The determination of serum cholesterol concentration with improved differential evolution algorithm based on ultraviolet-visible absorption spectrum

Mingcheng Gui[1,a], Weihua Zhu[1,b], Feng Zhu[2,c], Ying Geng[3,d], Weihao Hua[1,e], Chunmei Tang[1,f] and Zhimin Zhao[4]*

[1] School of Science, Hohai University, Nanjing 210098, China;

[2]CCCC AIRPORT INVESTIGATION AND DESIGN INSTITUTE Co., Ltd. 510000, China;

[3]CCCC-FHDI ENGINEERING CO., Ltd. 510000, China;

[4]School of Science, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China.

[a]mc_gui01@163.com, [b]weihua_zhu@126.com, [c]zf199095@126.com

[d]gengyingwilla@163.com, [e]997185774@qq.com, [f]tcmnj@163.com

* Correspondence author. E-mail: zhaozhimin@nuaa.edu.cn

Keywords: Cholesterol; ultraviolet-visible absorption spectrum; Partial least square; Differential evolution; Multiple correlation

Abstract. This article uses the partial least square algorithm and the improved differential evolution algorithm to preprocess the ultraviolet-visible absorption spectral data and build a model for detection of cholesterol concentration in human serum. The result indicates that the wavelet decomposition and the partial least square algorithm can effectively reduce the multiple correlation coefficients. The absorption spectrum is investigated by different basis functions with various pre-processed spectra, and produces the best result with the prediction error of 0.12mmol/L. This study shows that the ultraviolet-visible absorption spectroscopy can offer a feasible research direction for detection and quantitative analysis of cholesterol concentration.

Introduction

As one of the most important component of human serum, cholesterol is indispensable for participating in the cytomembrane formation and is the raw material of bile acid, vitamin D and steroid hormones. Most tissues are permitted to gain hormone from the precursor substance, which is exactly cholesterol in serum. The concentration of serum cholesterol outside the normal range may be an indicator of a medical condition. In recent years, hypertension, hyperlipidemia and hyperglycemia have caught much attention in modern life because of its high incidence and difficulty to control. The rising incidence of three-hypers has promoted the prosperity of the market of blood component detection. Traditionally, the components has been detected by chemical and biochemical methods, however, the process were relatively complicated, time-consuming and yielded results with low accuracy. Thus, it is highly desirable to build an effective and convenient model of serum cholesterol concentration to offer a dependable way for detection of serum cholesterol level[1].

As an efficient way for detecting the concentration of serum components, the spectrographic analysis is a potential alternative method for a rapid measurement of cholesterol concentration in human serum, which can obtain the information of cholesterol concentration quickly and easily. Lan et al. has studied the absorption characteristics of normal human serum and hypercholesterolemia serum, the result displays the apparent difference of spectral curve shape with variational cholesterol concentration[2]. Zhu et al has done research on the corresponding wavelengths to the cholesterol concentration[3]. This offers the possibility of quantitative analysis for cholesterol component in serum by absorption spectrum. The usually used methods are major based on the infrared spectroscopy, near-infrared spectroscopy and so on[4,5]. This paper used the

ultraviolet-visible absorption spectrum ranged from 300nm to 700nm from the human serum samples, which covers the spectral region of ultraviolet and visible light.

Many researchers are devoted to chemometrics by the method of spectroscopy, which catch significant attention in clinical field, but there are some problems remained for practical application[6,7]. Various methods of spectral data pre-processing and background removing, as well as proper variables selection and model calibration methods need to be further studied dealing with above problems.

Materials and methods

Instrumentation. The *UV-3600* ultraviolet-visible spectrophotometer made by Japan's *SHIMADZU* Corporation and several personal microcomputers were adopted as the experimental apparatus. The parameters of the spectrophotometer were set as follows: the wavelength coverage for test is from 300~700nm with step-length of 1nm. The slit width is 0.2nm, the scanning speed is set as fast, of which the interval is 0.5nm.

Samples and Reagents. 36 human serum samples obtained from the *Hospital of Nanjing University of Aeronautics and Astronautics* are analyzed in this work. Blood donors who are not permitted to intake food before drawing blood are selected randomly. The concentration of the serum cholesterol in these samples is measured by traditional ultracentrifugation method, which ranged from 3.53 to 8.4mmol/L, with the mean value of 5.5289mmol/L. The third and the fourth detail wavelet-decomposed signals of absorption spectrum are obtained by programming process. The samples are divided randomly into three sets: calibration set containing 24 samples, validation set containing 6 samples, and test set containing 6 samples. The calibration set is used for building the model. The validation set is used for model construction and convergence by the fitness function of differential evolution algorithm. The test set produces the results to test the prediction ability of the model.

