Hindawi
Journal of Spectroscopy
Volume 2022, Article ID 2141872, 8 pages
https://doi.org/10.1155/2022/2141872

Research Article

A Variable Selection Method Based on Fast Nondominated Sorting Genetic Algorithm for Qualitative Discrimination of Near Infrared Spectroscopy

Hubin Liu,[1] **Na Liu**,[2] **Yuhui Yuan**,[1] **Cihai Zhang**,[2] **Longlian Zhao**,[1] **and Junhui Li**[1]

[1]*College of Information and Electrical Engineering, China Agricultural University, Beijing 100000, China*
[2]*Technology Center of China Tobacco Guizhou Industrial Co. Ltd., Guiyang 550009, China*

Correspondence should be addressed to Junhui Li; caunir@cau.edu.cn

Received 12 March 2022; Revised 29 May 2022; Accepted 2 June 2022; Published 23 June 2022

Academic Editor: Thomas Walther

A reliable and effective qualitative near-infrared (NIR) spectroscopy discrimination method is critical for excellent model building, yet the performance of models built by these methods is highly dependent on valid feature extraction. The goal of feature selection is to associate the selected variables with the property of interest, which many have done successfully. However, many of selection methods focus only on strong association with the analytes or properties of interest, neglecting correlations between variables. A variable selection method based on a fast nondominated-ranking genetic algorithm (NSGA-II) was proposed in this paper for qualitative discrimination of NIR spectra. The method had two objective functions: (1) maximizing the sum of ratios of interclass variance to intraclass variance, (2) minimizing the sum of correlation coefficients between the selected variables. FT-NIR spectra of a total of 124 tobacco samples from different origins and parts in Guizhou Province, China, were used as the experimental objects, and the part-grade discrimination models of tobacco leaves were established by combining this method with partial least squares-based discriminant analysis (PLS-DA), and compared with PLS-DA model based on the full spectrum. The results showed that the performance of PLS-DA model with the NSGA-II was improved, with a comparable or better correct discrimination rate and reasonable discrimination rate, and could discriminate different parts of the tobacco leaves well. It indicates that the NSGA-II can select a few and effective feature variables to build a high-performance qualitative discrimination model and is proved to be a promising algorithm. In addition, the method is not designed exclusively for spectral data. It is a general strategy that could be used for variable selection for other types of data.

1. Introduction

Near-infrared (NIR) spectroscopy technology is a powerful analysis tool to obtain feature information of the hydrogen-containing groups (O-H, N-H, C-H) in organic substances by recording the NIR spectra of samples, which has the characteristics of speed, accuracy, and destructiveness and has been widely applied for qualitative discrimination in agriculture [1, 2] and food industries [3]. A reliable and effective qualitative near-infrared spectroscopy discrimination method is critical for excellent model building. At present, NIR spectroscopy combined with pattern recognition technique has become an important type of nondestructive discriminant method [4].

The commonly used pattern recognition techniques include partial least squares-based discriminant analysis (PLS-DA) [5], soft independent modeling of class analogy (SIMCA) [6], support vector machines (SVM) [7], and linear discriminant analysis (LDA) [8, 9]. However, the performance of these models highly depends on valid feature extraction and therefore are often combined with variable selection methods to improve performance, such as successive projections algorithms (SPA) [10], uninformative variables elimination (UAE) [11], competitive

adaptive weighted sampling (CARS) [12, 13], and intelligent optimization algorithms [4, 14]. Huang et al. [15] selected two variable selection methods, genetic algorithm and successive projection algorithm, to acquire the feature variables of the spectra, and applied partial least squares-based discriminant analysis and support vector machine algorithms to establish the grading discrimination models of Chinese Dianhong black tea based on NIR spectroscopy. Moreira et al. [16] employed the successive projection algorithm for variable selection and applied linear discriminant analysis to establish the model for cigarette brand classification. The results suggested that the proposed methodology is a promising alternative for assessment of cigarette authenticity. The whole procedure of feature selection is to associate the selected variables with the property of interest, which many have done successfully such as regression coefficient (RC) [17], variable importance in projection (VIP), and the interval PLS (iPLS) [18]. Moreover, elimination of collinearity between variables is also noted by some methods such as successive projection algorithm (SPA), principal component analysis (PCA) loadings [19]. However, many of selection methods focus only on strong association with the analytes or properties of interest, neglecting correlations between variables [20, 21]. In fact, combining the above two objectives may select more effective variables.

