Hindawi
Journal of Spectroscopy
Volume 2022, Article ID 2190893, 9 pages
https://doi.org/10.1155/2022/2190893

Research Article

A Progressive Combined Variable Selection Method for Near-Infrared Spectral Analysis Based on Three-Step Hybrid Strategy

Hongmin Sun,[1] Fanze Kong,[1] Cheng Xiu,[2] Weizheng Shen,[1] and Yan Wang[1]

[1]*College of Electrical and Information, Northeast Agricultural University, Harbin 150030, China*
[2]*The 49TH Research Institute of China Electronics Technology Group Corporation, Harbin 150028, China*

Correspondence should be addressed to Yan Wang; wangyan_neau@126.com

Received 10 March 2022; Revised 11 April 2022; Accepted 25 April 2022; Published 10 May 2022

Academic Editor: Daniel Cozzolino

A specific variable selection method was proposed based on a three-step hybrid strategy for near-infrared spectral analysis. By analyzing functions of each step and characteristics of various variable selection methods, synergy interval partial least squares, iterative variable subset optimization, and bootstrapping soft shrinkage were chosen for three steps. To test the effect of the three-step hybrid method, it was applied to corn and soil spectral data and compared to other common methods. Results for oil content in corn data showed that the three-step hybrid variable selection method selected 1% variables of full spectrum, calibration determination coefficient, and prediction determination coefficient reached 0.998 and 0.993 where the explained variance was increased by 27.30%. It could effectively extract variables related to the tested substance and provide a new variable selection method for near-infrared spectral analysis.

1. Introduction

Near-infrared spectrum analysis technology has been continuously studied and developed since it was recognized in the 1930s. Its advantages are fast analysis speed, simple instrument operation, no pollution to samples, low cost, and high accuracy. And has achieved satisfactory results in the sample detection in many fields [1–5]. Variable selection is necessary and important for the accuracy and stability of quantitative analysis model by near-infrared spectrum [6]. The meaning of stability is generally different on different occasions. Here, standard deviation and root mean square error were used to explain. Through the variable selection method, a small amount of variables or wavelengths are used to represent the whole spectrum to participate in the modeling. The model is simplified and time-consuming is reduced to improve efficiency. More importantly, the influence of irrelevant or nonlinear variables is eliminated, to obtain a model with stronger stability, interpretability, and prediction ability.

Many variable selection algorithms appear and perform better effects. Some methods choose variables by parameters from the partial least squares (PLS) model. Algorithms based on this idea include competitive adaptive reweighted sampling (CARS) and iterative variable subset optimization (IVSO) [7,8]. Some algorithms select variables based on the model cluster analysis strategy. Variable combination population analysis (VCPA), iteratively retaining informative variables (IRIV), and bootstrapping soft shrinkage (BOSS) are typical algorithms [9,10]. Some methods do not select wavelength points but choose wavelength interval. Mainly including interval partial least squares (iPLS) and synergy interval partial least squares (SiPLS) [11]. Strategy combined with two or several variable selection methods were focused on and performed better effects, especially the three-step hybrid strategy [12–15]. This strategy can take advantage of multiple methods and extract useful variables by rough selection, fine selection, and optimal selection.

Based on the advantage of the three-step hybrid strategy proposed by Yu et al [15], variable selection methods with

different principles were analyzed, and algorithms were determined for each step. The SiPLS-IVSO-BOSS variable selection method was proposed in this paper, which was introduced by analyzing the oil spectral data of corn in detail. And the verification was carried out on the starch spectral data of corn and organic matter spectral data of soil. Results of SiPLS-IVSO-BOSS were compared with CARS, VCPA, IRIV, single method from them, and two methods among them. CARS, VCPA, and IRIV were used to replace one of the algorithms of the combination strategy to illustrate the irreplaceability of the three algorithms in the strategy.

2. Materials and Methods

2.1. Combined Variable Selection Method. In the process of variable selection, the usual one-step method is difficult to meet the needs of models in practical applications. The three-step hybrid strategy can optimize the variables step by step from different angles. Whether this idea applies to many newly developed and not widely used algorithms is worth studying. Therefore, based on this strategy and some proposed three-step hybrid variable selection methods, this study selected two algorithms, IVSO and BOSS, to explore whether the superiority of their separate modeling could be reflected and improved after mixing. The most important thing about the three-step hybrid strategy is that the first two steps should appropriately reduce the variable space, in which the first step generally retains 10% to 20% of the variables, and it is necessary to use the band selection method. The second step continues to extract to about 5%, and it is suitable to select the algorithm that can appropriately control the number of variables based on continuous optimization of the variable space in the wavelength selection method. The last step is to compress the variables to dozens or even several, which is enough to build a model that can accurately predict the unknown content. SiPLS equalizes the wavelength range and can select 2 to 4 subintervals. The continuity of the wavelength in the interval makes the variables continue to be optimized. Due to its warmth, IVSO usually retains the number of variables at the same level as the band selection method, so it is suitable for the second step, which can provide a sufficient number of variables for the third step through different optimization ideas. BOSS can highly compress the variable space, but when used alone, the sampling process is cumbersome and the running time is long, which requires the relatively stable variable space given by the previous algorithm. These three algorithms have three-step hybrid strategy implementation conditions in theory. Therefore, the SiPLS-IVSO-BOSS variable selection method was proposed based on a three-step hybrid strategy, to select variables for near-infrared spectrum analysis. It included rough selection, fine selection, and optimal selection. Results of the former step would be the basis of the next step, so the first step becomes important. The rough selection should include enough variables related to the tested substance so that the later method could effectively improve the performance of the model and avoid the risk of falling into overfitting or variable selection being wrong repeatedly.

