The Digging Signal Identification by the Random Forest Algorithm in the Phase-OTDR Technology

Jiao Wang*, Yanzhu Hu and Yanchao Shao

Automation College, Beijing University of Posts and Telecommunications, China, Beijing, 100876

*Corresponding author e-mail: wangjiao0516@bupt.edu.cn

Abstract. Phase optical time domain reflectometry (Phase-OTDR) is a new distributed detection technology. Its work on signal analysis is not perfect. Random forest algorithm is generally effective for classification problems. A method of digging signal identification based on random forest algorithm is proposed in Phase-OTDR technology. The method has nice identification accuracy. Firstly, the Phase-OTDR signal is pre-processed by the filter method. Secondly, the digging signals and normal signals are extracted respectively in time domain and frequency domain. Thirdly, the random forest algorithm is regarded as a classifier to identify and classify signals. Lastly, the signal is divided into test samples and training samples to experiment. The experiments prove the method the effectiveness.

1. Introduction

Phase-OTDR or Φ-OTDR (phase sensitive optical time domain reflect meter) serves as a typical technique for monitoring distributed vibrations which is put forward by Taylor and Lee in 1993. It has a wide range of applications in the field of large building structure health monitoring, [2] the perimeter security of important places [3] and so on. Digging behavior is a typical abnormal behavior. Whether stealing or digging, it is easy to cause negative effects on work and life. The digging behavior includes illegal excavation, basement expansion and so on.

In recent years, distributed optical fiber sensors gradually replace traditional vibration sensors in the field of safety monitoring. Because it has the advantages of high sensitivity, long detection distance, anti-electromagnetic interference, corrosion resistance, long life and so on. [4] When digging behavior is directly or indirectly applied to optical fiber, the light signal in the optical fiber will change with it. [5] The digging signals can be identified through signal processing and data digging. It has far-reaching significance for finding illegal behavior in time.

A method of digging signal recognition based on random forest algorithm is proposed in the paper. [6] Firstly, the signal of Phase-OTDR is excluded from the outliers. It is filtered and de-noised in order to reduce the effect of noise. [7] Secondly, the signal is extracted features in the time domain and frequency domain. [8] These features are constructed feature vector. Lastly, the random forest algorithm is optimized as a classifier. The digging signal is extracted. The experiments prove recognition accuracy of the method effective.
2. The Digging Signal of Phase-OTDR

2.1. Signal Preprocessing Based on Smooth Filtering

Because it has the characteristics of high sensitivity, tiny signal changes can be detected. [9] When the propagation distance is far away and the signal intensity is weak, the signal is easily drowned by noise. [10] First of all, the signal should be de-noised. Signal de-noising method is usually wavelet de-noising. [11] After the experiment, it was found that the de-noising effect was not obvious for the excavation signal. A smoothing filtering method is adapted to de-noise the digging signal. There are two main points in smoothing filtering, filter window size and filter algorithm selection. The size of filter window can be set according to the actual processing data and processing requirements. Filtering algorithms include mean filtering, median filtering and so on. Three steps in the smoothing filter are Low Pass Filtering, Sliding Window Superposition with 10 Data Points and Repeating the Second Step after the Absolute Value is taken. The smoothing filter and the wavelet filter are shown in the Figure 1 and Figure 2.

![Figure 1. Smooth filter effect diagram.](image1)

![Figure 2. Wavelet filter effect diagram.](image2)
In Figure 1 and Figure 2, the upper part is the signal with weak SNR and the lower part is the signal with strong SNR. The left part is the filtered signal and the right part is the original signal. The performance of the method in the paper is better than that of wavelet filter.

2.2. The Feature Extraction in the Time Domain and Frequency Domain

Signal feature extraction is a necessary step in signal classification. Signal feature is extracted from two aspects in time domain and frequency domain. [12] In terms of time domain feature extraction, the shape of time-domain signal will change due to digging behaviour. The signal is shown in the Figure 3 and Figure 4.

