Science of the Total Environment 714 (2020) 136765

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Kernel functions embedded in support vector machine learning models for rapid water pollution assessment via near-infrared spectroscopy

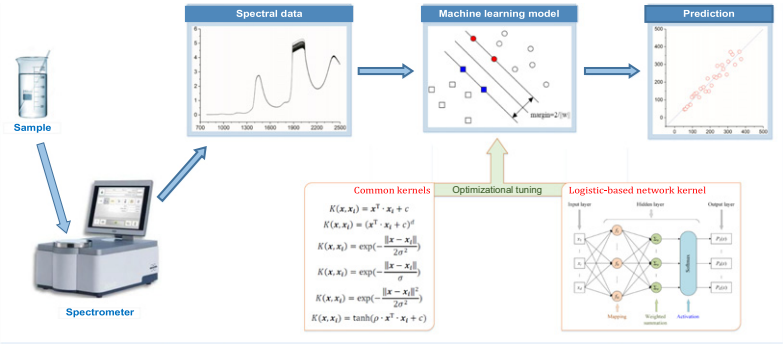
Huazhou Chen [a,b], Lili Xu [c], Wu Ai [a,b], Bin Lin [a,b], Quanxi Feng [a,b], Ken Cai [d,*]

[a] College of Science, Guilin University of Technology, Guilin 541004, China
[b] Center for Data analysis and Algorithm Technology, Guilin University of Technology, Guilin 541004, China
[c] College of Marine Sciences, Beibu Gulf University, Qinzhou 535011, China
[d] College of Automation, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China

HIGHLIGHTS

- Machine learning for NIR rapid assessment of water pollution by detecting COD value
- Parameter optimization for kernels embedded in support vector machine algorithm
- A novel type of kernel was proposed using logistic-based neural network.
- To explore a deep learning way for model optimization by the proposed network kernel

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 24 November 2019
Received in revised form 15 January 2020
Accepted 15 January 2020
Available online 17 January 2020

Editor: Damia Barcelo

Keywords:
Water pollution
Near-infrared spectroscopy
Least squares support vector machine
Logistic-based network
Kernel functions

ABSTRACT

Water pollution is a challenging problem encountered in total environmental development. Near-infrared (NIR) spectroscopy is a well-refined technology for rapid water pollution detection. Calibration models are established and optimized to search for chemometric algorithms with considerably improved prediction effects. Machine learning improves the prediction capability of NIR spectroscopy for the accurate assessment of water pollution. Least squares support vector machine (LSSVM) algorithm fits parameters to target problems in a data-driven manner. The modeling capability of this algorithm mainly depends on its kernel functions. In this study, the LSSVM method was used to establish NIR calibration models for the quantitative determination of chemical oxygen demand, which is a critical indicator of water pollution level. The effects of different kernels embedded in LSSVM were investigated. A novel kernel was proposed by using a logistic-based neural network. In contrast to common kernels, this novel kernel can utilize a deep learning approach for parameter optimization. The proposed kernel also strengthens model resistance to over-fitting such that cross-validation can be reasonably utilized. The proposed novel kernel is applicable for the quantitative determination of water pollution and is a prospective solution to other problems in the field of water resource management.

* Corresponding author at: College of Automation, Zhongkai University of Agriculture and Engineering, Zhongkai Road 501#, Guangzhou 510225, China.
E-mail address: kencaizhku@foxmail.com (K. Cai).

https://doi.org/10.1016/j.scitotenv.2020.136765
0048-9697/

1. Introduction

Water is a basic resource for the growth of agri-environmental systems and social economic environments. Water also controls the various elements of the cycle associated with regional agricultural and sustainable bioenvironmental development. Water safety is a demanding problem that directly affects animal survival and reproduction and human daily life (Briscoe, 2015; Giuliano, 2003). The precise evaluation of water safety provides a scientific basis for the rational development and utilization of regional water resources for sustainable social development. With the development of social production, industrial factories continuously discharge wastewater, which contains numerous chemical elements, thus increasing the risk of water pollution. Severe water pollution will destroy the life-critical water system and consequently cause serious harm to humans (Elleuch et al., 2018; Novotny and Hill, 2007).

Water pollution is mainly characterized by the excessive concentration of several chemical components in water. Polluted water can be considered a single complex analyte given that the internal molecule structures and functional groups of its chemical components are complex and codependent. Thus, difficulty arises in directly quantifying each target component (Ling et al., 2018; Roebeling et al., 2015). In water pollution treatment, chemical oxygen demand (COD), in accordance with the knowledge of environmental chemistry, is preferably used as an index of the amount of oxygen that can be consumed during reactions (Pasztor et al., 2009). COD test can be used to quantify the amount of oxidizable pollutants in surface water or wastewater. The conventional method for COD testing is oxidization, which detects the oxidized contribution from both organic and inorganic components. This method is time-consuming and highly influenced by manual operations (Lee, 2017). Therefore, an effective technology is required for the rapid and precise detection of COD for polluted water treatment.

Recently, studies have focused on the development of near-infrared (NIR) spectroscopy. NIR spectra result from the weak and broad overtones and the combined bands of fundamental vibrational transitions associated with functional groups (Cozzolino and Moron, 2004; El-Mesery et al., 2019). NIR spectroscopy technology excels in various fields as a sensitive, rapid, and nondestructive analytical technique for the detection of component properties in target analytes through the combined use of statistical modeling and chemometric methods. This method is widely used in the environmental, agricultural, biomedical, pharmaceutical, and food industries, mainly contributing to quantitative determination of organic targets (Borràs et al., 2015; Chen et al., 2015c; de Almeida et al., 2018). Considering that the organic component predominates in polluted water, NIR spectroscopy technology shows promise for the detection and precise quantification of COD in polluted water treatment.

