Network Traffic IMF Selection Method Based on EMD

Xiuli Du[1,2, a], Yasheng Sun[1,b] and Shaoming Qiu[2,c]
[1]Communication and Network Laboratory, Dalian University, Dalian, 116622, China
[2]College of Information Engineering, Dalian University, Dalian, 116622, China
[a]22811623@qq.com, [b]sys_job@126.com, [c]17896015@qq.com

Keywords: EMD, average mutual information, nonlinear correlation coefficient, IMF selection.

Abstract. Empirical mode decomposition is an advanced time-frequency analysis method, but it will produce a few invalid components in the process of decomposition. For this problem, the paper proposes a nonlinear correlation coefficient method that is based on the average mutual information. The method can get different the Intrinsic Mode Function by using EMD decomposition of the original signal, and then calculate the nonlinear correlation coefficient about each Intrinsic Mode Function (IMF) and the original signal, and find invalid component through comparing with the threshold. The experimental result shows that the method is a more effective way to select the IMF, which can obviously distinguish the real components and the invalid components.

Introduction

In 1998, Huang et al. from NASA proposed a new signal processing method - empirical mode decomposition (EMD) [1]. With step screening, the method decomposes the different fluctuations or trends of the signal from large to small according to the frequency, resulting in a series of IMF-scale components with different characteristics. EMD is effective method to analyze the non-linear, non-stationary data, which has been widely used in seismic signal, structural analysis, mechanical fault diagnosis and speech enhancement and other fields. However, due to the limitations inevitable of end effects and filtering criteria, the low-frequency part of the intrinsic mode function decomposed by EMD will generate surplus invalid component, and the high-frequency part will generate the noise component, so as to impact the real component selecting [2].

Currently, for the problem of removing invalid component from the IMF, Lin et al proposed the improved algorithm of EMD [3]. They removed the invalid component with the correlation coefficient between IMF and the original signal. The results show that the method has good adaptability, but the differential is small and has not obvious different, especially when the coefficient difference in the vicinity of the threshold value boundary is less, the method may get wrong judgment. Han et al proposed the invalid component identification which is based on K-L divergence method, the method calculate K-L (Kullback Leibler divergence, KL) divergence values between the original signal and the IMF components, then identify the fake components according to the divergence values[4]. The results show that K-L divergence method has large differential. But its adaptability is poor, and the computational is complexity. Besides, the method includes the kernel function selecting. In order to solve it effectively, a IMF selection method based on the average mutual information is proposed, and verified by obtaining network traffic signal. The results show the adaptability and effectiveness of this method.

The IMF Invalid Component Cause Analysis. The empirical mode decomposition determine the envelope of the original signal by using spline curves, while the endpoints are not always the extreme points in the fitting process of the envelope, it is difficult to ensure the accuracy of the interpolation near the endpoint, the endpoint will have the inevitable divergence phenomenon. It will produce excess invalid components in the low frequency portion and generate noise components in the high frequency portion which will have an influence on the real signal selecting[4].

IMF Selection Method Based on the Average Mutual Information

Nonlinear Correlation Coefficient and Its Algorithm. In order to be more convenient and intuitive to view the correlation between two variables, making it have the advantage that the average mutual information can be sensitive to the relationship between two variables, and like the correlation coefficient to use the closed interval [0, 1] to characterize the strength of the degree of correlation, where 0 represents the weakest correlation, 1 indicates the strongest correlation, we improve the definition of the average mutual information. The correlation coefficient is defined as:

$$I^r(X;Y) = H^r(X) + H^r(Y) - H^r(X,Y)\ . \tag{1}$$

Further inference can be drawn, the nonlinear correlation coefficient:

$$I^r(X;Y) = 2*(-\sum_{i=1}^{a}\frac{1}{a}\log_a\frac{1}{a}) + \sum_{i=1}^{a}\sum_{i=1}^{a} p_{ij}\log_a p_{ij}\ . \tag{2}$$

Where p_{ij} is N data of probability distribution in the two-dimensional state, observation of the formula, you will find the data N of probability distribution of the two-dimensional matrix can represent the relationship between two variables in statistical sense [9][10].

IMF Selection Process. The method mainly measures the relevance about the IMF component and the original signal by nonlinear correlation coefficient. The smaller coefficient, the less relevant, the IMF components is invalid, which should be ruled out; the larger coefficient, the more relevant, the IMF component will need to reserve. The specific steps are as follows:

i) The method gets several IMF components by EMD decomposing the original signal;

ii) Calculating the nonlinear correlation coefficient about each IMF component and the original signal;

iii) Setting a reasonable threshold d, if the nonlinear correlation coefficient is less than d, it will be thought the IMF component to be invalid component.

