

文章编号: 1672-8785(2011)12-0039-06

基于可见 – 近红外光谱技术的润滑油 酸值无损检测方法研究

张 瑜¹ 吴 迪² 何 勇² 谈黎虹¹ 蒋璐璐¹

(1. 浙江经济职业技术学院, 浙江杭州 310018;

2. 浙江大学生物系统工程与食品科学学院, 浙江杭州 310058)

摘要: 研究了基于可见 – 近红外光谱技术的润滑油酸值无损检测方法。获得了润滑油在 475 ~ 975 nm 范围内的可见 – 近红外光谱。采用最小二乘支持向量机 (LS-SVM) 建立可见 – 近红外光谱检测模型, 并通过将无信息变量消除算法 (UVE) 与连续投影算法 (SPA) 相结合进行光谱有效波长选取。通过 UVE-SPA 法进行变量选择计算, 最终将原始光谱所包含的 500 个光谱变量减少到了 8 个 (分别为 489 nm、553 nm、591 nm、874 nm、893 nm、910 nm、935 nm 和 951 nm)。基于这 8 个变量建立的 LS-SVM 模型获得了预测集确定系数为 0.9546、误差均方根为 0.0081 和剩余预测残差为 4.5663 的预测结果, 说明可见 – 近红外光谱技术可以用于润滑油酸值无损检测。与酸值测定标准方法相比, 该方法具有快速、无损和成本低等优点。同时, UVE-SPA 法是一种有效的光谱变量选择方法。

关键词: 可见 – 近红外光谱; 润滑油; 酸值; 最小二乘支持向量机; 无信息变量消除 – 连续投影算法 (UVE-SPA)

中图分类号: TP391 ; TK418 **文献标识码:** A **DOI:** 10.3969/j.issn.1672-8785.2011.12.008

Non-invasive Measurement of Acid Value of Lubricant Using Visible and Near Infrared Spectroscopy

ZHANG Yu¹, WU Di², HE Yong², TAN Li-hong¹, JIANG Lu-lu¹

(1. Zhejiang Technology Institute of Economy, Hangzhou 310018, China;

2. College of Biosystems Engineering & Food Science, Zhejiang University, Hangzhou 310058, China)

Abstract: The non-invasive lubricant acid value measurement method based on a visible and near-infrared spectroscopy is studied. The visible and near-infrared spectra of the lubricant in the region from 475 μm to 975 μm are obtained. A visible and near-infrared spectral detection model is established by using a least square support vector machine. By combining the uninformative variable elimination with the successive projection algorithm, effective spectral wavelengths are selected. Among the 500 variables, only eight wavelength variables namely 498 nm, 553 nm, 591 nm, 874 nm, 893 nm, 910 nm, 935 nm and 951 nm are selected. The least square support vector machine model based on those eight wavelength variables has obtained a prediction set of 0.9546, a root mean square error of 0.0081 and a residual predi-

收稿日期: 2011-11-02**基金项目:** 十一五国家科技支撑项目 (2006BAD10A0403); 浙江省教育厅科技项目 (20071275)**作者简介:** 张瑜 (1977-), 女, 浙江丽水人, 讲师, 主要从事光谱检测与多光谱成像技术及应用研究。

E-mail: zy7739@126.com

ctive deviation of 4.5663. This prediction result shows that the visible and near-infrared spectroscopy can be used to measure the acid value of lubricant non-invasively. Compared with the standard acid value measurement method, it has the advantages of fastness, non-invasion and low cost. Meanwhile, it is also an effective algorithm for spectral variable selection.

Key words: visible and near infrared spectroscopy; lubricant; acid value; least-square support vector machine; uninformative variable elimination; successive projection algorithm

0 引言

发动机润滑油的酸值是一个重要的品质指标，它不仅可以用于衡量润滑油是否符合规定标准，而且还可以用于反映润滑油在储存与使用过程中的氧化变质情况。目前，国内的酸值测定标准方法主要采用指示剂法和电位滴定法。但是上述方法具有耗时费力，检测过程需要破坏样本，检测成本高以及不能实时反映润滑油酸值等缺点。因此，人们需要一种可以快速、无损地检测润滑油掺水量的方法。

