Mixture gas component concentration analysis based on support vector machine and infrared spectrum

Peng Bai (白 鹏)[1,2] and Junhua Liu (刘君华)[1]

[1] *School of Electrical Engineering, Xi'an Jiaotong University, Xi'an 710049*

[2] *Institute of Science, Air Force Engineering University, Xi'an 710051*

Received September 27, 2005

A novel quantitative analysis method of multi-component mixture gas concentration based on support vector machine (SVM) and spectroscopy is proposed. Through transformation of the kernel function, the seriously overlapped and nonlinear spectrum data are transformed in high-dimensional space, but the high-dimensional data can be processed in the original space. Some factors, such as kernel function, range of the wavelength, and penalty coefficient, are discussed. This method is applied to the quantitative analysis of natural gas components concentration, and the component concentration maximal deviation is 2.28%.

OCIS codes: 300.6340, 200.4740, 070.4790.

Infrared spectroscopy is a non-destructive real time analytical method, preferring to online monitoring for the properties of mixture gas. The classical spectrum quantitative analysis methods include K-matrix and P-matrix, which are useful for mixture gas that has fewer components and non-overlapping spectrum characteristic absorption bands[1]. The quantitative analysis methods of multi-component mixture gas concentration including multi-linear regression (MLR) and partial least square (PLS) regression[2] proposed in 1970s are effective in dealing with multi-component mixture gas whose component gas spectrum characteristic absorption bands do not overlap. However, the performances of all these methods are unsatisfactory when used in dealing with natural gas with components like CH_4, C_2H_6, C_3H_8, C_4H_{10} (isobutane), and C_4H_{10} (normal butane) whose spectrum characteristic absorption bands seriously overlap.

Owing to its strong approaching ability, artificial neural network (ANN) has been introduced to the quantitative analysis of multi-component mixture gas[3−5]. But, the practical application of ANN is limited by two factors. First, the more the training samples we provide, the better the ANN performs. In fact, the number of training samples is fairly limited in practical applied work. Second, if the dimension of input spectrum data is set too large, the computation quantity would multiply geometrically, thus the dimension of the spectrum data has to be lowered.

Support vector machine (SVM) is a novel machine learning method based on the statistical learning theory (SLT) presented by Vapnik[6,7]. On the basis of structural risk minimization (SRM) principle, SLT minimizes both the sample error and the structural risk, thus improving the generalization ability of the model and providing a new method for the quantitative analysis of multi-component mixture gas concentration. In order to deal with the serious overlapping of multi-component mixture gas characteristic absorption spectrum bands, the kernel function in SVM is used to map the serious overlapping absorption spectrum bands into high-dimensional space, after transformation the high-dimensional data can be processed in the original space. This is the key idea of this paper.

Originating from the problem of two classes in SLT, the basic principle of SVM finds an optimal hyperplane to make the margin between it and the samples on both sides maximum, so SVM can be applied in regression analysis[8−10]. There is only one kind of samples in SVM regression analysis and the optimal hyperplane is not to separate the two kinds of samples, but to make the margin between all samples and optimal hyperplane minimum. Under similar thought, the quantitative analysis method of the multi-component mixture gas concentration based on SVM makes use of the acquired multi-dimensional spectrum data that are mapped into high-dimensional space through nonlinear mapping function ϕ, and regression analysis can be performed in the high-dimensional space. Finally, a connection between multi-dimensional spectrum and component concentration of mixture gas is established by function ϕ.

Because it is difficult to express the connection between spectrum and component concentration of mixture gas by function, the SVM regression analysis is defined as follows. The sample set is a nonlinear set containing n data points $\{\mathbf{x}_i, y_i\}$, $i = 1, \cdots, n$, with input $\mathbf{x}_i \in \mathbf{R}^d$ and output $y_i \in \mathbf{R}$, the regression algorithm is

$$f(x) = \boldsymbol{\omega} \cdot \phi(\mathbf{x}) + b, \tag{1}$$

where $\phi(\mathbf{x})$ maps the input data into a higher dimensional space, $\boldsymbol{\omega}$ is a vector in higher dimensional space, b is the bias. $\boldsymbol{\omega}$ and b are obtained by solving an optimization problem

$$\min_{\omega,\xi} \frac{1}{2} \|\boldsymbol{\omega}\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{2}$$

