TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.4, April 2014, pp. 2683 ~ 2689
DOI: http://dx.doi.org/10.11591/telkomnika.v12i4.4060

Comparision of Several Preprocessing Algorithms Based on Near Infrared Spectroscopic Measurement of Glucose in Aqueous Glucose Solutions

Yan Zhang*, Yawen Deng, Jinwei Sun, Chunling Yang, Guoliang Zhang, Dan Liu
School of Electrical Engineering and Automaton, Harbin Institute of Technology
92, West Dazhi Street, NanGang District, Harbin 150001, China
*Corresponding author, e-mail: zyhit@hit.edu.cn

Abstract

Glucose concentration measurement is the basis of noninvasive detection of blood glucose concentration. It is significant in scientific research. In this study, Near Infrared Spectroscopy (*NIRS*) *and regression analysis methodology were combined to measure the glucose concentration. The spectrum of glucose solutions was obtained with the Fourier Transformed Infrared Spectrometer, and then the data was used for regression analysis. In addition, the method of Partial Least Squares* (*PLS*) *was used to achieve principle components and various spectral preprocessing methods were discussed. During PLS modeling, the Savitzky-Golay could improve the Prediction Residual Error Sum of Squares* (*PRESS*) *within 6%. The experiment results demonstrate that NIRS has the potential for the measurement of glucose solution.*

Keywords*: near-infrared spectrum, partial least squares, spectral preprocessing, predicted residual error sum of squares*

1. Introduction

Glycuresis is one kind of global illnesses and it seriously harms the healthy of human beings. Seriously diabetic patients must measure their blood glucose contents several times one day in the presently used therapy. Instruments now used for the self-monitoring of blood glucose are almost all invasive types that require a drop of blood to be withdrawn from a fingertip or other measurement site on the body by a needle puncture. This requires the diabetic patient to suffer pain and also involves a risk of infection. More frequent or continuous blood glucose monitoring is necessary for distinct blood glucose control, which would more effectively reduce the risk of complications from diabetes mellitus. For this purpose, a noninvasive method for blood glucose monitoring is highly desired.

Near infrared spectroscopy (NIRS) has been known to have the potential to realize noninvasive blood glucose monitoring, and there have been many trials for monitoring blood glucose contents using NIRS over these years. In the stoichiometric analysis field, the researchers have taken up the related research. Gary W.Small team from America [1] and Kasemsumran S team from Japan [2] and other researchers strive to be the first one to dig deep in the application area of near-infrared spectrum. The Chinese researchers like Xu Ke-Xin in the Tianjin University [3], Huang Lan in Shanghai [4] also have some relate studies in developing instruments based on near infrared spectroscopy. However, there are some problems which are difficult to desovle and this impede the development of non-invasive measurement of blood glucose.

In the present study, near infrared spectra of different glucose concentrations were collected with a FT-IR spectrometer. Then several spectral preprocessing methods and the PLS algorithm were used to analyse the data. Finally, the experimental results were evaluated with the Prediction Residual Error Sum of Squares.

Received August 6, 2013; Revised October 18, 2013; Accepted November 7, 2013

2. Theory

2.1. Absorption Characteristics of Chemical Bond

A molecular bond vibration absorsb near infrared light and it is the principle of application of NIRS. In this study, the most notable point is a frequency-doubled and combination-tone bond vibration of methyl in glucose molecules. As a result, the absorption peak of water should be avoided. This absorption band is initially selected over the 4200 to 4800 cm^{-1} spectral range with samples maintained at room temperature. The different preprocessing algorithms were evaluated by judging the ability to determine glucose concentrations from a set of prediction spectra.

2.2. Lambert-Beer Law

The continuous wave spectroscopy was used to realize glucose concentration measurement, Lambert-Beer Law is the macro base of the aborption process. The Lambert-Beer Law [5] could be defined as:

$$A = \sum_{i=1}^{n} \alpha_i(\lambda) c_i L \quad (1)$$

Where A is the vector of absorbance, n is the number of various solutes being observed, α is the coefficient related to the specific wavelength λ and L is the optical distance.

According to the Lambert-Beer Law, glucose concentration can be acquired once the spectrum is obtained. The absorbance of near infrared spectroscopy is positively related to glucose concentration, which is the basis of the measurement of glucose concentration.

3. Experiments

3.1. Instruments

Near-infrared spectroscopic data were measured with a JASCO FT-IR spectrometer. Samples were contained in 1mm quartz colorimetric dish.The electronic balance Sartorius BS 224S was used to quantify the glucose.

