Research Article

Application of Data Denoising and Classification Algorithm Based on RPCA and Multigroup Random Walk Random Forest in Engineering

Renchao Wang, Yanlei Wang, and Yuming Ma

State Key Laboratory of Hydraulic Engineering Simulation and Safety, Tianjin University, Tianjin 300350, China

Correspondence should be addressed to Yanlei Wang; tjuwyl@tju.edu.cn

Received 27 June 2019; Revised 30 October 2019; Accepted 30 November 2019; Published 24 December 2019

1. Introduction

Data classification algorithms are often used in the engineering field, but the data measured in the actual engineering often contains different types and degrees of noise, such as vibration noise caused by water flow when measuring the natural frequencies of aqueducts or other hydraulic structures, which will affect the accuracy of classification. In reality, these noises often appear disorganized and stochastic and some existing algorithms exhibit poor performance in the face of these non-Gaussian noise. Therefore, the classification algorithms with excellent performance are needed. To address this issue, a hybrid algorithm of robust principal component analysis (RPCA) combined multigroup random walk random forest (MRWRF) is proposed in this paper. On the one hand RPCA can effectively remove part of non-Gaussian noise, and on the other hand MRWRF can select a better number of decision trees (DTs), which can effectively improve random forest (RF) robustness and classification performance, and the combination of RPCA and MRWRF can effectively classify data with non-Gaussian distribution noise. Compared with other existing algorithms, this hybrid algorithm has strong robustness and preferable classification performance and can thus provide a new approach for data classification problems in engineering.
identification of a concrete frame structure shows that the combination algorithm with kernel principal component analysis (KPCA) and the SVM has a certain antinoise ability [14]. A multigroup particle swarm optimization algorithm based on real coding is proposed and applied to the classification of the damage data of large structures under environmental excitation, and the results show that the algorithm has better noise immunity [15].

When solving various engineering problems with various algorithms, noise effects are unavoidable in engineering data acquisition systems. Thus, advanced noise reduction algorithms should be applied to minimize the impact of data corruption in problem solving.

Many data denoising algorithms have emerged in the early time and these include wavelet transform, principal component analysis (PCA), independent component analysis (ICA), the adaptive filtering, neural networks (NNs), and empirical mode decomposition (EMD) [16]. These methods and the variations of these methods are widely used in signal data denoising and image data denoising [17, 18]. Besides, some new noise reduction algorithms are proposed and applied in new engineering fields in recent years, e.g., wavelet packet transform (WPT) is used to extract data from acoustic emission signals containing noisy data [19], noise reduction for desert seismic data using spectral kurtosis adaptive bandpass filter [20], the 1D undecimated discrete wavelet transform (UDWT) has been acquired to attenuate random noise and ground roll [21], a new denoising method was proposed for the simultaneous noise reduction and preservation of seismic signals based on variational mode decomposition (VMD) [22], PCA + linear discriminate analysis (LDA) is first used to extract and denoise the original data, and then nearest neighbor (NN) is used to classify the processed data [23]. These methods are usually for Gaussian distribution noise when performing robustness analysis. In practice, these seemingly chaotic noises do not necessarily obey Gaussian distribution.

Therefore, we firstly try to use robust principal component analysis (RPCA) [24] to denoise the signal data containing non-Gaussian noise in engineering field, which has excellent performance in the field of image noise reduction. It is worth mentioning that the data denoising algorithm is only to reduce the influence of noise as much as possible and cannot completely eliminate the noise. Therefore, we also need excellent classification algorithms to improve classification accuracy. To address this issue, multigroup random walk random forest (MRWRF) is proposed to classify RPCA-purified data. RPCA can effectively take out part of non-Gaussian noise from raw data, and MRWRF can select a better number of decision trees (DTs), which can effectively improve random forest (RF) robustness and classification performance. The hybrid of RPCA and MRWRF can effectively classify data with non-Gaussian distribution noise, and a detailed introduction will be given in the following sections.

