STUDY ON HIERARCHICAL THRESHOLD DE-NOISING METHOD BASED ON NEAR INFRARED SPECTRUM DATA

Pengfei Pan[1], Houguang Sun[1], Hankang Zhang[2], Yachun Mao[2*], Hui Luan[1]
1. Anqian Mining Co.Ltd, No. 11, Qiwang Road, Anshan 114043, China
2.College of Resources and Civil Engineering, Northeastern University, No. 3-11, Wenhua Road, Shenyang 110819, China;

KEY WORDS: near infrared spectrum data; model precision; hierarchical threshold; wavelet de-noising

ABSTRACT: In recent years, the research of modeling method based on near infrared spectrum data has become one of the main methods for the analysis of mineral composition, however, due to the influence of various factors, there is a lot of noise in the near infrared spectrum data, which causes serious influence on the precision of the model and model robustness. In this paper, the method of using wavelet detail coefficients of autocorrelation for hierarchical threshold de-noising is proposed to eliminate the noise contained in near infrared spectrum data of hematite. First, the maximum decomposition layer is determined according to the minimum frequency of effective signal. Second, the signal is decomposed to the maximum degree, and the threshold value is determined according to the correlation of the coefficients of each decomposition layer and the noise. Then, the near infrared spectrum data is processed by the calculated threshold value. Results show that the method not only can eliminate the noise in the data effectively but also can maximize the retention of feature information in the data, which improve the precision and robustness of the model effectively, and an effective de-noising method was provided for the models establishment based on near infrared spectrum data of hematite.

1. INTRODUCTION

In the field of remote sensing, the characteristics, variation law and influencing factors of rock and mineral spectrum (reflection, emission) are the basis and foundation of mineral identification (Su and Du, 2006; Liu et al, 1999; Hunt et al, 1978). The rock ore near infrared spectral data has the characteristics of low SNR, high volatility and spectral image peak overlap (Chu et al, 2006). So a reliable data source is the foundation of the modeling (Wang et al, 2011). Due to the influence of various factors, in the process of collecting the sample data by using near infrared spectrometer (Kahte and Goetz, 1983), as shown in picture 1, a lot of noise is contained in the test data. The noise will cause serious impact on the accuracy and robustness of the model. Therefore, data preprocessing is very important and necessary in the process of modeling based on near infrared spectroscopy data (Gao et al, 2004). The noise contained in the near infrared spectrum data is eliminated through data pretreatment, so the goal of improving the accuracy of the modeling is achieved. In the conventional method, the global threshold de-noising method is usually used to deal with the noise in near infrared spectrum data (Cohen et al, 1999; Wu et al, 2015; Li et al, 2010; Wang et al, 2009), but as the threshold value calculated by the global threshold de-noising method (Donoho, 1995; Donoho and Johnstone, 1994; Donoho and Johnstone, 1995; Mallat, 1989; He and Yu, 1997) is usually too large, the high frequency part of the signal is all eliminated, and it causes the phenomenon of the loss of useful signals, which causes the signal distortion. In this paper, the method of using wavelet detail coefficients of autocorrelation for hierarchical threshold de-noising is proposed. First, the maximum decomposition layer is determined according to the minimum frequency of effective signal. Second, the signal is decomposed to the maximum degree, and the threshold value is determined according to the correlation of the coefficients of each decomposition layer and the noise. Then, the near infrared spectrum data is processed by the calculated threshold value. Results showe that the method can not only eliminate the noise in the data effectively but also can maximize the retention of feature information in the data, which improve the precision and robustness of the model effectively, and an effective de-noising method was provided for the models establishment based on near infrared spectrum data of hematite.

Fig.1 Iron ore samples and acquisition of spectral data

2. TEORY OF WAVELET THRESHOLD DE-NOISING ALGORITHM

Suppose a one dimensional observation signal is:

$$f(t) = s(t) + n(t) \quad (1)$$

where, $s(t)$ is original signal, $n(t)$ is gauss white noise, and σ^2 is its variance , and obey the distribution $N(0, \sigma^2)$.