A variable selection method based on a fast nondominated ranking genetic algorithm (NSGA-II) was proposed in this paper for qualitative discrimination of NIR spectra, which both correlated the analytes with the selected variables and largely reduced the linear correlation between the selected variables. The method had two objective functions: (1) maximizing the sum of ratios of interclass variance to intraclass variance, (2) minimizing the sum of correlation coefficients between the selected variables. The former was mainly used to select the feature information for sample categories, while the latter was mainly used to eliminate a large amount of redundant information with linear correlation properties. In this study, FT-NIR spectra of a total of 124 tobacco samples from different origins and parts in Guizhou Province, China, were used as the experimental objects, and the part grade discrimination models of tobacco leaves were established by combining this method with PLS-DA, and compared with PLS-DA model based on the full spectrum. It was found that the NSGA-II could select fewer variables and the model built had better performance. The correct discrimination rate and reasonable discrimination rate were comparable or better compared to the model built on full-spectrum, indicating that the algorithm can select effective feature variables to build high-performance qualitative discrimination models and is proved to be a promising variable selection method. It is notable that the method focuses on the designed objective function rather than on the optimization algorithm itself. The fast nondominated ranking genetic algorithm used can be replaced by other multiobjective optimization algorithms such as particle swarm algorithms (PSO) [22]. It should be pointed out that the NSGA-II is not designed exclusively for spectral data. It is a general strategy that can be used for variable selection for other types of data, such as selecting the key chemical components that affect the quality of tobacco leaves.

2. Materials and Methods

2.1. Sample Preparation. A total of 124 raw tobacco leaves of different origins and parts were collected from Guizhou Province, China. Of which, the number of upper tobacco leaves was 33, the number of middle tobacco leaves was 30, and the number of lower tobacco leaves was 30. The samples were baked in an oven at 40°C for about 2 hours and were then cooled to room temperature. To improve efficiency and ensure rapid analysis, the tobacco leaves were ground into powder and passed through a 40-mesh sieve (425 μm) and were placed in the same temperature and humidity environment to collect FT-NIR spectra. Additionally, a fixed weight was placed on top of the sample to ensure that it was naturally pressed.

The contents of chemical components in different parts of the same plant have some differences, which provide a basis for discriminative analysis of tobacco parts by NIR spectroscopy [23]. Hua et al. [24] studied nitrogen accumulation in different parts of tobacco plant using 15 N isotope labeling technique. Results indicated that N content decreased after transplanting, with descending order of suckers > middle leaves > lower leaves > stem and root. Wang et al. [25] determined the contents of reducing sugar, water soluble total sugar, total alkaloids, chlorine, potassium, and total nitrogen of 889 tobacco samples collected from eight counties and cities in Honghe tobacco growing area, and the content of main chemical components in tobacco leaves from different parts, varieties, and production areas were analyzed by cluster analysis. The results showed that the content characteristics of chemical components in upper and middle leaves were similar, while those in lower leaves were different from those in other positions. Moreover, tobacco leaves from different origins may also have some differences [26, 27], affecting the accuracy of part discrimination.

2.2. NIR Spectra Measurements. The data were acquired by Antaris II Fourier transform near infrared spectrometer from Thermo Nicolet. The NIRS instrument was used to record the diffuse reflectance spectra of the samples between the wavenumbers of 10,000–4000 cm^{-1} at 8 cm^{-1} resolution by 64 scans. Each sample was repeatedly measured three times, and the average spectrum was taken as the spectral data of that sample.

2.3. Spectral Data Preprocessing. The raw spectra obtained from FT-NIR spectrometer are easily affected by the physical properties of the sample, background information, and noise interference. A reasonable preprocessing of the raw spectra can reduce the noise information and retain the valid information. First derivative (FD) and Savitzky–Golay (SG) smoothing algorithms [2, 28] were chosen as preprocessing approaches in this study. The FD can eliminate baseline drift

Table 1: Number of calibration and prediction samples in each class.

Class	Calibration set	Prediction set
B	33	10
C	30	12
X	30	9
Total	93	31

Note. B denotes upper tobacco leaf, C denotes middle tobacco leaf, and X denotes lower tobacco leaf.

and smooth out the effects of background interference, providing higher resolution and sharper profile changes in the spectrum than the raw spectrum. However, noise is also amplified, so the SG smoothing algorithm is used to smooth the spectrum to eliminate high-frequency noise.

The samples were divided into calibration and prediction sets by applying the classic Kennard–Stone (KS) uniform sampling algorithm [29] to the FT-NIR spectra. The number of samples from different parts of the tobacco in the calibration and prediction set was presented in Table 1.

2.4. Objective Functions of the Fast Nondominated Sorting Genetic Algorithm. To select feature variables, it is necessary to maintain a large degree of relevance between the selected variables and a single analyte property, while reducing the linear correlation between the variables.

The fundamental idea of Fisher's criterion [8] was applied to evaluate the relevance of NIR spectral features to sample categories. The samples were classified according to different categories of tobacco leaves, and the ratio of interclass variance to intraclass variance was maximized to enable maximizing the interclass distance and minimizing the intraclass distance, so as to select the feature variables with the best discrimination ability in different categories. The objective function was as follows:

$$\max J_F = \frac{\sum_{k=1}^{K} N_k (m_k - m)(m_k - m)^T}{\sum_{k=1}^{K} \sum_{n \in C_k} (x_n - m_k)(x_n - m_k)^T}. \quad (1)$$

where K denotes the number of categories, N_k denotes the number of samples in the k th class, m_k is the mean value of samples in the k th class, and m is the mean value of total samples.

