Fault Diagnosis of Wind Turbine’s Gearbox Based on Improved GA Random Forest Classifier
Hao GAN* and Bin JIAO
Shanghai Dianji University, Shanghai, China
*Corresponding author

Keywords: Gearbox, Fault diagnosis, PCA, GA random forest.

Abstract. In recent years, there were many studies on intelligent fault diagnosis of wind turbines based on vibration signals. There are also many algorithms for classification of failure categories. Such as SVM, ELM and Random Forest. Random forest is a new ensemble algorithm. In this paper, the dimension reduction of feature vector using PCA algorithm is proposed for gearbox vibration signals, which makes the training time of the model reduced. And then, it used GA to optimize the number of decision trees and the number of attributes in the split attribute set in the random forest combined classifier. Through comparative experiments, the effectiveness of the proposed method is proved.

Introduction

Recently, the research on vibration monitoring and fault diagnosis technology for wind turbines has been gradually deepened, and more and more fault diagnosis techniques and algorithms have been applied to the monitoring of wind turbines [1][2]. Gearbox is an important part of wind turbines that drives the energy to generator. In the event of failure of rotating machinery and equipment, the results often result in serious economic losses and safety accidents. It is very important to predict the failure of gearbox.

Vibration sensors are usually used to collect the vibration signal of the gearbox. And then, they extracted the time domain features, frequency domain features, or time-frequency features as feature vectors of the gearbox [3]. Many experts apply extracted feature vectors to fault recognition algorithms, such as Zhen-You [4] used ANN to detect the fault of wind turbine. And Zhao R [5] used soft label SVM to predict fault of wind turbine.

Random forest is an efficient algorithm invented and proposed by Leo [6] in 2001. It is composed of many single tree classification regression trees (CART). Finally, the voting method determines the classification result. While ensuring the effectiveness of a single classification tree, the correlation between the classification trees is reduced, and the performance of the combined classifier is improved. The random forest combination classifier contains a plurality of randomly generated decision tree classifiers, and there is no association between the decision trees. When the data to be classified is input into the random forest combination classifier, each decision tree is classified separately, and the classification results of multiple decision trees are integrated to perform vote evaluation, thereby obtaining the final classification result.

For random forests, there are too many features in the input samples and the structure is too complex. The training of decision trees will become tedious and complicated. The Principal Component Analysis (PCA) algorithm can do a good job of reducing attributes.

In this paper, fault diagnosis of wind turbine’s gearbox is a multi-classification problem. For vibration signals, there are various characteristics in the time-frequency domain that reflect gear failures. How to choose the right feature value for fault diagnosis is very important. PCA is a good algorithm for selecting feature values and the number of decision trees in the random forest combination classifier and the number of attributes in the split attribute set need to be set in advance. The correct choice of parameters will affect the final classifier classification effect. So, this paper proposed a global optimization algorithm to select these two indicators. GA Random forest algorithm
was proposed. The experiment compares the improved random forest algorithm with the traditional classifier and it proves that it is superior to other classifier in fault diagnosis.

**Proposed Algorithm**

**Basic Random Forest Algorithm**

Random forest is a classifier composed of multiple decision trees. \( \{ h(x, \Theta_k), k = 1, 2, 3, \ldots \} \) where \( \{ \Theta_k \} \) is a random vector that is independent and identically distributed. Eventually, the decision result is determined by the comprehensive voting of all decision trees.

For example, there are \( K \) classifiers: \( h_1(x), h_2(x), h_3(x), \ldots, h_k(x) \). \( X, Y \) is the random vectors. Edge function is defined as follow:

\[
mg(X, Y) = av_y I(h_k(X) = Y) - \max_{j \neq Y} av_y I(h_k(X) = j)
\]

(1)

Where \( I(x) \) is Characteristic Function, The edge function characterizes the degree to which the average number of votes for the correct classification of vector \( Y, X \) exceeds that of any other class average. It can be seen that the greater the edge function value is, the higher the confidence of the classification is.

The Bagging algorithm is an important anthropological approach proposed by Leo Breiman [8]. Given some weak learning algorithm and the training set samples \( T \{ (x_1, y_1), \ldots, (x_k, y_k) \} \), return training samples for training set. Then for each base classifier, a training subset that is the same as in number, but different from the original training set is generated, thereby we can get different basic classifiers. Nearly 37% of all training samples will not appear in every training subset that is called OOB(out of bag).

The random forest algorithm flow is as follows:

1. We assume that the size of the forest to be built is \( K \). So, \( k \) training sample sets are generated by bagging in the training sample set, (2) Construct a classification tree for each training sample set. Assume that the number of features of the sample is \( M \). The growth of a single classification tree is as follows: Select \( M_i \) features at random from the \( M \) features of each internal node of the classification tree as candidate features \( (M_i < M) \). And then, select an optimal feature from the \( M_i \) candidate features to split the nodes by the minimum principle of node purity. Repeat the above operation until the stop extreme is reached.