For discrete sampling $f(t)$, discrete signal $f(n)$ is achieved, $(n = 0,1 \cdots N-1)$, the wavelet transform of which is:

$$Wf(j,k) = 2^{-j/2} \sum_{n=0}^{N-1} f(n) \cdot \psi(2^{-j} - k) \quad j,k \in Z \tag{2}$$

$Wf(j,k)$ is wavelet coefficients. Because the calculation of (2) is more complicated, the commonly used wavelet coefficients to obtain the recursive are expressed as follows:

$$Sf(j+1,k) = Sf(j,k) \cdot h(j,k) \tag{3}$$
$$Wf(j+1,k) = Sf(j,k) \cdot g(j,k) \tag{4}$$

where h and g are the low pass and high pass filters respectively corresponding to the scaling function ϕ（t) and the wavelet function ψ（t）; $Sf(0, k)$ is the original signal; $Sf(j,k)$ is the scale coefficients; $Wf(j,k)$ is wavelet coefficients. The corresponding reconstruction formula is:

$$Sf(j-1,k) = Sf(j,k) \cdot \tilde{h}(j,k) + Wf(j,k) \cdot \tilde{g}(j,k) \tag{5}$$

where $\tilde{h}$ and $\tilde{g}$ are correspond to the reconstruction of low pass and high pass filters.

Let $w_{j,k} = Wf(j,k)$, because the wavelet transform is a linear transformation, the wavelet coefficients $w_{j,k}$ is obtained from discrete transform $f(k) = s(k) + n(k)$ .The $w_{j,k}$ consists of two parts, one part is the wavelet coefficients $W_s(j,k)$ corresponding to signal $s(k)$, and the other part is the wavelet coefficients $W_n(j,k)$ corresponding to signal $n(k)$.

The theory of wavelet threshold de-noising method is that when the wavelet coefficient $w_{j,k}$ is less than a certain critical threshold, $W_{j,k}$ is mainly composed of noise, it should be abandoned; when the wavelet coefficient $w_{j,k}$ is larger than a certain critical threshold, the $w_{j,k}$ is mainly composed of signal. Put this part of the value shrink to zero according to a fixed amount, namely soft threshold de-noising method, then the de-noised signal is achieved by wavelet reconstruction according to new wavelet coefficients.

3. THE HIERARCHICAL THRESHOLD DE-NOISING METHOD BASED ON HEMATITE NEAR INFRARED SPECTRUM DATA

In this paper, the near infrared spectrum data of hematite ore collected in Anqian Mine are used as the data source, the laws of reflectivity variation with wavelength are showed in Figure 2. It can be clearly seen from Figure 2 that the sample data include a certain amount of noise in the range of 750~1000nm and 1750~2000nm.

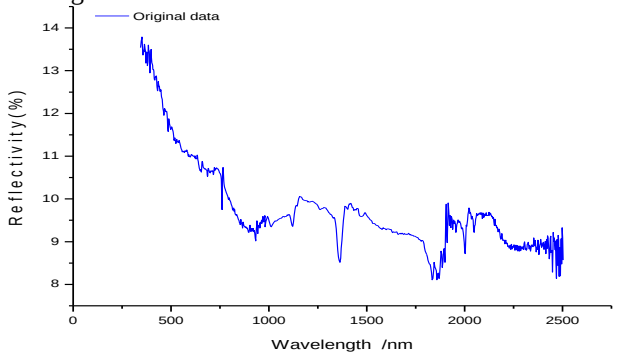

Fig.2 Original near infrared spectrum data

Sym4 wavelet basis functions as its better regularity is used in this paper, and the near infrared spectrum data is de-noised by soft threshold based on hierarchical threshold de-noising method. The results are compared with the data dealt with global threshold de-noising method, the applicability for hierarchical threshold method to hematite near infrared spectral data is further showed.

3.1 The determining of optimal decomposition layer of wavelet based on near infrared spectrum data

Decomposition layer has a great effect on the de-noising effect. If the number of decomposition layer is too small, the noise signal in low frequency coefficient cannot be eliminated, which will not achieve effective noise elimination effect. If the number of decomposition layers is too large, the calculation quantity will increase, which leads to the slowdown of the processing speed and the decreasing of the signal to noise ratio. At the same time, the phenomenon of excessive noise elimination will be produced as well.

