Nondestructive Identification of Tea (*Camellia sinensis* L.) Varieties using FT-NIR Spectroscopy and Pattern Recognition

Quansheng CHEN[1], Jiewen ZHAO[1], Muhua LIU[2] and Jianrong CAI[1]

[1]School of Food & Biological Engineering, Jiangsu University, Zhenjiang, P. R. China;
[2]Engineering College, Jiangxi Agricultural University, Nanchang, P. R. China

Abstract

Chen Q., Zhao J., Liu M., Cai J. (2008): **Nondestructive identification of tea** (*Camellia sinensis* L.) **varieties using FT-NIR spectroscopy and pattern recognition**. Czech J. Food Sci., **26**: 360–367.

Due to more and more tea varieties in the current tea market, rapid and accurate identification of tea (*Camellia sinensis* L.) varieties is crucial to the tea quality control. Fourier Transform Near-Infrared (FT-NIR) spectroscopy coupled with the pattern recognition was used to identify individual tea varieties as a rapid and non-invasive analytical tool in this work. Seven varieties of Chinese tea were studied in the experiment. Linear Discriminant Analysis (LDA) and Artificial Neural Network (ANN) were compared to construct the identification models based on Principal Component Analysis (PCA). The number of principal components factors (PCs) was optimised in the constructing model. The experimental results showed that the performance of ANN model was better than LDA models. The optimal ANN model was achieved when four PCs were used, identification rates being all 100% in the training and prediction sets. The overall results demonstrated that FT-NIR spectroscopy technology with ANN pattern recognition method can be successfully applied as a rapid method to identify tea varieties.

Keywords: green tea; variety; identification; FT-NIR spectroscopy; pattern recognition

Nowadays, most commercial tea-leaves (*Camellia sinensis* L.) have many varieties in the market, and tea varieties differ not only from botanical standpoints but also in terms of quality attributes. The differences between the tea varieties are recognised commercially and appreciated by the consumers, therefore, the identification of tea varieties is still focused on at present. Usually, the conventional identification of a tea variety is performed by sensory evaluation. Tea sensory evaluation is still dependent on inspection of its sensory appearance, smell, flavour, and taste using oral examination to determine its quality and variety. Sensory evaluation of tea is one of the most difficult tasks in the overall tea attributes assessment. It relies on information provided by selected and trained tasting panels, whose members may be influenced by physiological, psychological, and environmental factors (Yan 2005). Therefore, the identification of tea varieties by sensory evaluation produces

Supported by the National Natur and Science Foundation of China for Youth Program (Grant No. 30800666), the Natural Science Foundation for Colleges and Universities in Jiangsu Province (Grant No. 08KJB550003), and the Advanced Talents Science Foundation of Jiangsu University (Grant No. 08JDG007).

inevitably subjective and inconsistent results. Nowadays, the identification of a tea variety is also performed according to various wet chemical methods such as high-performance liquid chromatography (HPLC) (Valera *et al*. 1996; Zuo *et al*. 2002), gas chromatography (GC) (Togari *et al*. 1995), capillary electrophoresis (Hideki *et al*. 1997), and plasma atomic emission spectrometry (Herrador & Gonzalez 2001). Compared with the sensory evaluation method, all these wet chemical methods mentioned above are precise, but they are all time-consuming in the identification of tea varieties.

FT-NIR spectroscopy has proved to be a powerful analytical tool for analysing a wide variety of samples used in the agricultural, nutritional, petrochemical, textile, and pharmaceutical industries, especially its use in the qualitative analysis of agricultural products and pharmaceutical samples has significantly increased during the last decade (McGlone *et al*. 2002; Esteban-Diez *et al*. 2004; Clark *et al*. 2005; Huck *et al*. 2005; Woo *et al*. 2005). The FT-NIR spectroscopy technique is a non-destructive analytical technique with the advantages of rapid sample analysis, simple operation, and small samples, and particularly the use of solid samples. Compared to conventional analytical methods, FT-NIR spectroscopy is a fast, accurate, and non-destructive technique that can be used as a replacement of conventional sensory evaluation methods and time-consuming chemical methods.

