基于可见反射光谱和遗传区间偏最小二乘法的血迹年龄预测研究

董永芳 孟耀勇 张平丽 文 玮 李 娜

华南师范大学生物光子学研究院激光生命科学教育部重点实验室, 广东 广州 510631

摘要 精确的血迹年龄预测具有重大的法医学价值。利用可见反射光谱技术与偏最小二乘法(PLS)相结合分析预测血迹年龄。遗传算法与偏最小二乘法相结合被用来选择有效光谱区间。与全光谱PLS模型相比较,建立在优化光谱区间的遗传区间偏最小二乘法(GA-iPLS)模型具有更好的预测能力。结果表明GA-iPLS能合理地选择有效光谱区间,提高预测能力。在考虑取自不同个体血迹特异性的情况下,建立在2.00~48.00 h时间段和48.00~1080.00 h时间段的GA-iPLS模型的相关系数(R_p)、预测标准误差(RMSEP)和剩余预测偏差(RPD)分别为0.9949/0.9924、1.59 h/43.56 h、10.32/8.42。两个GA-iPLS模型的结合可代替建立在2.00~1080.00 h时间段的GA-iPLS模型精确预测2.00~1080.00 h时间段的血迹年龄。结果表明可见反射光谱与GA-iPLS模型在法医学领域可成为一种可靠的精确预测血迹年龄的方法。

关键词 光谱学; 年龄预测; 遗传区间偏最小二乘法; 可见反射光谱; 法医学

中图分类号 O433.4 **文献标识码** A

doi: 10.3788/AOS201535.0830001

Accurate Age Estimation of Bloodstains Based on Visible Reflectance and Genetic Algorithm Interval Partial Least Squares

Dong Yongfang Meng Yaoyong Zhang Pingli Wen Wei Li Na

Key Laboratory of Laser Life Science, College of Biophotonics, South China Normal University, Guangzhou, Guangdong 510631, *China*

Abstract Accurate age estimation of bloodstains can provide enormous forensic value. Visible reflectance spectroscopy technique combined with partial least squares (PLS) is applied to accurately estimate the age of bloodstains for forensic purposes. Genetic algorithm (GA) combined with PLS is used to select the most efficient spectral intervals. The genetic algorithm interval partial least squares (GA-iPLS) models built in the optimal intervals, present better predictive capability than full-spectrum PLS model. GA-iPLS can validly select desirable intervals and improve predictive ability. Considering the effect of the specificity of bloodstains on models, the GA-iPLS models built in age period from 2.00 h to 48.00 h and in age period from 48.00 h and 1080.00 h are achieved with correlation coefficient (R_p) of 0.9949/0.9924, root-mean-square error of prediction (RMSEP) of 1.59 h/43.56 h , residual predictive deviation (RPD) of 10.32/8.4243 respectively, which can be used to accurately predict the age of bloodstains from 2.00 h to 1080.00 h instead of GA-iPLS model in 2.00~1080.00 h. The results demonstrate that visible reflectance spectroscopy combined with GA-iPLS will be a reliable tool to accurately estimate the age of bloodstains for forensic practical applications.

Key words spectroscopy; age estimation; genetic algorithm interval partial least squares; visible reflectance spectroscopy; forensic

收稿日期: 2015-03-25; **收到修改稿日期**: 2015-05-04

基金项目: 国家自然科学基金(60878063)、广东省中医药项目(2008233)

作者简介: 董永芳(1990—),男,硕士研究生,主要从事反射光谱及拉曼光谱数据处理方面的研究。

E-mail: my101202@sina.com

导师简介: 孟耀勇(1963—),男,研究员,博士生导师,主要从事反射光谱及拉曼光谱技术应用及分析等方面的研究。

E-mail: yaoyongmeng@sina.cn(通信联系人)

