可见-近红外反射光谱用于疾病快速筛查

李 刚[1] 赵 静[1] 李家星[1,2] 林 凌[1] 张宝菊[3]*

([1] 天津大学精密测试技术及仪器国家重点实验室，天津 300072
[2] 天津科技大学海洋科学与工程学院，天津 300072；[3] 天津师范大学物理与电子信息学院，天津 300387)

摘要 为了探讨基于舌诊的疾病快速筛查，运用可见和近红外光谱仪，采集 149 名志愿者舌尖的反射光谱并且进行反射率归一化处理。根据临床诊断结果将样本分为 4 组：健康组、高粘血症倾向组、脂肪肝患者组和冠心病患者组。运用主成分分析(PCA)结合人工神经网络(ANN)方法、偏最小二乘(PLS)方法和间隔偏最小二乘(iPLS)方法 3 种方法建立分类预测模型。预测准确率分别为 75%，75%和 85%。实验结果表明，在 3 种建模方法中，iPLS 预测效果最好，与可见光波段相比，近红外波段含有更多与疾病分类相关的光谱信息。实验的结果表明，光谱法用于某些疾病的快速诊断具有较高的可行性。

关键词 光谱学；疾病诊断；主成分分析；人工神经网络；偏最小二乘法；间隔偏最小二乘法

中图分类号 O433.4 **文献标识码** A **doi**：10.3788/AOS201131.0317001

Visible-Infrared Reflectance Spectroscopy Applied in Rapid Screen of Diseases

Li Gang[1] Zhao Jing[1] Li Jiaxing[1,2] Lin Ling[1] Zhang Baoju[3]

([1] *State Key Laboratory of Precision Measurement Technology and Instruments, Tianjin University, Tianjin* 300072, *China*
[2] *College of Marine Science and Engineering, Tianjin Uinversity of Science and Technology, Tianjin* 300072, *China*
[3] *College of Physics and Electronic Information, Tianjin Normal University, Tianjin* 300387, *China*)

Abstract To screen disease which based on tongue inspection rapidly, the reflection spectrum on the tongue tips of 149 volunteers were collected by visible and near-infrared spectrometer and then the normalized reflectivity was calculated. Samples were divided into four classes according to the clinical diagnosis information: healthy, hyperviscosity, fatty liver, and coronary heart disease groups. Spectra were then subjected to three different analysis methods: principle component analysis (PCA) combined with artificial neural network (ANN), partial least squares (PLS), and interval PLS (iPLS). The classification accuracy of each model are 75%, 75%, and 85%, respectively. The results show that iPLS method sees more robust than the others. And the results also show that near-infrared region including more disease information than visible region. Experimental results show that the application of the spectra for disease diagnosis is promising.

Key words spectroscopy; disease diagnosis; principal component analysis (PCA); artificial neural network (ANN); partial least square (PLS); interval partial least square (iPLS)

OCIS codes 170.0170; 300.6170; 240.0240

1 引 言

随着社会的进步，人们饮食的结构也发生了变化，现代人类摄取了更多的脂肪和糖分，而纤维素的摄取量相对减少[1]。这种饮食上的变化，以及快节奏的现代生活带给人类的压力导致了一些慢性疾病的发生。如脂肪肝、高粘血症以及近年越来越受到

收稿日期：2010-08-11；**收到修改稿日期**：2010-10-27

基金项目：国家自然科学基金(30973964)和天津市应用基础及其前沿技术研究计划(10JCYBJC00400)资助课题。

作者简介：李 刚(1959—)，男，博士，教授，博士生导师，主要从事精密机械与测试技术及生物医学信号检测与处理等方面的研究。E-mail：ligang59@tju.edu.cn

* **通信联系人**。E-mail：cetty3190@163.com

重视的冠心病，这些疾病严重威胁着人类的健康，所以早期诊断和治疗对于控制疾病的发展显得极为重要。现在，对高粘血症的主要诊断手段是血流变各项指标检查；对脂肪肝的主要诊断手段是B超、计算X射线层析(CT)和核磁共振；而对冠心病的主要诊断手段是心电图、运动平板和冠状动脉造影[2,3]。这些方法中B超、心电图运用得最为普遍，然而B超只能对肝脏形状和大小进行检测，诊断结果受到图片质量和医师经验的影响。心电图方法虽然简洁，然而诊断准确率不高；其他诊断方法多少都会给人体带来一些痛苦或伤害。

