Discriminate of Moldy Chestnut based on Near Infrared Spectroscopy and Feature Extraction by Fourier Transform

Zhu Zhou[1], Xiaoyu Li[1*], Peiwu Li[2], Yun Gao[1], Jie Liu[1], Wei Wang [1]

[1]College of Engineering, Huazhong Agricultural University, Wuhan, P. R. China

[2]Oil Crops Research Institute of China Agricultural Science Research Institute, Wuhan, P. R. China

zhouzhugcxy@webmail.hzau.edu.cn, lixiaoyu@mail.hzau.edu.cn, peiwuli@public.wh.hb.cn, gaoyun@webmail.hzau.edu.cn, liujie11028@yahoo.com.cn, wangwei@mail.hzau.edu.cn

Abstract. As near infrared spectra has the characters of multi-variables and strong correlations, to solve the problem, Fourier transform (FT) was used to extract feature variables of shelled chestnuts spectra. FT coefficients and the status of 178 chestnuts were selected as inputs and outputs of the back-propagation neural network (BPNN) classifier to build a recognition model. For comparison, principal component analysis (PCA) was utilized to compress the variables, which then was introduced as input of the neural network model. The results demonstrate that FT is a powerful feature extraction method and is better than PCA as a feature extraction method when employed together with BPNN. When the preprocessing method of standard normal variate transformation(SNV) was carried out and the first 15-point FT coefficients were used as the input, an optimal network structure of 15-6-1 was obtained, where discriminating rates of qualified chestnut, surface moldy chestnut and internal moldy chestnut in prediction set are 100%、100% and 92.31%, respectively.

Keywords: Near infrared spectroscopy, Fourier Transform, feature extraction, BP neural network, chestnut

* Corresponding Author

1 Introduction

Chestnut is one of the most popular nuts in the world and China is the biggest chestnut producer. It is reported that the annual yield of chestnuts in China is ca. 9.25×105 metric tons (in 2007) which accounts for 75.61% of the total world yield [1]. But chestnuts, which are rich in carbohydrates and low in fat, are susceptible to getting moldy after harvest. In China, manual sorting or brine floatation is the primary method to pick out moldy and spoiled chestnuts, which proves low sorting efficient and high misjudgment rate [2]. Therefore, finding a fast, effective and applicable method to sort moldy chestnut is urgently required.

Near infrared (NIR) spectroscopy can record the response of the molecular bonds (e.g. C–H, N–H and O–H) of chemical constituents to near infrared radiation and thereby build a characteristic spectrum that performs as a fingerprint of the sample [3, 4]. Being nondestructive, simply applicable and fast, it requires minimal sample processing prior to analysis [5]. NIR spectroscopy has become a rapid and well-established technique for the quantitative and qualitative analysis of agricultural products. However, NIR spectra typically consists of broad, weak, non-specific, and extensively overlapped bands, and may have hundreds or thousands of wavelength variables [4, 6]. The use of all variables for classification purposes is not an adequate strategy because it produces the so-called “curse of dimensionality” [7]. Moreover, some of these variables may include useless or irrelevant information for calibration model like noise and background, which can worsen the predictive ability of the whole model [8]. Therefore, the data dimensionality needs to be reduced.

Principal component analysis (PCA) [9-12] and Fourier transform (FT) [13-15] can be applied to reduce the dimensionality of the NIR data. PCA is data set dependent, whereas FT is independent of the data set. It means that with PCA a whole data set is simultaneously treated, while with FT each spectrum is treated individually. If changes occur in one spectrum, this does not affect the FT of the other spectra, but it does affect in PCA [13, 14].

In our previous work[2], NIR spectroscopy and PCA were used to discriminate moldy chestnut. This work aims to study the use of FT to reduce data dimensionality for discriminate classifier and to compare the results with that of PCA method.

2 Materials and methods

2.1 Sample preparation

Chestnuts used in the experiment were from Macheng, Hubei Province in China. The weight scope of chestnuts was between 8.50g ~ 20.41g. After purchased, they were stored according to Chinese commercial profession standard SB/T10192-1993. Samples were divided into two categories: qualified chestnut and moldy chestnut which include surface moldy chestnut and internal moldy chestnut, the judgment of which is made in accordance with GH / T 1029-2002 requirements. To determine internal moldy chestnut, they should be hulled after spectral measurements. All the samples were laid at room temperature (25 ℃, 60% relatively humidity) for 24 h to equilibrate to experiment environmental before spectra collection. Finally, 69 qualified chestnuts, 66 surface moldy chestnuts and 43 internal moldy chestnuts were analyzed.