2.2. Rough Selection Method. The rough selection was the first step and basis of the fine selection. It will directly affect the final variable selection result. In order to extract variables related to the tested substance as many as possible, there should be including enough variables with strong explanatory power, greatly reducing the variable space, and retaining enough important information variables for the next selection after rough selection. Interval selection methods select a group of variables by dividing wavelength points into several groups. This kind of algorithm will be fitful for rough selection. Here, SiPLS and the most basic iPLS method in the interval selection algorithm are compared. SiPLS selects a combination of multiple interval intervals, while iPLS selects a single interval. Compared with iPLS, when the variables related to the tested substance are not continuous, the performance of SiPLS will be better than iPLS. SiPLS considers the combination effect and combines the intervals instead of simply selecting several local optimal intervals. So SiPLS was used as the first step for rough selection in this paper.

2.3. Fine Selection Method. The second step was to further select variables based on SiPLS. Redundant information in variables, noninformation, and weak information variables need to be filtered out from the wavelength interval. Strong information variables need to be retained. The variable space needs to be optimized and narrowed to achieve the fine selection of variables, providing sufficient quantities of variables for the next step. IVSO uses PLS regression coefficient to represent the importance of variables and uses weighted binary matrix sampling (WBMS) and sequential addition to eliminate useless information variables in a competitive manner. It can gently eliminate the redundancy caused by continuity in the selected bands of SiPLS and retain sufficient variables for the implementation of the next algorithm. IVSO meets the above objectives, so it was used for fine selection.

2.4. Optimal Selection Method. After the first two steps of selection, most noise and interference variables had been filtered, just to solve the continuity between them. By calculating the regression coefficients of multiple submodels, BOSS determines the weight of variables. The greater the weight, the greater the probability of being selected. Weighted guided sampling (WBS) was used to optimize the weight and simplify the shrinkage variable space. The optimal variables were determined by extracting the variable set with the minimum root mean square error of cross-validation (RMSECV) in the submodel. BOSS can remove the collinearity between variables, which can find the optimal variable combination in the variable space with a small wavelength span. So it was a better choice.

The core idea of the SiPLS-IVSO-BOSS method is shown in Figure 1. The variable selection method can accurately locate the relevant information variables from the complex

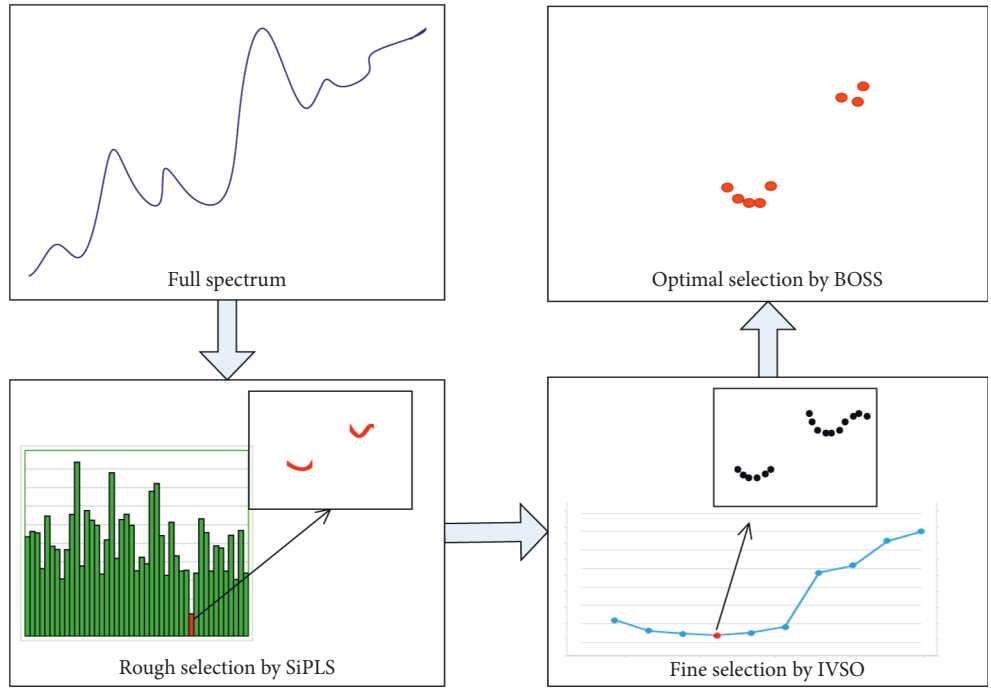

Figure 1: Process of variable selection method combined synergy interval partial least squares, iterative variable subset optimization, and bootstrapping soft shrinkage. Several variable groups were firstly collected in rough selection by synergy interval partial least squares. Then some variables were extracted in fine selection by iterative variable subset optimization. Finally, less variables were retained in optimal selection by bootstrapping soft shrinkage.

and huge variable space and realize the rapid and efficient processing of data. After analyzing the effectiveness of the method, in theory, the actual effect was verified and compared with other algorithms.

2.5. Data Set. Near-infrared spectral data for corn published by Eigenvector Research, Inc. was used in this paper. There were 80 samples with 700 wavelength variables in the range of 1,100 to 2,498 nm in the corn data set. And soil data set was also used; it contained 108 samples and 1,050 wavelength points in the range of 400 to 2,500 nm [16].

Spectral data usually need to be preprocessed before modeling. Preprocessing can remove background noise in the spectrum and interference of specific physical factors and generally improve the correlation between spectrum and chemical components. The existing researches show that it has a good processing effect on the corn data set [17,18]. Centralization eliminates the adverse effects of large-scale differences. Therefore, this article used centralization to preprocess the corn and soil data. Kennard–Stone (KS) algorithm was used to divide the data set as training set and test set with 3:1 [19]. The determination coefficient (R^2), root mean square error of calibration (RMSEC), and root mean square error of prediction (RMSEP) of PLS models were used to compare the effects of different variable selection algorithms. In this paper, the prediction of oil content in corn was analyzed in detail, and the content prediction of the other two substances gave only modeling results. All calculations were carried out in MATLAB R2014a (MathWorks, Inc.).