In the Figure 3 and Figure 4, the noise signal and the digging signal are different. The amplitude of the signal will suddenly increase. The signal will also increase at short time. The short-time zero crossing rate of the signal will increase. The cycle of the signal will change. The number of zero crossing will change. Therefore, the time domain characteristics of the signal include peak valley value, mean value, median, short time energy, short time average crossing zero rate, zero crossing times, up to down peak to average ratio and signal energy.

In terms of frequency domain feature extraction, the energy of frequency-domain signal will change due to digging behaviour. The signal is shown in the Figure 5 and Figure 6.

In the Figure 5 and Figure 6, the noise signal and the digging signal are different. The amplitude of the signal in the frequency domain will have an increase. The ratio of the maximum amplitude to the frequency is shifted and the ratio of the main peak to the secondary peak decreases.
2.3. The Construction and Optimization of Random Forest Algorithm

Random forest is an important classification algorithm in machine learning. [13] It is a classifier that contains multiple decision trees. Its output class is determined by output classes on each tree. In each decision tree, we determine the splitting property and measure according to the sample characteristics. Data samples will be split into sub sets and the category of each subset should be consistent. The training of decision trees in random forests is parallel computing. This will increase the speed of calculation.

The random forest algorithm is optimized by the grid method. [14] The two parameters to be optimized are divided into grids in a certain spatial range. The optimal parameters are searched through the ergodic grid. It traverses all the combinations of the parameters. It is found that its operation time is long. A threshold is set according to the result of the test. The traversal is stopped when the accuracy of recognition is higher than this value. In the words, the features include the frequency, the minimum of the signal, the signal energy, the maximum value of the signal, the maximum amplitude of the frequency domain, the ratio of peak to zero, the median of the signal and the ratio of the main peak value and second peak value.

3. Experiments

3.1. Data Acquisition and Data Processing

(1) Data Acquisition

In the experiment, all the signals were collected from the real environment. In a 14.5m*15.5m lawn, the distributed optical fiber is embedded around the lawn. The depth is about 20cm. Signal acquisition is based on the technology of Phase-OTDR technology. Digging behaviour is an external force to optical fiber. When external force acts on optical fiber, the shape of optical fiber will change. The shape change will lead to the distance and angle of refracted beam propagation in the fiber changes. Finally, the phase of the propagating beam changes in the optical fiber. The device collects optical signals according to the set sampling frequency, sampling range and sampling time. The collected optical signal is converted to the voltage of corresponding intensity. Its unit is volt. The measurement unit of the voltage is mV. After setting up the device parameters, the digging experiments are began and the data is collected. The data is digging signal. Any behaviour is not carried out and the noise signal is collected. The 500 pieces of digging signal and the 1500 pieces of noise signal are collected as the sample of experiments.

(2) Data Processing

Firstly, the abnormal signal points in samples are cleaned and processsed. Secondly, the signal is de-noised by the smooth filtering method. The signal features are extracted in the time domain and frequency domain. The eigenvector of the signal is stored. The specific representation of the features is shown in Table 1.

Table 1. Feature names and symbols.

| No. | Feature Name                                  | Feature Symbol    |
|-----|----------------------------------------------|-------------------|
| 1   | The maximum value of the signal              | peak_signal       |
| 2   | The minimum value of the signal              | trough_signal     |
| 3   | The maximum amplitude of the frequency domain| max_fft           |
| 4   | The frequency                               | fft_signal        |
| 5   | The median of the signal                     | par               |
| 6   | The signal energy                           | power_signal      |
| 7   | The ratio of peak to zero                    | zcr_signal        |
| 8   | The median signal                           | median_signal     |
| 9   | The ratio of the main peak value and second peak value | rate_signal   |
The distribution of feature values is shown in Figure 7.

Figure 7. Digging signal feature distribution.

Figure 8. Noise signal feature distribution.
It can be seen in the Figure 8 that there are great differences between the values of digging signals and noise signals in each feature.