Especial attention, threshold d will have certain changes in different situations, so it is necessary to adjust the threshold d in dealing with a different signal. The threshold d is set up 2.0 in this paper[5].

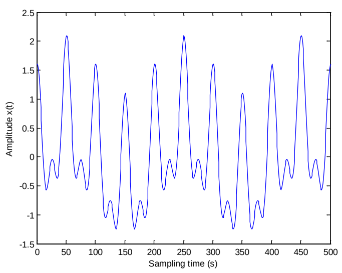
The Simulation Analysis

In order to prove the IMF selection method of nonlinear correlation coefficient which is based on the average mutual information is correctness, here it sets a simulation signal, and its expression is:

$$X(\mathrm{t}) = \cos(\frac{\pi}{2.5}\times t) + 0.6\times\cos(\frac{2\times\pi}{2.5}\times t) + 0.5\times\sin(\frac{\pi}{10}\times t)\ . \tag{3}$$

The signal is made up of three compounded cosine signals, the waveform is shown in figure 1.

Firstly, using the EMD decomposition for overlapping signal, suppressing the end effect with the method of waveform matching endpoint continuation, the result is shown in figure 2. It got the coefficient size of the original signal and each IMF component through using the method of nonlinear correlation coefficient which is shown in table 1.



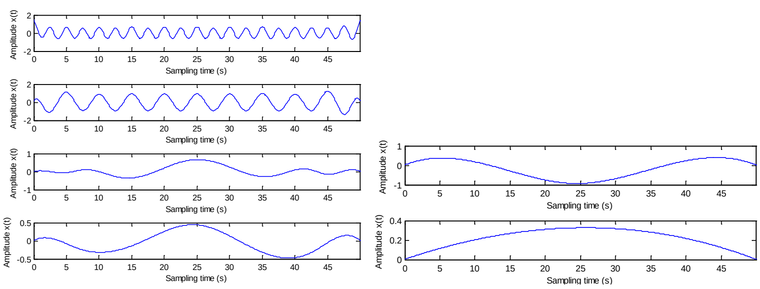


Fig.1: The simulation waveform   Fig. 2: The result of the EMD decomposition original signal

Table 1: The nonlinear correlation coefficient

IMF1	IMF2	IMF3	IMF4	IMF5
0.7680	0.6872	0.5620	0.1023	0.0698

From table 1, it shows that the nonlinear correlation coefficient of the top three IMF component is greater than the threshold d=0.2. so we think the component is the real component of original signal, While IMF4 and IMF5's coefficient is smaller than the threshold d, namely as invalid component, which accords with the actual approval. so, using nonlinear correlation coefficient method can choose the effective component of the IMF appropriately , removing the invalid component, which verified the selection effective of the low-frequency IMF component.

Following, the original signal adds the white gaussian noise whose mean is 0 and variance is 1, and the waveform of which added with noise is shown in figure 3. The figure 4 is the result of EMD decomposition.

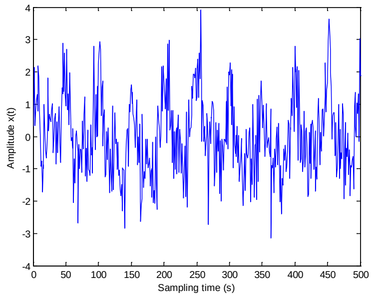


Fig. 3: The add noise of the original signal

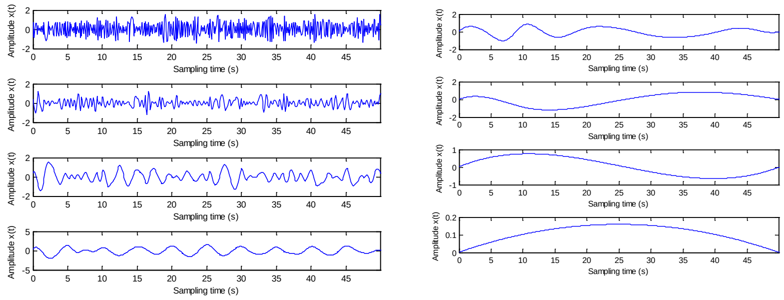


Fig. 4: The result of the EMD decomposition network signal

The same, the coefficient about the original signal and each IMF component by using nonlinear correlation coefficient method are also shown in table 2.

Table 2: The nonlinear correlation coefficient of Each IMF and the original signal

IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7
0.1202	0.1032	0.5320	0.6875	0.6210	0.1102	0.0901

It can be seen that the nonlinear correlation coefficient of IMF1 and IMF2 is less than the threshold d from the table 2. So it could be concluded that the IMF1 and IMF2 are invalid noise component of the IMF. The results of simulation show that the nonlinear correlation coefficient not only can remove the low frequency invalid component but also can be a good choice for high frequency.

