文章编号：1001－9014(2016)02－0200－06

DOI:10.11972/j.issn.1001－9014.2016.02.014

Origin identification of Shandong green tea by moving window back propagation artificial neural network based on near infrared spectroscopy

ZHUANG Xin-Gang[1,2,3], WANG Li-Li[1,3*], WU Xue-Yuan[4], FANG Jia-Xiong[1,3]

(1. Advanced Research Center for Optics, Shandong University, Jinan 250100, China;
2. School of Information Science and Engineering, Shandong University, Jinan 250100, China;
3. State Key Laboratories of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China;
4. State Key Laboratory of Tea and Agricultural Products Detection (Huangshan), Huangshan 245000, China)

Abstract: Near infrared (NIR) spectroscopy was used to identify the origins of two representative of Shandong green tea (Laoshan green tea and Rizhao green tea) rapidly and non-destructively. Several preprocessing methods of NIR, such as smoothing, first- and second-derivative were compared. Moving window back propagation artificial neural network (MW-BP-ANN) was used to select characteristic spectral variables. It was found that the first-derivative and MW-BP-ANN processing techniques improved the predictive abilities of the support vector machine (SVM) classification models. The best estimated identification accuracy can be improved to 98.33%, which demonstrates that the spectral variables selection method is significant for the predictive ability of origin identification models.

Key words: near-infrared spectroscopy, support vector machine, green tea, origin identification
PACS: 42.30.Sy

基于近红外的移动窗口BP神经网络实现山东绿茶产地溯源

庄新港[1,2,3], 王丽丽[1,3*], 吴雪原[4], 方家熊[1,3]

(1. 山东大学 光学高等研究中心,山东 济南 250100;
2. 山东大学 信息科学与工程学院,山东 济南 250100;
3. 中国科学院上海微系统与信息技术研究所 传感技术联合国家重点实验室,上海 200050;
4. 国家茶叶及农产品检测重点实验室(黄山),安徽 黄山 245000)

摘要:将近红外光谱分析技术用于对山东省代表性绿茶(崂山绿茶和日照绿茶)进行快速、无损伤产地溯源.对平滑处理、一阶微分和二阶微分等几种不同的光谱预处理方法进行了系统性对比和研究创新.提出移动窗口BP神经网络(MW-BP-ANN)算法用于选择特征光谱变量.实验发现,一阶微分和移动窗口-BP神经网络可以大幅提高支持向量机(SVM)分类模型的预测能力.经预处理后,分类模型的最优鉴别准确率可达98.33%.研究结果表明,该光谱变量选择方法对提高产地溯源模型的预测能力起到至关重要作用.

关 键 词:近红外光谱分析技术;支持向量机;绿茶;产地溯源

中图分类号:O657.33 **文献标识码**:A

Introduction

Green tea is one of the most popular beverages worldwide. In recent years, it has been subjected to intensive scientific and medical studies, due to a variety of its associated health benefits[1-5]. The yearly output of Chinese green tea reaches hundreds of thousands of tons, and the tea comes primarily from the central and southern regions in China. According to China market research,

Received date: 2015-03-31, **revised date**: 2015-12-12 **收稿日期**:2015-03-31, **修回日期**:2015-12-12
Foundation items: Supported by the State Key Laboratory of Sensor Technology Fund (SKT1202), China Postdoctoral Science Foundation (2012M521319) and thecrosswise project "Application of micro NIR spectrograph in the wireless sensornetwork" (2015-1-1273)
Biography: ZHUANG Xin-Gang (1989-), male, Rizhao, China, Ph. D. Research area is near infrared spectroscopy and spectral-sensing internet of things. E-mail: zhuangxingang@mail.sdu.edu.cn
* **Corresponding author**: E-mail: wanglili1983@sdu.edu.cn

the production of green tea is growing annually at a rate of approximately 7.2% worldwide. The concomitant problem is that the amount of adulterated green tea in the market has also increased dramatically, as reported by the local media and other studies[6]. Furthermore, traditional chemical analysis methods in laboratories are too expensive and complicated to discern such adulteration. Hence, developing an easy-to-use and economic on-line identification approach for green tea is quite urgent.