3. Results

This section discussed the verification results of oil spectral data of corn, and the implementation process of the SiPLS-IVSO-BOSS was introduced in detail. Centralization was used for spectral preprocessing. PLS was used to construct a quantitative analysis model. The results were compared with other variable selection methods. In addition, to further check its practicality and stability, starch spectral data of corn and organic matter of soil spectral data were researched and only gave the prediction results.

The 80 corn samples were divided into 60 training set samples and 20 test set samples using the KS algorithm. Table 1 lists the statistical results of oil and starch content in corn and organic matter content in the soil. It can be seen that the oil content of the test set was within the range of the training set, indicating that the sample set used for modeling could well represent the overall sample.

3.1. Rough Selection by Synergy Interval Partial Least Squares. When using SiPLS to screen the full spectrum bands, the whole spectrum was equally divided into 10, 20, 30, and 40 intervals. Based on different intervals, 2 to 4 subintervals were combined to establish PLS models. The calculation time for each interactive verification was 0.5 to 1 minute when the number of intervals was small; while the model verification time increased exponentially with the increase of the variables, the execution time of the 4 subintervals increased significantly when the interval number was 30. When the interval number was 40 and 3 subintervals were combined,

Table 1: Statistical results of oil and starch content in corn samples and organic matter content in soil samples.

Substance	Sample set	Number of samples	Minimum	Maximum	Mean	Standard deviation
Oil in corn	Training set	60	3.088	3.832	3.482	0.187
	Test set	20	3.264	3.766	3.547	0.136
	Total set	80	3.088	3.832	3.498	0.177
Starch in corn	Training set	60	62.826	66.472	64.737	0.812
	Test set	20	63.021	65.808	64.571	0.856
	Total set	80	62.826	66.472	64.696	0.821
Organic matter in soil	Training set	81	42.910	95.850	84.768	11.853
	Test set	27	64.610	93.010	87.408	6.656
	Total set	108	42.910	95.850	85.428	10.822

The training set and the test set were divided as 3:1 ratio by Kennard–Stone algorithm.

the number of model calculations reached 27,405; the number of model calculations reached 91,300 for 4 combinations; and the calculation time surged to 24 minutes. Therefore, this experiment did not count the variable optimization results under the condition of the interval number that was 40. The joint interval with the smallest root mean square error (RMSE) value was selected through interactive verification.

Table 2 lists the results of different intervals and combinations. The bands of 5 and 9 subintervals and 9, 10, 17, and 18 subintervals selected by the total number of 10 and 20 intervals were completely consistent (denoted as band I). The RMSE was second to the bands of 13, 14, 25, and 26 subintervals selected by the total number of 30 intervals (denoted as band II), and the difference was about 16.6%. However, band I completely contained band II and covered more variables; coverage increased by 34.3%, which can theoretically provide more information for further wavelength selection and help improve the diversity of selection. Considering both the error and coverage, the bands 1,660 to 1,798 nm and 2,220 to 2,358 nm were selected as the initial variable sets for subsequent variable selection.

The position of the selected subinterval in the whole spectrum using the 10 intervals is shown in Figure 2. The selected wavelength points were 140, accounting for 20% of the whole spectrum. SiPLS effectively eliminates a large number of useless information and retains enough variables for the implementation of the SiPLS-IVSO-BOSS method.

3.2. Fine Selection by Iterative Variable Subset Optimization. The characteristic bands selected by SiPLS screened the interference information to a certain extent and selected the bands with relatively rich information. However, due to the limitation of continuous wavelength in the interval, it still contained redundant information. The selected band I was further screened by IVSO, and the parameters were set as follows: the maximum latent variable number of cross-validation was set to 10, the cross-validation fold was set to 10, and the WBMS number was set to 1000. The number of cycles was set to 50. When the number of variables sampled by WBMS was equal to the number of combinations of variable subsets, the cycle stopped, and the variable subset with the lowest error value was selected as the selection result of IVSO.

Fine selection results by IVSO are shown in Figure 3. Figure 3(a) shows the images of the RMSECV value of oil content cross-validation prediction model. Figure 3(b) shows the number of selection variables changing with the number of iterations during the IVSO operation. It can be seen from Figure 3(a) that the lowest RMSECV was obtained in the fourth iteration. A large number of weak information variables are mainly eliminated in the first three iterations, and the main information variables are eliminated after the fourth iteration, resulting in a trend that the curve fell first and then rose. In Figure 3(b), the number of variables selected at four iterations was 52, and the variable space was further reduced by 62.86% compared with band I, continuing to reduce the spectral dimension.

3.3. Optimal Selection by Bootstrapping Soft Shrinkage. After IVSO optimization, a large number of wavelengths were concentrated in 1,660 to 1,740 nm and 2,234 to 2,324 nm, which continued to show certain continuity. The effect on the spectrum was that there were some segments of wavelength, not scatters. In addition, there was still a correlation between variables. A variable could be expressed by the linear combination of other variables, which also had collinearity. Therefore, the wavelengths can be streamlined and optimized. The BOSS algorithm was used to screen the optimal wavelength variables in the final step. The cross-validation fold was five; the number of WBS runs was 1,000; the best model ratio was 10%; and the iteration ran 50 times until the number of new subset variables generated by WBS was 1. The subset with the smallest RMSECV value was used as the final feature variable selected by BOSS.