2. Related Work

Many applications in engineering are characterized by large quantities of very high-dimensional data. Although these data often lie in very high-dimensional observation spaces, these many dimensions may express only a few intrinsic degrees of freedom [24]. It can be indicated in the following form:

\[ M = A + E, \]  

where \( M \) denotes the contaminated data matrix, \( A \) is low-rank, and \( E \) is a noise term.

Statistically, the above problem is equal to exploring the principal components of the data. When \( N \) follows Gaussian distribution with small variance, the PCA can handle the above problem well. However, the performance of PCA is limited due to lack of robustness to gross corruptions [25]. To overcome the disadvantages of PCA, some robust principal component analysis methods [24, 26] have emerged in recent years. In particular, Zhang et al. [27] established a RPCA method which is a powerful tool for various applications [27–29]. Its noise reduction process can be described by the following optimization problem:

\[
\begin{array}{l}
\min_{A,E} \|A\|_\star + \lambda \|E\|_1 \\
\text{s.t. } M = A + E,
\end{array}
\]  

where \( \| \cdot \|_\star \) and \( \| \cdot \|_1 \) are the kernel norm and 1 norm of the matrix, respectively. If the singular vector of \( A \) is not related to the standard base, when \( \lambda = n^{-1/2} \), the convex problem of (2) can be better solved [30]. Subsequently, \( \lambda \) was modified in [30] and it proves that RPCA is still robust when the noise is not so sparse. This paper uses the formula given in [31] as follows:

\[ \lambda = C_1 \left( 4 \sqrt{1 - \rho} + \frac{9}{4} \right)^{-1} \sqrt{\frac{1 - \rho}{\rho n}}, \]  

where \( C_1 > 0 \) is a constant and \( 0 < \rho < 1 \) is a sparsity parameter, which is the ratio of the number of nonzero entries of the noise matrix \( E \) to the total number in the original matrix.

The solution method of the RPCA algorithm mainly includes the following: the iterative thresholding approach (ITA), the accelerated proximal gradient approach (APGA), the dual approach (DA), and augmented Lagrangian multipliers (ALM) [32]. This paper uses the inexact augmented Lagrangian multiplier method (Inexact ALM) that was proposed in [32]. Its main formula is as follows.

First, construct augmented Lagrangian function for (2):

\[
\Gamma(A, E, Y, \mu) = \|A\|_\star + \lambda \|E\|_1 + \langle Y, M - L - S \rangle \\
+ \frac{\mu}{2} \|M - L - S\|_F^2,
\]  

where \( Y \in R^{m \times n} \) denotes the Lagrange multiplier, \( \mu > 0 \) denotes a positive scalar, and \( \langle \cdot \rangle \) denotes the standard inner product.

Then, iterate according to the following formula:
The rest of the paper is organized as follows. Section 3 presents the proposed method which was tested and verified using a concrete aqueduct’s structural damage data. Finally, Section 6 concludes the paper.

3. The Improved Algorithm

3.1. Random Forest Based on the Thought of Random Walk (RWRF). RF is a classifier with multiple DTs, and its basic thought is to extract T samples from the original training set using Bootstrap sampling. Then, T decision tree models are established for T samples and T classification results are obtained. Finally, according to the T classification results, each record is voted to determine its final classification. Its core function is

\[ T(X) = \arg \max_y \sum_{i=1}^{n} I(t_i(X) = Y), \]

where \( T(X) \) denotes the combined classifier model, \( t_i(X) \) is single DT classification model, \( Y \) is the output variable, and \( I(\cdot) \) is the voting decision function.

RF can obtain the upper bound of the generalization error according to the law of large numbers, as shown in the following:

\[ \lim_{k \to \infty} PE^* = P_{XY}\left\{ P(t(X) = Y) - \max_{j \neq Y} P(t(X) = j) < 0 \right\}, \]

where \( k \) denotes the number of DTs. As the number of DTs increases, the generalization error of RF will gradually tend to the upper bound of the above formula, which shows that RF has a good ability to prevent overfitting.

Random walk (RW) [44] is one of the most basic processes in dynamics. It has the ability to globally patrol and can be widely used, such as in image segmentation [45], biology, and electron transport. Its core thought is that the individual moves from the current state to the next state is random, that is, there are same probability of reaching the other locations in the next step. When moving to the next position, the individual will use the current position as the starting point and repeat the above process. In 2-dimensional space, random walk can be explained by Figure 1.