可见-近红外光谱法是一种简便、快速且成本低的非破坏性分析技术，已被广泛应用于食品、医药以及石油化工等行业^[1-4]。在石油化工领域中，该方法已用于汽油和柴油的测试及其成分的定量分析与研究^[5]。但是目前还没有关于润滑油酸值测定的相关报道。近红外光谱主要是倍频和合频的吸收，其光谱信息重叠严重。因此，可见-近红外光谱技术必须从海量光谱信息中剔除冗余信息，建立定量数学模型。在进行建模分析的同时，由于全波段光谱包含数百个光谱变量，建立模型的效率会受到影响，因此通过减少模型的输入变量个数，可以提高模型的校正速度和建模效率。

连续投影算法 (Successive Projections Algorithm, SPA) 是一种新的变量提取方法^[6-7]。它能够利用向量的投影分析，寻找含有最低限度的冗余信息的变量组，并使变量之间的共线性达到最小和建模所用变量的个数大大减少，从而提高建模的速度和效率。然而对全波段进行 SPA 计算，将会增加计算量。因此，在进行 SPA 计算之前最好对全波段变量剔出无信息变量，以提高信噪比并减少 SPA 法的计算量。1996 年，Centner

等人提出了一种基于 PLS 的无信息变量消除算法 (Uninformative Variable Elimination, UVE)^[8]。这是一种基于 PLS 回归系数分析建立的波长选择算法，它可以消除不提供信息的变量，减少模型中包含的变量数，降低模型的复杂性。

本文采用可见-近红外光谱技术，通过无信息变量消除算法进行光谱信息筛选。然后进一步采用 SPA 法提取有效波长作为最小二乘支持向量机 (Least Square-support Vector Machine, LS-SVM)^[9] 的输入，并建立光谱检测模型，从而实现润滑油酸值的快速、无损检测。

1 材料与方法

1.1 样品准备和酸值测定

实验选用 5 个常用的发动机润滑油品牌，它们分别为加德士、美孚、壳牌喜力、长城和广本专用润滑油。每个品牌的润滑油都被密封保存在 25 ℃ 的实验室内。用每个品牌分别制备 30 个样本，共计 150 个样本。从每个品牌样本中选择 20 个样本 (共 100 个) 作为建模集，并将剩余的每个品牌的 10 个样本 (共 50 个样本) 作为预测集。润滑油的酸值采用《石油产品酸值测定法 GB264-83》国家标准测定。

1.2 样品准备和光谱采集

实验使用的光谱测定范围为 325 ~ 1075 nm (采用美国 ASD 公司生产的 Handheld FieldSpec 光谱仪)。光源采用与光谱仪配套的 14.5 V 卤素灯。光谱扫描采用漫反射法进行光谱采样，光程为 2 mm。每个样本扫描 30 次，保存 3 条光谱，并以其平均光谱作为最终光谱。光谱分析采用 Unscramble V9.7 和 MATLAB 7.6 软件进行。

1.3 光谱数据预处理与化学计量学分析

1.3.1 光谱数据预处理

由于光谱数据在采集范围的首端与末端存在噪音, 选取 476~975 nm 波段内共 500 个变量的光谱数据进行分析。在将光谱数据转换成吸光度值之后, 将根据原始光谱与采用不同窗口大小的 Savitzky-Golay 平滑法预处理后的光谱所建的模型的效果进行比较。

1.3.2 无信息变量消除算法

下面介绍无信息变量消除算法的实现步骤。在 PLS 模型中, 光谱矩阵 X 与浓度矩阵 Y 之间存在以下关系:

$$Y = Xb + e \quad (1)$$

式中, b 为系数向量, e 为误差向量。无信息变量消除算法就是把具有一定变量数目的随机变量矩阵加入光谱矩阵中, 然后通过交叉验证建立 PLS 模型, 得到系数矩阵 B 。分析系数向量 b 的平均值和标准偏差的商 C 的稳定性为

$$C_i = \frac{\text{mean}(b_i)}{S(b_i)} \quad (2)$$

式中, $\text{mean}(b)$ 为系数向量 b 的平均值; $S(b)$ 为系数向量 b 的标准差; i 表示光谱矩阵中的第 i 列向量。根据 C_i 的绝对值大小确定是否要把第 i 列变量用于最后的 PLS 模型中。具体算法见参考文献 [8]。

1.3.3 连续投影算法

下面介绍连续投影算法的实现步骤^[10]。记 $x_k(0)$ 为初始迭代向量, N 为需要提取的变量个数。光谱矩阵为 j 列:

