Study on Pretreatment Algorithm of Near Infrared Spectroscopy

Xiaoli Wang Guomin Zhou
Agricultural Information Institute of CAAS
wxl608@126.com zhougm@mail.caas.net.cn

Abstract

Pretreatment of near-infrared spectral data is the basis of feature extraction, quantitative and qualitative analysis and establishment of models, it plays a significant role in obtaining the data and get reliable results. The purpose of the paper is *compared the advantages and disadvantages of the S-G, derivative and multiple algorithm methods of spectral preprocessing through the example of apple leaves. S-G algorithm can smooth the data relatively better, but we must according to the specific circumstances of the case while chose the width of the window and the order of polynomial; Kernel smoothing is better than S-G in two-ends data processing, but its processing speed is slower than S-G. Derivative algorithm can get more stable reflectance, but it is sensitive to noise, so it need to be used with the smoothing algorithm. Multiple scatter correction can be used effectively to eliminate the translation and offset of baseline. All of above algorithms have been applied in the system of near infrared spectroscopy processing system of leaves and satisfactory result was obtained.*

Keywords：Near Infrared Spectroscopy；Spectral pretreatment；Smoothing algorithm；Derivative algorithm；Multiple scatter correction algorithms

1. Introduction

Near infrared spectroscopy is electromagnetic waves whose wavelength is between 780 ~ 2500nm, and it is intermediate between visible light and infrared, NIR refers primarily to low-energy electronic transitions and the clusters containing hydrogen atoms. Various objects' components information can be obtained through the near-infrared spectroscopy.

However, when we survey and evaluate the components information of the objects, the measurements may be quite different for the difference of the weather, environment, humidity, temperature and human factors, as well as the existence of dark current. Near infrared spectroscopy is mainly the spectral information of the measured object, but also includes all sorts of interference and noise from outside. Therefore, the spectral data must be preprocessed before establish the model to reduce noise and interference information, lay the roots for simplifying the operational process of modeling and improve the accuracy of the analysis.

2. Materials and Methods

2.1 experimental materials

We used FieldSpec 3 portable spectroradiometer as the instruments which can be applied to many areas; it's spectral range: 350-2500nm; data interval: 1nm; wavelength accuracy: + /-1nm; It works through the

measurement of light reflectance, transmittance, emissive or radiation rate.

Samples are collected from fruit Research Institute of Liaoning Xingcheng, 60 apple leaves are collected and its spectral information are collected, in the dark room with black cloth as the background, artificial lighting； 40 samples are used to calibrate and 20 samples used to predict.

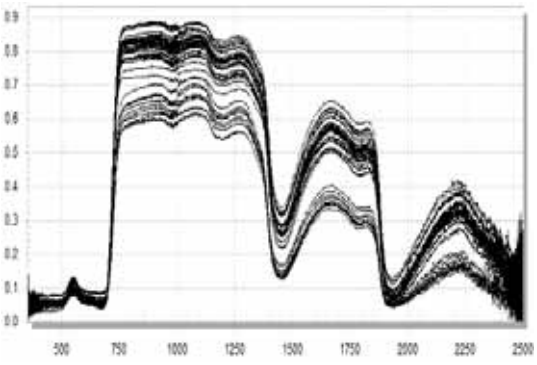

Figure1. Near infrared spectroscopy of apple leaves

2.2. Data smoothing algorithm

2.2.1. Savitzky-Golay smoothing algorithm. The main idea of Savitzky-Golay [1-3] (later referred to as S-G) filter algorithm is that use higher order polynomial to approximate fit the fixed number of data (also called the active window). For each point *i*, use the least square method, fit all the data of this activity window to a polynomial expression. Then replace the original value with the value of polynomial (g_i) at point *i*. When the window moves to the next point *i*+1, we use least-squares fitting algorithm to fit the original data rather than the data of the polynomial in the new window. Using the S-G algorithm requires that horizontal axis x_i have a uniform spacing (x_{i+1}-$x_i \equiv \Delta x$).

The core formula of the S-G algorithm:

$$g_i = \sum_{n=-n_l}^{n_r} c_n f_{i+n} ,$$

$$c_n = e_{M+1}{}^T (A^T A)^{-1} A^T$$

$$(-n_l \le n \le n_r) ;$$

Here f_i is the original data, g_i is the data after smoothing, e_{M+1} is M+1 dimensional unit vector, n_r is the number used to the right, i.e., later than i, n_l is the number of points used to the left of a data point i, i.e., earlier.

