Research and application of SVM analysis model for spectral fingerprint of mixed gas

Zh MuRong¹, J Sh Guo¹, Y X Zhang², Y Jiang³, Zh G Yu⁴ and P Bai¹

¹Air Force Engineering University, Xi’an Shannxi 710051, China
²China Oil & Foodstuffs Corporation, Beijing 100020, China
³China National Petroleum Corporation, Beijing 100007, China
⁴Beijing Energy Holding Co., Ltd, Beijing 100022, China)

Email: friendmu@gmail.com

Abstract. The infrared spectrum analysis method has the characteristics of non-contact, high-speed and no consumption of sample. There is some useless interference information that is not sensitive to the analysis results in the infrared spectrum data sample. Combining support vector machine with infrared spectrum analysis, the band of spectrum data that is sensitive to gas concentration is selected by experiment. The useless, insensitive and disturbing information is removed, and the useful information is retained. And then the infrared spectrum analysis model based on support vector machine is used to complete the output of gas concentration of mixed gas components. Based on Lambert's theorem and using the known spectrum data samples as input in experiment, the sensitivity analysis of the mixed gas spectrum data samples is carried out by using the variable sliding window technology which is similar to the center wavelength and band-pass filter, and the start point and band width of the band are both variable. The band with small error is selected as the spectrum data fingerprint series to edit and splice, and then the mixed gas data sample with the fingerprint characteristics of the mixed gas component gas spectrum data is formed. On this basis, a support vector machine based infrared spectrum analysis model of mixed gases is established. The analytical model includes two processes: training and verification. The preconfigured mixed gas samples that are known concentration is fingerprint edited and processed after scanning by the spectrometer. Then the gas samples are formed the spectral data samples which can be considered as the input samples of the analytical model. The model is trained to determine the support vector and corresponding weight of the analytical model based on the spectral data samples; The fingerprint characteristic data samples of the unknown-concentration mixed gas spectrum are tested by the analytical model that has determined parameters. Then, the gas concentration of the mixed gas components is obtained. The aspects of spectral data’s fingerprint feature extraction and the selection of the model’s main parameters are researched, through experiment, the methods and processes of fingerprint feature editing and splicing extraction of spectral data are studied, and the influence of the model’s parameters on the analysis results is optimized. The experimental results show that under the condition of optimizing the model’s parameters, selecting the fingerprint features of the spectral data reasonably and constructing a new spectral data sample with the fingerprint features of the mixed gas component can help eliminate the measurement cross sensitivity caused by the cross...
overlap of the absorption spectra of the mixed gas component, and improve the calculation efficiency and the analysis accuracy, which has theoretical and practical value.

1. Introduction.
Infrared spectroscopic analysis has the advantages of non-contact, fast speed and no sample consumption, and is widely used in physics, astronomy, meteorology, remote sensing, biology, medicine, oil exploration and other fields [1]. The conventional linear mixed gas infrared spectroscopy models are K-matrix method, P-matrix method, Least Squares Regression (LSR), etc. [2], based on the Lambert-Beer absorption law, for single-component or two-component gas mixtures with few data points or no overlapping characteristic absorption spectral lines; the nonlinear analysis models are Multiple Linear Regression (MLR), Principal Component Analysis (PCA), Partial Least Squares (PLS) regression, etc., for two-component gas mixtures with many spectral data points or overlapping characteristic absorption spectral lines of component gases [3]. With the development of new technologies, new information processing methods such as Artificial Neural Network [4] (ANN), Support Vector Machines [5], wavelet analysis [6] and so on have started to intervene, opening up new paths for spectral analysis, and how to improve the ability of spectral models to identify mixed gas features is the key to achieve mixed gas component concentration detection.

Support Vector Machine (SVM) is a new machine learning method based on Statistical Learning Theory (SLT) [7], which has been applied and achieved some achievements in infrared spectral mixed gas analysis [8], atmospheric pollutant concentration calculation and other fields [9].

Since the spectral data samples contain some useless, insensitive and interfering information, using the good generalization ability of SVM in the case of small samples, the SVM is combined with infrared spectroscopy to select the spectral data bands sensitive to the gas concentration by experimental methods, remove the above useless information and retain useful information. Therefore, how to select the spectral data fingerprint characteristics of the mixed gas and perform the optimization of the SVM analysis model is the focus of the research in this paper.

2. Fingerprinting principle for spectral data
Based on the introduction of the concept of spectral data fingerprinting, the editorial splicing feature selection of spectral data fingerprinting, and the construction of SVM spectral analysis model, then the mixed gas spectral analysis, based on the experimental data, focus on the mixed gas spectral data fingerprinting feature selection process and analysis model construction.

2.1. Fingerprinting of spectral data
Each gas has its own unique absorption spectrum, and the spectral data bands that reflect the concentration information of the contained gas are selected as the spectral data fingerprint for that gas, as shown in figure 1.
The selected spectral data fingerprints are edited and stitched to form a new spectral data sample, which is then entered into the SVM spectral analysis model for analysis.

