CHEMICAL ENGINEERING TRANSACTIONS
VOL. 59, 2017
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Zhuo Yang, Junjie Ba, Jing Pan

ISBN 978-88-95608- 49-5; **ISSN** 2283-9216
DOI: 10.3303/CET1759106

Study on Computer-assisted Infrared Spectroscopy for Identification of Chemical Structure System

Daping Deng[a], Yameng Bai[b], Xiaohong Deng[a], Xiaoyun Xie[a], Jiuming Yan[a]

[a]College of Applied Science, Jiangxi University of Science and Technology, Ganzhou 341000, China
[b]College of information engineering, Jiaozuo University, Jiaozuo 454000, China
apd21105@21.cn

In the past decades, people are trying to search the way to analyze the infrared spectra. Along with the computerization of the commercialized infrared spectroscopy, there are many computer-assisted identification methods of infrared spectroscopy. For decades, people have been exploring the empirical analysis of infrared spectroscopy. These methods can be divided into three categories: expert system; spectrum retrieval system and pattern recognition method. The most commonly used pattern identification methods are artificial neural network and partial least squares. The literature shows that the prediction accuracy of the structural fragments is not very high, and the neural network is still unstable, easy to fall into the local optimal and slow convergence and other issues. In this paper, the support vector machine is used to analyze the sub-structure of infrared spectroscopy. The vector machine is a good machine learning algorithm for small sample system. For most of the substructures, the predictive ability of support vector machines is better. The support vector machine also has the advantages of stability and fast training speed. It is a good tool for assistant analysis of infrared spectrum.

1. Introduction

Infrared test technology is gradually built and developed in 1800 by the physicist W. Herschel after found the infrared radiation (Zhang, et al., 2016). In the early 1950s, the infrared spectrometer came out, and the infrared spectroscopy was widely developed, which opened a new stage in the identification of organic structure. A wealth of infrared spectral data was accumulated till the late 1950s. The infrared spectroscopy has been the most important method to identify the organic compounds till the mid-70s (Mecozzi & Sturchio, 2017). In recent decades, the advent of Fourier transform infrared spectroscopy and the emergence of some new technologies (such as emission spectra, photoacoustic spectroscopy, color-red combination, etc.) have made the infrared spectrum more widely used. The infrared spectrum has a wide adaptability to the samples, regardless of solid, liquid or gaseous samples. In addition, infrared spectroscopy has the characteristics of fast, high sensitivity and less sample amount. Therefore, it has become the most commonly used and indispensable tool for modern structural chemistry and analytical chemistry (Allen, et al., 2016).

The wave length of infrared light covers 0.76μm~1000 μm, and the corresponding wave number is in the range of 13330~10 cm-1. Usually, the infrared region is divided into near infrared region (13330~4000 cm-1), middle infrared region (4000~650 cm-1) and far infrared region (650~10cm-1). Because the vibration frequency of most of the organic compounds is in the mid infrared region. The study on the mid infrared spectra is the most. The data collection, collation and induction of the absorption peak area has become quite perfect.

The application of infrared spectroscopy in chemistry is various. It can be used not only for basic research structure, such as determining the molecular space structure, and calculating chemical bond force constants, bond lengths and angles. It is also widely used in the qualitative and quantitative analysis of compounds and chemical reaction mechanism research. The widest use of infrared spectroscopy is the structural identification of unknown compounds.

Please cite this article as: Daping Deng, Yameng Bai, Xiaohong Deng, Xiaoyun Xie, Jiuming Yan, 2017, Study on computer-assisted infrared spectroscopy for identification of chemical structure system, Chemical Engineering Transactions, 59, 631-636 DOI:10.3303/CET1759106

The infrared spectrum is very complex. The different atomic mass of compounds, different chemical bond properties, and different order and spatial location of the atoms will cause the difference of infrared spectrum. In recent years, people have been looking for a better pattern identification method to analyze the structure of infrared spectrum. Vapnik et al proposed support vector machine (SVM) based on the Statistical Learning Theory (SLT) in 1995. According to the limited sample information, it found the best compromise between model complexity and learning ability in order to obtain the best generalization ability. In this paper, support vector machine (SVM) is used to analyze the sub-structure of infrared spectrum, and compared with back propagation artificial neural network. Support vector machine (SVM) is similar to multilayer feedforward network in form, and can also be used for pattern recognition and nonlinear regression (Nguyen, et al., 2016).

