第30卷第5期
2011年10月

红外与毫米波学报
J. Infrared Millim. Waves

Vol. 30, No. 5
October, 2011

文章编号:1001-9014(2011)05-0458-05

BiPLS结合模拟退火算法的近红外光谱特征波长选择研究

石吉勇[1], 邹小波[1*], 赵杰文[1], 毛罕平[2]
(1. 江苏大学 食品与生物工程学院,江苏 镇江 212013;
2. 江苏大学 现代农业装备与技术重点实验室,江苏 镇江 212013)

摘要:为了简化近红外光谱模型,提高对草莓可溶性固形物含量的预测精度,将反向偏最小二乘法(BiPLS)与模拟退火算法(Simulated annealing algorithm,SAA)相结合优选特征波长,建立了多元线性回归可溶性固形物光谱模型.原始光谱经过预处理后,用反向偏最小二乘法优选出4个特征子区间(分别为第8、13、16、17);对所选的特征子区间,进一步用模拟退火算法选择可溶性固形物的特征波长.在SAA选择出的7565 cm^{-1}、7706 cm^{-1}、8289 cm^{-1}、8489 cm^{-1}、8499 cm^{-1}、8724 cm^{-1}、8807 cm^{-1} 7个特征波数点的基础上建立了预测模型.模型的预测均方根误差为0.428,优于偏最小二乘法、向后区间偏最小二乘法建模结果.研究结果表明:反向偏最小二乘法结合模拟退火算法可以有效选择近红外光谱特征波长.

关 键 词:近红外光谱;模拟退火算法;反向偏最小二乘法;波长选择
中图分类号:O657.33 **文献标识码**:A

Selection of wavelength for strawberry NIR spectroscopy based on BiPLS combined with SAA

SHI Ji-Yong[1], ZOU Xiao-Bo[1*], ZHAO Jie-Wen[1], MAO Han-Ping[2]
(1. College of Food and Biological Engineering, Jiangsu University, Zhenjiang 212013, China;
2. Key Laboratory of Modern Agricultural Equipment and Technology, Jiangsu University, Zhenjiang 212013, China)

Abstract: To improve and simplify the NIR prediction model of the soluble solid content (SSC) of strawberry, backward interval partial least squares (BiPLS) and simulated annealing algorithm (SAA) were combined to select the efficient wavelengths. The strawberry spectra were divided into 21 intervals, among which 4 subsets, i.e. No. 8, 13, 16 and 17 were selected by BiPLS. Then SAA was used to select variables in these informative regions, which were used for regression variables of MLR model. Finally, 7565 cm^{-1}, 7706 cm^{-1}, 8289 cm^{-1}, 8489 cm^{-1}, 8499 cm^{-1}, 8724 cm^{-1} and 8807 cm^{-1} were used to build a MLR model. The MLR model performs well with root mean standard error of prediction (RMSEP) of 0.428 for SSC, which out performs models using PLS and BiPLS. This work proved that the BiPLS-SAA could determine optimal variables in NIR spectra and improve the accuracy of model.

Key words: NIR spectroscopy; Backward interval PLS; Simulated annealing algorithm; Wavelength selection
PACS: 07.05Kf

引言

近红外光谱分析技术是近年来快速发展起来的分析技术,同传统的化学检测方法相比,近红外光谱分析技术具有不破坏样品、速度快、效率高、成本低、重现性好、运用范围广等特点,被广泛的运用到食品、药品行业,如对梨[1-2]、苹果[3-4]、桃[5]及杨梅[6]等的内外品质指标进行检测.

国内外学者对近红外光谱分析技术的研究主要集中于波段/波长选择方法和建模方法两方面.现有的波段/波长选择方法主要有间隔偏最小二乘法(iPLS)、前向偏最小二乘法(FiPLS)、反向偏最小二

收稿日期:2010-12-27,**修回日期**:2011-06-18 **Received date**: 2010-12-27, **revised date**: 2011-06-18
基金项目:国家863项目(2008AA10Z208);国家自然科学基金(60901079);江苏省六大人才高峰,青蓝工程项目;国家博士后基金;优秀博士论文基金;江苏省普通高校研究生科研创新计划项目资助(CX10B_277Z)
作者简介:石吉勇(1984-),男,湖南株洲人,博士生,主要研究方向为农产品品质无损检测研究. E-mail:stoneboy_2007@sohu.com.
* **通讯作者**:E-mail:zou_xiaobo@ujs.edu.cn.

