Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

Functional groups prediction from infrared spectra based on computer-assist approaches

Zhimeng Wang, Xiaoyu Feng, Junhong Liu, Minchun Lu, Menglong Li*

College of Chemistry, Sichuan University, Chengdu, Sichuan 610064, China

ARTICLE INFO

Keywords:
Spectroscopy
Support vector machines
Artificial neural networks
Infrared spectra
Functional groups prediction

ABSTRACT

The support vector machine was used as a tool for mining the structure information of infrared spectra. A total set of 823 compounds from the OMNIC Fourier transform infrared database was used for training and testing with respect to the presence or absence of 16 functional groups. The results were evaluated by a variety of indices and compared favorably with those obtained by using artificial neural networks method. The trained support vector machines can identify the 16 functional groups with an average prediction accuracy of 93.3% for the presence and 99.0% for the absence of functionalities.

1. Introduction

Approaches to structure elucidation are usually based on the interpretation of spectral data as the source of structural information, including modern infrared (IR), [1–5] ultraviolet,[6] nuclear magnetic resonance [7] and mass spectra.[8] Among various spectra from which organic chemists derive structural information, infrared spectra present the greatest challenge for automated machine interpretation. The infrared spectra of most organic compounds are quite complicated, containing a large number of absorption bands that could be attributed to many different functional groups. Manual analysis by skilled chemists is time-consuming and error-prone.[9] The neighboring molecular features and environmental conditions always have more or less influence on the functional groups. Furthermore, the presence or absence of a certain functional group is not only based on the presence or absence of a single spectra band, but also on complicated spectral regions. So the correlation between infrared spectral feature and structural feature is very sophisticated, which has been attracting experts from different fields of spectroscopy and chemometrics to develop automatic interpretation.[1–5] The automatic structure elucidation of infrared spectra generally falls into three groups: library search, knowledge-based system, or pattern recognition. Among the last group of methods, artificial neural networks (ANNs) [10–17] and partial least squares (PLS) [18–20] are most frequently used. Different PLS and ANN methods have been developed for non-linear calibration models in IR spectroscopy.[15] Although deep neural network has recently been applied to accurately identify functional groups of unknown compounds,[21] additional mass spectrometry was needed in this method. It was therefore not discussed in this study. ANNs have several major drawbacks: unsteadiness, local minima and low speed of convergence. In order to solve these problems, the support vector machine (SVM) was introduced as a tool for structure elucidation of infrared spectra. SVM is a kind of learning machine and is known as a very good tool for classification problems with excellent generalization ability and can model non-linear boundaries through the use of kernel functions.[22–27] Olivier Devos[28] proposed a grid search method to guide the regularization and kernel *meta*-parameters adjustment in the SVM model, which is indispensable for controlling risk of overfitting and the complexity of the boundary. The results of support SVMs were evaluated by a variety of indices and compared with those by ANNs.

2. Theory and algorithm

2.1. Data sets

The spectral data were obtained from the OMNIC FTIR spectral library. The original spectrum was ranging from 449 cm^{-1} to 4000 cm^{-1}. In this paper, 823 spectra were chosen from the database. A cross-validation procedure was used. Each time one-fifth of the spectra (164 spectra) were used as the test set and the remainders as the training set.

The functional groups used in this work and percentages of them in the 823 spectra are listed in Table 1. These 16 functional groups are defined on basis of the infrared absorption frequencies. Some of these groups are of general representation (e.g., hydroxyl and carbonyl) while others are less general (e.g., carboxyl and ketone).

* Corresponding author.
E-mail address: liml@scu.edu.cn (M. Li).

https://doi.org/10.1016/j.microc.2020.105395
Received 24 June 2020; Received in revised form 5 August 2020; Accepted 6 August 2020
Available online 09 August 2020
0026-265X/

Table 1
16 functional groups and their percentages in the 823 spectra.

