International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.6 (2015), pp.187-194
http://dx.doi.org/10.14257/ijmue.2015.10.6.19

Research on Analysis Model of Soybean Straw Component

Weizheng Shen[1], Jianbo Wang[1], Qingming Kong[1],Jing Guan[1,] Jin Cui[1] and Ziqing Liu[2]

1.School of Electronic Engineering and Information, Northeast Agricultural University, Harbin, 150030, China
2.Agricultural Power Company of Huachuan Electric Power Bureau, Jiamusi, 154300, China
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, PredictionR^2(P-R^2) reaches 0.9179266; And the best modeling band of the hemicellulose content is 5615-5731cm^{-1} ,the P-R^2 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; interval Partial Least Squares; 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 hm^2 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 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.* [2-6]. Domestic and foreign scholars have demonstrated through extensive experimental study its feasibility in the detection of crop residues [7-13]. However, when using the bands of full spectrum in

ISSN: 1975-0080 IJMUE
Copyright ⓒ 2015 SERSC

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 [14] method, combining with Wang Yuwan's improved method [15] and Contreras Lara et al [16] modified nylon bag 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. Statistics of Content Distribution in the Soybean Straw Samples

Sample components	Number	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 is shown in Figure 1. NIR spectra with multiple absorption peaks can be used as the basis for quantitative analysis. From 4000-12000cm^{-1} within the range of the soybean spectrum, it can be seen that the absorption peak of water is at 6896cm^{-1}(1450nm) and 5181cm^{-1}(1930nm). The main absorption peak of cellulose and hemicellulose in the soybean straw is at 5617 cm^{-1}(1780nm) and 1780cm-1(2336nm).

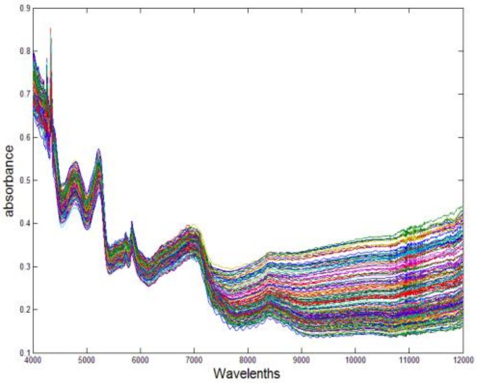


Figure 1. Spectrum of Soybean Straw

Due to the impact of high frequency random noise, baseline drift, uneven samples, surface scattering and other factors on the modeling, the original spectra collected need the necessary mathematical preprocessing like smoothing and derivation and so on so as to filter noise, improve signal interference noise ratio and eliminate interference of baseline drift. 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 (2nd-Der). By using Matlab programming modeling effect, Multiplicative Scatter Correction removing scattering interference between samples is better. Then the effect of 1st-Der 21 points for noise elimination is better than that of 2nd-Der 21 points. Finally SG elimination of signal burr of 1st-Der makes the spectrum more smoothing. The application of 1st-Der and S-G smoothing filter can bring more satisfactory results. In Figure 2 a is the original spectrum after MSC, b is the 1st-Der 21 points to eliminate noise after a and c is the smoothing 7 points after b.

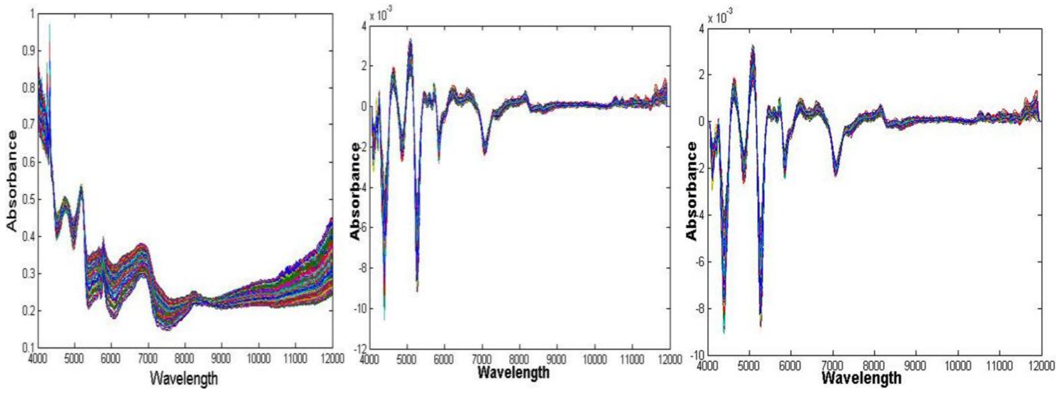


Figure 2. The Spectrum after Removing Noise

3.2 Interval Partial Least Squares Band Selection

To further determine the range of optimal spectral band, reducing redundant information in spectral bands uses iPLS to select the optimal spectral region. Full-spectrum region is divided into a number of subinterval of equal width by iPLS and interval PLS models are developed on them. The comparison is mainly based on the validation parameter Root Mean Squared Error of Cross-Validation (RMSECV). Select the interval of the smallest RMSECV as the optimal [18].

