embedding dimension. The accuracy results are strong proofs to demonstrate the embedding dimension is better, only embedding dimensionality is 23, we can obtain the maximum classification accuracy.

4.3 The Result

As displayed in Table 4, there are 148 total samples for the random forests test set, and the number of five classes is 30, 35, 27, 21 and 35, respectively. Furthermore, the random forests method has correctly predicted instances 30, 35, 27, 21 and 33 of P1, P2, P3, P4 and P5, respectively. While we totally have 148 samples for KNN test set, and the number of five classes is 31, 30, 31, 29 and 27, respectively. From Table 5, it can be observed that KNN can correctly predict 31, 29, 30, 27 and 26 instances of P1, P2, P3, P4 and P5, respectively. Then, we have 148 samples for DBN test set. The number of five classes is 34, 30, 29, 28 and 27, respectively. From Table 6, it can be observed that KNN can correctly predict 22, 10, 26, 23 and 20 instances of P1, P2, P3, P4 and P5, respectively. The comparison of the LE-RF, LE-KNN and LE-DBN is shown in Table 7.

Table 4. Confusion matrix for LE-RF method

	P1	P2	P3	P4	P5
P1	30	0	0	0	0
P2	0	35	0	0	0
P3	0	0	27	0	0
P4	0	0	0	21	0
P5	0	0	0	2	33

Table 5. Confusion matrix for LE-KNN method

	P1	P2	P3	P4	P5
P1	31	0	0	0	0
P2	1	29	0	0	0
P3	0	0	30	0	1
P4	0	0	0	27	2

P5	0	0	0	1	26
----	---	---	---	---	----

Table 6. Confusion matrix for deep LE-DBN-BP method

	P1	P2	P3	P4	P5
P1	22	11	1	0	0
P2	20	10	0	0	0
P3	1	0	26	1	1
P4	0	0	1	23	4
P5	0	0	0	7	20

Table 7. Comparison between random forests and KNN

	LE-RF	LE-KNN	Deep LE-DBN
Correctly classified instances	146(98.6486%)	143(96.6216%)	101(68.2432%)
Incorrectly classified instances	2(1.3513%)	5(3.3784%)	47(31.7568%)
Total number of instances	148	148	148

Figure 9. Out-of-bag classification error

The out-of-bag classification error is shown in Figure 9, the selected samples are the results of the LE dimensional reduction method. It can be observed that the error of classification remains steady and a relatively low level when the number of trees is larger than the threshold number.

4.4 The test set validation

The test pump (Pump 6) is added as the test set for the system. Pump 6 is also in the batch above,

the only difference between Pump 6 and other pumps is running time. The running time of P6 is about 300h, and within the range of P1.

The signals from P6 are processed by feature extraction and dimension reduction. The processed data is $T^{100 \times 23}$, and classified by the trained model of random forest. A vector of predicted labels T_s is returned after random forest method. We calculated the frequency of each label, and the highest frequency of occurrence label is returned as label L_{si} ($L_{si} \in [1,2,3,4,5]$). The states of P6 can be determined by L_s . In this time, the P6 is predicted as a certain condition C_{si} .

However, considering the initialization randomness of data, the predicted label L_s may be influenced. For the data from P6, 100 groups of processed data ($T^{100 \times 23}$) are collected. The prediction process has been recycled for 100 times. The predicted labels L_{si} are combined into the set L_s . The probability of correct prediction (P1) in set L_s is 77%, and the conclusion has been reached that P6 is within the range of 0~500 h (the first stage of degradation).

5. Conclusions

A method based on LE-RF forests is proposed to assess piston pump health. The results show that compared with different manifold learning methods, the LE-RF method is proposed with better clustering and fewer classification errors, which successfully reduces the high-dimensional features into a minimal one and improves accuracy of health state recognition. The method has also verified by test set, and successfully used in degradation assessment. The proposed method plays an important role for piston pump health condition evaluation, which is helpful for predicting potential risks in factory manufacturing.

In the future research, our effort will be devoted to make the proposed method more reliable for practical use. To achieve this, a knowledge database corresponding to different health conditions of mechanical components is considered. Moreover, the database is modified if the predicted health states are inconsistent with results of the actual health states, and our health assessment algorithm can be applied to more mechanical equipment with better accuracy.

Acknowledgements

This research is supported by NSFC Project (Grant No. 51975356). The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of this manuscript.

