文章编号：1005-5630(2020)04-0061-06 **DOI**: 10.3969/j.issn.1005-5630.2020.04.010

基于近红外光谱法快速鉴别转基因油研究

朱建国[1,3,4]，王雅静[2,3,4]，尹知沁[3,4]，谢雷英[3,4,5]，王 娜[3,4,6]，曹 铎[2]

（1. 上海理工大学 材料科学与工程学院，上海 200093；
2. 上海师范大学 数理学院，上海 200234；
3. 中国科学院上海技术物理研究所 红外物理国家重点实验室，上海 200083；
4. 上海节能镀膜玻璃工程技术研究中心，上海 200083；
5. 上海科技大学 物质学院，上海 200120；
6. 复旦大学 信息科学与工程学院，上海 200433）

摘要：采用近红外光谱法对转基因油/非转基因油的混合溶液进行研究。对采集到的原始光谱分别进行多元散射校正(MSC)、一阶导数(FD)、移动窗口平滑(MWS)、Savitzky-Golay 平滑一阶导数(SG1)预处理。研究比较了不同预处理方法对转基因油/非转基因油支持向量机(SVM)建模判别分析的影响，其中 MSC 预处理后的模型预测效果最好，准确率为 91.6%。为了进一步提高模型的精度与稳定性，采用连续投影算法(SPA)对全波长进行特征波长筛选。利用筛选后的 15 个特征波长输入到 SVM 中，预测准确率提高到 98.3%。实验结果表明，采用近红外光谱法，可以实现对转基因油/非转基因油快速检测，不仅适用于纯转基因油的鉴别，也适用于非转基因油中掺入转基因油的鉴别。

关键词：转基因油；近红外光谱；支持向量机(SVM)
中图分类号：O 657.3 **文献标志码**：A

Research on fast identification of transgenic oil based on near infrared spectroscopy

ZHU Jianguo[1,3,4], WANG Yajing[2,3,4], YIN Zhiqin[3,4], XIE Leiying[3,4,5], WANG Na[3,4,6], CAO Duo[2]

(1. School of Materials Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;
2. Department of Physics, Shanghai Normal University, Shanghai 200234, China;
3. State Key Laboratory of Infrared Physics, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China;
4. Shanghai Engineering Research Center of Energy-Saving Coatings, Shanghai 200083, China;
5. School of Material Science, Shanghai University of Science and Technology, Shanghai 200120, China;
6. School of Information Science and Engineering, Fudan University, Shanghai 200433, China)

收稿日期：2020-03-09
基金项目：国家重点研发专项(2017YFC0111400)；国家自然科学基金(11874376)；上海市科学与技术委员会项目(18590712600、18DZ2282200、19DZ2293400、19ZR1465900)
作者简介：朱建国 (1994—)，男，硕士研究生，研究方向为红外光谱应用。E-mail：1277218957@qq.com
通信作者：曹 铎 (1988—)，男，讲师，研究方向为光学工程仪器。E-mail：dcao@shnu.edu.cn

Abstract: Near-infrared spectroscopy was used to study the mixed solution of transgenic oil/non-transgenic oil. The acquired original spectra were pretreated with multiple scattering correction (MSC), first-order derivative (FD), moving window smoothing (MWS) and savitzky-golay smoothing first-order derivative (SG1). The effects of different pretreatment methods on the discriminant analysis of transgenic oil/non-transgenic oil support vector machine (SVM) modeling were compared. The model after MSC pretreated has the best prediction effect, and the accuracy rate is 91.6%. In order to further improve the accuracy and stability of the model, the successive projections algorithm (SPA) is used to select the characteristic band. We input the 15 feature wavelengths after SPA selection into the SVM, and the prediction accuracy is 98.3%. The experimental results show that the near-infrared transmission spectrum combined with stoichiometry can achieve rapid and non-destructive detection of transgenic oil/non-transgenic oil, not only suitable for the identification of pure genetically modified oils, but also for the identification of non-transgenic oils that are mixed with genetically modified oils.