In the process of iPLS calculation, first of all, the intervals are set as 10, 20, 40, 70, and 100 over the entire spectral range. The iPLS uses these interval spectral regions as input

Copyright ⓒ 2015 SERSC

variables to establish PLS model respectively. The comparisons between the optimal RMSECV values of each interval area in the spectrum region of the model established allows us to better determine which interval way to get the better spectrum. The Optimal RMSECV in each region is as shown in Table 2. It can be seen, in the process of interval area selection, the RMSECV of full spectrum is the highest. The RMSECV trend of the intervals from 10-70 is decreasing, when subdivided, that is, from 70-100, it appears an increasing trend. And when intervals 200 is used, over-dense sampling have a greater impact on modeling, models are not representative and have destructive effect. So a lower RMSECV of intervals (such as 70 or 100) is mainly chosen for bands selection.

By establishing PLS models, the models established by the lower RMSECV bands at each interval region selected are compared with the full spectrum model. The characteristic bands selected can improve the model fitting ability, and it is the best when the interval of Cellulose and hemicellulose spectrum is at 70 (5615-5731 cm^{-1} and 4231-4342 cm^{-1}). The RMSECV of models in this region are 0.9308331 and 0.8943536 respectively. At the same time, these two bands are near the absorption peak we mentioned above.

Table 2. Optimal Range of Different Intervals

	Method	Factors	Interval (cm-1)	RMSECV	R2
Hemicellulose	PLS(full-spectrum)	3	4000-12000	1.027049	0.8161405
	iPLS(10 intervals)	7	4771-5538	1.3273025	08097639
	iPLS(20 intervals)	6	4000-4381	1.0800382	0.8243063
	iPLS(40 intervals)	8	4188-4381	1.0730224	0.8367089
	iPLS(70 intervals)	**6**	**4231-4342**	**0.8943536**	**0.8550375**
	iPLS(100intervals)	7	4227-4304	0.983099	0.8375943
Cellulose	PLS(full-spectrum)	3	4000-12000	1.317049	0.8751405
	iPLS(10 intervals)	4	4000-4767	1.2273025	0.9097639
	iPLS(20 intervals)	5	4385-4767	1..3400382	0.9043063
	iPLS(40 intervals)	5	4188-4381	1.030224	0.9167089
	iPLS(70 intervals)	**6**	**5615-5731**	**0.9308331**	**0.9325952**
	iPLS(100 intervals)	7	7081-7158	0.983099	0.9275943

3.3 Evaluation of Prediction Models

To evaluate the effectiveness of the regression model, the experiment applies the model Calibration coefficient R^2 (C-R^2), Root Mean Square Error of Calibration (RMSEC), Prediction coefficient R^2 (P - R^2), Root Mean Square Error of Prediction (RMSEP) and other indicators as the basis to make a comparative analysis of the model and evaluate the model predictions [18].

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 [19]. When the internal variables are high linear correlation, the PLSR is more effective. Moreover, the PLSR solve the problems such as sample number less than the number of variables better. The modeling of PLSR analysis has integrated the characteristics of methods like Principal Components Analysis, canonical correlation analysis and linear regression analysis so as to provide a more reasonable regression model.

Copyright ⓒ 2015 SERSC

The analysis and establishment of experimental models are based on the Matlab software. First of all, the samples 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 calibration set and 50 samples as a validation set. Then, the optimal band spectrum is separated in the order according to the calibration and validation sets, and the separate spectra are processed by the above-mentioned preprocessing methods. Finally, the processed data are treated as independent variables and build predictive models of cellulose and hemicellulose content respectively based on PLSR Cross Validation methods. The modeling and predictive effects of models are shown in Figure 3 and Figure 4.