OCIS codes 300.6170; 300.6550; 170.1470

1 引 言

在犯罪现场发现的血迹具有重大的法医学价值，比如脱氧核糖核酸(DNA)分析[1]和犯罪时间预测[2]。精确的血迹年龄预测可以用来推断犯罪发生时间。当血迹是犯罪现场唯一的证据时，血迹年龄预测显得尤其重要。预测的血迹年龄可用于核实证人的证词、为不在场设置参照物或者判断血迹是否与犯罪相关[2]。近年来，多种光谱技术，包括电子顺磁共振光谱[3]、红外光谱[4]、高光谱成像[2,5]、拉曼光谱[6-7]和反射光谱[8-10]等，都被用来预测血迹年龄。然而，大部分光谱技术存在设备昂贵、测试条件苛刻、破坏样品等缺点，这些缺点限制了光谱技术在法医学领域的应用与推广。可见反射光谱技术由于具有无损、无接触、简单、价格低廉等优势，受到了越来越强烈的关注。2011年，Li 等[10]用一个微型光谱仪采集血迹的反射光谱，并建立线性判别分析模型预测血迹年龄。结果表明，当一个新的取自相同个体的血迹样本用于预测集时，分类正确率由91.5% 降为37.3%。可见，取自相同个体的血迹特异性[7]对模型影响较大。Brememe r等[8]通过运用线性最小二乘法拟合反射光谱中三种血红蛋白衍生物的比例来预测血迹年龄。结果显示，真实年龄为35 d的血迹，预测年龄范围为25~55 d。虽然以上技术都能预测血迹的年龄，但是这些技术显然都没有考虑取自不同个体的血迹特异性[7]对模型的影响，并且都具有巨大的年龄预测误差。而相关的研究表明，采用化学计量学方法在提高模型预测能力上能够发挥非常重要的作用[11-12]。偏最小二乘法(PLS)是目前最广泛应用于光谱数据分析的一种化学计量方法，具有很强的抗干扰能力[11-16]。利用遗传区间偏最小二乘法(GA-iPLS)建立模型可有效剔除不相干变量，提高预测能力[17-18]。

本文采用遗传区间偏最小二乘法对建模区域进行优化，建立分段血迹预测模型，并与全谱范围建立的偏最小二乘法模型进行比较分析，进而得到预测能力最佳的血迹年龄可见光预测模型。

2 实验与方法

2.1 样品采集

8个血液样品采集自8个健康捐献者，采集时间为上午8:00~8:20。分别吸取15 μL血液滴在玻片上，制得8个血迹样品。8个血迹样品储存于37 ℃的恒温环境下，相对湿度为10%。8个血迹样品按照3:1的比例随机分成两部分，其中6个血迹作为校正集，2个血迹作为预测集。

2.2 反射光谱采集

采用美国海洋光学公司生产的USB-4000微型光纤光谱仪，光谱测定范围为200~1100 nm。光谱采集时，以漫反射方式测量样品光谱。先用标准白板校准光谱，再对血迹样品进行扫描，每扫描30次自动平均为一个光谱。所有的光谱采集均在暗室中进行。所有样品分别在2.00, 4.00, 8.00, 12.00, 24.00, 36.00, 48.00, 72.00, 168.00, 240.00, 360.00, 720.00, 1080.00 h测得反射光谱，共获取104幅光谱。

2.3 多元数据分析及模型建立

为去除来自测量首位处的高频随机噪声，选取400~800 nm光谱范围进行分析，并采用标准正态变换对光谱进行预处理。在Matlab平台中建立不同偏最小二乘法预测模型。留一交叉验证法用于确定模型建立所需的最佳主成分因子(LVs)。采用相关系数 R、校正均方根误差(RMSEC) C_{RMSE}、预测均方根误差(RMSEP) P_{RMSE}、剩余预测偏差(RPD) D_{RP}指标对模型进行评价。相关系数越接近1，RMSEC和RMSEP值越小且越接近，模型的预测能力越好[19]。当RPD值大于3时表示模型有较好的稳健性，且RPD值越大稳健性越好[20]。

3 结果与讨论

3.1 光谱预处理

所有光谱均进行标准正态变换(SNV)预处理。光谱预处理结果如图1所示。图1(a)为血迹年龄为8.00 h的8个血迹的原始反射光谱。从图1可知，受到基线平移和散射作用的影响，同样年龄的血迹光谱间存在较大差异[21]。经过光谱预处理后，这些差异显著地减小，如图1(b)所示。因此对比分析图1(a)和图1(b)可知，光谱预处理技术能有效地消除基线平移和散射作用所带来的光谱差异。