Results of the BOSS run for the informative wavelength parameter for prediction of oil content in corn samples are shown in Figure 4. It can be seen from Figure 4(a) that the RMSECV of the submodel reached the minimum when it was iterative to four times. Figure 4(b) is the weight values of each feature variable at the optimal iteration. As iterations progress, important variables will gradually take up large weights. Variables with the strongest explanatory power for oil content information were selected when the optimal iteration was reached. The variable space was reduced again, and the number of variables selected was only 7. The selected characteristic wavelengths were 1,660 nm, 1,682 nm, 1,688 nm, 1,708 nm, 1,730 nm, 2,250 nm, and 2,288 nm.

Table 2: Results of synergy interval partial least squares selection subintervals.

Number of intervals	Number of combinations	Interval	nLV	Corresponding band (nm)	RMSE (%)
10	2	5, 9	5	1,660–1,798, 2,220–2,358	2.165
	3	4, 5, 9	4	1,520–1,798, 2,220–2,358	3.714
	4	4, 5, 8, 9	4	1,520–1,798, 2,080–2,358	4.219
20	2	9, 18	6	1,660–1,728, 2,290–2,358	3.747
	3	9, 17, 18	8	1,660–1,728, 2,220–2,358	2.360
	4	9, 10, 17, 18	5	1,660–1,798, 2,220–2,358	2.165
30	2	12, 26	7	1,626–1,670, 2,270–2,314	4.075
	3	13, 14, 26	9	1,672–1,762, 2,270–2,314	2.525
	4	13, 14, 25, 26	6	1,672–1,762, 2,224–2,314	1.806

Notes. nLV – number of latent variables. "nm" – nanometer. RMSE – root mean square error. The synergy interval partial least squares method divided the whole spectrum into 10, 20, and 30 intervals. Based on three division methods, 2, 3, and 4 subintervals were combined, respectively. The RMSE value in each case was obtained. The spectral bands corresponding to the subinterval were the selected variables.

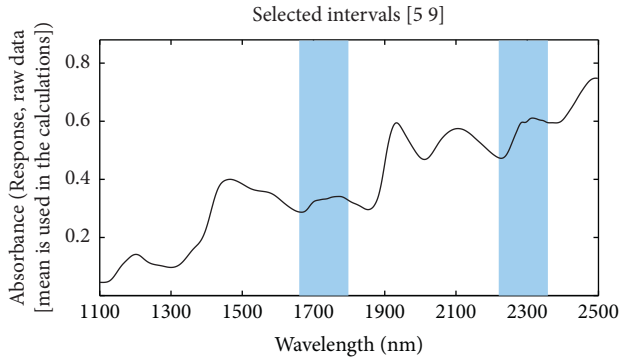

Figure 2: Results of rough selection by synergy interval partial least squares for oil spectral data in corn. The curve was average spectral data of oil in corn and divided into 10 intervals from 1,100 to 2,500 nanometers. Two columns were selected as variable groups in rough selection.

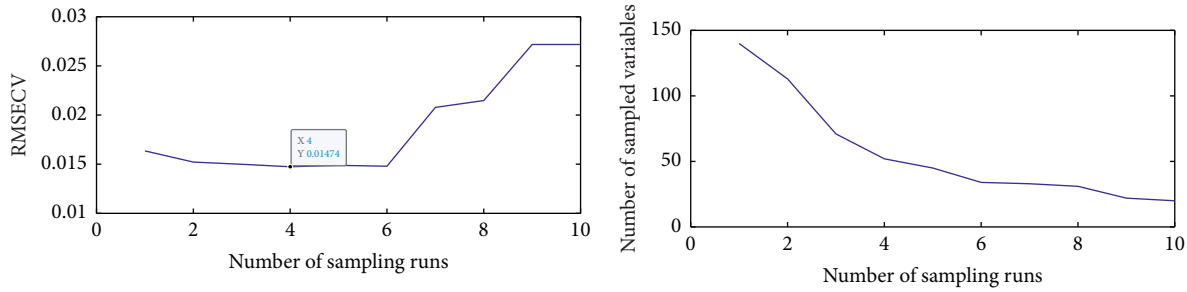

Figure 3: Results of fine selection by iterative variable subset optimization for oil content prediction in corn: (a) the root mean square error of cross-validation varying with the number of sampling and (b) the number of variables selected varying with the number of sampling. The lowest root mean square error appeared in the fourth iteration with 52 variables selected.

4. Results Analysis

After the SiPLS-IVSO-BOSS variable selection method, the number of wavelengths decreased from 700 to 7 for oil content prediction model, and the standard deviation was smaller than other methods. This method greatly reduced the complexity of the model and improved stability and interpretability. The distribution of variables selected for each step is shown in Figure 5. The number of variables was gradually reduced. Most of the variables selected were located in the region from 1,600 to 1,750 nm. The result was consistent with the vibration absorption wavelength of C = C in oil [20]. This proves that the SiPLS-IVSO-BOSS method can effectively screen out the information related to oil in the corn spectrum.