Wavelet Decomposition. As a kind of data decomposition method with the advantage of flexibility, wavelet decomposition is used to denoising and obtain detailed signals of spectrum information in this paper[8]. Because the absorptance of blood cholesterol in human serum is relatively less than other component in absorption spectrum, the detailed signals of wavelet decomposition are chosen to extract information for content of cholesterol, which has proven to be feasible in the analysis of human serum spectrum.

Partial least square regression algorithm. Partial least squares regression is a superior tool in spectroscopy analysis, which is composed of multiple linear regressions, traditional correlation analysis and principal component analysis. It creates orthogonal components in order to extract the useful information that can explain the functional relationship between explanatory variables and corresponding outputs. Partial least squares regression can obtain limited chemically interpretable spectral information[9]. It has been proved that it is significant to be used when the number of explanatory variables is greater than the number of observations and there is high multicollinearity among the existing variables[10,11].

The essential point to make the model with a better ability for both calibration and prediction is the determination of the number of principal components, which are used for model construction. Less number of principal components will lose the capacity of calibration of inputs, while the more will cause a substantial risk of "over-fitting", which means a well-fitting and complex model with little prediction ability[12]. To solve this, the *Monte Carlo* cross validation was considered to ascertain the complexity of the PLS model.

Considering the nonlinearity and multicollinearity of the spectral data, nonlinear transformation is required to optimize the model. The model with kernel function can map nonlinear data in a higher dimensional feature space, in which the spectral data can be quasi-linear. And this extension of model will contribute to the information extraction and help to build a model with better prediction ability.

The equation (1) presents equation for the regression model.

$$y = \beta_0 + \sum_{j=1}^{p} \sum_{l=1}^{M_j+2} \beta_{j,l} \Omega_k \left(\frac{x_j - \xi_{j,l-1}}{h_j} \right) + \varepsilon \quad (1)$$

$\Omega_k(x)$ is the basis function (linear, cubic B-spline, Gaussian), which preprocesses the original spectral data to be the new inputs in the model. $\beta_{j,l}$ is the coefficient for the first basis function. M_j 、 h_j and $\xi_{j,l-1}$ are the number of interpolation nodes with equal intervals, the length between the adjacent nodes, and the points on the interpolation nodes. The three basis function are showed in equation (2), (3), and (4) in the following:

$$\Omega_1(x) = x \quad (2)$$

$$\Omega_2(x) = \begin{cases} 0, (|x| \geq 2) \\ \frac{1}{2}|x|^3 - x^2 + \frac{2}{3}, (|x| \leq 1) \\ -\frac{1}{6}|x|^3 + x^2 - 2|x| + \frac{4}{3}, (1 < |x| < 2) \end{cases} \quad (3)$$

$$\Omega_3(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2} \quad (4)$$

Improved Differential Evolutionary Algorithm. Band selection requires an algorithm that can find the global optimal result in the whole spectrum. Differential evaluation algorithm (DE) is an effective strategy to select the wavelengths among the whole spectrum according to the result of PLS algorithm. Here we use the improved differential evaluation algorithm which is called self-adaptive differential evolution with global neighborhood research (NSSDE) proposed by Guo et al. for reference[13]. For the sake of improvement of the DE algorithm, proper determination of parameters is fundamental and expected. However, determination of fundamental parameters after experimental attempts are fairly time-consuming in traditional calculation procedure. The paper proposed the parameter self-adaptive property in order to save time cost. The essential control parameters are mainly scale factor F and cross rate CR in the traditional DE algorithm, thus they are determined after every iteration course. Also, considering the slow convergence when facing a complex problem with DE algorithm, the paper has put forward the global neighborhood research strategy to improve the efficiency of optimization process. From the result shown in the reference article, convergence speed is apparently accelerated when solving optimization problems when compared to traditional DE algorithm and other reference improved DE algorithm.

There are a certain number of wavelengths derived from the spectral range of 300~700nm are used as a free solution in the first generation. These wavelengths are chosen randomly, due to the ignorance of initial solution with improved differential evolution algorithm. In the process of every iteration course, the PLS model is implemented for each individual of the generation. The reciprocal of root-mean-square error is the fitness set for evaluating the results of calibration, validation and test set. The mutation of individual is according to the randomly selected three other individuals in the population, of which the detailed procedure explanation can be seen in the Guo's paper.