Pearson correlation coefficient [30] was applied to evaluate the correlation between the selected variables. Minimizing the sum of correlation coefficients of all permutations between two variables in the selected variables could make the selected variables more effective by selecting the variables with lower linear correlation. The associated objective function was as follows:

$$\min R_p = \sum_{a=1}^{M-1} \sum_{b=2}^{M} \frac{\sum_{i=1}^{N} \left(x_i^a - \overline{x^a}\right)\left(x_i^b - \overline{x^b}\right)}{\sqrt{\sum_{i=1}^{N} \left(x_i^a - \overline{x^a}\right)^2 \sum_{i=1}^{N} \left(x_i^b - \overline{x^b}\right)^2}} \quad (2)$$

$$(a \neq b).$$

where M is the number of selected variables, N is the number of samples, x_i^a denotes the reflectance of the i th sample at the a th variable, and $\overline{x^a}$ denotes the mean value of reflectance of all samples at the a th variable.

2.5. Process of the Fast Nondominated Sorting Genetic Algorithm. NSGA-II is an evolutionary multiobjective optimization (EMO) methodology based on the Pareto optimal solution theory, which is one of the popular biological heuristics. [31, 32] The algorithm has paid much attention to scholars for its fast convergence, robustness, and better approximation to the real Pareto optimal frontier [33].

To begin this process, the initial population P_0 of size N was randomly generated, and the offspring population Q_0 was generated by three basic operations of selection, crossover, and mutation of genetic algorithm after nondominated sorting. Elite strategy was then introduced, which implied that the parent population and the offspring population were combined for fast nondominated sorting, and each nondominated layer F_i was sorted by applying the crowding distance operator, and the best N individuals were selected to form a new parent population P_{n+1}. Finally, a new offspring population Q_{n+1} was generated by the basic operation of the genetic algorithm, and so on, until the end condition of the process was satisfied.

2.6. Procedure for Determining Optimal Feature Variables. The minimum root mean square error of cross-validation (RMSECV) [1, 34] was used as an evaluation criterion for optimal feature variables. NSGA-II can generate multiple nondominated solutions, which are called Pareto fronts. The multiple nondominated solutions obtained by NSGA-II were, respectively, modeled for discriminant analysis by partial least squares. The formula for RMSECV was as follows:

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-1}}. \quad (3)$$

2.7. Variable Selection Based on CARS. CARS [12, 17] is a new and novel variable selection method based on PLSR and "survival of the fittest," the principle of Darwin's Theory of Evolution. The main feature of this algorithm is the calculation of PLS regression coefficients and reweighted sampling. N subsets of variables were selected by N sampling runs in an iterative manner and finally choose the subset with the lowest RMSECV value as the optimal subset. In each sampling run, CARS works in four successive steps including Monte Carlo model sampling, enforced wavelength reduction by EDF, competitive wavelength reduction by ARS, and RMSECV calculation for each subset. Of these, EDF-based wavelength reduction in combination with competitive wavelength reduction by ARS is a two-step procedure for wavelength selection.

2.8. Partial Least Squares-Based Discriminant Analysis. PLS-DA is a linear discrimination method based on PLS regression is widely applied for qualitative analysis [6, 35]. To achieve the classification, a PLS regression model was established between the matrix of independent variables (X) and the matrix of dependent variables (Y). Y was coded in a