As can be seen from the figure, the model established with cellulose content, when the PLS factor is 4, has a higher coefficient of determination and the minimum Root Mean Square Error, its C-R^2 and P-R^2 are 0.9355296 and 0.9179266 respectively and its RMSEC and RMSEP are 0.9252655 and 0.956332 respectively. However, the model established with Hemicellulose content, when the PLS factor is 3, the C-R^2 and P-R^2 are 0.8550375 and 0.8308892 respectively and its RMSEC and RMSEP are 0.8943536 and 1.1714014 respectively. The predictive ability and modeling results of the model established with Hemicellulose content do not meet our expectation. The following we will attempt to establish based on BP Neural Network prediction models of cellulose and hemicellulose content.

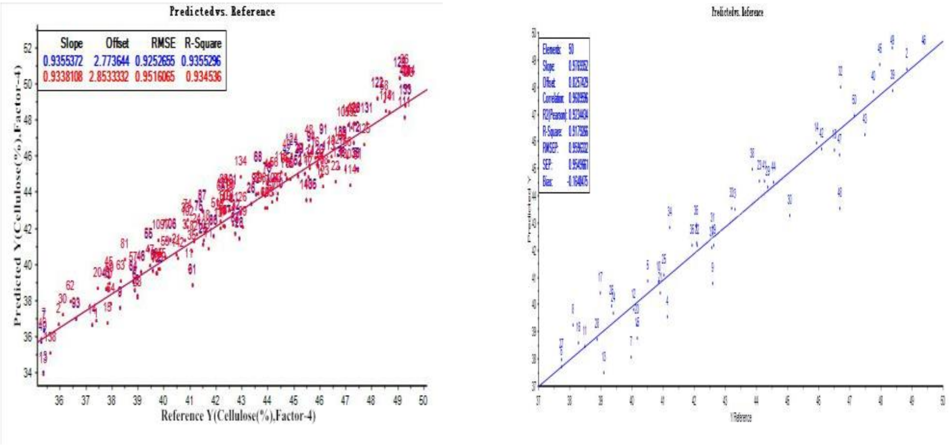


Figure 3. Cellulose Modeling and Prediction

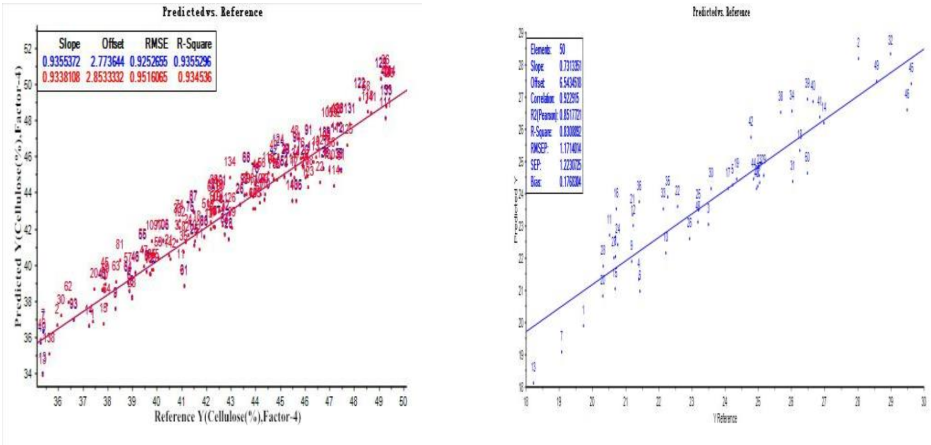


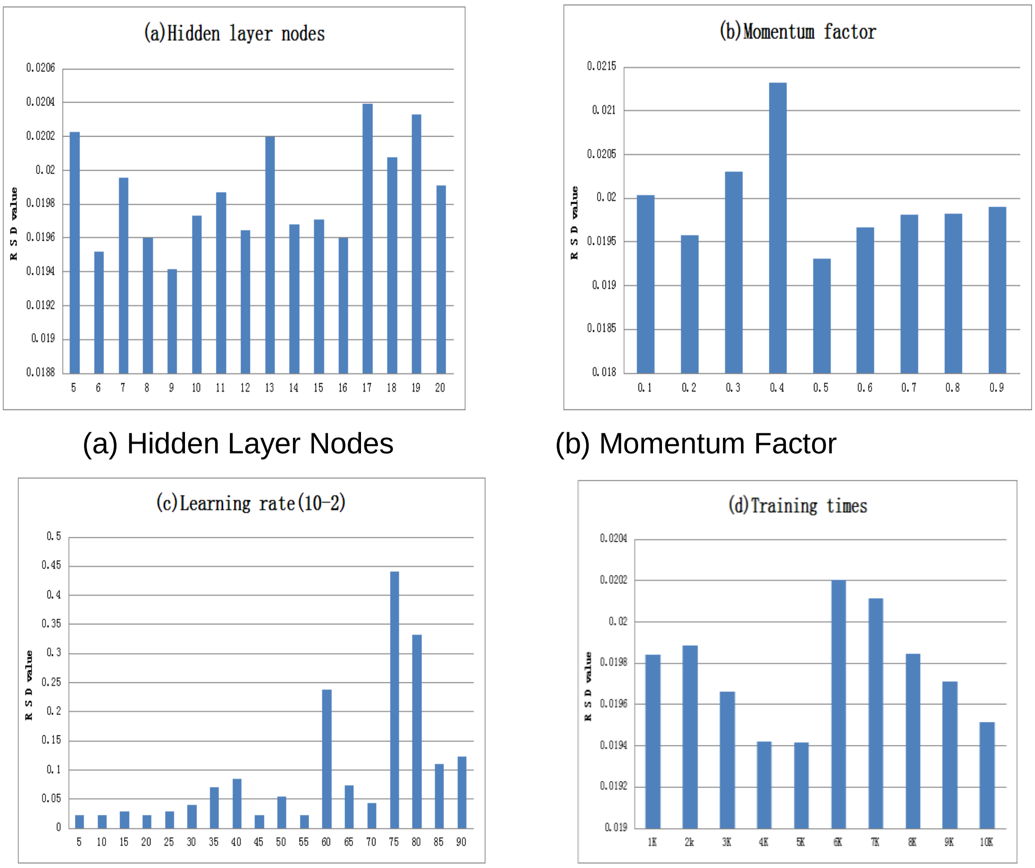
Figure 4. Hemicellulose Modeling and Prediction

Copyright ⓒ 2015 SERSC

3.4 BP Neural Network Modeling

Currently, artificial neural network is a frequently used nonlinear model. Back Propagation Neural Network (BPNN) based on error back propagation algorithm is widely used. Due to a significant nonlinear ability to process information, it has been widely applied in all areas [20].

The establishment of BP Neural Network model needs a further discussion for the optimal network parameters of model. All the parameters selected will be in accordance with the Relative Standard Deviation (RSD). As can be seen from the Figure 5, hidden layer nodes of 9, momentum factor of 0.5, learning rate of 0.05, training times of 5000 and precision aimed of 0.01 are finally selected as the optimal network parameters for modeling. Make predictions on a validation set of 50 samples of cellulose and hemicellulose content, the results are shown in Figure 6.



(a) Hidden Layer Nodes

(b) Momentum Factor

(c) Learning Rate (10-2)

(d) Training Time

Figure 5. Optimal Network Parameters of BP Model

For the model established with cellulose content, its C-R^2 and P-R^2 are 0. 92580 and 0.90696 respectively, and its RMSEC and RMSEP are 0.71135 and 0.85511 respectively. The prediction results can be seen the P-R^2 of BPNN and PLSR is not much different, but RMSEC and RMSEP have been obviously improved. Then for the model established with hemicellulose content, its C-R^2 and P-R^2 are 0.927889 and 0.920407 respectively, and 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.

