Detection of chemical oxygen demand in water based on UV absorption spectroscopy and PSO-LSSVM algorithm*

ZHOU Kunpeng[1], LIU Zhiyang[2]**, CONG Menglong[1], and MAN Shanxin[3]**

1. College of Engineering, Intelligent Manufacturing Technology Key Laboratory, Inner Mongolia Minzu University, Tongliao 028000, China

2. Jingzhou University, Jingzhou 434000, China

3. Alxa League Meteorological Bureau, Alxa 750300, China

(Received 10 September 2021; Revised 1 November 2021)

A method of detecting chemical oxygen demand (COD) of water based on ultraviolet (UV) absorption spectra is proposed. The modeling and analysis of the standard samples and the actual water samples are carried out respectively. For the standard solution samples, the univariate linear models based on single wavelengths and the partial least square (PLS) model based on synergy interval partial least square (SiPLS) and moving window partial least square (MWPLS) are established. For the actual water samples, different pre-processing methods are used. SiPLS and MWPLS are used to select the characteristic bands. The least squares support vector machine algorithm optimized by particle swarm optimization (PSO-LSSVM) algorithm is used to establish the prediction model, and the prediction results of various models are compared. The results show that the optimal model is PSO-LSSVM which uses SiPLS to select the characteristic bands of the first derivative spectra (preprocessing method). The determination coefficient of the prediction set is 0.963 1, and the root mean square error of prediction (*RMSEP*) is 2.225 4 mg/L. PSO-LSSVM algorithm has good prediction performance for the analysis of COD in actual water samples by UV spectra. This paper provides a new design idea for the research and development of water quality detection optical sensor.

Document code: A **Article ID:** 1673-1905(2022)04-0251-6

DOI https://doi.org/10.1007/s11801-022-1143-5

The quality of water resources has been declining and the water environment has been deteriorating, which has seriously threatened the sustainable development of society and the survival of human beings. In recent years, spectral analysis method has been widely used in food[1,2], medical[3] and health[4], biological[5], environmental safety monitoring[6] and other fields. Chemical oxygen demand (COD) is an important comprehensive index in the field of water quality detection, which reflects the degree of water pollution by reducing substances[7]. The standard methods of COD determination usually adopt wet chemical methods, such as dichromate method and potassium permanganate method. These methods have the characteristics of relatively reliable determination results and good reproducibility, but there are still many shortcomings, such as long heating reflux time, high toxicity of reagents and easy to cause secondary pollution, high operation cost and relatively complex operation. Spectral method has the characteristics of fast detection, green pollution-free and good reproducibility. This method has become the most effective one in addition to the national standard method. It is suitable for rapid on-line monitoring of COD in environmental water samples. Therefore, it has become one of the important development directions of water quality detection technology.

At present, ultraviolet-visible (UV-Vis) absorption spectrum is mainly used to detect COD in water[8], while near-infrared (NIR) absorption spectrum[9] and Raman spectrum[10] have also been reported to detect COD in water, but it is still in the laboratory research stage. Fluorescence spectrometry is mainly used to detect dissolved organic matter (DOM) in water[11]. Some researchers also use different types of spectral fusion methods to detect organic pollutants in water[12], but the detection results show that the effect of different types of multispectral fusion methods to detect water quality parameters needs to be further improved.

In order to ensure the integrity of the experimental process and improve the universality of the optimal prediction model, in this paper, the UV absorption spectrum is used to detect the COD of the experimental water

* This work has been supported by the National Natural Science Foundation of China (No.61963031), the Inner Mongolia Autonomous Region Natural Science Foundation (No.2019MS06017), and the Scientific Research Projects of Colleges and Universities in Inner Mongolia Autonomous Region of China (No.NJZY20122).

** E-mails: kunpeng032@126.com; zhiyangliu@dingtalk.com

samples, which are divided into standard samples and actual water samples. The synergy interval partial least square (SiPLS) and the moving window partial least square (MWPLS) algorithms are used as the feature extraction method and the least squares support vector machine algorithm optimized by particle swarm optimization (PSO-LSSVM) regression models are established. The present paper provides a new research idea and solution for the rapid detection of COD of water quality.

