The study on denoising model of time series in big data

Xiaoming Guo[1,a], Xingwang Zhang [2], Jianming Cui [3]

[1] *Guilin University of Technology ,Guilin, 541004,China*

[2] *Guilin University of Technology ,Guilin, 541004,China*

[3] *Guilin University of Technology ,Guilin, 541004,China*

[a]*gxm8389@163.com*

Abstract.

A model which includes wavelet analysis and windows Fourier transform has been designed to resolve huge time series and interferential data in big data. In the model, the mass data firstly has been clustered as static or dynamic data. The static data has been processed by windows Fourier transform, and the dynamic data has been processed by wavelet analysis. After testing and simulation, the computing speed and denoising effect have been improved.

Keywords: big data, windows Fourier transform, wavelet analysis, denoising, cluster

The time series in big data

The big data, also called as large or huge data, refers to be so much that it can't be obtained, managed, processed and integrated in reasonable time into useful message which would help enterprises to make some marketing or business decision. When the big data has been known by us, four characters which include Volume, Velocity, Variety and Veracity are called 4V for short.

Generally speaking, time series refers to a related data set which has been arranged by time in big data time. Time series is a kind of complicated data object, and it exists as one data form in all kinds of databases from business

system, medical treatment, building works and social science. Time series has an great effect on each field of our life such as share price, foreign exchange rate, merchandise sales, service, weather data and so on. Large time series could truthfully record all information happened at each time.If there is a modified and efficient data process method to find relation of time series, it will increase widely awareness and understand in order to forecast and control effectively to improve our present condition. How do we could do this? Time series data mining supply us new method. TSDM refers that some useful information should be extracted from large time series to guide people's activities. TSDM has a great significance on human society, development of technology and economy.

Processing the noise in time series

Usually, the time series obtained directly is useless, and we need to get rid of useless and noise part. The normal way to optimize time series has usually been adopted as picture 1. Althoughthe processing model in figure 1 contains data pretreatment, the noise in data has not been get rid of as we expect.

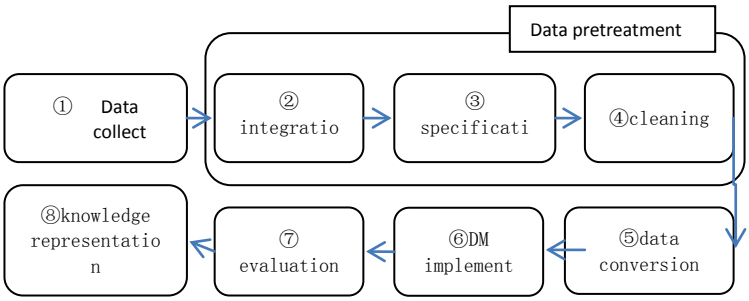

figure1. TSDM

The normal way of denoising is windows Fourier transform or wavelet analysis, as is shown in figure 1. However, the normal way has some disadvantages. For example, it is difficult for Fourier transform to separate low and high frequency signals, and it is more worse that the high frequency signal would be filtered with noise together. In figure 2, the sample, which fluctuates in [0,2], has been mixed into 6.5 noise, and then the signal would has much distortion, especially

the high-frequency signal. In figure 3, to the signal whose frequency is [0,2],SNR is 6.5, more noise has been retained in low-frequency, when the same threshold has been adopted to process signals with different layering. It happens even if the way of wavelet analysis could avoid the problem the Fourier transform causes.

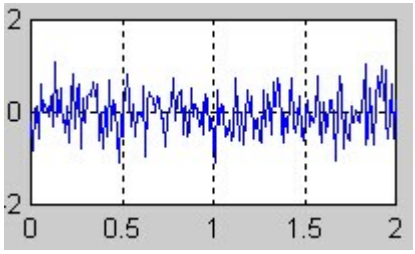

Figure 2. Fourier transform

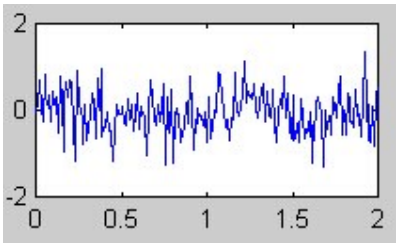

Figure 3. wavelet analysis

A new model has been given as shown in figure 6 according to big data. In the new model, k-means algorithm has been adopted to classify large data because it is difficult to forecast time series and it is easy to process large data fast for k-means algorithm. According to clustering algorithm in the model, the samples in dataset X given has d attributes which include $A_1,A_2\ldots A_d$, and data samples $x_i=(x_{i1},x_{i2},\ldots x_{id})$, $x_j=(x_{j1},x_{j2},\ldots x_{jd})$.Furthermore, $x_{i1},x_{i2},\ldots x_{id}$ and $x_{j1},x_{j2},\ldots x_{jd}$are the values of $x_i$and x_j's attributes $A_1,A_2\ldots A_d$.If the distance $d(x_i,x_j)$ between x_i and $x_j$is shorter,the sample x_i and x_j will be more similar. On the contrary, the distance is bigger, and samples are more different. The distance formula 1 as follows.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d}(x_{ik} - x_{jk})^2} \quad (1)$$