2.2. Spectral data fingerprint feature editing splicing and extraction

(1) Extraction of spectral data fingerprint characterization. Though an experimental approach, a sample of mixed gas spectral data is sensitively analyzed using a "variable slide window" technique with the known spectral data samples, and the smaller error bands selected as spectral data fingerprint series in preparation for subsequent analytical model calculations. "Variable sliding window" technology refers to a band-pass filter that selects a band with a variable starting point and band width, similar to a band-pass filter with a variable center wavelength and bandwidth. The spectral data fingerprint bands are different for each gas, for example, the main spectral data fingerprint bands for methane in hydrocarbon-containing mixtures are around 3017 cm\(^{-1}\) and 1305 cm\(^{-1}\), and the width of the sliding window is best chosen from the experimental data.

(2) Model construction. The idea of this paper is to build a model for each component gas, which can improve the analysis accuracy. For example, the experiment was used for a hydrocarbon-containing mixture of five gases, such as methane, ethane, propane, butane and n-butane gas, which require five analytical models, each with different parameters.

(3) The influence of parameters on the results of the analysis. Analyze the impact of spectral data samples and model parameters after fingerprinting and splicing of the spectral data on the analysis results.

2.3. SVM Analysis Model

The data samples used for the experiment, the sample is \(\{(x_i, y_i), i = 1, ..., n\}\), \(x_i \in \mathbb{R}^d\) is i sample,
\( \mathbf{x}_i = (x_1, x_2, \ldots, x_i) \) is the spectral data for the \( l \) sample mixed gas, and \( K(s, \mathbf{x}) \) is the weights of the \( s \) support vectors, \( K \) is the kernel function, and \( \mathbf{y}_i = (y_1, y_2, \ldots, y_m) \) is the \( m \) mixed gas composition gas concentration, corresponding to the structure of the SVM analysis model [7] is shown in figure 2.

Figure 2. Schematic diagram of SVM analysis model

In the figure, the input is \( \mathbf{x}_i = (x_1, x_2, \ldots, x_i) \); the output is \( \mathbf{y}_i = (y_1, y_2, \ldots, y_m) \), the corresponding component gas concentration.

The SVM analysis model analysis process consists of two parts, training and testing process, as shown in figure 3.

Figure 3. Flow chart of SVM analysis model

The model is trained to determine the SVM support vector and corresponding weights and other parameters. The SVM analysis model is used to examine the spectral data samples of unknown
concentrations of mixed gases to obtain the mixed gas component gas concentration output.

2.4. Sample production of mixed gas spectral data for experiments
The infrared spectrometer used in the experiment is BRUKER TENSOR-27. 200 mixed gas samples containing hydrocarbons were prepared from five gases: pure methane, ethane, propane, butane, n-butane gas, etc. The wavelength range of the spectrometer is 4000cm⁻¹-400cm⁻¹, the scanning interval is 12nm, as shown in table 1.

| Concentration range and interval | Methane | Ethane | Propane | Butane | N-butane |
|---------------------------------|---------|--------|---------|--------|----------|
| Minimum concentration /%        | 20.0    | 5.0    | 3.0     | 2.5    | 2.5      |
| Maximum concentration /%        | 80.0    | 35.0   | 17.0    | 8.0    | 8.0      |
| Concentration interval /%       | 1.0     | 1.0    | 1.0     | 1.0    | 1.0      |

3. Results and discussion
The effects of spectral data editing and splicing feature extraction on the analysis results are investigated based on the determination of the SVM analysis model kernel functions and penalty factors.

3.1. Fingerprinting band editing and splicing of spectral data
A group of gases in a mixture absorbs infrared light strongly in multiple bands, i.e., spectral data fingerprinting bands.

The 200 spectral data are used for mixed gas analysis, and the error is too large by moving the variable width selection window to calculate the analysis results, indicating that the spectral analysis band is not suitable for mixed gas SVM analysis, the calculation of the analysis results of the smallest error of one or several spectral data, editing and splicing processing, as the input of the SVM analysis model, the kernel function of the SVM analysis model to select a linear kernel function, the penalty factor takes the value $C = 30$.

Table 2 shows the spectral data samples of methane in hydrocarbon-containing gas mixtures, selected different spectral data analysis bands to form characteristic spectral data samples, and the error situation of the analysis using the SVM analysis model.
Since the gas mixture contains hydrocarbon gases, and the gases contained are methane, ethane, propane, butane and n-butane gases, the primary and secondary characteristic absorption spectral lines of these gases overlap. It was experimentally verified that the spectral data fingerprint of the remaining four gases, ethane, propane, butane and n-butane, are similar to methane, so the spectral data between 2070cm\(^{-1}\)-410cm\(^{-1}\) and 2360cm\(^{-1}\)-900cm\(^{-1}\) were selected as the spectral data characteristic fingerprint, which has better effect.

