Advanced Science and Technology Letters
Vol.77 (UNESST 2014), pp.106-111
http://dx.doi.org/10.14257/astl.2014.77.20

Research of Soybean Straw Cellulose and Hemicellulose Near Infrared Model

Shen Weizheng,Wang Jian-bo, Kong Qingming, Guan Jing, Wang Jingjing
School of Electrical and Information Northeast Agricultural University Harbin, 150030, China
E-mail: wzshen@neau.edu.cn

Abstract. To achieve the rapid detection of soybean straw component, the key lies in establishing a quantitative analysis model with higher prediction accuracy which is rapid, stable and reliable. In order to establish the optimal Near-infrared (NIR) analysis model of cellulose and hemicellulose content in soybean straw, this paper uses NIR transmission technology by applying interval Partial Least Squares (iPLS) on the optimization of characteristic spectrum range of cellulose and hemicellulose spectrum. During the optimized characteristic spectrum range, prediction models of Partial Least Squares Regression (PLSR) and the Back Propagation Neural Network (BPNN) are built in the cellulose and hemicellulose contents respectively. The results show that the best modeling band of the Cellulose content is 5615-5731cm-1, and the optimal coefficient of determination of prediction model, PredictionR2(P-R2) reaches 0.9179266; And the best modeling band of the hemicellulose content is 5615-5731cm-1 ,the P-R2 is 0.920407. After the selection of iPLS optimal band, the quantitative analysis model of cellulose and hemicelluloses established by adopting the PLSR and BP Neural Network is more concise and has higher prediction accuracy and faster data computing speed. It also provides a theoretical basis for the optimization of characteristic spectrum range for the design of small dedicated NIR analytical instruments.

Keywords: Near-infrared spectroscopy; Soybean straw; Partial Least Squares Regression; Back Propagation Neural Network

1 Introduction

In recent years, China's increasing energy consumption has caused increasingly serious energy shortage. Countries around the world are actively seeking and developing new energy and renewable energy. As an agricultural superpower, biological resources are extremely rich in our country. Heilongjiang Province has a long history in planting soybean with geographical, ecological and economic advantages. With Perennial soybean planting area of approximately 3.664 million hm2 accounting for one third of the national soybean acreage, Heilongjiang Province is the main producing area in our country [1]. According to the statistics, every year at least half of the soybean straw is burned or discarded. The straw is used as the main raw material for making fuel ethanol and biodiesel. People has been using traditional chemical detection technology with low detection efficiency, long period and high

ISSN: 2287-1233 ASTL

cost for the quality analysis of straw, which largely affected the yield and quality of bio-fuels. NIR spectroscopy analysis technology as a new direction and a new method of rapid detection, has been widely applied in many fields such as agriculture, food, pharmaceutical, chemical, etc. [2]. Domestic and foreign scholars have demonstrated through extensive experimental study its feasibility in the detection of crop residues [3]. However, when using the bands of full spectrum in the model building for analysis, a lot of redundant information contained will have a significant impact on model performance, and raise higher demand for hardware devices in the future research and development of portable devices. This paper, based on the advantages of iPLS algorithm, focuses on the research of establishing and improving the calibration model, reducing the difficulty of analytical models, accelerating the rate of analytical models and improving the modeling.

2 Materials and Methods

2.1 Sample collection and preparation

211 Soybean straw samples for experiments are from around the Heilongjiang Province which cover different regions, different climates, different soil types and different varieties in the province. After the straw samples being dried 48 hours under natural condition to remove the surface moisture of them, they are crushed by 9FQ-360 hammer mill and through a 60 mesh sieve. Samples are put in sealed bags and stored at room temperature in the dark for spectral acquisition and laboratory chemical analysis.

2.2 Spectral acquisition

Antaris II NIR Spectrometer of Thermo Company has been used in experiments to scan the spectrum of soybean straw samples with an integrating sphere scanning, Scanning range of 4000-12000cm-1(780 ~ 2500nm), Resolution of 4cm-1, air as a comparative object, at room temperature(20 to 22℃). The background scanning is set as 64 times, and scanning frequency is set as 64 times in the experiments. The abscissa is the wave number from 4000 to12000cm-1, the ordinate is the absorbance, and the data is stored as log 1/R.

2.3 Chemical Analysis

The chemical analysis of cellulose and hemicellulose in the Soybean straw is determined by the principle of Van Soest [4] method. Three parallel samples of each sample are measured and averaged. The contents of cellulose and hemicellulose are represented as %. The distribution of the content of the sample is as shown in Table 1.