Results and discussion

To build the model of blood cholesterol level detection, three aspects that need to be noted are data nonlinearity, multi-correlation of absorption spectrum and waveband selection. Firstly, the information of concentration of substance does not always present linear relationship with the corresponding spectral absorptance. To solve this, two basis functions (B-spline function and Gaussian radial function) for nonlinear transformation of absorption spectrum data are selected in this study, together with the linear condition for comparison. The utility of basis function is evident to improve the prediction ability of serum cholesterol concentration.

Secondly, the nonlinear relationship also exists among a wide range of wavelengths. That is, the absorption of each wavelength is correlated with a series of nearby or relative wavelengths, causing the strong overlapping of absorption spectrum, which is called multiple correlations. Correlation of regressors can dramatically interfere the performance of a regression model, and results in large Multiple variances and covariance for the estimators of least squares of the regression coefficients[10,12]. The wavelet decomposition was chosen to reduce the multiple correlation among the spectrum with traditional *Daubechies* function as generating function[14], which can process the absorption spectral data to construct three derived spectra. The original, the 3rd as well as the 4th detail signal of spectral data are studied. Another used method is the Partial least square regression algorithm that can effectively eliminates the multiple correlation among the spectrum. Based on the higher multiple correlation among the spectra and the nonlinear factors, the nonlinear model is built between the absorption spectrum and concentration of triglyceride by using B-spine and Gaussian interpolation basis function[15].

Thirdly, the analysis of whole spectra does not always yield optimal results due to the interference of irrelevant signals in some regions, therefore, the waveband selection and irrelevant variates elimination is essential for the model for concentration prediction. Researchers have put great effort in diverse variates selection and elimination methods, among which significant and improved model for robustness and accuracy are obtained eventually. Plenty of researches demonstrate that proper methods or combination of them can effectively balance the model size reduction and capability of prediction performance[16-19].

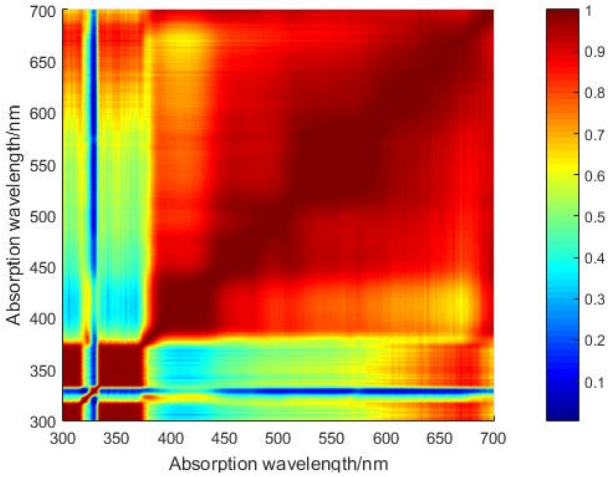

Figure 1. Absolute value cloud of multiple correlation coefficient of original ultraviolet-visible absorption spectra

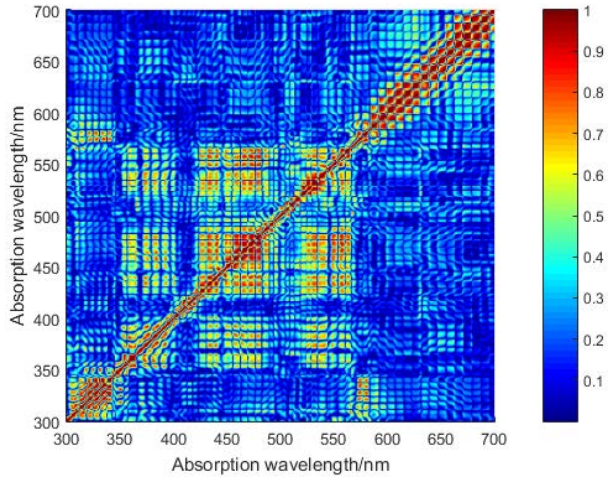

Figure 2. Absolute value cloud of multiple correlation coefficient of the 3rd detailed absorption spectra