The different tea varieties have chemical characters which are due to different tea processes and different original tea-leaves. Therefore, the spectral features of each tea variety are reasonably differentiated in the NIR region, the spectral differences providing sufficient qualitative spectral information for the identification. Since 1990s, attempts have been made to predict simultaneously alkaloids and phenolic substances in green tea leaves using near infrared spectroscopy (Hall *et al*. 1988; Schulz *et al*. 1999). Some studies on the quantitative analysis of total antioxidant capacity in green tea by NIR have been also reported by Luypaert *et al*. (2003) and Zhang *et al*. (2004). Recently, some researchers applied near infrared spectroscopy to analyse simultaneously the contents of free amino acids, caffeine, total polyphenols, and amylose in green tea (Sun *et al*. 2004; Luo *et al*. 2005; Chen *et al*. 2006a, b).

In the works mentioned above, near infrared spectroscopy techniques have often been used to analyse quantitatively the valid tea composition using some chemometrics methods, for example, Partial Least Squares (PLS) regression and Artificial Neural Net (ANN). In this work, FT-NIR spectroscopy was used to identify seven varieties of tea coupled with the pattern recognition method. Linear Discriminant Analysis (LDA) and Artificial Neural Network (ANN) will be used comparatively to construct the identification model based on Principal Component Analysis (PCA), and the number of principal components will be optimised in the constructing models.

MATERIAL AND METHODS

Sample preparation. All tea samples of seven varieties came from five different provinces in China. All tea materials were purchased at the local super-market from April to June in 2006, and they were all stored in air-tight containers for 4 months. About 10 ± 0.1 g air-dried tea-leaves were weighed randomly as one sample. The tea varieties, origins, numbers, and categories are shown in Table 1.

Table 1. Varieties, categories, and origins of tea samples

Tea varieties	Sample No.	Tea origin	Tea category
Maofeng	30	Anhui	green tea
Biluochun	27	Jiangsu	green tea
Tieguanyin	33	Fujiang	oolong tea
Maojian	30	Henan	green tea
Cuilan	21	Anhui	green tea
Longjing	21	Zhenjiang	green tea
Queshe	27	Jiangsu	green tea

Spectra collection. The NIR spectra were collected in the reflectance mode using the Antaris^{TM} near-infrared spectrophotometer (Thermo Electron Co., USA) with an integrating sphere. Each spectrum was the average spectrum of 32 scans. The spectra ranged from 10 000 cm^{-1} to 4000 cm^{-1}, and the data were measured in 3.856 cm^{-1} intervals, which resulted in 1557 variables.

The standard sample accessory holder was used for performing the tea spectra collection. The sample accessory holder is a sample cup specifically designed by Thermo Electron Co. For each tea sample, 10 ± 0.1 g of dry tea-leaves were filled into the sample cup by the standard procedure

depending upon the bulk density of the material. The corresponding amount of dry tea-leaves was densely packed into the sample cup and then compressed by closing it. Each tea sample was collected three times. The average of the three spectra which were collected from the same tea sample was used in the next analysis. The temperature was kept around 25°C and the humidity was kept at a steady level in the laboratory.

Software. All algorithms were implemented in Matlab V7.0 (Mathworks, USA) under Windows XP in data processing. Result Software (Thermo Electron Co., USA) was used in NIR spectral data acquisition.

RESULTS AND DISCUSSION

Design of the experiments

Rough spectral data are needed to conduct spectral preprocessing because of the light scatter in tea-leaves. In this work, three spectral preprocessing methods (Standard Normal Variate Transformation (SNV), Mean Centering (MC), and Multiplicative Scatter Correction (MSC)) were applied comparatively. SNV is a mathematical transformation method of the log (1/R) spectra used to remove the slope variation and to correct scatter effects. Each spectrum is corrected individually by first centering the spectral values, and then the centered spectrum is scaled by the standard deviation calculated from the individual spectral values. MC is to calculate the average spectrum of the data set and subtract the average from each spectrum. MSC is another important procedure for the correction of the scatter light performed, on the basis of different particle sizes. This technique is also used to correct the additive and multiplicative effects in the spectra.

