太赫兹光谱结合特征谱区筛选算法在发动机润滑油含水量定量分析中应用研究

陈孟秋[1]，何明霞[1*]，李 萌[2]，曲秋红[2]

1. 天津大学测试计量技术及仪器国家重点实验室，天津 300072
2. 莱仪特太赫兹(天津)科技有限公司，天津 300019

摘 要 发动机润滑油是保障汽车发动机持久且稳定运转的基石，准确评定发动机润滑油各项性能指标是其在生产到使用全过程必不可少的步骤。发动机润滑油在一段时间的使用后会因为多种原因引起油品变质，发动机润滑油变质的指标可以用其中非磁性颗粒物浓度、金属屑含量、pH值、粘稠度、含水率等表述。关于发动机润滑油含水量的检测，传统的检测方法存在操作复杂，及时性差等缺点。太赫兹对水吸收强烈，适合用于对样品中微水含量的分析。通过透射式太赫兹时域光谱系统获得1.0～3.5 THz下的六种不同水含量的发动机润滑油的吸收系数谱线，对谱线进行Savitzky-Golay(SG)平滑去噪，剔除奇异样本后，采用Kennard-Stone算法划分样品集，尝试常规区间偏最小二乘法(iPLS)、向后区间偏最小二乘法(BiPLS)和联合区间偏最小二乘法(SiPLS)对其太赫兹时域光谱特征谱区间进行筛选，着重研究区间间隔数、PLS组件数、最佳主因子数和区间选择等因素对PLS模型属性的影响，并且对不同含水量的润滑油建模分析，对不同模型比较选优，建立最优定量分析模型。建模结果表示特征谱区筛选可以提高建模性能、降低模型复杂性，特征谱区筛选算法通过剔除发动机润滑油太赫兹吸收系数谱线中非线性或者无关变量的方式，使建模结果更好的表达吸收系数谱线与其含水量的关系。结果表明：采用BiPLS模型用于发电机润滑油中微量水含量的定量分析时建模效果最佳，模型区间数为26，入选区间为[18 10 4 3 8 12 5 11 24 13 16 21 2]，主因子数为10，最优模型的交互验证均方根误差RMSECV为0.003 5，预测均方根误差RMSEP为0.004 6，校正集相关系数r为0.919 3，预测集相关系数r为0.865 7。由此可见，可以采用反向区间偏最小二乘法(BiPLS)用于发动机润滑油水含量的测定，且实验过程简单，建模计算速度快，效果理想，可以适用于非接触式油品含水量的定量分析。

关键词 太赫兹时域光谱技术；特征谱区筛选算法；发动机润滑油；水含量检测
中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)05-1393-05

引 言

发动机润滑油，主要成分是碳氢化合物，主要功能为对发动机起到润滑防磨、冷却降温、减震缓冲、防锈蚀防漏等作用[1]。新出厂的发动机润滑油是不含水分的，但在存储、运输和使用的过程中，可能会因为各种原因混入水分。按照国家标准，发动机润滑油中允许的含水量应在0.03%以下，若含水量超过标准，润滑油中会产生酸类物质，这些酸类物质会增加对发动机的腐蚀，引起发动机抱轴、烧瓦等严重事故。

针对发动机润滑油中水含量的检测现在常用方法有重量法、红外光谱分析法、蒸馏法、卡尔·费休法等。这些方法均已有成熟的测试步骤，但仍存在各自的不足，如：当样品中水分含量高时采用重量法会在烘干过程中发生飞溅，影响测量精度；红外光谱分析法会受到基础油类别、润滑油劣化程度等因素影响[2]；蒸馏法则需要的样品量较多，耗时较长；卡尔·费休法虽然应用最广泛，但是这种方法副反应较多，且测量使用的化学试剂具有毒性[3]。

太赫兹(Terahertz，THz)波是指波长在0.03～3 mm之间，频率在0.1～10 THz，介于红外和微波之间的电磁波[4]。

收稿日期：2020-04-24，**修订日期**：2020-07-19
基金项目：国家自然科学基金项目(61675151)资助
作者简介：陈孟秋，1966年生，天津大学精密仪器与光电子工程学院硕士研究生 e-mail：cmq1996@126.com
*通讯作者 e-mail：hhmmxx@tju.edu.cn

水在太赫兹频段拥有独特的分子键振动模式，使得水对太赫兹具有强烈的吸收性[5-6]。太赫兹光谱技术已被用于测量变压器油、原油、生物组织及细胞中的微水含量[7]。

本文利用太赫兹时域光谱技术对发动机润滑油中水含量进行检测并结合特征谱区筛选算法进行定量分析，对润滑油中水含量这一指标进行建模分析，对不同模型比较选优，建立最优定量分析模型。以期寻找一种检定润滑油含水量的新方法。

