2011 International Conference on Computer Communication and Management
Proc .of CSIT vol.5 (2011) © (2011) IACSIT Press, Singapore

Data Mining of Ultraviolet Spectra of Environment Pollutants

Shouxin Ren and Ling Gao

Department of Chemistry ，Inner Mongolia University ，Huhhot, Inner Mongolia, China

Abstract. This paper suggests a novel method named DOSC-WT-GRNN based on generalized regression neural network (GRNN) with direct orthogonal signal correction (DOSC) and wavelet transform (WT) as a pre-processing tool for the simultaneous spectrophotometric determination of o-nitro-aniline, m-nitro-aniline and p-nitro-aniline. DOSC was applied to remove structured noise that is unrelated to the concentration variables. Wavelet representations of signals provide a local time–frequency description, thus in the wavelet domain, the quality of noise removal can be improved. GRNN was applied for overcoming the convergence problem met in back propagation training and facilitating nonlinear calculation. In this case, by optimization, the number of DOSC components, tolerance factor, wavelet function, decomposition level and the width (σ) of GRNN for DOSC-WT-GRNN were selected as 1, 0.001，Coiflet 1, 5 and 0.4 respectively. The relative standard errors of prediction (RSEP) for all components with DOSC-WT-GRNN, WT-GRNN and GRNN were 4.37%, 4.93% and 6.56% respectively. The proposed method has been successfully applied to analyze overlapping spectra and was proven to be better than other techniques.

Keywords: data mining, generalized regression, neural network, direct orthogonal signal correction, wavelet transform

1. Introduction

The main drawback of ultraviolet-visible (UV-VIS) is its poor selectivity because in many cases UV-VIS spectra display strong overlaps, especially in complex organic samples. Those who are working in the new and exciting field of data mining [1-3], are actively seeking solution for the above mentioned problems. By using data mining, valuable knowledge can be extracted from huge data sets, the systematic relationships between variables can be determined, and effective predictions can made. Data mining is a cross-subject discipline and seems to be one of the most intriguing fields in information technology. Data mining provides effective ways of analyzing data by first reducing or eliminating noise and then reducing dimensions of the data set, eliminating collinearity and redundancy, transforming the data, extracting feature values to obtain valuable information, and constructing data models that having intelligent methods for making predictions. Some of the data mining methods in chemometrics such as neural computation, domain transformation, machine learning, data fusion, and data analysis, have been proven to be extremely useful [4]. Artificial neural network (ANN) is a form of artificial intelligence that mathematically simulates biological nervous systems. Multilayer feedforward networks (MLFNs), which are trained with back-propagation (BP) algorithms, are the most popular type of ANN. However, the BP-MLFN method is slow, exhibits temporal instability during training, and tends to fall into local minima. Much attention has been paid to solve this problem and to facilitate the training process into the global minimum. Generalized regression neural network (GRNN) [5], which is a kind of normalized radial basis function network, is applied in this paper. GRNN is a feedforward network based on non–linear regression theory, it does not require iterative training procedures and trains itself in a significantly shorter time than the BP based training. The GRNN training algorithm uses only one adjustable parameter i.e. the width σ of Gaussian RBF. The quality of the spectra could be improved by appropriate data pretreatment and feature extraction. In order to eliminate noise and irrelevant information, direct orthogonal signal correction (DOSC) and wavelet transform (WT) denoising method were used as a preprocessing steps to remove noise and irrelevant information. WT represents relatively recent mathematical developments, and can offer a successful time-frequency signal for enhanced information localization [6]. These characteristics of WT make it possible to perform data reduction, feature extraction, and denoising [7]. In order to avoid removing relevant information for prediction, Wold and

