applied sciences

MDPI

Article

An Effective Prediction Approach for Moisture Content of Tea Leaves Based on Discrete Wavelet Transforms and Bootstrap Soft Shrinkage Algorithm

Min Zhang [1], Jiaming Guo [1], Chengying Ma [2], Guangjun Qiu [3], Junjie Ren [1], Fanguo Zeng [1] and Enli Lü [1,*]

1 College of Engineering, South China Agricultural University, Guangzhou 510640, China; mimi1208@stu.scau.edu.cn (M.Z.); jmguo@scau.edu.cn (J.G.); renjacky@stu.scau.edu.cn (J.R.); tsvanco@stu.scau.edu.cn (F.Z.)

2 Tea Research Institute, Guangdong Academy of Agricultural Sciences/Guangdong Provincial Key Laboratory of Tea Plant Resources Innovation & Utilization, Guangzhou 510640, China; machengying@tea.gdaas.cn

3 Public Monitoring Center for Agro-Product of Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China; qiuq16@scau.edu.cn

* Correspondence: enlilv@scau.edu.cn; Tel.: +86-020-8528-2860

Received: 20 June 2020; Accepted: 13 July 2020; Published: 14 July 2020

Abstract: The traditional method used to determine the moisture content of tea leaves is time consuming and destructive. To address this problem, an effective and non-destructive prediction method based on near-infrared spectroscopy (NIRS) is proposed in this paper. This new method combines discrete wavelet transforms (DWT) with the bootstrap soft shrinkage algorithm (BOSS). To eliminate uninformative or interfering variables, DWT is applied to remove the noise in the spectral data by decomposing the origin spectrum into six layers. BOSS is used to select informative variables by reducing the dimensions of the sub-layers' reconstruction spectrum. After selecting the effective variables using DWT and BOSS, a prediction model based on partial least squares (PLS) is built. To validate effectiveness and stability of the prediction model, full-spectrum PLS, genetic algorithm PLS (GA-PLS), and interval PLS (iPLS) were compared with the proposed method. The experiment results illustrate that the proposed prediction model outperforms the other classical models considered in this study and shows promise for the prediction of the moisture content in Yinghong No. 9 tea leaves.

Keywords: near-infrared; moisture content; discrete wavelet transforms; bootstrap soft shrinkage algorithm; partial least squares

1. Introduction

Processed leaves and leaf buds of tea tree are used to produce tea, which are popular in many parts of the world [1]. Traditional tea making is complicated; the drying of fresh leaves is the primary and indispensable stage of this process [2] and moisture content is a key index in the drying process [3]. However, improper handling may lead to inaccurate measurements when determining moisture content. Therefore, an accurate and rapid detection approach would be indispensable for determining the moisture content of tea leaves during tea making [4].

Many attempts have been made to determine moisture based on near-infrared spectroscopy (NIRS). Moisture measurements are commonly recorded by detecting mass loss after heating to evaporate moisture. However, this procedure damages the samples and is time consuming. In contrast, direct determination of moisture by NIRS is fast, only requiring the acquisition of the sample's reflection spectrum [5]. However, the disadvantages of NIRS include broad overlapping, difficultly interpreting

the attribute absorption bands, and noise [6]. The effective selection of wave bands is used to address these problems. It is difficult to select effective variables that peaks are unresolved and important features cannot be recognized [7]. Therefore, it is crucial to eliminate noise and avoid losing spectral details in the spectral prediction model of determining the moisture content.

The current common spectral denoising methods include moving average, Savitzky–Golay filtering, and median computing [8]. Xie et al. proposed a tailoring noise frequency spectrum technique based on the Savitzky–Golay filter and obtained a satisfying result [9]. Morgan et al. used the moving weighting algorithm to estimate soil organic carbon content fixing the spectrum bias [10]. Although these methods can remove the noise in spectral data, useful signals may be lost during the process of denoising. To avoid losing effective variables, a discrete wavelet transform (DWT) of spectral signals was developed. The moisture content (MC), soluble solids content (SSC), pH, and hardness of Gala apple samples were tested non-destructively within 350–2500 nm using the wavelet transform pretreatment of raw spectral data [11]. The use of DWT successfully further simplified the genetic algorithm-the partial least squares (GA-PLS) model by reducing variables by 40–44% without reducing the prediction accuracy [12]. Other experimental results [13] showed that the DWT-support vector regression (DWT-SVR) multivariate regression model, having good robustness, can measure protein, starch, and fat contents in corn simultaneously, demonstrating that DWT can effectively remove noise from corn NIRS spectral data. However, an unsolved problem is that DWT cannot reduce the dimensions of huge data, which leads to a redundancy in the data volume during model building. In summary, wavelet decomposition is an effective method for removing noise without reducing data dimension.

As NIRS produces a large amount of data, considerable residual redundant noise and irrelevant data remain after spectra denoising. Therefore, variables must be selected before building a prediction model. The benefits of variable selection can be summarized into three aspects: (1) eliminating uninformative or interfering variables, (2) selecting informative variables, and (3) reducing the dimensions of the data [14]. The common selection methods can be divided into three types: (1) single variable selection, where some use different variable ranking criteria such as regression coefficients and variance analysis [9]; (2) random variable selection such as uninformative variable elimination (UVE) [15], genetic algorithm (GA) [16], random forest (RF) [17], etc.; and (3) interval variable selection such as interval partial least squares (iPLS) [18] and synergy iPLS (SiPLS) [19]. A new variable selection method called bootstrap soft shrinkage algorithm (BOSS) was proposed, which was derived from the idea of weighted bootstrap sampling (WBS) and model population analysis (MPA) [20]. In BOSS, WBS is used to generate sub-models based on the weights, and MPA is used to analyze the sub-models and update the weights of the variables [21,22]. Yan et al. used the BOSS method with mid-infrared (MIR) spectroscopy to determine chlorantraniliprole in abamectin, and obtained the highest coefficient of determination of cross-validation (R^2_{cv} = 0.9998) and coefficient of determination of the test set (R^2_p = 0.9989) [23]. Zhang et al. showed that BOSS can improve prediction performance and markedly reduce features, and had the best accuracy in calibration and prediction with the correction determination coefficient (R^2_c) of 0.9907, the root-mean-square error of calibration (RMSEC) of 0.4257 mg/kg, R^2_p of 0.9821, and the root-mean-square error of prediction (RMSEP) of 0.6461 mg/kg [24]. From the above research, the BOSS algorithm not only improves the prediction accuracy of the model but also effectively reduces the number of variables to speed up the calculation of the model. However, BOSS directly processes the original spectral data, which includes processing irrelevant noise information. Therefore, noise elimination steps must be added.

