Near Infrared Spectroscopic Quantification Using Firefly Wavelength Interval Selection Coupled with Partial Least Squares

Xihui Bian[1,2,3(✉)], Zizhen Zhao[1], Hao Sun[1], Yugao Guo[1], and Lizhuang Hao[3]

[1] State Key Laboratory of Separation Membranes and Membrane Processes, School of Chemical Engineering and Technology, Tiangong University, Tianjin 300387, China
bianxihui@163.com

[2] Key Lab of Process Analysis and Control of Sichuan Universities, Yibin University, Yibin 644000, Sichuan, China

[3] State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xining 810016, China

Abstract. Firefly algorithm (FA) combined with partial least squares (PLS) are developed for near infrared (NIR) spectral interval selection and quantitative analysis of complex samples. The method firstly segments the near-infrared spectra into a number of intervals. Vectors with 1 and 0, which represent the interval selected or not, are used as the inputs of the FA. The RMSEP value predicted by PLS model is used as the fitness function of the FA. The number of spectral intervals, the population number, environmental absorbance and the constant of FA are optimized. With the optimal parameters, FA-PLS model is established and applied to predict protein, hemoglobin and cetane number in wheat, blood and diesel fuel samples, respectively. The results show that FA-PLS can significantly improve the prediction accuracy compared with full-spectrum PLS model.

Keywords: Variable selection · Firefly algorithm · Multivariate calibration · Partial least squares · Near infrared spectroscopy

1 Introduction

Analysis of complex samples is a challenging task in analytical chemistry and industries [1, 2]. In addition, traditional separation methods are difficult and time-consuming. Therefore, it is necessary to find an appropriate method to analyze complex samples. Spectral analysis provides a simple method for the analysis of complex samples due to its advantages of simplicity, rapidity and non-destructiveness [3–5]. It is widely used in agricultural commodities [6], medical [7, 8], food [9] and tobacco [10], *etc.* Nevertheless, the spectral peaks are overlapping severely in complex samples, which can reduce the prediction performance. Thus, multivariate calibration is required in the quantitative analysis of complex samples.

X. Chu et al. (Eds.): ICNIR 2021, *Sense the Real Change: Proceedings of the 20th International Conference on Near Infrared Spectroscopy*, pp. 274–282, 2022.
https://doi.org/10.1007/978-981-19-4884-8_30

The commonly used multivariate calibration methods include multiple linear regression (MLR) [11], principal component regression (PCR) [12], partial least squares (PLS) [13, 14], artificial neural network (ANN) [15] and extreme learning machine (ELM) [16], *etc.* Among these multivariate calibration methods, PLS is the most popular technique in multivariate calibration. However, with the development of modern analytical instruments, spectra data contain an enormous number of wavelength. In some cases, some wavelength consist of irrelevant information. These irrelevant wavelength usually degrade the prediction performance of model. Therefore, wavelength selection is required before multivariate calibration.

At present, many wavelength selection methods have been developed. They mainly includes individual wavelengths and spectral interval selection based on single index [17], statistics [18, 19] and swarm intelligence optimization algorithms [20, 21]. Among these methods, the swarm intelligence optimization algorithm has attracted increasing attention, especially the genetic algorithm (GA) [22, 23]. GA is inspired by natural evolution, which simulates the phenomena of crossover and mutation that occur in natural selection. This process is repeated continuously and finally obtains the optimal individual in population. However, this method has some disadvantages such as slow convergence speed and easily trapping into local optimum. Therefore, a series of new swarm intelligence optimization algorithms have been proposed.

Inspired by the flashing behavior of fireflies, Yang [24] proposed the firefly algorithm (FA). In FA, the less bright fireflies can follow the brightest firefly by attraction. The group of fireflies can gradually close to the area where the brightest firefly is located. Furthermore, the brightest individual firefly is considered as the optimal solution. The position iteration is realized by this process. Although FA has been widely used in other fields, relative few research about FA are carried out in spectral analysis fields [25, 26].