Keywords: transgenic oil；near infrared spectroscopy；support vector machine（SVM）

引 言

自从 1996 年第一个转基因生物（GMO）批准入市以来，引入市场的转基因农作物数量急剧增加[1]。采用转基因技术可以将抗虫基因[2]、抗病基因[3]和抗除草剂基因[4] 等优良基因引入到农作物中，以此来改善农产品的品质、缩短生长周期，缓解由于人口快速增加和可用耕地减少而带来的粮食危机。然而转基因技术的潜在安全性仍然存在着争议，比如：转移基因表达的蛋白质对生态环境的非预期影响[5]，外源基因逃逸对其他作物的潜在影响[6]，以及由基因转移引起的食物中毒、过敏反应和耐药性对人体的有害影响[7]。因此，如何快速鉴别是否为转基因产品是非常必要的。

目前，聚合酶链反应、酶联免疫吸附分析、二维电泳和微阵列分析是转基因产品和作物最常用的检测方法[8]。这些方法在大多数情况下都具有良好的特异性和敏感性，但是检测过程过于繁琐，检测时间长达数个小时，无法满足人们想要实时检测转基因产品的需求。而近红外光谱则是一种快速、无损、可实时在线检测的技术，不需要对转基因样品进行任何处理就能表征基因结构变化所带来的构型变化，进而可以通过 C—O 键、C—H 键、C—N 键等数据变化看出基因表达的差异[9]。2010 年翟亚峰等[10] 采用近红外光谱技术实现了对不同品种的 9 个小麦转基因种子样品的准确鉴别。2013 年 Luna 等[11] 用近红外光谱对非转基因大豆油和转基因大豆油进行独立识别，识别率分别为 100% 和 90%，由于是对纯的转基因油与纯的非转基因油样本进行识别，实际应用价值不高。

本文则对不同品牌的转基因油和非转基因油进行混合，构成不同混合比例的转基因油样本，并采用近红外光谱技术对这些油样本进行分析。通过研究不同预处理方法对光谱预测模型的影响，提高了光谱预测模型准确性，实现了对纯的转基因油以及非转基因油中掺入转基因油的有效鉴别。

1 实验部分

1.1 样品制备

购置不同品牌的转基因大豆油、转基因玉米油和非转基因大豆油、非转基因玉米油若干瓶。将转基因油与非转基因油按 1:1，1:2，1:3，…，1:20 等比例混合得到不同体积分数的转基因油样本 102 份，同时将不同品牌的非转基因油按 1:1，1:2，1:3，…，1:20 混合得到不同体积分数的非转基因油样本 102 份，混合好后的样品放在超声清洗机中用超声波使之充分混合。

1.2　近红外光谱采集

实验的转基因油与非转基因油的近红外光谱由傅里叶变换红外光谱仪 vertex70（Bruker, Germany）采集，分辨率为 2 cm^{-1}，光谱区域为 4 000～12 500 cm^{-1}，扫描 16 次。探测器为 InGaAs。具体的扫描次数由信噪比决定，若信噪比较差可适当增加扫描次数。每次采集光谱时，先以空的比色皿测试以便扣除系统背景。

2　结果与讨论

2.1　样本集划分

在建立光谱预测模型过程中，校正集样本与预测集样本的选择至关重要，而 Kennard Stone（KS）算法是一种应用广泛的样本集划分方法[12-13]。KS 算法以光谱间的欧氏距离为基础，选择代表性强，分布范围广的样品作为转换集样品[12]。根据 KS 法，我们选取转基因样本 72 个、非转基因样本 72 个，共计 144 个样本作为校正集，余下的 60 个样本作为预测集。

2.2　光谱与光谱预处理

对于不同种类的油其理化性质差别不大，主要脂肪酸都是棕榈酸、硬脂酸、油酸、亚油酸等，只是在含量上有所差别[14]。近红外光谱能够表征基因结构变化所带来的构型变化，进而可以通过 C—O 键、C—H 键、C—N 键等数据变化看出基因表达的差异[9]，而 C—O 键、C—H 键、C—N 键等在近红外波长的吸收峰又是不同的，因此可以通过观察近红外光谱吸收峰的位置和强度来找出转基因油与非转基因油之间的差异，如图 1 所示。从图 1 可以看出，转基因油与非转基因油在近红外波段差异不大，在 1 550～1 650 nm 和 1 800～2 100 nm 范围光谱强度有一定的区别。这是由于转基因油与非转基因油为同源性物质，这两种物质因化学键含量的不同而表现在光谱强度上有所差别。