迭代开始前, 任选光谱矩阵的 1 列 j , 把建模集的第 j 列赋值给 x_j , 记为 $x_k(0)$;

把未选入的列向量位置的集合记为 s , $s = \{j, 1 \leq j \leq J, j \notin \{k(0), \dots, k(n-1)\}\}$;

分别计算 x_j 对剩下列向量的投影: $Px_j = x_j - (x_j^T x_{k(n-1)}) x_{k(n-1)} (x_{k(n-1)}^T x_{k(n-1)})^{-1}, j \in s$;

记 $k(n) = \arg(\max(\|Px_j\|)), j \in s$;

令 $x_j = Px_j, j \in s$;

$n = n + 1$ (若 $n < N$, 则回到 b 循环计算);

最后, 提取出的变量为 $\{x_{k(n)} = 0, \dots, N-1\}$ 。对应于每一个 $k(0)$ 和 N , 循环一次后进行多元线性回归分析 (MLR), 得到验证集的验证标准偏差 (RMSEV)。最小的 RMSEV 值所对应

的 $k(0)$ 和 N 就是最优值。

1.3.4 光谱建模方法与模型评价指标

最小二乘支持向量机 (LS-SVM) 是一种新型建模方法。该方法利用拉格朗日乘子对各个变量求偏微分, 得到最优解。在 LS-SVM 计算中, 采用非线性映射函数, 将输入变量映射到高维特征空间中, 从而将优化问题转化成等式约束条件。用求解一组线性方程代替经典 SVM 中复杂的二次优化问题, 加快了计算速度, 降低了计算的复杂性。该方法是用于解决“过学习”和“维数灾”的有效工具之一。LS-SVM 算法中的结构风险最小化特征能够提高其泛化能力, 从而可较好地解决小样本、非线性、高维数以及局部极小等问题。

本文中的光谱模型采用以下参数进行评估: 建模集和预测集的确定系数 (r_c^2 和 r_p^2)、剩余预测残差 (Residual Predictive Deviation, RPD)、建模均方根误差 (Root Mean Square Error for Calibration, RMSEC) 以及预测均方根误差 (Root Mean Square Error for Prediction, RMSEP)。效果好的模型需要有较低的均方根误差和较高的 r^2 与 RPD 值。一般认为, 当 RPD 值大于 3 时, 模型适合于实际应用^[11]。

2 试验结果与分析

2.1 预处理方法的选择

采用原始光谱和不同窗口大小的 Savitzky-Golay 平滑法预处理方法对全波段光谱进行处理, 并建立 LS-SVM 模型。当不对光谱进行预处理即采用原始光谱建模时, LS-SVM 模型的建模集和预测集的确定系数都达到了 0.94 以上(见表 1), RPD 值也在 4 以上, 说明模型获得了较好的结果。当采用 Savitzky-Golay 平滑法进行预处理时, 处理窗口大小从 6、11、16 至 76 不等。图 1 所示为不同模型的 RPD 值分布。从图 1 中可以看出, 当平滑窗口为 41 时, 模型的效果最好, 其 RPD 值为 4.75(见表 1)。用 Savitzky-Golay 平滑法处理后建立的光谱模型的效果比没有进行光谱预处理的模型的好。此外, 在用 Savitzky-Golay 平滑法处理后, 所建光谱模型的建模集和预测集的确定系数之间以及均方根误差之间的差值

表1 基于平滑窗口为41时 Savitzky-Golay 平滑法处理和不经过光谱
预处理建立的 LS-SVM 模型的效果比较

预处理	r_c^2	RMSEC	r_p^2	RMSEP	RPD
无	0.9671	0.0092	0.9443	0.0085	4.1646
Savitzky-Golay 平滑法	0.9594	0.0100	0.9574	0.0078	4.7494

注: r_c^2 为建模集的确定系数, RMSEC 为建模均方根误差, r_p^2 为预测集的确定系数, RMSEP 为预测均方根误差, RPD 为剩余预测残差。