In this study, we constructed a novel variable selecting method based on DWT and BOSS. GA [25] and iPLS [26,27] were compared with the new proposed method, which are categorized as a random variable selection method and interval variable selection method, respectively [28,29]. As classical methods, many studies selected variables to improve prediction ability. Jiang, H monitored yeast concentrations of *Saccharomyces cerevisiae* cultivations with NIRS and compared the results with different variable selection methods. The GA model was built on fewer data points than that based on full spectra, which ranges from 1557 to 71 points, with R^2_p ranging from 0.9777 to 0.9806 [30]. Sousa Sampaio

optimized rice amylose determination using NIRS with the iPLS method. The full spectrum was split into 10, 20, 25, and 50 intervals, and the optimal model was obtained for the Savitzky–Golay filter (R_p^2 = 0.92 and RMSEP = 2.133), which was better than the full-spectrum PLS model [7]. Yang et al. used different regression methods such as PLS, iPLS, and SiPLS with multiple pretreatment methods. The Al_2O_3 models obtained using the iPLS algorithm had R_p^2 values of 0.8273 to 0.9196 [31].

PLS, which can improve the prediction ability by selecting informative variables or eliminating uninformative variables, was used to build a prediction model in this paper [32–34]. DWT and BOSS were combined as a new variable selection method. After previous variable selection, DWT-BOSS-PLS, GA-PLS, and iPLS models were established, which are based on the NIRS data and moisture content. By comparison, three variable selection methods are discussed to choose the best one.

2. Materials and Methods

2.1. Trial Introduction

For the trial, we used Yinghong No. 9 variety tea leaves, which was carried out on 4 December, 2019 at the Yingde Yinghong No. 9 base of the Tea Research Institute of Guangdong Academy of Agricultural Sciences (Yingde, Qingyuan, Guangdong, China). The tea leaves were picked randomly within the tea garden. At 12:00 p.m., 100 kg of tea leaves were picked and placed in a withering trough, and the leaves were about 4 cm thick. Samples were taken every hour from withering trough. At the normal time of withering, there was a total of 15 h. The fresh tea leaves were taken in 5 samples and, in the other 14 h, the tea leaves were taken 10 samples. In total, 145 samples were obtained in this test.

The tea leaves reflectance spectra were measured using a Thermo Antaris II Fourier transform near-infrared (FT-NIR) spectrometer (Thermo Scientific Co., Waltham, MA, US) with a diffuse reflection of the integrating sphere at a spectral range of 12,000–3800 cm^{-1} (833–2630 nm). The resolution was 4 cm^{-1} and the diameter of the sample cup rotator was 20 cm. The number of sample scans were 64 (can rotate a circle). Each sample was covered with 25 g in an integration sphere. Three spectra were taken from each sample, and then the average spectra were taken as the spectra of the corresponding samples.

2.2. Moisture Content Acquisition

Tea leave samples were tested for moisture content immediately after the spectral experiments, and measured for moisture content according to GB/T 8304-2013 in Chinese. From 12:00 p.m., the moisture content was recorded every hour. Each spectrum corresponds to a moisture content, so the number of moisture content was 145. The average moisture content per hour over a range of 15 h is shown in Figure 1. As the withering time increases, the moisture content gradually decreases.

$$w = \frac{m_1 - m_2}{m_1} \times 100\% \quad (1)$$

where w is the moisture content, m_1 is the leaves of fresh weight, m_2 is the leaves of dry weight.

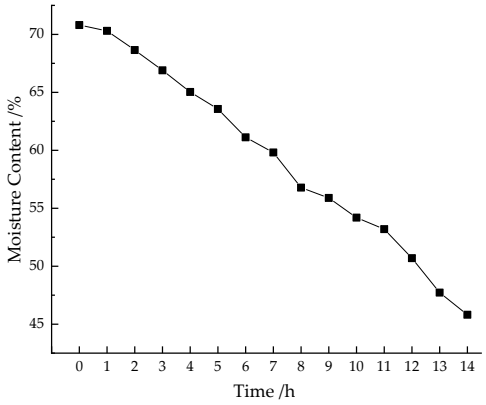

Figure 1. The moisture content of tea leaves changes with withering for 15 h.

2.3. Sample Set Partitioning Based on Joint X-Y Distance (SPXY)

The SPXY algorithm is a method of dividing the sample set considering both X and Y variables. It evolved from the Kennard–Stone (KS) algorithm and divides the samples into training and test sets by calculating the distance between samples [35]. In this paper, X indicates tea leaves' spectral data and Y indicates moisture content.

2.4. Analysis of PLS Model

The partial least squares organically combines the model and cognitive methods. Under regression modeling (multiple linear regression), data structure simplification (principal component analysis) and correlation analysis between two sets of variables can be performed simultaneously [36]. In this study, the prediction model was built on the PLS algorithm.

In this study, the following parameters were selected to evaluate the accuracy of the model: correction determination coefficient (R_c^2), cross-validated determination coefficient (R_{cv}^2), prediction determination coefficient (R_p^2), the root mean square error of calibration (RMSEC), the root mean square error of cross-validation (RMSECV), and the root mean square error of prediction (RMSEP). The larger the R^2, the more accurate the predictive ability of the mode, and the RMSE represents the stability of the model [37]. The lower the value of these three values, the higher the reliability of the model. The correlation coefficient (R) is used to measure the correlation between two variables; the closer R is to 1, the higher the correlation. In this paper, when $R > 0.8$, the corresponding variables are defined as strongly correlated variables.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})(y_i - \overline{y})^2}{(n-1)\sum_{i=1}^{n}(y_i - \hat{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2} \quad (2)$$

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

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_{i*})^2}{n}} \quad (4)$$

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}{m}} \quad (5)$$

where $\hat{y}_i$ is the value predicted by the calibration model, y_i is the reference value, $\hat{y}_{i*}$ is the value predicted by the cross-validation model, $\overline{y}$ is s the mean of the reference values, $\overline{\hat{y}}$ is the mean of the predicted values, n is the number of samples in the calibration or validation steps, and m is the number of predicted samples.