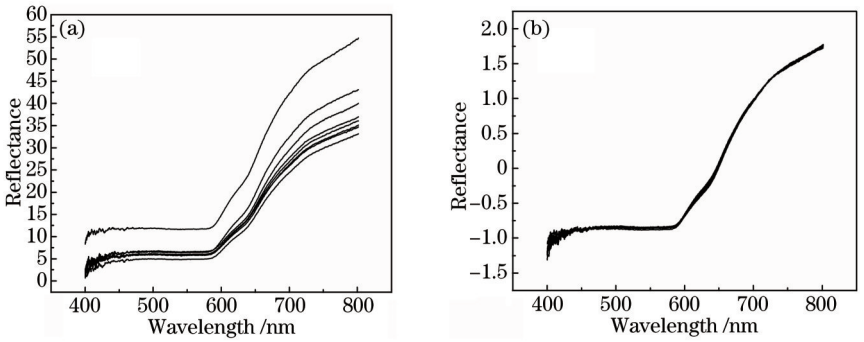

图1 年龄相同的8个血迹的光谱。(a) 原始反射光谱; (b) SNV预处理后光谱

Fig.1 Spectra of eight bloodstains with the same age. (a) Raw reflectance spectra; (b) spectra after SNV pretreatment

3.2 偏最小二乘法建模

为避免模型过拟合,用于建立PLS模型所需的主成分因子均不大于10[22]。通过综合考虑交叉验证均方根(RMSECV)和预测均方根误差来确定最佳LVs的数目。所取PLS LVs的数量对RMSECV和RMSEP的影响如图2所示。从图中可知,RMSECV的值随着LVs值的增大而减小,但RMSEP先随LVs值增大而减小,后随LVs增大而增大,表明LVs值超过6将导致模型过拟合。可见最佳LVs值应为6。该模型的性能指标如表1所示,其中D_{RP}=3.19,P_{RMSE}=101.68 h,表明该模型具有较好的稳健性及预测能力,但预测精度仍需提高。通过分析可知,导致该模型低精度的一个重要原因是一部分包含不相关信息的光谱削弱了模型的性能。因此,剔除不相关信息可以建立具有高精度的模型。众所周知,GA-iPLS算法可以用来优化建模区域,剔除不相关信息[13-14]。利用GA-iPLS算法改进上述模型的预测精度。

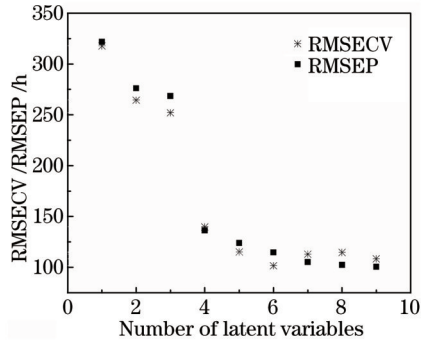

图2 LVs数量与RMSECV和RMSEP值的关系

Fig.2 Relationship between the number of latent variables and RMSEP, RMSECV values

3.3 遗传区间偏最小二乘法建模

首先,在2.00~1080.00 h时间段建立GA-iPLS模型。在模型中,将全光谱划分成20个等宽子区间。通过遗传算法寻找子区间组合,建立相应的模型。通过比较各模型的RMSECV,选择具有最优预测性能的模型。所得最佳GA-iPLS模型性能指标如表1所示,与PLS模型相比,RPD值显著增大,而P_{RMSE}减小了54.23 h,表明GA-iPLS能有效去除无效区间,进而提升预测能力。校正集和预测集的血迹年龄预测结果如图3所示,R_P=0.9888,表明预测值与实际值存在较好的线性关系,从总体上看,该模型具有很好的稳健性和预测能力,但是在2.00~48.00 h时间段,相对预测误差较大,预测能力不足,该时间段的预测精度需要提高。

表1 模型性能

Table 1 Performance of models

Model	Time interval	RMSEC /h	RMSEP /h	R_C	R_P	RPD
PLS	2 h~45 d	82.84	101.68	0.9656	0.9476	3.19
GA-iPLS	2 h~45 d	37.64	47.45	0.9930	0.9888	6.84
GA-iPLS	2 h~48 h	1.30	1.59	0.9968	0.9949	10.32
GA-iPLS	48 h~45 d	31.72	43.56	0.9960	0.9924	8.42