2. Forecasting OOB Data Using Built-in Multiple CART Tree Classifiers. The classification result is decided according to the simple majority vote method Eq.2 of each tree classifier voting result. At the same time, pruning is performed on the expanding binary tree according to the classification result.

\[
C_p = \arg \max \sum_{i=1}^{N} I(\frac{n_h - c}{n_h})
\]

(2)

**Genetic Algorithm**

The genetic algorithm is designed to achieve self-learning and optimization by simulating the process of biological evolution. It is operating through genetic operators such as selection, crossover and mutation in each optimization process to generate a new generation of candidate solutions. In this paper, when optimizing the number of decision trees and the number of attributes in the split attribute set in the random forest combined classifier, the fitness function is set to Eq.3.
\[
F = \frac{\sum_{i=1}^{V} \sum_{j=1}^{e} R_{ij}}{eV}, R_{ij} = \begin{cases} 
1, & O(T_{ij}) = O(y_{ij}) \\
0, & \text{else} 
\end{cases}
\]  

(3)

Where \(O(T_{ij})\) and \(O(y_{ij})\) are the actual mechanical failure category and predicted category. \(V\) is the cross-validation. \(e\) is the number of samples included in each data set.

**Principal Component Analysis**

PCA algorithm is proposed to reduce the original set of characteristic information to only a few by appropriate calculation and screening. And can represent the original information independently and comprehensively.

**Proposed GA Random Fores Algorithm**

For the vibration signal of the wind turbine’s gearbox, the signal contains a variety of characteristic values. That include approximately 34 eigenvalues in time-frequency domain and wavelet energy. In this paper, the first 8 characteristic values with the highest contribution rate to the fault signal are selected by the PCA algorithm to form the feature vector. The number of decision trees \(n_{tree}\) in the random forest and the number of attributes \(M_i\) in the split attribute set need to determine the optimal value. So this paper use GA algorithm to get the best value of \(n_{tree}\) and \(M_i\). The algorithm flow is as follows:

1. Using binary coding to encode different individuals (that is \(n_{tree}\) and \(M_i\)). The number of decision trees in each combined classifier and the number of attributes in the split attribute set make up a pair of alleles.
2. Using roulette operators to perform individual selection operations, so that adaptable individuals contribute one or more descendants to the next generation with a higher probability.
3. Generate new individuals through crossover and mutation
4. Repeat step (3) until the default number of seeks is reached to obtain the best individual. So we can get the best value of \(n_{tree}\) and \(M_i\)

**Simulation Experiment**

In this article, we have designed a test system that simulates the operation of wind turbines as the Fig.1 shows.

![Figure 1. The sensor installation diagram of simulation test device for wind turbine’s gearbox](image)

Seven vibration sensors are connected to the acquisition card to collect vibration signal. In this paper, a large number of fault feature parameters are extracted after the noise reduction processing of these vibration signals to form the original fault feature set. There are 3 frequency-domain features, 10 time-domain features, and 8 wavelet decomposition energy features.
Table 1. The feature set before PCA.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency domain</td>
<td>RMS value, Center of gravity frequency, Frequency variance</td>
</tr>
<tr>
<td>Time Domain</td>
<td>Absolute value, Maximum, Mean, variance, Waveform factor, Standard deviation, Peak factor, Skewness, Kurtosis, Margin factor, Square Root Amplitude,</td>
</tr>
<tr>
<td>Time-frequency domain</td>
<td>After the wavelet packet decomposition, the energy value corresponding to the eight nodes of the third layer. $E_{3-1},...,E_{3-8}$</td>
</tr>
</tbody>
</table>

Table 2. The feature set after PCA.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency domain</td>
<td>RMS value, Frequency variance</td>
</tr>
<tr>
<td>Time Domain</td>
<td>Peak factor, Skewness, Kurtosis, Margin factor, Square Root Amplitude,</td>
</tr>
<tr>
<td>Time-frequency domain</td>
<td>$E_{3-3}, E_{3-6}$</td>
</tr>
</tbody>
</table>

As we can see, after the PCA algorithm, we have reduced the characteristics of failure categories from 21 to 8. So, the feature vector dimension is greatly reduced, which makes the calculation of intelligent fault classification quick and easy.

We use the features obtained from dimensionality reduction as input of the improved GA random forest algorithm. There are 200 groups of vibration data. 150 of them are training data and 50 of them are test data. And in the GA algorithm, population size is 12, the default number of evolution is 100. We can get the number of corresponding optimal decision trees and the number of attributes in the best split attribute. GA algorithm iterative optimization fitness value transformation as shown below:

![Figure 2. Update of fitness value.](image)

The best $n_{tree}$ is 52, the best $M_i$ is 5. To prove the effectiveness of the proposed method, We compare the improved GA Random Forest algorithm with the traditional classification algorithm. Here are the results.

![Figure 3. The result of GA random forest.](image)  
![Figure 4. The result of random forest.](image)
Table 3. Result of accuracy and training time (50 test samples).

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>Random Forest</th>
<th>GA Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>86%</td>
<td>90%</td>
<td>94%</td>
</tr>
<tr>
<td><strong>Training time before PCA</strong></td>
<td>1.2764</td>
<td>1.5796</td>
<td>1.6745</td>
</tr>
<tr>
<td><strong>Training time after PCA</strong></td>
<td>0.9756</td>
<td>1.121</td>
<td>1.2613</td>
</tr>
</tbody>
</table>

In summary, it can be seen that the PCA algorithm can shorten the training time and increase the diagnosis speed. And the proposed GA random forest algorithm of fault diagnosis accuracy is 94%.

**Conclusion**

This paper reduces the calculation of feature data by the PCA algorithm and it also has a good effect on the training time of the intelligent fault diagnosis model. This paper also uses genetic algorithm to find the optimal parameters which makes the random forest algorithm improved the classification accuracy of fault diagnosis.

**References**