Functional group	Training Set(411)		Testing Set(412)	
	NP	P%	NP	P%
–NH	36	8.76%	36	8.74%
C-N	39	9.49%	41	9.95%
C = C (double band)	109	26.5%	113	27.4%
C≡N (nitrile)	19	4.62%	21	5.10%
–OH (alcohol)	124	30.2%	125	30.3%
Ar-OH (phenol)	24	5.84%	26	6.31%
–OH (hydroxyl)	192	46.7%	192	46.6%
C_6 aromatic (phenyl)	51	12.4%	54	13.1%
C-O-C (ether)	25	6.08%	25	6.07%
C(CO)C (ketone)	43	10.5%	40	9.71%
(CO)H (aldehyde)	25	6.08%	25	6.07%
(CO)OH (carboxylic acid)	87	21.2%	87	21.1%
(CO)OR (ester)	19	4.62%	17	4.13%
(CO)NH (amide)	14	3.41%	16	3.88%
(CO)Cl	5	1.22%	6	1.46%
(CO) (carbonyl)	187	45.5%	185	44.9%

2.2. Support vector Machines

The SVM is a kind of learning machine presented by Vapnik.[23] In the latest years it became extremely popular in variety of classification areas. The training set is assumed as $\{(x_i, y_i), i = 1,2, \cdots,l\}$, in which $x_i \in R^N$ is the input value with corresponding labels $y_i \in R$, l is the number of samples. SVM is to find the maximal margin hyper-plane separating the two sets. The hyper-plane (determined by coefficient αi and b) can be obtained by solving the following convex quadratic programming (QP) optimization problem:

$$Max \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j \mathrm{K}(x_i, x_j) \tag{1}$$

$$Subject\ to \sum_{i=1}^{l} a_i y_i = 0, (0 \leqslant a_i \leqslant C) \tag{2}$$

where C is a parameter that controls the trade-off between margin and classification error. K (xi, xj) is a kernel function[23–24]. In this study, radial basis function (RBF) is selected and given as follows:

$$\mathrm{K}(x_i, x_j) = \exp\left\{\frac{-1}{2\sigma^2}||x - x_i||^2\right\} \tag{3}$$

where σ is the width of kernel function.

The optimization function is:

$$f(x) = Sgn\left\{\sum_{i=1}^{l} y_i a_i \mathrm{K}(x_i, x) + b\right\} \tag{4}$$

For SVM classifiers, 1 for presence and −1 for absence for substructures were encoded as output vectors.

2.3. Artificial neural networks

The ANN is designed on the basis of the biological neutrons capable for parallel signal processing. ANN has the intelligence of calculation and is able to efficiently approximate a function of any form. Back-propagation artificial neural networks (BP-ANNs) were used in this work. In these networks, the elementary processing units are stratified in different layers, in which each unit is connected to all the units in the previous and the next layer. The purpose of the input layer is to receive the input data and to distribute it to the next layer, while the purpose of the output layer is to give the outputs of the network. Other layers are called hidden layers. Information is propagated from input layer through the hidden layers to the output layer. [29–30]

3. Results and discussion

3.1. The training of SVM

Input vector:The original spectrum ranging from 449 cm^{-1} to 4000 cm^{-1} was divided into 307 points of equal interval. The complete spectrum was presented as a set of intensities, in the form of a list of numerical values. Each value represents the absorbance at a frequency determined by its position in the list.

The SVMs used in this work were trained using the sequential minimal optimization (SMO) approach [31]. [19] Parameters (C and Sigma) influencing SVM's training and prediction of functional groups from infrared spectra were scrutinized for the maximal achievable present prediction rate. 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 slightly in order to gain a higher accuracy of the absent prediction. At first, C and Sigma were set with exponentially growing sequence (C = 2^{-1},2^0,2^1,2^2,2^3,…,2^{10}, Sigma = 2^{-1},2^0,2^1,2^2,2^3,…,2^{10}). For all C, Sigma pairs, the SVM was trained with training set and evaluated for test set to get the optimal prediction rate. After one training–testing cycle, a coarse region of C and Sigma can be obtained. Then a finer grid search on that region was conducted for the better prediction rate. After several cycles, the best parameters can be determined.

3.2. The training of ANN

The ANNs used were trained using back-propagation algorithm with momentum method and learning rate based on self-adaptive modulation. The input vector was the same as those used by SVM. Structures were encoded using 0 for absence and 1 for presence. An output value 0.5 was set as the threshold. The output value < 0.5 was interpreted as an absent functionality, and a present one otherwise.

The networks were designed with only one hidden layer, with a varying number of neurons between 1 and 30 for the hidden layer. The classical sigmoid transfer function was used. The initial learning rate and momentum were optimized by a trial and error procedure.

3.3. Model performance

In this work, several indices were used to show various aspects of the quality of SVMs and ANNs.

