Chinese Chemical Letters Vol. 16, No. 10, pp 1354-1356, 2005
http://www.imm.ac.cn/journal/ccl.html

Substructure Prediction from Infrared Spectra by Using Support Vector Machines

Jun Hong LIU, Min Chun LU, Fu Sheng NIE, Xiao Yu FENG, Meng Long LI*

College of Chemistry, Sichuan University, Chengdu 610064

Abstract: The potential of support vector machines (SVMs) for the substructure elucidation of infrared spectra have been investigated. The trained SVMs can identify the 16 substructures with high accuracy.

Keywords: Infrared spectra, substructures, support vector machines.

In the past decades, several methods have been applied to the automatic interpretation of infrared spectra. These methods can be classified into several groups: spectral libraries search, knowledge-based systems, pattern recognition techniques. The pattern recognition techniques were widely used in the past decades, dominated by artificial neural networks (ANNs)[1-7]. However, the prediction accuracy of present substructures is not very satisfying, and only for a few substructures can exceed 90%[2,5].

In this paper, support vector machine (SVM) was proposed as a tool for the substructure elucidation of infrared spectra. The SVM solution of Vapnik is known as a very good tool for classification problems with excellent generalization ability[8-15]. In distinction to the classical neural networks, SVM always seeks global optimum and can avoid over-fitting[8]. In recent years, it has demonstrated excellent performance in a variety of pattern recognition problems, such as text classification, face detection and protein fold recognition.

A set of 823 compounds from the OMNIC Fourier transform infrared (FTIR) database was used for training and testing with respect to the presence or absence of 16 substructures. These 16 substructures were defined on the basis of infrared absorption frequencies. The training and test sets were selected simply by taking the even-numbered samples and the odd-numbered samples, and were consisted of 411 and 412 FTIR spectra, respectively. The present percentage ranges from 1.22% to 46.7% in the training set, and 1.46% to 46.6% in the test set.

The spectra ranging from 449 to 4000 cm^{-1} were divided into 307 points of equal interval. The 307 data points were used as input vector. 1 for presence and -1 for absence for substructures were encoded as output vector. When the output is 1, it means that the substructure is present, else means that the substructure is absent.

* Email: liml@scu.edu.cn

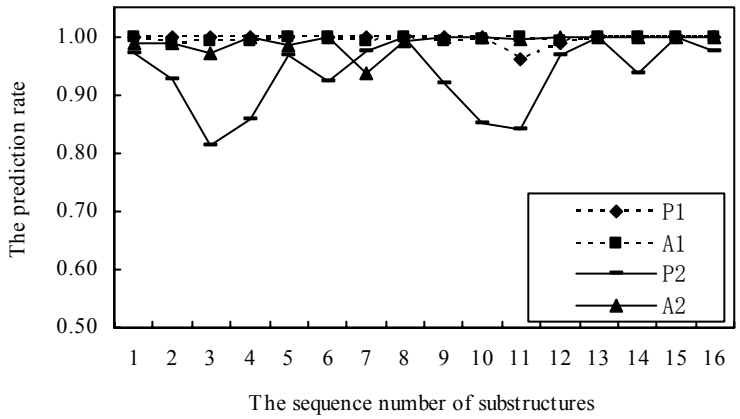
The SVMs used in this paper were trained using the sequential minimal optimization (SMO) approach[16]. The SVM algorithm was implemented in MATLAB 6.5. Radial Gaussian kernel was used in this paper[10,11]. Because both training and test sets can learn strongly toward absent groups, the prediction of absent groups did not pose problems[5]. Parameters (C and Sigma) influencing SVMs' training and prediction of substructures from infrared spectra were scrutinized for the maximal achievable present prediction rate. C is the regularization parameter and Sigma is the width parameter of the kernel. When the present prediction rate is very high but the absent prediction rate is very low, the striving for the best possible present prediction is relaxed a little in order to gain a higher accuracy of the absent prediction. For all C, Sigma pairs, the SVM was trained with training set and evaluated for test set to get the optimal prediction rate. Monitoring the best prediction rates during training allowed the SVM with the highest generalization capabilities.