The comparison of the results obtained by the three preprocessing methods revealed that SNV preprocessing method is as good as MSC, and much better than MC. This is because dry tea-leaves are solid particles which scatter light easily; while SNV and MSC spectral preprocessing methods can remove the slope variation and correct the light scatter due to different particle sizes. Therefore, SNV spectral preprocessing method was applied in this work.

In raw NIR spectra of tea-leaves, water absorption bands occur around 5155 cm^{-1} and 7000 cm^{-1} corresponding to O-H stretching + O-H deformation. These were excluded during the analyses along with some regions exhibiting a high noise level (e.g. 10 000–9000 $cm^{-1}$and 5000–4000 cm^{-1}).

The most intensive band in the spectrum belonged to the vibration of the 2nd overtone of the carbonyl group (5352 cm^{-1}), followed by the C-H stretch and C-H deformation vibration (about 7212 cm^{-1}), the -CH_2 (about 5742 cm^{-1}), and the -CH_3 overtone (about 5808 cm^{-1}). The vibration of the carbonyl group and the -C-H and -CH_2 vibrations are caused by some ingredients such as polyphenols, alkaloids, protein, and volatile or non-volatile acid. In general, the water content in the dry tea leaves amounts to 4–6% (w/w), therefore, the effect of water must be considered. To keep away from the water absorption band, the spectral regions between 5300 cm^{-1} and 6500 cm^{-1} were selected, because there is a great deal of information from organic substances in this NIR spectroscopy region according to the spectral investigation.

All NIR spectra of seven tea varieties were used for the PCA. The behaviour of PCA can indicate

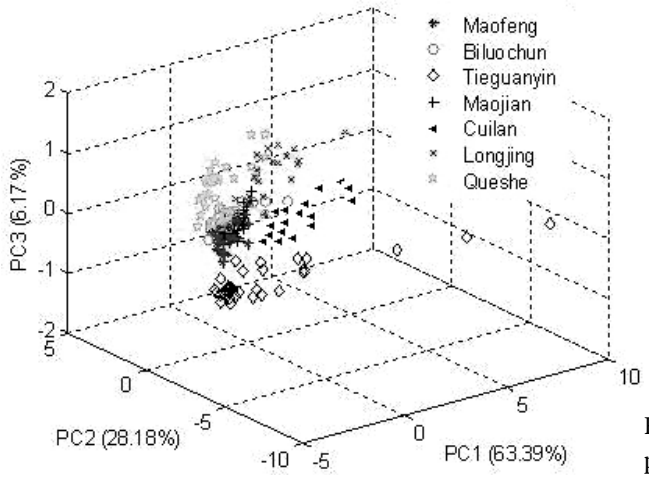

Figure 1. Score cluster plot of the top three principal components (PCs) for all samples from seven tea varieties

the data trend in visualising the dimension spaces. For visualising the data trends coming from NIR spectra of the seven tea varieties, a scatter plot of samples using the top three principal components (PCs) issued from PCA of the data matrix (also called scores plot) was obtained as shown in Figure 1.

Figure 1 shows that there is a neat separation of the seven tea varieties in the 3-dimensional space represented by PC1, PC2, and PC3 scores vectors. Such a good classification in this 3-dimensional space can be explained by the chemical background of tea and PCA methods. The different tea varieties can exhibit considerable differences in their botanical, genetic, and agronomic characteristics, combined also with different tea processes and different origins. The differences detected in the chemical composition of different tea varieties can be reasonably differentiated in the NIR spectroscopy region. Therefore, in the NIR spectroscopy region, the spectral differences provided enough spectral information for further qualitative analysis. In addition, through PCA, the total variance contribution rate was over 97% for the top three PCs, so PC1, PC2, and PC3 scores vectors can almost explain fully the chemical composition information in the NIR spectroscopy region. Thus, the 3-dimensional space represented by PC_1, PC_2, and PC_3 scores vectors from all samples can fully express the information that all samples are distributed in ultra-dimensional space.