NIR spectroscopy is an indirect analytical technique (Olumegbon et al., 2017). This procedure requires a set of training samples with known COD values as prior knowledge. Calibration models will be constructed on the basis of the regression of the relationship between the measured spectral data and the prepared COD values. Afterward, the models are evaluated by using a set of test samples with unknown COD values. In such an analysis, the chemoinformatic algorithm runs in a data-driven and event-triggered mode. The study of machine learning methods is a major task, especially in industrial fields, needed to solve chemometric problems for precise quantitative determination.

The machine learning algorithms used for rapid NIR determination build mathematical models on the basis of fidelity to the measured data. Logistic regression, support vector machine (SVM), and neural network are common machine learning methods used for NIR analysis (Chen et al., 2018a; Spetale et al., 2016; Uwadaira et al., 2015). Machine learning combined with spectral data mining focuses on making prediction decisions and discovering unknown properties in spectra (Fuentes et al., 2018; Richter et al., 2019). Least squares SVM (LSSVM) is a popular method for tuning kernel mapping functions. The fundamental concept of LSSVM is the mapping of original data onto a high-dimensional space by utilizing a type of kernel function; this process is succeeded by linear regression between the dependent variable and high-dimensional data (Andrecut, 2017; Chen et al., 2015a; Jayadeva et al., 2008). LSSVM includes a global optimum and exhibits model accuracy in addressing nonlinear and nonstationary data. The distribution of feature samples in high-dimensional space depends on the selection of the kernel and the use of its parameters. A suitable kernel function can determine resistance to collinear effects, which inherently exist between spectral data at different wavelengths.

This study aimed to determine the usefulness of different kernel functions embedded in LSSVM models as optimal machine learning-type calibrations when using NIR technology for the quantitative determination of the COD in polluted water samples. The calibration models in the NIR analytical field were established in accordance with the Beer–Lambert law (Pasquini, 2018; Pawar and Pratape, 2017), which defines the NIR spectrum as a linear response to a pure chemical component; the regression module is formulated as follows:

$$\boldsymbol{y} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} = \{y_i | i = 1,2\ldots n\}$ denotes the concentration of the target component for n samples, and $\boldsymbol{x} = \{x_i | i = 1,2\ldots n\}$ is the NIR spectral data. x_i is a p-dimension vector of the i-th sample. $\boldsymbol{w}^{\mathrm{T}}$ and $\boldsymbol{b}$ are regression coefficient vectors, and $\boldsymbol{\varepsilon}$ is the regression error vector.

Polluted water is a complex analyte that contains numerous chemical components. The target COD value is a comprehensive indicator of water pollution level. A small possibility indicates an inherent linear relationship between the COD and the spectra. A LSSVM algorithm with a properly operational kernel function can construct a high-dimensional function space wherein the original nonlinear relationship can be mathematically transformed into a linear one. The investigation of kernel functions is expected to provide the linear relationship defined by the Beer–Lambert law and is suitable for the prediction of the comprehensive COD value based on multicomponent response NIR spectral data. A kernel generated with a layered network is feasible to perform deep learning so as to strengthen the model resistance to over-fitting (Ghazi et al., 2017).

2. Theory and methodology

2.1. Theory of LSSVM

The LSSVM methodology employs a kernel function $\varphi(\cdot)$ to transform the original data into a high-dimensional space (recorded as a feature space); then, a set of linear equations is constructed to reduce the complexity of optimization associated with the support vectors (Tian et al., 2018).

In this process, the kernel function $\varphi(\cdot)$ constructs the corresponding feature data, which are transformed from the original spectral data, in the feature space and constructs a decision function Q by minimizing the weights that counter the regulation of prediction errors, that is:

$$Q = \min\left(\frac{1}{2}\|\boldsymbol{w}\|^2 + \gamma\|\boldsymbol{\varepsilon}\|^2\right) \text{ s.t.} \boldsymbol{y} = \boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}) + b + \varepsilon,$$

where γ is a parameter for regulation. Over-fitting can be prevented by tuning γ. Then, this convex optimization problem can be converted into a Lagrangian multiplier form as follows:

$$L(\boldsymbol{w}, \varepsilon, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \gamma\|\varepsilon\|^2 + \alpha(\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}) + b + \varepsilon - \boldsymbol{y}),$$

where α is the Lagrange multiplier.

We could obtain $\boldsymbol{w} = \alpha \cdot \varphi(\boldsymbol{x})$ and $\varepsilon = \alpha \cdot \frac{1}{2\gamma}$ by solving the Lagrangian function. Consequently, the predicted value of COD (denoted as $\hat{\boldsymbol{y}} =$

$\{\hat{y}_i | i = 1, 2 \ldots n\}$) is determined in the following manner:

$$\hat{\boldsymbol{y}} = \alpha \cdot K(\boldsymbol{x}, \boldsymbol{x_i}) + b,$$

where $K(\boldsymbol{x},\boldsymbol{x_i})$ refers to the transformed kernel function, which is defined as $K(\boldsymbol{x},\boldsymbol{x_i}) = \varphi(\boldsymbol{x}) \cdot \varphi(\boldsymbol{x_i})$; $\boldsymbol{x}$ represents the NIR spectrum of the tested sample with an unknown COD; $\boldsymbol{x_i}$ correspond to the transformed datum, which is a linear combination of the spectral matrix of the training samples and weighted on the basis of COD values. Here, α depends on the regularization parameter γ, that is, $\alpha = (\boldsymbol{x_i^T x_i} + \frac{1}{2\gamma})^{-1}$ (Chen et al., 2018b). Therefore, the prediction output by the LSSVM model is determined by tuning the regularization parameter γ and the kernel function.