First of all, Global quality was introduced:

$$\mathrm{GQ(Global\ quality)} = \text{the number of correct responses/the number of total compounds}$$

From the definition of GQ we know that GQ can be used as an indicator of the general performance of the algorithms, but it can give an overoptimistic view of algorithms' performance when the percentage of molecules containing the structure are very low or very high. [10] So we compared it with the statistical chance of giving a correct response which can be calculated on the basis of percentage of the samples, P_i, as follows: $S_t = (1- P_i)^2 + P_i^2$.[10] For very poorly or very highly represented groups, the probability of a chance correct characterization can be very high. The improvement in the quality of response over chance can be expressed as the extra statistical quality index (EQr): [10]

$$\mathrm{EQr} = \text{the difference between experimental and chance of correcting assignments} \\ \text{/maximum possible improvement} = (\mathrm{GQ}-\mathrm{S_t})/(1 - \mathrm{S_t})$$

From the definition of EQr, we know that, the closer the EQr to 1, the better performance of the algorithm. The results shown in Table 2 and Table 3 suggested that the SVMs' EQr of (CO)Cl is the same as that of ANN, while others are higher than those of ANN. From the values of EQrs we know that the SVMs' overall performance is better than ANNs.

Table 2
Prediction performance by SVM on training set and testing set in parentheses.

Functional group	GQ	EQr	P_c	A_c	Q_{pr}	Q_{ar}
–NH	0.998(0.988)	0.988(0.925)	1.000(0.972)	0.997(0.989)	0.973(0.897)	1.000(0.997)
C-N	0.988(0.983)	0.934(0.905)	1.000(0.927)	0.987(0.989)	0.8863(0.905)	1.000(0.992)
C = C	0.993(0.927)	0.985(0.817)	1.000(0.814)	0.990(0.970)	0.9733(0.911)	1.000(0.932)
C≡N	0.993(0.993)	0.922(0.928)	1.000(0.857)	0.992(1.000)	0.864(1.000)	1.000(0.992)
–OH alcohol	0.998(0.980)	0.996(0.953)	1.000(0.968)	0.997(0.986)	0.992(0.968)	1.000(0.986)
Ar-OH	0.998(0.995)	0.982(0.958)	1.000(0.923)	0.997(1.000)	0.960(1.000)	1.000(0.995)
–OH hydroxyl	0.995(0.954)	0.993(0.908)	1.000(0.974)	0.991(0.936)	0.990(0.930)	1.000(0.976)
C6 Aromatic	0.998(0.993)	0.991(0.969)	1.000(1.000)	0.997(0.992)	0.981(0.947)	1.000(1.000)
R-O-R	0.995(0.993)	0.957(0.939)	1.000(0.920)	0.995(0.997)	0.926(1.000)	1.000(0.995)
C(CO)C	0.995(0.983)	0.975(0.903)	1.000(0.850)	0.995(0.997)	0.956(0.971)	1.000(0.984)
(CO)H	0.995(0.985)	0.958(0.868)	0.960(0.840)	0.997(0.995)	0.960(0.913)	0.997(0.990)
(CO)OH	0.990(0.990)	0.974(0.970)	0.989(0.966)	0.991(0.997)	0.966(0.988)	0.997(0.991)
(CO)OR	1.000(1.000)	1.000(1.000)	1.000(1.000)	1.000(1.000)	1.000(1.000)	1.000(1.000)
(CO)NH	1.000(0.995)	1.000(0.933)	1.000(0.938)	1.000(0.997)	1.000(0.938)	1.000(0.997)
(CO)Cl	1.000(1.000)	1.000(1.000)	1.000(1.000)	1.000(1.000)	1.000(1.000)	1.000(1.000)
(CO)	1.000(0.988)	1.000(0.976)	1.000(0.973)	1.000(1.000)	1.000(1.000)	1.000(0.978)
Average Scores	0.996(0.984)	0.978(0.934)	0.997(0.933)	0.995(0.990)	0.964(0.960)	1.000(0.988)

Table 3
Prediction performance by ANN on training set and testing set in parentheses.

