Bridge anomaly data identification method based on statistical feature mixture and data augmentation through forwarding difference

Yang Qiu¹, Liang Jing¹ and Sheng Li²*
¹School of Information Engineering, Wuhan University of Technology, Wuhan, Hubei, 430070, China
²National Engineering Laboratory for Fiber Optic Sensing Technology, Wuhan University of Technology, Wuhan, Hubei, 430070, China
*Corresponding author’s e-mail: lisheng@whut.edu.cn

Abstract. Identifying abnormal data in the structural health monitoring system is of vital importance for correctly evaluating the structural service status. For the monitored data of a long-span cable-stayed bridge, this paper proposed a method to identify abnormal data, primarily including data augmentation through forwarding difference, and statistical feature hybrid. The average prediction results of the test set showed that the proposed method can significantly improve the classification accuracy of anomaly data compared to directly training the original samples. Besides, the comparison results of the confusion matrix illustrated that the prediction results based on classifiers of random forest and decision tree were more robust, and using the former as the classifier can gain better recognition performance.

1. Introduction
Monitoring and tracking the evolution of the bridge structure is critical to secure the bridge. For a large-scale cable-stayed bridge, vibration signal is an important indicator for diagnosing and evaluating structural service status [1]. Tracking the monitoring results and trends thereof, and judging whether the values are within the safety threshold can be used to evaluate the safety of the structure. However, vibration monitoring data are often affected by interferences, such as the long-term stability of on-site sensors, noises due to signal transmission. Therefore, to accurately evaluate the vibration state of the bridge structure, the abnormal data [2] mixed in the original signal must be identified and eliminated.

At present, there have been similar research reports for sensor fault diagnosis [3, 4]. However, the pattern and size of abnormal monitoring data are still relatively limited compared with actual engineering. In addition, the impact of the sample size imbalance between different patterns is usually ignored in current research. With the improvement of computer hardware performance, the classification problem in structural health monitoring can be handled by machine learning and deep learning technologies. The research reports concerning identifying abnormal data through deep learning based on computer vision were reported in [5]. However, the feature extraction method that transforms time-series signals into images consumes computational resources extremely. Although the feasibility of the artificial intelligence method has been verified for recognizing the abnormal sensor signal, sample label annotation is still lack of automated means when confronting the complex signal pattern of the actual engineering sensor network.
To address the above problems, this paper proposed a method for identifying abnormal monitoring data of a long-span cable-stayed bridge. The abnormal data recognition process forms the second part of the paper, which is followed by detailed information about data augmentation and feature extraction. For an actual engineering monitoring data set, the performance of the proposed method was assessed under four typical classifiers commonly used in machine learning.

2. The abnormal data recognition method

The process method for abnormal data recognition based on supervised learning contained four steps, including data augmentation [6], statistical feature calculation, feature importance ranking, and statistical feature hybrid.

![Diagram of the proposed identification method for abnormal data](image)

The original vibration data set was represented as \( X = \{x_1, x_2, \ldots, x_n\} \), where \( x_i \) denoted the raw data. The forward difference was used to achieve data augmentation of the raw data set. Then, define the differentiated data set as \( X^T = \{x^T_1, x^T_2, \ldots, x^T_n\} \), where \( x^T_i \) denoted the \( x_i \) derived from the forward difference.

To quantify the correlation between different statistical features and the original sample, the importance index of \( m \) types of the statistical features was calculated according to the RF-based permutation importance index (PIM) [7]. Moreover, the importance of statistical features was sorted based on the PIM index, so \( sta_{isort} = \{x_{isort1}, x_{isort2}, \ldots, x_{isortm}\} \), \( sta^T_{isort} = \{x^T_{isort1}, x^T_{isort2}, \ldots, x^T_{isortm}\} \) were obtained.

