BP-ANN Application to the Model Establishment of Determination Wheat Protein Using Near Infrared Spectroscopy

H C Chen¹,²,³, X D Chen² and Q P Lu²

¹ China Jiliang University, Hangzhou, 310018 Zhejiang, China
² State Key Laboratory of Applied Optics, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, 130033, Changchun, China
E-mail: Huacaichen@cjlu.edu.cn

Abstract. Near infrared spectroscopy calibration models of determination to the wheat protein concentrations were developed based on the back-propagation artificial neural network (BP-ANN). The spectra of 160 wheat samples were pretreated with detrend, normalization, then subjected to principal component analysis (PCA) to identify the principal components (PCs), the scores of the PCs were used as ANN input variables. Calibration models were established with the training set (80 samples) using various input variables and hidden nodes. The root mean square errors of prediction (RMSEP) of the models to the prediction set (80 samples) were used to optimize the models. The RMSEP became stabilized when the input variables were up to 5, but changed little with varying hidden nodes. The optimal model with 9 input variables, 1 hidden node lead to the lowest RMSEP of 0.2869% and the highest correlation coefficient (R) value of 0.980 for the prediction set. Comparison of the calibration models developed with training sets of various sizes found that the simulation degree of model decreased slightly but prediction capacity improved with the increase of the training set data size and the optimal training set size was 80 samples.

1. Introduction
Near-infrared (NIR) spectroscopy represents a very rapid and accurate method for the simultaneous measurement of different constituents in various food and agricultural products. Current application of NIR spectroscopy in agricultural products include the quantitative determinations of moisture, fat, oil, sugars, and protein in a wide range of products such as rice, wheat, and maize [1–5]. The main advantages of NIR spectroscopy are its speed, chemical-free, and no or little sample pretreatment [6]. Furthermore, NIR spectroscopy technology may be an efficient tool for real-time quality test of agricultural products and on-line or at-line industry processing control [7].

In quantitative analysis, multiple linear regression (MLR), principal component regression (PCR) and partial least-squares regression (PLSR) are the most widely used multivariate calibration methods. These calibration approaches assume that there were a linear relationship between the measured

³ To whom any correspondence should be addressed.
parameters for the sample and the intensity of its absorption band. With PCR and PLSR, the small deviation from linearity may be acceptable as they can readily be suppressed by pretreatments such as principal components analysis, de-trend, or standard normalization vector to the spectra. But in the presence of substantial non-linearity e.g., those that arise from scattered light or intrinsic non-linearity in the absorption bands, the models tend to give large prediction errors and call for mathematical processing of spectra or the use of alternative calibration procedures to correct non-linearity.

Artificial neural networks (ANNs) are the most widely used mathematical algorithms for overcoming non-linearity in NIR calibration which were more and more widely applied during the past several years. The networks are straightforward mathematical descriptions of what is currently known about the physical structure and mechanism of biological learning and knowledge. The ANN used in this work was a perceptron multilayer network with error back propagation as training scheme and the generalized delta rule for weighting. The topology of this network type affords the use of a variable number of layers. Each layer can contain one or more neurons or nodes, which can act with a linear or non-linear transfer function. The input layer contains as many neurons as variables are to be handled, the output layer, as many as parameters are to be determined. Between the input and output layer, a variable number of hidden layers can be inserted containing also a variable number of neurons. The most usual way of describing the architecture of a network is by using the notation \((i, h, o)\) where \(i\) is the number of nodes or neurons in the input layer, \(h\) is the number of neurons in the hidden layers and \(o\) is the number of neurons in the output layers. The number of neurons in the input and the hidden layers, and the transfer function should be optimized [8–11].

In this work, the BP-ANN was used to develop calibration models for determination the protein contents in wheat powder with NIR spectroscopy. Detrend and PCA were used to pre-treat the NIR spectra for getting the input variables of ANN. The number of neurons in the input and the hidden layers were optimized by the root mean square error of prediction (RMSEP) and the correlation coefficient of the prediction set. Furthermore, the calibration models in the context of varying calibration data size were investigated to see whether the relative performance of NIR calibration depends on these factors.