Figure 1 shows the absolute value cloud of multiple correlation coefficient of original ultraviolet-visible absorption spectra that ranges from 300nm to 700nm. Apart from some regions

that has a low correlation coefficient, high multiple correlation is scattered over two thirds of the whole spectrum. This surely cause a negative effect on the quantitative spectral analysis. The ultraviolet-visible absorption wavelengths ranges from 380nm to 700nm related significantly to that with each other, of which the correlation coefficient is over 0.8. The area that mainly distributed around the X-Y 45° line has the correlation coefficient approaching 1, giving rise to the multiple correlations among the nearby wavelengths. Traditional *Daubechies* function based wavelet decomposition is the way we choose for reducing multiple correlation among the variates. The detail signal derived from original spectrum demonstrates the detailed information in absorption spectrum and eliminates the background spectral interference. Figure 2 presents the absolute value cloud of multiple correlation coefficient of the 3rd absorption spectral signal. Apparently the multiple correlation is much lower than that of the original spectrum. The correlation coefficient of most areas is below 0.5. This is the illustration that the original absorption spectrum has been successfully exacted detailed signals which has a fairly low multiple correlation.

The PLS factors represent for the complexity of the modeling and the dimensions for projections by specified dependent variable. So it decides the time cost for the whole calculation. If it is too low, there is not enough extraction ability for useful information. Otherwise, there may be “over-fitting” in modeling process when the factors are too high. Over-fitting means that the training course extract the information overmuch and do not yield the same good prediction result. In the self-adaptive property, root-mean-square error is not the only criterion for evaluation after a round of calculation, instead, evaluate the result with both root-mean-square error and the difference value of correlation coefficient of training set and test set. Lower error for prediction does not always account for truly good result when the prediction result is much better than that of the training set. This is probably due to contingency caused by the one-goal convergence calculation strategy. Thus the property added consider to find the lowest prediction error with the avoiding of over-fitting. The PLS factor is determined in every round according to the previous calculation results. It is decided as the factor plus one of the previous best evaluation result. As figure 3 shows, the PLS factors gradually become stable in the 3rd detail spectra modeling process, which is the best determination for PLS factors in the corresponding situation.

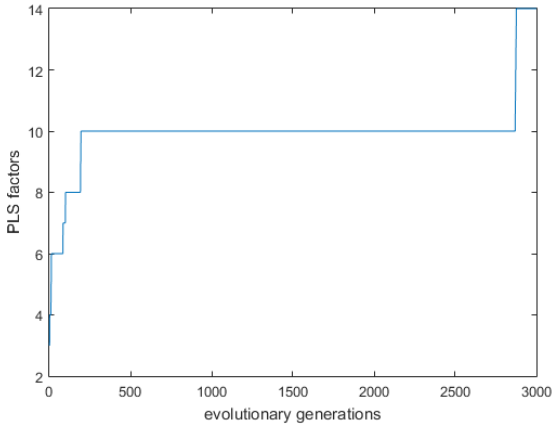

Figure 3. Variation tendency of PLS factors with the self-adaptive property

Table 1 gives the modeling results calculated by the PLS-NSSDE model with different derived spectral data by various basis functions. RMSEC signifies for the root-mean-square error of calibration samples, RMSEP for the correlation coefficients of test samples and R^2 for the coefficient of determination between independent and dependent variables in test set. And the factors represent for the number of principle components in PLS model.

Table 1. Modeling results of the original and derived ultraviolet-visible absorption spectra with different basis functions.

	linear				spline				gauss			
	RMSEC	RMSEP	R^2	FACTORS	RMSEC	RMSEP	R^2	FACTORS	RMSEC	RMSEP	R^2	FACTORS
original	0.56	0.43	0.89	13	0.27	0.37	0.89	11	0.12	0.27	0.97	14
3rd detail layer	0.72	0.90	0.71	16	0.01	0.12	0.97	10	0.23	0.27	0.91	11
4th detail layer	0.59	0.45	0.82	11	0.10	0.28	0.95	8	0.11	0.34	0.89	11

It is shown that the linear condition yielded pretty bad results with any of the three spectra or derived spectral data, which may be the result of nonlinearity correlation of the spectra and the concentration of cholesterol. The RMSEP value is unsatisfying for prediction of test samples. The concentration of cholesterol proves to be nonlinear relationship with absorptance in the model. Similar to that, the other condition yielded results with the RMSEP value over 0.25mmol/L in all the occasions with different basis functions, except for the spline function based condition with the 3rd detail layer spectra. The 3rd signal of spectra with spline function produces fairly good results with the R^2 value of 0.98. That demonstrates that the capability of wavelet decomposition in extraction of information of cholesterol content is feasible to research. And one thing to be addressed is that the results of detail signal spectra is more convincing in modeling performance than that of the original, this is also perhaps due to the background signals removing and noise elimination. The 3rd detail signal with spline basis function perform the best in all the occasions. The 3rd detail signal with spline basis function yield the RMSEP of 0.12mmol/L and the R^2 of 0.98, as figure 4 shows. Figure 4 shows the straight fitting-line of the training results and test results. The predictive concentration is close to the reference concentration, and the fitting-line approaches to the reference line.

