XVIIth World Congress of the International Commission of Agricultural and biosystems Engineering (CIGR)

Hosted by the Canadian Society for Bioengineering (CSBE/SCGAB)
Québec City, Canada June 13-17, 2010

GRADE CLASSIFICATION AND PROTEIN CONTENT DETERMINATION IN MATCHA BASED ON VIS/NIR

XIAOLEI ZHANG [1], RENTAO HE[2], YONG HE[3]

[1] X. Zhang, College of Biosystems Engineering and Food Science, Zhejiang University, 268 Kaixuan Road, Hangzhou 310029, China. xiaoleizhang@zju.edu.cn
[2] R. He, lanina@zju.edu.cn
[3] Y. He, yhe@zju.edu.cn

CSBE101481 – Presented at Section VI: Postharvest Technology and Process Engineering Conference

ABSTRACT In order to analyze matcha green tea powder grade classification and predict the protein content of matcha with near infrared spectra quickly, 240 samples of 3 matcha brands were collected for Vis/NIR spectroscopy on 325-1075 nm using a field spectroradiometer. The spectral data was processed by chemometrics which was integrated with partial least squares (PLS), principal component analysis (PCA) and back propagation neural network (BPNN) models. 180 samples (60 with each brand) were selected to build training model, the remaining 60 samples (20 with each brand) were applied as the prediction set. Firstly, PLS models were developed with comparison of different spectral pre-processing by smoothing, standard normal variant (SNV), multiplicative scatter correction (MSC), first and second derivative. The best PLS models were obtained by SNV both for the grade classification and protein measurement. Secondly, the selected principal components (PCs) from pre-processed spectra by SNV or original spectra were used as the inputs of back propagation neural networks (BPNN) models. The prediction results showed that PCA-BPNN models with original spectra were better than PLS models. The recognition ratio of 100% was achieved in validation set for matcha samples of three different brands. Moreover, an excellent precision was obtained in validation set for predicting the protein content, resulting in correlation coefficient (R), RMSEP and bias of 0.954, 1.114 and −0.123, respectively. The overall results indicated that Vis/NIR combined with PCA-BPNN was successfully applied for the grade classification and protein content determination in matcha.

Keywords: Vis/NIR spectroscopy, matcha, partial least squares analysis, back propagation neural network

INTRODUCTION Green tea is a “non-fermented” tea that has undergone minimal oxidation during processing, and contains more catechins than black tea or oolong tea. Several studies have reported that green tea extract has antioxidant, antibacterial, antiviral, anticarcinogenic, and antimutagenic functions (Higdon adn Frei, 2003; Lin et al., 2008). Matcha, a type of powdered green tea grown in the shade like gyokuro and which is traditionally used in the Japanese tea ceremony, is well known to be richer in some nutritional elements and epigallocatechin 3-O-gallate than other green teas(Noriko et al.,2009). Weil et al. (2002), have indicated that there must be a greater amount of nutrients, like protein and amino acid, consumed as a result of drinking matcha than other green teas since the powdered leaves themselves are ingested. Mactha is becoming a more and more popular ingredient in sweets and drinks to flavour and dye foods in recent years. There are many grades of matcha from food grade used in various recipes to a large number of ceremonial grades. Types of matcha are commonly graded depending on the quality and the parts of the plant used as well as how they are processed (Heiss, 2007). Additionally, the protein content is higher in fine matcha than other forms of green tea like sencha because of the shade-process before tea harvest (Haraguchi et al., 2003). So the price of top grade matcha can be extremely expensive. To pursuit the sudden huge profits, some underground factories mix brands of green tea powder of different grades, such as by packing a common green tea powder just belonging to konacha and containing low protein in the packaging of a high-grade matcha brand. These illegal behaviors harm the fair competition and badly infringe on the rights and interests of the consumers. While the prevailing method for discrimination of matcha quality is through organoleptic way at home and abroad, the results are subjective and not reliable all the time. Moreover, the traditional analyses of protein content in food products, such as Kjeldahl (Barbano et al., 1991), Lowry methods and reversed-Phase HPLC (Garcia, 2000), are time-consuming, making it disagree with the demand of modern production and trade market. Therefore, it is quite necessary to develop an accurate, rapid and less expensive method for determination of the grade and protein content in matcha.