The model prediction error can be estimated by applying the definition of the root mean square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_i (\hat{y}_i - y_i)^2},$$

where $\hat{y}_i$ is the i-th element of vector $\hat{\boldsymbol{y}}$ and represents the NIR predictive COD value of the i-th water sample, and $y_i (i = 1, 2 \ldots n)$ indicates the prior measured value of the corresponding i-th sample. We aimed to determine the appropriate values for the regression parameters ($\boldsymbol{w}^T, \boldsymbol{b}$) by minimizing RMSE. Alternatively, we tuned the regulation parameter γ together with the transformed kernel function $K(\boldsymbol{x},\boldsymbol{x_i})$. γ was tuned by assigning continuously changing, equal-interval changing, or logarithmic continuously changing values. Specifically, the selection of kernel function $K(\cdot)$ played a decisive role in the optimization of LSSVM models.

2.2. Common kernels for LSSVM optimization

Different kernel functions will generate diverse forms of feature data in various kinds of high-dimensional spaces (Espinoza et al., 2005). Thus, kernels must be optimized when LSSVM is applied as a special machine learning calibration method in NIR analysis. Hereafter, we studied the transformed kernel function $K(\cdot)$ instead of $\varphi(\cdot)$. We also discussed five commonly used kernel functions as shown in Table 1.

The linear kernel is the simplest kernel function. Such kernel is given by the inner product of the data plus an optional constant c. The linear kernel is normally suitable for fitting originally linear data (Ojeda et al., 2008).

The polynomial kernel is a global function, and it works efficiently when applied to orthogonal normalization data (Gorjaei et al., 2015). A constant c should be preset for this kernel, and the degree of polynomial (denoted as d) needs to be tuned.

Radial basis functions are the most commonly used kernels for SVM learning (Ramedani et al., 2014). The radial basis can be expressed in exponential, Laplacian, or Gaussian form, in which Euclidean distance is commonly used to calculate $\|\cdot\|$, and parameter σ represents the degree of generalization (also called the kernel width). In particular, the Gaussian radial basis function is regarded as the blind choice for LSSVM optimization when the kernel is excluded from the investigation.

The sigmoid kernel was originally generated by studies on neural networks for data mining (Huang, 2015), which often uses the sigmoid function as an activation function for artificial neurons. This kernel is popularly applied for machine and deep learning. The slope ρ and intercept constant c are specific parameters for tuning.

2.3. Logistic-based network kernel for deep learning

The commonly used kernel functions exhibit interesting characteristics when applied, that is, the LSSVM model prevents over-fitting and exhibits moderate robustness and stability if the appropriate kernel, such as the Gaussian kernel, is used (Boulkaibet et al., 2018; Razavi et al., 2019). On this basis, we attempted to study if a novel kernel function can be constructed by embedding a simple perceptron network into the LSSVM models to enable deep learning for model optimization. The logistic function is easily operated (Kudryashov, 2015) and can be used in network formation.

Suppose a dataset $\{(x_i, y_i)\}_{i=1}^n$ exists, where x_i is the feature data, y_i is a binary class tag, and $y_i \in \{\theta_1, \theta_2\}$. The simplest logistic regression formula is developed as follows:

$$P_1(x) = \frac{1}{1 + \exp\left(-\left(\beta^T x + \beta_0\right)\right)},$$

where $P_1(x_i)$ denotes the probability of $y_i = \theta_1$. $\beta \in R^d$ and $\beta_0 \in R$ are the regression coefficients which can be estimated by maximizing the conditional log-likelihood:

$$\mu(\beta, \beta_0) = \sum_{i=1}^{n} y_{i1} \ln P_1(\mathrm{x}_i) + (1 - y_{i1}) \ln (1 - P_1(\mathrm{x}_i)),$$

$$\mathrm{s.t.} y_{i1} = \begin{cases} 1, \text{if } y_i = \theta_1 \\ 0, \text{if } y_i = \theta_2 \end{cases}$$

For multinomial logistic regression wherein $y_i \in \{\theta_1, \theta_2 \ldots \theta_K\}$, the posterior probability of class θ_k can be expressed as follows:

$$P_k(x) = \frac{\exp\left(\beta_k^T x + \beta_{k0}\right)}{\sum_{l=1}^{K} \exp\left(\beta_l^T x + \beta_{l0}\right)},$$

where the coefficients (β_k, β_{k0}), $k = 1, 2 \ldots K$ can be estimated by maximizing the conditional likelihood similar to a binomial case. The transformation from the linear combinations of features $\beta_k^T x + \beta_{k0} \in R$ to probabilities in $[0, 1]$, as described by the formula of $P_k(x)$, is often referred to as softmax transformation.

Herein, we assumed that the performance of nonlinear logistic regression is equivalent to that of perceptron neural networks, which are applied for deep learning. In the logistic model, we constructed a nonlinear function $f(x)$ to map data from R^d to R. Thus, $f(x)$ is considered

Table 1
Common kernel functions for LSSVM models.