为了尽可能去除来自外界或者系统的随机噪声、光散射等对转基因油与非转基因油透射光谱的影响以及提高光谱与待测组分之间的相关性，

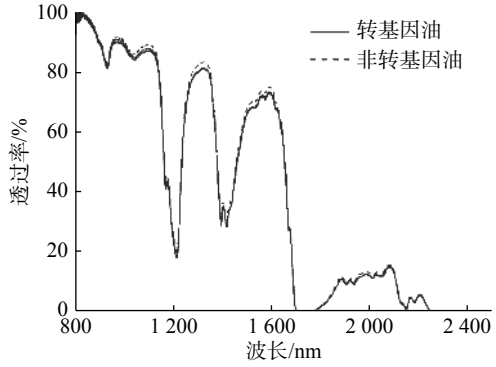

图 1　转基因油与非转基因油在近红外波段的原始光谱

Fig. 1　Spectra of genetically modified oil and non-transgenic oil in the near-infrared region

我们利用 MATLAB 2016a 软件，分别采用多元散射校正（MSC）、一阶导数（FD）、移动窗口平滑（MWS）、Savitzky-Golay 平滑一阶导数（SG1）等方法对原始光谱数据进行了预处理。光谱预处理结果如图 2 所示，其中多元散射校正可以有效去除散射对样品光谱的影响，移动窗口平滑则可以提高分析信号的信噪比及消除仪器的随机噪声，一阶导数和 Savitzky-Golay 平滑一阶导数可消除基线漂移、强化谱带特征和克服谱峰重叠[15]。

2.3　转基因油与非转基因油支持向量机模型的建立

支持向量机（SVM）是一种新的基于统计学习理论的机器学习方法。SVM 利用结构风险最小化原则避免过拟合问题，在最小化经验风险下所得结果优于传统的神经网络算法，而且在小样本、高维度数据情况下具有优异的建模能力[16]。我们选择 SVM 作为建模方法[17]，将预处理后的光谱数据分别输入到 SVM 中建立转基因油与非转基因油预测模型。通过对比不同预处理方法，建立模型后预测集样本的预测结果，选择最优预处理方式，预测结果如表 1 所示。在 SVM 模型参数选择中，选用径向基 RBF 核函数作为本次预测模型的核函数，并通过网格参数寻优和交叉验证获得最佳的惩罚因子系数（C）和核函数的参数系数（G）。

由表 1 可知，采用多元散射校正预处理方法预测准确率最高，达到了 91.6%，其他 3 种预处理方法准确率均不高于 75%。因此我们把多元散射校正定为转基因油和非转基因油后续其他建