2. Application of support vector regression for carbon black process modeling

2.1 Carbon black

Carbon black is produced by the incomplete combustion or pyrolysis of hydrocarbons (solid, liquid or gaseous. Carbon black is an important reinforcing agent and filler for rubber products (mainly tires). It not only can improve the strength of the rubber products, but also can improve the technical performance of the rubber material, and can endow the products with the advantages of wear resistance, tear resistance, heat resistance, cold resistance, oil resistance, etc., and can prolong the service life of the product. About 75% of the carbon black is used to make all kinds of tires. Therefore, the production of carbon black is closely related to the development of the automobile industry. In addition, carbon black can also be used for ink, plastic and batteries and so on (Moriya, et al., 2016; Torrado et al., 2016; Torrado et al., 2016).

At present, the carbon black which the annual production capacity is 10000 tons mostly realized automatic control, such as DSC (Dynamic Stability Control) control system, being able to adjust the technological parameters on the computer. But how to adjust the technical process according to the carbon black products information is not clear, even some factories still mainly depends on the experience. This is not beneficial to improve the quality of the carbon black to better meet the development of rubber industry and other related industries. Especially after China joined WTO, foreign carbon black entered the Chinese market, such as the United States, Japan, Western Europe, they have established carbon black production base in China, which occupied the market share, and China's carbon black industry is facing strong competition and challenge. Therefore, it is necessary to introduce advanced science and technology to optimize the technological parameter which plays a decisive role in the production of carbon black, and establish a reliable prediction model between the carbon black product index and process parameters, improve the carbon black production technology, and improve the ability of producing high quality carbon black. Therefore, the significance and necessity of carbon black process modeling are summarized into three points:

(1) To change the current status that carbon black production mainly depends on the long-term accumulated experience.

(2) To meet the requirements of high quality carbon black for the development of rubber industry.

(3) To enhance the competitive ability of carbon black at home and abroad after China's accession to the WTO.

2.2 Data processing and hardware and software equipment

2.2.1 Data sources and pretreatment

Data is from the original records of carbon black production experiment workshop of carbon black industry research and design institute of China Rubber Group, a total of 112 samples. According to the experience that the operating variable participated in optimal control totally have 7 (Ullah, et al., 2016), which are natural gas flow, raw oil flow, the first chilled water flow, the second emergency water flow rate and dosage of carbon black, the dryer outlet temperature and granulating machine power. Learned from the production experience, the iodine absorption value and the DBP oil absorption value are mainly detected in the carbon black production workshop. The data was processed with standardization, and the treated variables were the same as the weight, and the mean value was 0, the variance was 1. The Cluster analysis method (CA) (Li and Liu, et al., 2016; Li and Hou et al., 2016), and Multi-discrimination vector (MDV) were used and combined with the actual production situation. 3 outliers were removed, remaining 109 samples.

2.3 Hardware and software equipment

The hardware and software environment of the experimental data processing:

Hardware: 60G disk, 1.0G Celeron CPU, 256M memory.

Software: Windows XP, Office XP, MATLAB 6.5.

2.4 Application of support vector machine (SVM) in carbon black process modeling

2.4.1 Application of support vector machine for carbon black process modeling

After 3 outliers are removed from 112 original samples, according to 10~20% of the number of prediction set samples to the number of calibration set samples, the original samples are divided into two groups: one group is the calibration set (training set), a total of 89 samples, which is used in the construction of carbon black production model; another is the prediction set, a total of 20 samples which is used for detecting model. The 20 prediction samples are randomly generated in 109 samples by using the rands function in MATLAB. The carbon black production process obtained from the previous work in our laboratory has a very strong nonlinear (Zampieri, et al., 2016). In this paper, support vector machine (SVM) is used to establish the model of carbon black process, and compared with the back propagation artificial neural network and radial basis function neural network modeling method.