A total of 52 samples of standard solution are used in the experiment, which are diluted with 1 000 mg/L mother solution of potassium hydrogen phthalate by heavy distilled water in proportion, and the concentration range is 1—100 mg/L. The other 53 are actual water samples, which are collected from coastal seawater and surface water in Qinhuangdao City of Hebei Province. After standing for 30 min, the upper liquid is collected for UV absorption spectra and COD measurement. The basic potassium permanganate titration method is used to measure the COD of seawater, the rapid digestion spectrophotometry method is applied to detect COD of surface water samples, the DRB200 thermostat and DR6000 spectrophotometer (HACH, American) are used to complete the digestion surface water samples and COD detection. The temperature of the thermostat is set at 150 °C and kept heating for 120 min. The UV spectrum acquisition system of the laboratory is shown in Fig.1. Avaspec-2048 UV-Vis fiber spectrometer (Avantes, Holland) is used as the UV spectrometer. The light is deuterium and halogen combined fiber light source. A quartz cuvette (10 mm×10 mm) is used as the sample cell. In the process of collecting spectrum, the integration time is set as 2 ms, and the average number of sampling is set as 200.

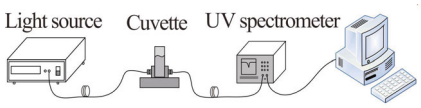


Fig.1 UV spectrum acquisition system

The UV absorption spectra of the sample solutions are shown in Fig.2. The data used in the modeling are all original without pretreatment.

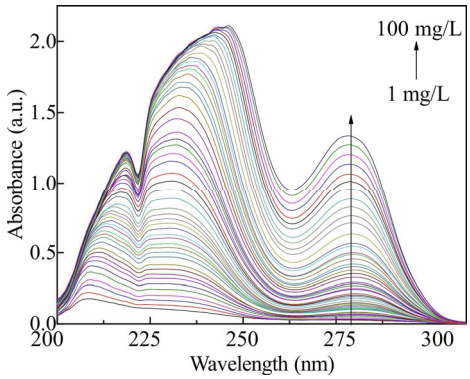


Fig.2 UV absorption spectra of standard solutions with different concentrations

It can be seen from Fig.2 that there are obvious absorption peaks in the UV absorption spectra of the sample solutions in the wavelength range of 250—300 nm, and there is no spectral saturation phenomenon. For the actual water samples, the range of UV spectra is 200—400 nm. The UV spectral curves of actual water samples are shown in Fig.3. The 37 actual samples are randomly selected as the calibration set to establish the calibration model, the remaining 16 samples (seawater and surface water account for 8 samples respectively) are used as the prediction set, which can test the prediction performance of the model.

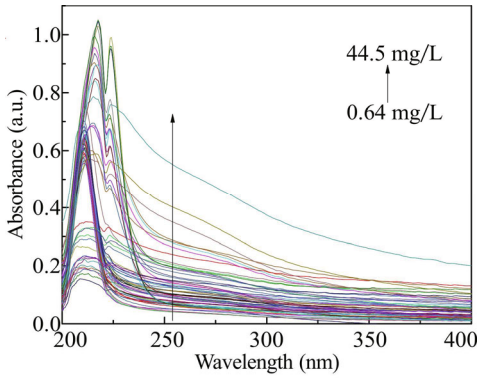
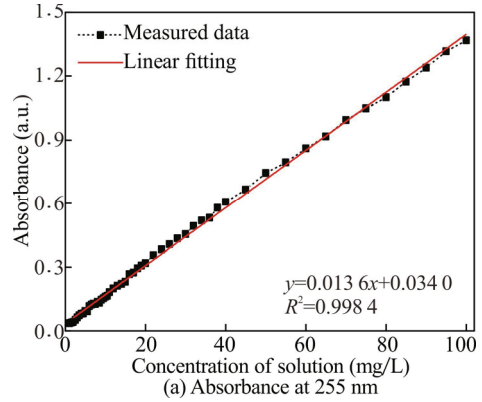


Fig.3 UV absorption spectra of actual water samples with different concentrations