1 实验部分

1.1 方法

实验使用的是日本 advantest 公司的 TAS7400SU 太赫兹光谱系统。光谱范围为 0.5～7.0 THz，频率精度±10 GHz，动态范围为 57 dB，频率分辨率为 7.6 GHz。该系统由三个主要部分组成，分别是飞秒激光器，太赫兹发射天线和接收天线。本实验中用的是其透射模块，其结构如图 1 所示。

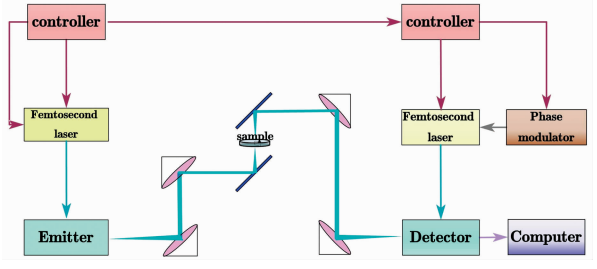

图 1 太赫兹时域光谱系统

Fig. 1 Schematic of THz-TDS

实验选用汽车发动机同型号不同老化程度的润滑油，利用卡尔·费休水分测定仪对其含水量进行测量，卡尔·费休法是利用了样品中的水与卡尔费休试剂中 SO_2 和 I_2 产生的氧化还原反应对其进行水含量的测量，每种润滑油分别测量 3 次，取平均值。其含水量分别为 0.039 2%，0.029 2%，0.026 1%，0.017 4%，0.015 8%和 0.013 3%，液体样品池采用光程为 10 mm 的 JGS1 级石英比色皿，样品需要干燥密封保存。

在实验中，以干燥空气作为背景信号，相同含水量的润滑油样品各准备 6 个样本，每个样本移动不同位置分别测量 5 次。得到每种润滑油各采集 30 组光谱数据，总共 180 组光谱数据。

为了降低系统及实验因素导致的干扰和噪声，使用 Savitzky-Golay(S-G)平滑预处理，考虑原光谱的特性，将平滑滤波器的拟合阶数设置为 3 阶，设置每 15 个点平滑一次。样品集的划分采用 Kennard-Stone(KS)算法，将所有样本均视为训练集候选样本，依次从中挑选样本进入训练集。通过 KS 算法，将样品中 150 组数据设为校正集，30 组数据设为预测集。

1.2 特征谱区筛选算法

常规区间最小二乘(iPLS)是一种较为常用的优选特征光谱区间的化学计量方法，由 Nørgaard 等提出。将数据集划分为 n 个子区间，分别建立每个子区间的 PLS 模型，取子区间交互验证均方根误差(RMSECV)最小时的因子数为最优因子数，以建立各个子区间的最优模型。向后区间偏最小二乘法(BiPLS)是每次排除根据 RMSECV 数值显示建模效果最差的子区间，使得在($n-1$)个子区间内建模，取 RMSECV 最小的区间组合为最优建模区间。联合区间偏最小二乘法(SiPLS)则是根据指定的组合区间个数将各个子区间随机组合，对每种组合的区间建立 PLS 模型，取 RMSECV 最小的区间组合为最优建模区间。

2 结果与讨论

2.1 吸收系数谱

经平滑处理后得到的 THz 吸收系数误差棒谱线如图 2 所示，光谱范围取 1.0～3.5 THz，频率间隔 7.6 GHz，每条谱线包含 328 个变量。从图中可以看出吸收系数谱线随含水量增加而升高，当频率大于 3.5 THz 时，由于受系统功率影响，出现了明显噪声，因此为了保证数据的可靠性，采用 1.0～3.5 THz 的数据作为定量分析的对象。

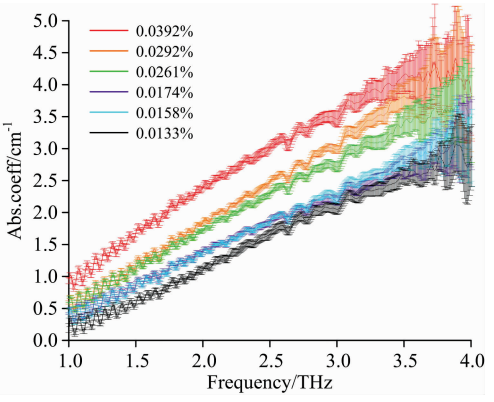

图 2 不同含水量润滑油的吸收系数误差棒谱线

Fig. 2 Dielectric constant spectra of lubricants with different water contents

2.2 水含量定量分析模型

2.2.1 iPLS 模型

将预处理过的 1.0～3.5 THz 范围的光谱区域划分为 10～30 个子区间，分别建立 iPLS 特征光谱区间筛选模型，比较不同模型的交互验证均方根误差(RMSECV)。选取所建立的回归模型中 RMSECV 最小时的子区间划分数、入选区间及主因子数建立润滑油水含量的定量分析模型，并以独立的预测集进行验证，比较预测模型的预测均方根误差(RMSEP)。