2.5. Wavelength Selection Method

2.5.1. DWT and BOSS Coupling Algorithm

DWT produces a multi-scale representation of digital signals using a series of high- and low-pass cutoff filters to classify signals according to their frequencies in the wavelength space of the spectrum [38–40]. In this study, the wavelet decomposition coefficient was extracted first, then the wavelet high-frequency coefficient (HC) and the wavelet low-frequency coefficient (LC) were extracted. Finally, the LC and the zeroing HC were combined to establish the wavelet reconstruction matrix.

BOSS is a method using collinearity to select effective features and using the information of the regression coefficient to flexibly shrink the information of interest. The BOSS algorithm is constructed using bootstrap sampling (BBS) and weighted bootstrap sampling (WBS) to generate random combination of

variables and sub-model, and by combining model population analysis (MPA) and the PLS algorithm to extract effective information from the sub-model [12].

In this paper, the DWT-BOSS algorithm is proposed by coupling DWT with BOSS to obtain the optimal band to establish the PLS prediction model. The process is as follows (shown in Figure 2).

(1) Obtain the maximum decomposition layer (L(Max)) of the wavelet transform. First observe the trend in the spectra image after decomposition, then according to the order of correlation coefficient to select the maximum number of layers;
(2) Use the BOSS algorithm to optimize the effective variables of each spectral data from L1 to L(Max) (L1-L(Max)), and the optimal variables set is obtained by superimposing the preferable variables of L1-L(Max). L1 is defined as the first decomposition layer.

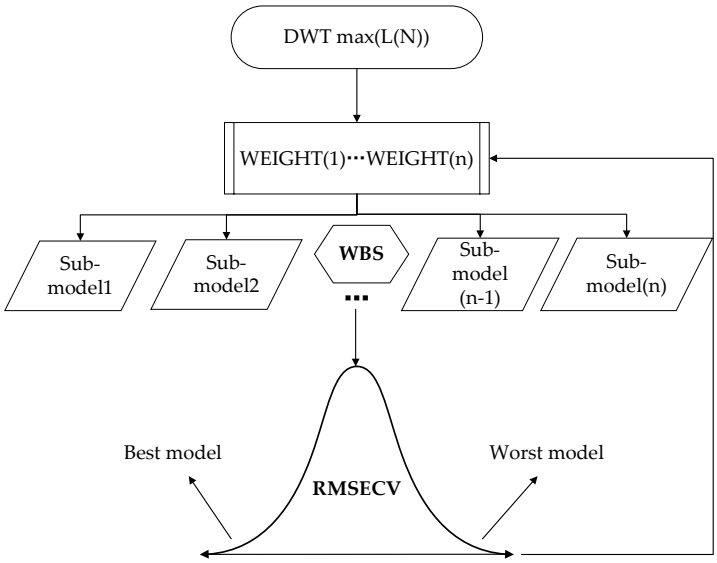

Figure 2. The flow chart of the discrete wavelet transforms (DWT) with the bootstrap soft shrinkage algorithm (BOSS) coupling algorithm. Note: N, the max layer of the DWT.; WEIGHT (*n*), the weight of the *n* variable; WBS, weighted bootstrap sampling; sub-model, generate random combination variables and sub-model; RMSEC, sub-models' root mean square error of cross-validation.

2.5.2. Correlation Coefficient (R)Method

The correlation coefficient method involves obtaining the correlation coefficient from the unknown sample and the reference sample to judge whether the unknown sample and the reference sample are consistent for a certain property. The higher the similarity, the closer the *R* value is to 1. The formula is as follows:

$$R = \frac{Cov(y_1, y_2)}{\text{delta\_}y_1 \times \text{delta\_}y_2} \quad (6)$$

where the y_1 is the absorbance corresponding to each wavelength point, y_2 is the water content. delta_y_1 means the y_1 of the standard deviation, delta_y_2 means the y_2 of the standard deviation. *Cov* is the covariance.

2.5.3. Genetic Algorithm

The GA is a global optimization method that can solve problems efficiently for which there are many possible solutions, such as variable selection. The core steps of GA are analogous to the process of Darwinian evolution, in which individuals are selected for the next generation through crossover, mutation, and survival of the fittest until a specific stopping criterion is reached [16]. The main GA parameters were set as follows: population size of 32, window width of 10, maximum generation of 100, and mutation rate of 0.005 in MATLAB R2016a (MathWorks, Natick, MA, USA).

2.5.4. iPLS

The iPLS method is a wavelength interval selection method. The method functions by dividing the whole spectrum into several intervals, and then expanding or decreasing the wavelength variables by the center of the interval [18]. The modeling setting of the iPLS method was as follows: the number of intervals was set to 20 in MATLAB R2016a (MathWorks, Natick, MA, USA).

3. Results and Discussion

3.1. Wavelet Transform and Maximum Decomposition Layer

In this study, the db4 wavelet-generating function was used in MATLAB R2016a (MathWorks, Natick, MA, USA) to decompose the eight layers wavelet of the original spectrum. The reconstructed signals of layers 1 to 8 are defined as L1 to L8(L1–L8), respectively. L0 indicates the origin spectrum.

The significant moisture absorption peaks around 1800 and 2400 nm, and weak peaks around 1200 and 2600 nm. There are three distinct areas of noise in L0, which were more obvious around 1600 (defined as noise1), 2200 (defined as noise2), and 2400 nm (defined as noise3) (Figure 3). The small burr phenomenon occurred in noise1, noise2, and noise3. Figure 4 depicts the noise spectral image around 1600 nm, the scope of which is disordered in L0. When decomposition was applied, the high-frequency signal was further removed, and the noise weakened. As shown in Figure 4, when the original spectrum was decomposed into the fifth sub-layer, the spectral curves became smoother. The spectral details were gradually removed and the spectral curve gradually tended to be smooth, so some absorption peaks representing the moisture characteristics of tea leaves disappeared. When the spectrum was decomposed into seven layers, the spectral curve was almost a straight line, but in this case, the spectral data considerably deviated from the original data and large amounts of effective information were lost. The loss of effective information was more serious at L8. Therefore, L7 and L8 completely deviated from L0 to L6.