In this study, the feasibility of near infrared (NIR) spectral interval selection by FA is discussed for multivariate calibration of complex samples. The number of spectral intervals is determined firstly. Then the population number, environmental absorbance and constant of FA are optimized, respectively. With the optimal parameters, FA is used for interval selection and then PLS model is established. Compared with the full-spectrum PLS model, FA-PLS has lower root mean square error of prediction (RMSEP) and higher correlation coefficient (R) for predicting samples in prediction set.

2 Theory and Algorithm

As a swarm intelligence algorithm, FA has advantages of fast convergence speed and strong global optimization ability. The realization of this algorithm is based on the following three assumptions (i) all fireflies are unisex (ii) the brighter fireflies are more attractive than the other fireflies, *i.e.*, attractiveness is proportional to the brightness. The attractiveness and brightness can decrease with the increasing of distance, the brightest firefly can move randomly (iii) the brightness of fireflies depends on objective function. I_i is the brightness of firefly i, $\vec{x}_i$ is the current location of firefly i. The equation of firefly brightness I_i at current location is as follows.

$$I_{\mathrm{i}} = \mathrm{f}(\overrightarrow{\mathrm{x_i}}) \tag{1}$$

where the brightness of firefly is equal to the value of objective function. Fireflies are attracted to fireflies that are brighter than themselves. The equation of relative brightness between two fireflies is as follows.

$$I_{ij}(r_{ij}) = I_i e^{\gamma \times r_{ij}^2} \tag{2}$$

where I_{ij} is the light intensity. γ is the environmental absorbance. r_{ij} is the distance between firefly i and firefly j. The distance formula of the standard FA is as follows.

$$r_{ij} = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2} \tag{3}$$

where d is the dimension of the solution. $x_{i,k}$ and $x_{j,k}$ are the kth dimension components of spatial coordinates x_i and x_j, respectively. The attractiveness is proportional to the brightness, the attractiveness β can be defined as

$$\beta_{ij}(r_{ij}) = \beta_0 e^{\gamma \times r_{ij}^2} \tag{4}$$

where β_0 is the maximum attraction. The distance is zero between firefly i and firefly j. The position update formula can be written as

$$\overrightarrow{x_j}(t+1) = \overrightarrow{x_j}(t) + \beta_{ij}(r_{ij})[\overrightarrow{x_i}(t) - \overrightarrow{x_j}(t)] + \alpha \overrightarrow{\varepsilon_j} \tag{5}$$

where t is the number of iterations, α is a constant and ε_j is the random number in Gaussian distribution. The schematic diagram of FA is shown in Fig. 1. The process FA is as follows. First of all, the parameters of algorithm need to be set, such as the population number, environmental absorbance and constant. The next step is to initialize

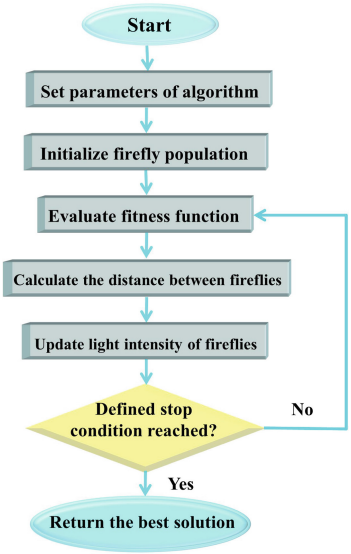

Fig. 1. The flowchart of FA algorithm.

firefly population. The third step is to evaluate fitness function. The better the fitness, the greater brightness of firefly is. The fourth step is to calculate the distance r_{ij} between fireflies. The fifth step is to update light intensity of fireflies. The last step is to output the best solution when the maximum number of iterations is reached. Otherwise, skip to the third step to continue iteration.

In the above process of FA, the population number n, environmental absorbance γ and constant α need to be optimized. Moreover, the number of wavelength interval is also an important parameter, which needs to be determined for the input of FA.

3 Experimental

Three NIR spectral datasets were used to evaluate the predictive performance of FA-PLS. Wheat dataset was contributed by P.C. Williams, which consists of visible-NIR spectra and six properties of 884 wheat samples. [27] The Vis-NIR spectra and the protein contents are used in this study. The spectra were scanned on a Foss Model 6500 over 1050 channels recorded in the wavelength range of 400–2498 nm with the digitization interval of 2 nm. The reference values of protein contents were determined at the Grain Research Laboratory, Winnipeg. The samples Nos. 680 and 681 are two outliers and have been deleted. Figure 2(a) displays the NIR spectra of the 882 samples.