Kernel	Function
Linear kernel	$K(\boldsymbol{x},\boldsymbol{x_i}) = \boldsymbol{x}^T \cdot \boldsymbol{x_i} + c$
Polynomial kernel	$K(\boldsymbol{x},\boldsymbol{x_i}) = (\boldsymbol{x}^T \cdot \boldsymbol{x_i} + c)^d$
Exponential kernel	$K(\boldsymbol{x},\boldsymbol{x_i}) = \exp(-\frac{\|\boldsymbol{x}-\boldsymbol{x_i}\|}{2\sigma^2})$
Laplacian kernel	$K(\boldsymbol{x},\boldsymbol{x_i}) = \exp(-\frac{\|\boldsymbol{x}-\boldsymbol{x_i}\|}{\sigma})$
Gaussian kernel	$K(\boldsymbol{x},\boldsymbol{x_i}) = \exp(-\frac{\|\boldsymbol{x}-\boldsymbol{x_i}\|^2}{2\sigma^2})$
Sigmoid kernel	$K(\boldsymbol{x},\boldsymbol{x_i}) = \tanh(\rho \cdot \boldsymbol{x}^T \cdot \boldsymbol{x_i} + c)$

a new feature extracted from data $\{x_i\}$. A neural network can be constructed if multiple functions are applied. This network comprises layers of elementary computing neurons (see Fig. 1). Each layer computes a vector of new feature as the function of the output from the previous layer. The output layer is typically a softmax layer with K output neurons. In the network, the new feature $f(x)$ and coefficients (β_k, β_{k0}) are simultaneously estimated by minimizing the cost functions defined in machine learning. Given that this logistic-based network is constructed as a kernel function for the optimization of LSSVM models, the number of output neurons is usually set to equal the number of inputs (i.e., $K = d$). If the network is built with multiple hidden layers (L), then the deep learning framework can be embedded into the LSSVM model for optimization. The network is designed for training with different numbers of L and neurons (H) in each layer. Network parameters, including the learning coefficient, momentum, and iteration number, are initialized with optional values at the initial training run. The network training results can be obtained afterward.

3. Data preparation

3.1. Data collection

A total of 83 polluted water samples were collected from wastewater disposed by chemical industries. The COD values of all samples were measured in the laboratory by using the potassium permanganate oxidation method (Ma et al., 2016). The values were applied as references for kernel LSSVM calibration based on NIR technology. The reference COD values for all samples ranged from 52 mg/L to 382 mg/L. The statistic average and standard deviations were 232.2 and 97.2 (mg/L), respectively.

The NIR spectra of target samples were recorded by using FOSS NIR Systems 5000 Grating Spectrometer equipped with an InGaAs detector (Foss NIRsystems Inc., Denmark). The surrounding temperature and humidity were maintained at 25 ± 1 °C and 46% ± 1% relative humidity, respectively. The full scanning range of the spectrum was set from 780 nm to 2500 nm with a resolution of 2 nm to generate 860 solid wavelength points at which the spectral data were acquired. Each sample was measured thrice, and the mean spectrum was calculated and prepared for modeling. Fig. 2 shows the NIR spectra of all 83 polluted water samples.

It is noted that the reference values involve the information of organic and inorganic components in polluted water, and NIR technology mainly calibrates on the organic targets. Thus there exist some unknown prediction errors related to the inorganic component. To take advantage of the rapid detection of NIR spectroscopy, we did not distinguish the organic and inorganic values using in-lab time-consuming chemical experiments. Instead, we launched to study novel chemometric methods to minimize the influence of this unknown bias. Therefore, the logistic-based network kernel is inevitably investigated for LSSVM model optimization based on the deep learning theory.

3.2. Sample allocation into training and testing sets

NIR spectral analytical systems require all samples to be split into two parts for training and testing to obtain model evaluation results instantaneously (Chen et al., 2015b). The test samples were assumed to be unknown, and their COD values were calculated by the training model. The model-predicted and prior measured values were used to calculate the RMSE (denoted as RMSET for the test samples) for the evaluation of prediction error to reveal the prediction performance of the training model. We selected 28 samples (~33% of all samples in the pool) for the test set to ensure that the test samples were objective and representative. These samples were randomly selected from the whole sample pool and were independent of training. The remaining 55 samples were used for training.

The training samples were used to establish the calibration models by utilizing stand-by machine learning methods. The NIR spectral data and the measured COD values were obtained. The calibration models were trained with basic fidelity to data to identify the optimal model parameters on the basis of internal predictive results. The regulation parameter γ should be tuned, and the kernel function $K(\cdot)$ should be investigated to optimize the LSSVM model. Sample training was conducted in running cross-validation for internal modeling. m-fold cross-validation was effectively used for regression training. The training sample pool was divided into m parts. One part was retained for training prediction, and the other m-1 parts were used to establish the LSSVM model. Cross-validation was repeated m times, with each of the m parts used exactly once for prediction. The m predictive results can then be averaged to produce a single RMSE calculation (denoted as RMSECV for cross-validation). The LSSVM models were optimized by adjusting γ and tuning the kernel parameters to determine the minimum RMSECV.

4. Results and discussion

The LSSVM method was applied for the NIR analysis of the COD values of 83 polluted water samples to evaluate the extent of water contamination. A total of 28 and 55 samples were randomly selected for testing and training, respectively. A fivefold cross-validation procedure was designed for the training. The results were discussed on the basis of model training. Finally, the model was evaluated by using the test samples.