$$A = \begin{bmatrix} (-n_l)^m & (-n_l)^{m-1} & \cdots & -n_l & 1 \\ -(n_l-1)^m & -(n_l-1)^{m-1} & \cdots & -(n_l-1) & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ (n_r-1)^m & (n_r-1)^{m-1} & \cdots & (n_r-1) & 1 \\ n_r{}^m & n_r{}^{m-1} & \cdots & n_r & 1 \end{bmatrix}$$

$$\in \mathrm{R}^{(n_l+n_r+1)\times(M+1)}$$

2.2.2. Kernel smoothing algorithm [4]. Same as S-G algorithm, suppose the data after smoothing is $g(i)$, window width is h, weighted function is $S(t)$, then the expression of fitting value in time t (such as the spectrum reservation time) is:

$$g(t) = \sum_{j=1}^{n} S_j(t) y(j)$$

The weighted function proposed Gasser and Muller is:

$$S_j(t) = \frac{1}{h} \int_{t^*_{(j-1)}}^{t^*_{(j)}} \ker n(\frac{u-t}{h})$$

In the two formulas above, kern (x) function is the core of smooth; it has three different weighted functions:

Uniform function：

$$\ker n(x) = \begin{cases} 0.5 & |x| \le 1 \\ 0 & otherwise \end{cases}$$

Quadratic function：

$$\ker n(x) = \begin{cases} 0.75(1-x^2) & |x| \le 1 \\ 0 & otherwise \end{cases}$$

Gaussian function：

$$\ker n(x) = (2\pi)^{-1/2} \exp(-u^2/2)$$

2.3. Derivative algorithm

Usually, derivative [5] spectrum use different methods to approximate calculate:

First derivative spectrum:

$$\rho'(\lambda_i) = \frac{[\rho(\lambda_{i+1}) - \rho(\lambda_{i-1})]}{2\Delta\lambda}$$

Second derivative spectrum:

$$\rho''(\lambda_i) = \frac{[\rho(\lambda_{i+2}) - 2\rho(\lambda_i) + \rho(\lambda_{i-2})]}{4\Delta\lambda^2}$$

2.4. Multiple scatter correction

The concrete account step of multiple scatter correction [6] is:

First of all, calculate the average spectrum of all samples as a standard spectrum; secondly, fit the spectral data for each sample with standard spectra through one-dimensional linear regression, and then get the offset (regression coefficient) and translational (regression constant) of each spectrum relative to standard spectrum; finally, the original spectra data of each sample minus regression constant then divided by regression coefficient and you get spectrum after correction. Specific process is as follows:

（1）calculate the average spectrum：

$$\overline{A_{i,j}} = \frac{\sum_{i=1}^{n} A_{i,j}}{n}$$

（2）the linear regression：

$$A_i = a_i \overline{A_{ij}} + b_i$$

（3）multiplicative Scatter Correction：

$$A_{i(msc)} = \frac{(A_i - b_i)}{a_i}$$

3. Experiment and Analysis

We use MATLAB software to process data. After programming and debugging，spectral data of apple leaves were processed and compared and we obtained the following results:

3.1. Smoothing

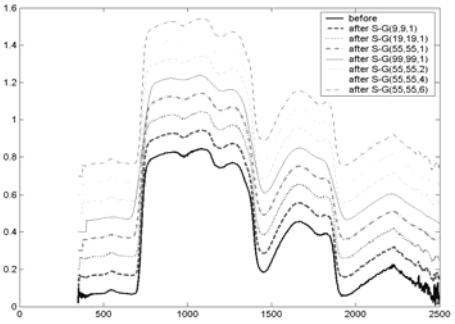

Figure 2. S-G smooth

Figure 2 is the original spectral and the spectrum which was treated by different S-G smoothing of parameters (except the original spectrum, in order not to overlap spectrum, ordinate value was added by 0.1 in proper order). We find that the original spectrum, in the 1000nm and 2000-2500nm, has relatively large noise. The spectrum also has large noise after S-G (9, 9, 1) processing. After S-G (99, 99, 1) processing, the smoothing effect is obvious, but many spectral information is lost. With the increase of the fitting polynomial order, the noise is also increased. (55, 55, 1) is the relatively ideal parameter. In practice, parameter selection is very important and we have to analyze specific issues.

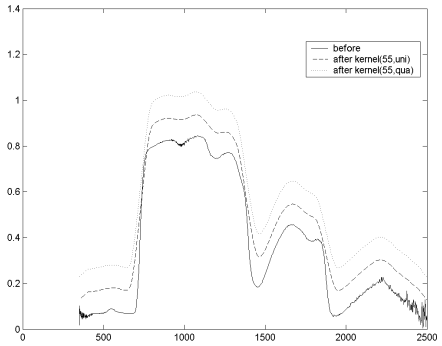

Figure 3. Kernel smooth

Figure 3 is the spectrum which was processed by kernel smoothing (except the original spectrum, in order not to overlap spectrum, ordinate value was added by 0.1 in proper order). We find, compared with the S-G algorithm, kernel algorithm has a better smooth result at the both ends of the data (Note the smoothing effect in about 400nm in Figure 2 and Figure 3). The disadvantage of kernel algorithms is the high time complexity. When the data is big, the processing speed is slow.