2.2 Spectral measurement

NIR diffuse reflectance spectra of chestnut samples were collected by a FT NIR spectrometer (Vector 33, Bruker Optics, German). The system consists of a gold-plated integrating sphere, a sample rotator, a 12 mm quartz glass and a PbS detector.

The FT NIR spectrometer was completely software-controlled by the OPUS software Version 5.0 which was provided by Bruker Optics. The spectra of chestnut samples were acquired between 12000 cm^{-1} to 4000 cm^{-1} at 8 cm^{-1} spectral resolution, taking the average of 64 scans and were analyzed at room temperature. Three replicates of each sample were taken and their mean values were calculated by using OPUS. Reference spectrum for air and dark spectrum were measured and stored prior to sample spectra measurement.

2.3 Fourier transform

Fourier transform is usually used in signal processing. An NIR spectrum is a signal measured in the wavelength domain. FT enables transitions between the wavelength and frequency domain [13].

If $f(1)$、$f(2)$,..., $f(N)$ represents the recorded spectral values at N equally spaced wavelengths, denoted by 1, 2, 3, λ, then, the discrete Fourier transform of the signal is defined as:

$$F(w)=\sum_{k=0}^{N-1} f(k)\cdot\exp(-\mathrm{j}2\pi wk/N)=\sum_{k=0}^{N-1} f(k)\cdot\left[\cos(\mathrm{j}2\pi wk/N)-\sin(\mathrm{j}2\pi wk/N)\right] \quad (1)$$

With $\mathrm{j}=\sqrt{-1}$ and the w =1, 2, ⋯, N.

After Fourier transform, the magnitude of $F(w)$ is defined as:

$$\left|F(w)\right|=\left|R^2(w)+I^2(w)\right|^{1/2} \quad (2)$$

where $R(w)$、$I(w)$ are the real and imaginary parts of $F(w)$, respectively.

In order to save computing time, the fast Fourier transform (FFT) can be used when the number of variables equals to 2^n, where n is a positive integer.

2.4 Back-propagation neural network

The back-propagation neural network (BPNN), owing to its excellent ability of non-linear mapping, generalization, self-organization and self-learning, it has been proved to be of widespread utility in pattern recognition [16-20]. BPNN is a three-layered feed forward architecture. The three layers are input layer, hidden layer and output layer. It is trained by repeatedly presenting a series of input/output pattern setting to the network. The network gradually “learns” the input/output relationship of the interest by adjusting the weights to minimize the error between the actual and predicted output patterns of the training set. The trained network is usually examined through a separate set of data called the train set to monitor its performance and validity. When the mean squared error (MSE) of the train set reaches a minimum, network training is considered complete and the weights are fixed [21]. There are many training algorithms for back propagation neural network, for example,

Gauss–Newton method, gradient descent algorithm and so on. However, an inappropriate algorithm can cause a wide variety of performance problems, e.g., divergence, slow convergence or local minimum trapping. Levenberg–Marquardt training (LM) algorithm was originally designed to serve as an intermediate optimization algorithm between the Gauss–Newton method and gradient descent algorithm, and it addressed the limitations of each of those techniques [2, 21].

In this paper, the training of the BPNN was done with LM algorithm. The transfer function of hidden layer was tansig function and the one of output layer was logsig function. The train function was trainlm. The goal error was set as 0.001. The time of training was set as 1000. The optimal architecture of neural network can be achieved by adjusting nodes of the hidden layer.

3 Results and analysis

3.1 Processing of NIR data

Figure.1 shows average spectrums of the qualified chestnut, surface moldy and the internal moldy chestnuts between the wave number range from 12 000 cm^{-1} to 4 000 cm^{-1}. As it can be seen from the figure, the spectral shape of three chestnut samples has little difference, and the spectrum of the qualified and internal moldy chestnuts overlap in the range of 12000~9000 cm^{-1}, which increases the difficulty of identifying the internal moldy chestnut. For reducing noise, offset and bias, 6 kinds of preprocessing techniques including smooth(Savitzky–Golay method, gap size = 9 data points), vector normalization (VN), max-min normalization (MMN), standard normal variate transform (SNV), first derivative(Savitzky–Golay method, gap size = 17 data points，FD) and no process(NP) were applied to the original spectrum respectively. The experiments were carried out on a range of 11895 cm^{-1} to 4000 cm^{-1} with a total of 2048 data points so that FFT could be used to the spectrums.