The first part of this section discusses the best LSSVM outputs under the optimal parameters using six common kernel functions. The second

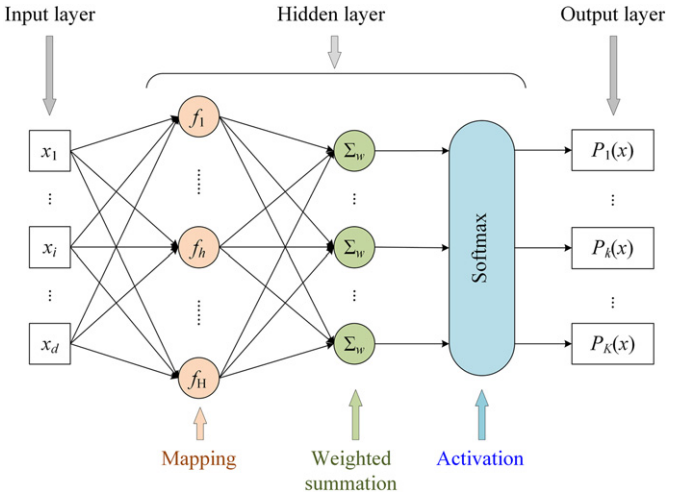


Fig. 1. The structure of perceptron neural networks with generalized logistic regression.

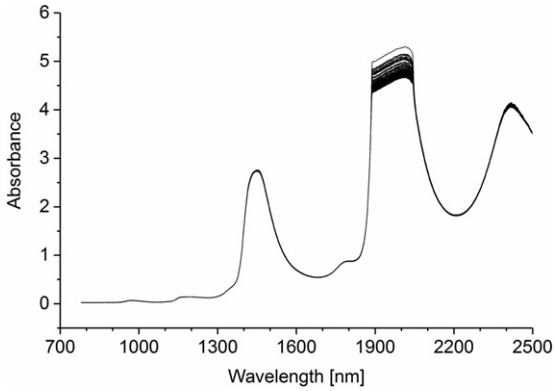


Fig. 2. The NIR spectra of all 83 polluted water samples.

part focuses on the investigation of the logistic-based network kernel with deep learning.

4.1. Predictions from common kernels

The LSSVM models were established for training samples in fivefold cross-validation mode. LSSVM optimization initially involved the tuning of regulation parameter γ. Then, the parameters in kernel functions were tuned. Grid search mode must be used to screen these parameters to obtain a smooth subarea to reduce the predictive error as well as to decrease the possibility of over-fitting.

The regulation parameter γ was changed from 10 to 300 with a step size of 10. This variable was optimized in combination with the kernel parameters. The degree of polynomial was set as 2, 3, 4, 5 and 6 in polynomial kernel tuning. The kernel width σ was changed from 1 to 20 in the radial basis kernels (exponential, Laplacian, and Gaussian) such that σ^2 would increase to 400. The slope ρ in the sigmoid kernel was set as $\rho = 1/n$, where n continuously changed from 1 to 55 (i.e., the number of training samples). The linear kernel lacked a parameter for tuning. The grid search LSSVR models were established with parameter tuning for the training data. The optimal parameters and predictive results were obtained in accordance with the acquired minimum RMSECV (see Table 2).

4.2. Predictions from the logistic-based network kernel

The constructed logistic-based network was used as a kernel function for the optimization of the LSSVM models in deep learning mode. In the network, information at all wavelengths should be considered for inclusion as the input data. The original measured spectrum contained 860 wavelength variables. The computation for deep learning might become overloaded if these variables were directly used as input nodes to the network. To solve this problem, we selected several variables that can represent most of the information in the 860 wavelengths. Principal components were selected before network deep learning. A total of 82 comprehensive components were identified with the fully transformed information (Fig. 3). Therefore, we started network deep learning with 82 input nodes, and the number of output nodes was set to be equal to that of the input to facilitate LSSVM kernel optimization.

In network training, the numbers of L and H in each layer were tuned. We tested the networks by changing L from 1 to 8 and changing H from 1 to 20. The learning coefficient (0.01–0.5), momentum (0.05–0.45), and iteration number (300–2400) were automatically selected in accordance with the error and the degree of approximation. The network was embedded during LSSVM optimization for each combination of L and H. The regulation parameter γ was tuned from 10 to 300 with a step size of 10. Thus, the LSSVM model with the logistic-based network kernel was optimized by screening combined parameters (L,H,γ). All possible models were established through fivefold validation, and the optimal parameters were identified in grid search mode to obtain the minimum RMSECV. Fig. 4 shows the optimal results corresponding to each value of γ. As inferred from Fig. 4, the minimum RMSECV was 20.19 mg/L (i.e., 9.02% of the average COD value) when $\gamma =$ 170. Additionally, several values of γ provided acceptable quasiminimum prediction error. When 10% of the average COD value was considered the metric for modeling performance, the γ values of 90, 100, 110, 160, and 210 indicated good performance. These values are marked by semisolid circles in Fig. 4. Their corresponding minimum RMSECV values were 21.46, 21.14, 21.27, 20.45, and 21.35 mg/L. The kernel performances based on these six γ values were studied, and the RMSECV varied in grid details in accordance with the changes in L and H. The training results are graphically presented by contour cool maps in Fig. 5. In the maps, a cool color indicates a small value of RMSECV. As illustrated in Fig. 5, the output results from the deep network training of L and H remained below 36 mg/L (~15% of the average COD value). The best optimal parameters (L,H) were identified as $(5,12)$, respectively, and numerous appropriate predictions were obtained at γ values of 160 and 170. These findings provide options for parameter design if the network kernel method is applied to detect water pollution by using computationally simple models.

In summary, deep learning optimization of LSSVM models based on the logistic-based network kernel depends on the values of the combined parameterizing procedure on (L,H,γ). The best effect for the rapid analysis of water COD through NIR spectroscopy was obtained when (L,H,γ) values were $(5,12,170)$, respectively, specifically, when the LSSVM transforming kernel was deeply trained by the network with 5 L and 12 nodes in each layer. Further training of this kernel-optimized LSSVM model with a regulation of $\gamma =$ 170 to the support vectors provided the training predictive RMSECV of 20.19 mg/L, which was less than 10% of the average COD value. The modeling results obtained from the logistic-based network-kernel LSSVM model were superior to those obtained with the common-kernel LSSVM.