3.2. Reagents

For the NIR data, reagent-grade crystalline glucose was dissolved in deionized water to configure 12 samples of different glucose concentration. Electronic balance has the range ability of 220g and the measurement accuracy is ±0.1mg. A 500mL volumetric flask was used for mother liquor and the 100mL ones for samples. During dilution, measuring cylinder with accuracy of ±0.2mL was used. Table 1 shows the concentration of samples.

Table 1. Number of Samples and Corresponding Concentration

number of samples	1	2	3	4	5	6	7	8	9	10	11	12
concentration（mg/dL）	100	400	420	550	350	250	270	30	450	150	80	300

3.3. Procedures

Firstly, spectrum of the empty infrasil glass cells and cells filled with samples were collected seperately. Secondly, spectrum data were imported to the computer and were calculated with the Matlab software. The band between 4200~4800cm^{-1} were selected and used to generate the target array. Finally, glucose absorption ability can be calculated by using average value from homologous data. For example, K_i is the average spectrum of the nth sample, K_{0i} is the average spectrum of the nth empty glass cell, then the corresponding glucose absorption can be defined as:

$$A_i = \ln\left(K_{0i} / K_i\right) \quad (2)$$

4. Spectral Data Preprocessiing

Near-infrared spectrum reflect the information about the chemical composition and the concentration of substances. It also can be affected by material viscosity, particle density and stray light, etc. Therefore, the elimination of these factors could improve the performance of the measurement. The original spectrum collected with the JASCO FT-IR spectrometer is shown in Figure 1. Several preprocessing methodologies were introduced to analyse the spectrum.

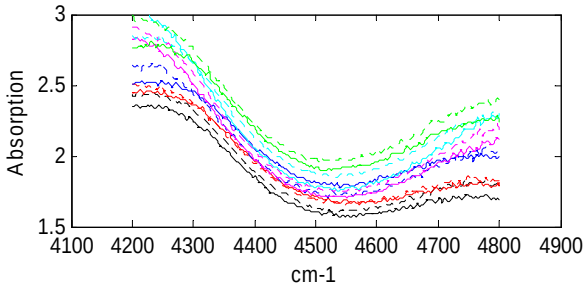

Figure 1. Original Spectrum

4.1. Standardization

Considering the obvious spectrum difference among the different wavelengths, the nonlinearity of the detector could cause different measurement error. Standardization is used to degrade its effects on the model. Autoscaling is one kind of standardization. Centering and normalization are two steps of this process.

During centering, the absorbance spectrum data at the same wavelength points but from different samples subtract the average value, indicated as expression (3):

$$X' = X - \mu \tag{3}$$

And the normalization result X_z can be expressed as:

$$X_z = \sigma X_i' = \frac{1}{\sqrt{\frac{1}{m-1}\sum_{j=1}^{m} x_{ij}}} \tag{4}$$

X_i' is the data on the *i*th wavelength, σ is the standard deviation of the *i*th wavelength, m is the number of samples, x_{ij} is the absorption of the *j*th sample at the *i*th wavelength points. After the standardized processing, the spectrum can be acquired. Figure 2 shows the spectra after standardization .

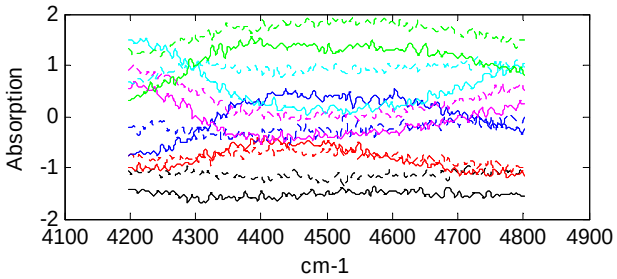

Figure 2. Spectra after Standardization

4.2. Savitzky-Golay

The application of Savitzky-Golay is based on the assumption that the noise contained in the spectrum is white noise, which can be degraded by calculating the spectral data of

adjacent wavelength points. During this processing, the data in the mobile window is smoothed by different polynomial sequences. However, the width of the moving window must be carefully choosed, otherwise some useful information could be lost and the processing method can not obtain the ideal results. Figure 3 shows the spectrum processed using window width of 10 and 100 wavelength points in turn.

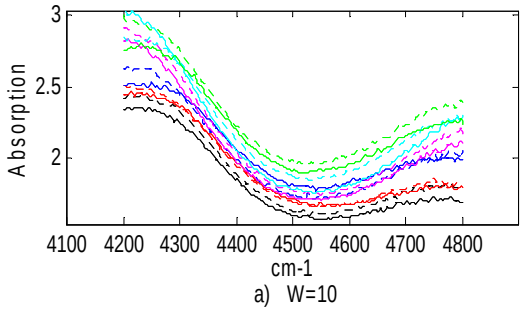

Figure 3. a) Spectrum processed using Savitzky-Golay with Window of 10 Wavelength points.