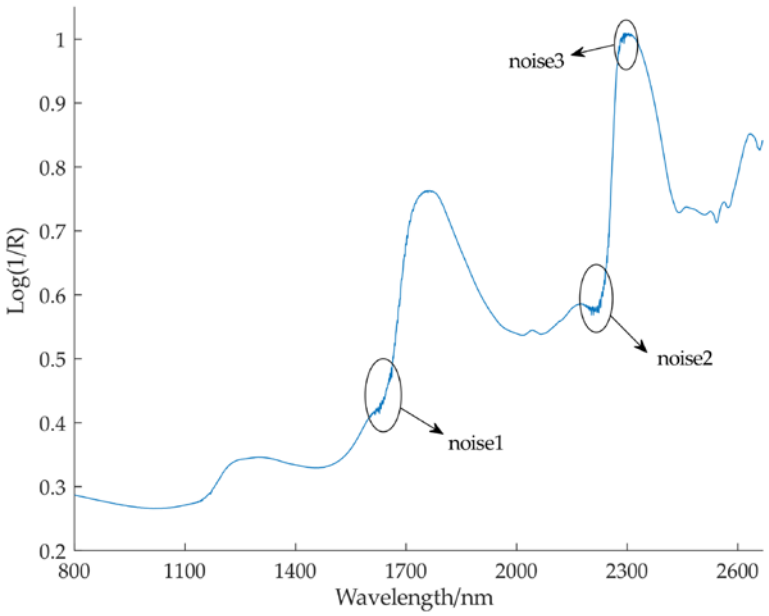

Figure 3. The average spectrum of L0. Note: L0 refers to the original spectrum, which has three distinct areas of noise around 1600, 2200, and 2400 nm here (defined as noise1, noise2, and noise3, respectively).

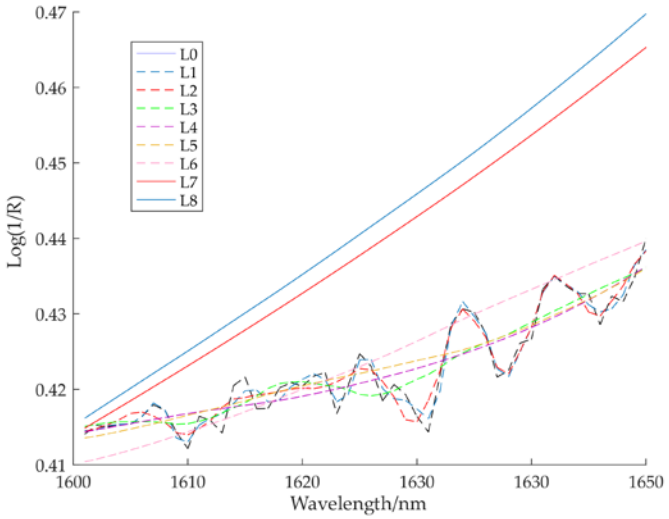

Figure 4. Zoom in on a noise spectral image around 1700 nm. Note: L0 refers to the original spectrum, L1 refers to the reconstructed spectral average after the layer wavelet transforms, L2 to L8 are analogous to L1. L0–L6 have similar trends under L7 and L8. L7 and L8 correspond to solid red and solid blue, respectively.

To further determine the appropriate maximum decomposition layer, the correlation coefficient method was used to measure the correlation between the spectral absorption and moisture characteristics of each wavelength point in the spectral matrix of L1–L8. By comparing the measured value (defined as R) with the threshold value, the preferable number of wavelength points was determined. In this study, the threshold value was set to 0.8. According to Figure 5, 259 points in L1 and L2 exceeded the threshold, 257 points in L3 and L4, 260 points in L5, 262 in L6, 175 in, and 199 in L8. Figure 5 shows that the numbers of points in L1–L6 passing threshold were similar, stable at 260 ± 3. About 30% less of the points in L7 and L8 passed the threshold than in L1–L6, gradually weakening the moisture characteristics of the spectrum. To ensure that enough moisture characteristics are preserved after the wavelet transform, the sixth decomposition layer was taken as the largest decomposition layer and the reconstruction spectrum of L1–L6 lost as few spectral details as possible and noise was relatively thoroughly removed. In the following, we used L1–L6 to replace the original spectrum.

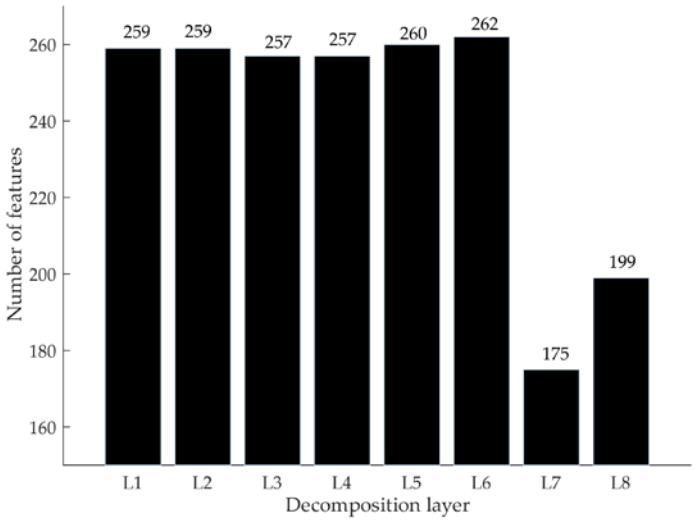

Figure 5. The number of wavelength points of L1 to L8 passing through the threshold, which was set to 0.8.

