Research on the prediction model of UV spectral water quality parameters based on INFO-LSSVM

Chao Li*, Wen Li, Xueke Luo, Zhuofan Yu, Yakai Han, Facheng Yu

School of Mechanical and Materials Engineering, North China University of Technology, Beijing China

Abstract: In order to meet the requirements of precision and rapid detection of water quality parameters, UV-visible absorption spectroscopy was used for the measurement and INFO-LSSVM was proposed for predictive modelling. The common nitrate nitrogen (NO_3-N) and nitrite nitrogen (NO_2-N) in water quality testing as the solution to be measured, the UV-visible absorption spectral data filtering, spectral data integration, the establishment of INFO-LSSVM nonlinear prediction model; comparison of GA-LSSVM, PSO-LSSVM and LSSVM algorithm models, the results show that the INFO-LSSVM prediction model is effective, and provides a good solution for water quality testing. LSSVM prediction model is effective and provides new research value for water quality detection.

1. Introduction

In recent years, with the rapid development of industry and economy in China, the problem of water quality pollution has become increasingly serious, and a large amount of nitrogen-containing wastewater is directly discharged into the water. Nitrogen-containing substances in the water body exceed the standard, the water body eutrophication[1], nitrogen-containing organic matter is finally converted into nitrate nitrogen (Nitrate nitrogen, NO_3-N), which is a key parameter for assessing the health of the water body, the degree of eutrophication[2], nitrogen-containing organic matter is converted into nitrate oxidation process, it will also produce nitrite nitrogen (Nitrite nitrogen, NO_2-N), when nitrite nitrogen content exceeds the standard, this indicates that there is still a water body. NO_2-N), when the nitrite nitrogen content exceeds the standard, this indicates that there is still incomplete oxidative decomposition of organic nitrogen substances in the water, while the pollution of nitrogen-containing organic matter is still going[3]. In order to achieve the governance and control of the water environment, nitrate nitrogen and nitrite nitrogen are important water quality testing parameters in water quality testing. The current methods of water quality testing are broadly divided into two, one is the traditional chemical analysis, ultraviolet spectrophotometry, ion-selective electrode method, ion chromatography, gas-phase molecular absorption spectrometry, etc.[4], the above methods require high experimental conditions, expensive experimental instruments, consumption of chemical reagents, and possible pollution of the newly generated substances.

At present, scholars at home and abroad conduct research on the modelling of spectral data prediction models, and the focus of the research is to improve the accuracy of prediction. Among them, SVM is a small-sample learning method[5] and has been widely used to solve the wastewater prediction problem, but the computational process is cumbersome and it is difficult to achieve large-scale training samples, in order to overcome these shortcomings, Deng W[6] used Least Squares Support Vector Machines, LSSVM, which improves the performance of the SVM algorithm by solving the linear programming instead of the quadratic programming. not quadratic programming to improve the performance of SVM algorithm and reduce the computational process. Chen Ying [7] regression prediction of nitrate concentration in water bodies based on LSSVM algorithm and constructed a mixture model to improve the prediction accuracy, but there are some unknown parameters in the kernel function of LSSVM that need to be selected in advance. To address this problem, population intelligence optimization algorithms have been widely studied, such as particle swarm optimization (PSO), genetic algorithm Genetic Algorithm (GA), etc., to find the optimal solution through the population intelligence of population optimization algorithms to perform a collaborative search mechanism to find the optimum for the penalty coefficients and kernel function parameters of the LSSVM. Zhou[8] based on PSO-LSSVM predictive modelling of chemical oxygen demand (COD) in actual water samples detected by UV spectroscopy obtained good prediction results. Wang Lingbin[9] optimised LSSVM for cyanobacterial bloom prediction by Genetic Algorithm (GA), the prediction error of the model decreased significantly, with higher accuracy and stability, both PSO and GA optimisation algorithms, etc. are based on the optimisation process in the population-based algorithm starting from a set of solutions and updating their positions during optimisation, and information sharing. sharing of information which helps them to search

* Corresponding author: 793280235@qq.com

better in difficult search spaces. However, the above algorithms require a lot of function evaluation during optimisation and are computationally expensive. Iman Ahmadi-Anfa et al [10] proposed The weighted mean of vectors algorithm (INFO). The algorithm calculates the weighted average of a set of vectors in the search space and converges quickly to find the optimal solution to improve the detection speed. The combination of vectors improves the algorithm's ability to explore and mine the search, increases the acquisition of valid information in spectral data, and reduces the processing of redundant information. While using global search to obtain the global optimal solution, local search is also set to avoid falling into the local optimum, which makes the prediction results more accurate and effectively improves the accuracy of detection.