3.2. Classification Testing
The 70% of the sample is treated as a training sample and the other 30% of the sample is treated as a testing sample. The number of training sample is 1400 and the number of testing sample is 600. The specific recognition results are shown in Table 2.

| Signal Category | The Correct Number | The Incorrect Number |
|-----------------|-------------------|---------------------|
| Digging Signal  | 149               | 1                   |
| Noise Signal    | 443               | 7                   |

It can be seen in the Table 2 that the random forest had nice performance. In the 150 sets of digging signals, 149 sets of signals are classified correctly and 1 set of signals are classified wrongly. In the 450 sets of noise signals, 443 sets of signals are classified correctly and 7 set of signals are classified wrongly. In the 600 sets of signal, the recognition accuracy is 98.67% and the recall rate is 95.5%. Experimental results show that random forest algorithm recognition has a good effect on signal classification. It has certain practical application significance.

4. Conclusion
A distributed optical fiber digging signal recognition method based on random forest algorithm is proposed in this paper. It chooses and improves the algorithm of signal de-noising, feature extraction and recognition. The smoothing filtering method can better restore the signal characteristics under the condition of severe noise. So the method is applicable to the actual environment. The random forest classification algorithm with better effect in machine learning is applied to signal recognition. The correct classification results can be obtained by selecting suitable features and parameters. Therefore, this method can effectively remove noise effects, extract feature vectors of signals, accurately identify digging signals and improve the accuracy of recognition.

Acknowledgments
This work was financially supported by the National Key R&D Program of China (No. 2016YFC0701309), National Natural Science Foundation of China (No. 61503034).

References
[1] Taylor H F, Lee C E. Apparatus and method for fiber optic intrusion sensing: U.S. Patent 5, 194, 847 [P]. 1993-3-16.
[2] Bucaro J A, Carome E F. Single fiber interferometric acoustic sensor [J]. Applied optics, 1978, 17(3):330-331.
[3] Juarez J C, Maier E W, Choi K N, et al. Distributed fiber-optic intrusion sensor system [J]. Journal of lightwave technology, 2005, 23(6):2081-2087.
[4] Kishida K, Li T, Lin S. Distributed optical fiber sensor: US, US 7719666 B2 [P]. 2010.
[5] Brunner N, Scarani V, Wegmüller M, et al. Direct Measurement of Superluminal Group Velocity and Signal Velocity in an Optical Fiber [J]. Physical Review Letters, 2004, 93 (20):203902.
[6] Cao J, Fan G. Signal Classification Using Random Forest with Kernels [C]// Sixth Advanced International Conference on Telecommunications. IEEE, 2010:191-195.
[7] Chen S, Sandryhaila A, Moura J M F, et al. Signal denoising on graphs via graph filtering [C]// Signal and Information Processing. IEEE, 2015:872-876.
WAVELET SCALORGRAM [J]. Journal of Sound & Vibration, 2002, 253(5):1087-1100.

[9] Shao Y M. Study and Programming on the Experimental Load Spectrum for the Fatigue of Motorcycle Frame [J]. Journal of Mechanical Strength, 2013.

[10] Ujigawa S, Nakata H, Shimakura S. The relationship between signal intensity and maximum observed frequency in long-distance propagation of VHF radio waves based on the reflection from the ionosphere [J]. Journal of Atmospheric Electricity, 2011, 31(1):63-70.

[11] Paraschivionescu A, Jutten C, Aminian K, et al. Source separation in strong noisy mixtures: A study of wavelet de-noising pre-processing [J]. 2002, 2:II-1681 - II-1684.

[12] Pra A L D, Passoni L I, Sendra G H, et al. Signal Feature Extraction Using Granular Computing. Comparative Analysis with Frequency and Time Descriptors Applied to Dynamic Laser Speckle Patterns [J]. International Journal of Computational Intelligence Systems, 2015, 8(sup2):28-40.

[13] Belgiu M, Drăguţ L. Random forest in remote sensing: A review of applications and future directions [J]. Isprs Journal of Photogrammetry & Remote Sensing, 2016, 114:24-31.

[14] Zhao F, Zhang T, Li K, et al. An Optimized Random Forest Based on Genetic Algorithm and its Application to Feature Selection for Gene Data [J]. Chinese Journal of Health Statistics, 2016.