NIR spectroscopy, which is considered to be a major alternative to the traditional chemical analysis methods (such as high performance liquid chromatography (HPLC), capillary electrophoresis and colorimetric measurements), becomes one of the most rapidly developed spectroscopic methods because of its rapid-response and non-destructive features[7-10]. In addition to food, agriculture and petrochemical industries, NIR spectroscopy has been successfully applied in green tea quality analysis[11-14]. For example, NIR in combination with principal component analysis (PCA), ANN and SVM, has been used in quantitative and qualitative analysis of green tea[15-17].

SVM is one of the most technologically advanced recognition methods. To avoid over-fitting, SVM fixes the classification decision function to the structural risk minimum mistake instead of the minimum mistake of the misclassification on the training set[18]. Therefore, SVM has the advantage of dealing with ill-posed problems and could lead to global models that are often unique. Furthermore, SVM is powerful for the problem characterized by small sample, nonlinearity and high dimension, with a good generalization performance[19]. Recently, SVM was successfully applied to build NIR spectroscopy classification models that generally have a recognition rate of 95%[16, 20-22]. However, almost all of the studies used tea powder as samples to identify green tea varieties, and the spectra were basically collected using a large, costly and high-resolution spectrometer that is not conducive to practical applications.

A rapid and non-destructive method to identify the origins of green tea was established using NIR spectroscopy and SVM. All spectra were collected using a low-resolution spectrometer. Preprocessing had been used to eliminate the difficulties caused by the heterogeneity of tea samples, which could affect the reproducibility of spectra. Then, several spectral preprocessing and spectral variable selection methods were investigated systematically. Laoshan and Rizhao green tea were collected from green tea plantations in their original producing areas to train and test the classification model.

1 Materials and methods

1.1 Sample preparation

200 green tea samples were selected for model building and analysis to well determine the mathematical relationship between the spectral variables and origins. 200 representative green tea samples (100 Laoshan and 100 Rizhao green tea samples) were collected from Laoshan and Rizhao. All samples were collected according to the proportion of regional and seasonal production, which could keep the uniform distribution of sample in time and space and enhance the predictive ability of the model. For experimental clarity, Laoshan green tea samples were labelled 1, and Rizhao green tea samples were labelled 2.

All 200 samples were randomly divided into two subsets at a ratio of 7:3 to ensure the training set and prediction set samples representative in quantity and variation range of properties. The first subset was the training set, which contained 140 samples (70 Laoshan and 70 Rizhao green tea samples). This set was used to build the calibration model. The remaining 60 samples (30 Laoshan and 30 Rizhao green tea samples) formed the prediction set to test the performance of the model. The tea variety, producing area, sample number and sample label are collected in Table1.

Table 1 The varieties, producing area, number and labels of tea samples

表 1 茶叶样本的种类、产地、数量和标签信息

Varieties	Producing area	Number of samples	Labels
Laoshan green tea	Wanggezhuang tea plantation, Qingdao, Shandong	70	1
	Shazikou tea plantation, Qingdao, Shandong	30	
Rizhao green tea	Houcun tea plantation, Rizhao, Shandong	33	2
	Beiguo tea plantation, Rizhao, Shandong	34	
	Jufeng tea plantation, Rizhao, Shandong	33	

1.2 Spectra collection

Ten spectra were collected for each sample in the reflectance mode using AvaSpec-NIR256/2.5TEC spectrometer (Avantes, Netherlands) with a fiber-optic probe. All samples were placed in 200 ml beakers and the spectra were directly collected from the beaker without grinding or any other sample pretreatments. The distance between the probe and tea was kept at 1 cm. Each spectrum was recorded as the average value from the spectra collected from 40 scans. The emission range of the spectrum was 1 050 ~ 2 500 nm, and the raw spectra had 227 variables at intervals of approximately 6.4 nm. For each sample, the mean of the 10 spectra was applied in the subsequent analysis. Additionally, both ends (1050 ~ 1300 nm and 2 300 ~ 2 500 nm) of the original spectra were excluded because the high level of noise prevented an accurate interpretation of the spectra. As a result, the spectral region of 1 300 ~ 2 300 nm was selected for further analysis and included 156 variables. The room temperature was kept at 25℃, and the humidity was kept at an ambient level in the laboratory.