As shown in the following figures (Fig.3to Fig5), the effect of different decomposition layers on the noise elimination effect can be clearly distinguished from the above three images. In Figure 3, noise elimination is not completed with the two layer decomposition of the signal de-noising, for there is still some noise contained in the processed signal and the purpose of eliminating noise is not achieved. From Figure 4 it can be found that the signal is too smooth. Although the noise is completely eliminated, the useful detail signal part is completely eliminated as well, which causes the loss of the signal and

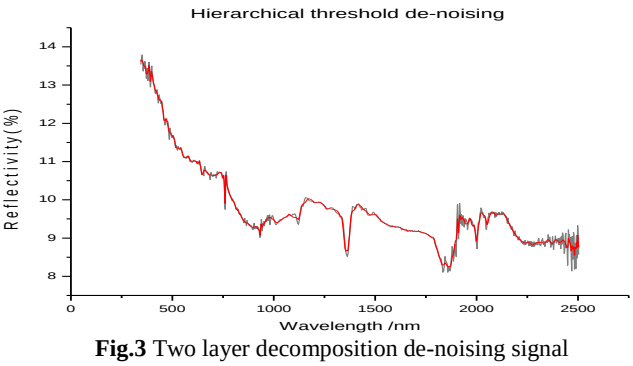

Fig.3 Two layer decomposition de-noising signal

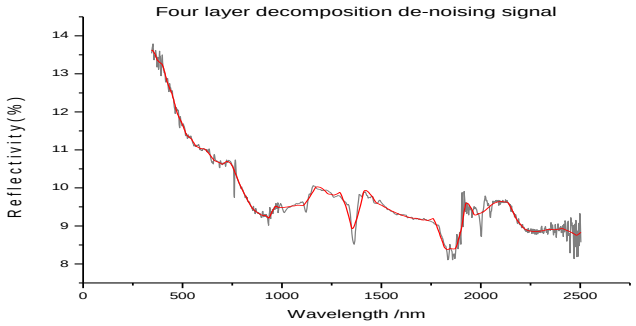

Fig.4 Three layer decomposition de-noising signal

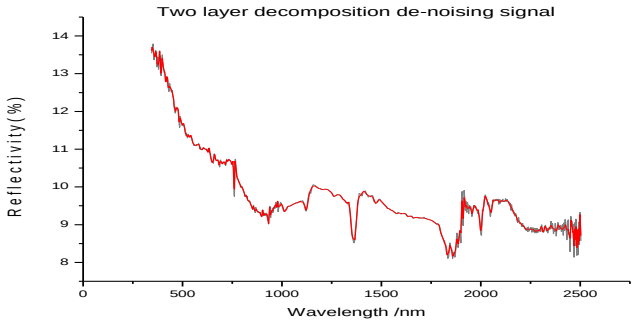

Fig.5 Four layer decomposition de-noising signal

is not desirable in the application. Therefore the three decomposition layer is the most appropriate in that it not only eliminates the noise in signal, but also ensures the authenticity and reliability of the signal.

The process of wavelet decomposition is the process of dividing the signal, because the required signal has the minimum frequency limit, thus the decomposition layer can be determined according to the minimum frequency of the signal decomposition. The determination of the maximum decomposition layer can be based on the corresponding relationship between the scale and frequency, the formula of which is as follow:

$$f_m^{ps} = \frac{f_c f_s}{m} \qquad (6)$$

where f_m^{ps} is the pseudo frequency corresponding to scale m, f_c is the center frequency of the corresponding wavelet, f_s is the sampling frequency. Then the correspondence of the useful signal frequency f_{sig} and pseudo frequency are as follows:

$$f_L^{ps} \le f_{sig} \le f_1^{ps} \qquad (7)$$

From formula (7), it is known that the minimum frequency of the useful signal is larger than or equal to the pseudo frequency of L, and the relationship is as follows:

$$f_L^{ps} \le \min\ f_{sig} \qquad (8)$$

The scale value L can be calculated from formula (8). The decomposition layer is to sample discretely to scale with two step size, so the formula of decomposition layer is determined as follows:

$$2^{i-1} \le L \le 2^i \qquad (9)$$

where i is the maximum decomposition layer of the signal.