2. Experimental component

2.1. Instruments and reagents

The experimental setup is shown in Figure 1. The instruments used in this experiment were: NSP01H industrial mini fibre-optic spectrometer (Qualicom), with a wavelength range of 190-380 nm, Czerny-Turner optical structure, grating spectroscopy and high-performance photodetectors; XYM 2020 pulsed xenon lamp (Qualitong), with a wavelength of 185-2500 nm and a pulse frequency of 72Hz, input current and voltage DC12/1.5V/A; high-throughput quartz cuvette (Ocean Optics), beakers, measuring cylinders, single-channel adjustable-range pipettes of Research plus from 0.1μL to 2μL and 0.5ml to 5ml.

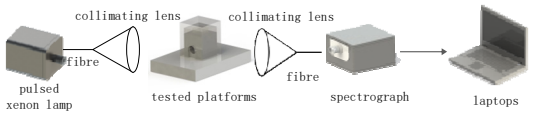

Fig. 1. Schematic diagram of experimental platform

Reagents used in the experiment: deionised water of grade EW-I in the National Standard for Deionised Water GB11446-1-2013; NO_3-N standard solution provided by the Beijing Research Institute of Chemical Industry as well as NO_2-N at a concentration of c(NO_3-N)=10mg·$L^{-1}$and nitrite nitrogen standard solution at a concentration of c(NO_2-N)= 2mg·L^{-1}

2.2. Detection principle

The UV absorption spectroscopy method used in this experiment is based on the Beer-Lambert Law (1),which reads the absorbance and the absorption range of the molecules of a substance in a liquid to obtain the concentration value of the liquid to be measured.

$$A = \lg\left(\frac{I_0}{I}\right) = KlC = -\lg\left(\frac{Sample-Dark}{Reference-Dark}\right) \times 100\% \quad (1)$$

Where A , I_0 , I are the absorbance, incident light intensity, outgoing light intensity, respectively;K,l,C are the absorbance coefficient, absorption light range, the concentration of the liquid to be measured, respectively; the right-hand side of the equation Sample is the photoelectric signal detected by turning on the spectral scanning of the light source for different concentrations of reagents.

2.3. Sample preparation and data processing

The nitrate concentration of general water body is 0.2~5mg·L^{-1}, by taking out 0.4, 0.6, 0.8, 1.2, 1.6, 2.0, 4.0, 6.0, 8.0, 10.0 ml of the nitrate nitrogen standard solution (10mg·L^{-1}) into a 20 ml measuring cylinder, diluting it by adding ultrapure water, and obtaining 0.2, 0.4, 0.6, 0.8, 1.0 after dilution, 2.0, 3.0, 4.0, 5.0 mg·L^{-1} of nitrate nitrogen standard solution;

Set the range of nitrite nitrogen0-2.0mg·L^{-1}, respectively, take out the nitrite nitrogen standard solution 2.0, 4.0, 6.0, 8.0, 10.0 ml to add 20 ml cylinder for dilution, add ultrapure water to dilute to the standard line, dilution to get the 5 groups of nitrite nitrogen standard solution of 0.2, 0.4, 0.6, 0.8, 1.0mg·L^{-1};

The solutions were transferred into a quartz cuvette for spectral scanning, respectively, with deionised water as the reference, the spectra were collected in the range of 190~400nm in steps of 0.6nm, and the measurements were repeated six times for each set of solutions, and the absorbance data were averaged. After filtering pre-processing as shown in Fig. 2, Fig.2.(a) shows the absorption spectrum of nitrate nitrogen, while Fig.2.(a) shows the absorption spectrum of nitrite nitrogen, the main absorption spectral region of NO_3-N solution is in the range of 200~240nm, and the absorption peak appears near 205m, and the absorption is almost 0 after 250nm; the absorption peak of NO_2-N solution appears near 210nm, and the absorption is almost 0 after 250nm, and the absorption is almost 0 after 250nm, in order to reduce the impact of the fluctuation of the signal of the spectrometer, the filtered In order to reduce the influence of the signal fluctuation of the spectrometer, the spectra after filtering were integrated, and the spectral integral value was taken as the spectral feature.