Standard deviation (Std Dev) is a measure of the degree of data dispersion, indicated the deviation of the data from the mean, it is also a target of variable stability. Range is the difference between the maximum and minimum data which described the dispersion of data. The range is bigger, the dispersion of data is bigger.

Table 1. The comparison of the raw data and the data which was processed by smoothing algorithm

	the raw data	S-G (55,1)	kernel(55, 'qua')	kernel(55, 'uni')
Std Dev	0.2937649	0.2919143	0.2887471	0.2861456
Range	0.8410000	0.8410000	0.0833749	0.8071000

By the comparison of Table 1, we can find, the discreteness of the data after S-G smooth is smaller than the original data; by the same smoothing window, kernel smoothing effect is better than the S-G smoothing; the effect of smooth is better when the kernel weighting function is not uniform but quadratic; and after repeated treatment, we find that variance and range change a little after 55 smoothing window. Thus, we selected kernel smoothing algorithm (the window width is 55, the polynomial weighted function is quadratic function) as the smoothing method of apple spectral data.

3.2. Derivatives

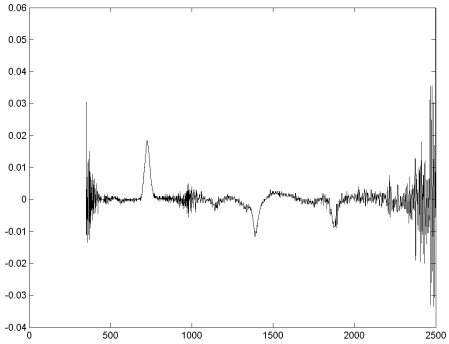

Figure 4. The first derivative spectra of the original data

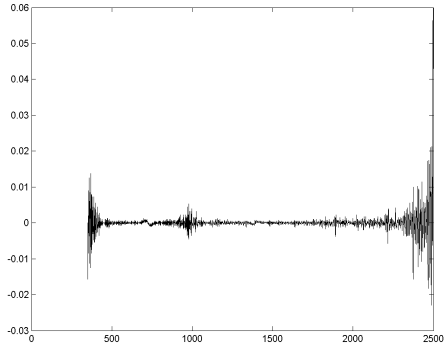

Figure 5. The second derivative spectra of the original data

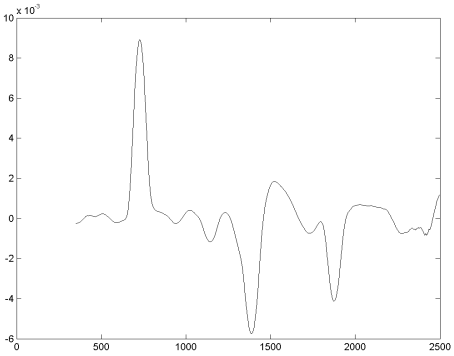

Figure 6. Kernel smoothing after the first derivative

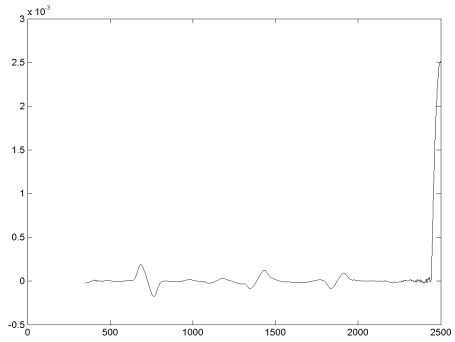

Figure 7. Kernel smoothing after the second derivative

Figure 4 and Figure 5 are the derivative spectrum of original spectrum, it has obvious absorption peaks, but without any smoothing, and very sensitive to noise, so, we have to smooth the spectral data before or after derivative. Figure 6 and Figure 7 are data which are smoothed by kernel after derivative, the spectral peaks are more obvious, and so it can do more complex spectral analysis work, and provide a good basis for the establishment of the stable model for the future. And compare with second derivative spectra, we find that the first derivative is relatively less sensitive to the noise impact than the second derivative. Therefore, the first derivative can be used for spectral preprocessing.