According to the sampling data of hematite near infrared spectrum, its sampling frequency is greater than 1000HZ, and the minimum of useful signal is about 200HZ, according to formula (8) and formula (9) it can be calculated as follows:

$$2^{i-1} \le 3.6 \le 2^i \qquad (10)$$

The decomposition layer can be determined to be 3.

3.2 Method to determine the hierarchical threshold based on hematite near infrared spectrum data

There are two kinds of de-noising methods included in threshold de-noising method, which are global threshold method and hierarchical threshold method. Usually, the threshold value confirmed by global threshold method is so large that all the high frequency signals in the data are eliminated, which causes signal distortion and SNR decrease. For the hierarchical threshold method, the threshold value is determined separately for each decomposition layer, which determines the appropriate threshold according to the characteristics of each layer data to ensure that the noise of each decomposition layer is eliminated and the useful signal to be effectively retained. So the details of the signal can be retained after dealt with hierarchical threshold method, thus the integrity and authenticity of the data can be retained as well.

The determination method of the hierarchical threshold is to obtain the main square error σ of each layer of high frequency. When the noise signal is large and the useful signal is few in high frequency layer, the coefficient is concentrated in [-3σ,3σ], and the probability of falling outside this interval is very small. So the condition that the coefficient is beyond the interval, namely the absolute value of coefficient is greater than 3σ is considered useful signal coefficient. In this case the threshold value should be between 3σ~4σ. But when the useful signal in high frequency layer is increased, the threshold value should be decreased in order to prevent the useful signal from being eliminated. If all the signals in the high frequency layer are useful signals, the threshold value should be 0. So the calculation formula of each layer threshold is as follow:

$$thr = k\sigma \quad (0 \le k \le 4) \qquad (11)$$

As shown in Figure 6, it is decomposition of hematite near infrared spectrum data, from Figure 6 it can be found that most of the noise is concentrated in the d1 decomposition layer, and useful signal is almost zero. The threshold value should be set large to eliminate the large number of noise signal, so, four times main square error

(4σ) is set as the threshold value in d1 layer. After the decomposition of the d1 layer, the noise contained in high frequency signal of d2 decomposition layer is relatively reduced, but a certain amount of noise signal is still contained in the decomposition layer. So in order to eliminate the noise while preserving the useful signal effectively in d2 decomposition layer, two times main square error (2σ) is set as the threshold value in d2 layer. In d3 decomposition layer only a small amount of noise is contained and the rest are useful signals. So as to eliminate the noise effectively and avoid the useful signal being eliminated, in d3 layer, threshold value is set to one time main square error (σ).

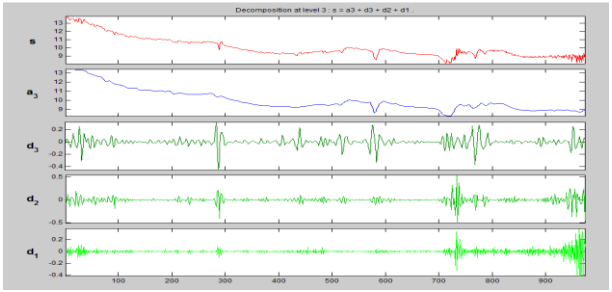

Fig.6 decomposition of near infrared spectrum data

The hierarchical threshold de-noising method is used to eliminate hematite near infrared spectrum data by using the above threshold value, and the result is shown in Figure 7. At the same time, the global threshold de-noising method is also used to eliminate hematite near infrared spectrum data by using "Rigrsure" threshold value, the result of which is showed in Figure 8.