乘法[7](BiPLS)、遗传算法[8](GA)、模拟退火算法[9](SAA)等,其中 iPLS、FiPLS、BiPLS 主要用于近红外特征波段选择;GA、SAA 既可以进行特征波段选择,也可以进行特征波长选择. 现有的建模方法主要有多元线性回归(MLR),逐步线性回归(SMLR),主成分回归(PCR),偏最小二乘法(PLS),人工神经网络(ANN)等. 中外学者的研究结果表明,上述方法均能有效用于波长选择和模型建立,然而每种方法都有各自的优劣,没有哪一种方法是万能的. 本文以近红外光谱技术的在线检测为出发点,采用简便、可靠的多元线性回归作为建模方法. 为了降低回归方程的复杂度和提高模型预测能力,在建立模型前对近红外光谱进行变量选择,从而移除那些同检测指标不相关和信噪比低的变量. 常用的 MLR 变量筛选方法为逐步回归分析方法,但该方法选取的变量间具有多重交互作用,模型中的一个变量可能会屏蔽其它变量并对结果产生影响,因此,逐步回归法选取的变量往往达不到最优. 模拟退火算法作为一种变量选择方法,是解决大规模组合优化问题,特别是 NP 完全类问题的有效近似优化算法. 同其它近似算法相比,具有描述简单、使用灵活、运行效率高和较少受到初始条件限制等优点. 为了避免初始搜索空间过大,提高算法运行的针对性,需要对波数点进行初步筛选. 反向偏最小二乘法可以剔除相关性较差的区间,对近红外光谱子区间进行初步定位,经过 BiPLS 选择出来的子区间可以作为 SAA 的最初解.

可溶性固形物主要指食品中所有溶容性糖类物质或者其它可溶性物质,包含的组分比较复杂,难以直接定位其对应的特征波长,故需要采用优化组合算法寻找光谱中最相关的信息. 本研究首先采用 BiPLS定位出若干光谱子区间,再用 SAA 在选出的特征子区间内优选特征波长,最终以选出的所有波长点对应的光谱信号作为 MLR 回归变量,对可溶性固形物含量建立 MLR 回归模型,同时考察校正模型对未知样本的预测能力.

1　原理与算法

1.1　反向偏最小二乘

iPLS 的原理是将整个光谱分割成 k 个等宽子区间,然后在每个子区间进行偏最小二乘回归. 采用留一交互验证法计算各个子区间的交互验证均方根误差(RMSECV),当 RMSECV 值最小时,对应的因子数为子区间最佳因子数,根据最佳因子数在各子区间建立局部最优 PLS 模型.

BiPLS 是在 iPLS 的基础上,依次减少信息量最差或共线性变量最多的 $i(i=0,1,2\cdots\cdots k)$ 个子区间,即去除 RMSECV 值最大的区间,在剩余的 $k-i$ 个区间上建立最优 PLS 模型,并给出相应的 RMSECV值. 当 RMSECV 最小时所对应的多个区间即为所优化的组合区间.

1.2　模拟退火算法

(1) 模拟退火算法基本原理

根据 Metropolis 准则[9],粒子在温度 T 时趋于平衡的概率 P 为

$$P = \exp[-\Delta E/(\mathrm{k}T)] \quad , \qquad (1)$$

式中 E 为温度 T 时粒子的内能,ΔE 为其改变量,k 为 Boltzmann 常数.

用固体退火模拟组合优化问题,先确定初始温度,随机选择一个初始状态并考察该状态的目标函数值(f);然后在当前解的领域中,以一定概率选择一个非局部最优解,并令这个解再重复下去,从而不会陷入局部最优. 算法由一个控制参数 T 决定,解经过大量迭代变换后,可求得给定控制参数 T 时优化问题的相对最优解. 然后缓慢减小控制参数 T,重复上诉迭代过程. 对温度为 T 时的所有迭代过程称为一个马尔科夫链,迭代次数称为马尔可夫链长度(L_k). 当计算完温度 T 对应的马尔科夫链时,温度 T 按一定冷却率(α)逐渐减小,重复上述过程直至温度 T 趋于 0 时,最终得到问题的全局最优近似解.