Near-infrared spectroscopy (NIRS) has been widely applied as a fast, low cost and nondestructive analytical method for discrimination the qualitative and quantitative analysis of various types of agro-food products including tea. Many researchers have applied Vis/NIR spectroscopy in tea industry, such as caffeine estimation in instant green tea powder (Sinija and Mishra, 2009), the determination of total antioxidant capacity in green tea (Zhang et al., 2004), and the qualitative identification of tea categories (Niu and Lin, 2009; He et al., 2007; Zhao et al., 2006; Chen et al., 2007). However, there are few reports on the fast grade classification and predicting the protein content in matcha using NIRS.

The object of the present work was to investigate the feasibility of using VIS/NIR spectroscopy for the rapid assessment of characteristic information such as grade and protein content of matcha, and to obtain the best prediction model after the comparison of partial least squares (PLS) and back prorogation neural networks (BPNN).

MATERIALS AND METHODS

Sample Preparation A total of 240 samples of 3 commercial matcha green tea powder brands were obtained from supermarkets and tea franchised stores in China, including Yijiachun, Yifutang and Uji import matcha. There were different grades among these different brands, which embodied the price from low to high (6 RMB/ 100g, 12 RMB/ 100g and 22 RMB/ 100g respectively). All samples were stored at a low temperature (4±1°C) without any chemical or biological preservative treatment. Before analysis, assay portions were dried in an oven at 60°C for 6 h in order to express the analysis results on a dry basis. 180 samples (60 with each brand) were randomly selected for the calibration set, the rest 60 samples (20 with each brand) were applied as the prediction set.

Spectral Collection And Reference Method For Protein Content The NIR spectra are acquired in the reflectance mode using a handheld FieldSpec Pro FR (325–1075 nm)/A110070 spectroradiometer with Trademarks of Analytical Spectral Devices, Inc. (Analytical Spectral Devices, Boulder, USA). The field-of-view (FOV) of the spectroradiometer is 25°. The light source consists of a Lowell pro-lam interior light source assemble/128930 with Lowell pro-lam 14.5V Bulb/128690 tungsten halogen bulb that could be used both in visible and near infrared region. Each sample of matcha green tea powder was poured into uniform glass containers (65mmdiameter, 14 mm in height). The light source was placed at a height of approximately 250 mm and 45°C angle away from the sample. To exploit the 10-degree field of view of the probe, the spectroradiometer was placed at a distance of approximately 100 mm and 45°C angle away from the measurement area. The reflectance spectrum of each sample was measured at 1.5 nm intervals with an average reading of 30 scans. Two spectra were collected for each sample and the average spectrum of these two measurements was transformed into ASCII format by using the ASD ViewSpecPro software. The pretreatment and calculation were progressed using "The Unscrambler V 9.8" (CAMO AS, Oslo, Norway), and DPS (Data Procession Systems for Practical Statistics). To avoid a low signal-to-noise ratio, only the wavelength region of 500 to 900 nm was used for the calculations. The temperature and humidity were kept a steady level in the laboratory.

The reference method for protein content detection was Dumas combustion method using Rapid N Cube. After complete combustion, reduction, purification and detection, the nitrogen content of fish feed was obtained through the Rapid N Software V 3.4.0. The protein content of the fish feed was calculated as the value of total N×6.25. This measurement was performed immediately after Vis/NIRS measurements.

Partial Least Square (PLS) In the applying of near infrared spectroscopy, partial least squares (PLS) is a widely utilized multi-analysis and regression method among many technologies (Ghafourian and Cronin, 2005; He et al., 2005). It takes the advantage of the correlation relationship that already exists between the spectral data and the constituent concentrations of matcha. That is because PLS considers simultaneously the variable matrix Y (variety matrix) and the variable matrix X (spectral data) (Cen et al., 2006). The number of significant PLS factors to build the model for each compositional trait was determined by cross-validation. According to the accumulative reliabilities, certain latent variables (LVs) which were found through PLS and could explain the variance of the original spectra data would be selected as the inputs of BPNN models.