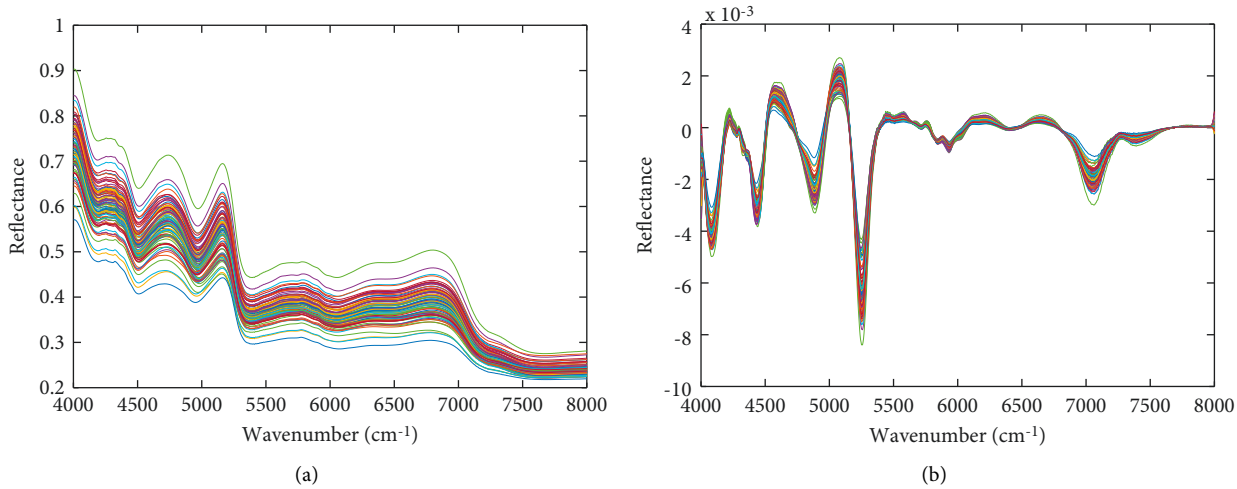

Figure 1: (a) Original and (b) preprocessed FT-NIR spectra of the samples. First derivative and Savitzky–Golay smoothing algorithms were chosen as preprocessing approaches.

binary way (1 or 0), where 1 indicates that the sample belongs to the class and 0 indicates that it does not. In this study, tobacco leaves were classified in the upper (B), middle (C), and lower (*X*) part categories; therefore, B, C, and *X* were coded in the form of 1.0.0, 0.1.0, and 0.0.1, respectively [4, 5]. The predicted values were then obtained by the PLS–DA model, and Bayesian statistics were adopted to calculate the classification threshold between categories and identify the category of each sample based on the calculated threshold value of each class. Finally, the class of each sample was identified based on the PLS-DA model.

2.9. Performance Evaluation. To evaluate the performance of the classification model, correct discrimination rate (CDR), reasonable discrimination rate (RDR), sensitivity (Sen), and specificity (Spe) were considered. The CDR was calculated referring to the ratio of the number of correct identified samples to the number of total samples. The RDR was the ratio of the number of samples correctly identified and identified as adjacent categories to the number of total samples. One of the reasons RDR was considered as a performance evaluation was the continuity of tobacco growth, which made it difficult to discriminate which category the junction between the upper and middle of the tobacco and the junction between the middle and lower of the tobacco belonged to. Therefore, the number of samples correctly identified as adjacent categories were both considered reasonable. Sensitivity (Sen) is the proportion of positives that are identified as such. Specificity (Spe) is the proportion of negatives that are correctly identified as such. The higher the values of these parameters are, the better the performance of the classification model is.

3. Results and Discussion

3.1. NIR Spectra. After a preliminary inspection of the spectra, those regions in which the detector was saturated or the signal-to-noise ratio was poor were discarded. As a result, the 8,000–4000 cm^{-1} interval was selected for the study. Figure 1(a) presents the raw FT-NIR spectra of the 124 tobacco samples in the range of 8,000–4000 cm^{-1}. As can be seen, the spectra are noisy and display systematic variations in the spectral baseline. These problems were circumvented by applying the FD and the SG with a 15-point window, as shown in Figure 1(b). Each spectrum had 1036 wavenumber variables.

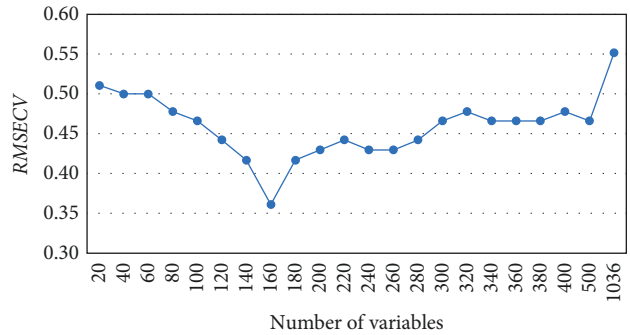

Figure 2: Variation of RMSECVs with the number of selected variables. RMSECV denotes root mean square error of cross-validation.

3.2. Variable Selection. To determine that the number of retained variables (*N*) has a significant effect on model performance which decides the stability and accuracy of the model. When the number of retained variables is too small, the robustness and accuracy of the model may be affected due to the loss of key informative variables. On the contrary, if the number of retained variables is too large, uninformative variables may be contained in the model and make its performance poor. The NSGA-II was applied for variable selection and the variation of the calibration set RMSECV with variable number *N* was investigated. The population size of the NSGA-II was set to 200, the number of iterations was set to 150, and the mutation probability and crossover