The PLS model was built for full spectral data and spectral data selected by SiPLS-IVSO-BOSS. The PLS model based on the SiPLS-IVSO-BOSS method performed with high accuracy and stability. Contrasted with the PLS model of raw spectral data, the determination coefficient of calibration (R_c^2) increased to 0.998 from 0.934; the determination coefficient of prediction (R_p^2) increased to 0.993 from

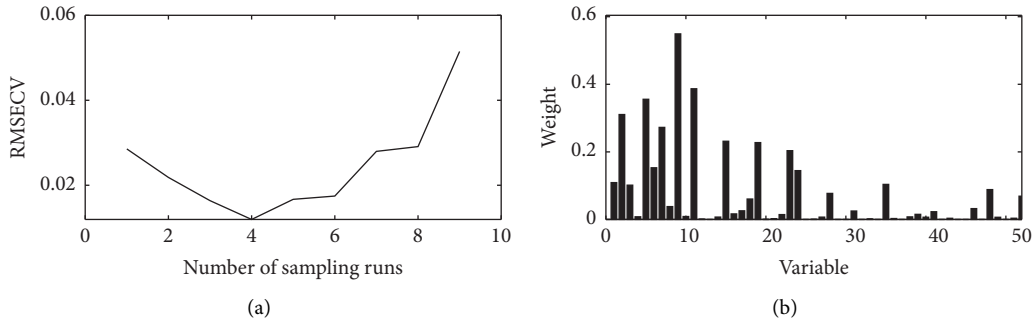

Figure 4: Results of optimal selection of variable by bootstrapping soft shrinkage used for the oil content prediction in corn: (a) the root mean square error of cross-validation varying with the number of sampling and (b) variable weights for sampling with the lowest root mean square error of cross-validation.

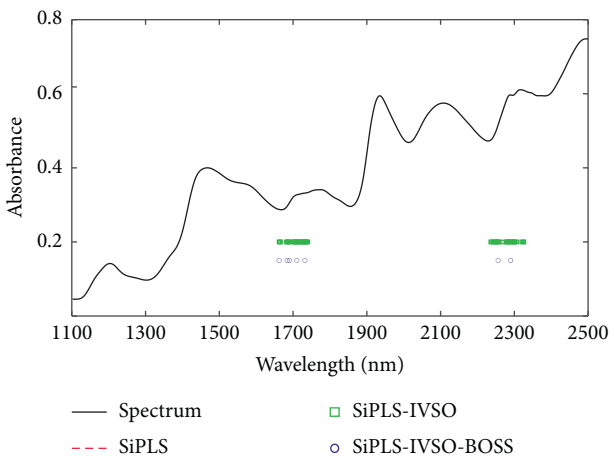

Figure 5: Variable selection results of oil spectral data in corn based on method combined synergy interval partial least squares, iterative variable subset optimization, and bootstrapping soft shrinkage. The curve was average spectral data of 80 corn samples with 700 variables from 1,100 to 2,498 nanometers. Variables selected in the first, second, and third steps were marked in the figure with different symbols. The distribution trend of variables could be seen by comparing the average spectrum.

0.720 that indicating 72.00% of information from the 700 variables; and RMSEC decreased to 0.88% from 4.77%, RMSEP decreased to 1.35% from 9.90%. The model had high accuracy, small error, and few variables. It proved that the SiPLS-IVSO-BOSS variable selection method could extract variables related to the tested substance. Seven hundred variables from raw spectra could explain the oil content of samples, of which 7 informative variables obtained from SiPLS-IVSO-BOSS could explain 99.30% of the information, indicating 27.30% explained variance increased, while 99% (693/700) of nonrelated or uninformative variables were eliminated.

In order to reflect the superiority of this method, compared with other methods, Table 3 is their prediction results. According to the results, the prediction effect of the IVSO-BOSS model was not stable when compared with the single method without the rough selection of the full spectrum band by SiPLS and sometimes even worse than the model accuracy of the single method. The reason is that there was a lack of band extraction method for preliminary selection of the full spectrum, resulting in the lack of guidance for the extraction of wavelength points and poor interpretability of the model. In addition, the prediction results of SiPLS-BOSS and SiPLS-IVSO were better than those of BOSS and IVSO, respectively. The reason is that SiPLS was used to select the band partition, which is convenient for filtering noninformation variables from the wavelength interval and retaining important variables. From two perspectives, it can be seen that the rough selection by SiPLS is essential for subsequent variable extraction. Then, IVSO can provide enough variables for the operation of BOSS, and if you first use BOSS and then IVSO, the initial variable set of IVSO is too small to continue to improve model performance. Therefore, the execution order of the algorithm is fixed.

When SiPLS was replaced by CARS, the modeling results were not even as good as the full spectrum. The main reason is that both CARS and IVSO select variables based on PLS regression coefficients, and it is easy to delete variables containing important information by using two algorithms with the same idea in turn. Another reason is that the number of variables screened by CARS was only 19, and the small number of variables contained rich information. If it continued to be selected, the importance of variables would be redistributed, resulting in those variables with slightly lower information content being eliminated, increasing the risk of information loss. When IVSO was replaced by VCPA, it generally selected fewer variables than CARS, and the above problems still occur. When BOSS was replaced by IRIV, IRIV would classify the variables selected from the previous two steps according to the strength of information, and removed the noninformation and interference information variables through multiple iterations. This classification method was used in the variable space after two optimizations, and there was still the risk of judging important variables as irrelevant variables. BOSS used WBS and general analysis technology to divide variable sets according to weights. Compared with IRIV, BOSS had a better optimization effect for smaller variable space. Therefore, these three algorithms cannot be replaced arbitrarily, and each of them has a role to play in their respective

Table 3: Prediction results of oil content in corn by different variable selection methods.