Functional group	GQ	EQr	P_c	A_c	Q_{pr}	Q_{ar}
–NH	1.000(0.988)	1.000(0.925)	1.000(0.944)	1.000(0.992)	1.000(0.919)	1.000(0.995)
C-N	1.000(0.980)	1.000(0.888)	1.000(0.878)	1.000(0.992)	1.000(0.923)	1.000(0.987)
C = C	0.995(0.900)	0.989(0.749)	0.982(0.761)	1.000(0.953)	1.000(0.860)	0.993(0.913)
C≡N	1.000(0.990)	1.000(0.897)	1.000(0.857)	1.000(0.997)	1.000(0.947)	1.000(0.992)
–OH alcohol	0.993(0.976)	0.986(0.943)	1.000(0.984)	0.990(0.972)	0.976(0.939)	1.000(0.993)
Ar-OH	1.000(0.993)	1.000(0.941)	1.000(0.923)	1.000(0.997)	1.000(0.960)	1.000(0.995)
–OH hydroxyl	0.983(0.951)	0.976(0.902)	0.995(0.974)	0.973(0.932)	0.970(0.926)	0.995(0.976)
C6 Aromatic	1.000(0.988)	1.000(0.947)	1.000(0.981)	1.000(0.989)	1.000(0.930)	1.000(0.997)
R-O-R	1.000(0.973)	1.000(0.763)	1.000(0.880)	1.000(0.979)	1.000(0.733)	1.000(0.992)
C(CO)C	0.998(0.973)	0.990(0.846)	1.000(0.825)	0.997(0.989)	0.977(0.892)	1.000(0.981)
(CO)H	0.998(0.983)	0.983(0.851)	1.000(0.800)	0.997(0.995)	0.962(0.909)	1.000(0.987)
(CO)OH	0.998(0.990)	0.995(0.970)	1.000(0.977)	0.997(0.994)	0.989(0.977)	1.000(0.994)
(CO)OR	1.000(0.998)	1.000(0.975)	1.000(1.000)	1.000(0.997)	1.000(0.944)	1.000(1.000)
(CO)NH	1.000(0.990)	1.000(0.867)	1.000(0.875)	1.000(0.995)	1.000(0.875)	1.000(0.995)
(CO)Cl	1.000(1.000)	1.000(1.000)	1.000(1.000)	1.000(1.000)	1.000(1.000)	1.000(1.000)
(CO)	1.000(0.988)	1.000(0.976)	1.000(0.984)	1.000(0.991)	1.000(0.989)	1.000(0.987)
Average Scores	0.998(0.979)	0.995(0.902)	0.999(0.915)	0.997(0.985)	0.992(0.920)	0.999(0.987)

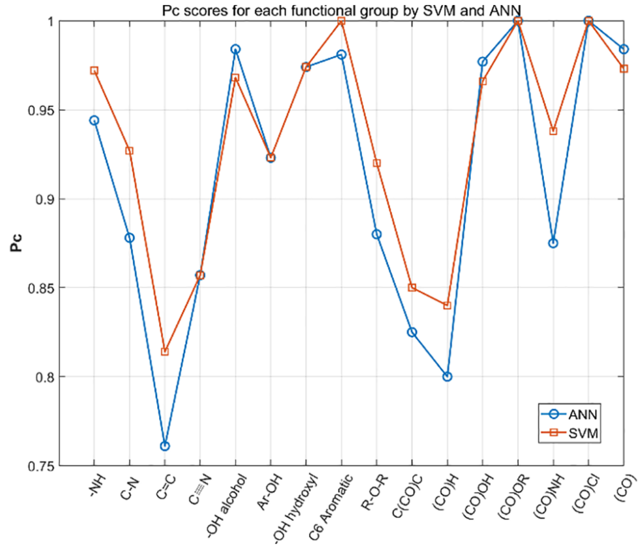

Fig. 1. the Pc scores for each functional group by SVM and ANN.