Table 1. Soybean straw samples distribution statistics

components	No.	Min(%)	Max(%)	Average(%)	SD(%)
cellulose	196	37.7410	49.45694	43.0962	3.00
hemicellulose	196	18.22959	29.61555	24.05298	3.16

3 Results and Discussion

3.1 Spectral Data Preprocessing

The NIR spectra of Soybean straw powder are shown in figure 1. Due to the high frequency random noise, baseline drift, samples uneven, surface scattering and other factors affect the modeling results[5]. The raw spectra collected need the necessary mathematical preprocessing like smoothing and derivative and so on. This study adopts the method of preprocessing with Mean-Centering Correction, Savitzky-Golay Smoothing(SG), Normalization, Multiplicative Scatter Correction (MSC), the First Derivative (1st-Der) and the Second Derivative(2st -Der).

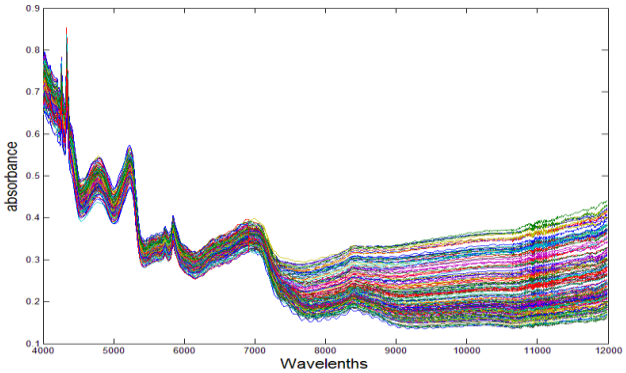

Fig.1 Spectrum of Soybean straw

3.2 Interval Partial Least Squares Band Selection

In the process of iPLS calculation,First of all, the intervals is set to 10,20,40,70,100 over the the entire spectral range. By establishing PLS models,the modeles establish by the lower Root Mean Squared Error of Cross-Validation(RMSECV) bands at each interval region selected compare with the full spectrum, select the characteristic bands,for the model fitting ability has certain upgrade. the intervals of Cellulose and hemicellulose spectrum in 70 cases (5615-5731 cm^{-1} and 4231-4342 cm^{-1}) works

best.the RMSECV of the model in this region are 0.9308331 and 0.8943536 respectively.

3.3 Evaluation of prediction models

Partial Least Squares Regression (PLSR) is a method of multivariate statistical data analysis, it mainly studies the regression modeling of multiple dependent variables and multiple independent variables. Analysis and the establishment of the experimental model are based on the Matlab software.First of all, the sample use Kennard-Stone algorithm to calculate the Euclidean Distance between sample spectra absorbance to choose the most representative samples as calibration set, 146 is chosen as the final calibration set an 50 samples as a validation set. Then, the optimal band spectrum separate in the order according to the calibration and validation sets, And the separate spectra were processed by the above-mentioned preprocessing methods. Finally, the data are treated as independent variables and build predictive models of cellulose and hemicellulose content respectively based on PLSR Cross Validation methods. The effects of predictive are shown in Figure 2.

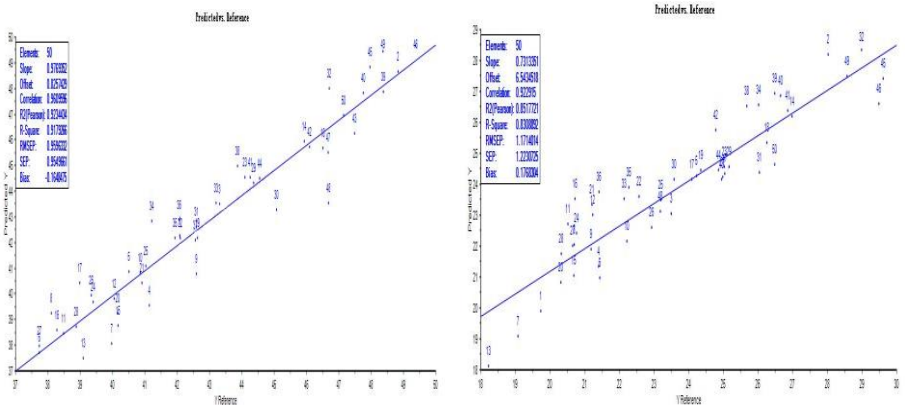

Fig.2. Cellulose modeling and prediction

As can be seen from the figure, model established with cellulose content,when the PLS factor is 4, has a high coefficient of determination and the minimum Root Mean Square Error,its Calibration coefficient R2 (C-R2) and Prediction coefficient R2(P-R2) are 0.9355296 and 0.9179266 respectively and its Root Mean Square Error of Calibration(RMSEC)and Root Mean Square Error of Prediction (RMSEP)are 0.9252655 and 0.956332 respectively.However, model established with Hemicellulose content, the C-R2 and P-R2 are 0.8550375 and 0.8308892 and its RMSEC and RMSEP are 0.8943536and 1.1714014 respectively.The predictive ability and model results of the Hemicellulose content model do not meet our expectations requirement.The following we will attempt to establish based on BP neural network prediction models of cellulose and hemicellulose content.

