NEW INDUCED MUTATION GENETIC ALGORITHM FOR SPECTRAL VARIABLES SELECTION IN NEAR INFRARED SPECTROSCOPY

X. G. Zhuang [1,2*], X. S. Shi [1,2], P. J. Zhang [1], H. B. Liu [1], C. M. Liu [1], H. F. Wang [1]

[1] The 41st Research Institute of CETC, Qingdao, China; e-mail: xingangzhuang@163.com
[2] Science and Technology on Electronic Test & Measurement Laboratory, Qingdao, China

In this paper, a new spectral variables selection method, induced mutation genetic algorithm (IMGA), is proposed for near-infrared (NIR) spectroscopy. Based on the idea of genetic algorithm (GA), the IMGA greatly simplifies the process of biological evolution, which not only inherits the advantages of global optimization of the GA, but also effectively improves the convergence speed. In this study, the IMGA is applied to the selection of characteristic spectral variables for green tea origin identification. After five times of genetic evolutions, 11 characteristic spectral variables are selected from 156 spectral variables. Based on the 11 characteristic spectral variables, the classification model is built by partial least squares (PLS), and both the sensitivity and specificity of classification model are raised to 1 for prediction set. The overall results indicate that the IMGA can be well applied to the selection of characteristic spectral variables and effectively improve the prediction accuracy and calculation speed of the near-infrared model.

Keywords: *induced mutation genetic algorithm (IMGA), NIR spectroscopy, spectral variables, green tea.*

АЛГОРИТМ ИНДУЦИРОВАННЫХ ГЕНЕТИЧЕСКИХ МУТАЦИЙ ДЛЯ ОТБОРА СПЕКТРАЛЬНЫХ ПЕРЕМЕННЫХ В БЛИЖНЕЙ ИНФРАКРАСНОЙ СПЕКТРОСКОПИИ

X. G. Zhuang [1,2*], X. S. Shi [1,2], P. J. Zhang [1], H. B. Liu [1], C. M. Liu [1], H. F. Wang [1]

УДК 535.34:575.24/.25

[1] 41-й Научно-исследовательский институт CETC, Циндао, Китай; e-mail: xingangzhuang@163.com
[2] Научно-техническая лаборатория электронных испытаний и измерений, Циндао, Китай

(Поступила 21 июня 2018)

Предлагается новый метод выбора спектральных переменных для ближней ИК спектроскопии — алгоритм индуцированных генетических мутаций (IMGA). Основываясь на генетическом алгоритме (GA), IMGA значительно упрощает процесс биологической эволюции, которая не только наследует преимущества глобальной оптимизации GA, но и эффективно улучшает скорость конвергенции. IMGA применен для выбора характерных спектральных переменных для идентификации происхождения зеленого чая. После пяти серий генетических эволюций из 156 спектральных переменных выбраны 11 характерных спектральных переменных, на основе которых с помощью метода частичных наименьших квадратов (PLS) построена классификационная модель, при этом для набора прогнозирования чувствительность и определенность классификационной модели повышаются до единицы. Показано, что IMGA можно использовать для выбора характеристических спектральных переменных и существенно улучшить точность прогнозирования и скорость расчета модели для ближней ИК области.

Ключевые слова: *алгоритм индуцированных генетических мутаций, ближняя ИК спектроскопия, спектральные переменные, зеленый чай.*

Introduction. Near-infrared (NIR) spectroscopy has been one of the fastest developing analytical techniques in recent years. It is a simple, fast and accurate method for nondestructive testing of material components. NIR spectroscopy has been widely applied in quality inspection and geographical origin identification of agricultural products [1–5]. The most intensive bands in the NIR spectral region belong to the fundamental frequency and overtone vibration of hydrogen-containing functional groups. While identifying agricultural products, it is difficult to determine their geographical origin. If the full spectrum data are applied to build identification model, unrelated spectral variables will be introduced into it. In previous study, the calibration models were mainly built by characteristic spectral variables for both qualitative identification and quantitative detection, which not only improves the prediction accuracy, but also improves the calculation speed [6–8].

