Analytical Biochemistry 405 (2010) 184–191

Contents lists available at ScienceDirect

Analytical Biochemistry

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

Prediction of nitrophenol-type compounds using chemometrics and spectrophotometry

Ling Gao, Shouxin Ren *

Department of Chemistry, Inner Mongolia University, Huhehot, Inner Mongolia 010021, People's Republic of China

ARTICLE INFO

Article history:
Received 26 March 2010
Received in revised form 31 May 2010
Accepted 19 June 2010
Available online 25 June 2010

Keywords:
Wavelet packet transform
Elman recurrent neural network
Least square support vector machines
Nitrophenol-type compounds
Chemometrics

ABSTRACT

Two chemometric methods, WPT–ERNN and least square support vector machines (LS–SVM), were developed to perform the simultaneous spectrophotometric determination of nitrophenol-type compounds with overlapping spectra. The WPT–ERNN method is based on Elman recurrent neural network (ERNN) regression combined with wavelet packet transform (WPT) preprocessing and relies on the concept of combining the idea of WPT denoising with ERNN calibration for enhancing the noise removal ability and the quality of regression without prior separation. The LS–SVM technique is capable of learning a high-dimensional feature with fewer training data and reducing the computational complexity by requiring the solution of only a set of linear equations instead of a quadratic programming problem. The relative standard errors of prediction (RSEPs) obtained for all components using WPT–ERNN, ERNN, LS–SVM, partial least squares (PLS), and multivariate linear regression (MLR) were compared. Experimental results showed that the WPT–ERNN and LS–SVM methods were successful for the simultaneous determination of nitrophenol-type compounds even when severe overlap of spectra was present.

Nitrophenol-type compounds are widely applied in industries such as chemistry, petrochemical, and pharmacy and are one of the most important raw materials for synthetic medicine, dye, herbicides, insecticides, resins, and explosives. Because nitrophenol-type compounds are cancer causing and highly poisonous, it is very important to test and analyze these compounds to avoid the harm and danger they may cause to humans, other living creatures, and biological systems. Simultaneous determination of nitrophenol-type compounds is very difficult because the similarity in their structures produces overlapping signals. Modern instruments are capable of generating huge data sets, but the acquired data are commonly characterized by a high level of redundancy. As a result, many problems are created by the noisy and highly collinear data, including poor prediction results, unstable models, and overfitting. Moreover, traditional methods are not able to perform direct determination without previous separation. In general, the number of objects is less than the number of variables; for ultraviolet–visible (UV–VIS)[1] spectrophotometry, the number of measurements is less than the number of wavelengths. In such a situation, multivariate linear regression (MLR) leads to an ill-posed inverse problem and causes a number of difficulties such as poor predictions, overfitting, and collinearity problems. In the application of MLR, the number of samples must be equal to or greater than the number of variables (i.e., absorbances in different wavelengths). Aiming to resolve these issues, chemometric methods [1–4] are being developed to eliminate irrelevant information contained in these raw data and reduce dimensionality of the data prior to calibration. Several chemometric methods, such as principal component regression (PCR) and partial least squares (PLS), have recently emerged as means to overcome this difficulty by eliminating the less important principal components or latent variables [5,6]. However, these methods are preferably applicable to linear systems. Overlap among signals and violations of the Beer–Lambert law can often cause nonlinearities. Other sources of nonlinearity in spectrophotometric measurements are nonlinear instrument responses and interactions between components. Artificial neural network (ANN) is one of the most broadly used mathematical algorithms for regression problems [7,8]. ANN is a mathematical model of which composition is inspired by the structure of the human brain. Recently, it has been proposed that ANN can be used to solve regression problems by acting as nonparametric calibration methods that have the ability to learn from a set of examples without requiring any knowledge of the model type and to generalize this knowledge to new situations [9–12]. ANN has the outstanding power for modeling both linearity and nonlinearity systems and has shown better prospects as a calibration model than as PLS and PCR methods in nonlinearity systems. Currently, the most widely used ANN is a multilayer feedforward

* Corresponding author. Fax: +86 471 4992984.
E-mail address: cersx@mail.imu.edu.cn (S. Ren).

[1] *Abbreviations used:* UV–VIS, ultraviolet–visible; MLR, multivariate linear regression; PCR, principal component regression; PLS, partial least squares; ANN, artificial neural network; MLFN, multilayer feedforward network; BP, backpropagation; ERNN, Elman recurrent neural network; RNN, recurrent neural network; FT, Fourier transform; DCT, discrete cosine transform; WT, wavelet transform; WPT, wavelet packet transform; WP, wavelet packets; WPT–ERNN, wavelet packet transform-based Elman recurrent neural network; SVM, support vector machines; LS–SVM, least square support vector machines; SVMR, support vector machine regression; FWPT, fast wavelet packet transform; DWT, discrete wavelet transform; SURE, Stein unbiased risk estimation; SEP, absolute standard error of prediction; RSEP, relative standard error of prediction; MSE, mean squared error; Db, Daubechies; NIPALS, nonlinear iterative partial least squares.

0003-2697/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2010.06.032

network (MLFN) with backpropagation (BP) algorithm. However, the BP–MLFN method often has the deficiency of slow convergence, is prone to the existence of many local minima during training, and tends to overfit. Much attention has been paid to solving these problems and to facilitating the training process into the global minimum. An ANN called Elman recurrent neural network (ERNN) was used in this case. It was introduced into the ANN literature by Elman in 1990 [13]. A recurrent neural network (RNN) has recurrent links between its layers and uses these links to provide networks with dynamic memory. In the Elman network, a so-called context layer, which provides the network with memory, is added to the conventional feedforward neural network. The RNN is able to tackle the linear and nonlinear relationships between spectra and concentrations and to reduce the computational complexity of the training procedure. Until now, RNN has rarely been applied to analytical chemistry [14].