本文研究采用光谱法对疾病进行快速诊断。光谱法能利用多波长下的光谱数据对物质进行分析，更全面、更客观地反映组织细胞的生理病理变化，探究不同个体之间的细微差别。由于光谱技术有操作简单、速度快、效率高和成本低等优势，近年来在食品[4,5]、农业[6~11]、石油化工和医学制药[12~16]等领域得到越来越广泛的应用。

舌诊属于中医望诊的一部分，对于诊断疾病有着非常重要的作用。同时从选取测量点角度分析，Burmeister等[17,18]用实验证明了舌体上布满血管，没有脂肪组织，比起指尖以及手臂等部位能更准确地反映人体内部微循环等方面的信息，并且证明了舌体的组织结构特征使得舌体成为疾病无创诊断很好的测量点。由物理学中物体的反射率原理可知，物体对于某一波长光的反射率应是物体自身的物理特性，不随光源光谱成分的变化而变化。基于这一原理，本课题组提出了一种基于光谱技术的舌诊客观化研究新方法，通过对数据进行反射率归一化，只考虑各个波长上光谱之间的相对关系，排除了背景噪声以及在操作过程中的不稳定因素带来的干扰，通过对不同舌象进行比较，取得了一些成果[19]。

为了进一步探究舌体所携带的生理病理信息，本文采用可见和近红外光谱仪采集舌尖反射光谱，运用主成分分析(PCA)和人工神经网络(ANN)的方法、偏最小二乘法(PLS)和间隔偏最小二乘法(iPLS)建立健康组、高粘血症组、脂肪肝组和冠心病组四组分类预测模型，三种方法取得了75%，75%和85%的预测正确率。实验结果验证了光谱法运用于疾病快速诊断具有一定的可行性。

2　材料和方法

2.1　来检者情况

149位来检者来自天津塘沽区永久医院中医体检科和石家庄医科大学第一附属医院心内科，年龄在20～80岁之间。根据医师临床诊断结果，149位来检者中41位健康，26位患有高粘血症倾向，48位患有不同程度的脂肪肝，34位患有冠心病。采集前1 h内来检者都未进食，所以舌体能够较客观地反应人体的真实病理信息。测量时，被测者为坐姿，自然将舌伸出口外，舌体自然放松，尽量将口张开。

2.2　光谱测定

采用Ocean公司的USB2000可见光谱仪和NIR512近红外光谱仪，并用配套的SpectralSuit软件进行采集，光源和光线采用定制的GY-30光纤耦合溴钨灯及其配套的光纤。装置连接如图1所示。采集数据前，光源先预测5 min，然后将光纤探头距离舌尖表面1 cm处垂直照射。由于软件有采集时间光谱功能，实验中近红外光谱仪NIR512的积分时间定为35 ms，同一点采集50次，采集到的数据经SpectraSuit软件转化为光强，在853.59～1737.26 nm可以采集512组数据。由于光谱仪特性的限制，853.59 nm处采集数据为零，最后进入运算的波长数为511组数据。可见光谱仪USB2000的积分时间定为3 ms，同一点采集50次，在462.87～1136.16 nm可以采集2048组数据。同样由于光谱仪特性的限制，462.87 nm处采集数据为零，最后进入运算的波长数为2047组数据。