subject to

$$\begin{aligned} & y_i - (\boldsymbol{\omega} \cdot \phi(\mathbf{x}_i) + b) \le \varepsilon + \xi_i, \\ & (\boldsymbol{\omega} \cdot \phi(\mathbf{x}_i) + b) - y_i \le \varepsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \ge 0, \qquad i = 1, \cdots, n, \end{aligned} \tag{3}$$

where C is the user-defined constant, which balances the algorithm complexity and approximation accuracy. The

slack variables ξ_i and ξ_i^* correspond to the size of this excessive deviation for positive and negative deviations. The optimization criterion penalizes data points whose y-values differ from $f(\mathbf{x})$ by more than ε.

This in turn leads to the dual optimization problem

$$\max \boldsymbol{\omega}(\alpha_i, \alpha_i^*) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) y_i - \frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle - \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) \varepsilon \tag{4}$$

subject to

$$\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0; \quad \alpha_i, \alpha_i^* \in [0, C].$$

In Eq. (4), α_i and α_i^* are Lagrange multipliers. $\langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle$ can be replaced with $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle$, and $K(\mathbf{x}_i, \mathbf{x})$ is named kernel function. Kernel functions enable $\langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle$ to be performed in high dimensional feature space by using low dimensional space data input without the transformation ϕ. Kernel functions must satisfy Mercer's condition that corresponds to the inner product of some feature space. The radial basis function (RBF), linear function, and polynomial function are commonly used for regression. Equation (1) can then be written as

$$f(\mathbf{x}) = \boldsymbol{\omega} \cdot \phi(\mathbf{x}) + b = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b. \tag{5}$$

For the variable b, it can be computed by applying Karush-Kuhn-Tucker (KKT) conditions which, in this case, imply that the product of the Lagrange multipliers and constrains has to equal to zero, thus b can be computed as

$$\begin{cases} b = y_i - \varepsilon - \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}), \alpha_i \in [0, C] \\ b = y_j + \varepsilon - \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}), \alpha_i^* \in [0, C] \end{cases}. \tag{6}$$

The structure of SVM regression algorithm is shown in Fig. 1. The input of SVM regression algorithm is spectrum data $\mathbf{x}_i = \{x_1, \cdots, x_l\}$, α and b are the solutions to Eq. (5); $K(\mathbf{x}_i, \mathbf{x})$ is the kernel function; and output is

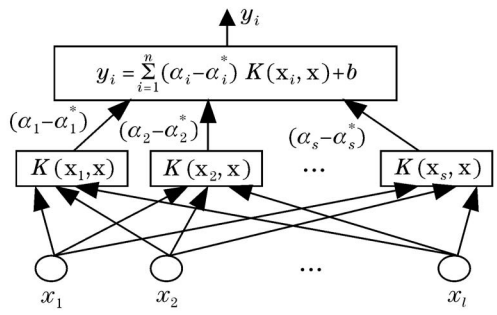

Fig. 1. The structure of SVM regression algorithm.

$y_i = \{y_1, y_2, ..., y_m\}$, the gas concentration of m components.

The flow chart of method based on SVM regression algorithm is shown in Fig. 2. Firstly, determine the parameters of SVM regression algorithm; and then take the multi-component mixture gas sample as the algorithm input and train the SVM regression algorithm. The concentration of the multi-component mixture gas sample is already known. After training, the support vector and weight value can be obtained to predict the unknown multi-component mixture gas. Finally the respective concentration of each component is the output.

The experiment process is as follows. 1) Prepare mixture gas sample of known component concentration, 2) use spectrometer to scan and obtain spectrum data, 3) unitarily process the data, 4) train and test the sample, 5) determine the parameters of SVM regression algorithm, 6) train the SVM regression algorithm, 7) test the SVM regression algorithm.

The experiment device consists of the automatic gas sampler, connection pipeline, spectrometer, data transmission cable, and computer. The spectrum of multi-component mixture gas scanned by the device mentioned above is shown in Fig. 3.

The tested multi-component mixture gas is natural gas consisting of CH_4, C_2H_6, C_3H_8, C_4H_{10}* (isobutane), and and C_4H_{10}** (butane). The characteristic absorption spectrum band of each component is listed in Table 1.