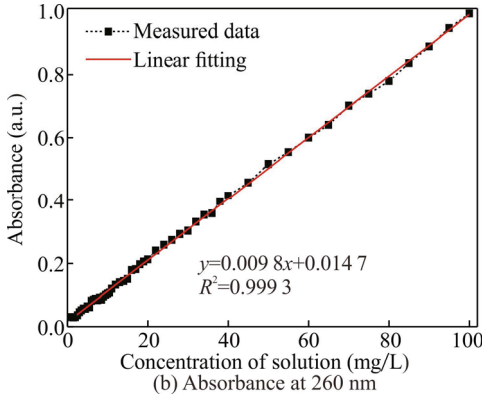
Due to the complex composition of the actual water samples, their UV absorption spectral curves are likely to be affected by the water environment, such as base-line drift, additive noise and so on. For this reason, it is necessary to pre-process the spectral data. It is found that after pre-processing, the spectral data retain the absorption characteristics of the original spectra. The spectral pre-processing methods used in this paper include Savitzky-Golay smoothing (SG smoothing), eliminating constant offset (ConOffEli), first derivative (1st-Der), second derivative (2nd-Der), multiple scattering correction (MSC), standard normal variate (SNV), min-max normalization (MinMaxNor) and vector normalization (VN). The PSO-LSSVM algorithm has the advantages of both particle swarm optimization (PSO) and least squares support vector machine (LSSVM) algorithm. The modeling method used in this paper gives full play to the advantages of feature extraction algorithm, PSO algorithm and LSSVM algorithm, and the prediction effect of the model is significantly improved.

The single characteristic wavelength and multi-characteristic wavelength/characteristic spectral interval methods are used to model the UV absorption spectrum of standard solution. Nine wavelength points are selected from Fig.2 as characteristic wavelengths (the spectral range is 255—295 nm, with a step size of 5 nm) to establish the UV single wavelength model. The univariate regression models are shown in Fig.4. These models are based on the UV absorbance data at the solution concentration and characteristic wavelength. It can

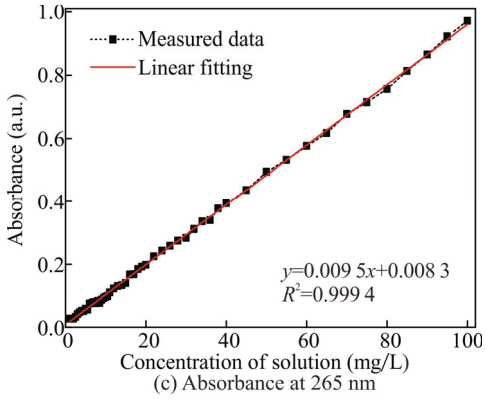
be seen from Fig.4 that there is a good linear positive correlation between the concentration of the standard solution and the UV absorbance corresponding to each single wavelength. The results of the two models are the optimal when the wavelength is at 270 nm and 275 nm. The determination coefficients (R^2) are 0.999 6.