As can be seen from Figure 1, the red dot represents the position of the individual at the current moment and it has the same transition probability to reach any other position in the plane at the next moment. When the next position is reached, the individual repeats the above process with the following position as the starting point. This unpredictable random walk gives the individual the ability to find the best advantage within the set area.

In fact, it is a nonlinear optimization problem that chooses the number of DTs that enables the random forest algorithm to achieve the best classification effect. The optimization problem is denoted as follows:

\[ \begin{align*}
\text{Find} & \quad k \in R \\
\text{Subject to} & \quad \max(P) \in [0, 1],
\end{align*} \]

where \( P \) denotes the classification accuracy.
In response to this problem, this paper proposes a random forest based on the thought of random walk. The specific algorithm steps are as follows:

Step 1: randomly select the initial number of DTs \( k_1 \). Set the initial walking step \( \lambda_1 \), the control precision \( \varepsilon \), the iteration control times \( N \), and the current number of iterations \( i = 1 \). Calculate the initial accuracy \( P_1 \), with equation (6).

Step 2: generate a random number \( u_i \) between \(-1, 1\) and calculate \( k_{i+1} = k_i + \lfloor u_i \lambda_i \rfloor \) (where \( \lfloor \rfloor \) denotes the largest integer no more than \( u_i \lambda_i \)). Complete a step walk.

Step 3: calculate the value of \( P_{i+1} \). If \( P_{i+1} > P_i \), reset \( i \) to 1 and change \( P \) to \( P_0 \). Otherwise, return to step 2.

Step 4: if no better value can be found after \( N \) consecutive iterations, it is considered that the optimal solution is centered on the current optimal solution. At this point, if \( \lambda_{i+1} < \varepsilon \), end the algorithm. Otherwise, set \( \lambda = \alpha \lambda_i \), where \( 0 < \alpha < 1 \) is the step reduction factor. Then, return to step 1 and start a new round of walks.

3.2. Multigroup Random Walk Random Forest (MRWRF)

In practice, we find that although the random walk algorithm is simple to operate, its performance depends on the initial step size and the choice of the initial number of DTs. This phenomenon easily leads to the algorithm falling into the local optimum. Therefore, we propose the multigroup random walk random forest (MRWRF). First, establish three groups of the same level where each group is responsible for different areas, and then the individuals with different velocities randomly walk from different locations in each group. Update the best point of each area in each walk. If no better value is found over multiple successive iterations, then the best point of the area is considered to be the current point, and end the search. When the three groups find their respective best advantages, the three compare again and select the point with the highest accuracy as the optimal number of DTs in the random forest. Key steps of the MRWRF algorithm:

Step 1: initialize the MRWRF algorithm, create three groups, and create three individuals in each group. Set each individual's range of walk and randomly select the initial step size \( \lambda_{mn} \) \( m,n = 1,2,3 \) and initial position of each individual \( k_{mn} \) \( m,n = 1,2,3 \). \( \lambda_{mn} \) denotes the initial step size of the \( n \)th individual in the \( m \)th group, and \( k_{mn} \) denotes the initial position of the \( n \)th individual in the \( m \)th group. Forming the following matrix:

\[
\begin{pmatrix}
\lambda_{11} & \lambda_{12} & \lambda_{13} \\
\lambda_{21} & \lambda_{22} & \lambda_{23} \\
\lambda_{31} & \lambda_{32} & \lambda_{33}
\end{pmatrix}
\]

Set the iteration control times \( N \), the control precision \( \varepsilon \), and the current number of iterations \( i = 1 \).

Step 2: calculate the classification accuracy of each individual's initial point \( P_{mn} \) \( m,n = 1,2,3 \), with equation (6). Generate random number \( u_{mn} \) \( m,n = 1,2,3 \) between \(-1,1\) and calculate \( k'_{mn} = k_{mn} + [u_{mn} \lambda_{mn}] \), \( i = 1,2,3, \ldots, N \). Complete one step walk.