Table 2
The optimal prediction results of the 5-fold cross-validation LSSVM models corresponding to different common kernel functions.

	γ	Kernel parameter	RMSECV (mg/L)
Linear kernel	240	–	46.4
Polynomial kernel	170	$d = 5$	41.6
Exponential kernel	190	$\sigma = 14$	32.9
Laplacian kernel	210	$\sigma = 17$	36.7
Gaussian kernel	160	$\sigma = 11$	29.4
Sigmoid kernel	130	$\rho = 1/34$	27.5

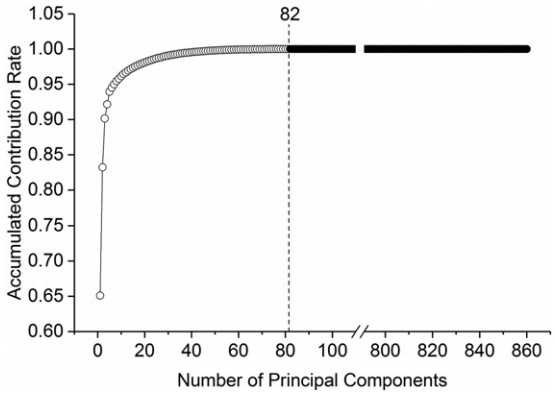


Fig. 3. The variance contributions of the principal components to the 860 wavelengths.

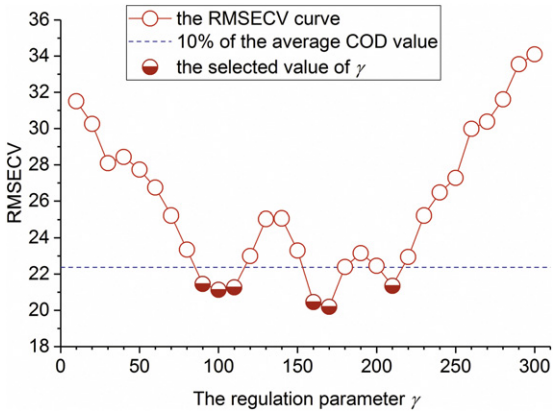


Fig. 4. The optimal RMSECV corresponding to each value of γ.

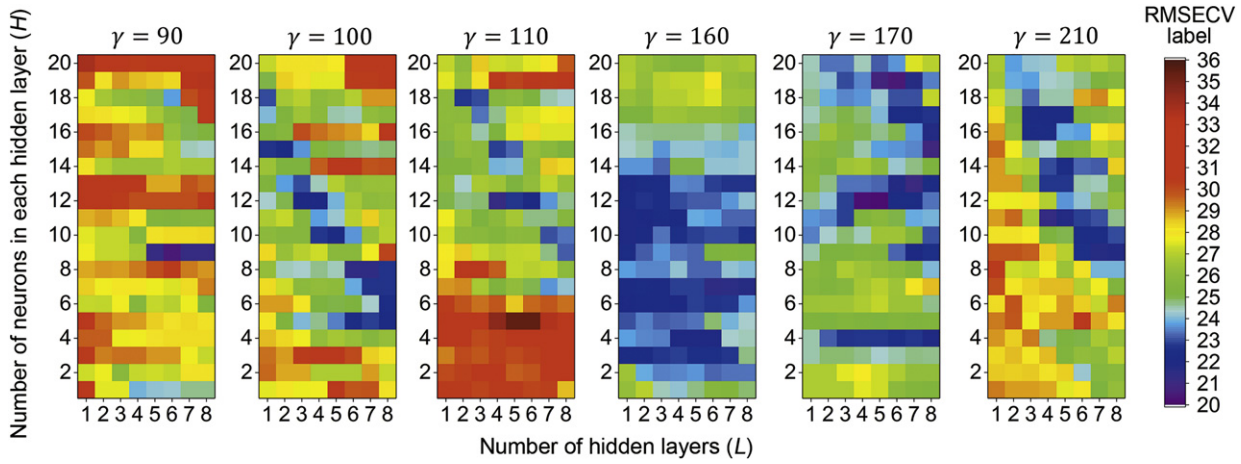


Fig. 5. The contour cool map presenting the optimal training results for each combination of *L* and *H*.

4.3. Model evaluation based on test samples

The predictive performances of the LSSVM regression method with the logistic-based network kernel was further evaluated by predicting 28 test samples, which were assumed to be "unknown" because they were independent of training. A common strategy for reliable evaluation involves adopting models and the optimal training parameters $(L,H,\gamma) = (5,12,170)$, respectively, to test samples and then calculating the prediction error (RMSET) of the target COD values. Fig. 6 displays the regression plots between the NIR-predicted and presaved measured values. As shown in Fig. 6, the predicted RMSET was 27.7 (i.e., ~13.5% of the average value). The prediction coefficient (R^2) was 0.912 (higher than 0.9) and the p-value was smaller than 1‰. It seems that the prediction result is not good enough as the predicted RMSET did not reach below 10% although it is the optimized best model obtained from training. But we have to remind that the NIR calibration on COD mainly targeted on the contribution of organic components, while the inorganic contribution that is included in the reference value was not able to be predicted. In this case, it is acceptable that the RMSET was a little higher than 10%, accompanied with the predictive R^2 is higher than 0.9. Thus we conclude that the model-predicted COD values for 28 test samples were close to the measured reference values. The test effect for the "unknown" test samples was satisfactory because the nonlinear LSSVM model was optimized by the deep learning network kernel. The testing results showed that over-fitting was avoided. Besides the best optimal values, numerous quasi-optimal values of (L,H,γ) were generated by network deep searching. Nevertheless, these results were expected as acceptable modeling options.