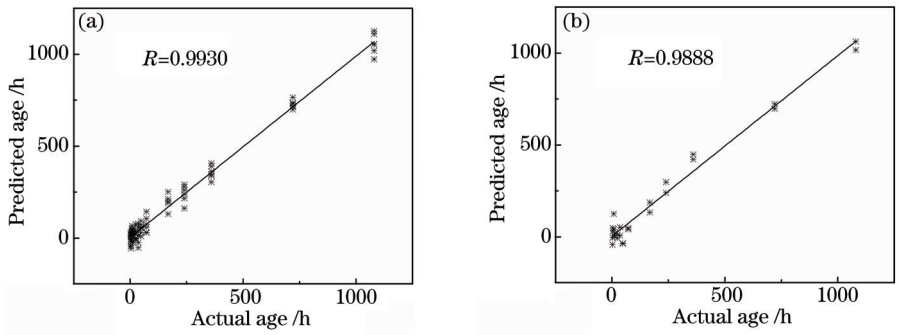

图3 2.00~1080.00 h时间段GA-iPLS模型预测结果。(a) 校正集; (b) 预测集

Fig.3 Prediction results of GA-iPLS model built in age period from 2.00 h to 1080.00 h. (a) Calibration set; (b) prediction set

为解决上述不足,分别在2.00~48.00 h和48.00~1080.00 h两个时间段建立了GA-iPLS模型,使用分段模型预测血迹年龄,模型性能如表1所示。对于2.00~48.00 h时间段的GA-iPLS模型,校正集和预测集的血迹年龄预测结果如图4所示,其性能指标分别为 P_{RMSE}=1.59 h, R_p=0.9949, D_{RP}=10.32,与全时间段GA-iPLS模型相比,预测值与实际值之间的相关性增强,模型更加稳定且精确度更高,预测能力显著提升,同时也解决了全时间段GA-iPLS模型不能精确预测2.00~48.00 h时间段血迹年龄的问题。该模型主要用于预测离案发时间较短的血迹年龄,预测偏差仅仅为1.59 h。此外,对于48.00~1080.00 h时间段的遗传区间偏最小二乘法模型,校正集和预测集的血迹年龄预测结果如图5所示,P_{RMSE}=43.56 h, R_p=0.9924, D_{RP}=8.42,表明该模型具有非常好的预测能力和稳健性,进一步提高了48 h~45 d时间段的预测精度。该模型主要用于预测离案发时间较长的血迹年龄,预测偏差仅仅为43.56 h。两个模型的RMSEC与RMSEP值相差较小,不仅能表明模型具有很好的稳健性,也表明遗传区间偏最小二乘法能减小血迹个体差异对模型的影响。上述两个模型所选择的子区间如图6所示。位于540 nm和576 nm附近的含氧血红蛋白(HbO_2)反射带以及分别位于600 nm和650 nm附近的高铁血红蛋白(met-Hb)和半血色质(HC)反射带都包含于所选子区间内[8]。可见反射光谱技术预测血迹时间的主要依据是血液中HbO_2, met-Hb, HC的转换[8-10]。可通过GA-iPLS算法选择有效光谱区间,去除不相关光谱变量,从而尽可能排除血液中其他不相关因素,如白细胞、血小板、蛋白质等会导致血迹产生个体特异性的物质,对其模型预测的影响。子区间选择结果表明在考虑血迹特异性的情况下,GA-iPLS可以合理地选择有效子区间,去除与血迹年龄不相关或包含大量噪声的光谱变量,尽可能减小血液特异性对模型预测能力的影响。