When using support vector machine regression, the parameter group which has a great influence on the training result is (σ, C), the error will be increased if it is too large or too small. ε is not sensitive to the loss function. If ε is too small, it is easy to produce the phenomenon of over fitting, and if too large, it is easy to produce less fitting. C is a penalty factor, which controls the degree of penalty for misclassification samples. The greater the C, the greater the penalty for the error. σ is the width of the kernel function. According to the experience, σ and C are adjusted respectively in 0.1-512, and the optimal parameter set (σ, C) is found, which make the regression have the best prediction ability. After debugging, the prediction effect of the iodine absorption is the best when ε is 0.01, C is 274 and ε is 3. When ε is 0.01, C is 3, σ is 1.1, the prediction effect of the oil absorption is the best when ε is 0.01, C is 3 and σ is 1.1. The fitting of the network output and actual production value of the carbon black iodine absorption and oil absorption value in the training set is shown in Figure 1.

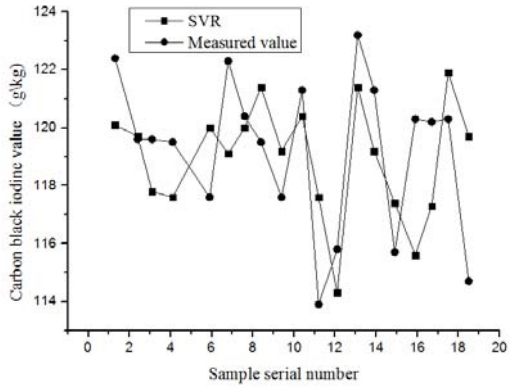

Figure 1: Fitting of the SVR prediction and actual production values in the training set

It can be seen from Figure 1 that the fitting of the predicted value and the actual production value is very good. The model is used to predict the prediction set, and the fitting of the predicted value and the actual production value is shown in Figure 2.

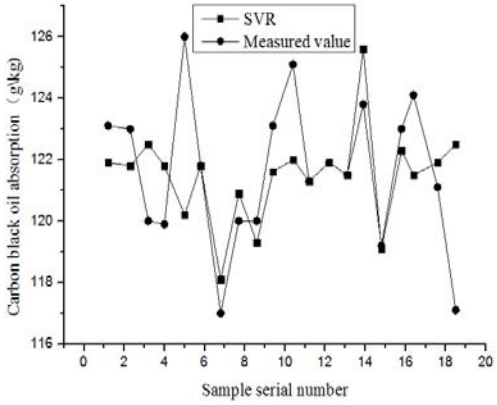

Figure 2: Fitting of SVR model prediction value and actual production value

It can be seen from Figure 2, the prediction of SVR on the carbon black iodine absorption and oil absorption value is better, and for the oil absorption value, the prediction error on the 5,11, 19 point is larger.

2.4.2 Application of neural network on carbon black process modeling

The training method of back propagation artificial neural network and radial basis function neural network is the same as the literature. The fitting of the network output and actual production value of the carbon black iodine absorption and oil absorption value in the two training sets is shown in Figure 3.