Geometrical exploration based on PCA score plots only gives the cluster trends in visualising the dimension spaces, therefore, in this study, the ultimate aim was to search an appropriate supervised pattern recognition to identify tea varieties after PCA. The supervised pattern recognition refers to such techniques in which the a priori knowledge about the category membership of samples is used for the classification. The classification model is developed on a training set of samples with categories. The model performance is evaluated by the use of a prediction set by comparing the predictive classification with the true categories (Roggo *et al.* 2003). Therefore, before building the identification model, all 189 samples were separated into two groups. One is called the training set and includes 126 samples (i.e. 20 Maofeng samples, 18 Biluochun samples, 22 Tieguanyin samples, 20 Maojian samples, 14 Cuilan samples, 14 Longjing samples, and 18 Queshe samples). The other is called the prediction set and includes the remaining 63 samples, and is used to evaluate the performance of the identification model.

The supervised pattern recognition methods are numerous, and the main problem is to choose the most accurate method. In this study, the linear (i.e. LDA) and the non-linear (i.e. ANN) supervised pattern recognition methods were compared. The identification rates in the training and prediction sets were used to evaluate the performance of LDA and ANN models as the important criteria.

The top principal components were extracted as the input of the pattern recognition by PCA. It goes without saying that the number of principal component factors (PCs) is crucial to the performance of the identification model. Therefore, PCs should be optimised in building the identification models.

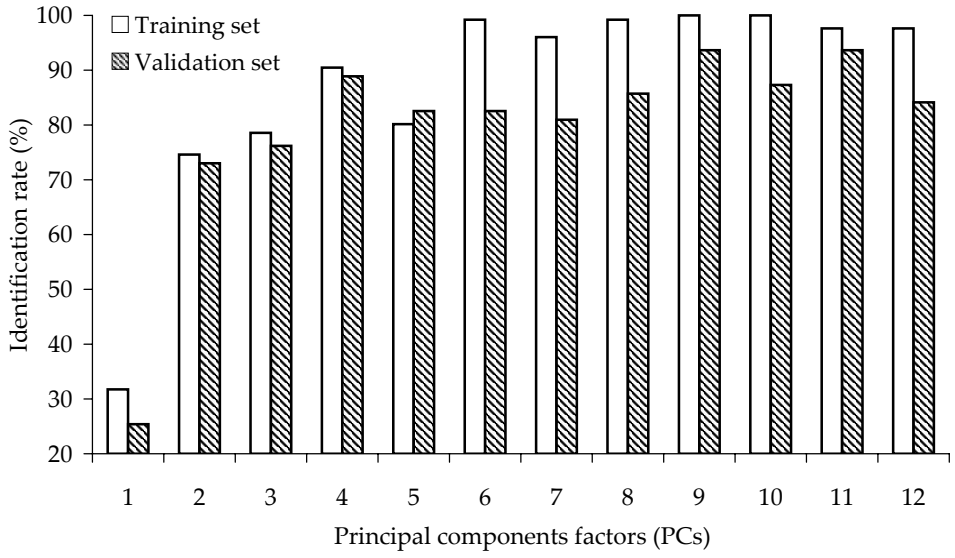

Figure 2. Identification rates of LDA model with different PCs in the training and prediction sets

Identification results of LDA model

Linear Discriminant Analysis (LDA) is a linear and parametric method with a discriminating character. LDA is focused on finding optimal boundaries between classes. Here, a brief introduction of LDA is presented in this paper, and readers can refer to other literature (Yang *et al*. 2005; Chen *et al*. 2008). The number of principal component factors (PCs) is crucial for the performance of the LDA identification model, and identification rates in training and prediction sets were used as criteria to optimise the number of PCs. Figure 2 shows the identification results of LDA model under 1~12 PCs for the training and prediction sets. As can be seen in Figure 2, the optimal LDA model is achieved when 9 PCs are included.

The identification results in the training and prediction sets are shown in Table 2. In the training set, all samples were identified correctly, and the total identification rate was 100%. In the prediction set, two Biluochun samples were identified wrongly as Maofeng group and Maojian group, respectively, one Maofeng sample was identified wrongly as Biluochun group, one Maojian sample was identified wrongly as Biluochun group, and all the remaining samples were identified correctly. The total identification rate was 93.65% in the prediction set.