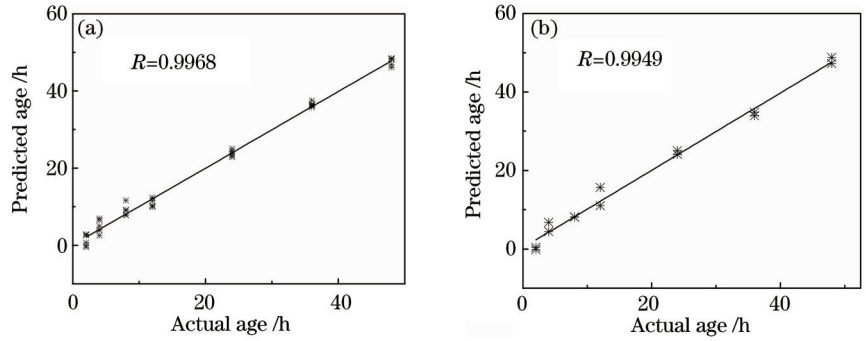

图4 2.00~48.00 h时间段GA-iPLS模型预测结果。(a) 校正集; (b) 预测集

Fig.4 Prediction results of GA-iPLS model built in age period from 2.00 h to 48.00 h. (a) Calibration set; (b) prediction set

3.4 比较分析

大多数关于血迹年龄的研究都没有对预测精度进行量化[3, 8-9],因而很难相互比较。但与其中预测精度最好的研究成果相比[5, 10],该研究所得精度显著提升。在Li等[10]的研究中,虽然采用了线性判别分析与可见反射光谱相互结合的方法预测血迹年龄,但是由于没有考虑取自相同个体的血迹的特异性对模型的影响,所以所得分类正确率仅为37.3%,模型的泛化能力不足,不能用于实际应用中。2013年,Li等[5]使用线性判别分析与高光谱成像技术相结合,建立线性判别分析模型预测血迹年龄,由于该模型考虑了取自同一个体血迹所存在的特异性,与之前研究成果相比,获得了更好的预测结果。但是新的模型并未考虑取自不同个体

的血迹的特异性对模型的影响。而研究表明,取自不同个体血迹之间的差异比取自相同个体的血迹之间的差异更大[7]。因此在本文中既考虑了同一个体血迹所存在的特异性,又考虑了不同个体的血迹的特异性,血液的特异性被充分考虑,为模型的实际运用提供了重要保证。根据Li等[5]论文中的血迹预测结果,计算了其D_{RP}、P_{RMSE}和R_p值,分别为3.55、60.48 h、0.9594。通过比较以上模型指标可以得知,在充分考虑血迹特异性的情况下,分段GA-iPLS模型具有更好的预测能力和稳健性。

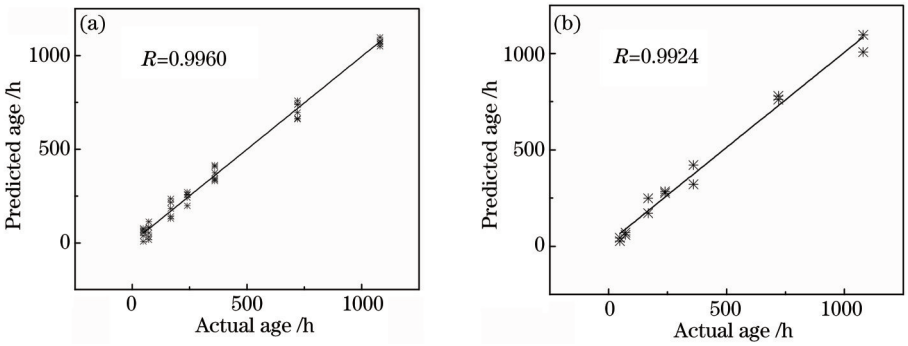

图5 48.00~1080.00 h时间段GA-iPLS模型预测结果。(a) 校正集; (b) 预测集

Fig.5 Prediction results of GA-iPLS model built in age period from 48.00 h to 1080.00 h. (a) Calibration set; (b) prediction set