Blood dataset was provided by Norris *et al.* [28], which includes the NIR transmission and reflection spectra and four properties of 231 blood samples. The NIR reflection spectra and hemoglobin concentrations were used for this study. The spectra were scanned by model 6500 spectrometer (NIR systems, Inc., Silver Springs, USA). Each spectrum is composed of 700 variables recorded in the wavelength range 1100–2498 nm with a 2 nm interval. Figure 2(b) displays the NIR reflection spectra of the 231 samples.

Diesel fuel dataset was provided by SWRI, San Antonio, TX through Eigenvector Research, Inc. (Manson, Washington) [29]. It consists of NIR spectra and six properties of 256 fuel samples. The NIR spectra and cetane number values were investigated for this study. The spectra were measured at Southwest Research Institute (SWRI) on a project sponsored by the U.S. Army. Each spectrum is composed of 401 variables recorded in the wavelength range of 750–1550 nm. The cetane numbers were independently measured by the American Society of Testing and Materials (ASTM) standard method. Figure 2(c) displays the NIR spectra of the 256 samples.

Before calculation, the three datasets were divided into the training and prediction sets as described on the website for model building and performance validation, respectively. For wheat dataset, 775 and 107 samples were used as the training set and prediction sets. For blood dataset, 173 and 58 samples were used as the training and prediction sets. For the diesel fuel dataset, 138 and 118 samples were taken as the training and prediction sets. For PLS modeling, the optimal latent variable (LV) number is determined as 10, 11 and 9 by Monte Carlo cross-validation combined with F-test for wheat, blood and diesel fuel datasets, respectively.

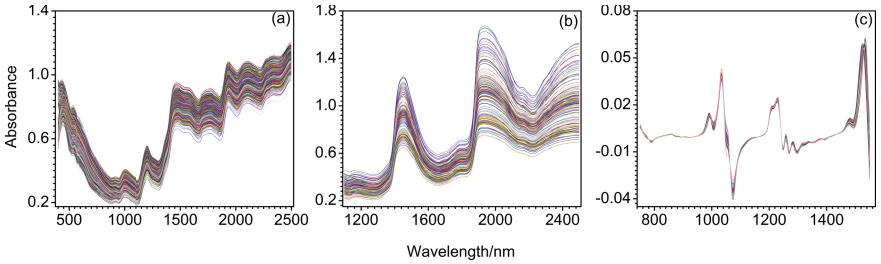

Fig. 2. NIR spectra for wheat (a), blood (b) and diesel fuel (c) datasets, respectively.

4 Results and Discussion

4.1 Determination the Interval Number

Interval number is a key parameter for FA-PLS. In order to get the optimal interval number, FA parameter takes default values, *i.e.*, the population number is 20, environmental absorbance is 1 and the constant is 0.5. The interval numbers are investigated in the range of 5–30 with an interval of 5. For each interval number, FA-PLS is performed and a RMSEP can be obtained. The variation of RMSEP with the number of intervals for wheat dataset is shown in Fig. 3.

As demonstrated in Fig. 3, with the increase of interval number, the RMSEP decreases before 10. After that, the RMSEP increases significantly. RMSEP represents predictive ability of the model, a model with a better parameter should have a lower RMSEP. Thus, the optimal interval number for wheat dataset is 10. Similarly, the optimal interval number is determined as 20 for both blood and diesel fuel datasets.

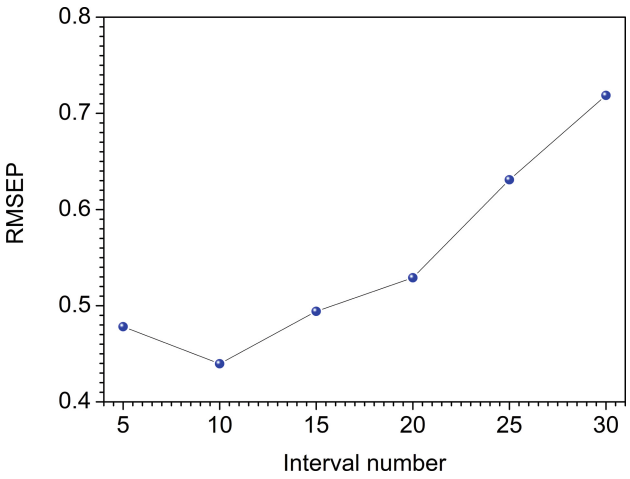

Fig. 3. Variation of RMSEP with interval number for wheat dataset.