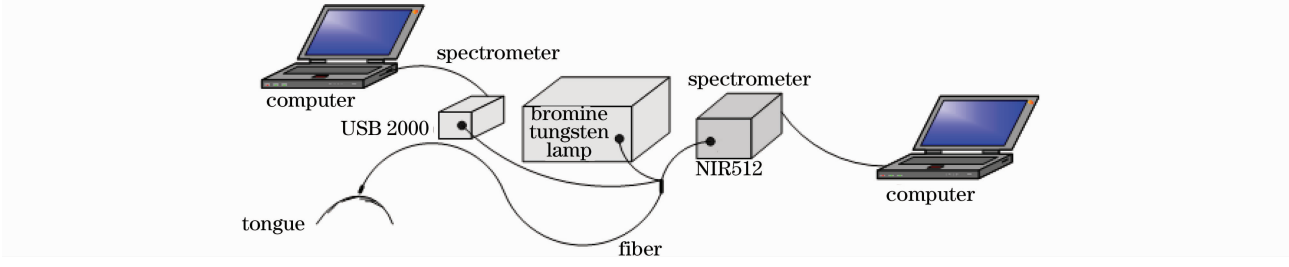

图1　实验装置图

Fig. 1　Sketch of experimental device

2.3　光谱数据预处理

首先采集光源和舌尖两部分的光谱数据。光源光谱通过多次测量获得平均值。通过光谱仪采集得到的数据为各点反射光强信息，这里将舌尖光谱数据在各个波长上计算反射率。

由于外界环境的干扰和测量时光纤探头与被测者舌体之间不同等因素的影响，将处理后的反射率进行归一化，只考虑各个波长上反射率数据之间的相对关系，从而消除照明和采集带来的影响。计算方法为

$$R_g = R/R_{max}, \tag{1}$$

式中 R_g 为归一化反射率，R 为反射率，R_{max} 为不同波长上的反射率最大值。

149 例样本反射光谱如图 2 所示，图 2(a)为 USB2000 所采集的样本反射光强曲线图，图 2(b)为 NIR512 采集的归一化反射率曲线图。

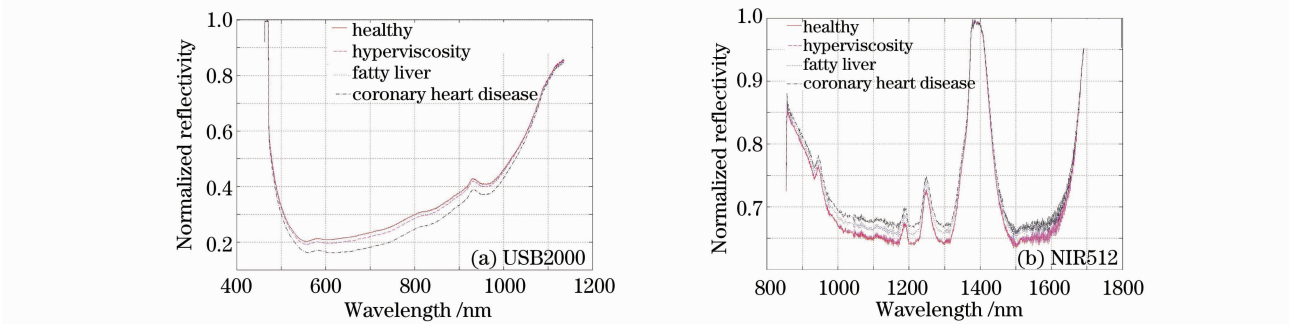

图 2　不同光谱仪测得的归一化反射率平均值

Fig. 2　Average of normalized reflectivity detected by different spectrometers

3　模型建立结果与分析

3.1　模型建立结果

3.1.1　PCA 结合 ANN

PCA 是多元统计中的一种数据挖掘技术，该方法能将数据进行压缩，起到数据降维的作用，以消除众多信息中相互共存重叠的部分。它将原变量进行变换，线性组合为少数几个新变量，而且新变量能够最大限度地表征原变量的数据结构特征，不丢失信息。对采集来的 149 例样本的舌尖近红外反射光谱数据进行 PCA，前 11 个主成分累计贡献率达到 98.98%。图 3 为被测者前两个主成分贡献得分图。图中横轴表示样本的第一个主成分得分(PC1)，纵轴表示样本的第二个主成分得分(PC2)。

NIR512 光谱仪在近红外波段共采集 512 个波长数据，如果将全部数据用来建立模型则数据量大。由于前 11 个主成分已经能代表原数据 98.98%信息，所以将前 11 个主成分作为 ANN 的输入，这样能够大大减少 ANN 的计算量。将 149 例样本分为训练集和预测集，其中训练集 129 例样本，预测集 20 例样本，样本均为随机抽取，建立 11-7-1 三层逆向传播神经网络(BPANN)。设定输出值为分类结果，设定健康组为 1，高粘血症倾向患者组为 2，脂肪肝患者组为 3，冠心病患者组为 4。各层传递函数分别为 sigmoid，sigmoid 和 purelin，训练函数为 traingdm，目标误差设定为 0.001，训练次数为 8000 次。训练集模型预测结果如图 4 所示。