1.3 Spectral preprocessing methods

The purpose of spectral preprocessing is to reduce spectral noise and enhance spectral details to improve the predictive performance of the model. In this work, three data preprocessing methods were applied: spectral smoothing, first- and second-derivative analyses. Because it is a moving window averaging method, spectral smoothing eliminates random white noise by averaging the testing values of repeated measurements[23]; thus, a wider win-

dow results in good smoothing performance. Compared with smoothing, first- and second-derivative analyses can enhance small spectral differences, thereby eliminating overlapped spectral lines and baseline drift. First-derivative processing is usually applied to eliminate a singular point, allowing the linear background to reach a fixed value. Second-derivative processing differentiates between two adjacent first derivatives.

1.4 Spectral variable selection method

MW-BP-ANN was used to select characteristic variables in this study. This is the first time combining MW with BP-ANN to select the spectral variable. The idea of the moving window has been implemented by a number of studies. The moving window method has a wide universality, and it can be globally optimized and easily operated[24-26].

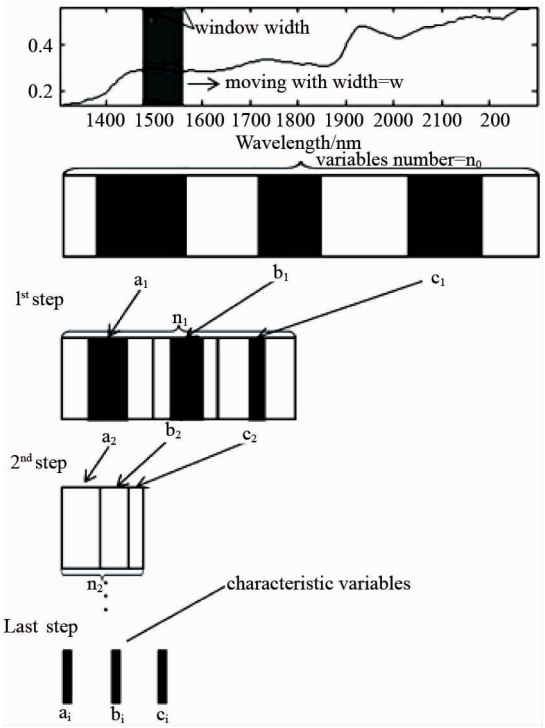

Fig. 1 The schematic of MW-BP-ANN for spectral variable selection

图1 移动窗口-BP神经网络选择特征波长流程图

MW-BP-ANN has two steps for spectral variable selection. The first step is setting the width of the window, which moved from the left side to the right side of the whole spectral range, as shown in Fig. 1. Then, the spectra covered in the window were used to build classification models by using a back propagation-artificial neural network (BP-ANN) as the window was moved. After one movement was completed, all of the prediction results of each model were calculated simultaneously. All of the windows were denoted by the sequence number of their intermediate variables. The second step was to select the optimal spectral variables according to the recognition rate. Finally, all of the selected variables were applied to build the classification model by SVM.

The optimization process of characteristic variables is shown in Fig. 1: the spectral window starts at the first spectral channel, and it ends at the last spectral channel, which moved (n-w) times. n is the number of spectral variables, w is the window width and it is also the main parameter which needed to be adjusted manually. In general, w is adjusted according to the discrimination of the final results. Too broad of a window would easily lead to a problem in which some non-informative variables are selected. In turn, it is difficult to obtain a consecutive range of spectral intervals if the window is too narrow. After the optimization, we got (n-w) BP-ANN models and (n-w) evaluation indices. To evaluate the optimization results, the optimized variables were used to build SVM classification model. The second optimization is based on the result of the first optimization. The process was repeated until getting the optimal SVM classification model.

2 Result and discussion

2.1 Spectra investigation

The mean spectra of two varieties for the raw and first derivative data are presented in Figs. 2(a) and (b), respectively. As presented in Fig. 2(a), the strongest absorption peak is at approximately 1940 nm caused by the stretching and deformation vibrations of O-H in water. The signals at 1 428 ~ 1 666 nm are mainly the stretching vibrations of O-H and N-H[27]. Other obvious absorption bands include the one at 1785 nm (the fundamental stretching of C-H) and 2 172 nm (C-O stretching vibration). The spectra of Laoshan green tea and Rizhao green tea have an obvious separation between 2 100 nm to 2300 nm, which is the strongest absorption band of amino acids (2132 nm and 2294 nm correspond to the stretching vibrations of N-H and C = O, and 2 242 nm is the absorption peak of $-NH_3$)[28].