Back Propagation Neural Network (BPNN) BPNN is one of the most popular neural network topologies (He et al., 2005; He et al., 2006). It is a supervised learning technique that uses a gradient descent rule which attempts to minimize the error of the network by moving down the gradient of the error curve. One can construct a non-linear mapping function from multiple input data to multiple output data within the network in a distributed manner through a training process. In the system identification based on ANNs, a set of samples, usually structural response or seismic input data, are used to train the network to continuously adjust the link weights between neurons (Li and Yang, 2007). The weights are adjusted after every trial using external information specifying the correct result until the weights converge and the errors are reduced to acceptable values. For the PCA-BPNN models, PCA was performed firstly to extract information from the whole spectral regions, and the few principal components were used to be the neurons of network input layer. Several network architectures were tested by varying the number of neurons in the hidden layer with different initial weights (at least 5 times). The optimal parameters of the target error, the training rate, and the momentum were determined by the least prediction error.

RESULTS AND DISCUSSION

Spectral Features and Statistics of Protein Content Fig.1 shows the reflectance spectra of three brands of matcha green tea powder. The trends of the spectral curves were quite similar since all the samples showed approximately green, but there were some crossovers and overlapping among these samples. In order to make good use of the spectra, further pre-process and treatment with the chemometric method should be applied to classify the different grades and predict the protein content in matcha. Table 1 shows the statistics of protein content of matcha green tea powder samples. A relatively wide range of protein content was covered due to different brands of matcha.

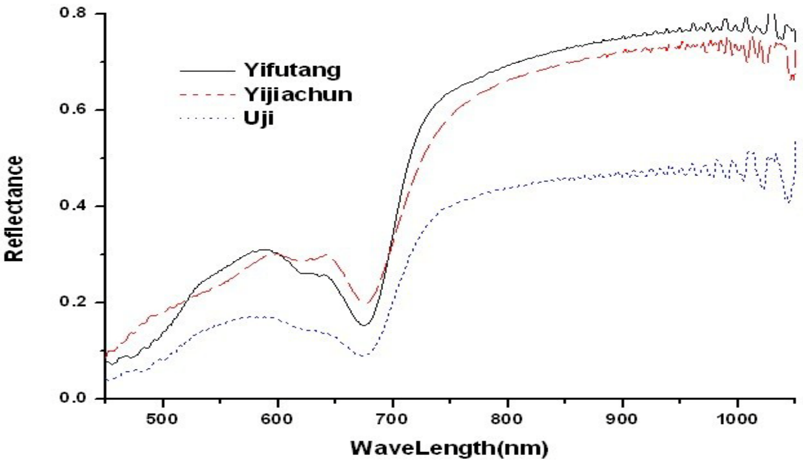

Figure 1. Vis/NIR reflectance of three brands matcha.

Table 1. Descriptive statistics of protein content of matcha in calibration and validation sets.

Data set	No.	Range	Mean	Standard deviation
Calibration	180	15.19- 25.59	22.03	2.392
Validation	60	15.19- 26.27	21.94	2.817
All samples	240	15.19- 26.27	22.01	2.500

Grade Classification

PLS Model and Analysis of PCs In order to enhance the characteristics of spectra and reduce the dimensionality of the spectra data matrix, partial least squares was used as the first treatment. Kinds of developed calibration models with deferent preprocessing methods were compared to obtain the best PLS model. The threshold error of recognition was set as ±0.4. Table 2 shows that the calibration model with SNV preprocessing method is the best one, and the recognition ratio for validation is 88.88%. So SNV was selected to be the pretreatment of PLS model. Simultaneously, another PLS model with raw spectra was developed for further comparison in BPNN model analysis.

Table 2. The effect of different pretreatment methods by PLS to discriminate brands.

preprocessing	Smoothing	SNV	MSC	First derivative	Second derivative
recognition ratio	82.44%	88.88%	87.65%	66.67%	68.89%

Several principal components (PCs) were extracted from the spectra of 240 samples. The superiority of using the PC score image is to pick up the characteristic information of the samples to interpret the clustering analysis. The cumulative reliabilities of the first 8 principal components could explain 98.402% of the total variance. The ninth principal component contributed only an additional 0.094%. Hence, the first 8 principal components could represent most of the feature information of the original spectra and were regarded as the inputs of the BP neural networks.. Figure 2 shows the score plot of the 3 varieties by using the first 3 principal components. From the scatter plot, it could be discovered that matcha of different brands distributed separately in the 3-dimension area.