We continue to verify selection effect of the nonlinear correlation coefficient method to the real network signal. The time interval is 0.4s in this paper. We collect 500 data totally. Here, the original flow should be EMD decomposed firstly. The original network signals is a kind of non-stationary random signal In figure 5.

To getting each IMF component, we use the EMD method to decompose the network traffic signal, it shows as figure 6.

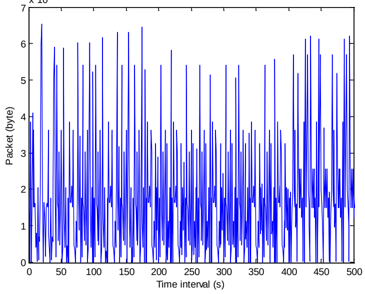


Fig. 5: The network signal

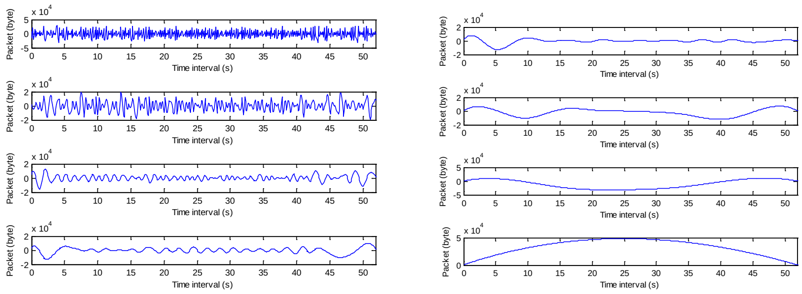


Fig. 6 The result of the EMD decomposition network signal

For comparing and analyzing, we use the correlation coefficient method, K - L divergence value method and the nonlinear correlation coefficient method to identify the invalid IMF signals, the

Comparison results is shown in table 3. (Correlation coefficient, CC; K-L divergence, K-L; nonlinear correlation coefficient, NCC)

Table 3: The IMF choose contrast table

Selection method	IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7
CC	0.086	0.5952	0.8964	0.7852	0.1235	0.0928	0.0628
K-L	0.0021	0.0002	0.0025	0.0012	0.0054	0.3326	0.3126
NCC	0.4368	0.6215	0.5662	0.6852	0.5124	0.1045	0.0625

For K-L divergence value method, it's threshold value is 0.01, the IMF components whose value is less than the threshold value are effective IMF components. We can obtain from the table 3 that the IMF1, IMF2, IMF3 and IMF4 are effective IMF components. At the same time, we can get the same conclusions through the using of nonlinear method of correlation coefficient. For the method of correlation coefficient, we usually set the threshold value as 1/10 of the maximum correlation coefficient, through this method, we find that IMF2、IMF3、IMF4、IMF5 and IMF6 are all effective components, it shows that this method gets the wrong calculation when the threshold value is around 0.089.

In summary, Three methods verify the real network signal to find, nonlinear correlation coefficient method can more effectively identify the effective IMF components and has stronger adaptability by compared with the K-L divergence method. so, the nonlinear correlation coefficient that is based on the average mutual information is obviously better than correlation coefficient and own more adaptability and simple calculation by comparing the K-L divergence method.

Summary

The paper proposes a nonlinear correlation coefficient method based on the average mutual information and identifies the invalid components by calculating nonlinear correlation coefficient about the IMF component and the original signal. In the end ,the nonlinear correlation coefficient method can choose the effective IMF component by verifying the real network traffic signal to prove the accuracy of this method and also further verify that the method has better adaptability by testing the add noise network signal.

Acknowledgements

The research work was supported by Space Information Network Routing and Transmission Protocol Optimization Theory and Technology Research: the National Natural Science Research Foundation Major Research Projects under Grant No. 91338104.

References

[1] D.Pawel, M. Miroslaw, Computer network traffic analysis with the use of statistical self similarity factor. Annales UMCS, Information,13(2),pp.69-81,2014.

[2] Lin Li, Yu Lun, Improvement on Empirical Mode Decomposition Based on Correlation Coefficient. Computer and digital engineer, 36(12), pp.28-30, December 2008.

[3] Han Zhong-He, Zhu Xiao-Xun, A false component identification method of emd based on kullback-leibler divergence.proceedings of the CSEE, 32(11), pp.112-116, 2012.

[4] Huang Xian-Xiang, Xie Jian, New approach to dealing with the end effect of empirical mode decomposition. Chinese Journal of Mechanical Engineering, 44(9),pp.1-5,2008.

[5] Han Jie-Ying, Network traffic anomaly detection using EMD and Hilbert-Huang transform. Western Carolina University,2013.