(2) 控制参数设计

根据模拟退火算法基本原理,要成功使用模拟退火算法解决实际问题,必须合理选择目标函数(f)对应退火过程中粒子的能量 E. 算法收敛速度取决于 T_k 和 L_k 的选择,因此如何合理选择一组控制算法进程的参数,使算法在有限时间内返回一个近似最优解,是该算法的关键. 这样的一组控制参数通常称为冷却进度表,它主要包括以下参数:(1)起始温度 T_0;(2)温度衰减函数 $\alpha(T_k=\alpha T)$;(3)终止温度 T_f;(4)马尔科夫链长度 L_k;下面将讨论如何设置这些控制参数.

(3) 目标函数的选取

在波长筛选过程中,常采用交互验证法来评价模型的预测能力,即采用交换验证均方根误差(RMSECV)、预测残差平方和(PRESS)、待测组分预测值与实测值之间的相关系数(r)等作为目标函数. 如采用 RMSECV 作为评价指标,RMSECV 的值越小,对应校正模型的预测能力越好,即目标函数 $f(x_k)$ 可以表示为:

$$f(x_k) = \min RMSECV \quad , \tag{2}$$

但该函数是一个求最小值的问题,同模拟退火算法操作相背,因此需把该函数变换成求最大值问题,即目标函数 $F(x)$ 可以表示为:

$$F(x) = \frac{1}{1 + f(x_k)} \quad , \tag{3}$$

式中 x_K 为优选出来的波长组合,$f(x_k)$ 为用 x_K 中波数点建立的 PLS 模型对应的 RMSECV 值.

(4) 冷却进度表的设计

冷却进度表的构造是基于算法的准平衡概念,其定义如下:设 L_k 是第 k 个马尔科夫链的长度,T_k 是相应的第 k 个温度控制参数值. 若第 k 个马尔科夫链的 L_k 次变换后,解的概率分布充分逼近 $T = T_k$ 时的平稳分布,则称模拟退火算法达到准平衡. 根据上面的准则,可以得到两个结论:(1) 只要 T 充分大,算法会立刻达到准平衡;(2) 控制参数 T_k 的衰减量越大,需要的马尔可夫链的长度 L_k 越长,才能恢复准平衡,通常选取 T_k 的小衰减量以避免过长的马尔科夫链. 同时有效的冷却进度表还要兼顾算法的收敛性和执行效率. 综合上面的结论以及参数设置的经验,本研究选中模拟退火算法使用如下冷却进度表:$T_0 = 200℃$;$T_k = 0.95T$;$T_f = 0℃$;$L_k = 50$.

设置好目标函数和冷却进度表后,执行模拟退火算法对近红外光谱优选波长. 本文采用的模拟退火算法具有记忆功能,即在退火过程中以一定的概率接受恶化解时,能够记住当前最优解,保证优化过程中最优解不会因为接受恶化解而退化,使算法具有一定的智能性.

2 材料与方法

2.1 试验材料

试验用草莓(奶油味)采于镇江长岗草莓园,共采摘了两批次. 第一批主要采摘完全成熟(成熟度依据草莓果的着色率判断)的草莓;第二批采摘八分熟左右的草莓. 试验当天,从草莓园采回试验所需草莓,带回实验室后挑选无碰伤、表面干净的草莓进行编号(共选出 300 个草莓,200 个为校正集,100 个为预测集),整个过程保持实验室温度为 25℃,湿度基本不变.

2.2 光谱数据采集

试验所用近红外数据采集设备为 Antaris Ⅱ 型傅立叶变换近红外光谱仪(Thermo Fisher,美国),光谱范围 10000 ~ 4000 cm^{-1},扫描次数 32 次,分辨率 4 cm^{-1},数据采样间隔为 1.928 cm^{-1}. 光谱数据采集前首先将草莓放在检测位置优化增益倍数,并采集背景光谱. 每个草莓采集 3 次光谱,采样点分布如图 1 所示. 图 1 中 a 为草莓样本图,b 为草莓赤道部位横截面图. b 中 1、2 和 3 为 3 个光谱采样点,它们彼此间夹角为 120°,将采集的 3 条光谱取平均值作为该样本的原始光谱. 由于仪器、样品背景、环境条件及其它因素的影响,近红外光谱中常出现噪声、谱图基线漂移和平移等现象,为了消除这些不利因素对建模的影响,经过多次测试与比较,采用标准正交变换(SNV)对原始光谱进行预处理,处理后的效果如图 2 所示.