The commonly used methods for spectral variables selection are mainly genetic algorithm (GA) [9], simulated annealing algorithm (SA) [10], successive projection algorithm (SPA) [11] and the methods based on moving window [12, 13]. It is worth noting that GA has attracted much attention due to their global optimization capabilities. Combining GA with partial least squares (PLS), R. Leardi, et al. [14] first proposed genetic algorithm-partial least squares (GA-PLS) to select characteristic spectral variables. The main drawback of GA-PLS is that the spectral variables cannot be too many, otherwise the algorithm converges with difficulty. To solve this problem, interval partial least square with genetic algorithm (iPLS–GA) is presented by Chen et al. [15]. The spectrum is firstly divided into several spectral intervals, and the characteristic intervals and variables are selected by iPLS–GA and GA-PLS, respectively. This method is suitable for large numbers of spectral variables, but easily causes mis-selection. All the above methods are based on the GA proposed by J. Holland [16] in 1975. The traditional GA mainly consists of three steps: selection, crossover and mutation, which has obvious shortcomings in convergence speed [17].

Inspired by mutation breeding, induced mutation genetic algorithm (IMGA), a new genetic algorithm, is proposed to select characteristic spectral variables. The new method consists of five steps: gene coding, gene decoding, individual fitness evaluation, gene mutation, and reproductive isolation. The IMGA gives full play to the advantages of the GA and effectively improves the convergence speed of the algorithm. In this paper, the IMGA is applied to select characteristic spectral variables for origin identification of green tea.

Theory. Compared with natural evolution, induced mutation is a random process, but it can produce new species in a short time by artificial selection. Mutation breeding is to artificially expose seeds to chemical or high radiation environment to produce mutants with desirable traits, which is commonly used to obtain new varieties with higher yield, larger size and higher quality. Based on the idea of mutation breeding, IMGA is proposed, but it is quite different from the traditional GA in the process of genetic evolution. By regarding each spectrum as a chromosome, the spectral variables are genes on the chromosomes. The genes carrying green tea origins information are called origin gene, and the others are irrelevant genes. Each gene might be dominant or recessive, and only dominant genes show their traits. Hence, the IMGA is the process of randomly changing each spectral variable from dominant to recessive, or from recessive to dominant. The mutation process and methods will be given detailed description in the following paragraphs. When most of the origin genes become dominant genes and others become recessive genes, a satisfactory mutation results will be got. So far, all the variables corresponding to dominant genes are called characteristic spectral variable, which can be well applied to identify green tea origins. The new spectrum consist of only characteristic spectral variables is called a new species.

The evolution of new species involves three processes: gene mutation, natural selection and reproductive isolation. Gene mutation provides the only raw information for species evolution. In the process of natural selection, the dominant gene is retained in the mutant gene, and the recessive gene will be discarded by reproductive isolation. Up to this point, new species will come into being. After several times of gene mutation, natural selection and reproductive isolation, the new species will only contain origin genes, which has the highest environment adaptability. In this study, genetic evolution is the process of selecting dominant genes related to origin. Genetic inheritance only exists between adjacent species, and there is no inheritance in gene mutation. Therefore, it greatly simplifies the genetic complexity of genetic algorithm and improves the convergence speed. Figure 1 presents the specific process of IMGA, which mainly includes the following steps: gene coding, gene decoding, individual fitness evaluation, gene mutation and reproductive isolation.

Gene coding. In this study, all genes are divided into dominant and recessive ones. Before the genetic evolutionary process, the raw genes are coded by binary numbers to judge the genetic traits. In the genetic evolutionary process, the gene frequency is set to P ($0 \le P \le 1$). After the gene mutation, each gene will be given a random number between 0 and 1. If the random number is greater than P, it is regarded as dominant gene, marked with 1; otherwise, it is recessive gene, labeled with 0. The genes labeled with numbers 1 and 0

indicate that they are selected and not selected, respectively. For example, the genetic code 11010010 indicates that the first, second, fourth, and seventh genes are selected, while the other genes are not selected. Therefore, the probability of a gene being selected can be adjusted by changing the value of gene frequency P. The lager the value of P, the fewer the number of dominant genes. In particular, all genes default to dominant genes when they are firstly encoded.