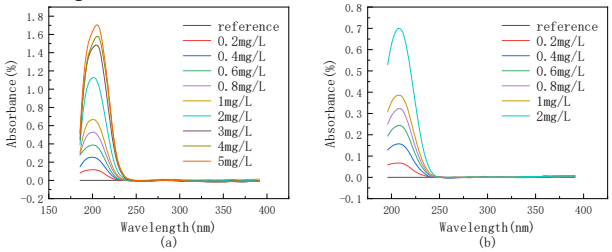

Fig. 2. Pre-treatment wavelength-absorbance curve

3. Modelling study methodology

3.1. Partial Least Squares Support Vector Machine

Least Square support vector machine[11], LSSVM is to extend the SVM inequality constraints into one equation constraints as shown in equation (2), which simplifies the computational process and improves the speed of

operation.

$$\begin{cases} min_{w,b,e} J(w,e) = \frac{1}{2} w^T w + 0.5\gamma \sum_{k=1}^{n} e_k^2 \\ y[w^T(x_k) + b] = 1 - e_k, e_k \geq 0, k = 1,2 \ldots n \end{cases} \quad (2)$$

The corresponding Lagrange function is:

$$L(w,b,\alpha,e) = J(w,e) - \sum_{k=1}^{n} \alpha_k [w^T \varphi(x_k) + b + e_k - y_k] \quad (3)$$

The system of equations (4) is obtained by taking the partial derivatives of w, b, α, e in equation (3) respectively and making them equal to zero.

$$\begin{cases} w = \sum_{k=1}^{n} \alpha_k \varphi(x_k) \\ \sum_{k=1}^{n} \alpha_k = 0 \\ \alpha_k = \gamma e_k \\ w^T \varphi(x_k) + b + e_k - y_k = 0 \end{cases} \quad (4)$$

Eliminating w and e_k gives:

$$\begin{pmatrix} 0 & e_n^T \\ e_n & \Omega + \gamma^{-1} I \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix} \quad (5)$$

where Ω and I are the kernel and constant matrices. Solving Eq. (5) yields the following predictive model function for LSSVM:

$$y(x) = \sum_{k=1}^{n} \alpha_k K(x, x_k) + b \quad (6)$$

Where $K(x, x_k)$ is the kernel function, the radial basis function (RBF) is selected as the kernel function, and its main parameter σ indicates the radial basis function width. Since the selection of parameters γ and σ determines the accuracy of LSSVM regression and prediction.

3.2. particle swarm optimization

Particle swarm optimization, PSO [12] is a stochastic optimization algorithm inspired by bird flock predation, the main idea of this algorithm is to find the individual optimal solution (p_{best}) and the population optimal solution (g_{best}) in the population through cooperation and information sharing between each individual, where each solution is represented by a random particle, the principle is as follows:

$$\begin{cases} x_{i,j}^{t+1} = x_{i,j}^{t} + v_{i,j}^{t+1} \\ v_{i,j}^{t+1} = v_{i,j}^{t} + c_1 r_1 (p_{best} - x_{i,j}^{t}) + c_2 r_2 (g_{best} - x_{i,j}^{t}) \end{cases} \quad (7)$$

Where x_i is the current position of the particle; v_i is the particle velocity; c_1 and c_2 are the acceleration factors; r is a random number between (0,1).

Where is the current position of the particle; is the particle velocity; and is the acceleration factor; is a random number between (0,1); is the inertia factor, which is used to balance the ability of global optimisation and local optimisation, and iterates continuously to determine the optimal solution of the particle.

