Vibration fault diagnosis based on Multi-scale EMD time-series similarity mining for hydroturbine

Y G Xue¹, H Wang²

¹ Lanzhou Institute of Technology, Lanzhou 730000, Gansu Province, China
² State Key Laboratory of Eco-hydraulics in Northwest Arid Region of China (Xi’an University of Technology), Xi’an 710048, Shaanxi Province, China

E-mail: xyg3024_cn@163.com

Abstract. Time-series similarity mining is an important method for vibration fault diagnosis of hydraulic turbine. In this paper, based on multiscalar EMD frequency fuzzy nearitude, a time-series similarity data mining algorithm is presented to solve the problem of similarity comparison between characteristic curves of vibration faults. Firstly, all high-dimension deformation data in time-series bank are pretreated by standardized multiscalar EMD. The stationarity of series is promoted and the detailed information is reserved. Then, Discrete Fourier Transformation is carried out for the obtained multiscalar IMF. Finally, the distances among time series are measured with fuzzy nearitude of IMF component series. The degree of similarity among time series is also described. To test its effectiveness, the method is applied to the prototype hydraulic turbine vibration fault series. As its result shows, multiscalar EMD tranquilizes the complex non-stationary time series. It conquers the problem of information loss in the process of data interception. At the same time, the method can help identify unit faults accurately, and classify different types of faults. Its discriminant accuracy is about 82.9%. Due to its low requirement for the number of data, and the efficiency in computing, the method is suitable for large-scale graphic series mining in hydraulic turbine fault diagnosis.

1. Introduction

When the hydraulic turbine is running, the shaft generates various vibrations. In general, monitoring data of the vibration status is a time-series. As the monitoring time increases, many large-scale data sets are formed[1-3]. The study of time-series similarity model mainly uses some certain specific algorithms to extract valuable knowledge from the data with the certain computing efficiency constraints[4]. Since the time series database often contains many kinds of deformation data, such as incompleteness, noise, high-dimensional and non-stationary data, directly using these data for mining is inefficient. Therefore, preprocessing data by standard is an important part of the time-series similarity search.

A common approach in the time-series similarity search is to compare the direct Euclidean distance of the time series or the time warping distances of the two sequences, describing the degree of similarity[5-7]. So far, there have been a variety of time-series similarity mining methods, such as Agrawal Rakesh et al.[8] proposed Discrete Fourier Transform (DFT) and its improved method, using DFT in the time domain and frequency domain of the Euclidean distances are equal. The dimension reduction is achieved by preserving the first few coefficients of the DFT, and the similarity of the
sequences is measured by the Euclidean distance. However, in the process of data interception, this method smoothes the local extrema of the original sequence and rejects high frequency components of the signal, resulting in the loss of important information. In addition, DFT is only an analysis method for stationary sequences, and it is not suitable for non-stationary sequences.

In the time-series of hydraulic turbine vibration condition monitoring, the sequence presents different manifestations due to different amplitudes and measuring points. It is impossible to describe the similarity of the two time-series with the traditional Euclidean distance and time bending distance. The Empirical Mode Decomposition (EMD) method is based on the local feature time scale of the original sequence. It decomposes the sequence into several Intrinsic mode functions (IMF) at the multiple scales, disintegrating all kinds of the multi-frequency interconnected component signal sequence. It has good adaptive multi-resolution characteristics. Because similar turbine faults have similar vibration frequencies, frequency characteristics can be used as a measure. Based on the above research, this paper proposes a time-series similarity data mining algorithm based on multiscale EMD frequency fuzzy nearitude. The method is simple, fast and does not require special time-series constraints, and can truly reflect the relationship between vibration characteristics. It is suitable for turbine vibration fault diagnosis under large-scale massive data sets.

