ISSN No. 0976-5697

Volume 5, No. 8, Nov-Dec 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Based on Random Forest Regression NIR Wavenumber Selection and BP Neural Network Modeling

Qiang Qin*
College of Science
Guilin University of Technology,
Guilin, China

Yinhe Bai
College of Foreign Language
Guilin University of Technology
Guilin, China

Qiwu Jiang
College of Science
Guilin University of Technology,
Guilin, China

Abstract: The paper uses near-infrared spectral analysis to predict fat content of 99 corn samples. At first random forests regression is used to building the model, and its variable importance index(VIP) is used to filter wavenumber. Then selecting 20 from the 390 wavenumbers. In order to judge whether the wavenumber of the selected are applicable to other models, three methods of regression (decision tree regression and BP neural network and the random forest regression) are used to build models respectively in the whole spectrum and 20 optimal wave number. Finally 6 kinds of model are established, and after VIP selecting, by comparing the optimal models what is left is BP neural network model :r^2 is 0.985，RMSEP is 0.089.

Keywords: Random forest VIP BP Decision tree regression.

I. INTRODUCTION

In recent years, near-infrared spectral analysis has been widely developed in fields like agricultural products, medicine, chemistry by its high efficiency, low cost and non-destructive analysis and many other features and advantages [1]. However, data analysis is extremely difficult for the features of the near-infrared spectra like complex information, the peak of spectra serious overlap as well as the weak absorption intensity [2]. So how to use mathematical methods to process the data and how to build the best model have become priorities in near-infrared technology.

The selection of independent variables(wavenumbers) would have a significant impact on the results. Some independent variables' contribution to the model are little, some are even negative ones or the presence of multicollinearity between the wavelength. All these can make the result poor. So It is necessary to select variables by some mathematical methods, in order to gain variables of "less but best". There are two methods to reduce wavenumbers: one of them is selecting optimal wavenumbers IPLS, MWPLS [3, 4], etc. The other one is choosing Discretely selected wavenumbers by Stepwise regression, Continuous projection, No information variable elimination and so on. This study attempts to use a new machine learning methods-random forest regression to build model, according to the variables importance to select the optimal variables to improve forecasting effect. In addition, different modeling methods are also important factors in the results of predicting slightly superior. This paper uses three different methods to model the best models respectively in the full spectra and VIP selected variables.

II. THE EXPERIMENT AND THE METHOD

A. The Experimental Data Collection and Processing

Firstly, preparing 99 corn samples. Secondly, making them into powder through physical methods. Thirdly Using Fourier infrared spectra analyzer to collect spectra data. The same sample with different frequencies of light have different response. Wave number range of 10000cm^{-1}~4000cm^{-1}(cm^{-1} is the wave number unit). There are totally 390 wave numbers. All the spectral response of the sample data is shown in figure 1. To complete the spectra modeling to realize rapid detection, we use conventional biochemical method to detect the fat content value of the corn samples. Its range is between 3.8% ~ 8.7%. All the ways of modeling are achieved by R language and MATLAB.

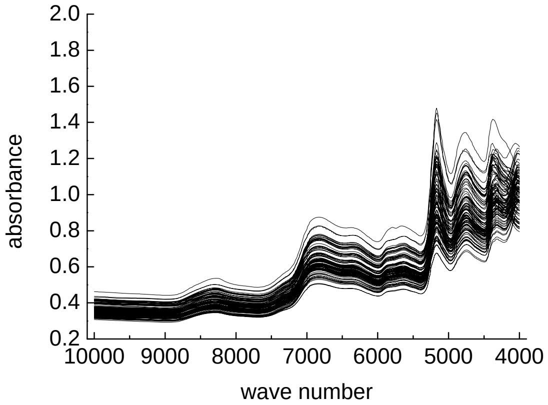

Figure 1. Near-infrared spectra of 99 corn samples

B. Random Forest Theory

Random Forest is a new machine learning method which is put forward by Leo Breiman on the basis of the theory of Decision trees in 2001.

The Algorithm of Random Forest [5,6,7]

(1) Generating n training sample by using random sampling techniques from the M samples of Calibration.

(2) Using these samples to generate n Decision trees.

(3) Finally, voting or averaging to form classification model system. The results of the paper obtained after averaging, the random forest regression prediction results.