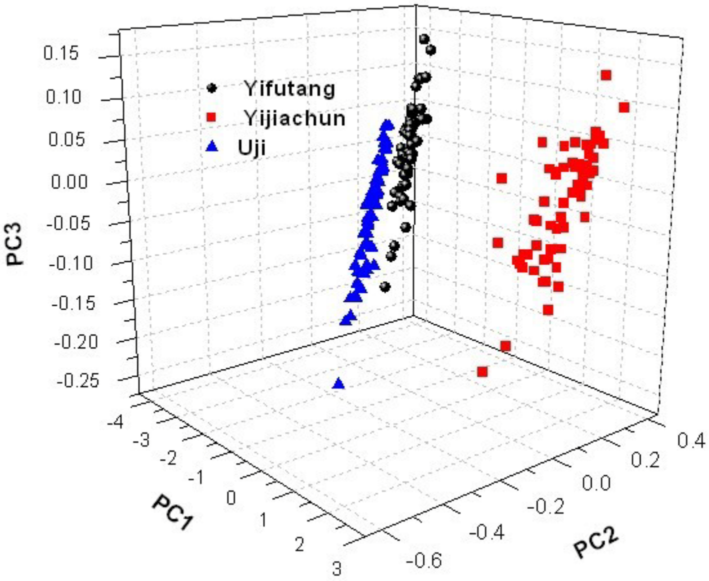

Figure 2. Scatter plots by PC1× PC2 × PC3 of 3 brands of matcha.

<u>BPNN Model</u> Firstly, a calibration model with SNV preprocessing method was developed, using the whole 240 spectroscopic data samples. Whereas, there was another calibration with no pretreatment. The matrix of training sets was composed of 180 samples and 8 variables. The performance of BPNN model was evaluated by the 60 unknown samples in the validation set. The discrimination results of prediction sets including calibration set with SNV pretreatment and without preprocessing were shown in Table 3. By comparison, the calibration with no pretreatment was regarded as the last matrix. The optimal parameters of this matrix in the modelling process were set as follows. After the adjustments of parameters, a three-layer BPNN model was achieved at last and the transfer function was sigmoid function. The nodes structure of input layer, hidden layer and output layer were 8-6-1. The dynamic parameter, the goal error and the times of training were set as 0.7, 0.00001 and 170, respectively. And the method for

treating data was set standardized transformation. All these parameters were selected after many trials of using different values. The one node of the output layer showed the results in the form of dummy variables 1, 2, and 3 (i.e., 1 for Yifutang, 2 for Yijiachun and 3 for Uji). The threshold error of recognition was also set as ±0.1. This criterion was more precise than that of setting in PLS method. The residual error of the prediction was $9.991x10^{-6}$. The prediction results demonstrated that the recognition ratio of PCA-BPNN model without pretreatment was 100% for validation sets.

Table 3. The discrimination results of different matcha brands in validation set.

Variety	Sample number	Reference value	Recognition ratio	
			SNV pretreatment	No pretreatment
Yifutang	01-20	1	95%	100%
Yijiachun	21-40	2	95%	100%
Uji	41-60	3	100%	100%
Totel	01-60	/	96.7%	100%

Protein Content Measurement

PLS Model Much the same as the process of grade classification, five deferent preprocessing methods for developing calibration models were used to obtain the best PLS model. The predictive capability of model was evaluated by the following standards: correlation coefficient (R), root mean square error of prediction (RMSEP), and bias. Generally, a good model should have high correlation coefficient values, low RMSEC, RMSEP and bias values. Table 4 shows that the calibration model with SNV preprocessing is the best one. The correlation coefficients, RMSEP and bias were 0.919, 1.167 and -0.078, respectively. So the PLS model with SNV pretreatment was also the selected method to predict the protein content of matcha.

Table 4. The effect of different pretreatment methods by PLS to predict protein content.

Preprocessing	R	RMSEP	Bias
Smoothing	0.899	1.249	- 0.090
SNV	0.919	1.167	- 0.078
MSC	0.903	1.169	- 0.079
First derivative	0.879	1.480	0.072
Second derivative	0.736	2.069	0.071

The selection of principal components here was the same as before. The first eight PCs were regarded as the inputs of the BPNN model as a comparison of PLS models.