3.2. An Optimal Variable Set Applicable to Moisture Characteristics of Tea Leaves

The BOSS method was used to optimize the variables of different layers of the wavelet reconstruction matrix. As shown in Figure 6, the optimal set of each layer was roughly distributed around the moisture absorption peak (1200, 1400, 2400, and 2600 nm). Due to the randomness of the variable selection, BOSS was repeated 30 times to reduce the statistical errors. Therefore, the top 10 variables with the highest occurrence frequency were taken as the preferable variables in each layer after 30 cycles in the test. After the combination of the preferable variables of L1–L6, the optimal variable set V was obtained. As the number of decomposition layers increased, some moisture features were optimized and some irrelevant information was eliminated. The optimal variables decomposed by L1–L6 were superimposed to obtain 55 optimal variables in the regions of 800–1000, 1100–1400, 1500–1700, 1900–2000, and 2300–2600 nm. The considerable number of variables are in the range of 800–1100, 1200–1400, and 1700–2000 nm. The wavelength ranges were mainly represented by the fundamental frequency vibration of the free –OH group, as well as the combination and octave vibration absorption.

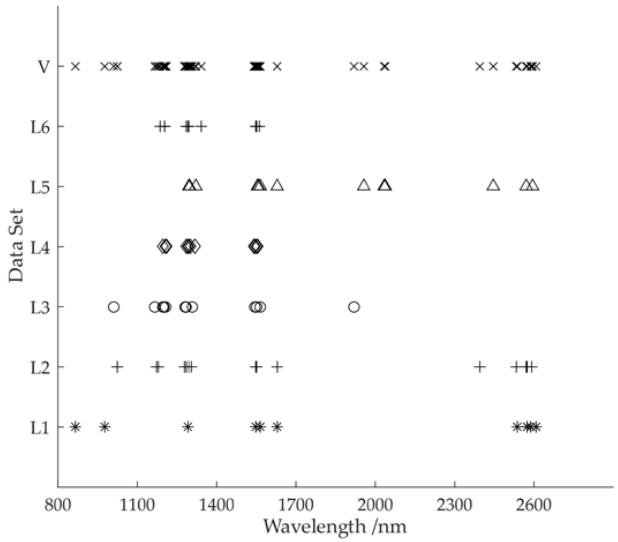

Figure 6. The variables selected by the DWT-BOSS coupling algorithm. Note: L1 refers to the reconstructed spectral average after one-layer wavelet transform, L2 to L6 are analogous to L1. V is the optimal variable set. Different graphs represent different levels of decomposition.

3.3. Establishment and Verification of the PLS Model Based on an Optimal Variable Set

The optimal variable set selected by DWT-BOSS was the independent variable of the tea leaves' moisture content prediction model, and the corresponding tea leaves moisture content was the dependent variable. The tea leaves moisture content prediction model (defined as L(i)-BOSS-PLS model, i = 1–6) was constructed. Due to the generation of random numbers, the model was run 30 times to verify the reliability of the model. In other words, 30 models were obtained in each layer. The optimal variable sets of L1–L6 were modeled, respectively, and the model of full-spectrum L0 was introduced for comparison. The V-PLS model was constructed to explore the model's accuracy and stability. By comparing the L(i)-BOSS-PLS algorithm with the V-PLS algorithm, we concluded that the accuracy and stability of the PLS models were improved. By analyzing the information in Table 1, we found the V-PLS model has the highest accuracy, with an R_c^2 of 0.9410, RMSEC of 0.2404, R_{cv}^2 of 0.9171, RMSECV of 0.2851, R_p^2 of 0.9513, and RMSEP pf 0.2236. In general, the L(i)-BOSS-PLS model produced a better effect than the L0-PLS model, with higher prediction accuracy and better model stability. The results obtained by running the program 30 times were within a reasonable range. The DWT-BOSS considerably reduces the amount of modeling computation and effectively improves the prediction ability of the model. In the V-PLS model established by the optimal variable set, 55 variables were

selected from 3112 variables for modeling, which greatly reduced the modeling time and improved the model accuracy. The method provides a reference for the selection of key bands for the near-infrared spectrum of Yinghong No. 9 tea leaves, providing an inversion of moisture content for other tea leaves.

Table 1. Results of different PLS models by different deposition layers.

Variable Set	R_c^2	RMSEC	R_{cv}^2	RMSECV	R_p^2	RMSEP	*n*_VAR
L0	0.9266	0.2713	0.8529	0.3855	0.9085	0.2349	3112
L1 + BOSS	0.9326 ± 0.0396	0.2588 ± 0.0689	0.9336 ± 0.0019	0.2581 ± 0.0037	0.9412 ± 0.0637	0.2355 ± 0.1064	9.8333 ± 2.8333
L2 + BOSS	0.9304 ± 0.0125	0.2641 ± 0.0229	0.9236 ± 0.0011	0.2768 ± 0.0020	0.9470 ± 0.0113	0.2247 ± 0.0230	16 ± 4
L3 + BOSS	0.9250 ± 0.0120	0.2742 ± 0.0213	0.9300 ± 0.0011	0.2650 ± 0.0021	0.9421 ± 0.0055	0.2351 ± 0.0110	10.5000 ± 3.5000
L4 + BOSS	0.9380 ± 0.0114	0.2494 ± 0.0220	0.9289 ± 0.0014	0.2672 ± 0.0011	0.9431 ± 0.0094	0.2328 ± 0.0199	14.1000 ± 2.9000
L5 + BOSS	0.9304 ± 0.0133	0.2641 ± 0.0244	0.9271 ± 0.0013	0.2705 ± 0.0012	0.9447 ± 0.0141	0.2292 ± 0.0298	14.1000 ± 2.9000
L6 + BOSS	0.9212 ± 0.0026	0.2811 ± 0.0047	0.9131 ± 0.0010	0.2953 ± 0.0017	0.9512 ± 0.0021	0.2158 ± 0.0047	14.9333 ± 2.9333
V	0.9410	0.2404	0.9171	0.2851	0.9513	0.2236	55

Note: *n*_VAR, number of variables; RMSEC, root mean square error of calibration; RMSECV, root mean square error of cross-validation; RMSEP, root mean square error of prediction; R_c^2, correction determination coefficient; R_{cv}^2, coefficient of determination of cross-validation; R_p^2, coefficient of determination of test set; statistical results are presented as mean value ± standard deviation for 30 runs.