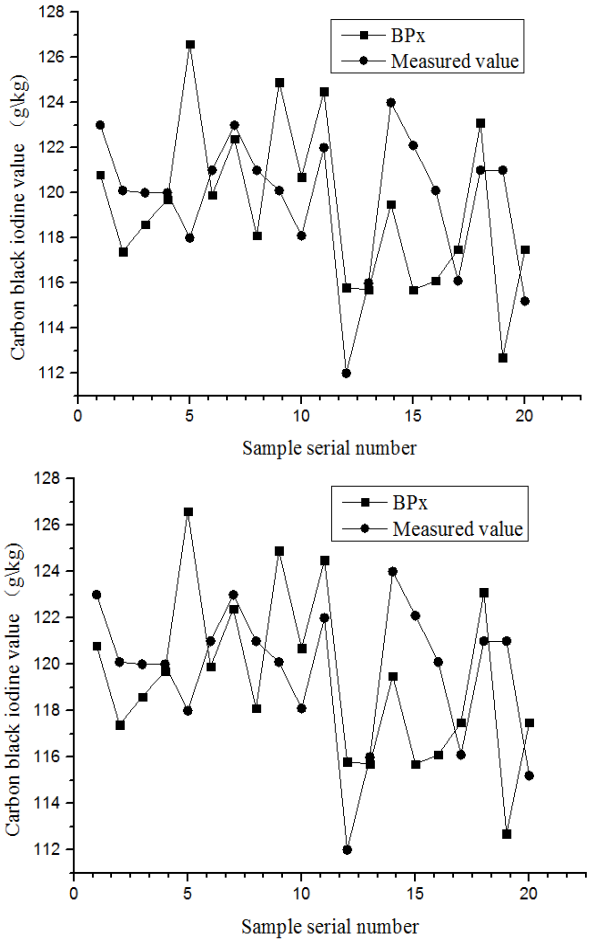

Figure 3: Fitting of forecast and actual production value of BPx and RBFN model of prediction set

As can be seen from Figure 3, the fitting of the two neural networks to the carbon black training set is also very good. Figure 4 is the the prediction value and the actual production value fitting of the prediction set of BPx and RBF to the two models.

From Figure 2 and Figure 4 it can be seen that the prediction ability of three models of carbon black oil absorption value is higher than the corresponding prediction ability of iodine absorption on carbon black. For the iodine absorption, the difference between the predictive value and the actual production value of BPx is the largest, and only a few points' prediction effect is good. The prediction of RBFN is better than BPx, but not as good as SVR. For the oil absorption value, the overall prediction effect of SVR and RBFN is better than BPx. The difference between SVR and RBFN two prediction model is not very large.

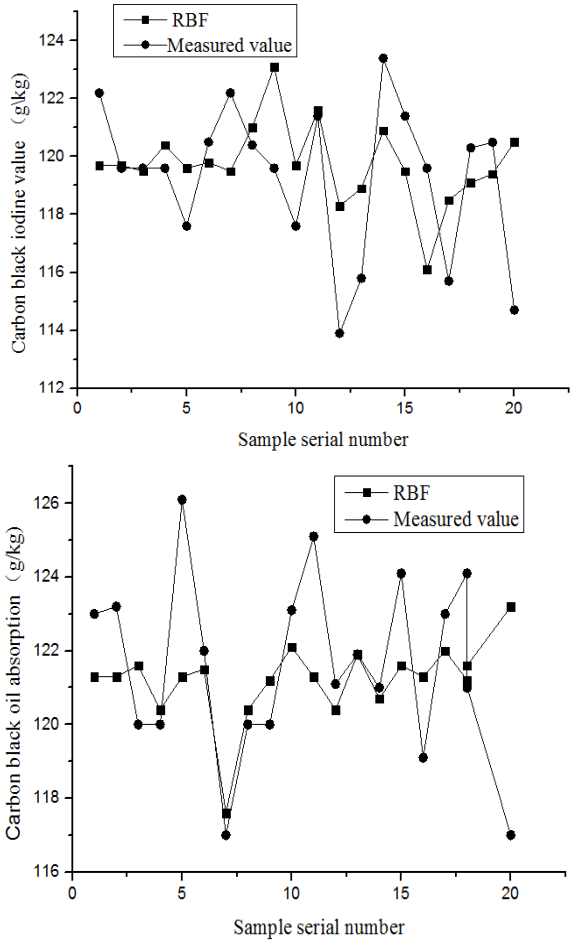

Figure 4: Fitting of prediction value and actual production value of BPx and RBFN model of prediction set

2.5 Comparison of prediction results support vector machine and neural network

In order to compare the modeling results of ANN-BPx, RBFN and SVR more intuitively, we use the average prediction error, the average relative error and the square error of the prediction set as the evaluation criteria of three indicators. The comparison results of the three models are shown in table 1 and table 2.