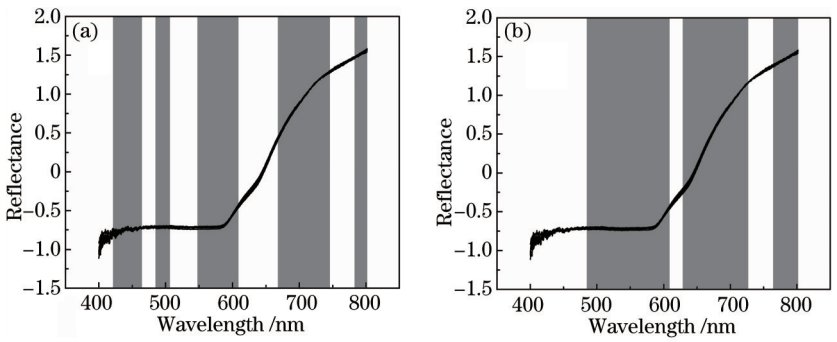

图6 不同模型优化区间选择结果。 (a) 2.00~48.00 h时间段GA-iPLS模型; (b) 48.00~1080.00 h时间段模型

Fig.6 Selected optimum intervals. (a) GA-iPLS model built in age period from 2.00 h to 48.00 h; (b) GA-iPLS model built in age period from 48.00 h to 1080.00 h

4 结 论

利用可见反射光谱结合遗传区间偏最小二乘法精确地预测了血迹年龄。结果表明,遗传区间偏最小二乘法可以有效地优化建模区域,使得所建模型具有良好的血迹年龄预测能力。建立在2.00~48.00 h时间段和48.00~1080.00 h时间段的GA-iPLS模型能代替建立在2.00~1080.00 h时间段的GA-iPLS模型预测血迹年龄。在充分考虑血迹特异性的前提下,与前人的研究结果相比,所建预测模型的预测能力显著提升,预测误差显著减小。因此,可见反射光谱结合GA-iPLS模型可以成为一种高精度的血迹年龄预测手段,将会在法医学领域中有重要应用价值。

参 考 文 献

1 S H James, P E Kish, T P Sutton. Principles of Bloodstain Pattern Analysis: Theory and Practice[M]. Boca Raton: CRC Press, 2005: 219-225.

2 G Edelman, T G van Leeuwen, M C Aalder. Hyperspectral imaging for non-contact analysis of forensic traces[J]. Forensic Science International, 2012, 223(1): 28-39.

3 Y Fujita, K Tsuchiya, S Abe, *et al*.. Estimation of the age of human bloodstains by electron paramagnetic resonance spectroscopy: Long-term controlled experiment on the effects of environmental factors[J]. Forensic Science International, 2005, 152(1): 39-43.

4 G Edelman, V Manti, S M van Ruth, *et al*.. Identification and age estimation of blood stains on colored backgrounds by near infrared spectroscopy[J]. Forensic Science International, 2012, 220(1): 239-244.

5 B Li, P Beveridge, W T O′Hare, *et al*.. The age estimation of blood stains up to 30 days old using visible wavelength hyperspectral image analysis and linear discriminant analysis[J]. Science and Justice, 2013, 53(3): 270-277.

6 W R Premasiri, J C Lee, L D Ziegler. Surface-enhanced Raman scattering of whole human blood, blood plasma, and red blood cells: Cellular processes and bioanalytical sensing[J]. Journal of Physical Chemistry B, 2012, 116(31): 9376-9386.

7 S Boyd, M F Bertino, S J Seashols. Raman spectroscopy of blood samples for forensic applications[J]. Forensic Science International, 2011, 208(1): 124-128.

8 R H Bremmer, A Nadort, T G van Leeuwen, *et al*.. Age estimation of blood stains by hemoglobin derivative determination using reflectance spectroscopy [J]. Forensic Science International, 2011, 206(1): 166-171.

9 E K Hanson, J Ballantyne. A blue spectral shift of the hemoglobin soret band correlates with the age (time since deposition) of dried bloodstains[J]. PLoS one, 2010, 5(9): 12830-12840.

10 B Li, P Beveridge, W T O′ Hare, *et al*.. The estimation of the age of a blood stain using reflectance spectroscopy with a microspectrophotometer, spectral pre-processing and linear discriminant analysis[J]. Forensic Science International, 2011, 212(1): 198-204.

11 Yu Xiaoya, Zhang Yujun, Yin Gaofang, *et al*.. Feature wavelength selection of phytoplankton fluorescence spectra based on partial least squares[J]. Acta Optica Sinica, 2014, 34(9): 0930002.
余晓娅, 张玉钧, 殷高方, 等. 基于偏最小二乘回归的藻类荧光光谱特征波长选取[J]. 光学学报, 2014, 34(9): 0930002.