coworkers developed a novel pre-processing technique for raw data called orthogonal signal correction (OSC) [8]. The goal of this algorithm is to discard information in the response matrix D, which is mathematically orthogonal and unrelated to concentration matrix C. Westerhuis and coworkers introduced an intriguing OSC method named DOSC [9]. DOSC always finds components that are orthogonal to matrix C and describes the largest variation of matrix D. Many preprocessing methods, when used alone, might degrade the data in certain aspects or lose some useful information. In general, combination of multiple data processing methods has the potential to extract information more completely than by using a signal processing method. This paper proposes a new concept of data processing, which is a hybrid method combining WT with DOSC; DOSC is especially helpful for removing the systematic structured variation that is independent of concentration. The hybrid DOSC–WT method inherits properties from the original methods for eliminating not only noise and background information but also other information irrelevant to concentration. This technique combines the advantages of the two methods and can efficiently remove both the systematic structured variation that is independent of concentration and the random variation. Thus, DOSC–WT could be applied as a preprocessor in the DOSC-WT-GRNN method.

A novel approach named DOSC-WT-GRNN tested here is the combination of GRNN with DOSC and WT to eliminate noise and model linear and non-linear information and is proposed for analyzing overlapping spectra. This seems to be the first application of a combined DOSC-WT-GRNN approach to multicomponent spectrophotometric determination.

2. Theory

2.1. Generalized regression neural network

GRNN is a kind of radial basis function network. Each GRNN consists of four layers: the input layer, the pattern layer, the summation layer and the output layer. The input layer does not process information; it serves only to distribute the input data among the pattern layer. The input and pattern layer are fully connected. Each node on the pattern layer represents a radial basis function. A general Gaussian kernel function is used in pattern layer:

$$H_i(X) = \exp\left(\frac{-(X-C_i)^T(X-C_i)}{2\sigma^2}\right) \tag{1}$$

where the vector X denotes the noisy input data, the vector C_i represents the training input vector for class i. The width σ of the Gaussian kernel function is a parameter controlling the smoothness properties of the function and is called smoothing factor σ. Processing for summation layer nodes is very simple. The dot product is performed between the weights W_i and the output signals $H_i(x)$ from the pattern layer nodes in such a way: $\sum_{i=1}^{N} W_i H_i(x)$ where N is the number of training pairs. The $\sum_{i=1}^{N} H_i(x)$ were calculated by a single special node in the summation layer. The summation and pattern layers are fully connected. The weights between pattern and summation layer are as same as the target vectors t_i. The summation layer consists of two types of nodes termed A and B. Node A computes the summation of each kernel function weighted by the known concentration: $\sum_{i=1}^{N} t_i H_i(x)$; while the node B simply computes the summation of the distance: $\sum_{i=1}^{N} H_i(x)$. The output layer, the final layer in GRNN, then performs the actual divisions. The output node divides B into A to provide the predicted concentration: $\hat{Y}(x) = \frac{\sum_{i=1}^{N} t_i H_i(x)}{\sum_{i=1}^{N} H_i(x)}$.

The summation layer always has exactly one more node than the output layer. The standard equation of the GRNN is given by:

$$\hat{Y}(X) = \frac{\sum_{i=1}^{N} t_i \exp\left(\frac{-(X-C_i)^T (X-C_i)}{2\sigma^2}\right)}{\sum_{i=1}^{N} \exp\left(\frac{-(X-C_i)^T (X-C_i)}{2\sigma^2}\right)} \quad (2)$$

2.2. Direct orthogonal signal correction

Westerhuis and coworkers introduced an intriguing OSC method named DOSC [9]. The DOSC algorithm is based on least squares steps to find components that are orthogonal to matrix C and account for the largest variation of D. DOSC ia a modified OSC method. Hence, like OSC, DOSC is also applied to discard structured noise that is unrelated to the concentration variables. The algorithm of DOSC was well described in reference 9 by Westerhuis.