3.3. Genetic Algorithm

Genetic Algorithm[13] is a stochastic search algorithm inspired by the evolutionary studies of living organisms, which seeks for the optimal solution by simulating the evolution of living organisms to carry out an iterative search. The computational framework of the Genetic Algorithm is divided into the following steps: coding, evaluation function, genetic operation, initialisation of the population, selection of the control parameters and the termination conditions. The genetic algorithm is the core step, which uses selection, crossover and mutation (genetic operators) to update and select the population, and repeats the process until a satisfactory solution is obtained.

3.4. The weighted mean of vectors algorithm

The weighted mean of vectors algorithm[14] is implemented in four phases: initialisation, rule updating, vector merging and local search. In the initialisation phase, as in other optimisation algorithms, the population individuals are randomly initialised in the search space. And k-fold cross validation is added in this phase to improve the accuracy of the model. In the Update Rule phase, INFO updates the current positions of the vectors by means of the established Mean Rule, which is extracted from the weighted average of a set of random vectors. Where (CA) is the convergence acceleration part of the algorithm, in which the global search capability is improved.

$$z1_l^g = \begin{cases} x_l^g + \sigma \times Meanrule + CA, rand < 0.5 \\ x_a^g + \sigma \times Meanrule + CA, rand \geq 0.5 \end{cases} \quad (8)$$

$$z2_l^g = \begin{cases} x_{bs}^g + \sigma \times Meanrule + CA, rand < 0.5 \\ x_{bt}^g + \sigma \times Meanrule + CA, rand \geq 0.5 \end{cases} \quad (9)$$

$$MeanRule = r \times WM1_l^g + (1-r) WM2_l^g \quad (10)$$

$$CA = randn \times \frac{(x_{bs} - x_{a1})}{[f(x_{bs}) - f(x_{a1}) + \varepsilon]} \quad (11)$$

Vector combination stage According to the following equation (12), INFO combines the two vectors ($z1_l^g$, $z2_l^g$) computed in the previous stage with the vector rand < 0.5 of the condition u_l^g, in order to generate a new vector.

$$u_l^g = \begin{cases} z1_l^g + \mu_l^g |z1_l^g - z2_l^g| \; rand1 < 0.5, rand2 < 0.5 \\ z2_l^g + \mu_l^g |z1_l^g - z2_l^g| \; rand1 < 0.5, rand2 \geq 0.5 \\ x_l^g \qquad\qquad rand \geq 0.5 \end{cases} \quad (12)$$

The local search phase INFO uses a local search phase to prevent falling into a local optimum. It improves the utilisation and convergence.

$$au_l^g = \begin{cases} x_{bs} + randn \times [Meanrule + V_a], \\ rand1 < 0.5, rand2 < 0.5 \\ x_{md} + randn \times [Meanrule + V_b], \\ rand1 < 0.5, rand2 \geq 0.5 \end{cases} \quad (13)$$

Figure 3 shows the flowchart of the algorithm.

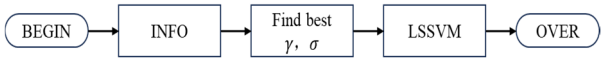

Fig 3. INFO algorithm flowchart

4. Results and discussion

4.1. Model validation

The determination coefficients (R^2) and root mean squared error (RMSE) were taken to measure the assessment of model accuracy. The closer the R^2 is to 1, the better the

model fits, and the smaller the RMSE is, the better the predictive ability of the model.

4.1.1 Comparison of NO3-N model

The 210-240 nm absorbance data of 54 groups of NO_3-N solutions were selected for predictive modelling, and 45 groups were randomly selected as training samples and 9 groups as predictive samples, and compared with the true values, and the comparison between the predicted and true values for each modelling of nitrate nitrogen was obtained as shown in Figure 4.

