
1. Multivariate scattering correction

Multi-scattering correction method is a data processing method commonly used in spectral standards. The spectral data obtained after the scattered correction can effectively eliminate the effects of scattering due to an uneven distribution of sample particles and particle size. Regarding spectral absorption information, light scattered shooting has been used in solid mansion, slurry transmission, and reflection spectra.

This method assumes that the changes in the spectrum and the contents in the sample have a direct linear relationship. The ideal spectrum of a sample must be established before use, and the spectrum of other samples should be corrected accordingly. In practical applications, the ideal spectrum can be difficult to obtain. Since this method is only used to correct the relative baseline translation and offset in the near -infrared spectrum of each sample, it is necessary to calculate all the spectral arrays of specific samples to calculate all the spectrum arrays of the whole sample. The average spectrum of spectrum is an ideal standard spectrum [1].

First, calculate the average spectrum of all samples near infrared spectrum,

$$\vec{y} = \frac{\sum_{j=1}^m y_{ij}}{m}$$

where y is the spectral value; $i = 1, 2, 3, \dots$ represents the number of bands; $j = 1, 2, 3, \dots, m$ represents the number of samples.

The average spectrum was taken as the standard spectrum, and the near-infrared spectrum of each sample was subjected to a linear regression operation with the standard spectrum to obtain the linear translation (regression constant) and tilt offset (regression coefficient) of each spectrum relative to the standard spectrum.

$$y_{j(S)} = k_j \vec{y} + b_j$$

By subtracting the linear shift from the original spectrum of each sample and dividing by the regression coefficient to correct the relative tilt of the baseline of the spectrum, the corrected spectrum was obtained.

$$y_{j(MSC)} = \frac{y_{j(S)} - b_j}{k_j}$$

2. First derivative

In the process of spectral data acquisition and data analysis, absorption spectrum baseline drift, translation, or background interference is common, which can be processed by derivation. With the increase in order, the sensitivity will be improved, the resolution will be reduced, and overlapping peaks will be separated. However, this can also result in noise problems. The increase in extreme values and spectral characteristics makes it difficult for smooth processing to obtain ideal signal-to-noise ratio, and the data enhancement effect may not be significant. Secondly, the selection of the experimental wavelength is also a key issue in derivative correction. If the experimental wavelength is small, the noise is also more obvious, which has a certain influence on the actual quantitative and qualitative analyses. If the experimental wavelength is large, the transformed wave pattern is relatively smooth, but some feature details may be overlooked. Therefore, the process is usually to calculate the first-order derivative or the second-order derivative of the spectrum. There are two methods to calculate the derivative of the spectrum: the direct-difference method and the convolution method. To avoid a reduction in the characteristics in the process, the wavelength is selected as 1, so the simple direct-difference method can be used for calculation [2].

First derivative formula:

$$\frac{d_y}{d_\lambda} = \frac{y_{i+1} - y_i}{\Delta\lambda}$$

where y is the spectral value; and $i = 1, 2, 3, \dots$ represents the band length.

3. Wavelet Inverse Transform

In recent years, wavelet theory has developed rapidly. Due to its multiresolution decomposition and good time frequency, it has been widely adopted. Wavelet transform decomposes the spectrum into wavelet functions in different frequency bands by multi-resolution decomposition at different scales according to the frequencies. These wavelet functions are the sub-signal functions obtained by translation and scaling of a mother wavelet function and then directly extracting the frequency band where the useful signal is located or setting the frequency band where the noise is located to zero for wavelet reconstruction, so as to focus on any part of the signal and achieve the purpose of complete extraction and denoising of the signal data. Therefore, the essence of the wavelet inverse transform is to project the spectrum onto

the wavelet to obtain simplified wavelet coefficients. The final signal was selected according to different needs to deal with different wavelet coefficients, and then the spectrum was obtained by inverse transform.

The most significant difference between the Fourier transform and the wavelet transform is that the Fourier transform has fewer available functions, typically only trigonometric functions, and the results are, therefore, very different. According to the environments and use requirements, different wavelet functions can be selected, and the experimental results closest to the ideal results can be obtained. Therefore, wavelet transform has significant advantages over Fourier transform for time-frequency analysis.

The key to the wavelet transformation is the choice of wavelets, followed by the selection of the threshold and the threshold function. The wavelet transformation has continuous wavelet transformation and discrete wavelet transformation. The discrete wavelet transformation is obtained from the scale of continuous wavelet transformation [3].

4. Enhanced data processing

Before analyzing the spectral data, data enhancement algorithms are typically used to reduce or eliminate some redundant information, which can increase the difference between samples and unify the dimensional problems, so as to improve the reproducibility and prediction ability of the model. It can also remove the unit limit of the data and transform it into dimensionless pure values, which is convenient for the comparison or weighting of varied units or magnitudes. Common algorithms include mean centralization, standardization, normalization, and so on. However, since the research content was proportional, it was necessary to generalize the water samples for horizontal data processing, that is, using the absorbance of pure aquifer samples to assess the absorbance in each proportion of mixed samples, so as to improve the method of data enhancement. We selected the commonly used min-max normalization method for improvement.