Identification results of ANN model

Artificial neural network (ANN) is a non-linear pattern recognition method. Many parameters exert to some extent certain influence on the performance of ANN models. These parameters include the number of neurons in the middle layer, scale functions, learning rate factor, momentum factors, and initial weights (Blanco *et al*. 1999; Mouwen *et al*. 2006). In this work, the most classical Back Propagation Artificial Neural Network (BP-ANN) with 3 layers construction was used to construct the identification model. These parameters of the BP-ANN model were optimised by cross validation as follows: the number of neurons in the hidden layer was set to 8, the learning rate factor and momentum factor were set to 0.1 each, the initial weight was set to 0.3, and the scale function was set as 'tanh' function.

It is crucial to select the appropriate number of PCs in constructing a BP-ANN identification model. Figure 3 shows the identification rates of

Table 2. Confusion matrix for the identification results of LDA model in the training and prediction sets

Subsets	Tea varieties	Sample number	Identification results							Identification rate (%)
			M1	B	T	M2	C	L	Q	
Training set	M1	20	20	0	0	0	0	0	0	100
	B	18	0	18	0	0	0	0	0	
	T	22	0	0	22	0	0	0	0	
	M2	20	0	0	0	20	0	0	0	
	C	14	0	0	0	0	14	0	0	
	L	14	0	0	0	0	0	14	0	
	Q	18	0	0	0	0	0	0	18	
Prediction set	M1	10	9	1	0	0	0	0	0	93.65
	B	9	1	7	0	1	0	0	0	
	T	11	0	0	11	0	0	0	0	
	M2	10	0	1	0	9	0	0	0	
	C	7	0	0	0	0	7	0	0	
	L	7	0	0	0	0	0	7	0	
	Q	9	0	0	0	0	0	0	9	

M1 – Maofeng; B – Biluochun; T – Tieguanyin; M2 – Maojian; C – Cuilan; L – Longjing; Q – Queshe

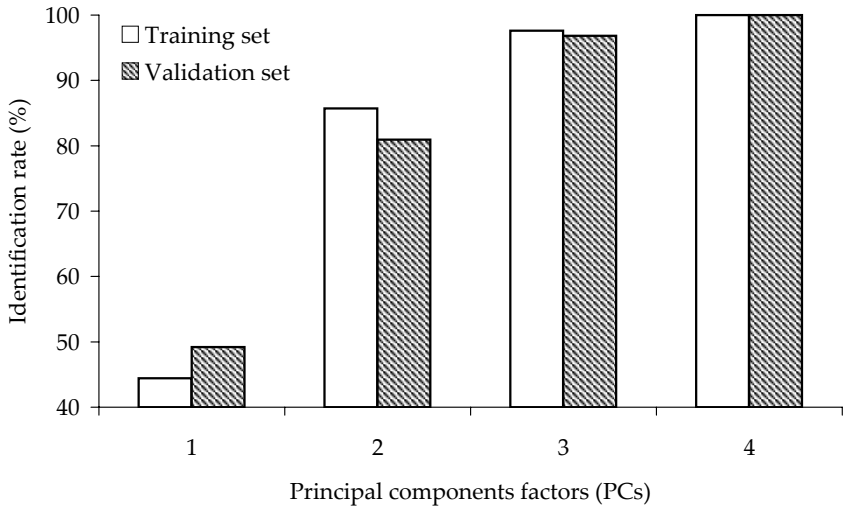

Figure 3. Identification rates of ANN model with different PCs in the training and prediction sets

BP-ANN model with different PCs in the training and prediction sets. As can be seen in Figure 3, the highest identification rates of BP-ANN model are 100% in the training and prediction sets, respectively, when 4 PCs are included. Therefore, the optimal ANN model was achieved with four PCs, and all samples were identified correctly in both the training and prediction sets.