3.4. Two Classical Methods Introduced to Establish PLS Models

To validate the prediction accuracy and stability of the prediction model, two classical algorithms for selecting variables based on the near-infrared spectrum were introduced for comparison with the performance of the proposed DWT-BOSS selection algorithm. Two classical variable selection methods are the GA and interval iPLS.

3.4.1. GA-PLS Prediction Model Built for Comparison with the Proposed Model

The main GA parameters were set as follows: population size of 32, window width of 10, maximum generation of 100, and mutation rate of 0.005. Due to the randomness of the GA, 30 modeling repetitions were used in this experiment for selecting the best results. As shown in Figure 7, the corresponding bands above the red dotted line were selected, for a total of 870 bands. As shown in Table 2, R_c^2 was 0.9318, RMSEC was 0.2617, R_{cv}^2 was 0.8908, RMSECV was 0.3287, R_p^2 was 0.9420, and RMSEP was 0.2421. Due to the complexity of the full-spectrum data, which contained redundant information and noise, the GA left the band closer to the moisture characteristics using the survival of the fittest rule, and the result was optimized and improved compared with the original spectrum. However, compared with the DWT-BOSS algorithm, the result still had redundant wavebands; the proportion of the number of the variables was about 1:16 (V:GA).

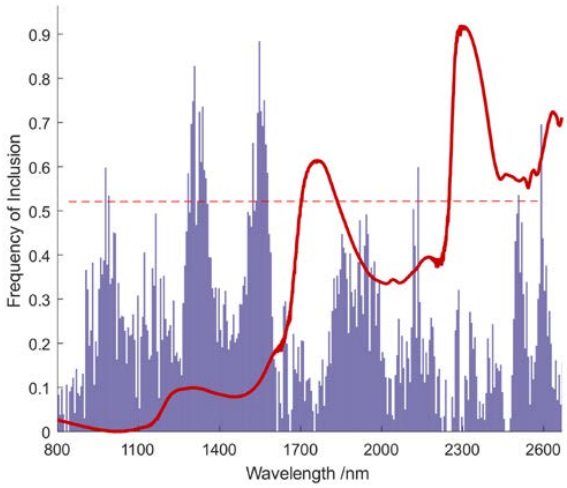

Figure 7. The frequencies of selected variables within 30 runs by the GA. Note: The selected wavelength is above the dotted line.

Table 2. PLS model was established by different wavelength selection methods.

Variable Set	R_c^2	RMSEC	R_{cv}^2	RMSECV	R_p^2	RMSEP	*n*_VAR
GA	0.9318	0.2589	0.8908	0.3287	0.9420	0.2421	870
iPLS	0.9294	0.2617	0.9021	0.3088	0.9232	0.2838	280
V	0.9410	0.2404	0.9171	0.2851	0.9513	0.2236	55

3.4.2. iPLS Prediction Model Built for Comparison with the Proposed Model

The modeling setting of the iPLS method was as follows: By moving windows, the interval size was set to 20. The result included 280 bands in total that were selected from 14 intervals, which were located near 850, 1200, 1350, 1600, 1800, 2200, 2400, 2500, and 2600 nm. As shown in Figure 8, about 280 variables were selected as the modeling objects to establish the iPLS model, whose R_c^2 was 0.9294, RMSEC was 0.2713, R_{cv}^2 was 0.9021, RMSECV was 0.3088, R_p^2 was 0.9232, and RMSEP was 0.2838. By selecting the interval, the bands with a stronger correlation with moisture characteristics were obtained, which increased the accuracy and stability of the model. However, the performance of the iPLS model was slightly worse than that of the DWT-BOSS-PLS model because the selected variables were still redundant bands.

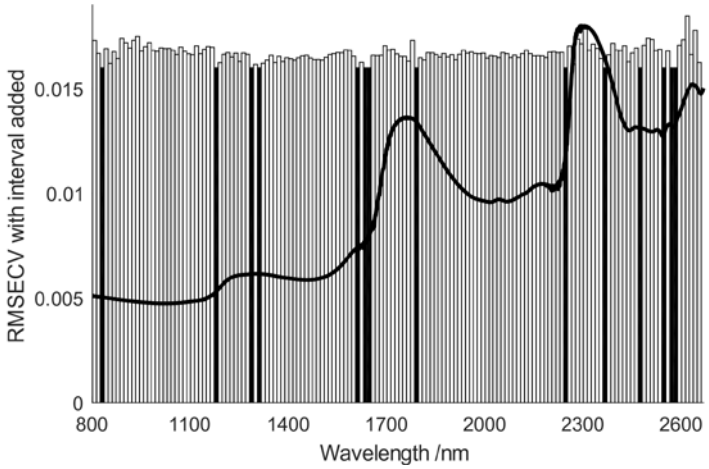

Figure 8. Several interval models (bars) and full spectrum models (solid lines) for cross-validation prediction errors (RMSECV). Note: The preferred interval is given in black.

As shown in Table 2, among the three different wavelength selection methods, the DWT-BOSS algorithm performed the best. The PLS model established using the DWT-BOSS algorithm not only had the best stability and prediction ability but also used the least number of wavelength points. In summary, the ranking of the number of selected variables was as follows: DWT-BOSS < iPLS < GA, whereas the ranking of the prediction accuracy was: DWT-BOSS > GA > iPLS.

4. Conclusions

In this study, a novel variable selecting algorithm based on DWT and BOSS was employed to select the optimal variable set of the moisture content of tea leaves for the Yinghong No. 9 variety. After selecting the optimal variables, a PLS prediction model was built. The prediction effect of this algorithm on the moisture content of tea leaves was explored. Some conclusions and contributions of this research are summarized as follows:

(1) In the DWT process, the noise was considerably removed. The band was calculated by the correlation coefficient method to select the maximum levels and the maximum levels of decomposition was found to be six. In general, the moisture-related spectrum of L6 was denoised but retained effective information.