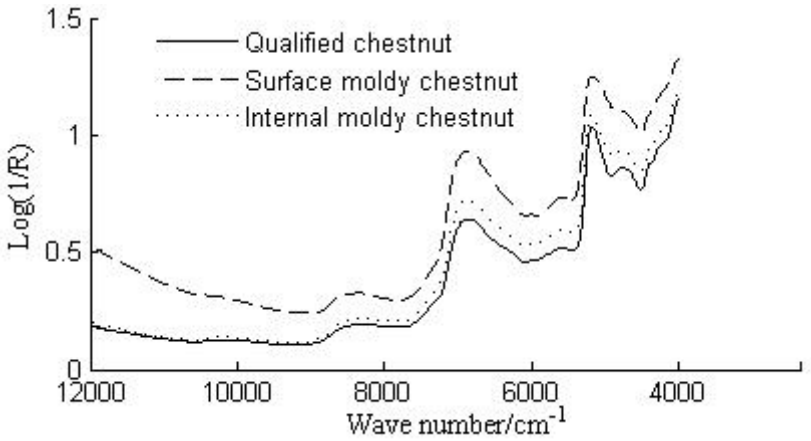

Fig. 1. Chestnuts mean spectra from raw data

3.2 Fourier feature extraction of Chestnuts NIR spectrum

When BP neural network applies to pattern recognition, and if the input has too many characteristic quantities, it will reduce the network training speed and efficiency, and lead to non-convergence in severe case. McClure [22] pointed out that if one transforms the NIR data, most of the information is in the range of the first 50 Fourier coefficients and the remainder can be discarded because it is mainly noise, so fewer Fourier coefficients can instead of the original spectral data, make the spectral dimension reduction.

In this paper, we applied FT to transforming the processed NIR data from the wavelength domain into the frequency domain, and we used the first 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 points of Fourier coefficients in the Fourier spectra respectively as the input of the BP neural network classifier.

3.3 The BPNN predicting model

19 qualified chestnuts、18 surface moldy chestnuts and 13 internal moldy chestnuts from each variety were selected randomly as the prediction set. The remaining 128 samples (50 qualified chestnuts、48 surface moldy chestnuts and 30 internal moldy chestnuts involved) were used as the training set to build the training model which was validated by the samples in the prediction set.

Fourier coefficients after different preprocessing were respectively used as the input of BPNN to build the model, with the figures 0 and 1 expressed the qualified chestnuts and moldy chestnuts separately. Deviation value was set to ±0.1, if the difference of true value and the predicted value was between ±0.1, it meant the correct identification, otherwise meant the mistake. Trial-and-error method were used to determine the number of node in hidden layer, when the highest correct discriminating rate reached, the best network structure were obtained.

Fig.2 shows the optimal results of BP neural network classifier with different processing methods and numbers of Fourier coefficients. Although the correct classification rates vary to processing and numbers of Fourier coefficients, the overall correct classification rates are higher than 80%. Since SNV and VN could avoid the effects caused by chestnut sizes and spectral scattering, the correct classification rates are higher than the other processing methods.

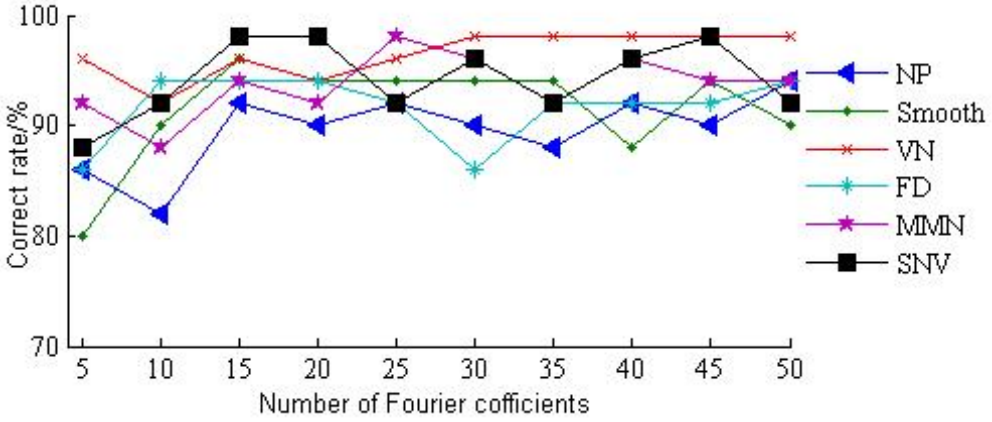

Fig. 2. The classification results of BPNN as a function of the number of the selected Fourier coefficients