2. Theoretical model

2.1. Time-series similarity and measurement

Let \( x = \{x_1, x_2, \ldots, x_n\} \) be a time-series. \( x_1, x_2, \ldots, x_n \) are the values on time axis \( t_1, t_2, \ldots, t_n \) respectively. \(|x|\) is the observation length of sequence \( x \). \( TB = \{x_1, x_2, \ldots, x_n\} \) is a time series database containing the sequence \( x_1, x_2, \ldots, x_n \). Then the time-series similarity problem can be expressed as: given a query sequence \( q \) and time sequence library \( TB \), similarity measure function \( \sim \cdot \), similarity search strategy \( \text{find}(\cdot) \), similarity search is to find the set of sequences \( R \) that is similar to \( q \) in the library \( TB \), which is:

\[
R = \{x \in TB | \text{find}(\sim(q, x), TB)\}
\]  

(1)

In the similarity measure, Euclidean distance is the most widely used. But since the Euclidean distance and the length of the sequence are not normative, similarity between different length sequences cannot be calculated, and it is difficult to specify the query distance. Moreover, each dimension in a multidimensional space is independent each other. It is easy to lose the mutual relationship between sequence elements, when calculating the distance. Determining the similarity failure according to the change pattern of the sequence.

2.2. Empirical model decomposition

The empirical mode decomposition algorithm can decompose the fluctuations or trends of different scales or frequencies step by step, resulting in a series of narrowband stationary data sequences with different feature scales called the intrinsic mode component functions. The process is as follows:  

1. Determining the local maxima sequence \( x_{\max} \) and local minimum sequence \( x_{\min} \) of the signal.  
2. According to \( x_{\max} \), \( x_{\min} \), it can determine the upper and lower envelopes of the signal \( x(t) \) and the local mean \( m(t) = \frac{x_{\max} + x_{\min}}{2} \).  
3. \( h(t) = x(t) - m(t) \) is the difference between the signal and the local mean.  
4. Due to the existence of asymmetric waves in nonlinear and non-stationary signals, \( m(t) \) is not a true local mean.

Therefore, instead of \( h(t) \) with \( x(t) \), cycle through the above steps until \( h_{k+1} \) satisfies the basic conditions of the IMF. Let \( c_1 = h_{ik} \), \( r_i = x(t) - c_1 \), available:
\[ x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \] (2)

Where \( c_i(t) \) is the EMD's decomposed IMF, and \( r_n(t) \) is the signal's residual trend item. At this point, \( x(t) \) is decomposed into \( n \) IMF components and the sum of trend residuals by EMD.

Theoretically, the IMF divides a group of narrow-band stationary sequences with high and low local order frequencies in strict accordance with different feature scales of the sequence. After the signal sequence is decomposed by EMD, the IMF components at different scales are obtained. This process is equivalent to the decomposition of the non-stationary original sequence into a smooth segmented sequence, while the details of the information are better preserved. Within any of the feature-scale intervals, changes in the signal sequence can be ascertained at several different scales. If each dimension is a source, the IMF component at each scale corresponds to the information sent by a source. With the multi-scale nature of empirical mode decomposition, the signal sequence is standardized and preprocessed, the feature of the component sequence is extracted at different scales, multi-scale similarity mining is performed, and similar time series or sub-sequences are searched.

### 2.3 Frequency fuzzy nearitude time-series similarity measurement algorithm

When a hydraulic turbine fails, its symptoms are mostly manifested in the frequency and energy of the signal sequence. The signal sequence of each frequency component contains a wealth of fault information. Therefore, the vibration time series similarity search can first decompose the vibration time series EMD, and then compare the obtained segmented multi-scale IMF sequences in the frequency domain to solve the similarity search between the vibration time series. That is to say, the similarity degree of the two sequence is excavated by comparing the approximate degree of the segmented IMF frequency domain. In the similarity degree, the closeness of the lattice in the fuzzy set is introduced to describe the distance relationship between two sequences.