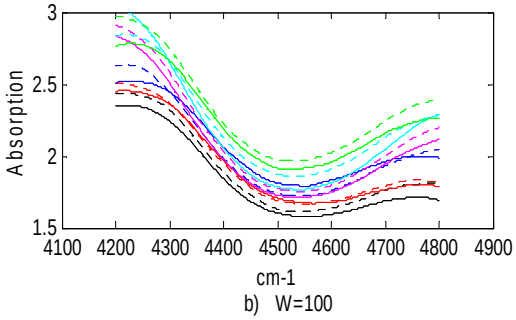

Figure 3. b) Spectrum processed using Savitzky-Golay with Window of 100 Wavelength points

4.3. Direct Orthogonal Signal Correction (DOSC)

The above spectral preprocessing methods carry out the data processing without computing density matrix, only relating to spectra data. During DOSC method, the spectral array vary correspondingly and is orthogonal to the concentration [6]. After the multivariate calibration, the model present more robust and prediction ability [7].

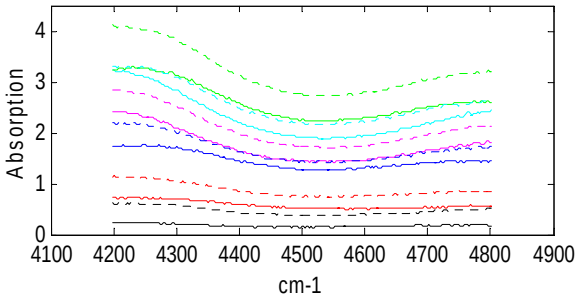

Figure 4. Spectra Processed by DOSC Method

If Y present the concentration and X present the spectral array, DOSC can be divided into the following four steps.

(1) Project Y on X, $\hat{Y} = P_X Y = X((X^{\mathrm{T}})^{-1})^{\mathrm{T}} Y$

(2) $Z = X - \hat{Y}((\hat{Y})^{-1} X)$, the step ensure that Z is orthogonal to Y and $\hat{Y}$.

(3) Z is processed by Principal Component Analysis(PCA) and the score matrix *t* can be acquired. The weight vector $w = X^{-1} t$, then we can calculate score vector $t_s = Xw$,and the loading vector $p_s = \frac{X^{\mathrm{T}} t_s}{t_s{}^{\mathrm{T}} t_s}$.

(4) $X_{\mathrm{DOSC}} = X - t_s p_s{}^{\mathrm{T}}$

In the above steps, there are two main components were used during PCA. After the DOSC, most information about the sample characteristics is lost and the spectrum about the concentrations can be arranged. The spectra processed by DOSC method is shown in Figure 4.

5. NIRS and PLS

5.1. PLS Theory

PLS is a method combining factor analysis and regression analysis. It has two steps.

Step 1: Factor analysis. Decompose X and Y.

$$X = VP^{\mathrm{T}} + E \tag{5}$$

$$Y = UQ^{\mathrm{T}} + F \tag{6}$$

Where superscript T means transposed matrix, V is the score matrix of X, U is the score matrix of Y, P is loading matrix of X, Q is loading matrix of X, E and F are the error matrix. It is important to note that T is orthogonal to P. And t_i reflects the information of spectrum matrix X when it was conveyed by vectors p_i.The remaining information is considered to be included in the error matrix [8-9]. In a similar way, Y is decomposed.

Step 2: Regression analysis. B is the correlation coefficient matrix.

$$U = VB \tag{7}$$

$$B = (V^{\mathrm{T}}V)^{-1}V^{\mathrm{T}}U \tag{8}$$

The predicted value of unknown concentration Y_p can be defined as:

$$Y_P = T_P BQ \tag{9}$$

Where T_P is got from the spectrum of unknown samples and the loading matrix P.

5.2. Determination the Number of PLS Components

During the PLS modeling, the number of components is an important element. At present, the most common method to determine the number of PLS components is Prediction Residual Error Sum of Squares (PRESS). The cross validation method is used to analyse PRESS [10].

Neglecting the ith wavelength points every time, build PLS model with h components by using the rest data. Then plug the ith wavelength points into regression equation and get $\hat{x}_{(i)j}(h)$. The forecasting error square sum of x_i can be defined as:

$$PRESS_j(h) = \sum_{i=1}^{n} (x_{ij} - \hat{x}_{(i)j}(h))^2 \tag{10}$$

Where j =1, 2, ……, p.