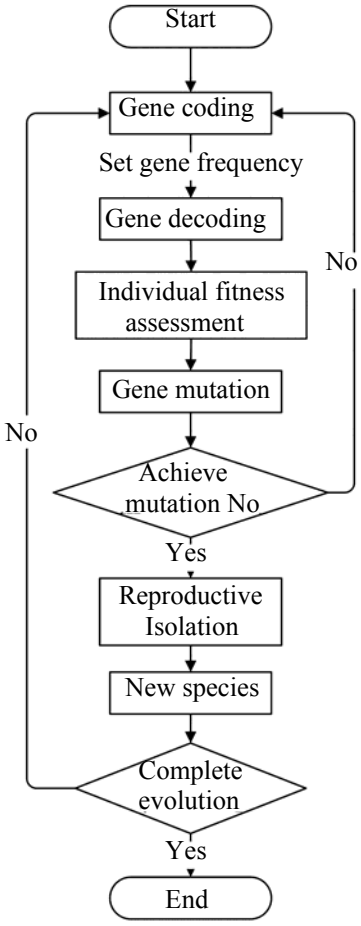

Fig. 1. The program flow chart of IMGA.

Gene decoding. The purpose of gene decoding is to pick out the spectral variables corresponding to dominant genes by referring to the values of random number and gene frequency (*P*). The selected spectral variables will be used for individual fitness evaluation in the next step. Instead, the spectra variables corresponding to recessive genes are considered as irrelevant variables.

Individual fitness evaluation. After gene decoding, we will get a new spectral matrix corresponding to dominant gene. Then it is the time to evaluate the fitness value of the selected spectral variables. All samples are randomly divided into the calibration set and the prediction set in proportion 7:3. After that, the calibration set is applied to build a classification model by PLS to identify the green tea origins. In the process of spectra variables selection, the average values of sensitivity and specificity are used as individual fitness values.

Gene mutation. Gene mutation provides the only raw material for species evolution. By using a random function, a one-dimensional array with the same length as the chromosome is created. Then the genes are encoded by binary encode according to the value of gene frequency. Therefore, each gene has the same probability of being selected as the dominant gene. With each mutation, the new gene are recoded and decoded to calculate the individual fitness value. After a complete optimization cycle, the gene combination with the highest fitness value is passed on to the next generation.

Reproductive isolation. According to the optimized gene combination, the characteristic spectrum variables are selected, and other spectrum variables corresponding to recessive genes are eliminated. As a result, a new spectral matrix is generated, which is considered a new species. The next genetic evolution will take place in new species.

Let us repeat the steps above until the evolution is completed. Finally, the best spectral variable related to the origin of green tea is obtained, which is called characteristic spectral variables.

Materials and methods. Two hundred representative Shandong green tea samples (100 Laoshan green tea samples and 100 Rizhao green tea samples) were collected directly from the origins. All spectra of green tea were collected shortly after production to exclude the effects of storage. For each sample, 30 ± 0.1 g tea leaf was filled into a 200 ml beaker, and the spectrum was collected with a standard diffuse reflection optical fiber probe. The distance between probe and tea leaf was kept at 10 mm. The spectra in the range 1050–2500 nm was collected by an AvaSpec-NIR256/2.5TEC spectrometer (Avantes, Netherlands) in the reflectance. Three spectra were collected for each sample from different places, and each spectrum was the average of 100 scans. For each sample, the average of the three spectra was used in the subsequent analysis. Considering the influence of noise in the edge of spectra, the region of 1300–2300 nm was selected for further analysis. The room temperature was kept at 25°C, and the humidity was kept at an ambient level. The raw spectra are presented in Fig. 2.

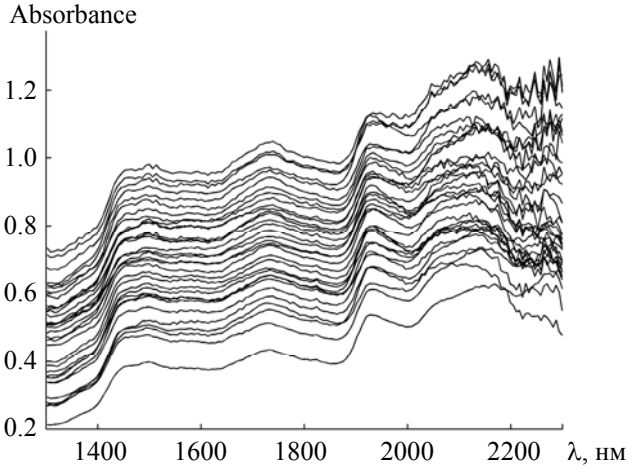

Fig. 2. The NIR spectra of green tea.

In order to identify the origin of green tea, Laoshan and Rizhao Green tea samples were labeled with label 1 (positive) and 2 (negative), respectively. Two hundred samples were randomly divided into the calibration set and the prediction set at a 7:3 ratio.