4.2 Parameter Optimization of FA

The population number n, environmental absorbance γ and constant α are three important parameters for FA. The three parameters are optimized successively for each dataset. For determining the optimal population number n, the interval number is used the optimal value determined above. Other parameters take default values, *i.e.*, environmental absorbance is 1 and the constant is 0.5. Population number changes from 10 to 60 with an interval of 10. For each population number, FA-PLS is performed and a RMSEP can be obtained. Figure 4(a) shows the variation of RMSEP values with the population number for wheat dataset. Clearly, the RMSEP decreases with the increase of population number before 30. Above 30, the RMSEP value has an increasing trend. The lowest RMSEP is located at 30. Thus, the optimal population number is 30 for wheat dataset. Because FA is swarm intelligence, it is easy to understand that too few and too many fireflies are not good for population performance. Similar analysis can be performed for blood and diesel fuel datasets and the optimal population number is set as 30 and 40, respectively.

Environmental absorbance γ is determined subsequently. To determine the optimal environmental absorbance, interval number and population number are used the optimal values determined above, the constant is used the default value 0.5. Environmental absorbance is investigated in the range of 0.1–1.2 with interval of 0.1. Figure 4(b) depicts the variation of RMSEP with environmental absorbance for wheat dataset. It can be seen that the RMSEP is comparatively large at the beginning. With the increasing of environmental absorbance, although it has some fluctuation, the overall trend of RMSEP is decreasing. When the environmental absorbance is 0.5, RMSEP reaches the lowest value. Accordingly, 0.5 is used as the optimal environmental absorbance for wheat dataset. For blood and diesel fuel datasets, the optimal environmental absorbance is 0.8 and 0.9, respectively.

The optimal constant α is the last parameter need to be optimized. The interval number, population number, environmental absorbance are set to 10, 30 and 0.5 determined above for wheat dataset, respectively. The constant is investigated in the range of 0.1–1 with an interval 0.1. From Fig. 4(c), it is obvious that the RMSEP tends to decrease with the increase of constant before 0.3. When it is above 0.3, the overall trend in RMSEP is increasing. Thus, the optimal constant is set as 0.3 for wheat dataset. Similarly, 0.1 and 0.5 are the optimal α for blood and diesel fuel dataset, respectively.

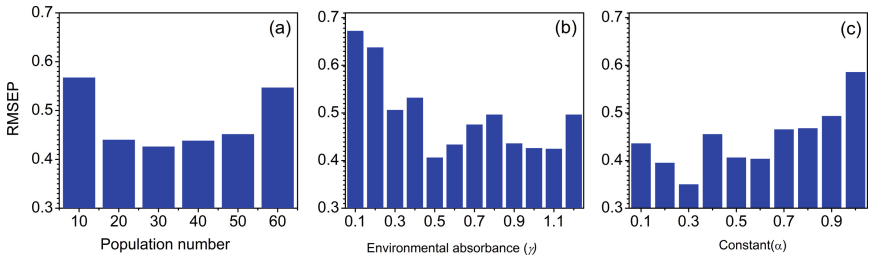

Fig. 4. Variation of RMSEPs with population number (a), environmental absorbance γ (b) and constant α (c) for wheat dataset.