Step 3: calculate the accuracy of the latest location \( P_{mn} \) with equation (6). If \( P_{mn} > P_{mn} \), change \( k_{mn} \) to \( k'_{mn} \) and change \( P_{mn} \) to \( P_{mn} \). \( k_{mn} \) is the best point of the \( n \)th individual in the \( m \)th group, \( P_{mn} \) denotes the optimal accuracy of the \( n \)th individual in the \( m \)th group, and reset \( i \) to 1. Otherwise, return to step 2.

Step 4: if no better value can be found for \( N \) consecutive iterations, it is considered that the optimal solution is centered on the current optimal solution. At this point, if \( \lambda < \varepsilon \), end the algorithm. Otherwise, let \( k'_{mn} = \alpha k_{mn} \), where \( 0 < \alpha < 1 \) is the step reduction factor; return to step 1 and start a new round of walks.

Step 5: after three individuals in a group find the best point of their respective regions, the three compare and select the best point of the group. When the best points of the three groups are found, the groups are compared and the overall optimal value is selected. Output the value of the accuracy \( P \) and the number of DTs \( k \), which ends the algorithm.

The difficulty of the algorithm is that when the individual randomly walks, the same position may be swept many times, which will reduce the efficiency of the algorithm. We trim the algorithm as follows. The individual will mark the current position and classification accuracy rate during each walk and will automatically skip the marked position in subsequent walks, thereby improving the computational efficiency of the algorithm.
3.3. Data Denoising and Classification Process of RPCA-MRWRF. RPCA-MRWRF is a hybrid learning algorithm that differs from traditional random forest algorithms in two aspects: (1) it preprocesses raw data using RPCA and (2) a better number of DTs can be selected to effectively improve the classification accuracy. The specific classification process of the RPCA-MRWRF algorithm is as follows.

As shown in Figure 2, first, the raw data contaminated by random noise is input into the RPCA algorithm for preprocessing to remove noise. Then, the processed data is input into MRWRF for classification, and the specific classification process is given in Section 3.2. Finally, the classification results of different data and the corresponding accuracy rate are output.

4. Numerical Simulation Case

Generally, large-scale building damage identification is mainly based on the monitoring data that are obtained by monitoring sensors that are embedded in the interior of the building and the test data that are obtained by routine and special tests. These data usually contain varying degrees of noise, which will affect the accuracy of the classification. Therefore, in this paper, a finite element model is established for an aqueduct and the relevant parameters are obtained. Then, random noise of different degrees and intensity is added to simulate actual environmental noise and used as raw data to test the performance of the RPCA-MRWRF algorithm.

An aqueduct in an empty tank overhaul state, the main body of which is a single-slot ribbed belt tie rod structure. The section size is 6.0 m * 5.4 m, the length of a single span is 30 m, the thickness of a sidewall is 0.60 m, and a 2.0 m wide sidewalk board is at the top. Side ribs and bottom ribs are added to the aqueduct body with widths of 0.5 m and heights of 0.7 m and 0.9 m, respectively. A tie rod is placed on the top of the transverse wall. Its section size is 0.3 m * 0.4 m. The rib spacing is 2.5 m. The density of the concrete material was set at 2550 kg/m³, the elasticity modulus was set at 34.5 GPa, and the Poisson’s ratio was set at 0.167. The analysis modeling was built using SOLID95 mechanics elements in ANSYS. The three-dimensional map and free meshing map of the aqueduct is shown in Figure 3.

Next, we simulate damage in the middle of the bottom plate. The crack width is 4 mm. The degree of aqueduct damage is expressed by η, which is the ratio of the depth of the crack to the thickness of the bottom plate. In this paper, η was 0, 5%, 10%, 20%, 30%, 40%, and 50%. In addition, four types of damage are defined: no damage (ND, η = 0), general damage (GD, η = 5%, 10%), heavier damage (HD, η = 20%, 30%), and serious damage (SD, η = 40%, 50%).

The input parameters consist of the first ten natural frequencies wi, i = 1, 2, . . . , 10, of the aqueduct before and after the damage is calculated by ANSYS, as shown in Table 1.