For the spectra collection, AvaSoft (AvaSpecNIR256/2.5TEC systems) was used. All data analysis was done using self-developed NIR analysis software (ARCO-NIR), which was developed in the MATLAB programming language by MATLAB 2010a (The math works Inc., Natick, MA).

Results and discussion. For the 200 green tea samples, the raw spectrum in the range of 1300–2300 nm has 156 spectral variables at an interval of 6.4 nm. Therefore, the dataset is a spectral matrix with 156 rows and 200 columns. The process of gene coding and decoding is presented in Table 1, which include three spectra and eight variables. First, eight random numbers between 0 and 1 are generated by random function. Then, eight variables are coded according to the dominant gene frequency P (P = 0.5). For the eight random numbers, only four numbers (the second, third, fourth, and seventh) are greater than 0.5. Therefore, the four genes are considered as dominant genes and labeled with 1. After that, the four spectral variables corresponding to dominant genes are selected as a new spectral matrix to build classification models. The decoded spectral matrix is collected in Table 1.

Following the results of gene decoding, a new spectral matrix is created. Then the 200 spectra in the new matrix are randomly divided into calibration set and the prediction set with a ratio of 7:3. Based on the calibration set, the classification model is built by PLS to identify the origins of green tea samples in the prediction set. The average value of sensitivity and specificity is used as the individual fitness value.

TABLE 1. The Process of Gene Coding and Decoding

No.	Random number	Gene	λ (nm)	Spectra		
1	0.012	0	1301.844	0.356	0.338	0.428
2	0.722	1	1308.733	0.359	0.337	0.431
3	0.942	1	1315.618	0.359	0.337	0.432
4	0.845	1	1322.499	0.358	0.337	0.435
5	0.041	0	1329.375	0.362	0.342	0.434
6	0.215	0	1336.248	0.370	0.347	0.443
7	0.671	1	1343.116	0.379	0.351	0.446
8	0.277	0	1363.695	0.399	0.369	0.471

No.	Random number	Gene	λ (nm)	Spectra		
1	0.722	1	1308.733	0.359	0.337	0.431
2	0.942	1	1315.618	0.359	0.337	0.432
3	0.845	1	1322.499	0.358	0.337	0.435
4	0.671	1	1343.116	0.379	0.351	0.446

Every time the fitness value is calculated, the process of gene mutation, coding, and decoding is repeated until the mutation number is reached. In this study, the number of mutations is set at 1000 for a full genetic evolution process. Figure 3 shows the individual fitness evaluation results of 1000 gene mutations during the first genetic evolution. The results show that the 473rd mutant had the highest fitness value, as shown in Fig. 3. Therefore, the spectral variables corresponding to the dominant gene of the 473rd mutant are regarded as the characteristic spectral variables. After reproductive isolation, the selected characteristic spectral variables will be considered as new species and other unrelated variables will be discarded. The next genetic evolution will be based on the new species.

Seventy-three spectral variables are selected in first generation of genetic evolution, as show in Fig. 4a. Based on the above methods, 11 characteristic spectral variables are selected after five genetic evolutions. All the evolution results are collected in Fig. 4. On the basis of the fifth generation of species, further genetic variation did not generate recessive genes, which means that the evolution is over. The 11 characteristic spectral variables are 1472.76, 1499.835, 1513.342, 1540.295, 1785.308, 1856.495, 1977.446, 1996.298, 2033.792, 2120.156, and 2287.833 nm.

The number of PLS components is the most important parameter for the PLS model. In general, too few PLS components cannot sufficiently reflect the relationship between the spectral data and the samples. In turn, too many PLS components will introduce more noise and uncorrelated spectral variables into the model, which will easily lead to overfitting.

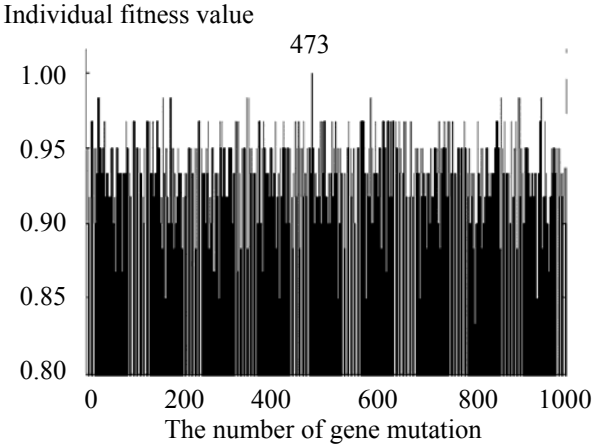

Fig. 3. The individual fitness assessment value of 1000 gene mutations.