2. Materials and methods

2.1. Experimental apparatus and software
Nicolet Nexus® 870 Fourier transform near infrared spectrometer (Thermo Corporation USA) with an InGaAs detector, OMNIC E. S. P. 5. 2a software, TQ analyst V6 chemometrics analysis software. BP-ANN programs were written with Matlab 7.01 software (Mathworks®).

2.2. Experimental samples
A total of 160 wheat samples were collected from different regions in China with protein contents of 3.64% to 8.32%(w/w). Protein content of each sample was measured by automatic Kjeldahl method as the reference values. The samples were rearranged according to the protein content increment and divided into a training set and a prediction set. In order to investigate the dependence of the calibration models on the training set data size, calibration models were developed with variable data sizes (n) and validated with the same prediction set. The selection of samples for calibration was alternatively with the protein content values to ensure the representative of the training set and the prediction set.

2.3. NIR spectra collection
Wheat samples were dry milled to less than 80 meshes in a Cyclone mill. NIR reflectance spectra of 160 wheat samples were collected over 4 000-10 000 cm-1 (1000-2500 nm) spectral region with a Nicolet NEXUS 870 FT-NIR spectrometer equipped with an InGaAs detector. The sample cup with a quartz window (50 mm in diameter and 3 mm in thickness) was set up on a revolving accessory with a
center 10 mm away from the incident light. The spectra were collected with 32 cm⁻¹ data interval. Each sample spectrum was obtained by averaging 30 scans. All spectra were recorded as log(1/R) with respect to a ceramic reference standard.

2.4. Data processing
The calibration models were constructed by using BP-ANN. Detrend treatment was applied to the spectra to avoid the variations arising from base drift. Because spectrum for each sample has 390 data points, it was too large for directly use these data as input layer in an ANN, spectra variables were compressed by principal component analysis (PCA) firstly to get the principal component (PC) scores. The vectors of the PC scores were used as the input variables of the ANN. In order to identify the best principal components (PCs) describing the data matrix, the network leading to the lowest root mean square error of prediction for the prediction set was adopted as the optimal.

\[
\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n-1}}
\]

Where \( \hat{y}_i \) denotes the NIR predicted value and \( y_i \) the chemical reference method value of wheat protein content.

3. Results and Discuss

3.1. Structure of ANN
In this work, a BP-ANN consisting of three layers (input layer, hidden layer and output layer) was established. The logsig function was used for the transfer function between the input layer and the hidden layer, and the purelin function for that between the hidden layer and the output layer. When using the BP-ANN to simulate NIR calibration model, the number of neurons in the input and the hidden layers should be optimized. Usually NIR spectra data of training sample set are used for the input variables of ANN, but if the number of spectra data is very large, the input nodes will be too much. It will be time consuming for network simulation and lead to model overfitting. In this work, each spectrum of 160 wheat samples contained 390 data. The spectra were pretreated with de-trend, normalization firstly, and then subjected to principal component analysis (PCA), computing the scores for the principal components (PCs) that describe the body of spectra. The first several PCs were used as ANN input variables. Calibration models with varying PCs (input neurons) and varying hidden nodes were developed. Figure 1 and figure 2 present RMSEPs and correlation coefficients (R) of the prediction set by the models established with varying input nodes and hidden nodes. The RMSEP decreased dramatically with an increase of the input node when the input variables were less than 5. The least RMSEP of 0.2869% was got with 9 input nodes. The RMSEP changed little with an increase of the hidden nodes. The models with 1 hidden node and 9 input nodes had the least RMSEPs. Therefore, the structure of the optimal model was (i, h, o) = (9, 1, 1). The predicted results of the optimal model to the prediction set, the training set and the total samples set were shown on Figure 3 to Figure 5, with RMSEPs of 0.2869%, 0.2646% and 0.2760% and correlation coefficients of 0.98, 0.982 and 0.981 respectively.
Figure 1. SEP of the prediction set by the methods established with varying input nodes and hidden nodes.

Figure 2. Correlation coefficients (R) of the prediction set by the models established with varying input nodes and hidden nodes.
Figure 3. Correlation between determinations of wheat protein contents in the prediction set by the reference method and NIR method.