In actual measurement, it is necessary to measure the natural frequencies of the structure multiple times to reduce the influence of environmental noise. Therefore, we performed the following processing on the data set in this simulation. η = 0, which means that no damage was recorded 184 times. η = 5%, 10%, 20%, 30%, 40%, and 50% of the data were recorded 92 times, respectively, forming a 736 * 10-dimensional data set. Non-Gaussian random noise [−σ, σ] of different degrees and intensity is added to it. \( \Delta = \sigma/\omega_i, i = 1, 2, \ldots, 10 \), is used to represent the strength of the random noise, and \( \omega_i \) is the input parameter when \( \eta = 0 \).

5. Algorithm Performance Verification and Result Analysis

In this section, we mainly test the classification performance of the RPCA-MRWRF algorithm and classify noise-contaminated aqueduct modal data. The main contents include the following: (1) discussing the influence of the number of DTs in RF on its classification performance and comparing the ability of multigroup random walk (MRW), FS, BS, and GA algorithm to find the better number of DTs in RF, (2) comparing the classification performance of the RPCA-MRWRF algorithm with RF algorithm under different degrees of random noise, and (3) comparing the classification performance of the RPCA-MRWRF algorithm with other existing classification algorithms under random noise with the same degree and intensity.

5.1. Discussing the Influence of the Number of DTs in RF and MRW Performance Verification. In this paper, 586 groups were randomly selected from the 736 sets of data as the training set and the remaining 150 sets of data were used as the test sets. The effect of the number of DTs on the RF's performance is tested when \( \Delta = 30\% \) and \( \rho = 0.5 \). As shown in Figure 4, the RF’s performance is best when the number of DTs is 98, but the RF’s performance is the worst when the number is 140. When the number of DTs increases to a certain extent, the classification accuracy of the RF is basically unchanged. Therefore, it can be stated that the number of DTs has an impact on the RF’s performance.

The better number of DTs found by MRW, FS, BS, and GA and the corresponding accuracy rates are shown in the Tables 2 and 3. It is worth mentioning that these are performed under the conditions of the RPCA processing the raw noise data.

When \( \rho = 0.5 \), the number of DTs sought by MRW corresponds to the highest accuracy rate except in the case of \( \Delta = 2\% \), and when \( \rho = 0.6 \), the number of DTs sought by MRW corresponds to the highest accuracy rate except in the case of \( \Delta = 10\% \) and \( \Delta = 30\% \). The better number of DTs searched by FS is less than MRW, and the better number of DTs searched by BS is more than MRW. There is a similar situation when \( \rho = 0.6 \). A likely reason is that the two methods are relatively easy to fall into local optimum [39]. GA has better performance than FS and BS, and it is not easy to fall into local optimum, but it has poor performance compared with MRW. The reason for this phenomenon may be that GA is more dependent on the choice of initial values.
and it is difficult to find the optimal or near-optimal number of DTs in a limited iteration [46].

The above analysis can show that MRW has better performance than FS, BS, and GA. Since the test set and the training set are randomly selected during each classification process, the better number of preferred DTs will be different each time. However, it is always able to find a better number of DTs to improve the classification accuracy of the data in each run.

5.2. Performance Comparison between RPCA-MRWRF and RF. This paper compares the classification performance of the RPCA-MRWRF algorithm and the RF algorithm when $\rho$ is 0.5 and 0.6, respectively. The following indicators are used to evaluate the performance of the algorithm: (1) Classification Precision, (2) Confusion Matrix, and (3) Cohen’s Kappa Coefficient.