According to the principle of minimum required network node, table 1 below listed the parameters when the highest correct classification rates(CCR) acquired under different processing methods and numbers of Fourier coefficients. It showed that, the qualified chestnut can be fully discriminated under NIR data processing. Different preprocessing methods such as smooth, SNV, MMN, VN, FD resulted in a greatly different discriminating rate. The highest discriminating rates of surface moldy chestnut and internal moldy chestnut were obtained by VN and SNV. The correct discriminating rates were 100%, 92.31%. When applying original spectrum，the

lowest discriminating rate of surface moldy chestnut was reached, which was 88.89%. The lowest discriminating rate of internal moldy chestnut obtained by FD method was only 76.92%.

Table 1. Parameters and CCRs of different preprocessing techniques under FT method

Method	Number of Fourier coefficients	Hidden layer nodes	Prediction set /%			Training set /%		
				moldy			moldy	
			qualified	surface	internal	qualified	surface	internal
NP	15	6	100	88.89	84.62	98	100	100
smooth	15	9	100	100	84.62	100	100	100
MMN	25	29	100	100	92.31	100	100	100
VN	30	17	100	100	92.31	100	100	100
FD	15	12	100	94.44	76.92	100	100	100
SNV	15	6	100	100	92.31	98	100	100

3.4 Comparison with PCA

In our previous work[2], six processing methods including smooth(Savitzky–Golay method, gap size = 9 data points), vector normalization(VN), min-max normalization(MMN), standard normal variate transformation(SNV), multiplication scattering correction（MSC）and first derivative(Savitzky–Golay method, gap size = 17 data points，FD), were processed to the original spectrum, then PCA method was used to reduce the dimension，and BPNN was utilized to establish the model. Table 2 below listed the recognition rate and parameters of BP neural network.

From a comparison of the results of table 1 and table 2, it was obvious that both PCA and FT achieve acceptable results, but FT has higher correct classification rates. By using PCA method, we can get the optimal predicting model when the NIR data was pretreated by the vector normalization (VN). The correct discriminating rates of qualified chestnut and internal moldy chestnut were 94.74%、94.74% and 92.31%, respectively. Obviously, the results were lower than the way that they were extracted by FT. However, the network structure of the BP model with principal component analysis is 7-4-1, which should be simpler than the 15-6-1 with Fourier feature extraction. Thus, for a further study, Fourier coefficients should be optimized by genetic algorithms (GA).

Table 2. Parameters and CCRs of different preprocessing techniques under PCA method

Method	Number of PCs	Hidden layer nodes	Prediction set /%			Training set /%		
				moldy			moldy	
			qualified	surface	internal	qualified	surface	internal
smooth	4	14	89.47	94.44	53.85	96	100	96.67
SNV	5	4	89.47	94.44	76.92	98	97.92	93.33
MSC	5	10	78.95	94.44	84.62	98	100	100
MMN	5	10	68.42	100	76.92	98	100	100
VN	7	4	94.74	94.44	92.31	98	100	100
FD	10	4	84.21	94.44	84.62	98	100	100

4 Conclusions

The application of BP neural network with NIR data, after different processing methods and transformation to FT coefficients, was studied. It was found that the preprocessing methods and the numbers of FT coefficients affect the CCR of the BPNN classifier. When preprocessing method of standard normal variate transformation was utilized and the first 15 point of FT coefficients were used as the input, an optimal network structure of 15-6-1 was obtained, where discriminating rates of qualified chestnut, surface moldy chestnut and internal moldy chestnut in prediction set were 100%、100% and 92.31%. It is better than the BPNN model which used vector normalization (VN) processing and PCA methods. As the Fourier feature extraction is not dependent on the spectral data set, only treats each spectrum individually, therefore, we recommend applying FT as a dimensionality reduction method in pattern recognition of NIR data.

Acknowledgement

The financial support provided by Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20090146110018) was appreciated.