The forecasting error square sum of X= (x_1, x_2, ……, x_p)$^{\mathrm{T}}$ can be defined as:

$$PRESS(h) = \sum_{j=1}^{p} PRESS_j(h) \tag{11}$$

At the same time, we build the PLS model with h components by using all data. $x_{p,ij}$ is the redicted value of the ith wavelength point. The forecasting error square sum of x_i can be defined as:

$$SS_j(h) = \sum_{i=1}^{n} (x_{p,ij} - x_{ij})^2 \tag{12}$$

The forecasting error square sum of x_p can be defined as:

$$SS(h) = \sum_{j=1}^{p} SS_j(h) \quad (13)$$

When the minimum PRESS(*h*) is achieved, the appropriate number of the PLS components can be determined. It is defined as:

$$Q_h^2 = 1 - PRESS(h) / SS(h) \quad (14)$$

If the condition that $Q_h^2 \geq 0.0975$ can be achieved by using h components, the computing process stoped.

6. Results

This paper discussed several methods for spectral preprocessing. Table 2 shows the experiment results. From the tables, it can be obtained that Savitzky-Golay diminished the root mean square residual and maximum relative error.

Table 2. Parameters of The First Prediction Set

Processing Method	Number of Components *r*	Maximum Relative Error	RMSE (mg/dL)
nothing	4	0.0539	11.85875
Autoscaling	5	0.0671	12.64839
DOSC	3	0.0725	17.77251
S-G	4	0.0524	11.59880

In this study, The DOSC and autoscaling methods did not exhibit good performance. This may be denpendent on the characteristics of the data. Savitzky-Golay showed the best result in our experiments when the width of moving window was 15 wavelength points and the number of PLS components was 4. The results presented in the table demonstrated that the maximum relative error is less than 8% and the maximum RMSE is less than 18mg/dL.

7. Conclusion

In this study, several preprocessing methods and PLS were combined to measure the glucose concentration.We discussed several preprocessing methods in advance. By using autoscaling, DOSC, Savitzky-Golay, the *PRESS* is calculated. The maximum relative error was confined to 8%. The experiment results could demonstrate that the appilcation of NIRS and PLS has the potential for the measurement and analysis of glucose solution.

Acknowledgements

The authors are grateful for the support from the National Science Foundation of China (No. 61201017, 61378046), China Postdoctoral Science Foundation (No. 2013M531027), Heilongjiang Postdoctoral Fund (No. LBH-Z12093), the Fundamental Research Funds for the Central Universities (No. HIT.NSRIF.2013010, No. HIT.NSRIF.201146).

References

[1] Mark AA, Gary WS. Determination of Physiological Levels of Glucose in all Aqueous Matrix with Digitally Filtered Fourier Transform Near-Infrared Spectra. *Analytical Chemistry.* 1990; 62(14): 1457–1464.
[2] Bai G. The Measurement and Analysis of Glucose Concentration Changes in Blood with Near-infrared Spectroscopy. Wuhan: HUST. 2008.

[3] Li QB, Wang Q, Xu KX, Wang B. Near-Infrared Spectroscopic Assay of Principal Milk Constituents. *Food Science.* 2002, 23(6): 121–124.
[4] Huang L, Ding HS, Wang GZ. The Preliminary Study on Noninvasive Detection Using NIR Diffusion Reflectance Spectrum for Monitoring Blood Glucose. Shanghai Institute of Metallurgy, Academy of Sciences of China. *Materials Physics and Chemistry.* 2000.
[5] Yuan HF, Lu WZ. Near infrared spectral analysis technology is rapidly marching into the petrochemical field. *Oil Refining and Chemical Industry*. 1998; 29(9): 47–50.
[6] Johan AW, S deJong, Smilde AK. Direct orthogonal signal correction. *Chemom. Intell. Lab. Syst*. 2001; 56: 13–25.
[7] Liu GJ, Dai DM, Gao HT. The algorithms of orthogonal signal correction and its application in spectra processing. *J. Shandong Univ. Arehit. Eng.,* 2005; 20(2): 85.
[8] Gao J, Zeng XP, Zhang X, Chen X, Zheng DT. Cavity Vertex Regeneration through Optimal EnergyModel for Restoration of Worn Parts. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2013; 11(5): 2490–2501.
[9] Zhang XS, Wang MH, Ma J. Sparse Representation for Detection ofMicrocalcification Clusters. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2013, 10(3): 545-550.
[10] Rajkumar P, Wang N, EImasry G, Raghavan GSV, Gariepy Y. Studies on banana fruit quality and maturity stages using hyperspectral imaging. *Journal of Food Engineering*. 2012; 108(1): 194–200.