都很小, 说明模型没有发生过拟合现象, 而且模型的鲁棒性较好。因此, 在后面的分析中均采用平滑窗口为41时经 Savitzky-Golay 平滑法处理后的光谱。

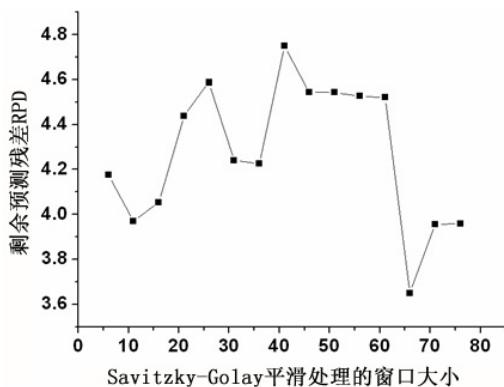


图1 基于不同窗口大小的 Savitzky-Golay 平滑处理后建立的 LS-SVM 模型的 RPD 值分布

2.2 基于全波段光谱进行 SPA 处理

首先采用 SPA 法对全波段光谱(共500个变量)进行变量选择。图2所示为基于选择得到的不同变量个数的验证标准偏差(RMSEV)值的变化情况。从图2中可以看出, 变量个数从1个到4个、8个到12个和16个到17个时的 RMSEV

值下降较快。当变量个数为17时, RMSEV 值为0.0238, 并且此后基本上不再减小。用 SPA 法对全波段光谱进行处理, 将光谱变量从600个减少到17个, 简化了 LS-SVM 模型。将得到的17个变量作为输入, 建立 LS-SVM 模型(结果见表2)。从预测结果中可以看出, 采用由 SPA 法选出的17个变量建立的 LS-SVM 模型的结果与用全波段变量建立的 LS-SVM 模型相比并没有明显下降, 说明 SPA 法能够从冗余的光谱数据中选择出少量最利于建模的光谱变量。但是, 对500个变量进行 SPA 处理的计算量较大。对此本文进一步提出在 SPA 计算之前对全波段变量进行无信息变量剔出, 以简化 SPA 处理。

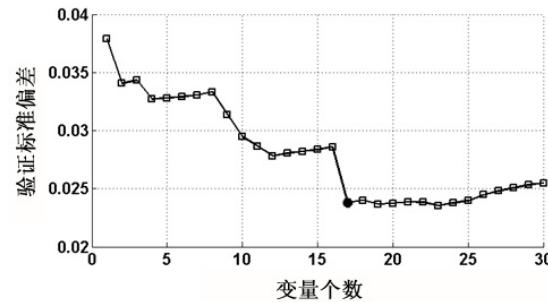


图2 用连续投影算法选取的波长数所对应的验证集预测标准偏差

表2 基于用不同变量选择方法得到的光谱变量建立的 LS-SVM 模型的建模和预测结果

变量选择方法	变量个数	r_c^2	RMSECV	r_p^2	RMSEP	RPD
无	500	0.9594	0.0100	0.9574	0.0078	4.7494
SPA	17	0.9618	0.0097	0.9598	0.0076	4.8902
UVE	197	0.9619	0.0097	0.9624	0.0075	4.9763
UVE+SPA	8	0.9597	0.0100	0.9546	0.0081	4.5663

注: r_c^2 为建模集的确定系数, RMSEC 为建模均方根误差, r_p^2 为预测集的确定系数, RMSEP 为预测均方根误差, RPD 为剩余预测残差。

2.3 基于 UVE 计算的全波段光谱无信息变量去除

采用 UVE 法对全波段中的 500 个光谱变量进行选择。将 UVE 中产生的随机变量个数同样设置为 500 个。图 3 所示为 UVE 变量的选择结果。其中, 黑色虚线的左侧为 500 个变量的稳定性分布曲线, 右侧为 UVE 中产生的 500 个随机变量的稳定性分布曲线。两条黑色水平线表示变量选择的阈值上下限。线外数值所对应的变量被保留, 线内数值所对应的变量不用于建模。阈值的选择标准为随机变量稳定性最大值的 99 %。通过 UVE 的变量选择, 最后得到 197 个波长。UVE 选择的波长在可见光和近红外区域内均有分布, 说明在检测润滑油酸值时需要考虑可见光光谱和近红外光谱。将得到的 197 个波长作为 LS-SVM 模型的输入变量(预测结果见表 2)。可以看出, 在去除了无信息变量信息以后, 模型的结果与全波段光谱模型相比有所改善。其中, 预测集的确定系数提高了 0.5 %, 预测均方根误差下降了 3.8 %, 剩余预测残差提高了 4.8 %。值得注意的是, 模型的输入变量数减少了 60 %。这将有助于在继续采用 SPA 计算时减少计算量。