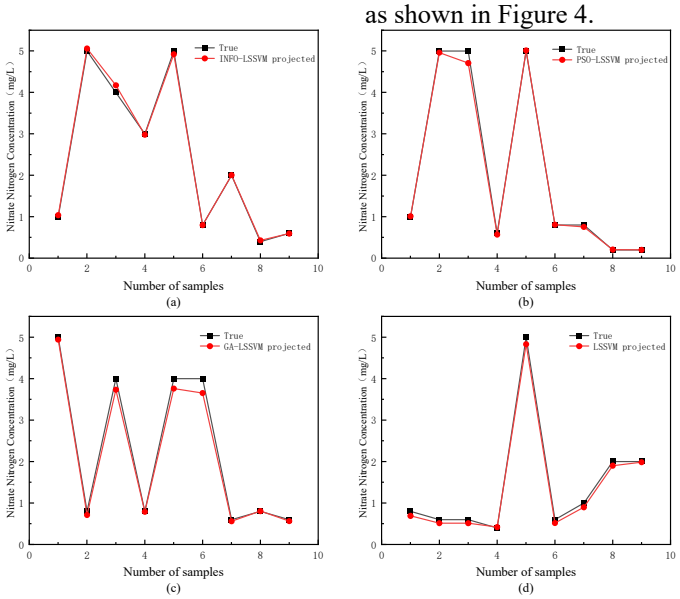

Fig. 4 Comparison of predicted and true values of the four models for NO_3-N

A comparison of the prediction performance of the three models for NO_3-N is shown in Table 1. As can be seen from the table, in the prediction of nitrate nitrogen, the root-mean-square error of the INFO-LSSVM model is smaller than that of the PSO-LSSVM, GA-LSSVM and LSSVM models; the coefficients of determination of the INFO-LSSVM model are larger than that of the PSO-LSSVM, GA-LSSVM and LSSVM; and a comparison of the indexes shows that the INFO-LSSVM has a good prediction effect relative to all three prediction models for modelling nitrate nitrogen. LSSVM has a good predictive effect on modelling the data of nitrate nitrogen relative to all three predictive models.

Table 1. Comparison of NO_3-N modelling results

NO_3-N model	results			
	R^2	RMSE	γ	σ
INFO-LSSVM	0.998	0.063	825.84	0.16
PSO-LSSVM	0.997	0.090	1000	0.20
GA-LSSVM	0.993	0.117	971.49	8.71
LSSVM	0.989	0.179	50	9.00

4.1.2 Comparison of NO_3-N model

Thirty-six sets of 210-240 nm absorbance data of NO3-N solutions were selected for predictive modelling, and 30 sets were randomly selected as training samples and 6 sets as prediction samples. The comparison of the four modelled predicted and true values of nitrate nitrogen (NO_2-N) solutions is shown in Fig.5.

A comparison of the prediction performance of the three models for NO_2-N is shown in Table 2. The RMSE of the INFO-LSSVM model is smaller than that of the PSO-LSSVM, GA-LSSVM and LSSVM models.

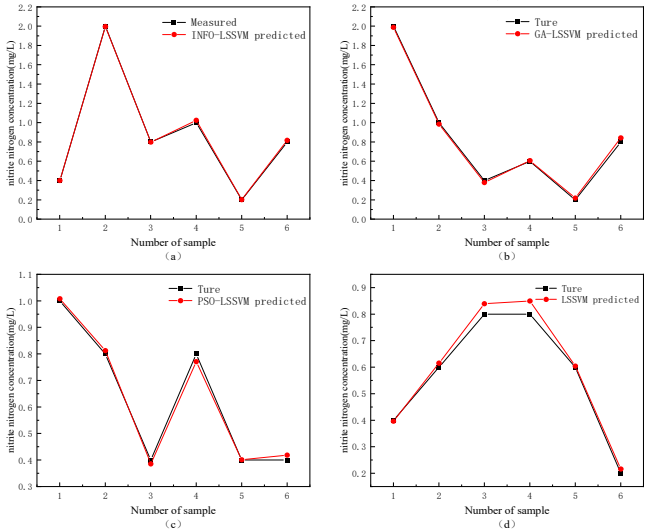

Fig. 5. Comparison of predicted and true values of the four models for NO_2-N

the R^2 of the INFO-LSSVM model is larger than that of, the PSO-LSSVM, the GA-LSSVM and the LSSVM, and the R^2 of the INFO-LSSVM is better than that of the LSSVM. data modelling of nitrate nitrogen has good results.