As can be seen from Figure 5, RPCA can effectively reduce the impact of noise on the data but cannot completely
Table 1: Input parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>( \eta = 0 )</th>
<th>( \eta = 5% )</th>
<th>( \eta = 10% )</th>
<th>( \eta = 20% )</th>
<th>( \eta = 30% )</th>
<th>( \eta = 40% )</th>
<th>( \eta = 50% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_1 )</td>
<td>11.055</td>
<td>11.05</td>
<td>11.052</td>
<td>11.051</td>
<td>11.05</td>
<td>11.05</td>
<td>11.05</td>
</tr>
<tr>
<td>( w_3 )</td>
<td>18.548</td>
<td>18.538</td>
<td>18.542</td>
<td>18.542</td>
<td>18.539</td>
<td>18.542</td>
<td>18.539</td>
</tr>
<tr>
<td>( w_4 )</td>
<td>18.85</td>
<td>18.854</td>
<td>18.857</td>
<td>18.857</td>
<td>18.85</td>
<td>18.853</td>
<td>18.855</td>
</tr>
<tr>
<td>( w_6 )</td>
<td>27.24</td>
<td>27.231</td>
<td>27.233</td>
<td>27.229</td>
<td>27.23</td>
<td>27.227</td>
<td>27.225</td>
</tr>
<tr>
<td>( w_7 )</td>
<td>27.581</td>
<td>27.566</td>
<td>27.569</td>
<td>27.566</td>
<td>27.563</td>
<td>27.564</td>
<td>27.566</td>
</tr>
<tr>
<td>( w_9 )</td>
<td>40.424</td>
<td>40.412</td>
<td>40.417</td>
<td>40.412</td>
<td>40.412</td>
<td>40.415</td>
<td>40.417</td>
</tr>
<tr>
<td>( w_{10} )</td>
<td>42.097</td>
<td>42.074</td>
<td>42.072</td>
<td>42.07</td>
<td>42.055</td>
<td>42.045</td>
<td>42.027</td>
</tr>
</tbody>
</table>

Table 2: The better number of DTs found by MRW, FS, BS, and GA and the corresponding accuracy rates when \( \rho \) is 0.5.

<table>
<thead>
<tr>
<th>( \Delta )</th>
<th>MRW DTs</th>
<th>MRW Accuracy (%)</th>
<th>FS DTs</th>
<th>FS Accuracy (%)</th>
<th>BS DTs</th>
<th>BS Accuracy (%)</th>
<th>GA DTs</th>
<th>GA Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1%</td>
<td>73</td>
<td>98</td>
<td>73</td>
<td>98</td>
<td>799</td>
<td>90.67</td>
<td>73</td>
<td>98</td>
</tr>
<tr>
<td>1%</td>
<td>93</td>
<td>93.33</td>
<td>69</td>
<td>88.67</td>
<td>535</td>
<td>85</td>
<td>151</td>
<td>91.33</td>
</tr>
<tr>
<td>2%</td>
<td>524</td>
<td>94</td>
<td>135</td>
<td>91.33</td>
<td>481</td>
<td>95.33</td>
<td>163</td>
<td>96</td>
</tr>
<tr>
<td>5%</td>
<td>172</td>
<td>92.67</td>
<td>91</td>
<td>86.67</td>
<td>267</td>
<td>90</td>
<td>97</td>
<td>90.67</td>
</tr>
<tr>
<td>10%</td>
<td>204</td>
<td>94.67</td>
<td>109</td>
<td>90.67</td>
<td>247</td>
<td>91</td>
<td>173</td>
<td>93.33</td>
</tr>
<tr>
<td>20%</td>
<td>206</td>
<td>94.67</td>
<td>55</td>
<td>82</td>
<td>385</td>
<td>89.33</td>
<td>537</td>
<td>88.67</td>
</tr>
<tr>
<td>30%</td>
<td>98</td>
<td>94.67</td>
<td>91</td>
<td>90.67</td>
<td>163</td>
<td>88</td>
<td>133</td>
<td>90.67</td>
</tr>
</tbody>
</table>

Table 3: The better number of DTs found by MRW, FS, BS, and GA and the corresponding accuracy rates when \( \rho \) is 0.6.