Variable selection methods	Number of variables	nLV	R_c^2	RMSEC (%)	R_p^2	RMSEP (%)
None	700	10	0.934	4.77	0.720	9.90
CARS	19	6	0.985 ± 0.003	2.27 ± 0.24	0.933 ± 0.010	3.34 ± 0.35
VCPA	8	6	0.985 ± 0.011	2.29 ± 0.37	0.954 ± 0.004	2.86 ± 1.01
IRIV	34	8	0.956 ± 0.013	3.87 ± 0.76	0.745 ± 0.035	6.69 ± 1.25
SiPLS	140	7	0.993 ± 0.002	1.41 ± 0.47	0.984 ± 0.007	1.68 ± 0.92
IVSO	89	6	0.996 ± 0.002	1.12 ± 0.36	0.977 ± 0.008	2.01 ± 1.14
BOSS	22	8	0.992 ± 0.003	1.69 ± 0.58	0.958 ± 0.008	2.73 ± 0.64
IVSO-BOSS	26	7	0.990 ± 0.002	1.38 ± 0.11	0.991 ± 0.003	1.63 ± 0.28
SiPLS-IVSO	52	8	0.996 ± 0.001	1.02 ± 0.28	0.994 ± 0.001	1.49 ± 0.79
SiPLS-BOSS	21	8	0.996 ± 0.001	1.02 ± 0.38	0.990 ± 0.001	1.83 ± 0.47
CARS-IVSO-BOSS	6	7	0.758 ± 0.021	6.58 ± 0.82	0.613 ± 0.047	12.54 ± 1.02
SiPLS-VCPA-BOSS	9	6	0.992 ± 0.004	1.64 ± 0.34	0.980 ± 0.005	2.30 ± 0.53
SiPLS-IVSO-IRIV	35	5	0.997 ± 0.003	0.93 ± 0.21	0.992 ± 0.007	1.48 ± 0.49
SiPLS-IVSO-BOSS	7	6	0.998 ± 0.001	0.88 ± 0.12	0.993 ± 0.001	1.35 ± 0.29

Notes. nLV – number of latent variables. R_c^2 – determination coefficient of calibration. R_p^2 – determination coefficient of prediction. RMSEC – root mean square error of calibration. RMSEP – root mean square error of prediction. Models were built by partial least squares method under different variable selection methods and no method. The statistical results were expressed as mean ± standard deviation of 50 runs.

positions, making it possible to gradually improve the model performance when used in conjunction.

4.1. Content Prediction Results of Starch of Corn and Organic Matter of Soil. The prediction results of the content of two substances by different methods are shown in Tables 4 and 5. Table 4 is the prediction results of starch of corn based on different variable selection methods. The SiPLS-IVSO-BOSS method compressed the number of variables to 1.14% of the original. Contrasted with the PLS model of raw spectral data, R_c^2 increased to 0.999 from 0.941; R_p^2 increased to 0.999 from 0.882; and RMSEC decreased to 2.40% from 19.89%; and RMSEP decreased to 2.80% from 28.05%. It can be seen that the prediction accuracy of the model was gradually increasing after the SiPLS-IVSO-BOSS method, which can achieve better results than a single method or two-step strategy. Table 5 is the prediction results of different variable selection methods for the organic matter content of the soil. The number of variables was reduced from 1,050 to 10 through the SiPLS-IVSO-BOSS method, and the spectral space was fully compressed. Taking RMSEC as a reference, after the gradual optimization of the SiPLS-IVSO-BOSS method, its value decreased from 184.25% to 129.99%. Overall, its modeling results were superior to other methods.

5. Discussion

The superiority of the SiPLS-IVSO-BOSS method was proved after the detailed analysis of the prediction of oil content in corn. Moreover, the above research showed that the SiPLS-IVSO-BOSS method also had the best prediction effect and could effectively extract relevant information of the tested components for starch spectral data and organic matter spectral data. In order to reflect the superiority of this method, the experimental results were compared with other studies. In some previous PLS-based models, references [21–23] used centralization as a pretreatment method. When the starch content of corn was predicted in reference [21], VCPA was used to select variables, and the RMSEC and RMSEP were 5.18% and 4.87%, respectively. When the oil and starch contents of corn were predicted in reference [22], BOSS was used to select variables, and their RMSEP were 2.32% and 19.10%, respectively. Reference [23] predicted the oil content of corn by using Fisher optimal subspace shrinkage to select variables; RMSEC was 1.06%; RMSEP was 1.61%; R_c^2 was 0.997; and R_p^2 was 0.979. Reference [24] used a standardized pretreatment method and used sparse coefficients wavelength selection and regression to select variables. The experiment used the organic matter content of soil for verification. The obtained R_c^2 was 0.948; R_p^2 was 0.977; RMSEC was 240.32%; and RMSEP was 172.96%. It can be seen that SiPLS-IVSO-BOSS has a better model performance. Admittedly, this combined strategy increases the computational task and makes the model more complex. However, it does not significantly increase the computation time, and the gradually shrinking variable space allows the variable selection algorithm to execute more smoothly and efficiently. When applied, the integration is packaged together without additional burden, and the process of running the program can be simple and convenient. In the prediction of the content of the three substances, the method improved the prediction error by at least 8% compared to the two-step strategy, which is an improvement that should not be overlooked in content detection studies, since small differences can cause different results in practical applications. Therefore, SiPLS-IVSO-BOSS has the advantage for variable selection of near-infrared spectral analysis from its principle to practice. It is a choice when dealing with high-dimensional data, which provides an algorithm fusion idea for researchers.

The method in this paper can theoretically realize the rough selection to the optimal selection of variables. Experiments were carried out on this basis. Through modeling on public data sets of corn and soil, the content of the three substances was predicted. The accuracy of the model had been improved, but the degree of improvement is different, which is related to the nature of the data itself.

Table 4: Prediction results of starch content in corn by different variable selection methods.