The prediction ability of the SVMs was evaluated by P_f and A_f:

P_f (recognition accuracy of presence) = the number correctly classified as present / the number present; A_f (recognition accuracy of absence) = the number correctly classified as absent / the number absent.

The P_f and A_f of the 16 substructures of the training and test set are shown in **Figure 1**. We can see from **Figure 1** that, the SVMs can recognize the 16 substructures in the training set with perfect accuracy: the P_fs of 14 substructures are 1, the A_fs of the 16 substructures are all closer to 1. The trained SVMs can predict the 16 substructures in the test set with high accuracy: most of the A_fs of the test set are as good as those of the training set; 12 out of 16 P_fs are above 0.900, and the other four are still higher than 0.800. The values of P_f and A_f in training and test set also reflect SVMs' excellent generalization ability and strong robust. From the high values of P_f and A_f we can come

Figure 1 The P_f and A_f of the substructures



(Note: P1: the P_f of the training set, P2: the P_f of the test set, A1: the A_f of the training set, A2: the A_f of the test set, 1 -NH, 2 C-N, 3 C=C, 4 C≡N, 5 -OH (alcohol), 6 Ar-OH, 7 -OH (hydroxyl), 8 C_6 aromatic, 9 C-O-C, 10 C(CO)C, 11 (CO)H, 12 (CO)OH, 13 (CO)OR, 14 (CO)NH, 15 (CO)Cl, 16 (CO).)

to the conclusion that SVM approach is a powerful tool for the interpretation of infrared spectra.

The discrimination ability of SVM is correlated with the infrared absorption of substructures. As expected, substructures that show very distinctive characteristic infrared absorptions could be discriminated by the SVM quite easily (*e.g.*, esters, carboxylic acids and C_6 aromatic). Substructures without distinctive infrared absorptions are less well recognized by SVM (*e.g.*, double band).

References

1. M. E. Munk, M. S. Madison, E. W. Robb, *Mikrochim. Acta (Wien),* **1991**, *2*, 505.
2. D. Ricard, C. Cachet, D. Cabrol-Bass, T. P. Forrest, *J. Chem. Inf. Comput. Sci.,* **1993**, *33*(2), 202.
3. C. Klawun, C. L. Wilkins, *J. Chem. Inf. Comput. Sci.,* **1996**, *36*(1), 69.
4. M. L. Li, M. L. Luo, Z. L. Sun, *et al*., *Journal of Fushun Petroleum Institute*, **1996**, *13*(3), 37.
5. M. Novic, J. Zupan, *J. Chem. Inf. Comput. Sci.,* **1995**, *35*(3), 454.
6. M. Meyer, K. Meyer, H. Hobert, *Anal. Chim. Acta,* **1993**, *282*(2), 407.
7. T. Kazutoshi, M. Takatoshi, T. Tadao, *et al*., *Applied Spectroscopy,* **2001**, *55*(10), 1394.
8. S. Haykin, 2, *Neural Networks: a Comprehensive Foundation*, Prentice-Hall, New Jersey, **1999**, 318.
9. R. Burbidge, M. Trotter, B. Buxton, S. Holden, *Computers and Chemistry*, **2001**, *26*(1), 5.
10. K. Brudzewski, S. Osowski, T. Markiewicz, *Sensors and Actuators B*, **2004**, *98*(2-3), 291.
11. A. I. Belousov, S. A. Verzakov, J. Von Frese, *Chemometrics and Intelligent Laboratory Systems,* **2002**, *64*(1), 15.
12. A. I. Belousov, S. A. Verzakov, J. Von Frese, *Journal of Chemometrics*, **2002**, *16*(8-10), 482.
13. L. Hu, L. A. Qiao, Y. D. Gong, *et al.*, *Acta Biophysica Sinica*, **2001**, *17*(4), 669.
14. S. X. Du, T. J. Wu, *Control and Instruments in Chemical Industry*, **2004**, *313*(3), 54.
15. H. B. Qu, X. X. Liu, Y. Y. Cheng, *Chem. J. of Chin. Univ.*, **2004**, *25*(1), 39.
16. F. Gray William, L. Steve, *Machine Learning*, **2002**, *46*(1-3), 271.

Received 3 December, 2004