Error variance and rule function has been adopted to evaluate computing results, dataset X, including described attributes not class attributes. Supposed dataset X had k clustering subsets x_1,x_2,…x_K.Each clustering subset's number of samples is n_1， n_2,…,n_k. Each clustering subset's mean value representative point is m_1， m_2,…,m_k.Error square and function variance shown as formula 2.

$$E = \sum_{I=1}^{K} \sum_{P \in X_I} ||P - m_i||^2 \quad (2)$$

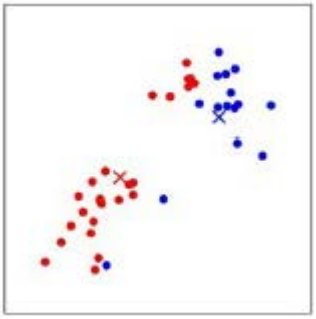

Figure 4.k-means

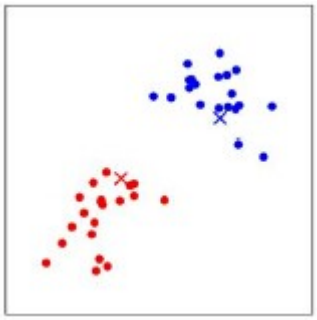

Figure 5. aftermodifying

After clustering data by k-means algorithm, the test result is shown as figure 4. After modifying the algorithm, by the evaluation result, we could use formula 2 to get result shown as figure 5. By comparison, the k classes could be divided as we expect, and other parts of process will not be introduced.

In short, if we firstly divide data into dynamic part and static part by analyzing Fourier transform and wavelet analysis, much calculated amount will be reduced

and it will have better effect on denoising.The dividing grade is vitality of data on the time axis. The high vitality has been seen as dynamic part, and the low vitality has been seen as static part.The parameter K is important. If there is no clustering, Fourier transform will take the high-frequency part. Otherwise, windows Fourier transform is generally fit for concentrated energy and low-frequency signals, in another words, once the size of Fourier windows has been chosen, it will not be changed and only suited to handle static data.Although wavelet analysis has good adaptability, it is better for wavelet analysis to process high-frequency part, dynamic data. The intention is to denoise in the low-frequency signals.

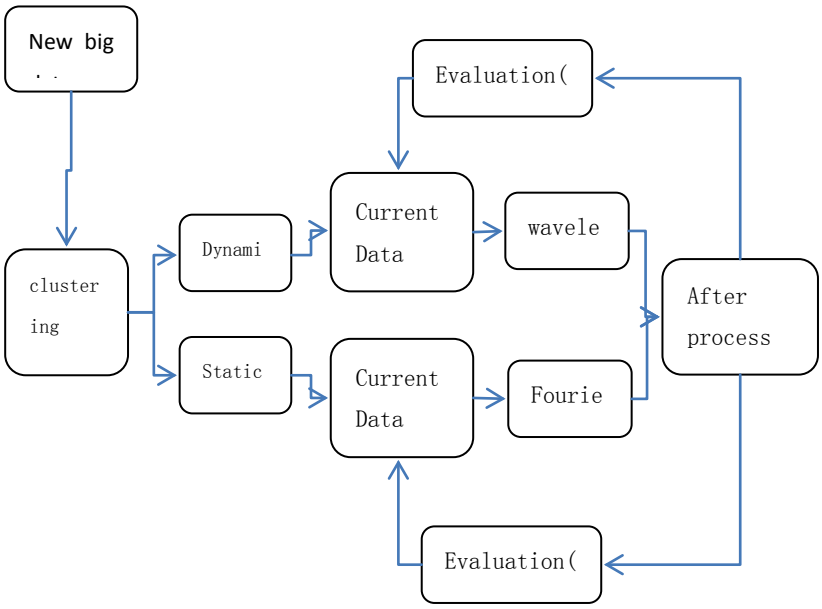

Figure 6.Mixed denoising model

Windows Fourier transform

Define Fourier transform: if f(t) is the periodic function of t which meets Dirichlet conditions, there is finite breaks on which functions have finite values;

there is finite extreme points in a period, and it is absolutely integrable. The formula 3 is Fourier transform.

$$F(\omega) = °F[f(t)] = \int_{-\infty}^{\infty} f(t)e^{-iwt}dt \quad (3)$$

For testing data in computer, function must be defined on discrete points not continuous region and the condition of finiteness and periodicity must be met. In that condition , the discrete Fourier transform of series $\{x_n\}_{N=}^{N-1}$is formula 4.

$$X[K] = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N} \quad (4)$$

The computational complexity defined by DFT is O(N2), and it will be O(nlogn) by Fourier transform. Otherwise, Fourier transform could do frequency analysis at the same time by doing quadrature between smooth function and signal f(t). The algorithm Mallet has been adopted to do decomposition and reconstruction to high-frequency data. Although the method of wavelet analysis could do decomposition to signals, we just need high-frequency part which is classified by algorithm k-means. So windows Fourier transform has been adopted to process [0,200] data, and the signals beyond 200Hz will be decomposed to the forth layer, which uses sym4 after getting threshold.