As shown in Fig. 2(b), the most intense bands in the spectrum belong to the vibration of the 2nd overtone of the carbonyl group (1 868 nm), the C-H deformation vibration (1 386 nm), the stretching vibrations of $-CH_2$ (1 741 nm) and $-CH_3$ (1 712 nm). The presence of carbonyl group, C-H and $-CH_2$ are mainly caused by polyphenols, alkaloids, proteins, volatile, non-volatile acids, and some aromatic compounds[20].

2.2 Spectral preprocessing

As presented in Fig. 3(a), the raw spectra of Laoshan green tea and Rizhao green tea samples are highly overlapped, so specific peaks are difficult to be isolated. To sharpen the poor peak resolution and improve the classification performance of the calibration models, the preprocessing methods of spectral smoothing and first- and second-derivative analyses were applied to process the raw spectra. All of the processed spectra in the training set are shown in Fig. 3. The red lines represent Laoshan green tea samples, and Rizhao green tea samples are indicated by green lines. The dark regions are the overlapping portions of two varieties of tea samples.

Based on the preprocessed spectra, five classification models were built by SVM, and the results of the predictions are shown in Table. 2. As determined from Table. 2, the model based on first-derivative processing has optimal recognition accuracy. Although smoothing can effectively eliminate the noise, it has the risk of los-

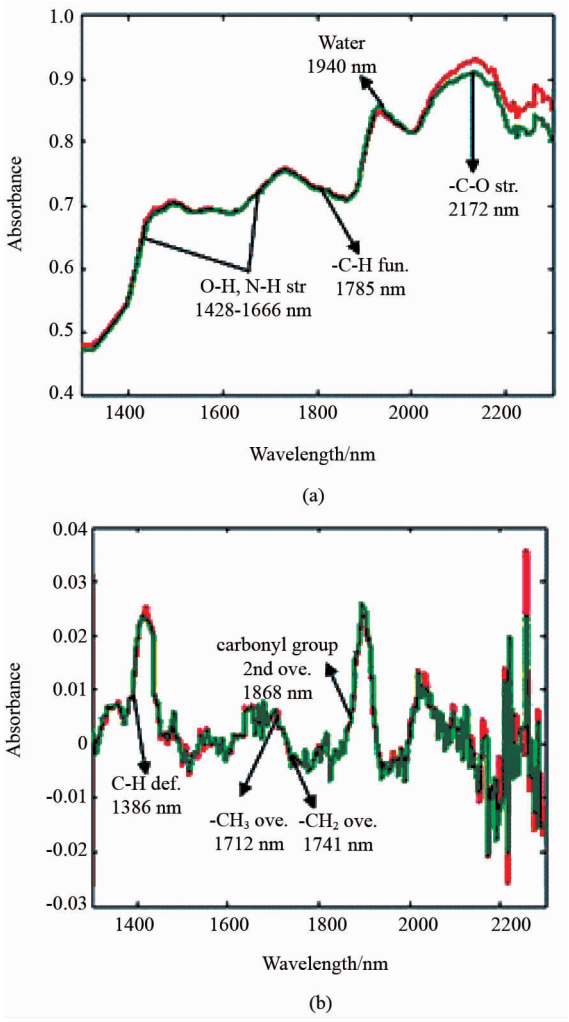

Fig. 2　Mean spectra for two varieties of green tea (red: Laoshan green tea, Green: Rizhao green tea) (a) raw spectra, and (b) first derivative spectra

图 2　两产地绿茶的平均光谱曲线(红色:崂山绿茶,绿色:日照绿茶)(a)原始光谱,(b)一阶微分光谱

ing some useful high-frequency information in the raw data. In addition, more points in the smoothing window will reduce the classification performance. In contrast, the first- and second-derivative spectra can enhance most of the detailed information, which enhances the difference between the Laoshan and Rizhao green tea samples. More importantly, the derivation spectra can remove the baseline and eliminate any influence caused by the heterogeneity of tea samples, thereby enhancing the reproducibility of the spectra.