The indices GQ and EQr only give an overall view of the algorithm's performance. To see the performance difference of the algorithms between the presence and absence of the functional groups, more detailed indices were used:

P_c = the number correctly classified as present/the number present

A_c = the number correctly classified as absent/the number absent

P_c and A_c are shown in Fig. 1 and Fig. 2, respectively. It can be seen from Fig. 1 that only two functional groups are ANN's P_c values slightly higher than SVM's, while others are the same or lower than SVM's. From Fig. 2 shows that only for two functional groups are ANN's A_c values higher than SVM's, while others are the same or lower than SVM's. Although P_c and A_c show the capability of the algorithm in detecting the presence or absence of the functional groups, none of them contains information about how reliable such predictions are. For example, if the fraction of absence found $A_c = 1$, it means that all compounds of the test set without the structure are predicted correctly, but it contains no information about the number of compounds incorrectly classified as absent. P_c is likewise. So two other indices [10] are introduced:

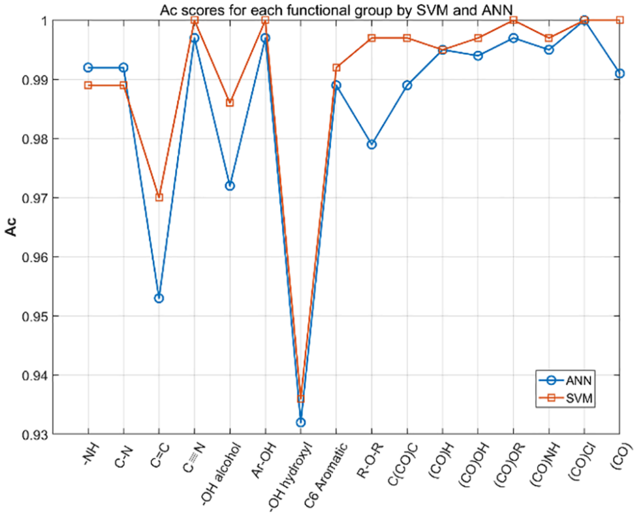

Fig. 2. the Ac scores for each functional group by SVM and ANN.

$$\begin{aligned} Q_{pr}\,&(\text{Quality of found present response}) \\ &= \text{the number correctly classified as present} \\ &\quad /(\text{the number correctly classified as present} \\ &\quad + \text{the number incorrectly classified as present}) \end{aligned}$$

$$\begin{aligned} Q_{ar}\,&(\text{Quality of found absent response}) \\ &= \text{the number correctly classified as absent} \\ &\quad /(\text{the number correctly classified as absent} \\ &\quad + \text{the number incorrectly classified as absent}) \end{aligned}$$

Where Q_{pr} and Q_{ar} are the quality of found present and absent, respectively, ranging from 1 to 0. They reflect the reliability of saying a functional group is present or not. Q_{pr} is related with the corresponding A_c, and only when the value of A_c is high, the Q_{pr} can be high. When Q_{pr} is 1, it reflects that compounds without the structure are all predicted correctively ($A_c = 1$). Q_{ar} is related with the corresponding P_c, and only the value of P_c is high, the Q_{ar} can be high. When Q_{ar} is 1, it reflects that compounds with the structure are all predicted correctively ($P_c = 1$). Q_{pr} and Q_{ar} are listed in the 6th and 7th columns of Table 2 and Table 3. 14 Q_{pr}s by SVM are higher than those by ANN. For most of the functional groups, the Q_{ar}s by SVM are slightly higher than those by ANN. It can be seen from Fig. 1, Fig. 2 that for most of the functional groups, SVMs' prediction rates are higher than ANNs. As to the SVMs, an overall prediction accuracy of 93.3% for the presence (P_c) and 99.0% for the absence (A_c) of functionalities, with an average extra statistical quality (EQr) of 93.4% achieved. The quality of found present response (Q_{pr}) and absent response (Q_{ar}) was 96.0% and 98.8%, respectively. As to the ANNs, an overall prediction accuracy of 91.5% for the presence (P_c) and 98.5% for the absence (A_c) of functionalities, with average extra statistical quality (EQr) of 90.2%, were achieved, respectively. The quality of found present response (Q_{pr}) and absent response (Q_{ar}) was 92.0% and 98.7%, respectively.

3.4. Correlation between discrimination ability and characteristic infrared absorption

Theoretically, both SVM and ANN can predict and find the structure features that show very distinctive characteristic infrared absorption at a unique position easily. Structure features with both shared common infrared bands and distinctive absorption at the same time could also be well recognized. While those structure features without distinctive absorption are less likely to be well recognized. Although ketone has a distinctive absorption band (CO stretching band) around 1715 cm^{-1}, neither SVM nor ANN can recognize it easily. It is probably because that the stretching band of CO is overlapped with those of esters, aldehyde and carboxylic acid which fall inside the carbonyl range. Although the stretching band of CO may be overlapped with other compounds containing carbonyl, the carboxylic acid still has other distinctive absorption bands such as the C-O and O–H stretching vibrations, so SVM and ANN can easily recognize the carboxylic acid and the P_c of them are all > 0.960. The double band only has a weak absorption band (C = C stretching band) around 1650 cm^{-1}, which is difficult for SVM or ANN to recognize, so its P_c is very low.