BPNN Model Two matrixes of calibration set were developed using BPNN, one of which was composed of 180 samples and 8 variables with preprocessed spectra by SNV, while another with raw spectra. The prediction results are shown in Table 5. One can see that the prediction accuracy of BPNN model without pretreatment was higher than that of the one with SNV preprocessing, and it could be the ideal . And the optimal parameters fitting this matrix of calibration without SNV pretreatment were set as follows. A three-layer BP model with the structure of 8-6-1 was achieved at last with the sigmoid transfer function. The dynamic parameter, the goal error and the times of training were set as 0.7, 0.00001 and 1000, respectively. It was discovered that the structure was the same for grade classification and protein content measurement. The reason might be that the quality of different band was related to the protein in, and the latent information and relationship of spectral data was almost the same to these properties.

Table 5. The prediction results of protein content in matcha samples.

Evaluating standard	SNV-PLS	BPNN	
		SNV pretreatment	No pretreatment
R	0.939	0.867	0.954
RMSEP	1.167	1.744	1.114
Bias	-0.078	-0.605	-0.123

The performance of BPNN models was validated by the 60 unknown samples in the validation set using the aforementioned parameters. The residual error was 4.1x10-3. The correlation coefficient (R), RMSEP and bias for validation set by BPNN model with no pretreatment, which was the best one by comparison, were 0.954, 1.114 and -0.123, respectively. Fig. 3 shows the reference versus predicted values plots in validation set by this BPNN model. The solid line was the regression line which shows the predicted values are in close proximity to the reference values. The results indicated an optimal prediction performance for protein content in different grades matcha green tea powder, which was a bit better than only using PLS.

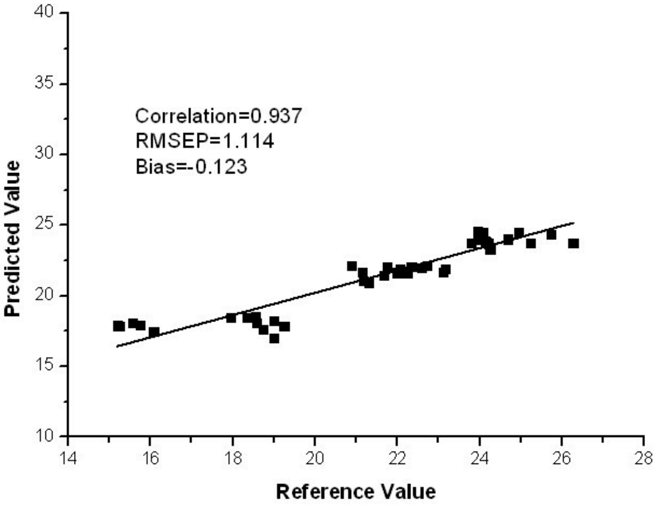

Figure 3. Measured vs. predicted values for protein content by BPNN model.

CONCLUSION VIS/NIR spectroscopy combined with chemometrics was successfully implemented for conducting grade identification and protein content determination of matcha green tea powder. In the PLS models, the preprocessed spectra by SNV achieved the best performance both for the grade classification and protein measurement. The correlation coefficients, RMSEP and bias were 0.919, 1.167 and -0.078, respectively in prediction of protein content, whereas the recognition rate of 88.88% was achieved in discrimination of different brands of matcha, which was an unacceptable prediction accuracy in practical applications. BPNN models were developed for comparison. The cut-off error of prediction was set as ±0.1, and the recognition ratio of BPNN model without pretreatment was 100% in the prediction for brands identification, which was improved greatly compared to that of PLS model in validation sets. Furthermore, the correlation coefficient (R), RMSEP and bias in validation set by BPNN model without pretreatment were 0.954, 1.114 and -0.123, respectively. Hence, the performance of BPNN was a little better than that of PLS models for the prediction of protein content in matcha. The results indicated that the prediction accuracy of BPNN without pretreatment was better than those of PLS models. The reason might be that raw spectra were more suitable to discriminating the brands and predicting the protein in matcha. Some latent useful information of spectra data might be removed in the preprocessed spectra by SNV. From another point of view, BPNN model could handle certain latent nonlinear information of spectral data, and the nonlinear information was contributed to the better performance of BPNN model. This was just a hypothesis that the nonlinear information played a key role in a better performance, more work would be done to discover the useful information or effective wavelength or wavelength bands for the non-destructive determination of grades and protein content in matcha green tea.