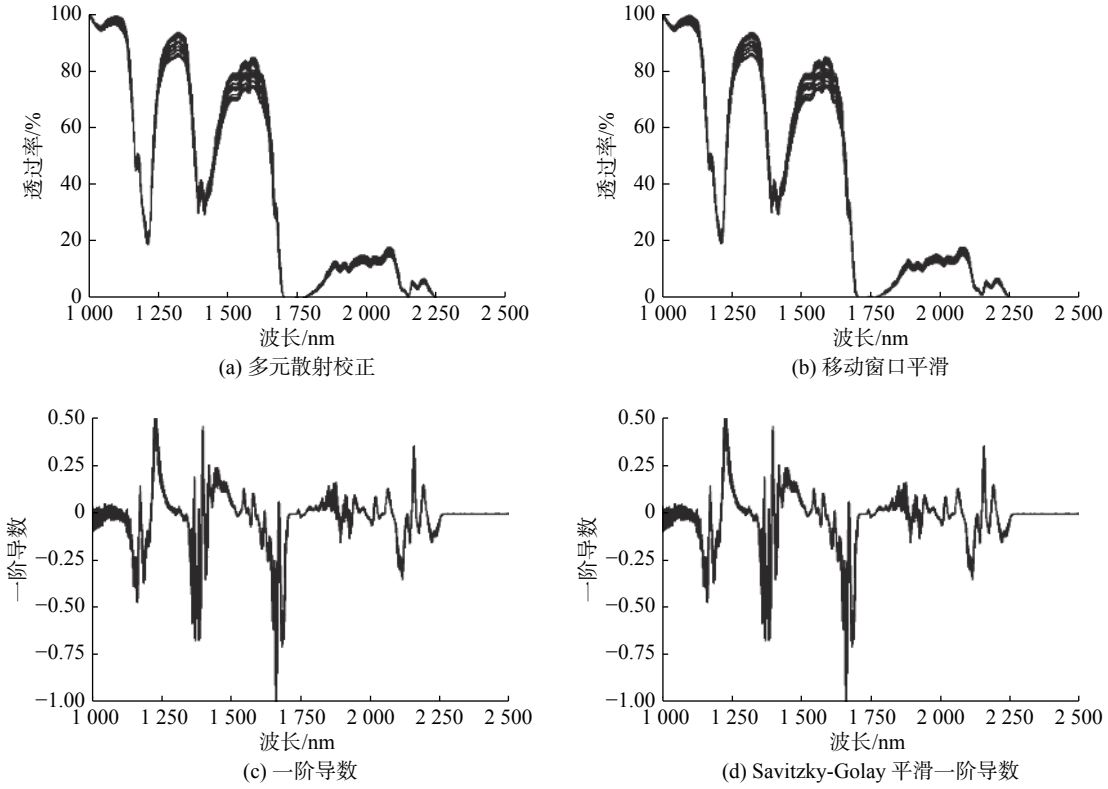

图 2　不同方法预处理后的光谱图

Fig. 2　Different methods preprocessed spectra

表 1　不同数据预处理方法的预测结果

Tab. 1　Prediction results of different pretreatment methods

预处理方法	准确率/%
多元散射校正	91.6
Savitzky-Golay平滑一阶导数	71.6
移动窗口平滑	65.0
一阶导数	75.0

模过程的光谱预处理方式，进而研究其他影响模型预测能力的因素。

2.4　变量波长的筛选

在光谱全波长建模中，虽然预测结果比较准确，光谱与待测性质表现出了很强的相关性。但光谱包含了 6000 多个数据点，其中包含了大量与待测性质无关的信息以及共线性变量。如果将这些冗余变量全部输入到模型中，不仅会增加模型的建立难度，而且还会降低模型的预测精度与稳定性。在近红外光谱分析中，特征波长筛选是非常重要的一步，通过光谱特征波长提取，可以有效地简化模型并提高模型的预测精度和稳定性。

连续投影算法(SPA)是一种向前变量筛选方法。通过选定一个初始波长，每一次迭代时加入新的波长，直至达到指定的波长数量。通过这种投影分析，从光谱矩阵中提取有效信息，并使光谱变量共线性达到最小[18]。通过 SPA 来提取特征波长可以有效地去除光谱数据间的冗余变量。图 3 显示了 SPA 的不同数量变量进行交叉验证的均方根误差(RMSE)趋势以及最终被选择的特征波长点。

从图 3(a)可以看出，当选择 15 个特征变量(1 152 nm、1 184 nm、1 210 nm、1 231 nm、1 410 nm、1 433 nm、1 660 nm、1 860 nm、1 895 nm、1 920 nm、1 935 nm、2 012 nm、2 038 nm、2 084 nm、2 102 nm)时，此时 RMSE 最小为 0.46。因此这 15 个特征波长点被输入到 SVM中，输入的特征波长如图 3(b)所示。