(2) BOSS selected the effective information from the spectrum with the noise removed by DWT. By superimposing layers L1–L6 optimized by the BOSS algorithm, 55 optimal variables were obtained as the optimal variable set V. A considerable number of variables fell within the range of 800–1100, 1200–1400, and 1700–2000 nm. Those wavelength ranges were mainly represented by –OH groups and the vibration absorption of combinations and octaves. The number of variables dropped from 3112 to 55 using BOSS.

(3) Compared with full spectral modeling, DWT-BOSS-PLS had higher accuracy and prediction accuracy, with R_c^2 of 0.9410, RMSEC of 0.2404, R_{cv}^2 of 0.9171, RMSECV of 0.2851, R_p^2 of 0.9513, and RMSEP of 0.2236. GA and iPLS algorithms were used for comparison with the proposed DWT-BOSS method; the DWT-BOSS results had higher stability and accuracy, with fewer bands used.

(4) We proposed a novel prediction model that is robust and effective for forecasting the moisture content of Yinghong No. 9 tea leaves.

However, tea making still has difficulties of extensive application of NIRS technology, such as expensive machinery and equipment, learning to use NIRS technology, and the production line design and so on. Thus, NIRS technology needs to be popularized in tea processing factories. Furthermore, spectral equipment needs some suitable designs for production, while the entrepreneur is willing to pay for technological transformation.

Author Contributions: All authors contributed to the conceptualization. M.Z. performed the formal analysis and original draft preparation; J.R. and F.Z. finished sample preparation and data visualization; J.G. and G.Q. reviewed and made relative edits; E.L. and C.M. contributed to the supervision and provided resources. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the special project from the Provincial governor of Guangdong Province [yuecainong[2019]170], the special project from the Dean of Guangdong Academy of Agricultural Sciences [201939], Enhance the tea technological capability of cities and counties to promote industrial development projects [890-2020-XMZC-2593-01-0001].

Acknowledgments: The authors are grateful for the support of South China Agricultural University and Tea Research Institute of Guangdong Academy of Agricultural Sciences. The authors also thank the anonymous reviewers for their critical comments and suggestions to improve the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gilbert, N. Drink tea and be merry. *Nature* **2019**, *566*, S9.
2. Chen, S.X.; Luo, J.L.; Li, P.W. Study on the effect of cleaning technology of fresh tea leaves on the quality of green tea. *Trans. Tech. Publ. Ltd.* **2020**, *984*, 160–167. [CrossRef]
3. Caliskan, G.; Dirim, S.N. The effect of different drying processes and the amounts of maltodextrin addition on the powder properties of sumac extract powders. *Powder Tech.* **2016**, *287*, 308–314. [CrossRef]
4. Wang, Y.J.; Li, L.Q.; Shen, S.S.; Liu, Y.; Ning, J.M.; Zhang, Z.Z. Rapid detection of quality index of postharvest fresh tea leaves using hyperspectral imaging. *J. Sci. Food Agric.* **2020**, *10*, 2259. [CrossRef] [PubMed]
5. Pasquini, C. Near infrared spectroscopy: A mature analytical technique with new perspectives e a review. *Anal. Chim. Acta* **2018**, *1026*, 8–36. [CrossRef] [PubMed]
6. Xue, J.; Ye, L.; Li, C.; Zhang, M. Rapid and nondestructive measurement of glucose in a skin tissue phantom by near-infrared spectroscopy. *Optik* **2018**, *170*, 30–36. [CrossRef]
7. Sousa, P.; Soares, A.; Castanho, A.; So, A.; Oliveira, J.; Brites, C. Optimization of rice amylose determination by NIR-spectroscopy using PLS chemometrics algorithms. *Food Chem.* **2018**, *242*, 196–204.
8. Zhang, R.; Li, Z.; Pan, J.; Zhang, R.; Li, Z.; Pan, J. Coupling discrete wavelet packet transformation and local correlation maximization improving prediction accuracy of soil organic carbon based on hyperspectral reflectance. *Trans. Chin. Soc. Agric. Eng.* **2017**, *33*, 175–181.
9. Xie, S.; Xiang, B.; Yu, L.; Deng, H. Tailoring noise frequency spectrum to improve NIR determinations. *Talanta* **2009**, *80*, 895–902. [CrossRef]
10. Morgan, C.L.S.; Waiser, T.H.; Brown, D.J.; Hallmark, C.T. Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy. *Geoderma* **2009**, *151*, 249–256. [CrossRef]