The establishment of BP neural network model, the Optimal network parameters of model should do further discuss.All the parameters selection will be in accordance with the Relative Standard Deviation (RSD) as the Standard. finally select hidden layer nodes 9, momentum factor 0.5, learning rate 0.05, training times 5000 and accuracy 0.01 as the Optimal network parameters to modeling. And a validation set of 50 samples of cellulose and hemicellulose content do the prediction, the results show in Figure 3.

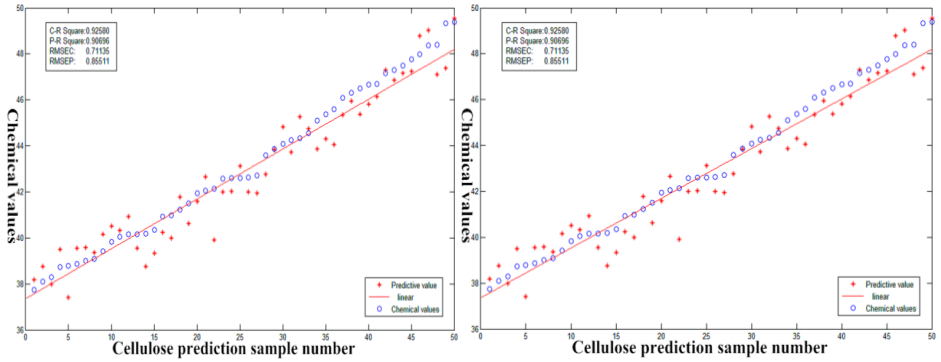

Fig.3. Prediction of Cellulose and Hemicellulose

Model established with cellulose content. its C-R2 and P-R2 are 0. 92580 and 0.90696 and its RMSEC and RMSEP are 0. 71135 and 0.85511 respectively. The prediction results can be seen the PR2 of BPNN and PLSR is not much different. Then look the model established with hemicellulose content. its C-R2 and P-R2 are 0.927889 and 0.920407,the RMSEC and RMSEP are 0.8363 and 0.8159 respectively. Compared with the PLSR model, we found that the model established by BP Neural Network has a higher coefficient of determination and a lower RMSEP,it also meet our expected results.

4 Conclusion

In order to establish the optimal Near-infrared(NIR) analysis model of soybean straw cellulose and hemicellulose content, using NIR transmission technology by applying interval Partial Least Squares (iPLS) on the cellulose and hemicellulose spectrum optimization of the characteristics spectrum. In the optimization region, the cellulose and hemicellulose contents are built Partial Least Squares Regression (PLSR) and the Back Propagation Neural Network (BPNN) prediction model respectively. The results show that this method can create high precision, low RMSEP NIR predictive model of cellulose and hemicellulose content.When the intervals is 70, the effect of PLSR prediction model established with cellulose content in the sub-intervals14 (5615-5731cm^{-1}) is the best and the predictive ability of P-R^2 reaches 0.9179266. Similarly When the intervals is 70, the effect of PLSR prediction model established

with hemicellulose content in the sub-intervals 3(5615-5731cm^{-1}) is the best and the predictive ability of P-R^2 even up to 0.920407.By iPLS selection of cellulose and hemicellulose spectrum characteristic absorption band，not only can establish more accurate calibration model, but also provides a theoretical basis for Small Near-infrared soybean straw composition analyzer.

References

1. Chuangzhi Wu, Zhaoqiu Zhou,Xiuli Yin, Biomass energy development present situation and the thinking in China,Transactions of the Chinese Society for Agricultural Machinery, 40(1):91-99(2009)
2. Kenneth P. Vogel, Bruce S. Dien, Hans G. Quantifying Actual and Theoretical Ethanol Yields for Switchgrass Strains Using NIRS Analyses. Bioenerg. Res.4; 96-110. (2011)
3. Lu Liu, X. Philip Ye, Alvin R. Womac. Variability of biomass chemical composition and rapid analysis using FT-NIR techniques.Carbohydrate Polymers. 81(4): 820-829.(2010)
4. Goering H K, Van Soest P J. Agric. Handbook 379. ARS. USDA. Washington D. C. 1970, 1.
5. Burns D A, Ciurczak E W. Handbook of Near 2 Infrared Analysis, Second Edition, Revised and Expanded. New York: Marcel Dekker. 431(2001)