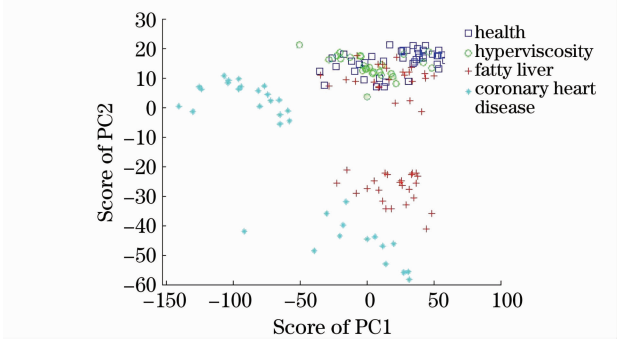

图 3　149 例样本主成分 1 和主成分 2 得分图

Fig. 3　PCA scores of PC1 and PC2 for 149 samples

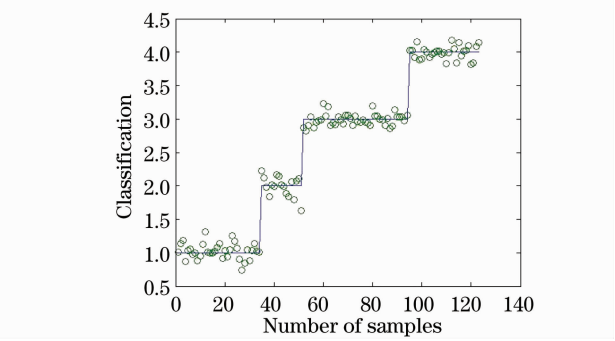

图 4　ANN 建模结果

Fig. 4　Result of ANN modeling

运用所见模型对 20 例预测样本集进行预测，预测准确率 75%。

3.1.2　PLS

为了充分比较分析反射光谱用于不同疾病快速诊断的可行性，又采用了光谱建模分析中最为常用的 PLS 进行分析[20]。PLS 将海量光谱数据进行压缩，通过提取新的特征变量来代表原始光谱数据的有效信息[21]。以模型的均方根误差（RMSE）E_{RMS} 为指标来选择最优的主因子数，其计算公式为

$$E_{\mathrm{RMS}}=\sqrt{\frac{\sum(y_{\mathrm{m}}-y_{\mathrm{p}})^2}{n}},\tag{2}$$

式中 y_{m} 为测量值，y_{p} 为预测值，n 为样本数。

建立 PLS 定量校正模型时，主因子数的选择直接关系到模型实际的预测能力。如果主因子数太少，重建光谱拟合就会不够，如果主因子数太多，对光谱重建又会产生过度拟合。本文以模型的交叉验证 RMSE 为指标，选择最优的主因子数。主因子数对 RMSE 的影响如图 5 所示。由图 5 可以看出，当主因子数为 14 时，RMSE 值最小。所以，最优主因子数为 14。运用所建 PLS 预测模型，对训练集的预测结果与临床诊断结果之间的相关系数 r 达到 0.9132，E_{RMS}=0.4589，训练集预测结果平均偏差为 −0.0137，建模结果如图 6 所示。