Before mining two time series, we first carry out DFT for each component IMF sequence to be compared\(^{[12]}\), transform the time-domain sequence into frequency domain sequence. Setting the length of IMF sequence \( X \) is \( T \), and the sampling interval is \( \Delta t \). There are:

\[
X(k\Delta f) = \sum_{r=0}^{N-1} x(r\Delta t)e^{-j2\pi kr/N} \left( \frac{T}{N} \right) \quad (3)
\]

\[
x(r\Delta t) = \sum_{k=0}^{N-1} X(k\Delta f)e^{j2\pi kr/N} \left( \frac{1}{T} \right)
\]

In view of the substitution relationship between the two, formula (3) can be transformed into the following equivalent formula:

\[
X_k = \frac{1}{N} \sum_{r=0}^{N-1} x_r e^{-j2\pi kr/N} \quad (4)
\]

\[
x_r = \sum_{k=0}^{N-1} X_k e^{j2\pi kr/N} \quad (5)
\]

After transforming the sequence, remove the high frequency information. The membership frequency of the feature frequency is obtained through the grid-fixed progress fuzzy of the fault feature frequency:

\[
u_k = x_k / \sum_{i=1}^{n} x_i \quad (6)
\]

Where \( n \) is the number of fault feature frequencies, \( k \) is each fault feature corresponding to each sequence, and \( u \) is the degree of membership at the specified feature frequency. The two time series similarity calculations are transformed into two time series discrete Fourier transform segments of the multi-scale IMF. Then the frequency of the fault features is blurred and the closeness of the two
fuzzy sets is compared. Let the time series $X$ and $Y$ transform the frequency domain sequence. After fuzzification, these two parameters are changed to $A$ and $B$, also $A, B \in R$. The outer product between them is defined as $A \circ B = \bigvee_{r \in R} (A(r) \wedge B(r))$. The inner product is defined as $A \wedge B = \bigwedge_{r \in R} (A(r) \vee B(r))$. The lattice closeness of the two sets is

$$N(A, B) = (A, B) = (A \circ B) \wedge (A \wedge B) \quad (7)$$

At this time, the distance of the segmented IMF sequence is described as:

$$L(X, Y) = N(A, B) = (A \circ B) \wedge (A \wedge B) \quad (8)$$

From the definition of the closeness degree, we can see that the closer $X$ and $Y$ are, the closer $N(A, B)$ is to 1 and the larger $X$ and $Y$ are, the closer to 0.

From the above theoretical analysis, we can see that the nature of the time-series similarity mining algorithm based on multi-scale EMD is to excavate the similarity relationship between the original sequences by using the similarity of the stationary IMF sequences in the single-scale single frequency segment after decomposing.

3. Hydraulic turbine vibration sequence EMD multi-scale similarity mining

3.1 Vibration sequence sampled collection and preprocessing

During the application test, a set of rotor imbalance failures and five sets of other oscillation conditions were collected by the monitoring system. These include the upper and lower guide swings and the upper bracket vibration signal sequence. The unit is the No. 2 turbine unit of Xinjiang Tajik Power Station. The type of the power station unit is HLA801-LJ-215, SF24.5-20/4250, the rated speed is 300 r/min (5 Hz), and the rated power is 24.5MW. Working head is 74 m. The measured signal is collected under the hydropower unit rated output of approximately 35% ~ 50% (8.5~12.25MW operating conditions), and the sampling frequency is 400 Hz. The eddy current sensor is used to measure the swing of the major axis in the upper and lower guides, and the water guide +X/-Y direction. The key phase sensors for measuring the rotational speed are set in the water guide -Y direction. The upper rack, lower rack and head cover (oil basin covered with water guide) set vertical and +X/-Y horizontal vibration measurement points.

This sample contains 6 sets of time series data, where sample 6 is the time series of rotor imbalance failures. What needs to be emphasized is that due to the effective amplitude values of the samples are not the same, the Euclidean distance cannot be used to characterize the similarity of time series. In addition, different sampling frequencies and different sampling points make it impossible to calculate the similarity by the time bending distance. Since the frequency characteristic of the graph shows a certain vibration failure characteristic, the approximation in the time domain is determined by comparing the degree of approximation in the frequency domain.