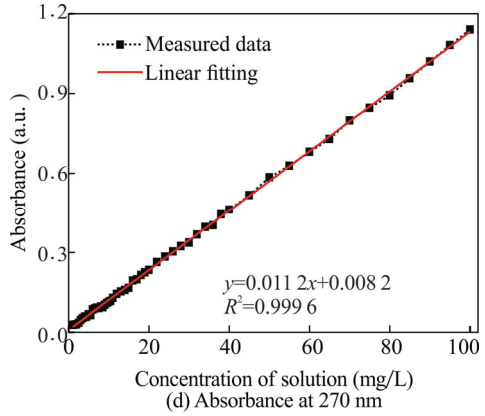
(a) Absorbance at 255 nm



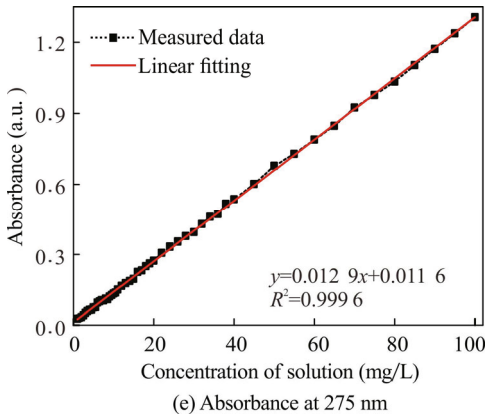
(b) Absorbance at 260 nm



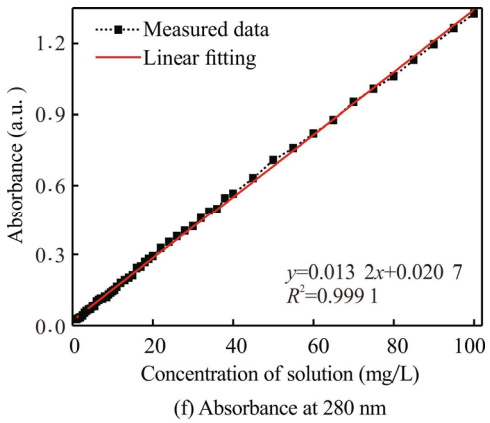
(c) Absorbance at 265 nm



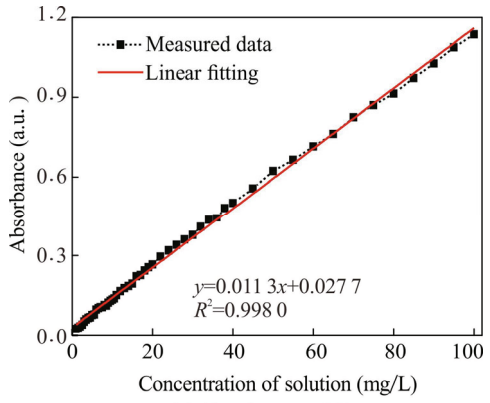
(d) Absorbance at 270 nm



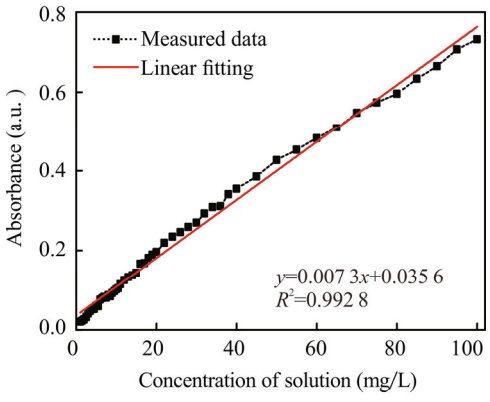
(e) Absorbance at 275 nm



(f) Absorbance at 280 nm



(g) Absorbance at 285 nm



(h) Absorbance at 290 nm

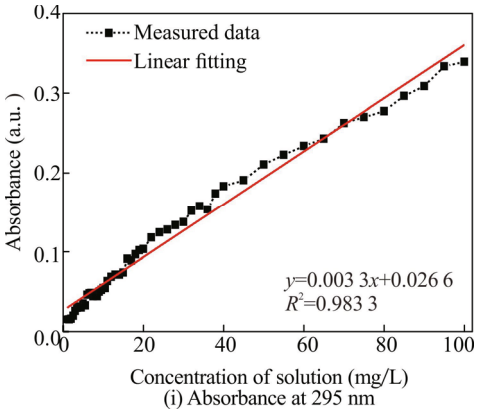


Fig.4 Single wavelength modeling results of UV absorption spectral data

For the standard solution, each UV single wavelength data model can reflect the relationship between the concentration and absorbance of the solution better, while the effect of 290 nm and 295 nm models is slightly worse, knowing the actual cause is that the absorbance at the two wavelengths is very low, the variation of UV absorbance caused by the variation of water sample concentration is very small, and the discrimination is very low, which leads to the increase of the detection error of the spectrometer, and it makes the linear correlation between the absorbance and the solution concentration worse. Therefore, the single wavelength method only uses 254 nm absorbance data to detect COD, which cannot fully reflect the UV absorption characteristics of water, and has the disadvantage of low detection accuracy. Multi-wavelength or wide spectrum method will improve the detection accuracy of the model.

At the same time, Fig.4 also reflects that when 270 nm and 275 nm are selected as the characteristic wavelengths, the single wavelength model is the best, but the two wavelengths are not included in the spectral ranges selected by the feature extraction algorithms. The reason is that models established by single wavelengths are only linear regression models based on the correlation between concentration and absorbance.

In the whole band range, the effective algorithm is used to select the characteristic spectra, and then the characteristic spectra are used to establish the calibration models. Compared with the single wavelength models, this model has better effect and stronger stability. The reason is that the redundant information is reduced after feature extraction, which makes the amount of calculation lessened and the efficiency of operation improved.

In this paper, SiPLS and MWPLS methods are used to extract the feature spectra in the full band (200—310 nm), and the regression models are established based on PLS algorithm.