References

- [1] Jardine, A.K.S.; Lin, D.; Banjevic, D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing* 2006, 20, 1483-1510.
- [2] Ebrahimkhanlou, A.; Salamone, S. Acoustic emission source localization in thin metallic plates: A single-sensor approach based on multimodal edge reflections. *Ultrasonics* 2017, 78, 134-145.
- [3] Gao, Y.; Zhang, Q.; Kong, X. Wavelet-based pressure analysis for hydraulic pump health diagnosis. *Transactions of the ASAE* 2003, 46, 969-976.
- [4] Zhang, Q.; Gao, Y. A wavelet packet and residual analysis based method for hydraulic pump health diagnosis. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* 2006, 220, 735-745.
- [5] Kong, F.; Chen, R. A combined method for triplex pump fault diagnosis based on wavelet transform, fuzzy logic and neuro-networks. *Mechanical Systems and Signal Processing* 2004, 18, 161-168.
- [6] Dong, M.; He, D. A segmental hidden semi-markov model (hsmm)-based diagnostics and prognostics framework and methodology. *Mechanical Systems and Signal Processing* 2007, 21, 2248-2266.
- [7] Chen, H.X.; Chua, P.S.K.; Lim, G.H. Fault degradation assessment of water hydraulic motor by impulse vibration signal with wavelet packet analysis and kolmogorov-smirnov test. *Mechanical Systems and Signal Processing* 2008, 22, 1670-1684.
- [8] Zhao, Z.; Jia, M.; Wang, F.; Wang, S. Intermittent chaos and sliding window symbol sequence statistics-based early fault diagnosis for hydraulic pump on hydraulic tube tester. *Mechanical Systems and Signal Processing* 2009, 23, 1573-1585.
- [9] Du, J.; Wang, S.; Zhang, H. Layered clustering multi-fault diagnosis for hydraulic piston pump. *Mechanical Systems and Signal Processing* 2013, 36, 487-504.
- [10] Roweis, S.T.; Saul, L.K. Nonlinear dimensionality reduction by locally linear embedding. *Science* 2000, 290, 2323-2326.

- [11] Belkin, M.; Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* 2003, 15, 1373-1396.
- [12] Tenenbaum, J.B.; De Silva, V.; Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science* 2000, 290, 2319-2323.
- [13] He, X.; Niyogi, P. In *Locality preserving projections*, Advances in Neural Information Processing Systems, 2004.
- [14] Jiang, Q.; Jia, M.; Hu, J.; Xu, F. Machinery fault diagnosis using supervised manifold learning. *Mechanical Systems and Signal Processing* 2009, 23, 2301-2311.
- [15] Yu, J. Local and nonlocal preserving projection for bearing defect classification and performance assessment. *Industrial Electronics, IEEE Transactions on* 2012, 59, 2363-2376.
- [16] Li, F.; Tang, B.; Yang, R. Rotating machine fault diagnosis using dimension reduction with linear local tangent space alignment. *Measurement* 2013, 46, 2525-2539.
- [17] Dietterich, T. Ensemble learning. The handbook of brain theory and neural networks, ma arbib. Cambridge, MA: MIT Press: 2002.
- [18] Breiman, L. Random forests. *Machine learning* 2001, 45, 5-32.
- [19] Verikas, A.; Gelzinis, A.; Bacauskiene, M. Mining data with random forests: A survey and results of new tests. *Pattern Recognition* 2011, 44, 330-349.
- [20] Yang, B.S.; Di, X.; Han, T. Random forests classifier for machine fault diagnosis. *Journal of Mechanical Science and Technology* 2008, 22, 1716-1725.
- [21] Li, C.; Sanchez, R.V.; Zurita, G.; Cerrada, M.; Cabrera, D.; Vásquez, R.E. Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals. *Mechanical Systems and Signal Processing* 2016, 76-77, 283-293.
- [22] Cerrada, M.; Zurita, G.; Cabrera, D.; Sánchez, R.V.; Artés, M.; Li, C. Fault diagnosis in spur gears based on genetic algorithm and random forest. *Mechanical Systems and Signal Processing* 2016, 70-71, 87-103.
- [23] Quinlan, J.R. Induction of decision trees. *Machine Learning* 1986, 1, 81-106.
- [24] Breiman, L. Bagging predictors. *Machine learning* 1996, 24, 123-140.
- [25] Bauer, E.; Kohavi, R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning* 1999, 36, 105-139.
- [26] Huang, Y.; Liu, C.; Zha, X.F.; Li, Y. An enhanced feature extraction model using lifting-based wavelet packet transform scheme and sampling-importance-resampling analysis. *Mechanical Systems and Signal Processing* 2009, 23, 2470-2487.

- [27] Liu, Q.; Chen, F.; Zhou, Z.; Wei, Q. Fault diagnosis of rolling bearing based on wavelet package transform and ensemble empirical mode decomposition. *Advances in Mechanical Engineering* 2013, 2013.
- [28] Wang, X.; Zi, Y.; He, Z. Multiwavelet construction via an adaptive symmetric lifting scheme and its applications for rotating machinery fault diagnosis. *Measurement Science and Technology* 2009, 20.
- [29] Merris, R. Laplacian matrices of graphs: A survey. *Linear Algebra and Its Applications* 1994, 197-198, 143-176.
- [30] Ivrii, V.Y. Second term of the spectral asymptotic expansion of the laplace-beltrami operator on manifolds with boundary. *Functional Analysis and its Applications* 1980, 14, 98-106.
- [31] Malik, Z.K.; Hussain, A.; Wu, J. An online generalized eigenvalue version of laplacian eigenmaps for visual big data. *Neurocomputing* 2016, 173, 127-136.
- [32] Fukunaga, K.; Narendra, P.M. A branch and bound algorithm for computing k-nearest neighbors. *IEEE Transactions on Computers* 1975, C-24, 750-753.
- [33] Lei, Y.; Zuo, M.J. Gear crack level identification based on weighted k nearest neighbor classification algorithm. *Mechanical Systems and Signal Processing* 2009, 23, 1535-1547.
- [34] Liaw, A.; Wiener, M. Classification and regression by randomforest. *R news* 2002, 2, 18-22.
- [35] Levina, E.; Bickel, P.J. In *Maximum likelihood estimation of intrinsic dimension*, Advances in Neural Information Processing Systems, 2005.