REFERENCES

Barbano, D.M., J.M. Lynch, and J.R. Fleming. 1991. Direct and indirect determination of true protein content of milk by Kjeldahl analysis. Anal.Chem. 74: 281.

Cen, H.Y., Y. He, and M. Huang. 2006. Measurement of soluble solids contents and pH in orange juice using chemometrics and Vis-NIRS. Journal of Agricultural and Food Chemistry 54: 7437–7443.

Chen, Q. S., J. W. Zhao, C. H. Fang, and D. M. Wang. 2007. Feasibility study on identification of green, black and Oolong teas using near-infrared reflectance spectroscopy based on support vector machine (SVM). Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. 66(3): 568-574.

Garcia, M.C., M. Marina, and M. Torre. 2000. Determination of the soybean protein content in soybean liquid milks by Reversed-Phase HPLC, J. Liq. Chromatogr. Relat. Technol. 23: 3165.

Ghafourian, T., and M. T. D. Cronin. 2005. The impact of variable selection on the modeling of oestrogenicity. SAR and QSAR in Environmental Research. 16: 171-190.

Haraguchi, Y., Y. Imada, and S. Sawamura. 2003. Production and characterization of fine Matcha for processed food. Journal of the Japanese Society for Food Science and Technology-Nippon Shokuhin Kagaku Kogaku Kaishi. 50(10): 468-473.

He, Y., H. Y. Song, P. A. Garcia, and G. A. Hernandez. 2005. A new approach to predict N, P, K and OM content in a loamy mixed soil by using near infrared reflectance spectroscopy. Lecture Notes in Computer Science. 3644: 859-867.

He, Y., S. J. Feng, X. F. Deng, and X. L. Li. 2006. Study on lossless discrimination of varieties of yogurt using the Visible/NIR-spectroscopy. Food Research International. 39: 645-650.

He, Y., X. L. Li, and X. F. Deng. 2007. Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model. Journal of Food Engineering 79: 1238–1242.

He, Y., Y. Zhang, and L.G. Xiang. 2005. Study of application model on BP neural network optimized by fuzzy clustering. Lecture Notes in Artificial Intelligence. 3789: 712-720.

Heiss, M. L., and R. J. Heiss. 2007. The story of tea: a cultural history and drinking guide. Library Journal. 132 (16): 90-90.

Higdon, J. V., and B. Frei. 2003. Tea catechins and polyphenols: Health effects, metabolism, and antioxidant functions. Critical Reviews in Food Science and Technology. 43: 89–143.

Li, H. N., and H. Yang. 2007. System identification of dynamic structure by the multi-branch BPNN. Neurocomputing. 70: 835-841.

Lin, S. D., E. H. Liu, and J. L. Mau. 2008. Effect of different brewing methods on antioxidant properties of steaming green tea. LWT – Food Science and Technology. 41: 1616–1623.

Niu, Z. Y. and X. Lin. 2009. Qualitative and Quantitative Analysis Method of Tea by Near Infrared Spectroscopy. Spectroscopy and Spectral Analysis. 29(9): 2417-2420.

Noriko, Y., S. K. Ki, M. H. Jong, and Y. Takako. 2009. Matcha, a Powdered Green Tea, Ameliorates the Progression of Renal. Journal of Medicinal Food. 12(4): 714–721.

Sinija, V.R. and H.N. Mishra. 2009. FT-NIR spectroscopy for caffeine estimation in instant green tea powder and granules. LWT - Food Science and Technology. 42: 998-1002.

Weil, A., and R. Daley. 2002. The Healthy Kitchen, Alfred A. Knopf, New York. 44.

Zhang, M.H., J. Luypaert, J.A. Fernández Pierna, Q.S. Xu, and D.L. Massart. 2004. Determination of total antioxidant capacity in green tea by near-infrared spectroscopy and multivariate calibration. Talanta. 62 (1): 25-35.

Zhao, J. W., Q. S. Chen, X. Y. Huang, and C. H. Fang. 2006. Qualitative identification of tea categories by near infrared spectroscopy and support vector machine. Journal of Pharmaceutical and Biomedical Analysis. 41: 1198–1204.