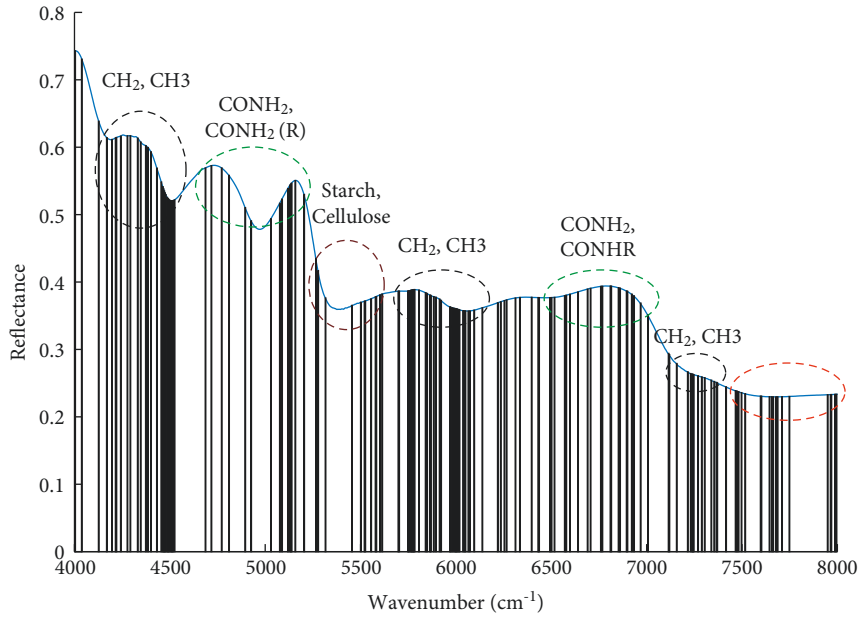

Figure 3: Results of variable selection obtained by the NSGA-II. The selected variables were mainly clustered in six chemically meaningful wavenumber bands and the band circled in red dashed lines was discussed.

probability were set to 0.02 and 0.8, respectively. The 10-fold crossover validation was employed to validate the performance of PLS-DA classification model. Figure 2 shows that the RMSECV was obtained with the number of variables N from 20 to 400 and a step of 20, and it was compared with the number of variables 500 and the full variables.

It can be seen that the RMSECV was large at the beginning and decreased rapidly with the increase of N. Apparently, the RMSECV reached the lowest at N of 160. After that, the RMSECV gradually increased with the increase of N, with certain fluctuations. It is indicated that fewer feature variables are beneficial for improving the model performance. However, the RMSECV also increased for too few variables, which were due to the fact that useful wavenumbers cannot be completely included, resulting in poor model quality. In contrast, irrelevant variables also affected the prediction results when more invalid variables were used. Therefore, $N = 160$ is used for further study. Notably, the RMSECVs of the PLS-DA with variable selection, although having some fluctuations, were both lower than the RMSECVs of the PLS-DA without variable selection, indicating that the NSGA-II can select effective feature variables and improve the model performance.

Figure 3 shows the results of variable selection obtained by the NSGA-II. It can be seen that most of the 160 selected variables were located in the absorption peaks of FT-NIR spectra, which were mainly clustered in six chemically meaningful wavenumber bands under 7370–7115 cm^{-1}, 7007–6397 cm^{-1}, 6138–5698 cm^{-1}, 5613–5266 cm^{-1}, 5158–4687 cm^{-1}, 4532–4170 cm^{-1}, which were corresponding to stretching and bending vibrations of C-H groups in the first overtone region, stretching vibrations of N-H groups and O-H groups in the first overtone region, stretching vibrations of C-H groups in the first overtone region, stretching vibrations of O-H groups in the combination and stretching vibrations of C-O groups in the first overtone region, symmetric and asymmetric stretching vibrations of N-H groups in the combination, stretching and bending vibrations of C-H groups in the combination, respectively. It is suggested that the selected variables contained key variables of chemical significance, indicating further the validity of the proposed method. It is worth noting that some of the variables were also selected in the 7995–7467 cm^{-1} band (already circled in red dashed lines), but there is almost no absorption of groups in this band. The possible reason was that some of the instrument-generated noise was selected and sometimes a certain amount of noise contributes to the stability of the model.

3.3. Comparison of the Classification Results from Different Variables Selection Methods. Table 2 shows the classification results of the full-spectrum-based PLS-DA, NSGA–II–PLS-DA, and CARS-PLS-DA. Compared with those of the full-spectrum-based PLS-DA model, the CCR, RDR, Sen, and Spe of the calibration set and prediction set in the NSGA–II–PLS-DA model have been increased obviously. Of which, the RDRs of both calibration and prediction sets reached 100%, indicating that the failure of assigning a reasonable category to a sample was solved. To further comparison of the performance with CARS-PLS-DA, the RDRs of calibration and prediction sets are 98.92% and 100%, respectively were and also adopted to identify the part grade. Though high reasonable discrimination rate was achieved, the classification performance of these models was still lower than that of the NSGA–II–PLS-DA model (the RDRs of both calibration and prediction sets are 100%), further validating the capability of the NSGA–II–PLS-DA model, indicating that the selected few variables by this

Table 2: Classification results of PLS-DA model based on different variables selection methods.