由表 1 中各模型的 RMSECV 值可知，在对应的 iPLS 谱区筛选模型的 21 个区间间隔划分模型中，当整个区间光谱间隔数为 10 个子区间，选择第 2 个子区间，对应 1.258 85～1.502 99 THz，主因子数为 8 时建模结果最佳。iPLS 最优模型的 RMSECV=0.004 8，RMSEP=0.006 0，校正集相关系数 R_c 为 0.848 2，预测集相关系数 R_p 为 0.761 8，对应频率范围为 1.258 85～1.502 99 THz。

表 1　不同区间划分数量时 iPLS 建模模型

Table 1　Results of iPLS model with different number of interval divisions

区间总数	入选区间	主因子数	r	RMSECV/%
10	**2**	**8**	**0.848 2**	**0.004 8**
11	2	8	0.829 2	0.005 0
12	2	8	0.805 6	0.005 3
13	5	6	0.819 6	0.005 2
14	5	3	0.801 7	0.005 4
15	3	6	0.805 6	0.005 3
16	3	5	0.790 0	0.005 5
17	6	5	0.805 5	0.005 3
18	3	10	0.807 3	0.005 3
19	3	10	0.804 1	0.005 4
20	7	5	0.801 1	0.005 4
21	4	7	0.800 9	0.005 4
22	3	6	0.786 9	0.005 6
23	8	6	0.786 8	0.005 6
24	2	7	0.770 8	0.005 7
25	2	7	0.770 8	0.005 7
26	9	6	0.780 0	0.005 6
27	9	5	0.776 2	0.005 7
28	5	5	0.780 3	0.005 6
29	5	5	0.780 3	0.005 6
30	4	5	0.780 5	0.005 6

2.2.2　BiPLS 模型

将预处理过的 1.0～3.5 THz 范围的光谱区域划分为 10～30 个子区间，分别建立 BiPLS 特征光谱区间筛选模型，以优选的光谱区间建立水含量定量分析模型并进行预测。

由表 2 可见，在对应的 BiPLS 谱区筛选模型的 21 个区间间隔划分模型中，当整个区间光谱间隔数为 26 个子区间，选择[18 10 4 3 8 12 5 11 24 13 16 21 2]子区间组合，主因子数为 10 时建模结果最佳。BiPLS 最优模型的 RMSECV＝0.003 5，RMSEP＝0.0046，R_c＝0.919 3，R_p＝0.865 7。

2.2.3　SiPLS 模型

将预处理过的 1.0～3.5 THz 范围的光谱划分为 10～30 个子区间，在区间间隔划分数相同的条件下，分别计算了 2 个、3 个和 4 个区间联合的模型，并以优选区间进行模型建立和预测。

由表 3 可得：当区间联合个数为 2 时，在全频段被划分成 28 个间隔，取第 2、第 19 区间，主因子数为 7 时建模，RMSECV＝0.003 9，RMSEP＝0.005 3，R_c＝0.900 2，R_p＝0.816 1。

当区间联合个数为 3 时，在全频段被划分成 23 个间隔，取第 1、第 3 和第 16 区间，主因子数为 7 时建模，RMSECV＝0.003 8，RMSEP＝0.004 6，R_c＝0.906 2，R_p＝0.862 0。

当区间联合个数为 4 时，在全频段被划分成 20 个间隔，取第 1、第 3、第 7 和第 14 区间，主因子数为 7 时建模，RMSECV＝0.003 7，RMSEP＝0.004 7，R_c＝0.913 7，R_p＝0.859 9。

综合考虑相关系数 r，RMSECV，RMSEP 以及计算时间等因素，采用区间联合个数为 3 时，全频段被划分成 23 个间隔，取第 1、第 3 和第 16 区间，主因子数为 7 时建模，对应频率范围为 1.007 1～1.113 9，1.236 0～1.342 8 和 2.655 0～2.754 2 THz。