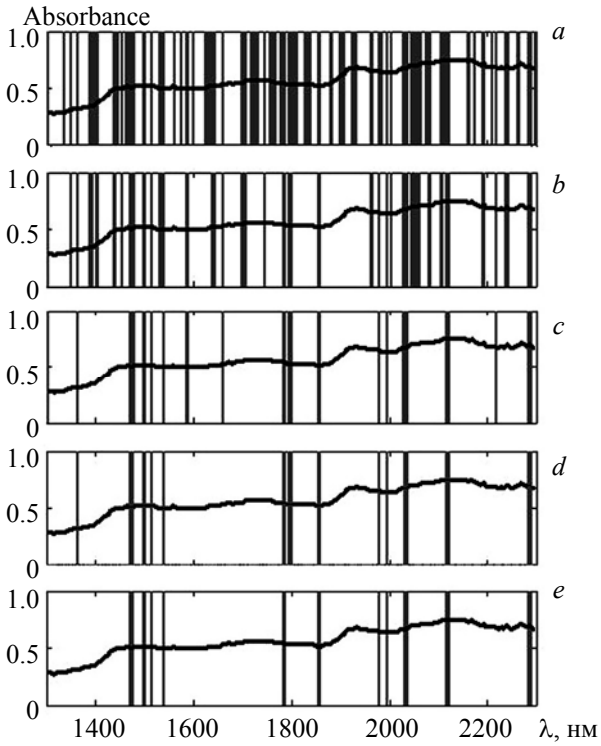

Fig. 4. The selection results of characteristic spectral variables.

In this study, the optimal number of PLS components is determined by *K*-fold cross-validation [18]. Figure 5 presents the value of root mean square error of cross validation (RMSECV) plotted as a function of PLS components. Generally, the smaller the value of RMSECV, the better the calibration model. It can be seen from Fig. 5 that the value of RMSECV decreased significantly with increase in the initial PLS components but slowly began to flatten and climbed up when more components are included. Finally, the optimal number of PLS components is set to 8 when RMSECV gets the minimum value.

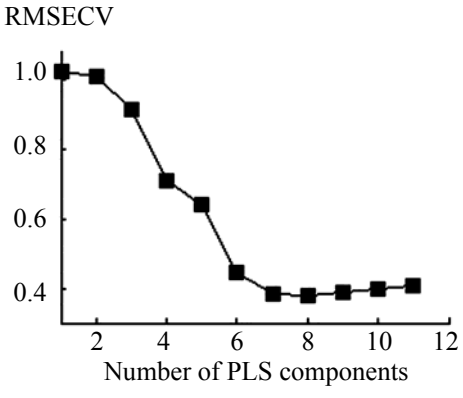

Fig. 5. Effect of PLS components on RMSECV.

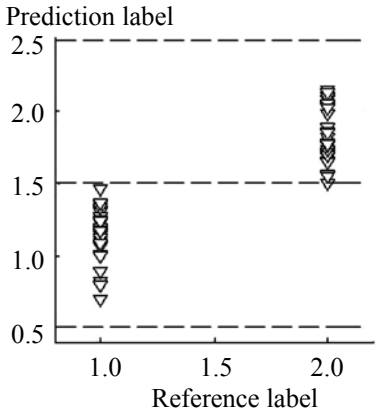

Fig. 6. The prediction results of PLS model.

All the evolution parameters, the number of spectral variables, and modeling results are collected in Table 2, which is the best result of 1000 genetic mutations. Compared with the model before and after evolution, the model variables are reduced from 156 to 11, and the sensitivity and specificity of prediction set are increased from 0.967 and 0.933 to 1, respectively, and kept constant. After selecting the characteristic spectrum variables, the calculation amount is reduced and the modeling speed is greatly improved. Experiments show that most of the characteristic spectral variables can be selected by 1000 gene mutations.