From Table 1 we can see that the interval between the primary and the secondary characteristic absorption spectrum bands of each component is very small and seriously overlapped, which makes it difficult to carry out further quantitative analysis.

The input of SVM regression algorithm is standardization spectrum data $\mathbf{x}_i$, whose wave number is in the range of 4000—400 cm^{-1}. The number of spectrum data is 2070. The output $y_i = \{y_1, y_2, \cdots, y_m\}$ is the concentration of m components, y_i is in correspondence with $\mathbf{x}_i$.

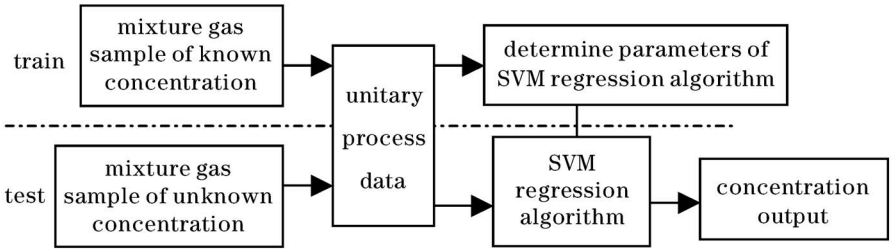

Fig. 2. The flow chart of method based on SVM regression algorithm.

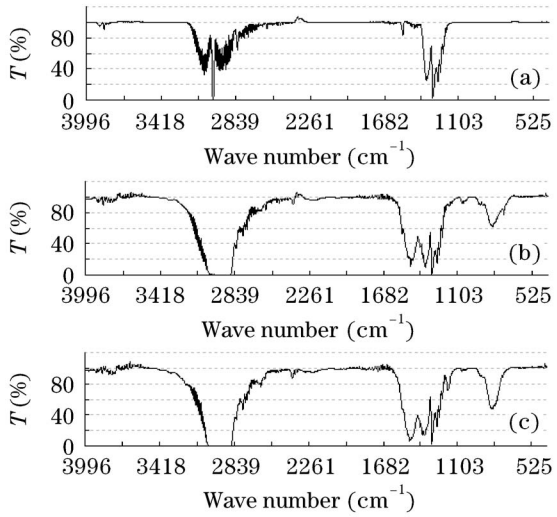

Fig. 3. Five component multi-dimensional spectra of natural gas scanned by spectrometer. (a) Volume fraction of CH_4 is 100%; (b) volume fractions of CH_4, C_2H_6, C_3H_8, $C_4H_{10}{}^*$, and $C_4H_{10}{}^{**}$ are 75%, 15%, 10%, 0%, and 0%, respectively; (c) volume fractions of CH_4, C_2H_6, C_3H_8, $C_4H_{10}{}^*$, and $C_4H_{10}{}^{**}$ are 45%, 30%, 20%, 0%, and 5%, respectively.

Table 1. The Primary and Secondary Characteristic Absorption Spectrum Bands of Natural Gas

Component Gas	Primary Wave Number (cm^{-1})	Secondary Wave Number (cm^{-1})
CH_4	3017	1305
C_2H_6	2965	1470
C_3H_8	2968	1472
$C_4H_{10}{}^*$	2967	1464
$C_4H_{10}{}^{**}$	2967	1477

The spectrum data are normalized according to

$$\mathbf{x}_l = \frac{\mathbf{x}_l}{\mathbf{x}_{\rm av}}, \quad \mathbf{x}_{\rm av} = \frac{1}{l}\sum_{k=1}^{l}\mathbf{x}_k. \tag{7}$$

As a result, calculations will be more concise and training rate of SVM regression algorithm is also increased.

In experiment, we use the software LIBSVM[11] which is a library for SVMs. Experiment parameters are selected on basic of cross-validation technique.

Different kernel functions will lead to different results, so linear kernel function and RBF kernel function are used in experiment. The type of SVM is $\varepsilon-$SVR; the kernel function is linear kernel function; the value of C is 10; value of ε is 0.1; the number of training samples is 39; and the number of testing samples is 78. The experimental result is shown in Table 2.

In Table 2, the maximum absolute error between computation and reality concentration is 43.719%. Comparing linear kernel function with RBF kernel function, we can find that the difference between the real gas concentration and the concentration calculated by the SVM regression algorithm varies considerably. Therefore, linear kernel function is more desirable in the SVM regression algorithm discussed in this paper.