Random forests can not only build regression model but select the most optimal variables to establish model according to the variable importance index [8]. The variable importance is calculated by the error increase of the bag outside [9]. Specific algorithm is as follows:

Calculating each Decision tree *i*'s errOOB(From the calculating of the rest samples of Random sampling). Random permutation to variable X_j(adding noise), calculating new error. The VIP $errOOB^*$ from the formula:

$$VIP(X_j) = \frac{1}{n}\sum_i (errOOB - errOOB^*)$$

III. RESULTS AND DISCUSSION

A. Data Processing Before Modeling

Before the setting up the near infrared spectrum model, the spectral preprocessing should be made. Spectral preprocessing methods usually be used are Data Enhancement Transformation, Smooth, Derivative, Standard Normal Variable Transformation, Multiple Scattering Correction, The Fourier Transform, The Wavelet Transform etc. These spectra pretreatment methods are just to process the data of spectrum itself, but does not take into account the influence of the density matrix. So when the preprocessing begin, it is likely to lose some useful information for setting up of calibration model, may be could not completely eliminate noise and affect the quality of the model. This study uses a deduction of the electromagnetic spectrum and concentration signal algorithm--OSC[10] to process the spectral data. Dividing 99 samples into calibration set and prediction set in a ratio of about 3:1 by KS [11], 74 samples of the calibration , 25 samples of prediction set. As shown in table 1, sample that can be seen that the selection of calibration setting the fat content of all the samples with the maximum and minimum values are consistent. The mean value and variance are very close. This shows that the chemical values of calibration set samples selected covers all the range of the sample.

Table 1: Statistics of fat content

Fat chemical measurements /%				
	MAX	***MIN***	***MEAN***	***VAR***
Cabibration set	8.700	3.700	4.654	0.430
Prediction set	8.500	3.600	4.889	0.421
All the samples	8.700	3.700	4.638	0.450

B. Random Forests Parameters Optimization

Before Random forests regression modeling, What the first should be done is studying the two important parameters of Random Forests : the number of decision trees and the influence of division number variable m_{try} to modeling to determine the optimal parameters [12]. Fixing splitting variable m_{try} number first, then selecting from 1 to 500 the number of decision tree, building 500 models. As can be seen from the figure 4, With the increase of the number of trees, Outside the bag error tends to a stable level. This shows that the random forests won't be over-fitting due to the increase of the trees. This is the advantage of random forests. So in the following models, selecting enough trees will be okay(This article selected 500 decision tree). And then doing research to divide the number of variables, fixing the number of decision tree 500, choosing from 1 to 401 divided variable. Seeing from figure 2, with the increase of split variable the error of bag outside firstly decreases and then increases . At 130, the minimum value will be achieved. It is close to 1/3 of the total number of variables, consistent with theory[13].

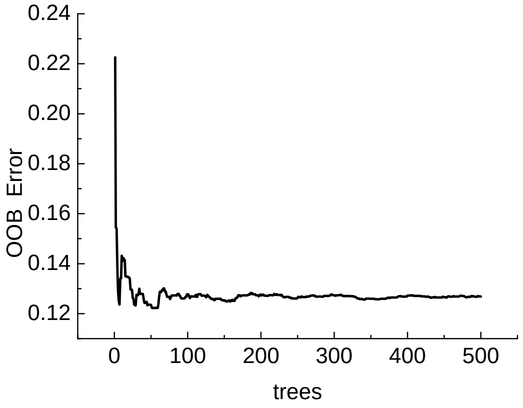

Figure 2. Relationship between RMSEP and nTree of RF model

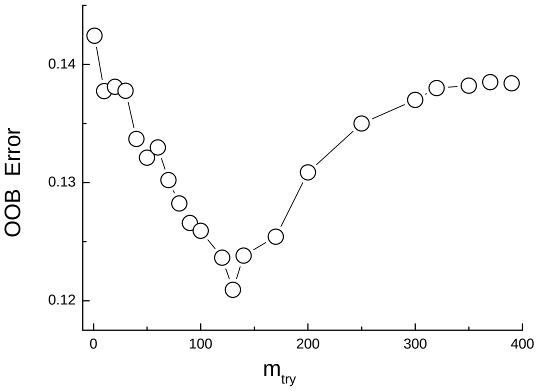

Figure 3. Relationship between RMSEP and mtry of RF model

C. Wavenumber Selection and Modeling

Setting the calibration by random forest regression modeling to get VIP value of each wave number. As shown in Figure 4, some of the wave number in modeling process are negative, others are positive. From the figure, VIP wave number greater than 2 mainly focus between 7000cm^{-1}-4500cm^{-1}. After comparing with the original spectrum, it can be seen that the spectral wave spectrum area fluctuates significantly, rich information content. Smaller values of the VIP area near the area of 4000cm^{-1} and 10000cm^{-1}-7000cm^{-1}.