12 Li Gang, Zhao Jing, Li Jiaxing, *et al*.. Visible-Infrared reflectance spectroscopy applied in rapid screen of diseases[J]. Acta Optica Sinica, 2011, 31(3): 0317001.
李 刚, 赵 静, 李家星, 等. 可见-近红外反射光谱用于疾病快速筛查[J]. 光学学报, 2011, 31(3): 0317001.

13 Wang Chunlong, Liu Jianguo, Zhao Nanjing, *et al*.. Quantitative analysis of laser-induced breakdown spectroscopy of heavy metals in water based on support-vector-machine regression[J]. Acta Optica Sinica, 2013, 33(3): 0330002.
王春龙, 刘建国, 赵南京, 等. 基于支持向量机回归的水体重金属激光诱导击穿光谱定量分析研究[J]. 光学学报, 2013, 33(3): 0330002.

14 Zhang Haidong, Li Guirong, Li Ruocheng, *et al*.. Determination of tea polyphenols content in Puerh tea using near-infrared spectroscopy combined with extreme learning machine and GA-PLS algorithm[J]. Laser & Optoelectronics Progress, 2013, 50(4): 043001.
张海东, 李贵荣, 李若城, 等. 近红外光谱结合极限学习机和GA-PLS算法检测普洱茶茶多酚含量[J]. 激光与光电子进展, 2013, 50(4): 043001.

15 Weng Shizhuang, Zheng Shouguo, Li Pan, *et al*.. Quantitative analysis of fenitrothion based on surface-enhanced Raman spectroscopy [J]. Chinese J Lasers, 2013, 40(8): 0815001..
翁士状, 郑守国, 李 盼, 等. 基于表面增强拉曼光谱的杀螟硫磷定量分析[J]. 中国激光, 2013, 40(8): 0815001.

16 Zhao Jiewen, Bi Xiakun, Lin Hao, *et al*.. Visible-near-infrared transmission spectra for rapid analysis of the freshness of eggs[J]. Laser & Optoelectronics Progress, 2013, 50(5): 053003.
赵杰文, 毕夏坤, 林 颢, 等. 鸡蛋新鲜度的可见-近红外透射光谱快速识别[J]. 激光与光电子学进展, 2013, 50(5): 053003.

17 Wang Jiahua, Pan Lu, Sun Qian, *et al*.. Nondestructive measurement of SSC in western pear using genetic algorithms and FT-NIR spectroscopy[J]. Spectroscopy and Spectral Analysis, 2009, 29(3): 678-681.
王加华, 潘 璐, 孙 谦, 等. 遗传算法结合偏最小二乘法无损评价西洋梨糖度[J]. 光谱学与光谱分析, 2009, 29(3): 678-681.

18 Li Yanxiao, Zou Xiaobo, Dong Ying. Near infrared determination of sugar content in apples based on GA-iPLS[J]. Spectroscopy and Spectral Analysis, 2007, 27(10): 2001-2004.
李艳肖, 邹小波, 董 英. 用遗传区间偏最小二乘法建立苹果糖度近红外光谱模型[J]. 光谱学与光谱分析, 2007, 27(10): 2001-2004.

19 Z Guo, Q Chen, L Chen, *et al*.. Optimization of informative spectral variables for the quantification of EGCG in green tea using Fourier transform near-infrared (FT-NIR) spectroscopy and multivariate calibration[J]. Applied Spectroscopy, 2011, 65(9): 1062-1067.

20 D Cozzolino, W U Cynkar, N Shah, *et al*.. Multivariate data analysis applied to spectroscopy: Potential application to juice and fruit quality [J]. Food Research International, 2011, 44(7): 1888-1896.

21 R J Barnes, M S Dhanoa, S Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra [J]. Applied Spectroscopy, 1989, 43(5): 772-777.

22 X Zou, J Zhao, Y Li. Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of ‘Fuji’ apple based on BiPLS and FiPLS models[J]. Vibrational Spectroscopy, 2007, 44(2): 220-227.

栏目编辑: 吴秀娟