Table 2. Comparison of NO2-N modelling results

NO_2-N model	results			
	R^2	RMSE	γ	σ
INFO-LSSVM	0.999	0.012	887.89	0.10
PSO-LSSVM	0.995	0.016	848.27	0.20
GA-LSSVM	0.998	0.022	1000	5.73
LSSVM	0.983	0.027	50	9.00

5. Conclusion

In this paper, four different predictive modelling studies were conducted for two common water quality parameters of nitrate nitrogen and nitrite nitrogen in water quality testing, the INFO-LSSVM model had the best predictive regression effect, and the coefficient of determination, R^2, and root-mean-square error, RMSE, for the predictive regression in NO_3-N modelling were higher for the INFO-LSSVM model compared to those of the PSO-LSSVM, GA-LSSVM, and LSSVM models. LSSVM, and LSSVM models, the coefficient of determination R^2 increased by 0.08%, 0.52%, and 0.95%, respectively; while the root mean square error RMSE decreased by 29.55%, 45.56%, and 64.46%, respectively. For the predictive modelling of NO_2-N, the INFO-LSSVM model predicted a regression with a coefficient of determination R^2 of 0.9995 and a root mean square error RMSE of 0.0124, which increased by 0.40%, 0.10%, and 1.60%, respectively, compared with the PSO-LSSVM, GA-LSSVM, and LSSVM models. The root mean square error RMSE was reduced by 23.93%, 28.06%, and 49.31%, respectively. The results show that the INFO-LSSVM model can achieve better regression prediction effect, has good prediction ability for the concentration of the parameters to be measured in water samples, has high model accuracy, and has fast detection speed, which provides a reference value for water quality testing.

References

1. Fan J, Zou L, Duan T, et al. Occurrence and distribution of microplastics in surface water and sediments in China's inland water systems: a critical review[J]. Journal of Cleaner Production, 2022, 331: 129968.
2. CHEN Xiao-wei,YIN Gao-fang,ZHAO Nan-jing,et al. Study on the detection method of nitrate concentration by ultraviolet derivative spectroscopy under turbidity interference [J]. Spectroscopy and Spectral Analysis, 2019,39(09): 2912-2916.
3. Wang Y, Yu W, Li X, et al. Electrocatalytic reduction of nitrogenous pollutants to ammonia[J]. Chemical Engineering Journal, 2023, 469: 143889.
4. Alahi M E E, Mukhopadhyay S C. Detection methods of nitrate in water: A review[J]. Sensors and Actuators A: Physical, 2018, 280: 210-221.
5. Kurani A, Doshi P, Vakharia A, et al. A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting[J]. Annals of Data Science, 2023, 10(1): 183-208.
6. Deng W, Yao R, Zhao H, et al. A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm[J]. Soft computing, 2019, 23: 2445-2462.
7. CHEN Ying, HE Lei, CUI Xing-ning, et al. A mixed prediction model for nitrate concentration in water based on ultraviolet spectroscopy [J]. Spectroscopy and Spectral Analysis，2019，39(05)：1489-1494.
8. Zhou K, Liu Z, Cong M, et al. Detection of chemical oxygen demand in water based on uv absorption spectroscopy and pso-lssvm algorithm[J]. Optoelectronics Letters, 2022, 18(4): 0251-0256.
9. Wang Ling-bin. Research on Intelligent Identification and Prediction of Algal Blooms in Lakes and Reservoirs [D]. Beijing Gongshang University,2016.
10. Ahmadianfar I, Heidari A A, Noshadian S, et al. INFO: An efficient optimization algorithm based on weighted mean of vectors[J]. Expert Systems with Applications, 2022, 195: 116516.
11. Guan S, Wu T, Yang H. Research on transformer fault diagnosis method based on ACGAN and CGWO-LSSVM[J]. Scientific Reports, 2024, 14(1): 17676.
12. Wang B, Gong W, Wang Y, et al. Prediction of the yield strength of RC columns using a PSO-LSSVM model[J]. Applied Sciences, 2022, 12(21): 10911.
13. Zhang C, Sun Q, Sun W, et al. An assembly tightness recognition method for bolted connection states with singular-value entropy and GA least-squares support vector machine[J]. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 2023, 237(14): 2240-2254.
14. Ahmadianfar I, Heidari A A, Noshadian S, et al. INFO: An efficient optimization algorithm based on weighted mean of vectors[J]. Expert Systems with Applications, 2022, 195: 116516.