Min-max normalization is a linear transformation of the original data, so that the results fall within the [0,1] interval. The improved conversion function is as follows:

$$\hat{y} = \frac{y - y_2}{y_1 - y_2}$$

where y_1 , y_2 is the sample data of pure water, and y is mixed sample data.

5. Extreme gradient boosting (XGBoost) algorithm

The basic principle of XGBoost [4,5]:

Create the base model of t decision tree:

$$f_t(x) = w_q(x), w \in R^T, q: R^d \{1, 2, \dots, T\}$$

where W is the decision tree leaf vector, q is the tree structure, and T is the number of leaves. Multiple decision tree models are composed of an additive formula for prediction. The initial prediction value:

$$\hat{y}_i^{(0)} = 0$$

Add a new tree to this and deduce the following formula:

$$\begin{aligned} \hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\ &\dots \\ \hat{y}_i^{(t)} &= \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \end{aligned}$$

where $\hat{y}_i^{(t)}$ is the final result and model predictive value of the t -round. When calculating this value, the model predictive value of the front wheel is retained, and a new tree function value is added.

To prevent the number of leaves from excessive leaf nodes, a single decision tree is overfit with XGBOOST, and the punishment items are introduced:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2$$

In the formula, Ω is the penalty term representing the complexity of the model, γ is the regularization parameter representing the number of leaves, λ is the regularization parameter representing the weight of leaves, and w is the value of leaf nodes. Therefore, the objective function is as follows:

$$Obj^{(t)} = \sum_{i=1}^n L(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^n \Omega(f_i)$$

With a sub -tree, it can also be expressed as:

$$Obj^{(t)} = \sum_{i=1}^n L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant$$

It is a convex loss function to measure the difference between predicted and real values. To improve the model, we needed to optimize this goal so that our objective function was as low as possible.

According to Taylor expansion:

L is a differential convex loss function to measure the difference between predicted value \hat{y}_i and real value y . To improve the model, we should find f_t to optimize this goal, so that our objective function is as low as possible.

According to Taylor expansion:

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2$$

Take $y_i, \hat{y}_i^{(t-1)}$ as the x , and $f_t(x_i)$ as Δx as a Tyler, and the target function is expanded:

$$Obj^{(t)} \approx \sum_{i=1}^n \left[L(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + constant$$

$$g_i = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} \quad h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)^2}}$$

Since the previous differential convex loss function is the sum of the previous trees, it can be considered as a fixed value. It is mentioned in the constant term and brought into:

$$Obj^{(t)} = \sum_{j=1}^T \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T + C$$

$$\text{Make } G_j = \sum_{i \in I_j} g_i, \quad H_j = \sum_{i \in I_j} h_i:$$

$$Obj^{(t)} = \sum_{j=1}^T \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T + C$$

Let the partial derivative of the function in brackets to w_j be zero to obtain the minimum objective function.

$$\frac{\partial J(f_t)}{\partial w_j} = G_j + (H_j + \lambda) w_j = 0$$

The new objective function is obtained by bringing it into the original objective function:

$$Obj^{(t)} = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T + C$$

Obj represents how much we reduce at most on the target when we specify the structure of a tree. We can call it a structural fraction, which is a function of scoring the tree structure. The smaller the score is, the better the structure of the tree. According to the objective function, when the tree structure is determined, the tree structure score is only related to the first-order and second-order reciprocals. When there are many feature nodes, we cannot enumerate all the possibilities of the tree structure, so we choose a greedy algorithm to start iterative splitting from a single leaf node to add nodes to the tree. The loss function of the divided nodes of the tree:

$$score = split(before) - split(after)$$

We used the following:

$$score = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma$$

The equation $\frac{G_L^2}{H_L + \lambda}$ is the score of the left sub-tree; $\frac{G_R^2}{H_R + \lambda}$ is the score of the right sub-tree; $\frac{(G_L + G_R)^2}{H_L + H_R + \lambda}$ is the node score that is not divided; γ is the complexity cost introduced by adding new leaf nodes, which can also be used as a threshold. When the gain after splitting is greater than this number, splitting is selected. Our goal is to find a cut score that maximizes the cut loss.

References

1. Guo, Q.Q. Research on Prediction Model of Soil Organic-Matter Based on the Near-Infrared Spectroscopy Technology. Henan Agricultural University: Zhengzhou, China, 2016.
2. Wang, X.M.; Zhu, B.Y.; Yin, C. Application of derivative spectrometry in pharmaceutical analysis. *Fujian Analysis & Testing* **2001**, *2*, 1431-1438.
3. Zhou, F.B.; Li, C.G.; Zhu, H.Q. Research on Threshold Improved Denoising Algorithm Based on Lifting Wavelet Transform in UV-Vis Spectrum. *Spectroscopy and Spectral Analysis* **2018**, *38*, 506-510.
4. Chen, T.Q.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, United States, 13/8/2016.
5. Tao, M.Q.; Liu, J.X.; Wu, Y.; Ning, Z.Q.; Fang, Y.H. Application of XGBoost in Gas Infrared Spectral Recognition. *Acta Optica Sinica* **2020**, *40*, 201-206.