Finally, the subsets of \( sta_{isort} \) and \( sta^T_{isort} \) were combined as input samples of four classifiers, including the K-neighbor method (KNN), Support vector machine (SVM), Decision tree (DT) and Random forest (RF). Here, the combined inputs were expressed as \( y_n = \{x_{isort1}, x_{isort2}, \ldots, x_{isortk}, x^T_{isort1}, x^T_{isort2}, \ldots, x^T_{isortk}\} \) (\( k \leq m \)), where \( n \) denoted different feature hybrid means.
3. Method verification

3.1. Engineering case description

The experimental data derived from one-month accelerometers of an actual long-span cable-stayed bridge [8] was used to assess the proposed method. The sampling frequency of the data was 20 Hz. Each sample recorded one-hour acceleration information and had a dimension of 1*72000. The labels with 7 different patterns were given in Table 1, in which the sample size was imbalanced for each pattern.

| label | pattern/label name | Size |
|-------|-------------------|------|
| 1     | Normal            | 13575|
| 2     | Missing           | 2942 |
| 3     | Minor             | 1775 |
| 4     | Outlier           | 527  |
| 5     | Square            | 2996 |
| 6     | Trend             | 5778 |
| 7     | Drift             | 679  |

Table 1. Classification and size of data samples

To reveal the original characteristic of seven patterns, two samples randomly selected from each pattern were depicted in Figure 2. From the time-domain waveform of the acceleration signals, some differences in the original samples with the same label can be observed.

3.2. Data set preparation

To avoid the effect of imbalanced sample size distribution on the prediction accuracy of supervised learning, as shown in Table 1, the sample size of Outlier label was seen as a benchmark. For the other six patterns, 527 samples in each label were randomly selected to overcome the imbalance. After the
equalization process, the data augmentation based on the forward difference was performed. Then, the statistical values, including maximum, minimum, mean value, median, standard deviation, range, effective value, mode, kurtosis, and skewness, were calculated respectively. The PIM sorting of ten types of statistical indicators was shown in Figure 3. Here, the top six statistical features were used to replace the \(sta_1\) and \(sta_2\), and they were used as the input sample of the supervised training. Table 2 showed each type of sample input expression with different size statistical features. Before training, the equalized \(X\) and \(X^T\) were divided into the training set \(X_{train}\), \(X^T_{train}\) and test set \(X_{test}\), \(X^T_{test}\) by the division ratio 7:3. Besides, the time series of the original sample was used as a comparison benchmark to compare the classification learning effect under a variety of sample input expressions.

**Table 2. The hybrid input expression of the sample**

| Mark  | Expression of sample input                                                                 |
|-------|------------------------------------------------------------------------------------------|
| \(y_0\) | Time series of the original sample                                                          |
| \(y_1\) | \{\(x_{isort 1}, x_{isort 2}, x_{isort 3}, x_{isort 4}, x_{isort 5}, x_{isort 6}\}\) |
| \(y_2\) | \{\(x_{isort 1}, x_{isort 2}, x_{isort 3}, x_{isort 4}, x_{isort 5}, x_{isort 6}\}\) |
| \(y_3\) | \{\(x_{isort 1}, x_{isort 2}, \ldots, x_{isort 6}, x_{isort 1}, x_{isort 2}, \ldots, x_{isort 6}\}\) |
| \(y_4\) | \{\(x_{isort 1}, x_{isort 2}, \ldots, x_{isort 7}, x_{isort 1}, x_{isort 2}, \ldots, x_{isort 7}\}\) |
| \(y_5\) | \{\(x_{isort 1}, x_{isort 2}, \ldots, x_{isort 8}, x_{isort 1}, x_{isort 2}, \ldots, x_{isort 8}\}\) |
| \(y_6\) | \{\(x_{isort 1}, x_{isort 2}, \ldots, x_{isort 9}, x_{isort 1}, x_{isort 2}, \ldots, x_{isort 9}\}\) |
| \(y_7\) | \{\(x_{isort 1}, x_{isort 2}, \ldots, x_{isort 10}, x_{isort 1}, x_{isort 2}, \ldots, x_{isort 10}\}\) |