Comparing with the spectrum, area near 4000cm^{-1} the noise is very large. Area near 10000cm^{-1}-7000cm^{-1} the absorption peak area is very weak. This suggests that using VIP value selected wavenumber selection process and practice of medium wave number are the same. So this mathematical method can be used to select the wave number.

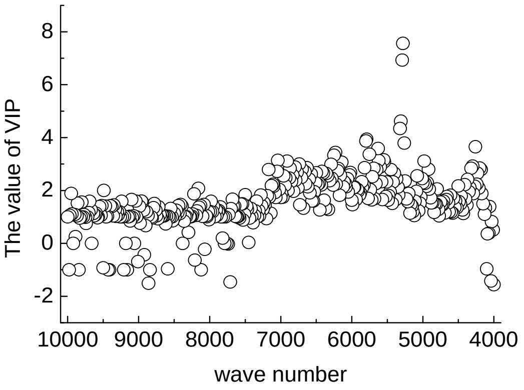

Figure 4. VIP value for each wave number

In order to get the optimal wave number for modeling, rearranging each wave number according to the VIP value size to get new spectrum matrix is necessary. By this way, the wave number which are more near the former are more important. Selecting the former n wave number for model to make the RMSEP's the smallest n former wave number are the optimal wave counts. In order to reduce the calculation, in this paper, the initial value 5, geometric sequence of common ratio 2 will be used to select the value of n. So n_1 = 5, n_2 = 10, n_3 = 20, n_4 = 40, n_5 = 80, n_6 = 160, n_7 = 320, n_8 = 640. As shown in figure 5, when n_3 = 20 , the modeling of the effect is the best. The selection of optimal wave number as shown in figure 6. According to the VIP value, the order of wave number : 5280cm^{-1}, 5290cm$^{-1,}$ 5310cm^{-1}, 5320cm^{-1}, 5790cm^{-1}, 5800cm$^{-1,}$ 5260cm$^{-1,}$ 4260cm^{-1}, 5630cm^{-1}, 6230cm^{-1}, 5750cm^{-1}, 6250cm^{-1} , 5550cm^{-1}, 7040cm^{-1}, 5620cm^{-1}, 4980cm^{-1}, 6910cm^{-1}, 5540cm^{-1}, 6140cm^{-1}, 6740cm^{-1}。 From these waves, chosen number near the strong absorption peak can be found clearly.

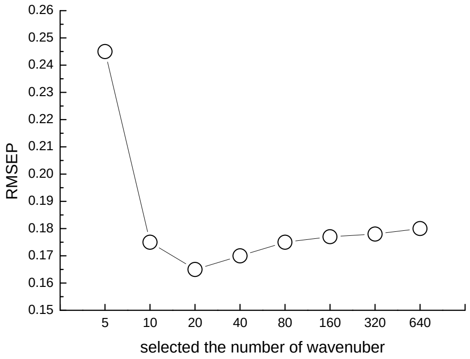

Figure 5. Relationship between RMSEP and selected the number of wavenumber

Table 2: different modeling methods

method		R^2	*RMSEC*	r^2	*RMSEP*
VIP selected wavelength	Random Forest	0.957	0.135	0.949	0.165
Full Spectra		0.955	0.137	0.941	0.18
VIP selected wavelength	BP	0.984	0.091	0.969	0.115
Full Spectra		0.908	0.258	0.834	0.353
VIP selected wavelength	decision tree	0.65	0.593	0.59	0.546
Full Spectra		0.6	0.665	0.589	0.546