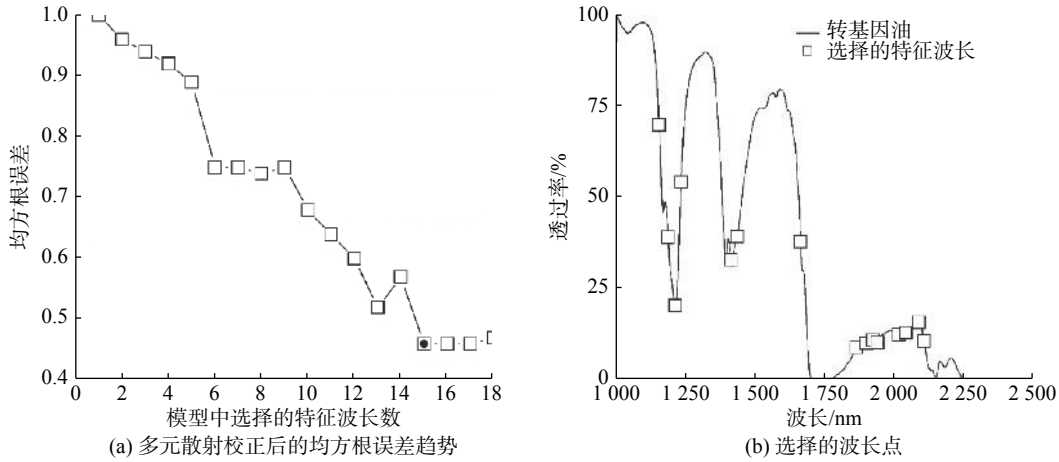

图 3　连续投影算法（SPA）特征波长选择结果

Fig. 3　Characteristic wavelength results selected by successive projections algorithm（SPA）

2.5　转基因油与非转基因油预测模型验证

与 Luna 等[11] 对转基因油与非转基因油进行鉴别的方法不同，我们是将不同的转基因油和非转基因油进行混合组成具有干扰性的转基因油样本进行分析。对混合后的转基因油与非转基因油放在一起进行预测，这样预测时的样本既可能是纯的转基因油，也可能是掺杂的转基因油，而不是纯的转基因油或者是纯的非转基因油，因而更接近实际应用情况。预测结果如图 4 所示，共对 60 个样本进行预测，其中 30 个非转基因油样本准确预测 29 个，准确率为 96.7%，仅有一个误判，而 30 个转基因油准确预测 30 个，准确率为 100%。

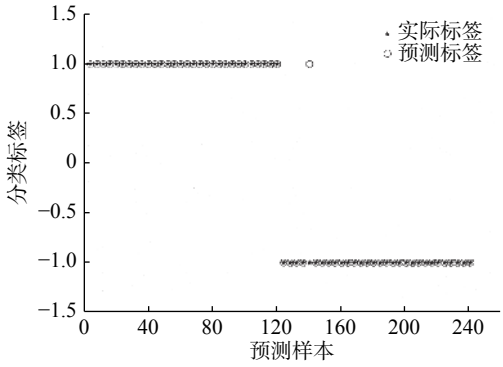

图 4　模型预测（标签为 1 的是转基因油，标签为−1 的是非转基因油）

Fig. 4　Model prediction（Label 1 is a genetically modified oil, and label −1 is a non-transgenic oil）

需要特别指出的是，我们所预测的 30 个转基因油中，只有 4 个是纯的转基因油，其余 26 个全部为非转基因油中掺入转基因油的样本，因此，只要食用油中有转基因油的存在就能被检测出来。与 Luna 等[11] 仅对纯转基因油样品进行判别的准确率（90%）相比，不但预测准确率更高，而且更有实用价值。该模型针对转基因油和非转基因油的整体预测准确率为 98.3%，相比于整体预测准确率为 91.6% 的 MSC-SVM 模型，MSC-SPA-SVM 模型提高了预测准确率，可以很大程度降低模型的复杂性，提高模型的预测精度。这也从侧面表明，样本光谱特征波长的提取对于提高模型预测精度、减少模型的复杂性发挥着至关重要的作用。同时通过对比不同预处理方式对预测结果的影响，可以发现，对于散射较为严重的样本，采用多元散射校正（MSC）预处理会大大增强光谱数据与待测性质之间的相关性，有利于光谱特征波段提取。

3　总　结

本文基于近红外光谱技术对转基因油和非基因油的鉴别进行了研究。通过 MSC 预处理方法，结合连续投影算法 SPA 和支持向量机 SVM 获得了很好的预测效果，准确率高达 98.3%。结果表明，转基因油与非转基因油基因表达在近红外波段有差异，从而可通过近红外光谱方法进行判别。通过筛选特征波长，可以有效地去除光谱数据间的冗余变量，提高模型的预测精度和鲁棒性。在我们所建立的预测模型中，只要食用油中

有转基因油的存在就能被检测出来。与其他传统检测方法相比，近红外光谱法操作简单、检测时间短、不破坏样品，可以满足消费者实时检测转基因产品的需求。后续研究将增加转基因油与非转基因油的种类，进一步扩大样本数和类型，以提高模型预测的普适性。

参考文献：

[1] BRARA Z, COSTA J, VILLA C, et al. Surveying genetically modified maize in foods marketed in Algeria[J]. Food Control, 2020, 109: 106928.