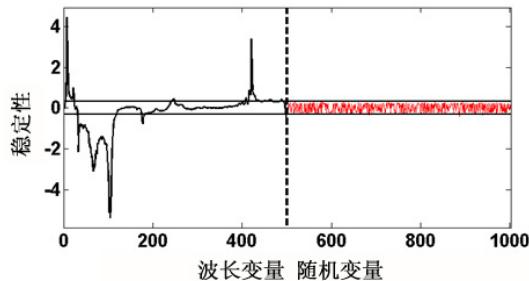


图 3 UVE 法的稳定性分布曲线。两条黑色水平线表示变量选择的阈值上下限

2.4 基于 UVE-SPA 选取有效波长

通过 UVE 选择将变量个数从全波段时的 600 个减少到 197 个。进一步采用 SPA 法进行变量选择, 得到共线性最小的 8 个有效波长(分别为 489 nm、553 nm、591 nm、874 nm、893 nm、910 nm、935 nm 和 951 nm)。图 4 所示为利用 UVE-SPA 得到的波长分布情况。其中, 黑色竖

线表示选中的波长。选中的波长在可见光光谱和近红外光谱范围内均有分布。将这 8 个波长作为输入变量建立润滑油酸值的 UVE-SPA-LS-SVM 模型(结果见表 2)。和仅采用 SPA 计算相比, UVE-SPA 的变量个数减少了 53 %, 但其预测集的确定系数仅下降了 0.5 %, 预测均方根误差增加了 6.2 %, 剩余预测残差下降了 6.6 %, 而且基于 UVE 得到的光谱变量进行 SPA 计算的工作量下降了 60 % (197 vs. 500)。因此, UVE-SPA 的计算结果是令人满意的。图 5 为利用 UVE-SPA 所得 8 个波长建立的 LS-SVM 模型的样本散点分布图。从图 5 中可以看出, 主要的样本点都基本分布在回归线附近, 实现了良好的建模和预测结果。

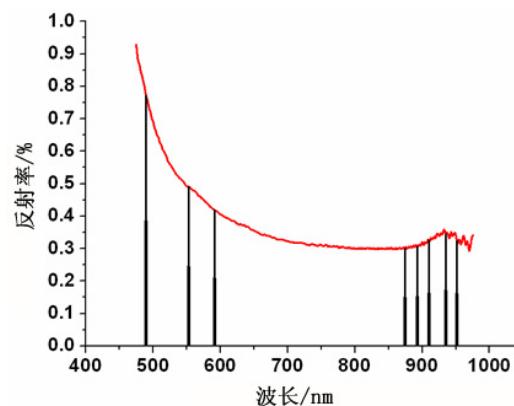


图 4 利用 UVE-SPA 选择的 8 个波长的分布图。其中, 黑色竖线表示选中的波长

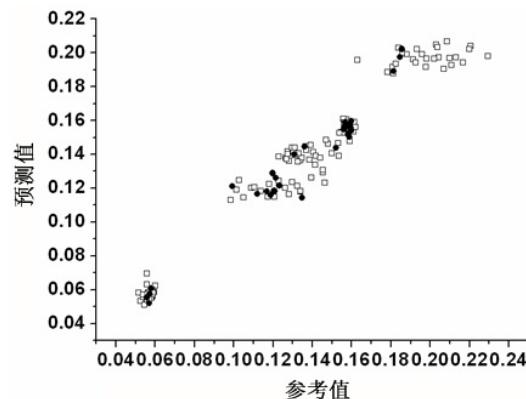


图 5 利用 UVE-SPA 选择的 8 个波长建立的 LS-SVM 模型对酸值的预测散点分布图。其中, □ 表示建模集样本, ● 表示预测集样本

3 结束语

本文利用可见 – 近红外光谱技术对发动机润滑油酸值的无损检测方法进行了研究。本文的具体结论包括：(1) 可见 – 近红外光谱技术可以用于对润滑油酸值进行无损、快速检测(与酸值测定标准方法相比, 可见 – 近红外光谱检测具有快速、无损, 操作程序简单以及检测成本低等特点); (2) 当平滑窗口为 41 时, Savitzky-Golay 平滑法的效果最好; (3) 通过在全波段光谱上进行 SPA 计算, 可以从冗余的光谱数据中选出少量最利于建模的光谱变量, 但其计算量较大; (4) 当采用 UVE 法对全波段光谱进行无信息变量去除时, 模型的结果与全波段光谱模型相比有所改善; (5) 采用 UVE–SPA 建模得到的结果与仅采用 SPA 所建模型的结果相比略有下降, 但其计算量仅有原来的 40 %。