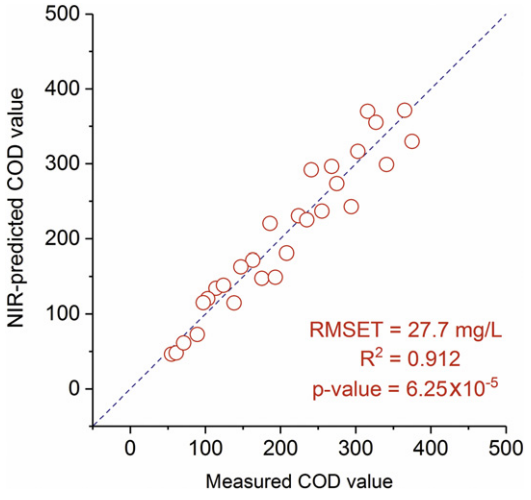


Fig. 6. The regression plots between and the measured COD values and the LSSVM-predicted values using the logistic-based network deep learning kernel.

5. Conclusions

The use of machine learning for the rapid detection of the COD of polluted water samples through NIR spectroscopy was studied. LSSVM kernels were investigated. A multilayer network was employed as the kernel for the optimization of LSSVM models to improve model prediction capability. The network was designed with interactively tuned numbers of *L* and *H*. The easily operated logistic function was used for activation. This logistic-based network kernel exhibited remarkable advantages over common kernels, such as linear, polynomial, radial basis, and sigmoid functions. The kernel presented advantages in three aspects: (1) deep optimization in the common parameter grid search mode, (2) achievement of global optimal results regardless of random or selected inputs, and (3) avoidance of over-fitting to enable cross-validation modeling.

The COD values of polluted water samples were predicted through NIR calibration to estimate water pollution level. The network kernel enhanced the performance of the LSSVM models. The best network structure included 5 hidden layers and 12 neurons in each layer and regulation parameter tuning. The optimal predicted RMSECV was 20.19 mg/L, which was less than 10% of the average of measured COD values. This appropriate training result was superior to the optimal predictions provided by common kernels. In model evaluation, the predictive RMSET of a little higher than 10% was tolerant as there are inorganic components not detected by NIR. Additionally, deep learning identified numerous available optimal models with different parameter values. Therefore, the logistic-based network is a novel kernel that can be used to optimize LSSVM through deep learning. This network also improves machine learning methods for the quantitative determination of water pollution and consequently provides suggestions for solving problems in water safety. Prospectively, the models are expected to be further improved if the reference value of COD is measured only involving the contribution of organic components.

Declaration of competing interest

The authors declare that they do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61505037, 61703117, 61763008), China Postdoctoral Science Foundation (2018T110880, 2017M620375) and the Natural Science Foundation of Guangxi Province (2018GXNSFAA050045).