TABLE 2. The Modeling Results and Parameters in the Process of Gene Evolution

No.	Gene frequency	Mutation times	Number of variables	PLS components	Prediction set	
					Sensitivity	Specificity
Original	–	–	156	13	0.967	0.933
1st	0.5	1000	71	15	1	1
2st	0.5	1000	38	12	1	1
3st	0.5	1000	19	10	1	1
4st	0.5	1000	13	8	1	1
5st	0.5	1000	11	10	1	1

Based on the 11 characteristic spectral variables, the calibration model is built by PLS, and the number of PLS components is 10. Figure 6 presents the final prediction results for Laoshan and Rizhao green tea origins. Each triangle represents a sample, plotted by the reference label on the horizontal axis and the prediction label on the vertical. According to the value of the prediction label, the green tea origins will be identified. For example, prediction labels more than 1.5 are considered Rizhao green tea, while the others are considered Laoshan green tea. Using Fig. 6, all samples in prediction set can be identified correctly.

Conclusions. The overall results show that the new induced mutation genetic algorithm (IMCA) is a feasible method for the selection of characteristic spectral variables in the application of NIR spectroscopy. Compared with the traditional genetic algorithms, the IMGA simplifies the process of genetic evolution, increases the probability of genetic variation, and provides the only raw material for species evolution. Therefore, the IMGA greatly improves the efficiency of characteristic spectral variables selection while inheriting the global optimization of genetic algorithms. In addition, the proposed method can be easily extended to other characteristic variables selection problems to improve the prediction ability and the calculation speed.

Acknowledgements. This work has been financially supported by the Key Research and Development Program of Anhui Province (No. 201904a07020073), Foundation of Science and Technology on Electronic Test & Measurement Laboratory (No. 6142001180307), and National Defense Basic Technology of China (No. JSJL2018210C003).

REFERENCES

1. V. R. Sinija, H. N. Mishra, *Food Bioproc. Technol.*, **4**, 136–141 (2011).
2. X. G. Zhuang, L. L. Wang, Q. Chen, X. Y. Wu, J. X. Fang, *Sci. China Technol.*, **60**, 84–90 (2017).
3. Q. S. Chen, J. W. Zhao, H. Lin, *Spectrochim. Acta, A*, **72**, 845–850 (2009).
4. D. Ono, T. Bamba, Y. Oku, T. Yonetani, E. Fukusaki, *J. Biosci. Bioeng*., **112**, 247–251 (2011).
5. Z. M. Guo, Q. S. Chen, L. P. Chen, W. Q. Huang, C. Zhang, C. J. Zhao, *Appl. Spectrosc*., **65**, 1062–1067 (2011).
6. Q. O. Yang, J. W. Zhao, Q. S. Chen, H. Lin, Z. B. Sun, *Anal. Methods*, **4**, 940–946 (2012).
7. M. Vohland, M. Ludwig, S. Thiele-Bruhn, B. Ludwig, *Geoderma,* **223-225**, 88–96 (2014).
8. H. Jiang, G. H. Liu, C. L. Mei, S. Yu, X. H. Xiao, Y. H. Ding, *Anal. Bioanal. Chem.*, **404**, 603–611 (2012).
9. B. M. Smith, P. J. Gemperline, *Anal. Chim. Acta*, **423**, 167–177 (2000).
10. J. Y. Shi, X. P. Yin, X. B. Zou, J. W. Zhao, S. G. Ju, *Chin. Soc. Agric. Mach.*, **41**, 99–103 (2010).
11. M. Ghasemi-Varnamkhasti, S. S. Mohtasebi, M. L. Rodriguez-Mendez, A. A. Gomes, M. C. U. Araújo, R. K. H. Galvão, *Talanta,* **89**, 286–291 (2012).
12. X. G. Zhuang, L. L. Wang, X. Y. Wu, J. X. Fang, *J. Infrared Millimeter Waves,* **35**, 200–205 (2016).
13. J. Jiang, R. J. Berry, H. W. Siesler, Y. Ozaki, *Anal. Chem.,* **74**, 3555–3565 (2002).
14. R. Leardi, A. L. Gonzalez, *Chemometr. Intell. Lab.*, **41**, 195–207 (1998).
15. Q. S. Chen, P. Jiang, J. W. Zhao, *Spectrochim. Acta, A*, **76**, 50–55 (2010).
16. J. H. Holland, *Adaptation in Natural and Artificial Systems*, *MIT Press,* 126–137 (1975).
17. L. Cséfalvayová, M. Pelikan, I. Kralj Cigić, J. Kolar, M. Strlič, *Talanta*, **82**, 1784–1790 (2010).
18. Y. F. Zhai, L. J. Cui, X. Zhou, Y. Gao, T. Fei, W. X. Gao, *Int. J. Remote Sens.*, **34**, 2502–2518 (2013).