参考文献

- [1] 严衍禄, 赵龙莲, 韩东海, 等. **近红外光谱分析基础与应用** [M]. 北京: 中国轻工业出版社, 2005.
- [2] Wu D, He Y, Shi J H, et al. Exploring Near and Mid-infrared Spectroscopy to Predict Trace Iron and Zinc Contents in Powdered Milk [J]. *Journal of Agricultural and Food Chemistry*, 2009, **57**(5): 1697–1704.
- [3] Wu D, Chen X J, Shi P Y, et al. Determination of Alpha-linolenic Acid and Linoleic Acid in Edible Oils Using Near-infrared Spectroscopy Improved by Wavelet Transform and Uninformative Variable Elimination [J]. *Analytica Chimica Acta*, 2009, **634**(2): 166–171.
- [4] Wu D, He Y, Feng S. Short-wave near-infrared Spectroscopy Analysis of Major Compounds in Milk Powder and Wavelength Assignment [J]. *Analytica Chimica Acta*, 2008, **610**(2): 232–242.
- [5] 周鑫. 在用车发动机油快速检测质量的研究 [J]. *汽车工艺与材料*, 2005, **20**(7): 39–45.
- [6] Araújo M C U, Saldanha T C B, Galvão R K H, et al. The Successive Projections Algorithm for Variable Selection in Spectroscopic Multicomponent Analysis [J]. *Chemometrics and Intelligent Laboratory Systems*, 2001, **57**(2): 65–73.
- [7] Galvão R K H, Araújo M C U, Fragoso W D, et al. A Variable Elimination Method to Improve The Parsimony of MLR Models Using The Successive Projections Algorithm [J]. *Chemometrics and Intelligent Laboratory Systems*, 2008, **92**(4): 83–91.
- [8] Centner V, Massart D L, Noord O E, et al. Elimination of Uninformative Variables for Multivariate Calibration [J]. *Analytical Chemistry*, 1996, **68**(21): 3851–3858.
- [9] Wu D, He Y, Feng S J, et al. Study on Infrared Spectroscopy Technique for Fast Measurement of Protein Content in Milk Powder Based on LS-SVM [J]. *Journal of Food Engineering*, 2008, **84**(1): 124–131.
- [10] 陈斌, 王忠. 用近红外透射光谱快速检测内燃机润滑油性能 [J]. *农业机械学报*, 2002, **33**(5): 17–19.
- [11] Wu D, He Y, Nie P C, et al. Hybrid Variable Selection in Visible and Near-infrared Spectral Analysis for Non-invasive Quality Determination of Grape Juice [J]. *Analytica Chimica Acta*, 2010, **659**(1–2): 229–237.

(上接第 30 页)

- [2] Kraus D, Marhefka J. Antennas: For All Applications [M]. Beijing: Publishing House of Electronics Industry, 2008.
- [3] Kang H-C, Nishimura T, Komori T, et al. 3D-integration of A Log Spiral Antenna onto A Dual Grating-gate Plasmon-resonant Terahertz Emitter for High-directivity Radiation [J]. *Journal of Physics: Conference Series*, 2008, **193**(1): 012070.
- [4] Yuan Y, Si L-M, Liu Y. Integrated Log-periodic Antenna for Terahertz Applications [C]. International

Conference on Microwave Technology and Computational Electromagnetics, 2009: 276–279.

- [5] Pastol Y, Arjavalangam G, Halbout J-M. Characterisation of An Optoelectronically Pulsed Equiangular Spiral Antenna [J]. *Electronics Letters*, 1990, **26**(2): 133–135.
- [6] Saijo H, Morimoto M, Yamashita M, et al. Terahertz Radiation from Log-periodic Antennas [C]. Technical Digest. IEEE, 1999: 451–452.
- [7] Dragoman D, Dragoman M. Terahertz Fields and Applications [J]. *Progress in Quantum Electron*, 2004, **28** (1): 1–66.