It is evident from the above six figures that by using multi-scale EMD to preprocess the high-dimensional deformation data, can improve the stationarity of the sequence, preserve the details of information, thus overcoming the frequency aliasing of the non-stationary sequence DFT process. Therefore, in any feature-scale interval, the change of the signal sequence can be determined at a plurality of different scales. After obtaining the frequency-domain characteristics through the DFT, the degree of membership in each frequency domain is calculated according to Equation (6). Select the $1/6~1/2$ (average value) of frequency doubling, frequency doubling, double frequency, 3 frequency multiplier, 50Hz frequency, 100Hz frequency membership degree to compare, as shown in table 1.
Figure 1. Time series EMD and frequency transform of No.1.

Figure 2. Time series EMD and frequency transform of No.2.

Figure 3. Time series EMD and frequency transform of No.3.
Figure 4. Time series EMD and frequency transform of No.4.

(a) Sample sequence 4 and multiscale decomposition of EMD

(b) Frequency transform of IMF

Figure 5. Time series EMD and frequency transform of No.5.

(a) Sample sequence 5 and multiscale decomposition of EMD

(b) Frequency transform of IMF

Figure 6. Time series EMD and frequency transform of No.6.

(a) Sample sequence 6 and multiscale decomposition of EMD

(b) Frequency transform of IMF
Table 1. Characteristic frequencies fuzzy subjection degree of samples.

| Sample No. | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\mu_4$ | $\mu_5$ | $\mu_6$ |
|------------|---------|---------|---------|---------|---------|---------|
| 1          | 0.0117  | 0.1596  | 0.0506  | 0.0623  | 0.0563  | 0.0298  |
|            | 0.0537  | 0.0921  | 0.1032  | 0.0536  | 0.0834  | 0.0214  |
|            | 0.1163  | 0.0834  | 0.0214  | 0.0181  | 0.0290  | 0.0204  |
|            | 0.0458  | 0.0318  | 0.0484  | 0.0410  | 0.0512  | 0.0101  |
| 2          | 0.0110  | 0.1339  | 0.0427  | 0.1011  | 0.0273  | 0.0175  |
|            | 0.0160  | 0.0975  | 0.0817  | 0.0799  | 0.0618  | 0.0210  |
|            | 0.1026  | 0.1027  | 0.0275  | 0.0735  | 0.0785  | 0.0672  |
|            | 0.0197  | 0.1358  | 0.0977  | 0.0925  | 0.0351  | 0.0449  |
| 3          | 0.0196  | 0.0931  | 0.0493  | 0.1559  | 0.0279  | 0.0257  |
|            | 0.0169  | 0.1345  | 0.0672  | 0.0928  | 0.0379  | 0.0527  |
|            | 0.0114  | 0.0735  | 0.0694  | 0.0785  | 0.0669  | 0.0103  |
|            | 0.0108  | 0.0664  | 0.1081  | 0.0992  | 0.1514  | 0.0150  |
| 4          | 0.0175  | 0.1386  | 0.0439  | 0.0719  | 0.0411  | 0.0313  |
|            | 0.1007  | 0.0743  | 0.0399  | 0.0294  | 0.0162  | 0.0102  |
|            | 0.0126  | 0.1026  | 0.1184  | 0.0896  | 0.0619  | 0.0254  |
|            | 0.0924  | 0.0437  | 0.0257  | 0.1393  | 0.0597  | 0.0311  |
| 5          | 0.0399  | 0.0560  | 0.0519  | 0.0564  | 0.1590  | 0.0159  |
|            | 0.0104  | 0.0907  | 0.1027  | 0.0947  | 0.0629  | 0.0201  |
|            | 0.0495  | 0.0526  | 0.0597  | 0.0537  | 0.1627  | 0.0150  |
|            | 0.0216  | 0.0775  | 0.0187  | 0.0249  | 0.0118  | 0.0231  |
| 6          | 0.0145  | 0.0597  | 0.1152  | 0.0868  | 0.1386  | 0.0198  |
|            | 0.0209  | 0.0891  | 0.1027  | 0.0372  | 0.0494  | 0.0399  |
|            | 0.0264  | 0.1212  | 0.0345  | 0.0716  | 0.0768  | 0.0142  |
|            | 0.0204  | 0.1007  | 0.1027  | 0.0847  | 0.0619  | 0.0211  |

3.2 Hydraulic turbine vibration sequence similarity mining

The time sequence under a known fault condition is shown in figure 7. The first row of figure 7. (a) is the original sequence, the last few lines are the IMF sequence waveforms without trend residuals, and figure. 7(b) is the IMF frequency spectrum at each scale.