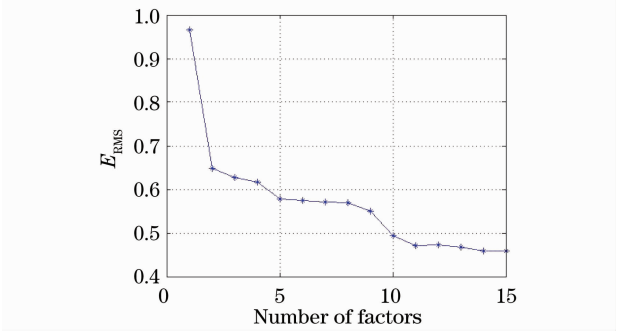

图 5　不同主因子数对 E_{RMS} 的影响

Fig. 5　Influences of different principal factors on E_{RMS}

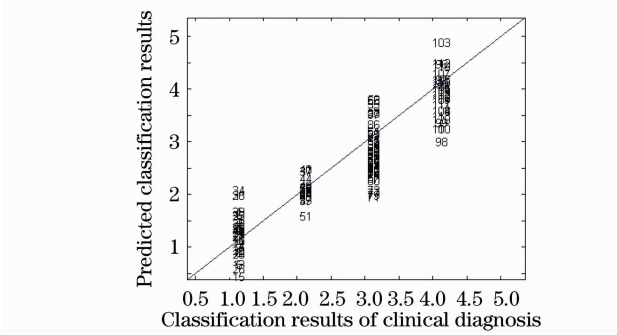

图 6　PLS 建模结果

Fig. 6　Result of PLS modeling

3.1.3　iPLS

为获得更多的光谱信息，使用了两个光谱仪，大大加大了需要处理的光谱数据量，给数据建模带来了困难。为了寻找与预测分类关系最密切的波段，采用了 iPLS。该方法能将整个波段分成若干小波段，再在每个小波段上分别建立 PLS 模型，根据建模的结果，来选择效果最好的波段或者几个较好波段的组合来建立模型[22]。

将整个光谱范围分为 15 个波段，通过比较发现第 14 个波段（波长区间 1188.07～1461.59 nm）在选择 PLS 成分数为 9 时建模效果最好，模型相关系数 r 达到 0.9339，E_{RMS}=0.4046，训练集预测结果平均偏差为 0.0012。第 14 个波段建模结果如图 7 所示。

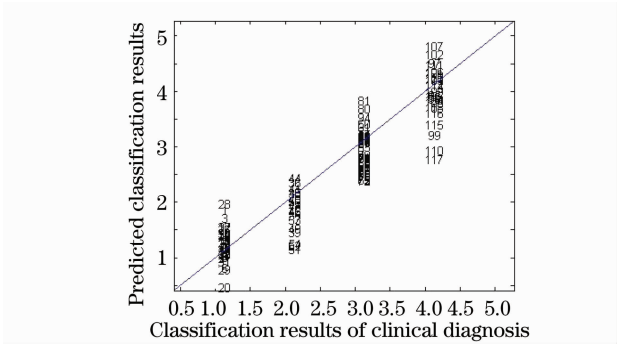

图 7　iPLS 建模结果

Fig. 7　Result of iPLS modeling

3.1.4　三种方法预测结果

分别运用三种方法对预测集 20 例样本进行预测。其中 PCA 结合 ANN 预测准确率 75%，PLS 预测准确率 75%，iPLS 预测准确率 85%。

3.2　实验结果分析

由图 3 前两个主分量的得分图可见，四类中除冠心病组外，其他三组之间都存在一定程度交叉，需要进一步建立模型进行分离。由图 4，6，7 和表 1 可见，运用 PCA 结合 ANN，PLS 以及 iPLS 均能建立比较好的分类预测模型，训练集临床诊断结果值和预测值之间相关度较高。

由表 1 可见，运用所建立的三个不同模型对同一预测集样本进行预测，三种方法比较，PCA 结合 ANN 虽然得到了好的建模效果，然而预测准确率不高，而 iPLS 方法得到了较高的预测相关系数和准确率。说明运用多波段的光谱数据，在获得更多信息的同时，也带来了更多的冗余信息，给模型的建立带来困难。由 iPLS 方法选取波段的实验结果可见，与可见光波段相比，近红外波段与疾病分类相关度更高，用来建立本实验所述疾病分类模型可以得到更高的预测准确率。这一结果为进一步的研究奠定基础。

表 1　三种方法建模效果

Table 1　Performance of the three different models

Method	Component numbers	Band range / nm	RMSE	Correlation coefficients of test samples	Correlation coefficients of prediction samples	Classification accuracy/%
PCA and ANN	11	463.25～1737.26	0.1189	0.994	0.846	75
PLS	14	463.25～1737.26	0.4589	0.9132	0.902	75
iPLS	9	1181.07.25～1461.59	0.4046	0.9339	0.932	85