Penalty coefficient C can have significant influence on the error, the relation between penalty coefficient C and the error are shown in Fig. 4. From it, we can find that the best penalty coefficient C is 25. In experiment, we also find that different ranges of the wave number also affect experimental result as shown in Table 3. The best range of the wave number is 1970—1000 cm^{-1}.

The SVM has demonstrated its success in regression. However, little work has been done for component concentration analysis of mixture gas. The feasibility of applying support vector regression in component concentration analysis of mixture gas is examined in this paper and three major merits of the SVM regression algorithm

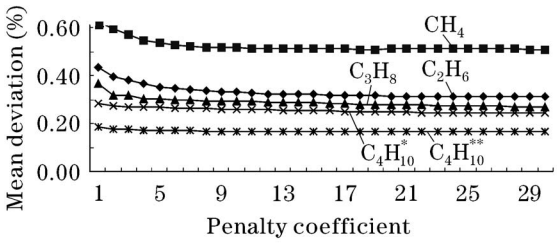

Fig. 4. The relation between penalty coefficient C and error.

Table 2. Experimental Results with Different Kernel Functions, Volume Fractions (%)

Component Gas	Concentration Interval	Concentration Range	Mean Deviation		Maximum Absolute Error	
			Linear	RBF	Linear	RBF
CH_4	5.0	100—30	0.295	8.870	1.452	43.719
C_2H_6	5.0	45—5	0.168	4.723	0.990	25.145
C_3H_8	5.0	25—0	0.136	3.569	0.907	10.477
$C_4H_{10}{}^*$	5.0	15—0	0.220	1.503	2.280	8.102
$C_4H_{10}{}^{**}$	5.0	10—0	0.169	1.147	1.303	5.106

Table 3. Experimental Results with Different Ranges of Wave Number, Volume Fractions (%)

Range of Wave Number (cm^{-1})	4000—400	3400—1000	2070—410	2360—900	3230—2600	1970—1000
Maximum Absolute Error	6.33	6.23	3.97	3.93	6.21	2.28
Mean Deviation	0.70	0.59	0.44	0.47	0.58	0.33

are concluded as follows.

1) The core of component concentration analysis of mixture gas is to set up the SVM regression algorithm between the spectrum data of mixture gas and the component concentration of mixture gas, so that SVM regression algorithm discussed in this paper can be used to deal with the serious overlapping of absorption spectrum.

2) There is no limit to the dimension of the input spectrum data when the multi-component mixture gas concentration SVM regression algorithm is used. Therefore, the problem of ANN that the computation quantity also multiplies geometrically when the input data dimension increases can be successfully solved.

3) Based on the SRM principle, the SVM regression algorithm is able to control the total expected error.

It should be noted that this study has examined only the feasibility of using SVM regression algorithm to analyze component concentration of mixture gas. Further studies on how to select SVM type, how to select kernel function and its parameters will be discussed later.

This work was supported by the National Natural Science Foundation of China under Grant No. 60276037. P. Bai's e-mail address is bai-peng410@sohu.com.

References

1. H. Yuan and W. Lu, Modern Scientific Instruments (in Chinese) (5) 6 (1998).
2. K. Hidajat and S. M. Chong, J. Near Infrared Spectrosc. **8,** 53 (2000).
3. X. Wan, F. Yang, and H. Wang, China Environmental Science (in Chinese) **23,** 110 (2003).
4. X. Sun, Y. Li, and J. Wang, Spectroscopy and Spectral Analysis (in Chinese) **23,** 739 (2003).
5. Y. Shi, Y. Lu, G. Xu, Y. Xu, Z. Xu, D. Cai, W. Lu, and J. Ma, Chin. J. Analyt. Chem. (in Chinese) **29,** 87 (2001).
6. V. N. Vapnik, *The Nature of Statistical Learning Theory* (in Chinese) X. Zhang (trans.) (Tsinghua University Press, Beijing, 2000).
7. V. N. Vapnik, IEEE Trans. Neural Networks **10,** 988 (1999).
8. N. Chen, W. Lu, C. Ye, and G. Li, Computers and Appl. Chem. (in Chinese) **19,** 691 (2002).
9. M. Ye, Chin. Opt. Lett. **3,** 205 (2005).
10. C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/∼cjlin/libsvm.