![Figure 7](image)

(a) Fault sample sequence and multi scale decomposition of EMD

(b) Frequency transform of IMF

Select the 1/6~1/2 (average value) of frequency doubling $\mu_1$, frequency doubling $\mu_2$, double frequency $\mu_3$, 3 frequency multiplier $\mu_4$, 50Hz frequency $\mu_5$, 100Hz frequency $\mu_6$ membership degree to compare, as shown in Table 2.

Table 2. Characteristic frequencies fuzzy subjection degree of fault sequence

| fault sequence | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\mu_4$ | $\mu_5$ | $\mu_6$ |
|----------------|--------|--------|--------|--------|--------|--------|
| sample         | 0.0503 | 0.1723 | 0.0836 | 0.0687 | 0.0210 | 0.0011 |
|                | 0.0362 | 0.1052 | 0.0139 | 0.0016 | 0.0054 | 0.0103 |
|                | 0.0834 | 0.0241 | 0.0079 | 0.0236 | 0.0034 | 0.0027 |
|                | 0.0659 | 0.0148 | 0.0064 | 0.0214 | 0.0026 | 0.0021 |

In order to make the close degree not exceed 1, the close degree of Haiming is adopted with the
data characteristics is combined, it is changed to Haiming distance:

\[ E(A, B) = \frac{1}{n} \sum_{i=1}^{n} |A(u_i) - B(u_i)| \]  

(9)

After calculating the close degree of the IMF component sequence of each scale, the similarity degree of each component is merged into the original sequence close degree, and the comparison is made. Then, the final closeness is obtained as:

\[ N_i(A, B) = 1 - \frac{\sum_{j=1}^{4} E_{ij}^2(A, B)}{\sum_{j=1}^{4} E_{ij}^2} \]  

(10)

Where \( i = 1, 2, ..., 4 \), \( j = 1, 2, ..., 4 \).

The lattice close-degree of the frequency of the fault sequence is calculated as shown in Table 3 below.

| No. | \( E_1(A, B) \) | \( E_2(A, B) \) | \( E_3(A, B) \) | \( E_4(A, B) \) | \( N(A, B) \) |
|-----|----------------|----------------|----------------|----------------|----------------|
| 1   | 0.0258         | 0.1701         | 0.0258         | 0.0259         | 0.4077         |
| 2   | 0.0290         | 0.1509         | 0.0512         | 0.0175         | 0.5260         |
| 3   | 0.0402         | 0.1024         | 0.0227         | 0.0197         | 0.6222         |
| 4   | 0.0318         | 0.0927         | 0.0194         | 0.0236         | 0.6317         |
| 5   | 0.0308         | 0.2101         | 0.0634         | 0.0165         | 0.4436         |
| 6   | 0.0443         | 0.0518         | 0.0601         | 0.0748         | 0.9627         |

From Table 3, it can be seen that the vibration sequence in the similarity mining is closest to the fault sample of No. 6, the result is 0.9627, and the close degree of other faults does not exceed 0.64. Through the excavation, it was found that the turbine vibration sequence may be the same as that contained in Sample of No. 6, which is the data collected from the unbalanced rotor fault condition of Unit 2 in Tajik Power Station. Therefore, the turbine vibration sequence contains rotor imbalance error information, the set should re-select the balance weight.