运用模型对预测集样本进行预测，三种模型预测准确率分别为75%，75%和85%。该实验结果证明了iPLS方法比另外两种方法更稳定，且预测效果更好。同时，表1的预测结果也证明了光谱法运用于疾病快速诊断的可行性。

通过对来检者临床病例分析，样本虽然被分为4类：然而个别类中又存在不同的情况，如脂肪肝组又分为轻度、中度和重度脂肪肝三种类型。冠心病组又分为心肌缺血、心绞痛和心肌梗塞等类型。现在还没有大量的数据样本建立更精细的分类模型。本研究将这些归为同一大类，是造成预测准确率不高的主要原因，需要补充数据改善分类。

4　结　　论

同时采集来检者舌尖部位可见-近红外反射光谱，根据临床诊断结果，运用三种方法建立分类预测模型：PCA结合ANN，PLS和iPLS。从实验结果可以看出，反射光谱法用于疾病快速诊断具有一定的可行性。虽然数据量等方面的支持有待加强，然而作为探索性研究，有可能为临床提供一种快速、简便的诊断手段。

参　考　文　献

1 C. W. C. Kendall, A. Esfahani, D. J. A. Jenkins. The link between dietary fibre and human health [J]. *Food Hydrocolloids*, 2010, **24**(1): 42～48

2 Wan Hongmei, Li Lanfang. New development of diagnosis the coronary heart disease [J]. *Contemporary Medicine*, 2008, **14**(20): 71～72
万红梅，黎兰芳. 冠心病诊断的新进展[J]. 当代医学，2008，**14**(20)：71～72

3 Yu Xuefang. Comparison between diagnosis of coronary heart disease by dynamic cardiogram and treadmill exercise test[J]. *J. Practical Electro Cardiology JS*, 2009, **18**(2): 122～123
于雪芳. 动态心电图诊断冠心病心肌缺血与平板运动试验的比较与评价[J]. 实用心电学杂志，2009，**18**(2)：122～123

4 He Yong, Li Xiaoli, Shao Yongni. Discrimination of varieties of apple using near infrared spectra based on principal component analysis and artificial neural network model[J]. *Spectroscopy and Spectral Analysis*, 2006, **26**(5): 850～853
何　勇，李晓丽，邵咏妮. 基于主成分分析和神经网络的近红外光谱苹果品种鉴别方法研究[J]. 光谱学与光谱分析，2006，**26**(5)：850～853

5 Pang Taotao, Yao Jianbin, Du Liming. Artificial neural networks for the indentification of infrared spectra of ilex kudingcha [J]. *Spectroscopy and Spectral Analysis*, 2007, **27**(7): 1336～1339
庞涛涛，姚建斌，杜黎明. 人工神经网络分类鉴别苦丁茶红外光谱[J]. 光谱学与光谱分析，2007，**27**(7)：1336～1339

6 Liu Huanjun, Zhang Bai, Wang Zongming *et al*.. Soil saline-alkalization evaluation basing on spectral reflectance characteristics [J]. *J. Infrared and Millimeter Vaves*, 2008, **27**(2): 138～142
刘焕军，张　柏，王宗明 等. 基于反射光谱特征的土壤盐碱化评价[J]. 红外与毫米波学报，2008，**27**(2)：138～142

7 Zhu Shiping, Wang Gang, Yang Fei *et al*.. Rapid detection method of the spicy components in zanthoxylum bungeagum maxim by near infrared spectroscopy [J]. *J. Infrared and Millimeter Vaves*, 2008, **27**(2): 129～132
祝诗平，王　刚，杨　飞 等. 基于近红外光谱的花椒麻味物质快速检测方法[J]. 红外与毫米波学报，2008，**27**(2)：129～132

8 Q. Fan, Y. Wang, P. Sun *et al*.. Discrimination of ephedra plants with diffuse reflectance FT-NIRS and multivariate analysis [J]. *Talanta*., 2010, **80**(3): 1245～1250

9 G. Steiner, T. Bartels, M. E. Krautwald Junghanns *et al*.. Sexing of turkey poults by fourier transform infrared spectroscopy [J]. *Analytical and Bioanalytical Chemistry*, 2010, **396**(1): 465～470