2.3　最佳模型优选

将采用上述三种方法所建立的模型进行比较，各模型预测结果如表 4。

表 2　不同区间划分数量时 BiPLS 建模模型

Table 2　Results of BiPLS model with different number of interval divisions

区间总数	入选区间数	入选变量数	主因子数	r	RMSECV/%
10	7	231	6	0.900 5	0.003 9
11	9	269	10	0.904 8	0.003 9
12	7	193	6	0.899 4	0.003 9
13	10	253	10	0.912 6	0.003 7
14	7	166	10	0.912 1	0.003 7
15	9	197	10	0.905 4	0.003 8
16	9	186	6	0.904 9	0.003 8
17	5	98	9	0.915 9	0.003 6
18	15	274	8	0.895 8	0.004 0
19	11	192	10	0.914 9	0.003 6
20	12	199	10	0.911 7	0.003 7
21	14	220	10	0.915 8	0.003 6
22	13	195	10	0.914 8	0.003 6
23	18	258	9	0.906 3	0.003 8
24	15	207	10	0.915 4	0.003 6
25	16	211	10	0.915 1	0.003 6
26	**13**	**166**	**10**	**0.919 3**	**0.003 5**
27	19	232	10	0.915 0	0.003 6
28	11	130	8	0.909 6	0.003 7
29	12	135	8	0.909 7	0.003 7
30	17	187	10	0.914 5	0.003 6

表 3　不同区间划分数量时 BiPLS 建模模型

Table 3　Results of SiPLS model with different number of interval divisions

联合区间个数	区间划分个数	入选区间	主因子数	R_c	RMSECV/%	R_p	RMSEP/%
2	28	[2 19]	7	0.900 2	0.003 9	0.816 1	0.005 3
3	**23**	**[1 3 16]**	**7**	**0.906 2**	**0.003 8**	**0.862 0**	**0.004 6**
4	20	[1 3 7 14]	7	0.913 7	0.003 7	0.859 9	0.004 7

表 4 不同光谱区间建模的优选模型

Table 4 Selected models with different spectral regions

模型类型	区间总数	入选区间数	主因子数	R_c	RMSECV/%	R_p	RMSEP/%
iPLS	10	1	8	0.848 2	0.004 8	0.761 8	0.006 0
BiPLS	26	26	10	0.919 3	0.003 5	0.865 7	0.004 6
SiPLS	23	3	7	0.906 2	0.003 8	0.862 0	0.004 6

由表 4 中数据可以得到，BiPLS 模型的 R_c 和 R_p 均高于 iPLS 模型和 SiPLS 模型，且运算速度远快于 SiPLS 模型。

本实验最后采用 BiPLS 模型用于润滑油中微量水含量的定量分析，模型区间数为 26，入选区间为[18 10 4 3 8 12 5 11 24 13 16 21 2]子区间组合，主因子数为 10，最优模型的 RMSECV=0.003 5，RMSEP=0.004 6，R_c=0.919 3，R_p=0.865 7，预测效果如图 3。

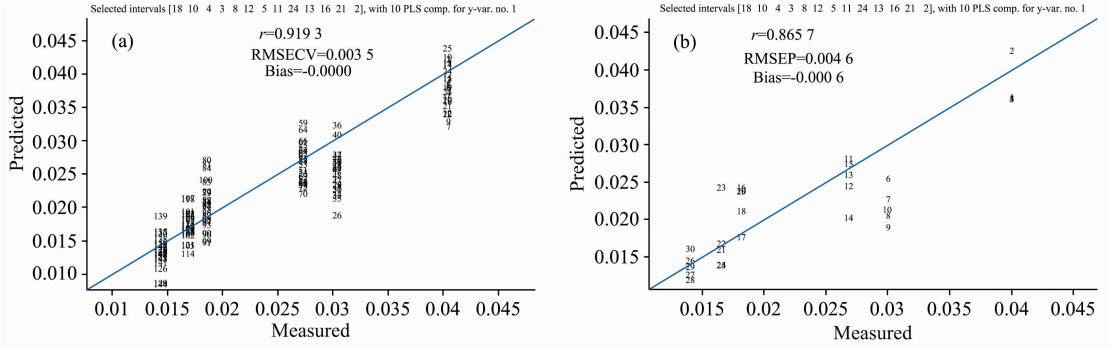

图 3 润滑油水含量的 BiPLS 模型(a)和最优预测结果(b)

Fig. 3 BiPLS model of lubricant water content (a) and optimal results of prediction models obtained by (b) BiPLS for water content of lubricating oil