3.5. Advantages of SVMs over ANNs

A fundamental difference between SVM and ANN is that SVM can overcome local minima. SVM is also more stable than ANN. Given the same training data, C and Sigma, SVM will always get the same answer. Because the training of ANN depends on the initial weight, the number of interactions for optimal configuration can differ considerably, and several training sessions are necessary to establish the optimum number. Then for a given set of parameters, ANN may get different answers. This is also the reason why the training of ANN is so time consuming.

4. Conclusions

In this work, we investigated the structure elucidation of infrared spectra. The trained SVMs can identify the presence or absence of the functional groups with the average prediction accuracy of 93.3% for the presence (P_c) and 99.0% for the absence (A_c) of functionalities. The quality of found present response (Q_{pr}) and absent response (Q_{ar}) was 96.0% and 98.8%, respectively. The average extra statistical quality (EQr) was 93.4%. The results were compared with those obtained by using ANNs methods, and showed that SVMs over-perform ANNs in most of the cases. From the results we can come to the conclusion that SVM approach is a powerful tool for the interpretation of infrared spectra. Moreover, it would be of great significance to extend such machine learning method to the practical application. For example, to develop web-server or tools available to the practically measured IR spectra in the future.

CRediT authorship contribution statement

Zhimeng Wang: Conceptualization, Investigation, Writing - original draft. **Xiaoyu Feng:** Data curation, Investigation. **Junhong Liu:** Visualization, Writing - review & editing. **Minchun Lu:** Methodology. **Menglong Li:** Supervision, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors wish to express their gratitude to anonymous for their contributive suggestions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.microc.2020.105395.

References

[1] G.N. Andreev, O.K. Argirov, EXPIRS, an expert system for generation of alternative sets of substructures, derived by infrared spectra interpretation, Anal. Chim. Acta. 321 (1996) 105–111.

[2] G.N. Andreev, O.K. Argirov, Implementation of human expert heuristics in computer supported infrared spectra interpretation, J. Mol. Struct. 347 (1995) 439–448.

[3] B.J. Debska, B. Guzowska-Swider, Knowledge discovery in an infrared database, Comput. Chem. 21 (1997) 51–59.

[4] M. Bos, J.A.M. Vrielink, The wavelet transform for pre-processing IR spectra in the identification of mono- and di-substituted benzenes, Chemom. Intell. Lab. Syst. 23 (1994) 115–122.

[5] A. Kai-man Leung, F.T. Chau, J.B. Gao, T.M. Shih, Application of wavelet transform in infrared spectrometry: spectral compression and library search, Chemom. Intell. Lab. Syst. 43 (1998) 69–88.

[6] Kai-man Leung, A. Chau, F. T. Gao. J. B. A review on applications of wavelet transform techniques in chemical analysis: 1989–1997. Chemom. Intell. Lab. Syst. 1998, 43, 165-184.

[7] D. Svozil, V. Kvasnicka, J. Pospichal, Introduction to multi-layer feed-forward neural networks, Chemom. Intell. Lab. Syst. 39 (1997) 43–62.

[8] W. Werther, W. Demuth, F.R. Krueger, J. Kissel, E.R. Schmid, K. Varmuza, Evaluation of mass spectra from organic compounds assumed to be present in cometary grains, Exploratory data analysis. *J. Chemom.* 16 (2002) 99–110.

[9] Coates, J., Interpretation of Infrared_Spectra, A Practical Approach. Encyclopedia of Analytical Chemistry, 2006.

[10] U.M. Weigel, R. Herges, Automatic interpretation of infrared spectra: recognition of aromatic substitution patterns using neural networks, J. Chem. Inf. Comput. Sci. 32 (1992) 723–731.

[11] P.N. Penchev, G.N. Andreev, K. Varmuza, Automatic classification of infrared spectra using a set of improved expert-based features, Anal. Chim. Acta. 388 (1999) 145–159.