4.3 Prediction Results

The optimal interval number, population number n, environmental absorbance γ and constant α for the three datasets are summarized in Table 1. With the optimal parameters, FA is used to select the spectral intervals of the samples in training set and then built PLS model. To validate the efficiency of FA method, the same spectral intervals of the samples in prediction set are selected and input into the model. The predicted values are obtained and used to calculate RMSEP and R for the three datasets, which are shown in Table 2. For comparison, the full-spectrum PLS results for the three datasets are also listed in Table 2. Obviously, a better model should have a lower RMSEP and a larger R. As shown in Table 2, with FA, the RMSEP of PLS decrease from 0.7763, 0.4310, 0.0023 to 0.3498, 0.3308, 0.0014 for wheat, blood and diesel fuel datasets, respectively. The R values of the three datasets all increase after FA interval selection. The results show that FA spectral interval selection can improve the prediction accuary of PLS obviously. To sum up, FA-PLS is an efficient method for NIR spectral interval selection and NIR spectral quantitative analysis.

Table 1. The parameter optimization results for different datasets

Datasets	Interval number	n	γ	α
Wheat	10	30	0.5	0.3
Blood	20	30	0.8	0.1
Diesel oil	20	40	0.9	0.5

Table 2. Prediction results of PLS before and after interval selection by FA

Datasets	Methods	RMSEP	R
Wheat	PLS	0.7763	0.9365
	FA-PLS	0.3498	0.9762
Blood	PLS	0.4310	0.9613
	FA-PLS	0.3308	0.9777
Diesel oil	PLS	0.0023	0.9744
	FA-PLS	0.0014	0.9901

5 Conclusion

FA interval selection coupled with PLS is used for NIR spectroscopic quantification of complex samples. In this approach, the parameters are firstly optimized and then wavelength intervals are selected by FA. With the optimal parameters, FA-PLS model is established and applied to predict unknown samples. In order to verify the validity of

the method, the contents of protein in wheat, hemoglobin in blood and cetane number in diesel fuel samples are predicted, respectively. The RMSEP and R of FA-PLS are compared with those of PLS on the three datasets. Result shows that FA can effectively improve the prediction performance of PLS.

Acknowledgments. This study is supported by Key Lab of Process Analysis and Control of Sichuan Universities (No. 2020001) and Opening Foundation of State Key Laboratory of Plateau Ecology and Agriculture (No. 2021-KF-07).

References

1. Wei, J.L., Huang, D.Y., Chen, Y.C.: Using gadolinium ions as affinity probes to selectively enrich and magnetically isolate bacteria from complex samples. Anal. Chim. Acta **1113**, 18–25 (2020)
2. Offermans, T., et al.: ENDBOSS: industrial endpoint detection using batch-specific control spaces of spectroscopic data. Chemometr. Intell. Lab. Syst. **209**, 104229 (2021)
3. Bian, X.H., et al.: Rapid identification of milk samples by high and low frequency unfolded partial least squares discriminant analysis combined with near infrared spectroscopy. Chemometr. Intell. Lab. Syst. **170**, 96–101 (2017)
4. Su, T., Sun, Y., Han, L., Cai, W.S., Shao, X.G.: Revealing the interactions of water with cryoprotectant and protein by near-infrared spectroscopy. Spectrochim. Acta A **266**, 120417 (2022)
5. Li, J.Y., Chu, X.L., Tian, S.B., Lu, W.Z.: The identification of highly similar crude oils by infrared spectroscopy combined with pattern recognition method. Spectrochim. Acta A **112**, 457–462 (2013)
6. Appell, M., Compton, D.L., Bosma, W.B.: Raman spectral analysis for rapid determination of zearalenone and alpha-zearalanol. Spectrochim. Acta A **270**, 120842 (2022)
7. Lin, L., et al.: A rapid analysis method of safflower (Carthamus tinctorius L.) using combination of computer vision and near-infrared. Spectrochim. Acta A **236**, 118360 (2020)
8. Bian, X.H., Lu, Z.K., van Kollenburg, G.H.: Ultraviolet-visible diffuse reflectance spectroscopy combined with chemometrics for rapid discrimination of Angelicae Sinensis Radix from its four similar herbs. Anal. Methods **12**, 3499–3507 (2020)
9. Urickova, V., Sadecka, J.: Determination of geographical origin of alcoholic beverages using ultraviolet, visible and infrared spectroscopy: a review. Spectrochim. Acta A **148**, 131–137 (2015)
10. Ma, Y.J., et al.: Rapid determination of four tobacco specific nitrosamines in burley tobacco by near-infrared spectroscopy. Anal. Methods **4**, 1371–1376 (2012)
11. Lemos, T., Kalivas, J.H.: Leveraging multiple linear regression for wavelength selection. Chemometr. Intell. Lab. Syst. **168**, 121–127 (2017)
12. Gemperline, P.J., Salt, A.: Principal components regression for routine multicomponent UV determinations: a validation protocol. J. Chemom. **3**, 343–357 (1989)
13. Wold, S., Sjostrom, M., Eriksson, L.: PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. Syst. **58**, 109–130 (2001)
14. Zhang, H., Hu, X.Y., Liu, L.M., Wei, J.F., Bian, X.H.: Near infrared spectroscopy combined with chemometrics for quantitative analysis of corn oil in edible blend oil. Spectrochim. Acta A **270**, 120841 (2022)
15. Fei, Q., Li, M., Wang, B., Huan, Y.F., Feng, G.D., Ren, Y.L.: Analysis of cefalexin with NIR spectrometry coupled to artificial neural networks with modified genetic algorithm for wavelength selection. Chemometr. Intell. Lab. Syst. **97**, 127–131 (2009)