Figure 4. Correlation between determinations of wheat protein contents in the training set by the reference method and NIR method.
3.2. Calibration models with varying training data size

NIR spectroscopy technique is an indirect method based on the NIR spectra and reference values of calibration sample set. Hence, the selection of calibration sample set is very important for robust model establishment. The calibration sample set should be representative and accurate and with enough samples. Usually the more samples in calibration set, the more robust the model is. But because the samples collecting is time consuming and expensive, the calibration set should be suitable. In this work, BP-ANN calibration models with the optimal structure were established with varying training data size. We predicted the training set, the same prediction set of 80 samples and the total sample set with these models. In terms of both RMSEP and R, the simulation degree of the models decreased, but the prediction capacity increase with the training data size increase. The optimal data size was 80 samples (shown on table 1). It suggested that the models would be overfitting with too few training samples and had poor prediction capacity. The simulation degree and the prediction capacity tend to be stable when the training sample set was up to 40. This result suggested that the suitable training set size were 40 to 80 samples.

| Calibration data size | RMSEP/R of training set | RMSEP/R of prediction set | RMSEP/R of total sample set |
|------------------------|-------------------------|---------------------------|-----------------------------|
| 10                     | 0.0228/1.000            | 0.7075/0.884              | 0.6864/0.884               |
| 20                     | 0.2344/0.986            | 0.4157/0.955              | 0.4383/0.958               |
| 40                     | 0.2817/0.979            | 0.3055/0.977              | 0.2905/0.979               |
| 54                     | 0.2557/0.984            | 0.2963/0.979              | 0.2887/0.979               |
| 80                     | 0.2646/0.982            | 0.2869/0.980              | 0.2760/0.981               |
4. Conclusion
The back-propagation artificial neural network is a very useful algorithm for establishment non-linear calibration models with near infrared spectroscopy. In this work, we established calibration models of determination to the wheat protein concentrations with BP-ANN. The spectra of 160 wheat samples were pretreated with detrend, normalization, then subjected to principal component analysis (PCA) to identify the principal components (PCs), the scores of the PCs were used as ANN input variables. Calibration models were established with the training set (80 samples) using various input variables and hidden nodes. The root mean square errors of prediction (RMSEPs) of the models to the prediction set were used to optimize the models. The RMSEP became stabilized when the input variables were up to 5, but changed little with varying hidden nodes. The optimal model with 9 input variables, 1 hidden node lead to the lowest RMSEP of 0.2869% and the highest correlation coefficients (R) values of 0.980 for the prediction set. Training set data size influences the simulation degree and prediction capacity of models. In terms of both RMSEP and R, the simulation degree of the models decreased slightly with the training set size, while the prediction capacity increased with the data size. This result suggested that the BP-ANN were a powerful algorithm for establishment calibration models for wheat protein determination with near infrared spectroscopy.

Acknowledgments
This study was funded by the Jilin Provincial Scientific and Technological Plan Key Project (No. 20040324-2) and the Research Project of Department of Education of Zhejiang Province(2005 0382).

References
[1] Wu J G and Shi C H 2004 Field Crops Research 8 13
[2] Cocchi M, Corbellini M, Foca G, Lucisano M, Pagani M A, Tassi L and Ulrici A 2005 Analytica Chimica Acta 544 100
[3] Lee K M, Herrman T J, Lingenfelser J and Jackson D S J 2005 Cereal Science 41 85
[4] Adamopoulos K G, Goula A M 2004 J. Food Engineering 63 199
[5] Chen B, Fu X G and Lu D L 2002 J. Food Engineering 5 373
[6] Laporte M F and Paquin P 1999 Agri. Food Chem. 47 2600
[7] Saeys W, Mouazen A M and Ramon H 2005 Biosystems Engineering 91 393
[8] Blanco M, Coello J, Iturriaga H, Maspoch S and Pages J 2000 Chemometrics and Intelligent Laboratory System 50 75
[9] Dou Y, Sun Y, Ren Y and Ren Y 2005 Analytica Chimica Acta 528 55
[10] Dou Y, Sun Y, Ren Y, Ju P and Ren Y 2005 J. Pharmaceutical and Biomedical Analysis. 37 543
[11] Zhang N R and Yan P F 1998 Neural Networks and Fuzzy Control (Beijing: Publishing House of Tsinghua University) 25