<table>
<thead>
<tr>
<th>( \Delta )</th>
<th>MRW DTs</th>
<th>MRW Accuracy (%)</th>
<th>FS DTs</th>
<th>FS Accuracy (%)</th>
<th>BS DTs</th>
<th>BS Accuracy (%)</th>
<th>GA DTs</th>
<th>GA Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1%</td>
<td>93</td>
<td>88</td>
<td>67</td>
<td>86</td>
<td>269</td>
<td>81.3</td>
<td>93</td>
<td>88</td>
</tr>
<tr>
<td>1%</td>
<td>195</td>
<td>86.67</td>
<td>95</td>
<td>83.33</td>
<td>395</td>
<td>79.33</td>
<td>263</td>
<td>84.67</td>
</tr>
<tr>
<td>2%</td>
<td>243</td>
<td>84</td>
<td>79</td>
<td>81.33</td>
<td>243</td>
<td>84</td>
<td>89</td>
<td>80.67</td>
</tr>
<tr>
<td>5%</td>
<td>166</td>
<td>84.67</td>
<td>153</td>
<td>84</td>
<td>197</td>
<td>82.67</td>
<td>684</td>
<td>77</td>
</tr>
<tr>
<td>10%</td>
<td>293</td>
<td>83.33</td>
<td>53</td>
<td>79.33</td>
<td>325</td>
<td>81.33</td>
<td>159</td>
<td>84</td>
</tr>
<tr>
<td>20%</td>
<td>97</td>
<td>85.33</td>
<td>97</td>
<td>85.33</td>
<td>651</td>
<td>75.33</td>
<td>463</td>
<td>81.33</td>
</tr>
<tr>
<td>30%</td>
<td>779</td>
<td>84</td>
<td>35</td>
<td>78.67</td>
<td>779</td>
<td>84</td>
<td>211</td>
<td>85.33</td>
</tr>
</tbody>
</table>

Figure 4: The influence of the number of DTs on the RF's performance.
eliminate it. Classification precision results are shown in Figures 6 and 7. When \( \rho \) is 0.5, the overall accuracy of the RPCA-MTWRF is above 93%, the lowest is 93.3%, and the accuracy of various types of damage is also above 85%. However, the classification accuracy of the RF is slightly different. When \( \Delta \) is small, its classification accuracy is higher, but as \( \Delta \) increases, the accuracy decreases rapidly. The overall recognition rate is basically below 85% except in the case of lower noise intensity (\( \Delta = 0.1\% \) or 1%), and the lowest is 81.33%. When \( \rho \) is 0.6, the overall recognition rate of RPCA-MRWRF is reduced, which is basically approximately 85%, and the lowest is 84%. However, the overall recognition rate of the RF is below 78% and the lowest is 68%.

The confusion matrix of RPCA-MRWRF and RF are shown in Figures 8–11. Most of the classification results of the two algorithms are distributed in the diagonal area of the confusion matrix, indicating that they all have certain classification ability. In the ideal state (\( \Delta = 0\% \)), both RPCA-MRWRF and RF have 100% accuracy. As \( \rho \) and \( \Delta \) increase, the number of RF classification errors increases significantly, but RPCA-MRWRF does not change much.

According to the confusion matrix of RPCA-MRWRF and RF, the Kappa coefficient under different conditions are calculated, as shown in Figure 12. The results show that the Kappa value of RPCA-MRWRF is always larger than RF under different levels of noise.

The above analysis shows that the RPCA-MRWRF algorithm has better noise immunity and better classification performance than RF.
Figure 7: The classification accuracy rates of (a) RPCA-MRWRF and (b) RF when $\rho$ is 0.6.

Figure 8: Continued.
serious damage.

20%, and 30%, respectively, where ND denotes no damage, GD denotes general damage, HD denotes heavier damage, and SD denotes serious damage.

Figure 8: When $\rho$ is 0.5, (a), (b), (c), (d), (e), (f), (g), and (h) are the confusion matrices of RPCA-MRWRF for $\Delta$ is 0, 0.1%, 1%, 2%, 5%, 10%, 20%, and 30%, respectively, where ND denotes no damage, GD denotes general damage, HD denotes heavier damage, and SD denotes serious damage.