10 C. Y. Wu, A. R. Jacobson, M. Laba *et al*.. Surrogate correlations and near-infrared diffuse reflectance sensing of trace metal content in soils[J]. *Water, Air & Soil Pollution*., 2009, **209**(1-4): 377～390

11 Chen Quansheng, Zhang Yanhua, Wan Xinmin *et al*.. Study on detection of pork tenderness using hyperspectral imaging technique [J]. *Acta Optica Sinica*, 2010, **30**(9): 2602～2607
陈全胜，张燕华，万新民 等. 基于高光谱成像技术的猪肉嫩度检测研究[J]. 光学学报，2010，**30**(9)：2602～2607

12 Tian Jian, Jin Lihong, Zhao Lihui *et al*.. Quantitative analysis of mixture of DNA bases with near-infrared diffuse reflectance spectrometry [J]. *J. Changchun University of Science and Technology (Natural Science Edition)*, 2008, **31**(4): 1～4
田　坚，金丽红，赵丽辉 等. 近红外漫反射光谱法对DNA碱基混合物的定量分析[J]. 长春理工大学学报(自然科学版)，2008，**31**(4)：1～4

13 Li Gang, Wang Yan, Li Qiuxia *et al*.. Theoretic study on improving noninvasive measurement accuracy of blood component by dynamic spectrum method [J]. *J. Infrared and Millimeter Vaves*, 2006, **25**(5): 345～348
李　刚，王　焱，李秋霞 等. 动态光谱法对提高近红外无创血液成分检测精度的理论分析[J]. 红外与毫米波学报，2006，**25**(5)：345～348

14 I. E. Bell, G. V. G. Baranoski. Reducing the dimensionality of plant spectral databases [J]. *IEEE Transaction on Geoscience and Remote Sensing*, 2004, **42**(3): 570～576

15 Su Rongguo, Liang Shengkang, Zhu Chenjian *et al*..

Fluorescence discrimination technology of bacillariophyta and pyrrophyta[J]. *Environmental Science and Technology*, 2008, **31**(3): 52～55
苏荣国, 梁生康, 祝陈坚 等. 硅藻和甲藻的荧光识别测定技术研究[J]. 环境科学与技术, 2008, **31**(3): 52～55

16 Feng Shangyuan, Chen Rong, Li Yongzeng *et al*.. Surface enhanced Raman spectroscopy of dangshen decotion [J]. *Chinese J. Lasers*, 2010, **37**(1): 121～124
冯尚源, 陈　荣, 李永增 等. 党参煎剂表面增强拉曼光谱[J]. 中国激光, 2010, **37**(1): 121～124

17 J. J. Burmeister, M. A. Arnold, G. W. Small. Noninvasive blood glucose measurements by near-infrared transmission spectroscopy across human tongues[J]. *Diabetes Technology and Therapeutics*., 2000, **2**(1): 5～16

18 J. J. Burmeister, M. A. Arnold. Evaluation of measurement sites for noninvasive blood glucose sensing with near-infrared transmission spectroscopy[J]. *Clinical Chemistry*, 1999, **45**(9): 1621～1627

19 Lin Ling, Xie Xin, Li Gang. Tongue manifestation based on spectral method [J]. *Spectroscopy and Spectral Analysis*, 2009, **29**(3): 707～710
林　凌, 谢　鑫, 李　刚. 基于光谱的中医舌色客观化方法初探[J]. 光谱学与光谱分析, 2009, **29**(3): 707～710

20 J. G. de Aguiar, A. Borin, R. J. Poppi. Determination of viscosity and solids in pressuresensitive adhesives by FTIR-ATR and multivariate calibration[J]. *J. Brazilian Chemical Society*, 2010, **21**(3): 436～440

21 Wang Zunyi, Jin Chunhua, Liu Fei *et al*.. Rapid discrimination of soil variety based on spectroscopic techniques[J]. *J. Zhejiang University* (*Agric. & Life Sci.*), 2010, **36**(3):282～286
王遵义, 金春华, 刘　飞 等. 基于光谱技术的土壤快速分类方法研究[J]. 浙江大学学报(农业与生命科学版), 2010, **36**(3): 282～286

22 L. Norgaard, A. Saudland, J. Wagner *et al*.. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy [J]. *Appl. Spectrosc.*, 2000, **54**(3): 413～419