References

Andrecut, M., 2017. Randomized kernel methods for least-squares support vector machines. Int. J. Mod. Phys. C 28, 1750015.
Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., Busto, O., 2015. Data fusion methodologies for food and beverage authentication and quality assessment–a review. Anal. Chim. Acta 891, 1–14.
Boulkaibet, I., Belarbi, K., Bououden, S., Chadli, M., Marwala, T., 2018. An adaptive fuzzy predictive control of nonlinear processes based on multi-kernel least squares support vector regression. Appl. Soft Comput. 73, 572–590.
Briscoe, J., 2015. Water security in a changing world. Daedalus 144, 27–34.
Chen, H.-Z., Ai, W., Feng, Q.-X., Tang, G.-Q., 2015a. FT-MIR modelling enhancement for the quantitative determination of haemoglobin in human blood by combined optimization of grid-search LSSVR algorithm with different pre-processing modes. Anal. Methods 7, 2869–2876.
Chen, H.-Z., Shi, K., Cai, K., Xu, L.-L., Feng, Q.-X., 2015b. Investigation of sample partitioning in quantitative near-infrared analysis of soil organic carbon based on parametric LS-SVR modeling. RSC Adv. 5, 80612–80619.
Chen, H.-Z., Tang, G.-Q., Ai, W., Xu, L.-L., Cai, K., 2015c. Use of random forest in FTIR analysis of LDL cholesterol and tri-glycerides for hyperlipidemia. Biotechnol. Prog. 31, 1693–1702.
Chen, H., Liu, X., Jia, Z., Liu, Z., Shi, K., Cai, K., 2018a. A combination strategy of random forest and back propagation network for variable selection in spectral calibration. Chemom. Intell. Lab. Syst. 182, 101–108.
Chen, H., Xu, L., Jia, Z., Cai, K., Shi, K., Gu, J., 2018b. Determination of parameter uncertainty for quantitative analysis of shaddock peel pectin using linear and nonlinear near-infrared spectroscopic models. Anal. Lett. 51, 1564–1577.
Cozzolino, D., Moron, A., 2004. Exploring the use of near infrared reflectance spectroscopy (NIRS) to predict trace minerals in legumes. Anim. Feed Sci. Technol. 111, 161–173.
de Almeida, V.E., de Araújo Gomes, A., de Sousa Fernandes, D.D., Goicoechea, H.C., Galvão, R.K.H., Araújo, M.C.U., 2018. Vis-NIR spectrometric determination of Brix and sucrose in sugar production samples using kernel partial least squares with interval selection based on the successive projections algorithm. Talanta 181, 38–43.
Elleuch, B., Bouhamed, F., Elloussaief, M., Jaghbir, M., 2018. Environmental sustainability and pollution prevention. Environ. Sci. Pollut. Res. 25, 18223–18225.
El-Mesery, H.S., Mao, H., Abomohra, A.E.F., 2019. Applications of non-destructive technologies for agricultural and food products quality inspection. Sensors 19 (1–23), 846.
Espinoza, M., Suykens, J.A.K., De Moor, B., 2005. Kernel based partially linear models and nonlinear identification. IEEE Trans. Automat. Contr. 50, 1602–1606.
Fuentes, S., Hernández-Montes, E., Escalona, J.M., Bota, J., Viejo, C.G., Poblete-Echeverría, C., Tongson, E., Medrano, H., 2018. Automated grapevine cultivar classification based on machine learning using leaf morpho-colorimetry, fractal dimension and near-infrared spectroscopy parameters. Comput. Electron. Agric. 151, 311–318.
Ghazi, M.M., Yanikoglu, B., Aptoula, E., 2017. Plant identification using deep neural networks via optimization of transfer learning parameters. Neurocomputing 235, 228–235.
Giuliano, L.J., 2003. Balancing priorities: the role of industry in water resource management. Water Sci. Technol. 47, xxi–xxv.
Gorjaei, R.G., Songolzadeh, R., Torkaman, M., Safari, M., Zargar, G., 2015. A novel PSO-LSSVM model for predicting liquid rate of two phase flow through wellhead chokes. J. Nat. Gas Sci. Eng. 24, 228–237.
Huang, G.-B., 2015. What are extreme learning machines? Filling the gap between Frank Rosenblatt's dream and John von Neumann's puzzle. Cognit. Comput. 7, 263–278.
Jayadeva, Khemchandani, R., Chandra, S., 2008. Regularized least squares support vector regression for the simultaneous learning of a function and its derivatives. Inf. Sci. (Ny) 178, 3402–3414.
Kudryashov, N.A., 2015. Logistic function as solution of many nonlinear differential equations. Appl. Math. Model. 39, 5733–5742.
Lee, B.-J., 2017. Assessment of biodegradable and refractory cod fractions using oxygen utilization rate and ultimate biochemical oxygen demand tests. J. Korean Soc. Water Sci. Technol. 25, 53–61.
Ling, M., Lv, C., Guo, X., 2018. Quantification method of water environmental value loss caused by water pollution based on emergy theory. Desalin. Water Treat. 129, 299–303.
Ma, Y., Tie, Z., Zhou, M., Wang, N., Cao, X., Xie, Y., 2016. Accurate determination of low-level chemical oxygen demand using a multistep chemical oxidation digestion process for treating drinking water samples. Anal. Methods 8, 3839–3846.
Novotny, V., Hill, K., 2007. Diffuse pollution abatement–a key component in the integrated effort towards sustainable urban basins. Water Sci. Technol. 56, 1–9.
Ojeda, F., Suykens, J.A.K., De Moor, B., 2008. Low rank updated LS-SVM classifiers for fast variable selection. Neural Netw. 21, 437–449.
Olumegbon, I.A., Oloyede, A., Afara, I.O., 2017. Near-infrared (NIR) spectroscopic evaluation of articular cartilage: a review of current and future trends. Appl. Spectrosc. Rev. 52, 541–559.
Pasquini, C., 2018. Near infrared spectroscopy: a mature analytical technique with new perspectives – a review. Anal. Chim. Acta 1026, 8–36.
Pasztor, I., Thury, P., Pulai, J., 2009. Chemical oxygen demand fractions of municipal wastewater for modeling of wastewater treatment. Int. J. Environ. Sci. Technol. 6, 51–56.
Pawar, S.B., Pratape, V.M., 2017. Fundamentals of infrared heating and its application in drying of food materials: a review. J. Food Process Eng. 40, e12308.
Ramedani, Z., Omid, M., Keyhani, A., Shamshirband, S., Khoshnevisan, B., 2014. Potential of radial basis function based support vector regression for global solar radiation prediction. Renew. Sust. Energ. Rev. 39, 1005–1011.
Razavi, R., Bemani, A., Baghban, A., Mohammadi, A.H., Habibzadeh, S., 2019. An insight into the estimation of fatty acid methyl ester based biodiesel properties using a LSSVM model. Fuel 243, 133–141.
Richter, B., Rurik, M., Gurk, S., Kohlbacher, O., Fischer, M., 2019. Food monitoring: screening of the geographical origin of white asparagus using FT-NIR and machine learning. Food Control 104, 318–325.
Roebeling, P.C., Cunha, M.C., Arroja, L., Van Grieken, M.E., 2015. Abatement vs. treatment for efficient diffuse source water pollution management in terrestrial-marine systems. Water Sci. Technol. 72, 730–737.
Spetale, F.E., Bulacio, P., Guillaume, S., Murillo, J., Tapia, E., 2016. A spectral envelope approach towards effective SVM-RFE on infrared data. Pattern Recogn. Lett. 71, 59–65.
Tian, Z., Li, S., Wang, Y., Wang, X., 2018. Mixed-kernel least square support vector machine predictive control based on improved free search algorithm for nonlinear systems. Trans. Inst. Meas. Control. 40, 4382–4396.
Uwadaira, Y., Shimotori, A., Ikehata, A., Fujie, K., Nakata, Y., Suzuki, H., Shimano, H., Hashimoto, K., 2015. Logistic regression analysis for identifying the factors affecting development of non-invasive blood glucose calibration model by near-infrared spectroscopy. Chemom. Intell. Lab. Syst. 148, 128–133.