<table>
<thead>
<tr>
<th>ND</th>
<th>GD</th>
<th>HD</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.7%</td>
<td>0.7%</td>
<td>0.7%</td>
<td>0.7%</td>
</tr>
<tr>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>97.3%</td>
<td>92.7%</td>
<td>92.3%</td>
<td>95.7%</td>
</tr>
<tr>
<td>2.7%</td>
<td>7.3%</td>
<td>7.7%</td>
<td>4.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ND</th>
<th>GD</th>
<th>HD</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.7%</td>
<td>0.7%</td>
<td>0.7%</td>
<td>0.7%</td>
</tr>
<tr>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>97.3%</td>
<td>95.1%</td>
<td>96.2%</td>
<td>91.3%</td>
</tr>
<tr>
<td>2.7%</td>
<td>4.9%</td>
<td>3.8%</td>
<td>8.7%</td>
</tr>
</tbody>
</table>

Figure 9: When $\rho$ is 0.5, (a), (b), (c), (d), (e), (f), (g), and (h) are the confusion matrices of RF for $\Delta$ is 0, 0.1%, 1%, 2%, 5%, 10%, 20%, and 30%, respectively, where ND denotes no damage, GD denotes general damage, HD denotes heavier damage, and SD denotes serious damage.
When $\rho$ is 0.6, (a), (b), (c), (d), (e), (f), (g), and (h) are the confusion matrices of RPCA-MRWRF for $\Delta$ is 0, 0.1%, 1%, 2%, 5%, 10%, 20%, and 30%, respectively, where ND denotes no damage, GD denotes general damage, HD denotes heavier damage, and SD denotes serious damage.
5.3. The Performance of Other Existing Algorithms. In addition, we tested the performance of other existing algorithms when $\rho$ is 0.5, including the following: multiwavelet and mutation particle swarm optimization algorithm (MW-MPSO) [5], the KPCA-SVM [14], the wavelet packet transform (WPT) analysis and fully connected deep neural networks.
network (WPT-FCDNN) [19], and the PCA-LDA-NN [23].

The classification accuracy of each algorithm is shown in Figure 13 and Table 4.

In the ideal state ($\Delta = 0\%$), several algorithms all have 100% accuracy. As the noise level increases, although the accuracy of several algorithms is decreasing, the accuracy of several methods is maintained above 80% except KPCA-SVM and PCA-LDA-NN. When $\Delta$ is small, KPCA-SVM has good performance, but when $\Delta$ is larger than 5%, the accuracy of KPCA-SVM is less than 70%. The accuracy of PCA-LDA-NN is lower than 80%, and it is maintained at around 75% when $\Delta\mu\lambda$ is greater than 20%. This may be due to the limited noise reduction capability of the PCA-LDA and the lack of robustness of the corresponding classifier. The lowest accuracy of RF, MW-MPSO, and WPT-FCDNN is more than 80% under different degrees of non-Gaussian noise pollution, which indicates that they all have certain antinoise ability. RF benefits from the integration of multiple decision trees to reduce the risk of misjudgment. The performance of MW-MPSO and WPT-FCDNN may be attributed to the favourable denoising capability and the preponderance of robustness of the corresponding classifier. Meanwhile, it can be distinctly observed that the accuracy of the developed algorithm is more than 93%, which is the most excellent in existing algorithms.

It is feasible to apply the RF, MW-MPSO, and WPT-FCDNN algorithm for the case when there is not a strict requirement for the classification accuracy and no serious environmental noise. However, the RPCA-MRWRF, the proposed algorithm in this paper, is the best choice for the case when there is a more stringent requirement for classification accuracy and more serious environmental noise. In

Figure 13: (a), (b), (c), and (d) are the classification accuracy rates of MW-MPSO, KPCA-SVM, WPT-FCDNN, and PCA-LDA-NN when $\rho$ is 0.5, respectively.
general, the RPCA-MRWRF algorithm has excellent anti-noise ability and can provide a new method for the identification of aqueduct structural damage and other engineering classification problems.

6. Conclusion

In this paper, a hybrid algorithm of RPCA-MRWRF is proposed. RPCA can reduce the impact of noise on the original data as much as possible, and MRWRF can choose the better DTS number to improve the classification accuracy. Experimental results show that RPCA-MRWRF has better classification performance than other existing classification algorithms and can thus provide a new approach for data classification problems in engineering.

However, there are still many shortcomings in this paper. For example, the time consumption of the RPCA-MRWRF algorithm is slightly longer than that of other types of algorithms, and it will be discussed as key points in subsequent research.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This article was funded by The National Key R&D Program of China (Grant number: 2018YFC0406902).

References