SiPLS splits the data set into a number of intervals (variable-wise) and calculates all possible PLS model combinations of two, three or four intervals. Two sub-interval models are selected for combinatorial calculation in this paper in order to ensure the balance between calculation time and model effect.

MWPLS calculates SiPLS models based on a moving window concept. For each variable, a PLS model is calculated with the given window size.

In this paper, the number of latent variables of PLS algorithm is set to 5, the sub-intervals number is 15 when the SiPLS works, and the final synergy interval is [8, 14] after optimization, corresponding to 251.464—258.586 nm and 295.900—302.993 nm respectively. The size of moving window set by MWPLS is 30, and the final optimal window number is 88 after optimization, corresponding to 251.464—268.667 nm.

The following performance indicators are commonly used to evaluate the prediction performance of the established model, including determination coefficient (R^2), root mean squared error of cross-validation (*RMSECV*) and root mean squared error of prediction (*RMSEP*). The equations of these performance indexes are shown as follows

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \overline{y})^2}, \tag{1}$$

$$RMSECV = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \tag{2}$$

$$RMSEP = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}, \tag{3}$$

where $\hat{y}_i$ is the predictive value of the calibration model, $\overline{y}$ is the mean of the actual value, y_i is the measured (actual) value based on the reference method, n is the number of calibration samples and m is the number of validation samples. The minimum *RMSECV* is used to select the optimal number of principal components.

The optimal wavebands and the results of PLS modeling are shown in Tab.1, where *Bias* represents the systematic deviation between a measured/predicted value and the true value. The smaller the value of *Bias*, the better the prediction effect of the model. For this paper, it is the overall average deviation between the measured value and the true value of the test sample.

It can be seen from Tab.1 that the prediction performance of the models can be significantly improved by extracting features by SiPLS and MWPLS, compared with the PLS model without feature extraction. The PLS modeling results based on different feature extraction algorithms are shown in Fig.5.

Tab.1 Evaluation indexes of PLS under different feature extraction methods

Feature extraction method	Selected band (nm)	Modeling results		
		R^2	*RMSECV* (mg/L)	*Bias* (mg/L)
Without extraction	200—310	0.993 30	27.629 0	−0.018 2
SiPLS	251.464—258.586 295.900—302.993	0.999 95	0.220 1	0.003 1
MWPLS	251.464—268.667	0.999 95	0.220 8	0.003 5

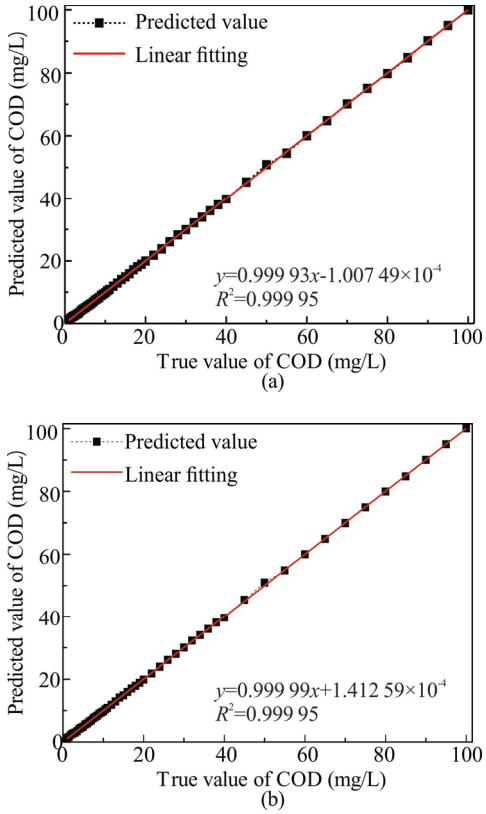


Fig.5 Modeling results of UV absorption spectra at 200—310 nm by (a) SiPLS and (b) MWPLS

In the process of PLS modeling by characteristic spectral intervals, the corresponding verification standard should be implemented for the selection of the results. Generally, the *RMSECV* is selected as the evaluation index, which has higher accuracy and stability than the model built by manually selected feature wavelength. Therefore, for the UV spectral data of the standard solution, the effect of the model based on the characteristic spectral intervals is better than that of the single wavelength models. The results suggest that the regression models based on the UV characteristic spectral intervals selected by the feature extraction algorithm can accurately predict the concentration of the standard solution.