2.3. The discrete wavelet transforms (DWT) and wavelet thresholding denoising

The discrete wavelet transform (DWT) has been recognized as a natural wavelet transform for discrete time signals. DWT can give a time-frequency analysis of signals. The fast discrete wavelet transform (FDWT) can be implemented by means of the Mallat's pyramid algorithm [6], which is more efficient than computing a full set of inner products. The theoretical background of FDWT has already been described in detail by the authors of this paper [7]. Original signals are always contaminated by noise. The main steps of signal denoising by DWT are:

1. Decompose measured data on a selected wavelet basis
2. Thresholding
3. Reconstruct the signals

According to these algorithms, three programs, PGRNN, PWTGRNN and PDOSCWTGRNN were designed to perform GRNN, WT-GRNN and DOSC-WT-GRNN calculations.

3. Experimental

A Shimadzu UV–240 spectrophotometer with optional model OPI–2 was used for all experiments. A Lenovo Pentium 4 microcomputer was used for all calculations. A series of mixed standard solutions containing various ratios of the o-nitro-aniline, m-nitro-aniline and p-nitro-aniline was prepared in 25 mL standard flasks, then diluted with distilled water to the mark and mixed well. Cuvettes with a path length of 1cm were used and the reagent blank absorbance was subtracted. Spectra were measured between 310 nm and 500 nm at 1 nm intervals. An absorption matrix D was built up from these data. According to the same procedures an absorption matrix for unknown mixtures D_u was built up.

4. Results and Discussion

4.1. Wavelet transform and wavelet denoising

Wavelets are localized in time and frequency, and possess the properties of multiresolution analysis. Here we selected the mean spectrum of D matrix as the original signals. A series of a^j and d^j was obtained for the mean spectrum. Approximation, a^1 and a^2, are similar to the original signal. However, when a^1 was converted to a^2, the lost information was stored in the detail signal d^2. In other words, the difference in the information contained in the approximation a^{j-1} and a^j was contained in the detail d^j. Thus, the approximation signal concentrates most energy of the source signal and has large magnitude coefficients located in low frequency parts, while the detail signal represents the change in the source signal and has small magnitude coefficients with high frequency. Therefore, most of approximation signals are likely larger than threshold, that is the signal-to- noise ratio in the approximations is usually much larger than that in the details. WT

allows the original signal f to represent series of coefficients of specified energy such as $f = a^5 + d^5 + d^4 + d^3 + d^2 + d^1$ when decomposition level equals 5. The noise can be easily reduced by removing all non-significant detail signals without substantially affecting the information content.

Eliminating noise with WT is usually done by thresholding, i.e omitting some coefficients with small magnitude, which contribute little to the total energy of the signal. The HYBRID soft thresholding method is applied in this case to select suitable coefficients in the wavelet domain for denoising. In the spectrophotometric measurements, the analytical signals usually center in low frequency parts, whereas the noise concentrates in high frequency parts. WT denoising aims to extract the desired signal from a complex instrument output, where the signal is present along with noise.

4.2. DOSC-WT-GRNN

The success at obtaining a reliable result by the DOSC-WT-GRNN method depends strongly on the judicious choice of relative parameters. Five parameters were optimized in the DOSC-WT-GRNN method: wavelet function, decomposition level (L), the number of DOSC factors, the tolerance factor and the width σ of GRNN. Each of the wavelet functions has different characteristics. In this case, the wavelet functions tested were Coiflet 1, 2...5, Daubechies 4, 6...20, and Symmlet 4, 5...8. The absolute and relative standard errors of prediction (SEP and RSEP) of total compounds [14] were computed. Computational results demonstrated that the Coiflet 1 outperformed others, so Coiflet 1 was selected in the paper. The efforts on optimizing the wavelet function and decomposition level are important for best representation of the original signal without significant loss of information. By optimization as mentioned above, Coiflet 1, L = 5, the number of DOSC components = 1, tolerance factor = 0.001 and the width σ of RBF = 0.4 were selected as optimal parameters. A training set of 16 samples formed by the mixture of o-nitro-aniline, m-nitro-aniline and p-nitro-aniline was designed according to a four-level orthogonal array design with the L_{16} (4^5) matrix. Experimental data obtained from the training set were arranged in matrix D.