In Figure 7, in order to make the comparison more obvious, the two signals are presented in the same picture: the red signal line is hierarchical threshold de-noising signal, and the gray signal line is original signal. With comparison of the red signal line and the gray signal line, it can be obviously found that the noise in 750~1000nm, 1750~2000nm and 2250~2500nm has been eliminated effectively with no the phenomenon of spectrum data loss. Not only the integrity of the signal is ensured, but also the accuracy of the reflectance of the near infrared spectrum is improved. From the comparison of Figure 7 and Figure 8, it can be found that the data dealt with global threshold de-noising method can cause the loss of useful signal, for example, in Figure 8, the phenomenon of spectrum loss comes up in 750nm and 2000nm and other wavelengths. This shows the limitations of global threshold de-noising method to eliminate near infrared spectrum data, while hierarchical threshold de-noising method can avoid this problem effectively with the elimination of spectrum data. Therefore, the hierarchical threshold de-noising method has better applicability to be used in eliminating the noise contained in near infrared spectrum data.

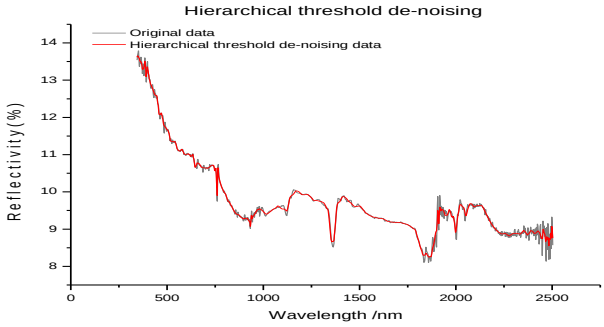

Fig.7 hierarchical threshold de-noising data

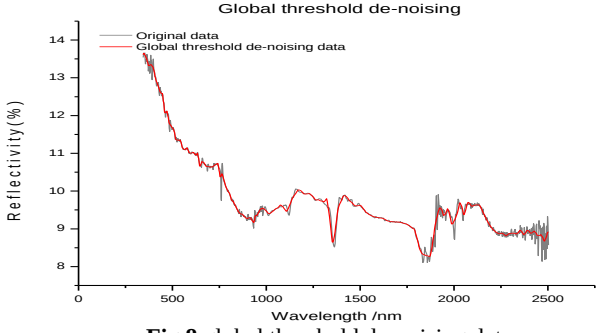

Fig.8 global threshold de-noising data

4. EFFECT TEST

Although the hierarchical threshold de-noising method has a better effect to eliminate the noise contained in hematite near infrared spectrum data, its accuracy still needs to be tested to improve the accuracy of modeling. Here, the de-noising near infrared spectrum data is used for random forest modeling to test the accuracy. The test method is to use original data, hierarchical threshold de-noising data, and global threshold de-noising data to establish the random forest model of the spectral curve and mineral components of the sample, then to calculate the model accuracy, according to which the effect of noise elimination can be determined. The model accuracy data is shown as Table 1.

Tab.1 Main square error of different de-noising method

De-noising method	Original data	Global threshold	Hierarchical threshold
Main square error	0.8573	0.7785	0.7154

It can be seen from Table 1 that the modeling accuracy of near infrared spectrum data has been greatly improved after de-noising. For different de-noising methods, the de-noising effect is different. the data accuracy of modeling can be improved by Hierarchical threshold de-noising method than that by global threshold de-noising method. So it can be judged that the hierarchical threshold de-noising method has a good applicability to improve the accuracy of the near infrared spectrum data model, and the de-noising data is more reliable than that of the global threshold method.

5. CONCLUSION

In this paper, hematite near infrared spectral data collected in Anqian open pit iron mine are used as the data source. Hierarchical threshold de-noising method is used to deal with the near infrared spectrum data, which shows good applicability in eliminating the noise contained in near infrared spectrum data. The main conclusions are as follows:

1) Decomposition of the signal layer should be appropriate, if too large or too small, the effect of noise elimination will be affected. According to the minimum frequency of the useful signal, the number of

decomposition layers can be determined, and the most effective decomposition layer for the near infrared spectrum data based on hematite is the third layer.

2) In d1 decomposition layer, with most of the noise contained in, the useful signal is almost zero. In order to eliminate the large number of noise signal, four times main square error (4σ) is set as the threshold value in d1 layer. After the decomposition of the d1 layer, the noise contained in high frequency signal of d2 decomposition layer is relatively reduced, but still contained a certain amount of noise signals in the decomposition layer. So two times of main square error (2σ) is set as the threshold value in d2 layer. In d3 decomposition layer only a small amount of noise is contained in and the rest are useful signals. In order to eliminate the noise effectively and avoid the useful signal being eliminated, in d3 layer, threshold value is set to one time of main square error (σ).