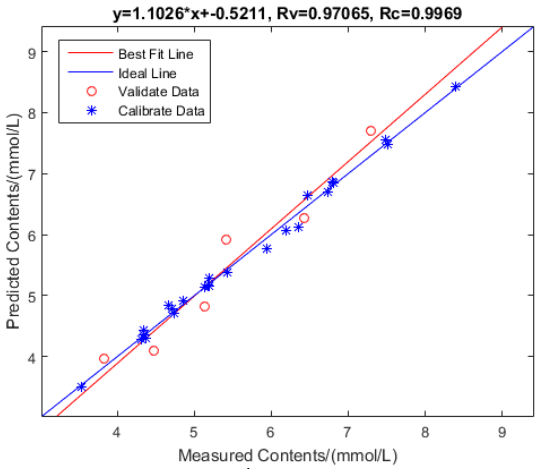

Figure 4. The best result yielded by the 3rd detail spectra with the spline basis function.

Each iteration of NSSDE algorithm produces a combination of wavelengths that participated in the model training, validation and test procedure. Every iteration course returns a solution with a better result, but not the optimal one. Due to that, the statistical analysis procedure needs to be considered. A concept called Spectral information density is proposed to record the selection times of wavelengths in all rounds of the algorithm. Each wavelength is probably chosen in each time, the executed total times of algorithm are 3000, which is enough to avoid the contingency that the selection frequency of irrelevant wavelength is high at the end. The figure 5 shows the information distribution of emission wavelength according to the selected frequency in the modeling process.

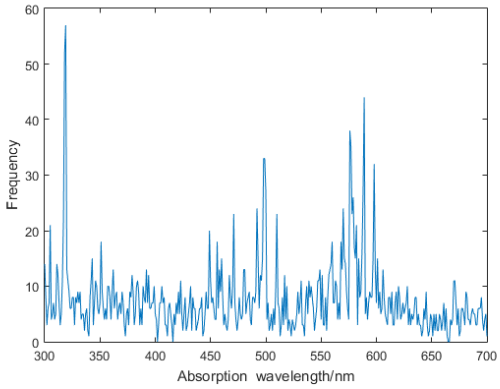

Figure 5. The selection frequency of the whole absorption wavelength range.

The optimal modeling waveband is made up of two parts: (1) The combination of wavelengths with minimum root-mean-square validation error. (2) The wavelengths that have a high peak in information density distribution. PLS algorithm is executed in the optimal modeling waveband to make the last model training and in the meanwhile, substitute the test samples in the model to give a convincing result for its capability of prediction. These ranges of spectra may benefit a lot in the model building, and they may also provide with the physicochemical properties through absorbance of the cholesterol in serum.

The result shows its powerful prediction ability of PLS in the aspect of information extraction in spectral analysis. NSSDE algorithm successfully picks out the wavelengths combination for modeling according to the training of calibration samples. The key points of prediction model are the method of variates selection and data pre-processing in the research of spectroscopic quantitative analysis. PLS-NSSDE model narrows the scope in the whole absorption spectra to predict the concentration of cholesterol with human serum samples, which also eliminates the interference of irrelevant and other noise signals. The other blood cholesterol research is mostly based on the methods of uninformative variates elimination for the promising calibration model to give better results. This needs proper pre-process of spectral data so that the residual variables may correctly correspond to the target variates for modeling. Considering the occasional errors for experimental accuracy caused by the health conditions of volunteers and other inevitable reasons, adequate samples with different conditions and exacted wavelengths for corresponding concentration information are needed.

Conclusions

Based on the ultraviolet-visible spectroscopy with human serum samples of 36 random persons, the partial least square regression and self-adaptive differential evolution algorithm with neighborhood research strategy were employed in the model building with different derived spectral data and various basis function. The built prediction model with a fairly good precision for prediction has the significant improvement for the cholesterol detection in human serum.

With the process of data pretreatment, the result demonstrates that the 3rd detail signals contains rich information of serum cholesterol content. It means that wavelet decomposition is a feasible method to reduce negative effects of multiple correlations. The model of the 3rd detail signal with spline basis function produces the best modeling result with the lowest root-mean square error in the whole calculation. The least RMSEP is 0.12mmol/L for prediction error, and with the RMSEC of 0.01mmol/L.