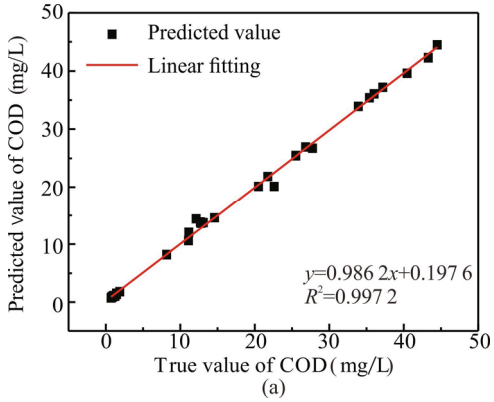
Since the standard solution contains only one substance (potassium hydrogen phthalate), and the preparation process of the solution is rigorous, the linearity of the corresponding relationship between the absorbance value and the concentration value of the obtained spectral line is good, and the effect of the model obtained with PLS algorithm is excellent. The composition of the actual water sample is very complex, and there are many interference factors in the process of spectral measurement. The effect will be worse if the linear model is used to evaluate the actual water sample. Therefore, in the subsequent analysis of actual water samples, two different feature extraction methods (SiPLS and MWPLS) are used for feature extraction and PSO-LSSVM algorithm is introduced for modeling because it has better performance with the modeling effect.

In the modeling of PSO-LSSVM for actual water samples, the optimal spectral characteristic intervals extracted by SiPLS and MWPLS are used as the input of LSSVM respectively. During the experiment, it is found that when SiPLS is used to extract the feature spectra, the optimal pre-processing method is 1st-Der, while MWPLS is used as the feature extraction method, and the optimal pre-processing method is MSC. The number of SiPLS subintervals is set to 18, the size of MWPLS moving window is set to 35, the c_1 and c_2 of PSO algorithm are set to 1.5 and 1.7, respectively, the population size is set to 30, and the evolution times is set to 300. The modeling results of PSO-LSSVM algorithm are shown in Tab.2, where, R_c^2 represents the coefficient of determination of calibration set and R_p^2 represents the determination coefficient of prediction set. The prediction results of calibration set and prediction set of the two models are shown in Fig.6 and Fig.7, respectively.

Tab.2 Modeling of UV spectra of actual water samples by PSO-LSSVM

Feature extraction method	Pre-processing method	Selected spectral interval (nm)	Calibration set		Prediction set	
			R_c^2	*RMSECV* (mg/L)	R_p^2	*RMSEP* (mg/L)
SiPLS	1st-Der	211.025—221.743 222.339—233.046	0.997 2	0.745 5	0.963 1	2.225 4
MWPLS	MSC	227.694—247.902	0.991 0	1.308 6	0.938 5	2.861 4

It can be seen from Tab.2 that the optimal prediction model is PSO-LSSVM which uses SiPLS to extract the characteristic spectral intervals of 1st-Der spectral data. For the optimal model, the coefficient R_c^2 is 0.997 2 and the *RMSECV* is 0.745 5 mg/L. The determination coefficient R_p^2 is 0.963 1, and the *RMSEP* of external test is 2.225 4 mg/L. Thereby, this method is recommended for practical application.



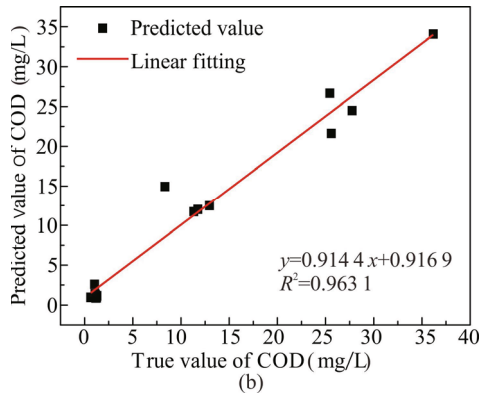
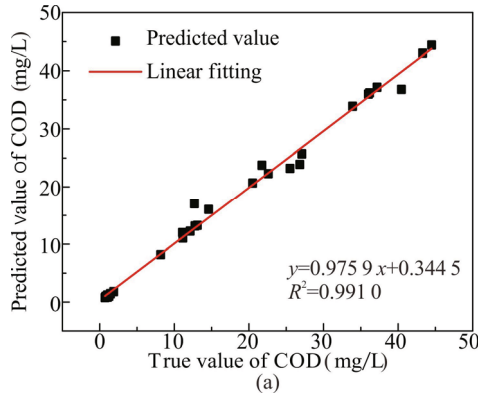