Model	Number of variables	Class	Calibration set				Prediction set			
			Sen	Spe	CDR (%)	RDR (%)	Sen	Spe	CDR (%)	RDR (%)
PLS-DA	1036	B	0.879	0.917	79.57	96.77	0.800	0.905	77.42	96.77
		C	0.667	0.905			0.583	0.947		
		X	0.833	0.873			1.000	0.818		
NSGA-II-PLS-DA	160	B	0.970	0.933	87.10	100	0.900	0.857	80.65	100
		C	0.633	0.984			0.583	0.947		
		X	1.000	0.889			1.000	0.909		
CARS-PLS-DA	91	B	0.970	0.917	84.95	98.92	1.000	0.857	80.65	100
		C	0.633	0.968			0.583	0.947		
		X	0.933	0.889			0.889	0.909		

Note "Sen," "Spe," "CDR," and "RDR" denote sensitivity, specificity, correct discriminant rate, and reasonable discriminant rate, respectively.

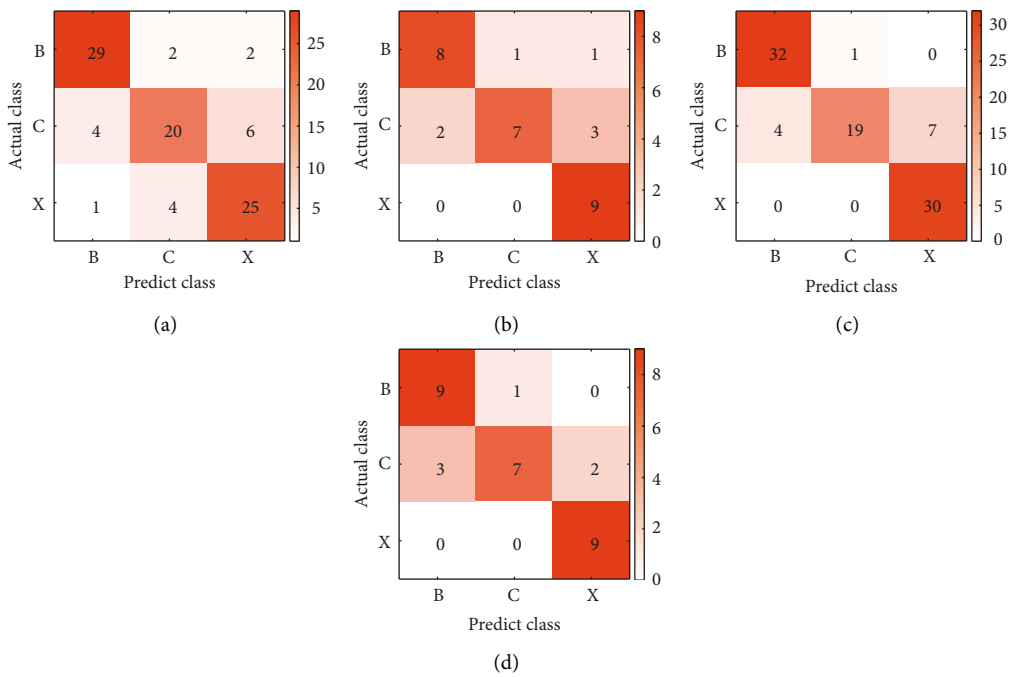

Figure 4: Confusion matrix description results of PLS-DA model and NSGA-II-PLS-DA model in the modeling and prediction process: (a) Calibration and (b) Prediction of PLS-DA model. (c) Calibration and (d) Prediction of NSGA-II-PLS-DA model.

method can well discriminate different parts of tobacco (upper, middle, and lower).

Figure 4 shows the results of the confusion matrix for the full-spectrum-based PLS-DA and NSGA–II–PLS-DA in the modeling and prediction process. It can be seen that the NSGA–II–PLS-DA calibration set and prediction set had a great improvement in the number of correct discriminations for the upper and lower tobacco leaves compared with the full-spectrum-based PLS-DA, in which only two samples of upper tobacco were misjudged as the middle tobacco leaves, and there was no false discrimination for the lower tobacco leaves. It is suggested that the selected variables can improve the performance of the model and enhance the discrimination for tobacco parts, further indicating that the proposed method is a promising variable selection method for qualitative discrimination. Notably, the misclassified samples were mainly distributed between adjacent classes, and more middle tobacco leaves were judged as upper and lower tobacco leaves, which may be due to differences in fertilization and plant densities of tobacco origins lead to some differences in the same tobacco part, or may be due to the continuity of tobacco growth, making the correct discrimination rate of the middle tobacco leaves lower.