### 3.3. Results and discussion

Figure 4 showed the recognition results of the four classifiers, which illustrated that the recognition effect can only reach an accuracy level of 11-15% without data enhancement and feature extraction. After the data processing based on the proposed strategy, the recognition accuracy under various classifiers increased by about 85%. Moreover, the overall average classification accuracy of the four classifiers had the same trend for the different feature hybrid. Further, RF and DT performed better than SVM and KNN with the same input sample expression and the average classification accuracies of the
former two classifiers were about 6% higher. When $y_1$ was selected as the input, the average accuracy of RF and DT for identifying normal data and abnormal data reached 96.11%. However, Figure 5 showed that the two classifiers can only provide 88.69% and 89.88% classification accuracies for normal data. In this situation, Normal label was easily identified as Outlier and Minor labels.

Figure 4 shows that when $y_5$ was selected as the input, the average recognition accuracies of the four classifiers were better than other input expressions. The recognition effect of RF was the best, reaching 97.10%. To further evaluated the recognition effect of the four classifiers on Normal data and six types of abnormal data, Figure 6 provided the results of the confusion matrix for the four classifiers on the test set when using $y_5$ as input. It can be seen that SVM and KNN cannot well balance the recognition accuracy of normal data and 6 types of abnormal data. Besides, the recognition effect of RF and DT was more robust, and the overall effect of the former was better. The recognition accuracies of each type of pattern exceeded 95% when using RF as the classifier combining with the proposed method. Especially, for Drift, Square, and Missing labels, the recognition accuracies were close to 100%. Furthermore, comparing the two input expressions of $y_1$ and $y_5$, the recognition accuracy based on RF for the Normal label increased from 88.69% in Figure 5(d) to 96.84% in Figure 6(d).

Figure 4. The classification accuracy of the four classifiers under different hybrid features

Figure 5. Confusion matrix of the test set under $y_1$ input mode for different classifiers

Figure 6. Confusion matrix of the test set under $y_5$ input mode for different classifiers

![Figure 4](image1.png)

![Figure 5](image2.png)

![Figure 6](image3.png)
4. Conclusion
This paper proposed an abnormal recognition method for the bridge monitoring data, including data augmentation by forwarding difference and statistical feature hybrid through calculating the RF-based PIM. The method can provide ideas to deal with impacts of imbalance of data sets and differences between samples of the same labels on actual engineering, which is beneficial to conducting the recognition classification for abnormal data based on supervised learning.

Acknowledgments
The research work reported in this paper was supported by the National Engineering Laboratory for Fiber Optic Sensing Technology, Wuhan University of Technology. This work was sponsored by the National Natural Science Foundation of China (Grant Number 61875155)

References
[1] Song, G., Wang, C., et al. (2017) Structural health monitoring (SHM) of civil structures. Multidisciplinary Digital Publishing Institute., 8:789.
[2] Yuan, S., Liang, D., et al. (2013) The bridge data diagnosis research based on structural health monitoring system. Journal of University of Electronic Science and Technology of China., 42:69-74.
[3] Yi, T., Li, H., et al. (2016) Detection of shifts in GPS measurements for a long-span bridge using CUSUM chart. International Journal of Structural Stability and Dynamics., 16:1640024.
[4] Liang, D., Zhang, Y., et al. (2011) Application of bridge data novelty diagnosis method for expansion joint monitoring. Journal of Data Acquisition and Processing., 26:579-584.
[5] Tang, Z., Chen, Z., et al. (2019) Convolutional neural network-based data anomaly detection method using multiple information for structural health monitoring. Structural Control and Health Monitoring., 26:2296.
[6] Shao, S., Wang, P., et al. (2019) Generative adversarial networks for data augmentation in machine fault diagnosis. Computers in Industry., 106:85-93
[7] Liaw, A., Wiener, M. (2002) Classification and regression by randomForest. R news., 2:18-22.
[8] Li, S., Li, H., et al. (2014) SMC structural health monitoring benchmark problem using monitored data from an actual cable-stayed bridge. Structural Control and Health Monitoring., 21:156-172.