Copyright ⓒ 2015 SERSC

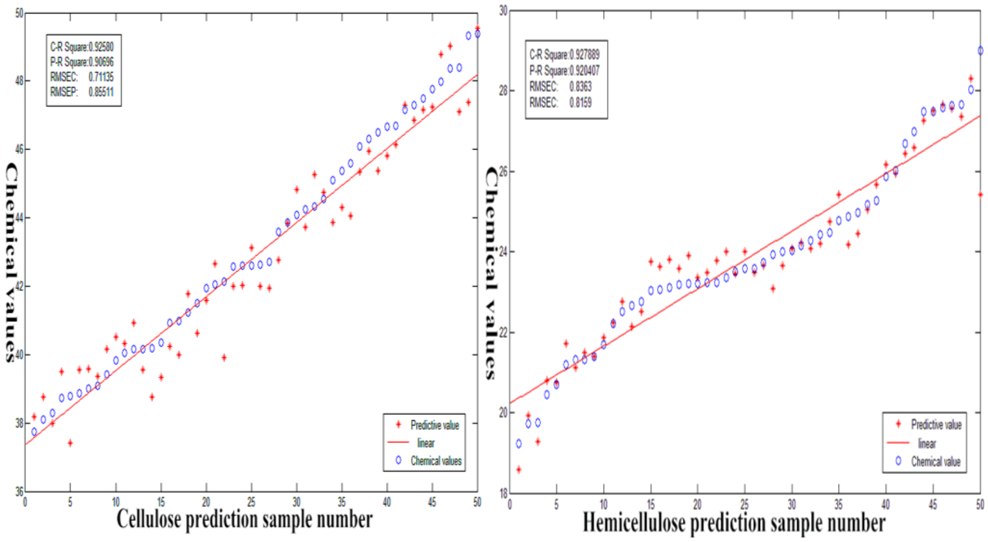


Figure 6. Prediction of Cellulose and Hemicellulose

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 NIR predictive model of cellulose and hemicellulose content with high precision and low RMSEP. When the interval 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 interval is 70, the effect of BP 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 reaches up to 0.920407. The optimization of absorption band of spectrum characteristics of cellulose and hemicelluloses can not only establish a more accurate calibration model, but also provide a theoretical basis for Small Near-infrared soybean straw composition analyzer.

Acknowledgements

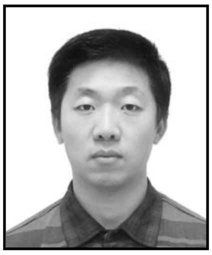
This research was supported by the National High-tech R&D Program of China (863 Program)(2013AA102303) and the Natural Science Foundation of Heilongjiang Province of China(F201402) and Northeast Agricultural University Innovation Foundation For Postgraduate(yjscx14002) and the Key Technologies R & D Program of Harbin （2013AA6BN010）.

References

[1] Chuangzhi W., Zhaoqiu Z. and Xiuli Y., “Transactions of the Chinese Society for Agricultural Machinery,” vol. 40 no. 91, **(2009)**.
[2] Kenneth P. V., Bruce S. D., Hans G. B., vol. 4 no. 96, **(2011)**.
[3] Gokhan H., Bismark L., Mark S., “Journal of Agricultural and Food Chemistry,” vol. 58 no. 702, **(2010)**.
[4] Michael M. Blanke,Springer-Verlag Berlin Heidelberg. 5, 19**(2013)**
[5] Yande L., Xudong S., Hailiang Z., “Computers and Electronics in Agriculture,” vol. 71 no. 10, **(2010)**.
[6] Haiqing Y., “Procedia Environmental Sciences,” vol. 10 no. 666, **(2011)**.
[7] Lu L., X. P. Ye, Alvin R., “Womac.Carbohydrate Polymers,” vol. 81 no. 820, **(2010)**.

[8] A. Belanche, M. R. Weisbjerg, G. G. Allison, “Journal of Dairy Science,” vol. 96 no. 7867, **(2013)**.
[9] Carina J. L., Mette H. T., Erik S. J., “Bioresource Technology,” vol. 101 no. 1199, **(2010)**.
[10] Feng X., Jianming Y., Tesfaye T., “Applied Energy,” vol. 104 no. 801, **(2013)**.
[11] F. Xu, Y.-C. Shi, D. Wang, “Carbohydrates Polymers,” vol. 88 no. 1147, **(2012)**.
[12] Hui J., Guohai L., Xiahong X., “Microchemical Journal,” vol. 102 no. 68, **(2012)**.
[13] Sander B., Jacob W. J., “Industrial Crops and Products,” vol. 31 no.321, **(2010)**.
[14] Goering H. K, Van S. P. J., “Agric. Handbook 379. ARS. USDA,” Washington D. C., **(1970)**.
[15] Wang Y. W., Xu W. Y., “Chinese Bulletin of Microbiology,” vol. 14 no. 8, **(1987)**.
[16] Contreras L., Gutié rrez C. D. L., Valdivia M. I., Arch. Z., vol. 48 no. 351, **(1999)**.
[17] Burns D. A., Ciurczak E. W., Marcel D., New York **(2001)**.
[18] Nφrgaard L., Saudland A., Wagner J., “Applied Spectroscopy,” vol. 54 no. 413, **(2000)**.
[19] Liang L., Yang M., Zhang L., “Transactions of the CSAE,” vol. 28 no. 162, **(2012)**.
[20] Wold S, Ruhe A., Wold H.S., “Journal on Scientific and Statistical Computing,” vol. 5 no. 735, **(1984)**.

Author



Weizheng Shen, (1977), male, Ph.D., professor, mainly engaged in the research and application of information technology in agriculture.