[2] BHOGE R K, CHHABRA R, RANDHAWA G, et al. Event-specific analytical methods for six genetically modified maize events using visual and real-time loop-mediated isothermal amplification[J]. Food Control, 2015, 55: 18 – 30.

[3] SUCHER J, BONI R, YANG P, et al. The durable wheat disease resistance gene *Lr34* confers common rust and northern corn leaf blight resistance in maize[J]. Plant Biotechnology Journal, 2017, 15(4): 489 – 496.

[4] MENKIR A, CHIKOYE D, LUM F. Incorporating an herbicide resistance gene into tropical maize with inherent polygenic resistance to control *Striga hermonthica* (Del.) Benth[J]. Plant Breeding, 2010, 129(4): 385 – 392.

[5] RAHMAN M, ZAMAN M, SHAHEEN T, et al. Safe use of *Cry* genes in genetically modified crops[J]. Environmental Chemistry Letters, 2015, 13(3): 239 – 249.

[6] YAN S, ZHU J L, ZHU W L, et al. Pollen-mediated gene flow from transgenic cotton under greenhouse conditions is dependent on different pollinators[J]. Scientific Reports, 2015, 5(1): 15917.

[7] BAWA A S, ANILAKUMAR K R. Genetically modified foods: safety, risks and public concerns-a review[J]. Journal of Food Science and Technology, 2013, 50(6): 1035 – 1046.

[8] LIU X D, FENG X P, LIU F, et al. Rapid identification of genetically modified maize using laser-induced breakdown Spectroscopy[J]. Food and Bioprocess Technology, 2019, 12(2): 347 – 357.

[9] 王晨光, 许文涛, 黄昆仑, 等. 转基因食品分析检测技术研究进展 [J]. 食品科学, 2014, 35(21): 297 – 305.

[10] 翟亚锋, 苏谦, 邬文锦, 等. 基于仿生模式识别和近红外光谱的转基因小麦快速鉴别方法 [J]. 光谱学与光谱分析, 2010, 30(4): 924 – 928.

[11] LUNA A S, DA SILVA A P, PINHO J S A, et al. Rapid characterization of transgenic and non-transgenic soybean oils by chemometric methods using NIR spectroscopy[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2013, 100: 115 – 119.

[12] 陈奕云, 赵瑞瑛, 齐天赐, 等. 结合光谱变换和 Kennard-Stone 算法的水稻土全氮光谱估算模型校正集构建策略研究 [J]. 光谱学与光谱分析, 2017, 37(7): 2133 – 2139.

[13] 李华, 王菊香, 邢志娜, 等. 改进的 K/S 算法对近红外光谱模型传递影响的研究 [J]. 光谱学与光谱分析, 2011, 31(2): 362 – 365.

[14] 于殿宇, 陈晓慧, 宋云花, 等. 转基因和非转基因大豆油理化性质比较研究 [J]. 东北农业大学学报, 2012, 43(11): 1 – 6.

[15] 陆婉珍. 现代近红外光谱分析技术 [M]. 2 版. 北京: 中国石化出版社, 2007.

[16] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer, 2000.

[17] 翟明阳, 赵远, 高浩, 等. 关节软骨的红外光谱成像及支持向量机定量研究 [J]. 分析化学, 2018, 46(6): 896 – 901.

[18] TANG R N, CHEN X P, LI C. Detection of nitrogen content in rubber leaves using Near-Infrared (NIR) spectroscopy with correlation-based Successive Projections Algorithm (SPA)[J]. Applied Spectroscopy, 2018, 72(5): 740 – 749.

（编辑：刘铁英）