In the program PDOSCWTGRNN, DOSC-WT operation was initially performed and one can treat each spectrum for a given mixture. Therefore, using the same method, each row vector of matrix D and D_u was corrected by DOSC-WT to remove structured noise not related to concentration, denoised by thresholding operation in wavelet domain, and then reconstructed by applying inverse FDWT operation. The reconstructed spectra can then serve as the input of GRNN, which can efficiently perform multivariate calibration. Lastly, the concentrations of o-nitro-aniline, m-nitro-aniline and p-nitro-aniline for a test set were calculated.

4.3. A comparison of BP-MLFN，PLS，GRNN, WTGRNN and DOSCWTGRNN

In order to evaluate the DOSCWTGRNN method, five methods were tested in this study with a set of synthetic unknown samples. The RSEP for the five methods (DOSC-WT-GRNN, WT-GRNN, GRNN, PLS and BP-MLFN) are given in Table 1. The RSEP values for all compounds using DOSC-WT-GRNN, WT-GRNN, GRNN, PLS and BP-MLFN were 4.37%, 4.93% ,6.56%, 6.74% and 25.6%, respectively. The results demonstrate that the DOSC-WT-GRNN method is a successful and promising technique and generates better results than the other methods.

5. Conclusion

A method named DOSC-WT-GRNN was developed for multicomponent spectrophotometric determination. This approach combines DOSC, WT and GRNN to enhance the ability of eliminating noise and unrelated information as well as the quality of regression method. Experimental results demonstrated the DOSC-WT-GRNN approach to be successful and more satisfying results were obtained comparing to the other methods.

TABLE 1 RSEP VALUES FOR ORGANIC COMPOUNDS SYSTEM BY FIVE METHODS

Method	RSEP (%)			
	I	II	III	Total compounds
DOSC-WT-GRNN	7.68	5.17	3.52	4.37
WT-GRNN	9.84	6.06	3.59	4.93
GRNN	9.89	6.74	6.15	6.56
PLS	2.45	1.72	8.31	6.74
BP-MLFN	32.0	35.4	18.5	25.6

I: p-nitro-aniline, II: o-nitro-aniline, III: m-nitro-aniline

Acknowledgmentsa

The authors would like to thank National Natural Science Foundation of China (21067006 and 60762003) and Natural Science Foundation of Inner Mongolia (2009MS0209) for financial support of this project.

References

[1] L. Mutihac, and R. Mutihac, "Mining in chemometrics," Anal. Chim. Acta, 2008, **612**:1-18.

[2] M. Daszykowski, B. Walczak, and D. L. Massart "Projection methods in chemistry," Chemometr. Intell. Lab. Syst., 2003, **65**: 97-112.

[3] J. Han, and M. kamber, Data Mining: Concepts and Techniques, 2nd ed., Elsevier Pte. Ltd.: Singapore, 2006.

[4] S. X. Ren, and L. Gao, "Resolve of overlapping voltammetric signals in using a wavelet packet transform based Elman recurrent neural network," J. Electroanal. Chem., 2002, **586**(1): 23-30.

[5] D. F. Specht, "A general regression neural network," IEEE Transactions on Neural Network, 1991, **2** : 568-576.

[6] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," IEEE Trans. Inform. Theory, 1992, **38** (2): 617-643.

[7] S. X. Ren, and L. Gao, "Simultaneous quantitative analysis of overlapping spectrophotometric signals using wavelet multiresolution analysis and partial least squares," Talanta, 2000 , **50**(6):1163-1173.

[8] S. Wold, H. Antti, F. Lindgren, and J. Ohman, "Orthogonal signal correction of near-infrared spectra," Chemometr. Intell. Lab. Syst., 1998, **44**:175-185.

[9] J. A. Westerhuis, S. de Jong, and A. G. Smilde, "Direct orthogonal signal correction," Chemometr. Intell. Lab. Syst., 2001, **56**:13-25.