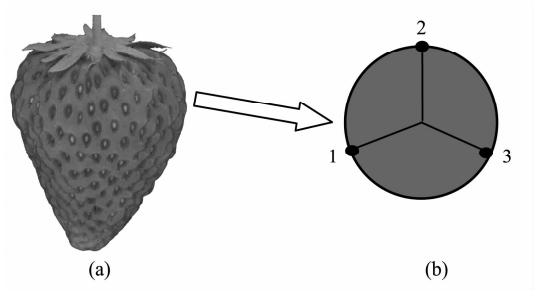

图 1 草莓近红外光谱采集位置示意图
Fig. 1 Sampling location of NIR

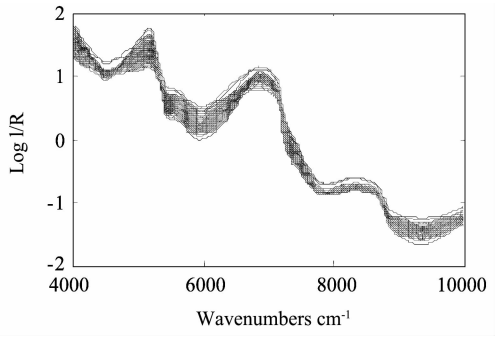

图 2 经过标准正交变换处理后的光谱
Fig. 2 Spectra after pre - processed by SNV

2.3 可溶性固形物含量测量

样本采集光谱后,在其对应 3 个部位分别用 WAY-2S 数字阿贝折射仪测量可溶性固形物含量值,对应平均值即为草莓样本的可溶性固形物含量值. 草莓样本的可溶性固形物含量值如表 1 所示.

2.4 数据处理软件

所用模拟退火算法由课题组在 Matlab7.4(The mathworks Inc., Natick, MA)平台下编程实现. 反向偏最小二乘法软件包由 Lars Norgaard 等人提供,通过 http://www.models.kvl.dk/免费获得.

表 1　草莓可溶性固形物含量值统计表
Table 1　Statistics of soluble solids content (SSC. °Brix) in strawberry samples

	样本数	平均值	最大值	最小值	标准偏差
校正集	200	7.12	10.5	3.9	1.3055
预测集	100	7.06	10.5	4.3	1.1898

3　试验结果与分析

3.1　BiPLS 选择特征子区间

草莓近红外光谱不仅包含了可溶性固形物信息,还包含了除可溶性固形物以外的组分信息. 同时,由于可溶性固形物在近红外光谱的多个波长处有吸收,且近红外光谱的谱峰较宽,致使近红外光谱在一个波长处有多个谱峰重叠. 在可溶性固形物特征子区间宽度不确定的情况下,为了使 BiPLS 准确定位包含可溶性固形物特征波长的子区间,需要对子区间划分总数进行优化.

将预处理后的光谱数据划分为 k 个子区间,k 的取值范围为 10 ~ 30. 当 k 取不同值时,采用 BiPLS 选择的特征子区间如表 2 所示. 从表 2 中可以看出,当光谱划分为 21(即 $k=21$)时,对应的交互验证均方根误差(RMESCV)最小,一共选出 4 个子区间(分别为第 8、13、16、17 区间),所选子区间包含 592 个波数点,对应的波数点范围为 6055 ~ 6289 cm^{-1}、7432 ~ 7716 cm^{-1}、8289 ~ 8572 cm^{-1}、8574 ~ 8857 cm^{-1}.

表 2　BiPLS 子区间优选结果
Table 2　Optimal spectra regions by BiPLS method

区间总数	入选区间数	RMSE	入选波数点数
10	5	0.4574	1555
11	4	0.4637	1132
12	5	0.4535	1296
13	7	0.4672	1674
14	7	0.4746	1555
15	8	0.4603	1659
16	8	0.4528	1555
17	7	0.447	1281
18	10	0.4573	1729
19	8	0.4531	1310
20	11	0.4668	1711
21	4	0.4031	592
22	8	0.4456	1130
23	12	0.4464	1621
24	7	0.4293	907
25	9	0.4664	1119
26	12	0.4421	1436
27	12	0.4462	1381
28	15	0.4438	1666
29	7	0.4523	751
30	11	0.4425	1141

3.2　SAA 选择特征波长

虽然根据 BiPLS 选择出来的 592 个波数点集合可以建立光谱模型,但是构建模型过程比较复杂. 以常规的间隔偏最小二乘法建模为例,首先对光谱矩阵 X 进行主成分降维处理(最大主成分数设置为 10),并根据 RMSECV 值确定主成分数为 8;然后根据光谱的前 8 个主成分和列向量 y 最终得到 PLS 校正模型. 用该 PLS 校正模型对未知光谱 x 进行预测时,首先需要提取光谱中的 592 个波数点对应的光谱值,其次,计算这 592 个光谱值对应的前 8 个主成分,最后将 8 个主成分代入 PLS 校正模型得到未知光谱对应的可溶性固形物含量. 从上述 PLS 校正模型建立和未知光谱预测过程可以看出,无论是校正模型建立和未知光谱预测,入选波数点太多均导致数据处理较复杂.