3 结 论

基于太赫兹时域光谱，采用特征谱区间筛选算法建模并优选最佳建模方式。最终选用向后区间偏最小二乘法(BiPLS)用于发动机润滑油中微量水含量的定量分析，所建模型具有较好的定量分析效果，且建模计算速度快，计算量较小。为测定发动机润滑油中微量水含量提供了一种较为快速简便的方式，也为检定发动机润滑油老化程度提供了一种新的思路。

References

[1] YU Xian-shu, GAO Lei, LU Gui-wu(于宪书，高 磊，卢贵武). Lubrication Engineering(润滑与密封), 2016, 41(12): 26.
[2] WANG Cheng-yong(王成勇). Plant Maintenance Engineering(设备管理与维修), 2013, (1): 60.
[3] JIANG Qiang, WANG Yue, WEN Zhe, et al(蒋 强，王 玥，文 哲，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(4): 1049.
[4] HE Ming-xia, GUO Shuai(何明霞，郭 帅). Journal of Electronic Measurement and Instrument(电子测量与仪器学报), 2012, 26(8): 663.
[5] Walrafen G, Chu Y, Piermarini G. Journal of Physical Chemistry, 1996, 100(24): 10363.
[6] Cecilie Rönne, Sören Rud Keiding. Journal of Molecular Liquids, 2002, 101(1): 199.
[7] JIN Wu-jun, ZHAO Kun, YANG Chen, et al(金武军，赵 昆，杨 晨，等). Applied Geophysics(应用地球物理), 2013, 10(4): 506.

Application of Interval Selection Methods in Quantitative Analysis of Water Content in Engine Oil by Terahertz Spectroscopy

CHEN Meng-qiu[1], HE Ming-xia[1*], LI Meng[2], QU Qiu-hong[2]

1. State Key Laboratory of Precision Measuring Technology and Instruments, Tianjin University, Tianjin 300072, China

2. LET Terahertz (Tianjin) Technology Co., Ltd., Tianjin 300019, China

Abstract Engine lubricating oil is the cornerstone to ensure the long-term and stable operation of automobile engines. Accurately evaluating various performance indicators of engine lubricating oil is an essential step in the entire process from production to use. Engine lubricating oil will deteriorate for a variety of reasons after being used for a while. The engine lubricating oil deterioration indicators can be expressed in terms of non-magnetic particulate matter concentration, metal filings content, pH value, viscosity, water content and so on. To detect water content in engine lubricating oil, the traditional detection methods have the disadvantages of complicated operation and poor timeliness. Terahertz has strong absorption of water and is suitable for analysing micro-water content in sample products. In this paper, the transmission coefficients of six engine oils with different water contents were used to obtain the absorption coefficient curve of 1.0～3.5 THz by the transmission terahertz time domain spectroscopy system. The spectroscopic data were preprocessed with Savitzky-Golay(SG). Then, the sample was divided into a calibration set and test set by the Kennard-Stone algorithm after rejecting the odd samples. The interval Partial Least Squares (iPLS), backward interval partial least squares (BiPLS), and synergy interval partial least squares (SiPLS) were used to screen their terahertz time-domain spectral characteristic spectral intervals. They were focusing on the impact of factors such as the number of intervals, the number of PLS components, the number of best principal factors, and the selection of intervals on the PLS model's properties. It also models and analyzes lubricants with different water contents, compares and selects different models, and establishes an optimal quantitative analysis model. The modeling results indicate that the feature spectrum region filtering can improve modeling performance and reduce model complexity. The characteristic spectrum region screening algorithm eliminates the non-linear or irrelevant variables in the terahertz absorption coefficient spectrum of engine lubricants so that the modeling results can better express the relationship between the absorption coefficient spectrum and its water content. The results show that the optimal model for quantitative analysis of trace water content in generator lubricants was obtained with BiPLS method that separated the whole spectra into 26 intervals and selected [18 10 4 3 8 12 5 11 24 13 16 21 2] intervals. The number of major factors is 10. The BiPLS model had a root mean standard error of cross-validation (RMSECV) of 0.003 5 and root mean standard error of prediction(RMSEP) of 0.004 6. The correlation coefficient (r) of the correction set is 0.913 9, and the correlation coefficient (r) of the prediction set is 0.865 7. Overall, BiPLS method could accurately predict the water content of engine lubricants, and the experimental process is simple, the modeling and calculation speed is fast, and the effect is ideal, and it can be applied to the quantitative analysis of the water content of non-contact oil products.

Keywords Terahertz time-domain spectrum; Intervalselection; Engine oil; Moisture content test

(Received Apr. 24, 2020; accepted Jul. 19, 2020)

* Corresponding author