In the model shown as figure 6, the preprocessing of time series data includes wavelet analysis for high-frequency part and Fourier transform for low-frequency part.Compared with original time series signals. There is three features after preprocessing. First, whole variant trend is not changed too much, and it is avoided to lost much data. Second, the loss of high-frequency data get less because this part filtered by wavelet analysis not Fourier transform. Third, reduce more noise, and enhance each signal.In the figure 7, it is the signal with noise, and new signal has been obtained after being processed by model shown in figure 6. The figure 8 show us the signal after preprocessing with wavelet analysis and Fourier transform.

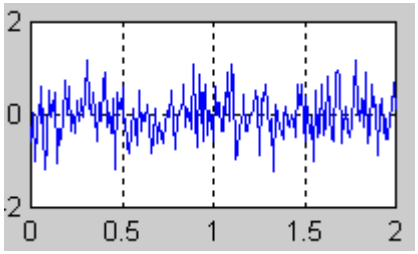

Figure 7.Before denoising

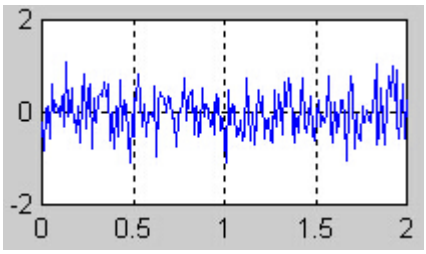

Figure 8.After denoising

According to results, SNR has been adopted to evaluate our model. Supported s_∂ is effective signal, and s_0 is noise,ε is the D-value between effective signal and denoising signal. So we could get SNR by formula 5 and 6.

$$\mathrm{SNR} = 20\log_{10}\frac{s_\partial}{||s_\partial - s_0||} \quad (5)$$

$$\partial = \sqrt{\frac{\varepsilon_1^2+\varepsilon_2^2+\cdots+\varepsilon_n^2}{n}} \quad (6)$$

Though analyzing figure 7 and figure 8, we could find the difference of SNR in table 1. To high-frequency data, the wavelet analysis is much faster than Fourier transform in processing data independently. Wavelet analysis will not lose high-frequency, but the result of the third layer is very close to the forth. The D-value is less 10^{-4}. Secondly, in processing the low-frequency, the loss of windows Fourier transform is 63% of wavelet analysis, but efficiency of denoising of low-frequency is 1.204 times of wavelet analysis. So in SNR view, wavelet analysis is better than Fourier transform, but the mixed model is better than wavelet analysis.

Table1 mixed denoising algorithm SNR compared

	Original data	Fourier transform	Wavelet analysis	Mixed method
SNR（db）	-8.3021	-3.2902	-2.4430	-2.0373

Acknowledgements

2014 Guangxi Department of Education Research Universities Project(YB2014152); 2013 Philosophy and Social Sciences of Guangxi "Twelve Five" Project(13CTQ001); 2014 Guangxi Natural Science Fund Project(2014GXNSFDA118032).

References

[1] Satirapod C, Ogaja C, Wang Jingling,et al. An Approach to GPS Analysis Incor-prationg Wavelet Decomposition[J]. Artificial Satellites Journal of Planetary Geodesy, 2001, 36(2):27-35.

[2]Yin hui, Zhu Feng.The evaluation standard and wavelet strategy in time series denoising.Journal of wuhan university information science. 11.2012.

[3]Si Zhenzhen. The application of Fourierand wavelet transform in signal denoising. Electronic design Engineering. 2.2011.

[4] Hu Changhua, Li Guohua, Zhou Tao. Wavelet analysis based on Matlab7.X's system analysis and design, version 3. Xi'an:Xi'an university of electronic science and technology press.2008.

[5]Du Yi. The study and application on time series data mining. USTC Doctoral Dissertation. 2007.5：45-46.

[6]Time series mining based on time phase curving distance[J]. Computer Research and Development. 2005.42(1):72-78.

[7]Chung F L, Fu T C, et al. An evolutionary approach to pattern-based time series segmentation[J]. IEEE Trans. Evolutionary Computation, 2004,8(5):471-489.

[8] FAN Deqin, ZHU Wenquan, PAN Yaozhong, JIANG Nan. Nose detection for NDVI time series based on Dixon's test and application in data reconstruction, Journal of Remote Sensing,2013.5.

[9]Zhang Bo. The study of denoising earthquake data base on sparse matrix transform. Zhejiang University:2013.1.25：66-78.

[10]Li Qing, Wang Runqiu, Huang Wen Feng. The method of form denoising in the processing earthquake data[J]. Petroleum Secience, 2005,(4):24-33.