Table 2　The classification results obtained by spectral preprocessing

表 2　光谱预处理后的产地分类结果

Preprocessing methods	Total No.	False No.	Accuracy rate/ (%)
Raw spectra	60	5	91.67
Three-point smoothing	60	6	90
Five-point smoothing	60	9	85
Seven-point smoothing	60	9	85
First derivative	60	2	96.67
Second derivative	60	5	91.67

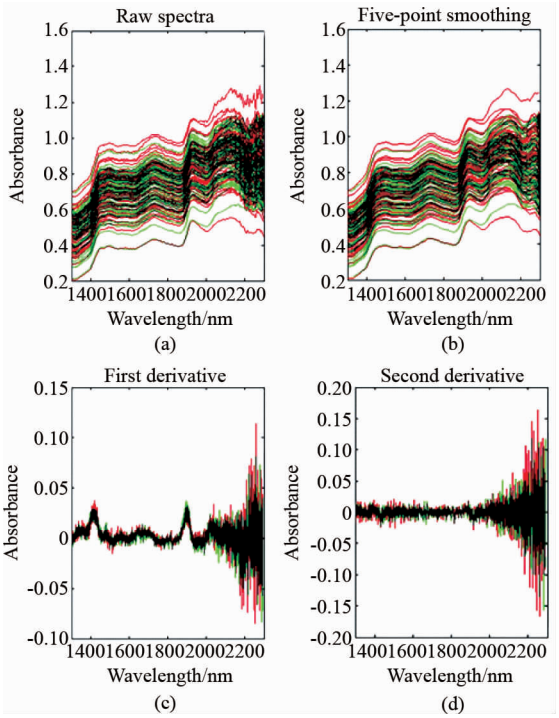

Fig. 3　The raw and processed spectra for two varieties of green tea samples (red: Laoshan green tea, blue: Rizhao green tea, black: overlap). (a) Raw spectra, (b) five-point smoothed spectra, (c) first derivative spectra, and (d) second derivative spectra

图 3　两产地绿茶样本的原始光谱和处理后的光谱(红色:崂山绿茶,绿色:日照绿茶,黑色:光谱重叠部分)(a)原始光谱,(b)五点平滑光谱,(c)一阶微分光谱,(d)二阶微分光谱

2.3　Spectral variable selection

Based on the spectra preprocessed by first derivative analyses, which has the best identification accuracy, the characteristic spectral variables were selected using MW-BP-ANN. The moving window size was set at 3 spectral variables (approximately 14.4 nm), which proved to be the optimal window size based on a series of experiments and calculations. Eventually, 82 characteristic variables were selected, as shown in Fig. 4. All of the accuracy rates corresponding to each window are marked by red boxes. The horizontal line in the graph is the mean value of the accuracy rates and was considered as the threshold. The variables above the line were selected as informative ones to build the classification models.

Most organic compounds in green tea such as polyphenols, amino acids, caffeine and polysaccharide, contain a variety of hydrogen groups. Moreover, the content and proportion of organic ingredients determine the quality of green tea. For example, the content of free amino acids and nitrogen is often related to the variety, region and season of green tea. For this reason, in addition to getting the content of various chemical constituents, a series of quality models can be built based on the chemical constituents, by analyzing the near-infrared spectroscopy of green tea. Taking into account that green tea origins are associated with a variety of chemical constituents, it

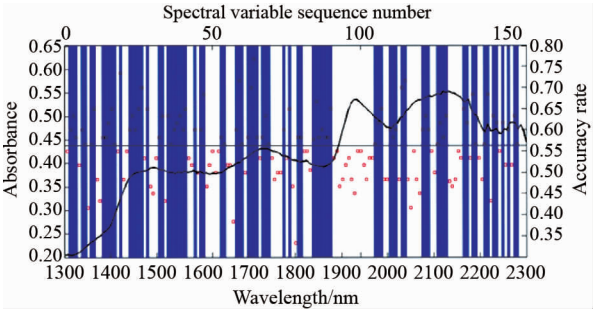

Fig. 4 The selection results of characteristic variables by MW-BP-ANN
图4 移动窗口-BP神经网络选择特征变量结果

is always extremely difficult to find out the variety, content and proportion of compounds, especially for non-professionals. MW-BP-ANN, starting from sample character to find the characteristic variables directly, has strong commonality and global optimization capability to find the characteristic variables directly from sample characters, as presented in Fig. 4.

Based on the 82 characteristic variables, another classification model was built by SVM. Compared with the previous models, the identification accuracy was further improved (accuracy rate = 98.33%), as shown in Fig. 5. Reference labels are indicated by +, and prediction labels are indicated by ○. As observed from Fig. 5, only one prediction sample is not correctly identified.