4. Conclusions

A variable selection method based on a fast nondominated ranking genetic algorithm was proposed in this paper for the qualitative discrimination of NIR spectra. The method selected variables that satisfy the objective function by the fast nondominated sorting genetic algorithm, which maximized the interclass variance and minimized the intraclass variance

while minimizing the correlation between the selected variables. The key variables were selected for the feature information of the sample classes, and a large amount of redundant information with linear correlation property was eliminated. Combining this method with PLS-DA to build a discrimination model of tobacco parts and comparing it with a full-spectrum partial least squares-based discrimination analysis model, the results showed that the algorithm can select a few and effective feature variables to improve the model performance, which can discriminate well between different parts of tobacco and obtain better classification results. The algorithm is demonstrated to be a promising algorithm for wavelength selection to build high-performance qualitative discriminant models.

It should be noted that the variable selection method proposed in this paper focuses on the designed objective function rather than the optimization algorithm, and NSGA-II can be replaced by other multiobjective optimization algorithms such as particle swarm algorithm (PSO). Moreover, the method is not designed only for spectral data. It is a general strategy that can be used for the variable selection of other types of data, such as the selection of key chemical components that affect the quality of tobacco leaves. In addition, it can be used to build qualitative discrimination models for different years, regions, or grades of tobacco, all of which are important for improving the quality of tobacco. Our future work will focus on these aspects.

Data Availability

The spectral data used to support the findings of this study are included within the supplementary materials.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported by the National Key Research and Development Program (Grant no. 2016YFD0700304) and Extraction and Application of Tobacco Quality Information Based on Near Infrared Cloud-Analysing Platform (Grant No. GZZY/KJ/JS/2017BY013-1).

Supplementary Materials

The spectral data: in the data file, column A gives tobacco part classification, column B gives sample ID, and the other columns show the spectral data. The first row presents the wavenumber value. (*Supplementary Materials*)

References

[1] V. Bisutti, R. Merlanti, L. Serva et al., "Multivariate and machine learning approaches for honey botanical origin authentication using near infrared spectroscopy," *Journal of Near Infrared Spectroscopy*, vol. 27, no. 1, pp. 65–74, 2019.

[2] L. Luan, Y. Wang, X. Li et al., "Application of multiple classifier fusion in the discriminant analysis of near infrared spectroscopy for agricultural products," *Journal of Near Infrared Spectroscopy*, vol. 24, no. 4, pp. 363–372, 2016.

[3] Y. Liu, Z. Xia, L. Yao et al., "Discriminating geographic origin of sesame oils and determining lignans by near-infrared spectroscopy combined with chemometric methods," *Journal of Food Composition and Analysis*, vol. 84, Article ID 103327, 2019.

[4] J. Zeng, Y. Guo, Y. Han et al., "A review of the discriminant analysis methods for food quality based on near-infrared spectroscopy and pattern recognition," *Molecules*, vol. 26, no. 3, Article ID 749, 2021.

[5] X. Chen, Y. Xu, L. Meng et al., "Non-parametric partial least squares-discriminant analysis model based on sum of ranking difference algorithm for tea grade identification using electronic tongue data," *Sensors and Actuators B: Chemical*, vol. 311, Article ID 127924, 2020.

[6] A. Biancolillo, P. Firmani, R. Bucci, A. Magrì, and F. Marini, "Determination of insect infestation on stored rice by near infrared (nir) spectroscopy," *Microchemical Journal*, vol. 145, pp. 252–258, 2019.

[7] C. Li, L. Li, Y. Wu, M. Lu, Y. Yang, and L. Li, "Apple variety identification using near-infrared spectroscopy," *Journal of Spectroscopy*, vol. 2018, Article ID 6935197, 7 pages, 2018.

[8] V. Elias de Almeida, D. Douglas de Sousa Fernandes, P. Henrique Gonçalves Dias Diniz et al., "Scores selection via Fisher's discriminant power in pca-lda to improve the classification of food data," *Food Chemistry*, vol. 363, Article ID 130296, 2021.

[9] K. C. Kaufmann, K. A. Sampaio, J. F. García-Martín, and D. F. Barbin, "Identification of coriander oil adulteration using a portable nir spectrometer," *Food Control*, vol. 132, Article ID 108536, 2022.

[10] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame, and V. Visani, "The successive projections algorithm for variable selection in spectroscopic multicomponent analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 57, no. 2, pp. 65–73, 2001.

[11] W. Cai, Y. Li, and X. Shao, "A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 90, no. 2, pp. 188–194, 2008.

[12] H. Li, Y. Liang, Q. Xu, and D. Cao, "Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration," *Analytica Chimica Acta*, vol. 648, no. 1, pp. 77–84, 2009.

[13] L. Li, S. Quan, D. Li, J. Wang, H. Zang, and L. Zhang, "Development of near infrared spectroscopy methodology for human albumin determination using a new calibration approach," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 217, pp. 256–262, 2019.

[14] F. Allegrini and A. C. Olivieri, "A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis," *Analytica Chimica Acta*, vol. 699, no. 1, pp. 18–25, 2011.