3.3. Multiple scatter correction

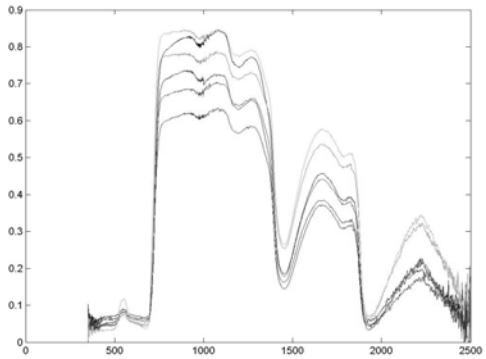

Figure 8. The original near infrared spectroscopy of apple leaves

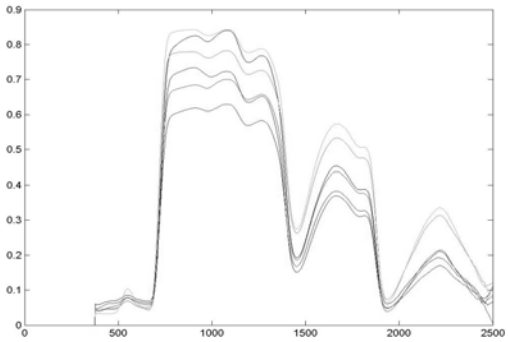

Figure 9. After S-G(55,55,1)

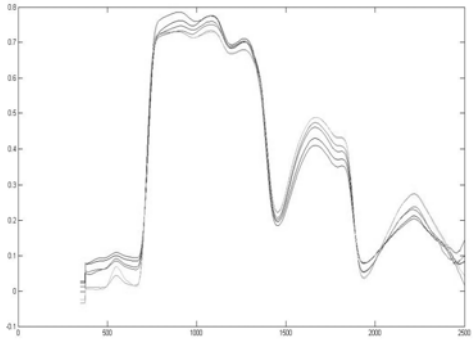

Figure 10. After multiple scatter correction

Figure 8 is a multiple spectrum of apple leaves, the impact of the apple leaves water, measuring environment and so on makes the baseline's translation and offset of the spectrum more serious and after processed by the S-G algorithm the data can be smoothed but can not be corrected(figure. 9). Figure 10 is spectral image after corrected by the multiple scatter correction, we find that the data after corrected by the multiple scatter correction have relatively consistent baseline, so it can improve the accuracy of the follow-up modeling.

4. Discussion and outlook

After the application of the above algorithm to the processing system of Near Infrared Spectroscopy in leaves, we found that:

1、S-G smoothing can be better to smooth the spectral data, while the window width and degree of polynomial selection is very important,

the more data points within the window, the more seriously the spectral resolution drops, also the more severely the spectrum distorted; But if too few data points, the smoothing result is not satisfied; the higher polynomial fitting degree, the larger data noise and smoothing effect is not very good, but the low-order polynomial fitting can not fit the data very good; so, the selection of the window width and the degree of polynomial need to be based on the actual situation. There, we choose the （55, 55, 1）as the smooth parameters for apple spectral data.

2、As the same parameters， after the kernel smoothing algorithm processing and then modeling analysis, the predicted result is more accuracy than S-G, but the time complexity is higher, therefore, we should select the smoothing algorithm according to the size of the data in practice.

3、Derivative algorithm processing results are not very satisfactory. A cause of this phenomenon is that the absorption peaks are clear, but its relatively narrow peaks are not conducive to model diagnosis.

4、Multiple Scatter Correction can improve the modeling accuracy better, and the accuracy of the results can greatly improve.

In summary, we can get the following conclusions:

The suitable pretreatment for apple leaf spectral data was found, and provided an important reference for the establishment of models.

But for the difference of the apple leaves' physical and chemical properties, spectroscopic may show some differences, so when we establish the models, we have to change the data preprocessing method at any time according to the different data.

References

[1] William H.Press,Saul A.Teukolsky, William T. Vetterling, Brian P.Flannery. *Numerical Recipes in C++: The Art of Scientific Computing, Second Edition* [M]. Cambridge University Press.2002, 2.

[2] Savitzky-Golay 法の大ざっぱな説明[EB/OL]. http://www.empitsu.com/pdf/sgd.20080718.pdf. 2008,7,18.

[3] Peter A. Gorry. General Least-Squares Smoothing and Differentiation by the Convolution (Savitzky-Golay) Method [J]. *Anal. Chem*. 1990, 62(6): 570-573.

[4] M.P.Wand,M.C.Jone. *Kernel Smoothing* [M]. Chapman and Hall/CRC. 1994,12.

[5] Zhen Ni, Chang-qin Hu, Fang Feng. Progress and effect of spectral data pretreatment in NIR analytical technique [J]. *Chin J Pharm Anal*. 2008, 28(5):824-829.

[6] Yong jun Lu, Yan ling Qu, Zhi qing Feng, Min Song. Research on the Method of Choosing Optimum Wavelengths Combination by Using Multiple Scattering Correction Technique [J]. *Spectroscopy and Spectral Analysis*. 2007, 1, 27(1):58-61.