Actually more research needs to be done in the future to improve the stability of the model. With the high precision of results obtained by PLS-NSSDE model for cholesterol level detection, wavelet decomposition for spectral data pre-processing and using NSSDE algorithm to search for the cholesterol information distribution can be considered to be a direction for clinical cholesterol determination.

Acknowledgment

Project supported by the Foundation of National Natural Science (Grant No. 61378037) and the National Innovation experiment program of China (Grant No. 2015102941132).

References

[1]. Chou L C S, Liu C C. Development of a molecular imprinting thick film electrochemical sensor for cholesterol detection[J]. Sensors & Actuators B Chemical, 2005, 110(2):204-208.

[2]. Lan X F, Ying L, Zhu T, et al. Spectroscopy Analysis of Total Cholesterol in Human Serum[J]. Acta Photonica Sinica, 2008, 37(3):547-551.

[3]. Zhu W H, Zhao Z M, Guo X. Study of Cholesterol Concentration Based on Serum ultraviolet-Visible Absorption Spectrum[J]. Guang pu xue yu guang pu fen xi = Guang pu, 2009, 29(4):1004-7.

[4]. Peuchant E, Salles C, Jensen R. Determination of serum cholesterol by near-infrared reflectance spectrometry.[J]. Analytical Chemistry, 1987, 59(14):1816-9.

[5]. Chen J, Arnold M A, Small G W. Comparison of combination and first overtone spectral regions for near-infrared calibration models for glucose and other biomolecules in aqueous solutions.[J]. Analytical Chemistry, 2004, 76(18):5405-13.

[6]. Arnold M A, Small G W. Noninvasive glucose sensing.[J]. Analytical Chemistry, 2005, 77(17):5429-39.

[7]. Chen X D. Possibility of noninvasive clinical biochemical examination by near infrared spectroscopy[J]. Guangxue Jingmi Gongcheng/optics & Precision Engineering, 2008, 16(5):759-763.

[8]. Santos R N F D, Galvão R K H, Araujo M C U, et al. Improvement of prediction ability of PLS models employing the wavelet packet transform: A case study concerning FT-IR determination of gasoline parameters[J]. Talanta, 2007, 71(3):1136-1143.

[9]. Haaland D M, Thomas E V. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information[J]. Analytical Chemistry, 1988, 60(11):1193-1202.

[10]. Rosipal R, Trejo L J. Kernel partial least squares regression in reproducing kernel hilbert space[J]. Journal of Machine Learning Research, 2002, 2(2):97-123.

[11]. Mehmood T, Martens H, Sæbø S, et al. A Partial Least Squares based algorithm for parsimonious variable selection[J]. Algorithms for Molecular Biology Amb, 2011, 6(1):: 27.

[12]. Kong X, Zhu W, Zhao Z, et al. Fluorescence spectroscopic determination of triglyceride in human serum with window genetic algorithm partial least squares[J]. Journal of Chemometrics, 2012, 26(1-2):25–33.

[13]. Guo Z, Liu G, Li D, et al. Self-adaptive differential evolution with global neighborhood search[J]. Soft Computing, 2016:1-10.

[14]. Zhang G J, Li-Na L I, Qing-Bo L I, et al. APPLICATION OF DENOISING AND BACKGROUND ELIMINATION BASED ON WAVELET TRANSFORM TO BLOOD GLUCOSE NONINVASIVE MEASUREMENT OF NEAR INFRARED SPECTROSCOPY[J]. Journal of Infrared & Millimeter Waves, 2009, 28(2):107-110.

[15]. Dan T N. Spectral Wavelength Selection Based on PLS Projection Analysis[J]. Spectroscopy & Spectral Analysis, 2009, 29(2):351-354(4).

[16]. Jiang J H, Berry R J, Siesler H W, et al. Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data.[J]. Analytical Chemistry, 2002, 74(14):3555-65.

[17]. And V C, Massart D, Noord O E D, et al. Elimination of Uninformative Variables for Multivariate Calibration[J]. Analytical Chemistry, 1996, 68(21):3851-8.

[18]. Santiago K S, Soares A S, Lima T W D, et al. Genetic algorithm for variable and samples selection

in multivariate calibration problems[J]. Journal of Computer Science, 2015.

[19]. Leardi R, Seasholtz M B, Pell R J. Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data[J]. Analytica Chimica Acta, 2002, 461(2):189-200.