Variable selection methods	Number of variables	nLV	R_c^2	RMSEC (%)	R_p^2	RMSEP (%)
None	700	8	0.941	19.89	0.882	28.05
CARS	25	6	0.980 ± 0.002	11.58 ± 0.52	0.971 ± 0.004	13.69 ± 0.97
VCPA	9	9	0.990 ± 0.001	8.04 ± 2.51	0.987 ± 0.002	9.28 ± 2.19
IRIV	16	6	0.964 ± 0.004	15.65 ± 1.68	0.920 ± 0.009	22.59 ± 1.97
SiPLS	182	8	0.991 ± 0.002	7.60 ± 1.63	0.976 ± 0.010	12.65 ± 1.89
IVSO	102	10	0.984 ± 0.002	10.26 ± 1.34	0.965 ± 0.006	14.92 ± 2.11
BOSS	13	9	0.991 ± 0.001	7.74 ± 0.85	0.990 ± 0.002	8.02 ± 1.09
IVSO-BOSS	31	8	0.987 ± 0.001	8.92 ± 0.49	0.974 ± 0.003	13.52 ± 0.67
SiPLS-IVSO	90	7	0.998 ± 0.001	4.05 ± 0.68	0.995 ± 0.001	5.55 ± 0.84
SiPLS-BOSS	27	6	0.999 ± 0.001	2.35 ± 0.57	0.997 ± 0.001	4.22 ± 0.52
CARS-IVSO-BOSS	8	5	0.812 ± 0.024	28.15 ± 3.54	0.0751 ± 0.054	34.02 ± 3.87
SiPLS-VCPA-BOSS	7	7	0.709 ± 0.042	9.31 ± 2.14	0.711 ± 0.032	9.81 ± 1.89
SiPLS-IVSO-IRIV	23	6	0.999 ± 0.001	2.52 ± 0.43	0.999 ± 0.002	3.01 ± 0.68
SiPLS-IVSO-BOSS	8	8	0.999 ± 0.001	2.40 ± 0.23	0.999 ± 0.001	2.80 ± 0.35

Table 5: Prediction results of organic matter content in soil by different variable selection methods.

Variable selection methods	Number of variables	nLV	R_c^2	RMSEC (%)	R_p^2	RMSEP (%)
None	1,050	10	0.976	184.25	0.749	318.79
CARS	72	8	0.985 ± 0.002	146.77 ± 10.15	0.916 ± 0.022	154.80 ± 22.36
VCPA	8	7	0.978 ± 0.016	151.41 ± 10.53	0.948 ± 0.031	122.40 ± 20.24
IRIV	11	9	0.983 ± 0.006	149.90 ± 12.13	0.933 ± 0.027	139.06 ± 15.47
SiPLS	140	8	0.985 ± 0.007	143.04 ± 15.89	0.876 ± 0.035	240.06 ± 24.19
IVSO	202	8	0.982 ± 0.009	152.96 ± 16.43	0.923 ± 0.029	148.41 ± 24.18
BOSS	50	10	0.987 ± 0.003	134.07 ± 24.73	0.901 ± 0.024	168.21 ± 25.12
IVSO-BOSS	26	6	0.985 ± 0.003	145.44 ± 21.31	0.891 ± 0.187	193.45 ± 22.74
SiPLS-IVSO	68	8	0.987 ± 0.001	137.22 ± 10.74	0.977 ± 0.168	95.41 ± 12.54
SiPLS-BOSS	13	7	0.988 ± 0.001	131.71 ± 11.85	0.971 ± 0.013	108.80 ± 20.57
CARS-IVSO-BOSS	6	8	0.901 ± 0.027	200.57 ± 20.64	0.682 ± 0.058	347.23 ± 24.13
SiPLS-VCPA-BOSS	9	6	0.984 ± 0.017	142.87 ± 13.54	0.941 ± 0.067	204.19 ± 27.10
SiPLS-IVSO-IRIV	15	8	0.983 ± 0.002	131.71 ± 12.64	0.977 ± 0.029	100.04 ± 16.47
SiPLS-IVSO-BOSS	10	9	0.988 ± 0.001	129.99 ± 10.07	0.979 ± 0.009	94.03 ± 11.76

This experiment shows that the application of SiPLS-IVSO-BOSS to variable selection is feasible and has a certain versatility. To achieve the best results in more data sets, further research is needed. The research will be carried out in future work.

6. Conclusions

Referring to the three-step hybrid strategy of near-infrared spectral analysis, SiPLS-IVSO-BOSS variable selection method was proposed. In the first two steps, the variable space was continuously contracted, and then the collinearity was removed from the remaining variable set, and the smaller variable combination was extracted from the large variable set. It was validated in spectral data sets of corn and soil and achieved better prediction results than a single algorithm, the combination of two algorithms, and CARS, VCPA, IRIV, and other mainstream algorithms. It provides an effective solution for dealing with high-dimensional data, avoids time-consuming and inefficient problems, and can provide a theoretical reference for the variable selection strategy of the spectrum. When used for other data, some parameters in the algorithm can be debugged to obtain the optimal model.

Data Availability

The website of corn spectral data is http://www.eigenvector.corn/data/corn/index.html. The website of soil spectral data is Quality & Technology.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the Research and Application of Key Technologies for Smart Pasture Production (2019YFE0125600), the China Agriculture Research System (CARS-36), and the Heilongjiang Post-Doctoral Subsidy Project of China (LBH-Z17020).

References

[1] G. H. Shen, Y. Y. Cao, X. Liu, J. H. Xu, J. R. Shi, and Y. W. Lee, "Identification of fusarium damaged kernels using near infrared hyperspectral imaging and characteristic bands selection," *Jiangsu Journal of Agricultural Sciences*, vol. 37, no. 02, pp. 509–516, 2021.

[2] M. Y. Huang, H. Y. Wu, H. Jin, G. M. Dong, Y. Y. Yang, and R. J. Yang, "Discrimination of adulterated milk based on near infrared transmission and diffuse reflectance spectroscopy," *Spectroscopy and Spectral Analysis*, vol. 40, no. S1, pp. 85-86, 2020.