Discussion on two identification results

Table 3 shows the identification results from LDA and ANN models. As can be seen in Table 3, the identification rates of LDA model are 100% in the training set and 93.65% in the prediction set when the optimal number of PCs is 9, and the identification rates of LDA model are 100% in the training set and 100% in the prediction set, respectively, when the optimal number of PCs is 4. Compared by these identification results, BP-ANN model is a little better than LDA model, in other words, the non-linear pattern recognition is a little better than the linear pattern recognition method. Compared by their optimal number of PCs, BP-ANN model (4 PCs) is less demanding than LDA model (9 PCs), in other words, BP-ANN model is simpler than LDA model. The high number of PCs in LDA model can explain the differences between the tea varieties, but too high a number of PCs might include 'specific information' in the calibrating model, and this 'specific information' might result in a worse generalisation performance of the identification model. This 'specific information' included in the training model might bring too much redundant information, which is an inevitable effect on the robustness of the model. Therefore, it will lead to 'bad' results when some new samples are predicted by this model. Generally, the non-linear pattern recognition method has better abilities than the linear pattern recognition method in self-organisation and self-learning (Zhao *et al.* 2006), therefore, the identification results obtained from ANN model are a little better than those from LDA model.

Table 3. Comparison of the identification results based on three models

Models	Optimal number of PCs	Identification results of models (%)	
		training set	validation set
LAD	9	100	93.65
ANN	4	100	100

CONCLUSION

The overall results sufficiently demonstrate that the FT-NIR spectroscopy technique coupled with the pattern recognition has a high potential to identify the tea varieties in a nondestructive way. Seven varieties of tea were studied in this work. Linear Discriminant Analysis (LDA) and Artificial Neural Network (ANN) were compared as the pattern recognition methods in the calibrating model. The number of principal components factors (PCs) was optimised in the building of the models. The performances of two supervised pattern recognition methods were compared. The experimental results showed that ANN model was better than LDA model, with both identification rates being

equal to 100% in the training and prediction sets when the optimal number of principal components factors (PCs) equaled 4.

It can be concluded that FT-NIR spectroscopy technique coupled with the pattern recognition has a high potential to estimate another foods quality in a nondestructive way. A reliable overall characterisation of a food product quality may be obtained at a low cost. It may be applied to the food quality control, process monitoring, and rapid classification in the industry. In comparison to subjective sensory assessing methods and time-consuming chemical methods, the results obtained by FT-NIR technique represent a considerable improvement in estimating the food quality.

References

Blanco M., Coello J., Iturriaga H., Maspoch S., Pagès J. (1999): Calibration in non-linear near infrared reflectance spectroscopy: a comparison of several methods. Analytica Chimica Acta, **384**: 207–214.

Chen Q.S., Zhao J.W., Huang X.Y., Zhang H.D., Liu M.H. (2006a): Simultaneous determination of total polyphenols and caffeine contents of green tea by near-infrared reflectance spectroscopy. Microchemical Journal, **83**: 42–47.

Chen Q.S., Zhao J.W., Zhang H.D., Wang X.Y. (2006b): Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration. Analytica Chimica Acta, **572**: 77–84.

Chen Q.S., Zhao J.W., Cai J.R. (2008): Identification of tea varieties using computer vision. Transcations of ASABE, **51**: 623–628.

Clark C.J., Mcglone V.A., Jordan R.B. (2005): Detection of Brownheart in 'Braeburn' apple by transmission NIR spectroscopy. Postharvest Biology and Technology, **28**: 65–71.

Esteban-Diez I., Gonzalez-Saiz J.M., Pizarro C. (2004): An evaluation of orthogonal signal correction methods for the characterization of Arabica and Robusta coffee varieties by NIRS. Analytica Chimica Acta, **514**: 57–67.

Hall M.N., Robertson A., Scotter C.N.G. (1988): Near-infrared reflectance prediction of quality, theaflavin content and moisture content of black tea. Food Chemistry, **27**: 61–75.

Herrador M.A., Gonzalez A.G. (2001): Pattern recognition procedures for differentiation of green, black and Oolong teas according to their metal content from inductively coupled plasma atomic emission spectrometry. Talanta, **53**: 1249–1257.

Hideki H., Toshihiro M., Katsunori K. (1997): Simultaneous determination of qualitatively important components in green tea infusions using capillary electrophoresis. Journal of Chromatography A, **758**: 332–335.