Table 1: Prediction results of iodine absorption value

Models	Average prediction error	Average relative error	Error sum of squares
BPx	3.05	2.54	292.99
RBFN	2.19	1.85	146.47
SVR	1.93	1.62	109.40

Table 2: Prediction results of oil absorption value

Models	Average prediction error	Average relative error	Error sum of squares
BPx	2.00	1.64	136.07
RBFN	1.68	1.38	105.30
SVR	1.60	1.31	97.79

From table 1 and table 2, relative prediction errors of iodine absorption value and oil absorption value of SVR on carbon black are 1.62% and 1.31%, the model prediction accuracy is significantly higher than that of ANN-BPx (2.54%, 1.64%), and is slightly better than that of RBFN (1.85%, 1.38%). The prediction ability of three models on the oil absorption value of carbon black are higher than the corresponding prediction ability of iodine absorption of carbon black, which is the same as the conclusion obtained in Figure 2 and Figure 3.

3. Conclusion

Support vector machine is a machine learning method of small sample theory. It can use the limited data to get the optimal solution. In this paper, it was applied to carbon black process modeling. Compared with RBFN and BPx-ANN method, the results show that the prediction accuracy of SVR models is better than that of ANN-BPx, and slightly better than the RBFN, solving the model building problems in the process of carbon black production, which has great significance for optimizing the operating conditions of carbon black production.

Reference

Allen F., Pon A., Greiner R., Wishart D., 2016, Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification. Analytical chemistry, 88(15), 7689-7697.

Li W., Liu Y., Sun H., Pan Y., Qian Z., 2016, Monitoring reduced scattering coefficient in pedicle screw insertion trajectory using near-infrared spectroscopy. Medical & biological engineering & computing, 54(10), 1533-1539.

Li Y., Hou S.H., Yao L.M., 2016, Profitability assessment using data envelopment with cluster analysis: a case for different types of gas stations, Chemical Engineering Transactions, 51, 727-732, DOI: 10.3303/CET1651122

Mecozzi M., Sturchio E., 2017, Computer Assisted Examination of Infrared and Near Infrared Spectra to Assess Structural and Molecular Changes in Biological Samples Exposed to Pollutants: A Case of Study. Journal of Imaging, 3(1), 11.

Moriya Y., Yamada T., Okuda S., Nakagawa Z., Kotera M., Tokimatsu T., Goto S., 2016, Identification of Enzyme Genes Using Chemical Structure Alignments of Substrate–Product Pairs. Journal of chemical information and modeling, 56(3), 510-516.

Nguyen S.C., Zhang Q., Manthiram K., Ye X., Lomont J. P., Harris C. B., Alivisatos A.P. (2016). Study of heat transfer dynamics from gold nanorods to the environment via time-resolved infrared spectroscopy. ACS nano, 10(2), 2144-2151.

Torrado D., Cuervo N., Pacault S., Dufour A., Glaude P., Murillo C., Dufaud O., 2016, Explosion of gas/carbon blacks nanoparticles mixtures: an approach to assess the role of soot formation, Chemical Engineering Transactions, 48, 379-384, DOI: 10.3303/CET1648064

Ullah I., Ahmad I., Nisar H., Khan S., Ullah R., Rashid R., Mahmood H., 2016, Computer assisted optical screening of human ovarian cancer using Raman spectroscopy. Photodiagnosis and photodynamic therapy, 15, 94-99.

Zampieri D., Vio L., Fermeglia M., Pricl S., Wünsch B., Schepmann D., Laurini E., 2016, Computer-assisted design, synthesis, binding and cytotoxicity assessments of new 1-(4-(aryl (methyl) amino) butyl)-heterocyclic sigma 1 ligands. European Journal of Medicinal Chemistry, 121, 712-726.

Zhang Z., Cao T., Liu H., Shu J., Li Z., 2016, September. A Computer-Assisted Learning System in the Teaching of Infrared Spectroscopy Course. In Educational Innovation through Technology (EITT), 2016 International Conference, 91-95. IEEE.