为了降低模型预测时的计算量,在 BiPLS 选出的 4 个子区间的基础上,进一步采用 SAA 进行特征波数点选择. 经过反复尝试,SAA 冷却进度表确定如下:$T_0=200℃$;$T_k=0.95T$;$T_f=0℃$;$L_k=50$. SAA 算法最终从 592 个波数点中优选出 7565 cm^{-1}、7706 cm^{-1}、8289 cm^{-1}、8489 cm^{-1}、8499 cm^{-1}、8724 cm^{-1}、8807 cm^{-1}共 7 个特征波数点.

3.3　MLR 建模

将 7565 cm^{-1}、7706 cm^{-1}、8289 cm^{-1}、8489 cm^{-1}、8499 cm^{-1}、8724 cm^{-1}、8807 cm^{-1}对应的光谱信号作为 MLR 的输入变量,对可溶性固形物含量进行 MLR 回归建模,该模型对应的回归方程见式(4),对应的校正均方根误差(RMSEC)为 0.403,预测集散点图如图 3 所示.

$$\begin{aligned} y = &-194.818X_{7565}+200.403X_{7706}-593.308X_{8829}\\ &+497.275X_{8489}+551.433X_{8499}-636.765X_{8724}\\ &+149.659X_{8807}-16.382 \end{aligned} \quad , \quad (4)$$

为了对建模效果进行比较,分别对全光谱和 BiPLS 优选出来的子区间进行 PLS 建模,结果如表 3 所示. 从表 3 中可以看出,BiPLS-SAA-MLR 建模波数点个数为 7 个,而 PLS 模型和 BiPLS 模型需要的变量数分别为 3112 个和 592 个;且 BiPLS-SAA-MLR 模型的精度也要优于其它两个模型.

表 3　不同建模方法效果比较
Table 3　Summary of model results after spectra being built by different methods

建模方法	变量数	r_c	*RMSEC*	r_p	*RMSEP*
PLS	3112	0.9026	0.544	0.9010	0.550
BiPLS	592	0.9445	0.416	0.9340	0.452
BiPLS-SAA-MLR	7	0.9478	0.403	0.9412	0.428

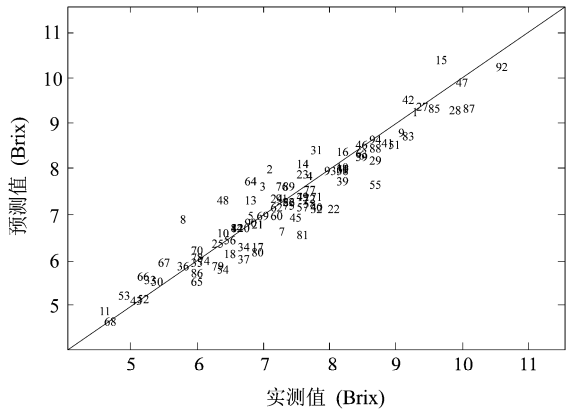

图3 预测集样本预测值和实测值之间的散点图
Fig. 3 Reference determination versus NIR prediction model

4 结语

借助先进的近红外光谱仪,研究者可以在短时间内快速获得海量光谱数据,如何对海量光谱数据进行有效筛选是国内外学者研究的重点.本研究根据样品通常在近红外光的某个或者某几个波段发生特征吸收,即光谱数据具有一定的连续性这一特点,采用特征子区间选择方法快速定位特征波长所在的子区间,然后采用特征波长选择方法得到草莓可溶性固形物对应的特征波长.