[15] J. Huang, G. Ren, Y. Sun et al., "Qualitative discrimination of Chinese dianhong black tea grades based on a handheld spectroscopy system coupled with chemometrics," *Food Sciences and Nutrition*, vol. 8, no. 4, pp. 2015–2024, 2020.

[16] E. D. T. Moreira, M. J. C. Pontes, R. K. H. Galvão, and M. C. U. Araújo, "Near infrared reflectance spectrometry classification of cigarettes using the successive projections

algorithm for variable selection," *Talanta*, vol. 79, no. 5, pp. 1260–1264, 2009.

[17] M. Kamruzzaman, D. Kalita, T. Ahmed, G. Elmasry, and Y. Makino, "Effect of variable selection algorithms on model performance for predicting moisture content in biological materials using spectral data," *Analytica Chimica Acta*, vol. 1202, Article ID 339390, 2022.

[18] M. L. da Silva Medeiros, J. P. Cruz-Tirado, A. F. Lima et al., "Assessment oil composition and species discrimination of brassicas seeds based on hyperspectral imaging and portable near infrared (nir) spectroscopy tools and chemometrics," *Journal of Food Composition and Analysis*, vol. 107, Article ID 104403, 2022.

[19] D. F. Barbin, A. T. Badaró, D. C. B. Honorato, E. Y. Ida, and M. Shimokomaki, "Identification of Turkey meat and processed products using near infrared spectroscopy," *Food Control*, vol. 107, Article ID 106816, 2020.

[20] C. Pasquini, "Near infrared spectroscopy: a mature analytical technique with new perspectives—a review," *Analytica Chimica Acta*, vol. 1026, pp. 8–36, 2018.

[21] Y.-H. Yun, H.-D. Li, B.-C. Deng, and D.-S. Cao, "An overview of variable selection methods in multivariate analysis of near-infrared spectra," *TRAC Trends in Analytical Chemistry*, vol. 113, pp. 102–115, 2019.

[22] F. Marini and B. Walczak, "Particle swarm optimization pso. a tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 153–165, 2015.

[23] J.-P. Laclau, J. C. R. Almeida, J. L. M. Goncalves et al., "Influence of nitrogen and potassium fertilization on leaf lifespan and allocation of above-ground growth in eucalyptus plantations," *Tree Physiology*, vol. 29, no. 1, pp. 111–124, 2008.

[24] D. Hua, S. Zhang, R. Wang, G. Huo, and W. Liu, "Accumulation of nitrogen from soil and ~15n-labeled fertilizer in different organs of flue-cured tobacco at different growth stage," *Acta Tabacaria Sinica*, vol. 19, no. 1, pp. 32–36, 2013.

[25] C. Wang, X. Chen, Y. Ruan et al., "Content characteristics of main chemical components in tobacco leaves in honghe tobacco growing area," *Southwest China Journal of Agricultural Sciences*, vol. 33, no. 12, pp. 2793–2799, 2020.

[26] D. F. Barbin, R. P. Sobottka, W. E. Risso, C. Zucareli, and E. Y. Hirooka, "Influence of plant densities and fertilization on maize grains by near-infrared spectroscopy," *Spectroscopy Letters*, vol. 49, no. 2, pp. 73–79, 2016.

[27] J. Olsen and J. Weiner, "The influence of triticum aestivum density, sowing pattern and nitrogen fertilization on leaf area index and its spatial variation," *Basic and Applied Ecology*, vol. 8, no. 3, pp. 252–257, 2007.

[28] A. Rinnan, F. V. D. Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TRAC Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009.

[29] H. Li, J. X. Wang, Z.-N. Xing, and G. Shen, "Influence of improved kennard/stone algorithm on the calibration transfer in near-infrared spectroscopy," *Spectroscopy and Spectral Analysis*, vol. 31, no. 2, pp. 362–365, 2011.

[30] H. Xu and Y. Deng, "Dependent evidence combination based on shearman coefficient and pearson coefficient," *IEEE Access*, vol. 6, pp. 11634–11640, 2018.

[31] A. Konak, D. W. Coit, and A. E. Smith, "Multi-objective optimization using genetic algorithms: a tutorial," *Reliability Engineering and System Safety*, vol. 91, no. 9, pp. 992–1007, 2006.

[32] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: nsga-ii," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[33] J. M. Lourenco and L. Lebensztajn, "Post-pareto optimality analysis with sum of ranking differences," *IEEE Transactions on Magnetics*, vol. 54, no. 8, pp. 1–10, 2018.

[34] S. Xu, Y. Zhao, M. Wang, and X. Shi, "Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by vis-nir spectroscopy," *Geoderma*, vol. 310, pp. 29–43, 2018.

[35] L. Meng, X. Chen, X. Chen et al., "Linear and nonlinear classification models for tea grade identification based on the elemental profile," *Microchemical Journal*, vol. 153, Article ID 104512, 2020.