3) Hierarchical threshold method not only can eliminate the noise contained in hematite near infrared spectrum data effectively, but also can avoid the drawback of spectral losing of the global threshold de-noising method in such a way that the useful detail signal is retained effectively and the result is more authentic and reliable. The accuracy of the random forest model calculated by hematite near infrared spectral data based on the hierarchical threshold de-noising method has been improved effectively. So it can be judged that hierarchical threshold de-noising method is feasible and practical.

ACKNOWLEDGMENT

The authors are grateful to the managers of the Anqian open iron pits of Liaoning Province for providing the experimental conditions used in this study. This research is funded by the National Natural Science Foundation of China (No. 41371437) and National Science and Technology Support Program of China (No. 2015BAB15B01).

REFERENCE

Chu, X.L., Yuan, H.F., Lu, W.Z., 2006. Research and applications of near infrared spectroscopy in China in recent years[J]. Analysis instrument, 2(2), pp.1-10.

Cohen, I., Raz, S., 1999. Malah.Translation – invariant denosingusing the mini mum description length criterion[J]. Signal Processing, 75(3), pp.201-223.

Donoho, D.L., 1995. De-noising by soft-thresholding[J]. IEEE Trans on Information Theory, 41(3), pp. 613-627.

Donoho, D.L., Johnstone, I.M., 1994. Ideal spatisl adaptation via wavelet shrinkage[J]. Biometrika, 81(12), pp.425-455.

Donoho, D.L., Johnstone, I.M., 1995. Adapting to unknown smoothness via wavelet shrinkage[J]. Journal of American Stat.Assoc, 12(90), pp.1200-1224.

Gao, R.Q., Fan, S.F., Yan, Y.L., et al, 2004. Preprocessing of Near Infrared Spectroscopic Data[J]. Spectroscopy and Spectral Analysis, 24(12), pp.1563-1565.

Hunt, G.R., Salisbury, J.W., Lenhoff, G.J., 1978. Visible and near-infrared spectra of minerals minerals and rocks: Oxides and hyoxides [J]. Modem Geology, 2(2), pp.195-205.

He, J., Yu, Y.L., 1997. Wavelet analysis and its application to signal processing[J]. Journal of University of Science and Technology, 4(3), pp.49-53.

Kahte, A.B., Goetz, A.F.H., 1983. Mineralogic Information from a New Airborne Thermal Infrared Multispectral Scanner [J]. Science, 222(4619), pp.24-27.

Liu, Q.S., Wang, Z.G., Jing, H.L., 1999. Correspondence Analysis of Laboratory Spectra of Rock[J]. Journal of remote sensing, 3(2), pp.68-73.

Li, H., Lin, Q.Z., Wang, Q.J., et al., 2010. Research on Spectrum Denoising Methods Based on the Combination of Wavelet Package Transformation and Mathematical Morphology[J]. Spectroscopy and Spectral Analysis, 30(3), pp.644-648.

Mallat, S.G., 1989. A theory for multi-resolution signal decomposition: the Wavelet Representation[J]. IEEE Trans.on PAMI, 11(7), pp.674-693.

Su, H.J., Du, P.J., 2006. Study on Feature Selection and Extraction of Hyperspectral Data[J]. Remote sensing technology and application, 21 (4), pp.288-293.

Wang, F.H., Zhu, H.L., Ge, Z.Y., 2011. Progress of Near-infrared Spectral Data Modeling Method[J]. Agricultural Engineering, 1(1), pp.157-163.

Wu, D.H., Wang, H.G., Zhang, P.L., et al., 2015. Denoising combined with spatial domain and neighborhood based on dual-tree complex wavelet packet transform[J]. J. HuazhongUniv. of Sci. & Tech. (Natural Science Edition), 43(4), pp.17-21.

Wang, X.S., Qi, D.W., Huang, A.M., 2009. Study on Denoising Near Infrared Spectra of Wood Based on Wavelet Transform[J]. Spectroscopy and Spectral Analysis, 8(29), pp.2060-2062.