本研究利用 BiPLS 对草莓近红外光谱初步定位出4个特征子区间,然后用 SAA 优选 7565 cm^{-1}、7706 cm^{-1}、8289 cm^{-1}、8489 cm^{-1}、8499 cm^{-1}、8724 cm^{-1}、8807 cm^{-1}共7个可溶性固形物对应的特征波数点,并采用 MLR 回归建立了草莓可溶性固形物光谱模型. MLR 模型对应的 r_c 和 r_p 分别为0.9478和0.9412,对应的 RMSEC、RMSEP 分别为 0.403 和 0.428. 同全光谱 PLS 模型,BiPLS-SAA-MLR 极大的减少了建模所需的波数点数,降低了模型复杂度,同时模型精度明显提高,同 BiPLS 模型相比较,BiPLS-SAA-MLR 在降低模型复杂度方面仍具有明显优势. 研究结果充分表明,将 BiPLS、SAA 和 MLR 相结合对草莓可溶性固形物含量进行建模,简化了模型复杂度,提高模型的预测精度和计算效率,为近红外光谱技术的在线运用提供了一套建模解决方案.

REFERENCES

[1] Tsai C Y, Chen H J, Hsieh J F, *et al*. Fabrication of a near infrared online inspection system for pear fruit[J]. *International Agricultural Engineering Journal*, 2007, **16**: 57 - 70.

[2] Walsh K B, Long R L, Middleton S G. Use of near infrared spectroscopy in evaluation of source-sink manipulation to increase the soluble sugar content of stonefruit[J]. *Journal of Food Engineering*, 2007, **78**(2): 701 - 707.

[3] Camps C, Guillermin P, Mauget J C, *et al*. Discrimination of storage duration of apples stored in a cooled room and shelf-life by visible-near infrared spectroscopy[J]. *Journal of near Infrared Spectroscopy*, 2007, **15**(3): 169 - 177.

[4] Lu R, Ariana D. A near-infrared sensing technique for measuring internal quality of apple fruit[J]. *Applied Engineering in Agriculture*, 2002, **18**(5): 585 - 590.

[5] WANG Jia-Hua, LI Peng-Fei, CAO Nan-Ning, *et al*. Study on the combination weight PLS model for determing ssc of peach based on the optimal information regions obtained from iPLS methods[J]. *J. Infrared Millim. Waves* (王加华,李鹏飞,曹楠宁,等. 基于 iPLS 原理最优化信息区间的桃糖度组合权重 PLS 模型研究. **红外与毫米波学报**), 2009, **28**(5): 386 - 391.

[6] HE Yong, LI Xiao-Li. Discrimination varieties of waxberry using near infrared spectra[J]. *J. Infrared Millim. Waves* (何勇,李晓丽. 用近红外光谱鉴别杨梅品种的研究. **红外与毫米波学报**), 2006, **25**(3): 192 - 194, 212.

[7] ZOU Xiao-Bo, ZHAO Jie-Wen, Li Yan-Xiao. Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of 'Fuji' apple based on BiPLS and FiPLS models[J]. *Vibration Spectroscopy*. 2007, **44**: 220 - 227.

[8] Guyer D, Yang X K. Use of genetic artificial neural networks and spectral imaging for defect detection on cherries [J]. *Computers and Electronics in Agriculture*, 2000, **29**(3): 179 - 194.

[9] Kirkpatrick S, Gelatt C D Jr, Vecchi M P. Optimization by simulated annealing[J]. *Science*. 1983, **220**: 671 - 680.

(上接405页)

[9] Shen Xue-Chu. Spectroscopy and Optical Properties of Semiconductors[M]. *Science Press* (沈学础. 半导体光谱和光学性质. **科学出版社**), 2002.

[10] Guo H C, Liu W M, Tang S H. Terahertz time-domain studies of far-infrared dielectric response in 5 mol % MgO: $LiNbO_3$ Ferroelectric single crystal[J]. *Journal of Applied Physics*, 2007, **102**: 033105(1-4).

[11] Thamizhmani L, Azad A K, Dai Jianming. Far-infrared optical and dielectric response of ZnS measured by terahertz time-domain spectroscopy[J]. *Applied Physics Letters*, 2005, **86**: 131111(1-3).

[12] Xu Jingzhou, Zhang Xicheng. Terahertz technology and application[M]. *Peking University press* (许景周,张希成. 太赫兹科学技术与应用. **北京大学出版社**), 2007.

[13] Li Biao. LPE Growth and Characterization of $Hg_{1-x}Cd_xTe$ Films[D]. PhD. thesis, Shanghai Institute of Technical Physics, Chinese Academy of Sciences. (李标. $Hg_{1-x}Cd_xTe$ 薄膜的 LPE 生长与特性分析. **中科院上海技术物理所博士论文**), 1995.