To further verify the validity and accuracy of the analysis, the fuzzy average closeness of the total close degree of the first five groups in Table 3 is calculated to be 0.5262, and the ratio of it to the calculated fuzzy average closeness of the total closeness of 0.9627 in the sixth group is 82.9%. The results show that lattice closeness algorithm using time series in the frequency domain is more accurate in identifying faults. The algorithm can also identify various fault types and find out the possibility of faults.

4. Conclusion

Aiming at the similarity comparison of hydraulic turbine vibration fault characteristic curves, this paper proposes a method of frequency fuzzy nearitude time-series similarity mining based on empirical mode decomposition. This method uses the EMD multi-scale decomposition feature to normalize and pretreat turbine vibration time series. It performs discrete Fourier transform on the IMF under each scale of decomposition, and uses the fuzzy closeness of the IMF component sequence to measure the distance between the original time series. Finally, it can describe the degree of similarity between time series.

The multi-scale EMD method smoothes the time series under the premise of adequately retaining the details of the sequence. At the same time, the fuzzy close degree algorithm in the frequency domain overcomes the constraints on the similarity mining by the length of the sample data. Take the prototype turbine vibration fault sequence as an example and apply the test. The results show
that this method can accurately identify the unit faults, occupy less storage space, and has high computational efficiency. It is suitable for the mining of the graphic sequence in turbine vibration fault diagnosis.

Acknowledgement
The research was supported by The National Natural Science Foundation of China (51769012).

References
[1] Xue Yangang, Wan Han, Luo Xingqi, et al. Investigation on turbine vibration signals based on down-sampling Hilbert-Huang transform [J]. Journal of Xi’an University of Technology, 2010, 26(2): 136-141.(in Chinese with English abstract).
[2] Bai Liang, Wang Han, Li Hui, Luo Xing-qi. Vibration Fault Diagnosis Based On Time-Series Similarity Mining for Hydropower Units [J]. Journal of Hydroelectric Engineering, 2010, 29(6): 232-239.(in Chinese with English abstract).
[3] Zhu Wenlong, Zhou Jianzhong, Xiao Jian. An ICA-EMD feature extraction method and its application to vibration signals of hydroelectric generating units [J]. Proceeding of the CSEE, 2013, 33(29): 95-101 (in Chinese with English abstract).
[4] Robert H. Shumway, David S. Stoffer. Time series analysis and its applications [M]. Beijing: World publishing corporation, 2009.
[5] Francesco Gullo, Giovanni Ponti, Andrea Tagarelli, et al. A time series representation model for accurate and fast similarity detection [J]. Pattern Recognition, 2009, 42(11): 2998-3014.
[6] Bachir Boucheham. Reduced data similarity-based matching for time series patterns alignment [J]. Pattern Recognition Letters, 2010, 31(7): 629-638.
[7] Li Mintong, Yang Qing, Song Meng. Fault diagnosis of diesel engine based on energy and time frequency domain characteristics of vibration signals IMF[J]. Transaction of the Chinese Society of Agricultural Engineering, 2012, 28(21): 37-43(in Chinese with English abstract)
[8] Agrawal Rakesh, Faloutsos Christos, Swami Arun, Efficient similarity search in sequence databases[C]. Proc. Of the 4th Conference on Foundations of Data Organization and Algorithms, Chicago, October, 1993.
[9] Norden. E. Huang, Z. Shen, S. R. Long, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceeding of the Royal Society of London Series A-Mathematical Physical and Engineering Sciences, 1998, 454(3): 903–995.
[10] Norden. E. Huang, Wu Zhaohua. A review on Hilbert-Huang transform: Method and its applications to geophysical studies [J]. Reviews of Geophysics, 46(2): 1-23.
[11] Li Juan, Zhou Donghua, Si Xiaosheng. Review of incipient fault diagnosis methods [J]. Control Theory & Applications, 2012, 29(12): 1517-1529(in Chinese with English abstract).
[12] Yang Shuzi, Wu Ya. Time series analysis in engineering application[M]. Huazhong University of Science and Technology Press, 1991 (in Chinese with English abstract).