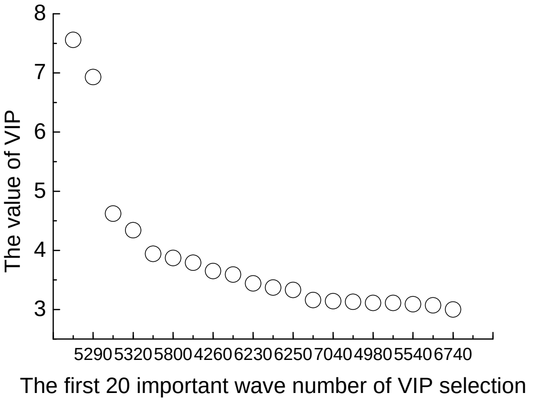

Figure.6 the 20 most important variables corresponding to the values of VIP

First getting the optimal wave number through the above ways, then comparing different regression methods to obtain the optimal model. In the next study, Random Forests Regression, The BP Neural Network Regression[13], The Decision Tree Regression[14], these three methods in optimal long point and within the scope of full spectrum of the wave number are used to establish model, and a total of 6 kinds of model. BP adopts three layers network, in the middle of the hidden layer nodes are eight. The results are in table 3. It can be seen that the decision tree regression is the worst regression model. This is because the decision tree is a kind of weak learning with simple structure, large prediction error. BP and Random Forest Regression modeling results are ideal. 20 optimal wave count by the results of three different regression modeling methods are better than in the whole spectral range. This shows that optimization of wave number is not only applicable to The Random Forest Return but also suitable for other regression method. Finally got the best model: The BP Neural Network Regression Model is on optimal wave count. Real value and predictive value of scatter diagram are shown in figure 8. r^2=0.969, RMSEP=0.115 .

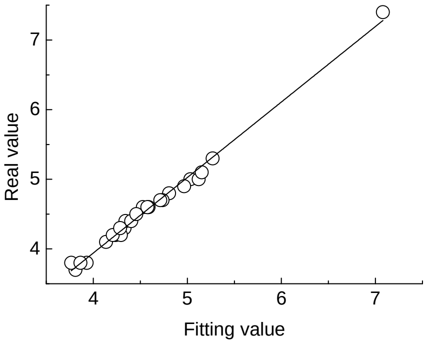

Figure 7. The scatter of true value and fitting value

IV. CONCLUSION

Random Forest regression is a good modeling method, and more importantly, by the importance of variables, it can reduce the number of variables, optimize models. And transferring this method to other models can also optimize other models. In this article, the application of BP neural network combined with random forests variable importance after modeling is the best of the effect of three kinds of modeling methods.

V. REFERENCES

[1] Yan YanLu,Zhao Long-Lian, Han Dong-Hai. Elements and application of near-infrared spectra analysis. Beijing: China Light Industry Press. pp.188—235, April 2005.

[2] LU Wanzhen. Modern near-infrared spectroscopy analytical technology (2nd ed).Beijing: China Petroch-emical Press, 2007.

[3] L.Norgaard. “Interval partial least-squares regression (iPLS): A comparative chemometrics study with an example from near-infrared spectroscopy”. Applied Spectroscopy, vol.3, 2000, pp.413-419.

[4] J.H.Jiang,R.J.Berry, H.W.Siesler, Y.Ozaki. “Wavel- ength interval selection in mulicomponent spectral analysis by moving windows partial least-squares regression with application to mid infrared and near infrared spectroscopic data”. Analytical Chemistr-y.vol.14, 2002, pp. 3555-3565.

[5] Breiman L. “Random Forests”. Machine Learning, vol.45, 2001, pp. 5-32.

[6] Dietterich T G. “An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization”. Machine Learning, vol.45, 2000, pp. 139-157.

[7] Fukuda S, Yasunaga E, Nagle M. Modelling the relationship between peel colour and the quality of fresh

[8] mango fruit using Random Forests [J]. Journal of Food Engineering, vol.131. 2014, pp.7-17.

[9] L. Zheng, D.G. Watson, B.F. Johnston, R.L. Clark, R. Edrada-Ebel, W. Elseheri, “A chemometric study of chromatograms of tea extracts by correlation optimization warping in conjunction with PCA, support vector machines and random forest data modeling”, Anal. Chim. Acta 642 , 2009, pp. 257–265.

[10] Fearn T. “On orthogonal signal correction. Chemometrics and Intelligent Laboratory Systems”, vol.50, 2 000, pp.47-52.

[11] Chen Huazhou, Ai Wu, Feng Quanxi. “FT-NIR Spectroscopy and Whittaker Smoother Applied to Joint Analysis of Duel-Components for Corn” . Molecular and Biomolecular Spectroscopy, vol.118, 2014, pp.752-759.

[12] Simon. Data mining tutorial. Beijing: Tsinghua university press, 2006.

[13] Liu Kea, Guo Wenyanb, Shen Xiaoli,“Research on the Forecast Model of Electricity Power Industry Loan Based on GA-BP Neural Network”. Energy Procedia,Vol.14, 2012, pp. 1918–1924

[14] Mao Guojun. Principle and algorithm of data mining . Beijing: Tsinghua university press, 2007.