References

1. Yao, Z.Y., Qi, J.H., Wang, L.H.: Equilibrium, kinetic and thermodynamic studies on the biosorption of Cu(II) onto chestnut shell. J. Hazard. Mater. 174, 137-143 (2010)
2. Zhou, Z., Liu, J., Li, X., Li, P., Wang, W., Zhan, H.: Discrimination of moldy Chinese chestnut based on artificial neural network and near infrared spectra. Transactions of CSAM. 40, 109-112 (2009)
3. Cozzolino, D., Liu, L., Cynkar, W.U., Dambergs, R.G., Janik, L., Colby, C.B., Gishen, M.: Effect of temperature variation on the visible and near infrared spectra of wine and the consequences on the partial least square calibrations developed to measure chemical composition. Anal. Chim. Acta. 588, 224-230 (2007)
4. Xie, L., Ying, Y., Ying, T.: Classification of tomatoes with different genotypes by visible and short-wave near-infrared spectroscopy with least-squares support vector machines and other chemometrics. J. Food Eng. 94, 34-39 (2009)
5. Chen, J., Arnold, M.A., Small, G.W.: Comparison of combination and first overtone spectral regions for near-infrared calibration models for glucose and other biomolecules in aqueous aolutions. Anal. Chem. 76, 5405-5413 (2004)
6. Blanco, M., Coello, J., Iturriaga, H., Maspoch, S., Pagès, J.: NIR calibration in non-linear systems: different PLS approaches and artificial neural networks. Chemom. Intell. Lab. Syst. 50, 75-82 (2000)
7. Acevedo, F.J., Jiménez, J., Maldonado, S., Domínguez, E., Narváez, A.: Classification of wines produced in specific regions by UV-Visible spectroscopy combined with support vector machines. J. Agric. Food Chem. 55, 6842-6849 (2007)
8. Han, Q., Wu, H., Cai, C., Xu, L., Yu, R.: An ensemble of Monte Carlo uninformative variable elimination for wavelength selection Anal. Chim. Acta. 612, 121-125 (2008)
9. Li, X., He, Y., Fang, H.: Non-destructive discrimination of Chinese bayberry varieties using Vis/NIR spectroscopy. J. Food Eng. 8, 357-363 (2007)
10. Liu, F., He, Y.: Classification of brands of instant noodles using Vis/NIR spectroscopy and chemometrics. Food Res. Int. 41, 562-567 (2008)
11. Cynkar, W., Dambergs, R., Smith, P., Cozzolino, D.: Classification of Tempranillo wines according to geographic origin: Combination of mass spectrometry based electronic nose and chemometrics. Anal. Chim. Acta. 660, 227-231 (2010)

12.Fernández-Ibáñez, V., Fearn, T., Soldado, A., Roza-Delgado, B.d.l.: Development and validation of near infrared microscopy spectral libraries of ingredients in animal feed as a first step to adopting traceability and authenticity as guarantors of food safety. Food Chem. 121, 871-877 (2010)
13.W. Wu, B.W., Penninckx, W., Massart, D.L.: Feature reduction by Fourier transform in pattern recognition of NIR data. Anal. Chim. Acta. 331, 75-83 (1996)
14.Pasti, L., Jouan-Rimbaud, D., Massart, D.L., Noord, O.E.d.: Application of Fourier transform to multivariate calibration of near-infrared data. Anal. Chim. Acta. 364, 253-263 (1998)
15.Xiao, W., Li, X., Li, P., Lei, T., Wang, W., Feng, Y.: Near-infrared spectral detection of soil moisture based on feature extraction of FFT. Transactions of CSAM. 40, 64-67 (2009)
16.Sadeghi, B.H.M.: A BP-neural network predictor model for plastic injection molding process. J. Mater. Process. Tech. 103, 411-416 (2000)
17.Panda, S.S., Chakraborty, D., Pal, S.K.: Flank wear prediction in drilling using back propagation neural network and radial basis function network. Appl. Soft Comput. 8, 858-871 (2007)
18.Archer, N.P., Wang, S.: Application of the back propagation neural network algorithm with monotonicity constraints for two-group classification problems. Decision Sci. 24, 60-75 (2007)
19.Lin, S., Tseng, T., Chou, S., Chen, S.: A simulated-annealing-based approach for simultaneous parameter optimization and feature selection of back-propagation networks. Expert Syst. Appl. 34, 1491-1499 (2008)
20.Bayar, S., Demir, I., Engin, G.O.: Modeling leaching behavior of solidified wastes using back-propagation neural networks Ecotoxicol. Environ. Saf. 72, 843-850 (2009)
21.Kermani, B.G., Schiffman, S.S., Nagle, H.T.: Performance of the Levenberg–Marquardt neural network training method in electronic nose applications. Senors Actuat. B-Chem. 110, 13-22 (2005)
22.Burns, D.A., Ciurczak, E.W.: Handbook of Near-Infrared Analysis. CRC Press. NewYork (2007)