11. Abasi, S.; Minaei, S.; Jamshidi, B.; Fathi, D. Rapid measurement of apple quality parameters using wavelet de-noising transform with Vis/NIR analysis. *Sci. Hortic. (Amst.)* **2019**, *252*, 7–13. [CrossRef]
12. Song, J.; Li, G.; Yang, X. Optimizing genetic algorithm-partial least squares model of soluble solids content in Fukumoto navel orange based on visible-near-infrared transmittance spectroscopy using discrete wavelet transform. *J. Sci. Food Agric.* **2019**, *99*, 4898–4903. [CrossRef] [PubMed]
13. Dun, L.; Chen, C.; Li, J.; Zhao, Y.; Chen, S.; Hou, Z. Research on robustness of support vector regression model base on near infrared spectroscopy of maize and Discrete wavelet transform. *J. Henan Inst. Sci. Technol. Sci. Ed.* **2017**, *45*, 43–47.
14. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection isabelle. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
15. Cai, W.; Li, Y.; Shao, X. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemom. Intell. Lab. Syst.* **2008**, *90*, 188–194. [CrossRef]
16. Jarvis, R.M.; Goodacre, R. Genome analysis Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. *Bioinformatics* **2005**, *21*, 860–868. [CrossRef] [PubMed]
17. Degenhardt, F.; Seifert, S.; Szymczak, S. Evaluation of variable selection methods for random forests and omics data sets. *Brief. Bioinform.* **2017**, *20*, 1–12. [CrossRef] [PubMed]
18. Rgaard, L.N.; Saudland, A.; Wagner, J.; Nielsen, J.P.; Unck, L.M.; Engelsen, S.B. Interval partial peast-squares regression (iPLS): A comparative chemometric study with an example from Near-Infrared Spectroscopy. *Appl. Spectrosc.* **2000**, *54*, 413–419. [CrossRef]
19. Jiang, H.; Liu, G.; Mei, C.; Yu, S.; Xiao, X.; Ding, Y. Measurement of process variables in solid-state fermentation of wheat straw using FT-NIR spectroscopy and synergy interval PLS algorithm. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2012**, *97*, 277–283. [CrossRef]
20. Deng, B.C.; Yun, Y.H.; Cao, D.S.; Yin, Y.L.; Wang, W.T.; Lu, H.M.; Luo, Q.Y.; Liang, Y.Z. A bootstrapping soft shrinkage approach for variable selection in chemical modeling. *Anal. Chim. Acta* **2016**, *908*, 63–74. [CrossRef]
21. Li, H.; Liang, Y.; Xu, Q.; Cao, D. Model population analysis for variable selection. *J. Chemom.* **2010**, *16*, 418–423. [CrossRef]
22. Wang, K.; Du, W.; Long, J. Near-infrared wavelength-selection method based on joint mutual information and weighted bootstrap sampling. *IEEE Trans. Ind. Inform.* **2020**, *3203*, 1–10. [CrossRef]
23. Yan, H.; Song, X.; Tian, K.; Chen, Y. Quantitative determination of additive Chlorantraniliprole in Abamectin preparation: Investigation of bootstrapping soft shrinkage approach by mid-infrared spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2017**, *191*. [CrossRef]
24. Zhang, Y.; Sun, J.; Li, J.; Wu, X.; Dai, C. Quantitative analysis of cadmium content in tomato leaves based on hyperspectral image and feature selection. *Appl. Eng. Agric. Vol.* **2018**, *34*, 789–798. [CrossRef]
25. Fei, Q.; Li, M.; Wang, B.; Huan, Y.; Feng, G.; Ren, Y. Analysis of cefalexin with NIR spectrometry coupled to artificial neural networks with modified genetic algorithm for wavelength selection. *Chemom. Intell. Lab. Syst.* **2009**, *97*, 127–131. [CrossRef]
26. Friedel, M.; Patz, C.D.; Dietrich, H. Comparison of different measurement techniques and variable selection methods for FT-MIR in wine analysis. *Food Chem.* **2013**, *141*, 4200–4207. [CrossRef] [PubMed]
27. Santos, B.; Fernandes, F.; Neto, G.; Kawakami, R.; Galva, H. NIR spectrometric determination of quality parameters in vegetable oils using i PLS and variable selection. *Sci. Direct* **2008**, *41*, 341–348.
28. Suhandy, D.; Yulia, M.; Ogawa, Y.; Kondo, N. Prediction of L-Ascorbic acid using FTIR-ATR terahertz spectroscopy combined with interval partial least Squares (iPLS) regression *. *Eng. Agric. Environ. Food* **2013**, *6*, 111–117. [CrossRef]
29. Li, X.; Sun, C.; Luo, L.; He, Y. Determination of tea polyphenols content by infrared spectroscopy coupled with iPLS and random frog techniques. *Comput. Electron. Agric.* **2015**, *112*, 28–35. [CrossRef]
30. Jiang, H.; Xu, W.; Chen, Q. Comparison of algorithms for wavelength variables selection from near-infrared (NIR) spectra for quantitative monitoring of yeast (*Saccharomyces cerevisiae*) cultivations. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *214*, 366–371. [CrossRef]
31. Yang, Z.; Xiao, H.; Zhang, L.; Feng, D.; Zhang, F.; Jiang, M.; Sui, Q.; Jia, L. Fast determination of oxides content in cement raw meal using NIR-spectroscopy and backward interval PLS with genetic algorithm. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *223*, 117327. [CrossRef]

32. Rizvi, T.S.; Mabood, F.; Ali, L.; Al-Broumi, M.; Al Rabani, H.K.M.; Hussain, J.; Jabeen, F.; Manzoor, S.; Al-Harrasi, A. Application of NIR spectroscopy coupled with PLS regression for quantification of total polyphenol contents from the fruit and aerial parts of citrullus colocynthis. *Phytochem. Anal.* **2018**, *29*, 16–22. [CrossRef]
33. Virginia, E.; Douglas, D.; Fernandes, D.S.; César, M.; De Araújo, U.; Henrique, P.; Dias, G.; Inês, M.; Maciel, S. Simultaneous determination of goat milk adulteration with cow milk and their fat and protein contents using NIR spectroscopy and PLS algorithms. *LWT Food Sci. Technol.* **2020**, *127*, 109427.
34. Ali, L.; Mabood, F.; Rizvi, T.S.; Rehman, N.U.; Arman, M.; Al-shidani, S.; Al-abri, Z.; Hussain, J.; Al-harrasi, A. Total polyphenols quantification in Acridocarpus orientalis and Moringa peregrina by using NIR spectroscopy coupled with PLS regression. *Chem. Data Collect.* **2018**, *13–14*, 104–112. [CrossRef]
35. Galvão, R.K.H.; Araujo, M.C.U.; José, G.E.; Pontes, M.J.C.; Silva, E.C.; Saldanha, T.C.B. A method for calibration and validation subset partitioning. *Talanta* **2005**, *67*, 736–740. [CrossRef]
36. Chemistry, P.A. Partial least-squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *186*, 1–17.
37. Costa, D.d.S.; Mesa, N.F.O.; Freire, M.S.; Ramos, R.P.; Mederos, B.J.T. Development of predictive models for quality and maturation stage attributes of wine grapes using vis-NIR reflectance spectroscopy. *Postharvest Biol. Technol.* **2019**, *150*, 166–178. [CrossRef]
38. Chen, D.; Chen, Z.; Grant, E. Adaptive wavelet transform suppresses background and noise for quantitative analysis by Raman spectrometry. *Anal. Bioanal. Chem.* **2011**, *400*, 625–634. [CrossRef]
39. Walczak, B.; Poppi, R.J.; Noord, O.E.D.; Massart, D.L.; Brussel, V.U.; Brussels, B. Application of wavelet transform to extract the relevant component from spectral data for multivariate calibration. *Anal. Chem.* **1997**, *69*, 4317–4323.
40. Ehrentreich, F. Wavelet transform applications in analytical chemistry. *Anal. Bioanal. Chem.* **2002**, *372*, 115–121. [CrossRef]