Fig.6 PSO-LSSVM modeling results of (a) calibration set and (b) prediction set based on SiPLS



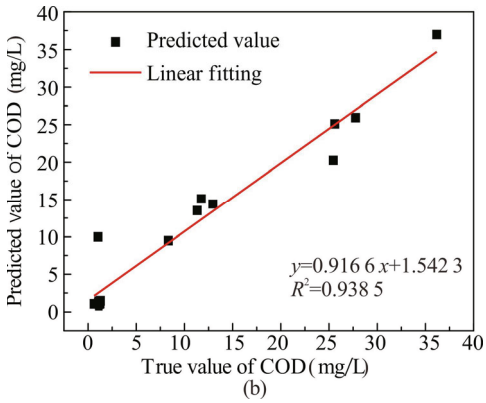


Fig.7 PSO-LSSVM modeling results of (a) calibration set and (b) prediction set based on MWPLS

Based on different feature extraction algorithms and modeling methods, the UV absorption spectral data of standard solution and actual water sample were modeled, and the prediction effects of various models were compared. The results show that the modeling method used in this paper is feasible, the prediction effect of the optimal model obtained from the experiment is excellent, and it can be used as an effective method for rapid detection of COD in water quality. This study was completed in the laboratory environment, but the substances in the natural water and the environment of the water are more complex, and the interference of the UV spectra needs to be further studied.

Statements and Declarations

The authors declare that there are no conflicts of interest related to this article.

References

[1] MENDES E, DUARTE N. Mid infrared spectroscopy as a valuable tool to tackle food analysis a literature review on coffee dairies honey olive oil and wine[J]. Foods, 2021, 10(2)：477.

[2] JAN U P, DIETMAR R K, REINHOLD C. On-line application of near infrared (NIR) spectroscopy in food production[J]. Trends in food science and technology, 2015, 46(2)：211-230.

[3] KUMAR A, JAIN S K. Development and validation of UV-spectroscopy based stability indicating method for the determination of fluoxetine hydrochloride[J]. Analytical chemistry letters, 2016, 6(6)：894-902.

[4] YUAN J Z, LU Q P, WU C Y, et al. Noninvasive human triglyceride detecting with near-infrared spectroscopy[J]. Spectroscopy and spectral analysis, 2018, 38(1)：42-48.

[5] MASSIE C, KNAPP E, CHEN K, et al. Improved prediction of femoral fracture toughness in mice by combining standard medical imaging with Raman spectroscopy[J]. Journal of biomechanics, 2021, 116(2)：110243.

[6] FERREIRO G M, AYUSO J, ÁLVAREZ J A, et al. Gasoline analysis by headspace mass spectrometry and near infrared spectroscopy[J]. Fuel, 2015, 153：402-407.

[7] ZHOU K P, BAI X F, BI W H. Detection of chemical oxygen demand (COD) of water quality based on fluorescence multi-spectral fusion[J]. Spectroscopy and spectral analysis, 2019, 39(3)：813-817.

[8] ZHOU K P, BI W H, ZHANG Q H, et al. Influence of temperature and turbidity on water COD detection by UV absorption spectroscopy[J]. Optoelectronics letters, 2016, 12(6)：461-464.

[9] DAHLBACKA J, NYSTROM J, MOSSING T, et al. On-line measurement of the chemical oxygen demand in waste water in a pulp and paper mill using near infrared spectroscopy[J]. Open journal of statistics, 2014, 2(4)：19-25.

[10] WU G Q, ZHAO W. Seawater chemical oxygen demand optical detection method based on Raman spectroscopy[J]. Journal of applied optics, 2019, 40(2)：278-283.

[11] HARRINGMEYER J P, KAISER K, THOMPSON D R, et al. Detection and sourcing of CDOM in urban coastal waters with UV-visible imaging spectroscopy[J]. Frontiers in environmental science, 2021, 9：647966.

[12] WU G Q, BI W H. Research on chemical oxygen demand optical detection method based on the combination of multi-source spectral characteristics[J]. Spectroscopy and spectral analysis, 2014, 34(11)：3071-3074.