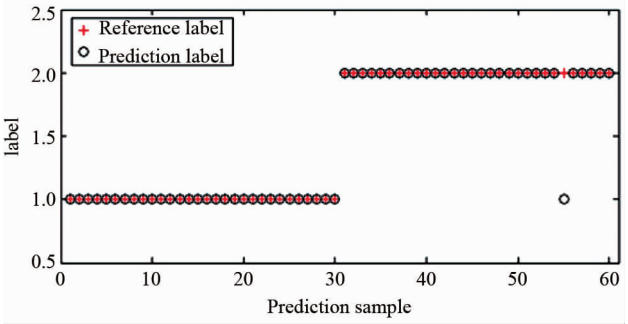

Fig. 5 The final prediction result obtained by SVM
图5 支持向量机分类模型最终预测结果

3 Conclusions

It can be concluded from the above results that Laoshan and Rizhao green tea origins can be well identified by low-resolution NIR spectroscopy and SVM. In the modeling process, preprocessing has a major impact on the final classification model. For example, first-derivative processing techniques can eliminate baseline drift, thereby reducing the poor reproducibility of spectra caused by the heterogeneity of tea samples. MW-BP-ANN was used as an effective method to select characteristic variables, which can further improve the prediction performance of the final classification models. Based on the processed spectra, the optimal classification model was built, and the identification accuracy was 98.33%. Compared with the results obtained by other NIR spectroscopy techniques, the method established in our study was shown to be relatively more practical and economical because it has lower requirements for spectral resolution and samples.

Acknowledgements

This work is supported by the State Key Laboratory of Sensor Technology Fund (No. SKT1202), China Postdoctoral Science Foundation (No. 2012M521319) and thecrosswise project "Application of micro NIR spectrograph in the wireless sensornetwork" (No. 2015-1-1273).

References

[1] Asif Siddiqui F, Naim M, Islam N. Apoptotic effect of green tea polyphenol (EGCG) on cervical carcinoma cells[J]. *Diagn Cytopathol*. 2011, **39**(7): 500-504.
[2] Zheng P, Zheng H M, Deng X M, *et al*. Green tea consumption and risk of esophageal cancer: a meta-analysis of epidemiologic studies [J]. *BMC Gastroenterol*. 2012, **12**(165).
[3] Stordrange L, Libnau F O, Malthe-S Renssen D, *et al*. Feasibility study of NIR for surveillance of a pharmaceutical process, including a study of different preprocessing techniques[J]. *J Chemom*. 2002, **16**(8-10): 529-541.
[4] Zaveri N T. Green tea and its polyphenolic catechins: Medicinal uses in cancer and noncancer applications[J]. *Life Sci*. 2006, **78**(18): 2073-2080.
[5] Zhang J Y, Liu S L, Wang Y. Gene association study with SVM, MLP and cross-validation for the diagnosis of diseases[J]. *Progress in Natural Science*. 2008, **18**(6): 741-750.
[6] Xu L, Zhou Y P, Wu H L, *et al*. Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration[J]. *Anal Chim Acta*. 2008, **616**(2): 138-143.
[7] Chen Q S, Zhao J W, Chaitep S, *et al*. Simultaneous analysis of main catechins contents in green tea (Camellia sinensis (L.)) by Fourier transform near infrared reflectance (FT-NIR) spectroscopy[J]. *Food Chem*. 2009, **113**(4): 1272-1277.
[8] Wei K, Wang L Y, Zhou J, *et al*. Comparison of catechins and purine alkaloids in albino and normal green tea cultivars (Camellia sinensis L.) by HPLC[J]. *Food Chem*. 2012, **130**(3): 720-724.
[9] Li P, Dong S Q, Wang Q J, *et al*. Analysis of trace ingredients in green tea by capillary electrophoresis with amperometric detection[J]. *Chinese J Chem*. 2008, **26**(3): 485-488.
[10] El-Hady D A, El-Maali N A. Determination of catechin isomers in human plasma subsequent to green tea ingestion using chiral capillary electrophoresis with a high-sensitivity cell[J]. *Talanta*. 2008, **76**(1): 138-145.
[11] Alishahi A, Farahmand H, Prieto N, *et al*. Identification of transgenic foods using NIR spectroscopy: A review[J]. *Spectrochim Acta A Mol Biomol Spectrosc*. 2010, **75**(1):1-7.
[12] Mouazen A M, De Baerdemaeker J, Ramon H. Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer[J]. *Soil Till Res*. 2005, **80**(1-2): 171-183.
[13] Roggo Y, Chalus P, Maurer L, *et al*. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies [J]. *J Pharm Biomed Anal*. 2007, **44**(3): 683-700.
[14] Luypaert J, Zhang M H, Massart D L. Feasibility study for the use of near infrared spectroscopy in the qualitative and quantitative analysis of green tea, Camellia sinensis (L.)[J]. *Anal Chim Acta*. 2003, **478**(2): 303-312.
[15] Chen Q S, Guo Z M, Zhao J W, *et al*. Comparisons of different regressions tools in measurement of antioxidant activity in green tea using near infrared spectroscopy[J]. *J Pharm Biomed Anal*. 2012, **60**: 92