During the experiment, the following patterns were also found: (1) without any editing and splicing, the analysis results were poorer using all the spectral data, as shown in columns 1 and 2 of Table 2; (2) using the selected fingerprint characteristics was too detailed and the number of spliced spectral segments was too large, which caused the accuracy of the analysis results to decrease, which can be called "over-editing and splicing", so two spectral fingerprint analysis bands were chosen.

### 3.2. SVM analysis model parameters

SVM analysis model kernel function: The kernel functions of the SVM analysis model are Linear kernel function, Polynomial kernel function and RBF kernel function, which have an impact on generalization ability [8].

Penalty Factor C: The degree of penalty for the sample beyond the error \(\varepsilon\) is the equilibrium parameter of the SVM's generalization ability. A large value of \(C\) tends to be the least empirical and ignores the consideration of structural complexity; conversely, the complexity of the problem is considered and the role of empirical data is ignored.

Parameter optimization process: (1) build the SVM model, select different kernel functions for training and testing, change the penalty factor \(C\), and determine the relationship between editing and splicing feature extraction spectral data sample data and error; (2) select the best kernel function and penalty factor \(C\) according to the above results.

Combining the above process, 200 samples of spectral data were configured for learning and training, the spectral data between 2070cm\(^{-1}\)-410cm\(^{-1}\) and 2360cm\(^{-1}\)-900cm\(^{-1}\) were selected as spectral data feature fingerprint data samples, 100 samples were used for learning and the other 100 samples were used for testing, and the analysis results are shown in table 3.

| Kernel function | Error | Methane | Ethane | Propane | Isobutane | N-butane |
|-----------------|-------|---------|--------|---------|-----------|---------|
| Linear kernel function | Minimum absolute/\% | 0.004 | 0.002 | 0.010 | 0.002 | 0.003 |
|                  | Maximum absolute/\% | 0.381 | 0.314 | 0.283 | 0.192 | 0.339 |
|                  | Average absolute/\% | 0.040 | 0.020 | 0.031 | 0.018 | 0.016 |
| Polynomial kernel function | Minimum absolute/\% | 0.008 | 0.007 | 0.005 | 0.003 | 0.007 |
|                  | Maximum absolute   | 0.574 | 0.202 | 0.134 | 0.113 | 0.257 |
When different kernel functions were selected, the effect of the penalty factor C on the results of the analysis was reduced when the value of the penalty factor C was greater than 30, and according to the data in Table 3, the effect of the linear kernel function was better.

4. Conclusion
The effect of the SVM analysis model parameters on the analysis results was investigated using the experimental data, and the selection results after the optimization of the analysis model parameters are shown in table 4.

| Processing links      | Parameters          | Select or take value |
|-----------------------|---------------------|----------------------|
| SVM analysis model    | Kernel function     | Linear kernel function |
|                       | Penalty factor C    | C=30                 |
| Spectral data samples | Spectral bands      | 2070cm⁻¹-410cm⁻¹     |

There is currently no theoretically valid method for the selection and optimization of the above SVM analysis model parameters. The analysis concluded that the selection of spectral analysis bands prior to the establishment of the SVM correction model is a very useful technical means to improve the operational efficiency and provides a basis for subsequent practical applications.

How to improve the ability of the system to identify the features of the mixture, how to extract the fingerprint features of the spectral data of different gases from the spectral data samples, and suppress the optical path noise are the key to the method of spectral information processing of mixed gases and are the focus of subsequent research.

Acknowledgments
This thesis was completed under a grant from the International Fund for Natural Science (IFNS) project (61502521).
References

[1] T. Davies. T 1998 Analysis 26(4) 17-9
[2] Lu, W.Z.. 2007 Modern near infrared spectroscopy techniques (2nd edition) Beijing: Sinopec Press
[3] Xu L, Shao Xeguang. 2004 Chemometrics methods (2nd edition). Beijing: Science Press
[4] Hu, Yuan-Bin et al. 2009 Neural networks Xi'an: Xi'an Electronic Technology University Press,
[5] V.N. Vapnik 1999. IEEE Trans on Neural Network 10(5) 988
[6] Gao-You Tian, Hong-Fu Yuan, Hui-Ying Liu, et al. 2003 Spectroscopy and Spectral Analysis 23(6) 1111-4
[7] V.N. Vapnik. 1995 The Nature of Statistical Learning New York: Springer
[8] Zhang L. D., Su S.G., Wang L.S., et al. 2005 Spectroscopy and spectral analysis 25(1) 33-5
[9] Ma, Xiaoguang, Hu, Fei. 2004 Progress in Natural Sciences 14(3) 349-53
[10] CHU Xiaoli, YUAN Hongfu, LU Wanzhen. 2004 Advances in Chemistry 16(4) 528-42