Huck C.W., Guggenbichler W., Bonn G.K. (2005): Analysis of caffeine, theobromine and theophylline in coffee by near infrared spectroscopy compared to high-performance liquid chromatography (HPLC) coupled to mass spectrometry. Journal of Pharmaceutical Biomedical Analysis, **538**: 195–203.

Luo Y.F., Guo Z.F., Zhu Z.Y., Wang C.P., Jiang H.Y., Han B.Y. (2005): Studies on ANN models of determination of tea polyphenol and amylose in tea by near-infrared spectroscopy. Spectroscopy Spectral Analysis, **25**: 1230–1233.

Luypaert J., Zhang M.H., Massart D.L. (2003): Feasibility study for the using near infrared spectroscopy in the qualitative and quantitative of green tea, *Camellia sinensis* (L). Analytica Chimica Acta, **487**: 303–312.

McGlone V.A., Jordan R.B., Seelye R., Martinsen P.J. (2002): Comparing density and NIR methods for measurement of kiwi fruit dry matter and soluble solids content. Postharvest Biology and Technology, **26**: 191–198.

Mouwen D.J.M., Capita R., Alonso-Calleja C., Prieto-Gómez J., Prieto M. (2006): Artificial neural network based identification of *Campylobacter* species by Fourier transform infrared spectroscopy. Journal of Microbiological Methods, **67**: 131–140.

Roggo Y., Duponchel L., Huvenne J.P. (2003): Comparison of supervised pattern recognition methods with McNemar's statistical test. Application to qualitative analysis of sugar beet by near-infrared spectroscopy. Analytica Chimica Acta, **477**: 187–200.

Schulz H., Engelhardt U.H., Wegent A., Drews H.H., Lapczynski S. (1999): Application of NIRS to the simultaneous prediction of alkaloids and phenolic substances in green tea leaves. Journal of Agricultural and Food Chemistry, **475**: 5064–5067.

Sun Y.G., Lin M., Lv J., Xu L.H. (2004): Determination of the contents of free amino acids, caffeine and tea polyphenols in green tea by Fourier transform near-infrared spectroscopy. Chinese Journal of Spectroscopy Laboratory, **21**: 940–943.

Togari N., Kobayashi A., Aishima T. (1995): Pattern recognition applied to gas chromatographic profiles of volatile component in three tea categories. Food Research International, **28**: 495–502.

Valera P., Pablo F., Gonzalez A.G. (1996): Classification of tea samples by their chemical composition using discriminant analysis. Talanta, **43**: 415–419.

Woo Y.A., Kim H.J., Ze K.R., Chung H. (2005): Near-infrared (NIR) spectroscopy for the non-destructive and fast determination of geographical origin of *Angelicae gigantis* Radix. Journal of Pharmaceutical Biomedical Analysis, **36**: 955–959.

Yan S.H. (2005): Evaluation of the composition and sensory properties of tea using near infrared spectroscopy and principal component analysis. Journal Near Infrared Spectroscopy, **6**: 313–325.

Yang H., Irudayaraj J., Paradkar M.M. (2005): Discriminant analysis of edible oils and fats by FTIR, FT-NIR and FT-Raman spectroscopy. Food Chemistry, **93**: 25–32.

Zhang M.H., Luypaert J., Xu Q.S., Massart D.L. (2004): Determination of total antioxidant capacity in green tea by NIRS and multivariate calibration. Talanta, **62**: 25–35.

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

Zuo Y.G., Chen H., Deng Y.W. (2002): Simultaneous determination of catechins, caffeine and gallic acids in green, Oolong, black and pu-erh teas using HPLC with a photodiode array detector. Talanta, **57**: 307–316.

Received for publication November 11, 2007
Accepted after corrections August 18, 2008

Corresponding author:

Dr. Quansheng Chen, Jiangsu University, School of Food & Biological Engineering, Xuefu Road 301#, Zhenjiang City, Jiangsu Province, 212013, P. R. China
tel.: + 86 511 887 903 18, fax: + 86 511 887 802 01, e-mail: q.s.chen@hotmail.com