-97.

[16] Zhao J W, Chen Q S, Huang X Y, *et al*. Qualitative identification of tea categories by near infrared spectroscopy and support vector machine [J]. *J Pharm Biomed Anal*. 2006, **41**(4): 1198-1204.

[17] Chen Q S, Zhao J W, Lin H. Study on discrimination of Roast green tea (Camellia sinensis L.) according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition[J]. *Spectrochim Acta A Mol Biomol Spectrosc*. 2009, **72**(4): 845-850.

[18] Hyun-Chul K, Shaoning P, Hong-Mo J, *et al*. Constructing support vector machine ensemble[J]. *Pattern Recognit*. 2003, **36**(12): 2757-2767.

[19] Li X L, He Y. Evaluation of least squares support vector machine regression and other multivariate calibrations in determination of internal attributes of tea beverages[J]. *Food Bioprocess Tech*. 2010, **3**(5): 651-661.

[20] Chen Q S, Zhao J W, Fang C H, *et al*. Feasibility study on identification of green, black and Oolong teas using near-infrared reflectance spectroscopy based on support vector machine (SVM)[J]. *Spectrochim Acta A Mol Biomol Spectrosc*. 2007, **66**(3): 568-574.

[21] Balabin R M, Safieva R Z, Lomakina E I. Near-infrared (NIR) spectroscopy for motor oil classification: From discriminant analysis to support vector machines[J]. *Microchem J*. 2011, **98**(1): 121-128.

[22] Wu D, Yang H Q, Chen X J, *et al*. Application of image texture for the sorting of tea categories using multi-spectral imaging technique and support vector machine[J]. *J Food Eng*. 2008, **88**(4): 474-483.

[23] CHU Xiao-Li, YUAN Hong-Fu, LU Wan-Zhen. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique[J]. *ProgChem*(褚小立,袁洪福,陆婉珍. 近红外分析中光谱预处理及波长选择方法进展与应用. **化学进展**),2004, **16**(04): 528-542.

[24] Navvab Kashani M, Aminian J, Shahhosseini S, *et al*. Dynamic crude oil fouling prediction in industrial preheaters using optimized ANN based moving window technique[J]. *Chem Eng Res Des*. 2012, **90**(7): 938-949.

[25] Du Y P, Liang Y Z, Jiang J H, *et al*. Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares[J]. *Anal Chim Acta*. 2004, **501**(2): 183-191.

[26] Kang N, Kasemsumran S, Woo Y, *et al*. Optimization of informative spectral regions for the quantification of cholesterol, glucose and urea in control serum solutions using searching combination moving window partial least squares regression method with near infrared spectroscopy [J]. *Chemometr Intell Lab*. 2006, **82**(1-2): 90-96.

[27] Xu L, Shi P T, Fu X S, *et al*. Protected geographical indication identification identification of a Chinese Green Tea (Anji-White) by near-infrared spectroscopy and chemometric class modeling techniques[J]. *J Spectrosc*. 2013, **2013**(501924): 1-8.

[28] SUN Yao-Guo, LIN Min, LV Jin, *et al*. Determination of the contents of free amino acids, caffeine and tea polyphenols in green tea by Fourier transform near-infrared spectroscopy[J]. *Chinese Journal of Spectroscopy Laboratory*(孙耀国,林敏,吕进,等. 近红外光谱法测定绿茶中氨基酸、咖啡碱和茶多酚的含量. **光谱实验室**), 2004, **21**(5): 940-943.