[12] T. Kazutoshi, M. Takatoshi, T. Tadao, S. Jiro, H. Shinnosuke, A. Miwako, O. Chisato, I. Shoji, U. Hiroyuki, T. Yasuhiro, Y. Kazushige, I. Tetsuya, M. Michiko, K. Shoji, M. Hisashi, K. Fumiko, J. Chihiro, O. Shuichiro, Identification of Chemical Structures from Infrared Spectra by Using Neural Networks, Appl. Spectrosc. 55 (2001) 1394–1403.

[13] M.E. Munk, Computer-Based Structure Determination: Then and Now, J. Chem. Inf. Comput. Sci. 38 (1998) 997–1009.

[14] Z. Wang, B. Xiang, Application of artificial neural network to determination of active principle ingredient in pharmaceutical quality control based on, near infrared spectroscopy, Microchemical Journal. 89 (1) (2008) 52–57.

[15] M. Blanco, et al., NIR calibration in non-linear systems: different PLS approaches and artificial neural networks, Chemometrics and Intelligent Laboratory Systems 50 (1) (2000) 75–82.

[16] R.M. Balabin, E.I. Lomakina, R.Z. Safieva, Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy, FUEL 90 (5) (2011) 2007–2015.

[17] M. Khanmohammadi, A.B. Garmarudi, K. Ghasemi, S. Garrigues, M. de la Guardia, Artificial neural network for quantitative determination of total protein in yogurt by infrared spectrometry, Microchemical Journal. 91 (1) (2009) 47–52.

[18] U. Thissen, M. Pepers, B. Ustun, W.J. Melssen, L.M.C. Buydens, Comparing support vector machines to PLS for spectral regression applications, Chemom. Intell. Lab. Syst. 73 (2004) 169–179.

[19] M. Kamruzzaman, S. Takahama, A.M. Dillner, Quantification of amine functional groups and their influence on OM/OC in the IMPROVE network, Atmospheric Environment 172 (2018) 124–132.

[20] K.E. Wilcox, E.W. Blanch, A.J. Doig, Determination of Protein Secondary Structure from Infrared Spectra Using Partial Least-Squares Regression, Biochemistry 55 (27) (2016) 3794–3802.

[21] J.A. Fine, A.A. Rajasekar, K.P. Jethava, G. Chopra, Spectral deep learning for prediction and prospective validation of functional groups, Chemical Science 11 (2020) 4618–4630.

[22] A.I. Belousov, S.A. Von Verzakov, J. Frese, A flexible classification approach with optimal generalisation performance: support vector machines, Chemom. Intell. Lab. Syst. 64 (2002) 15–25.

[23] Vapnik, V. N. Inc. Statistical Learning Theory, John Wiley and Sons press, 1998. (translated into Chinese: Publishing House of Electronics Industry; Peking, 2004; 293-379).

[24] K.W. Lau, Q.H. Wu, Online training of support vector classifier, Pattern Recogn. 36 (2003) 1913–1920.

[25] X.-Y. Feng, Q.-Q. Wang, J. Zhang, F.-S. Nie, M.-L. Li, Studying aromatic compounds in infrared spectra based on support vector machine, Vibrational Spectroscopy. 44 (2) (2007) 243–247.

[26] Nalla, R., Pinge, R., Narvaria, M., Chaudhury, B. Priority based functional group identification of organic molecules using machine learning. in Proceedings of the ACM India Joint International Conference on Data Science and Management of Data - CoDS-COMAD ’18, 2018, 201–209.

[27] X. Niu, Z. Zhao, K. Jia, X. Li, A feasibility study on quantitative analysis of glucose and fructose in lotus root powder by FT-NIR spectroscopy and chemometrics, Food Chemistry. 133 (2) (2012) 592–597.

[28] O. Devos, et al., Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation, Chemometrics and Intelligent Laboratory Systems 96 (2009) 27–33.

[29] Z. Ramndan, P.K. Hopke, M.J. Johnson, K.M. Scow, Application of PLS and Back-Propagation Neural Networks for the estimation of soil properties, Chemom. Intell. Lab. Syst. 75 (2005) 23–30.

[30] B. Kim, S. Kim, Plasma diagnosis by recognizing in situ data using a modular back propagation network, Chemom. Intell. Lab. Syst. 65 (2003) 231–240.

[31] G.F. William, L. Steve, Efficient SVM regression training with SMO, Mach. Learn. 46 (2002) 271–290.