The quality of multivariate calibration is dependent, to a great degree, on the quality of the spectra. The quality of UV–VIS spectra is worsened by overlap, noise, collinearity, nonlinearity, and sensitivity to external conditions such as change of temperature, pressure, apparatus, and physical condition of the samples. The quality of the spectra could be improved by appropriate data pretreatment and feature extraction. Discrete transform techniques are important ways of improving the quality of the spectra. The most widely used transform techniques are the Fourier transform (FT), discrete cosine transform (DCT), Hankel transform, Hartley transform, and Hadamard transform. Wavelet transform (WT) [15–17] and wavelet packet transform (WPT) [18,19] have also received considerable attention. WT and WPT have the ability to provide information in the time and frequency domain, so they can be used to convert data from their original domain into the wavelet domain, where the representation of a signal is sparse and it is easier to remove noise from the signal. These characteristics of WT and WPT make them potential techniques to perform data reduction, feature extraction, and denoising [20–25]. The wavelet functions are localized in both time and frequency. Wavelet packets (WP) is a generalization of wavelets and particular linear combinations of wavelets [26–29]. WP inherits the property of time–frequency localization from wavelets but offers more flexibility than wavelets in representing different types of signals. In our research, a wavelet packet transform-based Elman recurrent neural network (WPT–ERNN) was developed by combining a regression model based on ERNN with data denoising based on a WPT. To the authors' best knowledge, this method that uses the advantages of combining WPT with an RNN approach has not been used for spectrophotometric multicomponent analysis outside our research group.

As compared with ANN, support vector machines (SVM), pioneered by Vapnik [30,31], are a kind of machine learning method founded on modern statistical learning theory and have notable attributes, including the absence of local minima and high generalization ability. Suykens et al. [32] introduced a modified version of SVM called least square support vector machines (LS–SVM) that requires solving a set of linear equations instead of a quadratic programming problem, making it computationally simpler than SVM. Thus, LS–SVM regression has the advantage of providing the capability of learning a high-dimensional feature with fewer training data and produces a global estimate of the concentration of multicomponents. The LS–SVM method not only has the ability to model the linear relationship between D and C but also can grasp the nonlinearity existing in raw data. SVM and LS–SVM represent relatively recent machine learning methods [33–38] and have found some applications in biological sample analysis, image analysis, and the classification and diagnosis of diseases [39–42].

WPT, ANN, and SVM are three of the most successful advances in the field of applied mathematics during the past few years. The aim of the current work was to use the advantages of these techniques and employ the LS–SVM and WPT–ERNN methods to perform spectrophotometric multicomponent analysis. In this study, these two proposed methods were applied to the simultaneous spectrophotometric determination of a system containing *p*-nitrophenol, *o*-nitrophenol, and 2,4-dinitrophenol.

Theory

Support vector machines

The original theory of SVM introduced by Vapnik was a valuable tool for solving pattern recognition and classification problems [30,31]. The basic idea of SVM is to map the data set X into a higher dimensional feature space F via nonlinear mapping Φ and then to perform linear regression in the hyperspace. Vapnik expanded the concept of SVM and developed support vector machine regression (SVMR) by introducing an alternative cost function. The objective of SVMR is to find a regression function that relates the input data to the desired output property. In general, SVMR involves a solution of a quadratic programming problem. With the help of the Lagrange multiplier method and a quadratic programming algorithm, the constrained optimization problem is solved. For the details of SVM and SVMR algorithm, refer to Refs. [30,35–37,40].

Least squares support vector machines

LS–SVM is a modified version of SVM proposed by Suykens et al. [32] and has an important advantage of requiring the solution of only a set of linear equations instead of a quadratic programming problem.

The optimization problem is to minimize the cost function (J):

$$J_{\text{LS–SVM}} = \frac{1}{2}\|w\|^2 + \gamma \sum_{i=1}^{n} e_i^2 \tag{1}$$

subject to

$$y_i - wx_i - b = e_i. \tag{2}$$

The first part of the cost function is weight decay, which is used to regularize weight sizes and penalize quadratically large weights to make them converge to smaller values so as to avoid deteriorating the generalization ability of SVM. The second part of the cost function is the regression error (e_i) for all of the n training objects. The parameter γ is the regularization parameter, which indicates the relative weight of the error term as compared with the first part and must be optimized by the user. Analyzing Eq. (1) and its restriction given by Eq. (2), a typical problem of convex optimization is formulated. Thus, the Lagrange function is used to solve this optimization problem:

$$L(w, b, e, \alpha) = \frac{1}{2}\|w\|^2 + \gamma \sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} \alpha_i (wx_i - b - e_i - y_i). \tag{3}$$

In Eq. (3), the first two parts are the cost functions as defined earlier. The third part is the Lagrange term, which is multiplied by the so-called Lagrange multipliers (α_i). Each Lagrange multiplier corresponds to a certain training point. To obtain the final LS–SVM solution, the partial first derivatives of this Lagrangian function are obtained and are set to zero. The weight coefficients w can be written as a linear combination of the Lagrange multipliers with the corresponding training objects (x_i). A set of linear equations instead of a quadratic programming problem can be obtained and is required to be solved. From the equation, the Lagrange multipliers α are calculated and α and w are put into the original regression equation. The resulting LS–SVM model can be expressed as

$$y_i = \sum_{i=1}^{n} \alpha_i k(x, x_i) + b, \tag{4}$$