[3] X. P. Li, H. Z. Jiang, X. S. Jiang, H. Y. Gu, and H. P. Zhou, "Advances on non-destructive quality detection of forest-fruit in the sort of woody grain and oil based on near infrared spectroscopy and hyperspectral imaging technology," *Food and Fermentation Industries*, vol. 48, no. 2, pp. 302–308, 2022.

[4] Z. Y. Han, C. M. Wang, W. Song, and C. G. Liu, "Study on pharmacognosy and near infrared identification technology of Toricellia Angulata Oliv.var.intermedia (Harms) Hu leaves," *China Forest Products Industry*, vol. 58, no. 04, pp. 59–63, 2021.

[5] D. J. Xie, C. L. Lv, M. Zu, and H. F. Cheng, "Research progress of bionic materials simulating vegetation visible-near infrared reflectance spectra," *Spectroscopy and Spectral Analysis*, vol. 41, no. 04, pp. 1032–1038, 2021.

[6] L. Yu, Y. S. Hong, Y. Zhou et al., "Wavelength variable selection methods for estimation of soil organic matter content using hyperspectral technique," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 32, no. 13, pp. 95–102, 2016.

[7] L. Yu, Y. X. Zhu, Y. S. Hong, T. Xia, M. X. Liu, and Y. Zhou, "Determination of soil moisture content by hyperspectral technology with CARS algorithm," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 32, no. 22, pp. 138–145, 2016.

[8] Y. J. Wang, M. H. Li, L. Q. Li, J. M. Ning, and Z. Z. Zhang, "Green analytical assay for the quality assessment of tea by using pocket-sized NIR spectrometer," *Food Chemistry*, vol. 345, Article ID 128816, 2021.

[9] S. Y. An, L. Zhang, X. Z. Shang, H. S. Yue, W. Y. Liu, and A. C. Ju, "Variable selection method in the NIR quantitative analysis model of total saponins in red ginseng extract," *Spectroscopy and Spectral Analysis*, vol. 41, no. 01, pp. 206–209, 2021.

[10] D. M. Sun and K. W. Huan, "Research on near infrared spectroscopy analytical methods of moisture content in wheat based on BOSS," *Journal of Changchun University of Science and Technology (Natural Science Edition)*, vol. 43, no. 05, pp. 1–6, 2020.

[11] Q. M. Kong, J. T. Gu, R. Gao, Z. D. Li, Z. Ma, and Z. B. Su, "Study on detection of crude protein in ammonified and alkalized corn straw by spectrum characteristic band selection method based on synergy interval partial least squares," *Journal of Instrumental Analysis*, vol. 39, no. 11, pp. 1334–1343, 2020.

[12] W. H. Su, C. Yang, Y. H. Dong et al., "Hyperspectral imaging and improved feature variable selection for automated determination of deoxynivalenol in various genetic lines of barley kernels for resistance screening," *Food Chemistry*, vol. 343, Article ID 128507, 2021.

[13] D. D. Silalahi, H. Midi, J. Arasan, M. S. Mustafa, and J. P. Caliman, "Robust wavelength selection using filter-wrapper method and input scaling on near infrared spectral data," *Sensors*, vol. 20, no. 17, p. 5001, 2020.

[14] W. Y. Yang, W. M. Wang, R. Q. Zhang et al., "A modified moving-window partial least-squares method by coupling with sampling error profile analysis for variable selection in near infrared spectral analysis," *Analytical Sciences*, vol. 36, no. 3, pp. 303–309, 2020.

[15] H. D. Yu, Y. H. Yun, W. M. Zhang et al., "Three-step hybrid strategy towards efficiently selecting variables in multivariate calibration of near infrared spectra," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 224, Article ID 117376, 2020.

[16] R. Rinnan and Å. Rinnan, "Application of near infrared reflectance (NIR) and fluorescence spectroscopy to analysis of microbiological and chemical properties of arctic soil," *Soil Biology and Biochemistry*, vol. 39, no. 7, pp. 1664–1673, 2007.

[17] W. T. Wang, Y. H. Yun, B. C. Deng, W. Fan, and Y. Z. Liang, "Iteratively variable subset optimization for multivariate calibration," *RSC Advances*, vol. 5, no. 116, pp. 95771–95780, 2015.

[18] S. B. Chen and Z. Hu, "Determination of corn fat based on NIRS and QPSO-LSSVM model," *The Chemical Engineer*, vol. 31, no. 08, pp. 31–35, 2017.

[19] 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," *Guang pu xue yu guang pu fen xi = Guang pu*, vol. 31, no. 02, pp. 362–365, 2011.

[20] J. Wu, B. S. Chen, J. H. Fang, and J. Wang, "Oxidation mechanisms of soybean biodiesel based on spectroscopic analysis," *Renewable Energy Resources*, vol. 31, no. 12, pp. 107–110, 2013.

[21] Y. H. Yun, W. T. Wang, B. C. Deng et al., "Using variable combination population analysis for variable selection in multivariate calibration," *Analytica Chimica Acta*, vol. 862, pp. 14–23, 2015.

[22] B. C. Deng, Y. H. Yun, D. S. Cao et al., "A bootstrapping soft shrinkage approach for variable selection in chemical modeling," *Analytica Chimica Acta*, vol. 908, pp. 63–74, 2016.

[23] Y. W. Lin, B. C. Deng, L. L. Wang, Q. S. Xu, L. Liu, and Y. Z. Liang, "Fisher optimal subspace shrinkage for block variable selection with applications to NIR spectroscopic analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 159, pp. 196–204, 2016.

[24] T. Lei and D. W. Sun, "A novel NIR spectral calibration method: sparse coefficients wavelength selection and regression (SCWR)," *Analytica Chimica Acta*, vol. 1110, pp. 169–180, 2020.