16. Bian, X.H., Li, S.J., Fan, M.R., Guo, Y.G., Chang, N., Wang, J.J.: Spectral quantitative analysis of complex samples based on extreme learning machine. Anal. Methods **8**, 4674–4679 (2016)
17. Wu, W., Walczak, B., Massart, D.L., Prebble, K.A., Last, I.R.: Spectral transformation and wavelength selection in near-infrared spectra classification. Anal. Chim. Acta **315**, 243–255 (1995)
18. Xu, Z.C., Liu, W.S., Cai, X.G.: Shao, A wavelength selection method based on randomization test for near-infrared spectral analysis. Chemometr. Intell. Lab. Syst. **97**, 189–193 (2009)
19. Han, Q.J., Wu, H.L., Cai, C.B., Xu, L., Yu, R.Q.: An ensemble of monte carlo uninformative variable elimination for wavelength selection. Anal. Chim. Acta **612**, 121–125 (2008)
20. Shamsipur, M., Zare-Shahabadi, V., Hemmateenejad, B., Akhond, M.: Ant colony optimisation: a powerful tool for wavelength selection. J. Chemom. **20**, 146–157 (2006)
21. Wu, X.Y., Bian, X.H., Yang, S., Xu, P., Wang, H.T.: A variable selection method for near infrared spectroscopy based on gray wolf optimizer algorithm. J. Instrum. Anal. **39**, 1288–1292 (2020)
22. Mehmood, T., Liland, K.H., Snipen, L., Saebo, S.: A review of variable selection methods in partial least squares regression. Chemometr. Intell. Lab. Syst. **118**, 62–69 (2012)
23. Mohammadi, M., et al.: Genetic algorithm based support vector machine regression for prediction of SARA analysis in crude oil samples using ATR-FTIR spectroscopy. Spectrochim. Acta A **245**, 118945 (2021)
24. Yang, X.S.: Firefly algorithms for multimodal optimization. Lect. Notes Comput. Sci. **5792**, 169–178 (2009)
25. Goodarzi, M., Coelho, L.D.: Firefly as a novel swarm intelligence variable selection method in spectroscopy. Anal. Chim. Acta **852**, 20–27 (2014)
26. Attia, K.A.M., Nassar, M.W.I., El-Zeiny, M.B., Serag, A.: Firefly algorithm versus genetic algorithm as powerful variable selection tools and their effect on different multivariate calibration models in spectroscopy: a comparative study. Spectrochim. Acta A **170**, 117–123 (2017)
27. Bian, X.H., et al.: Robust boosting neural networks with random weights for multivariate calibration of complex samples. Anal. Chim. Acta **1009**, 20–26 (2018)
28. Kuenstner, J.T., Norris, K.H., McCarthy, W.F.: Measurement of hemoglobin in unlysed blood by near-infrared spectroscopy. Appl. Spectrosc. **48**, 484–488 (1994)
29. Soyemi, O.O., Busch, M.A., Busch, K.W.: Multivariate analysis of near-infrared spectra using the G-programming language. J. Chem. Inf. Comput. Sci. **40**, 1093–1100 (2000)