

太赫兹光谱在转基因菜籽油鉴别中的应用： 基于改进蜉蝣算法的支持向量机模型*

陈涛[†] 李欣

(桂林电子科技大学电子工程与自动化学院, 桂林 541004)

(2023 年 9 月 27 日收到; 2023 年 11 月 22 日收到修改稿)

为实现对转基因和非转基因菜籽油的快速准确鉴别, 结合太赫兹时域光谱技术, 提出了一种基于改进蜉蝣优化算法的支持向量机模型. 以两种转基因和两种非转基因菜籽油为研究对象, 应用太赫兹时域光谱技术获取其光谱信息, 发现相比于非转基因菜籽油, 转基因菜籽油在太赫兹波段具有更强的吸收特性, 同时它们的吸收光谱极为相似, 难以通过观察法进行准确区分. 为此, 提出一种基于改进蜉蝣优化算法的支持向量机模型, 通过采用蜉蝣优化算法对支持向量机参数进行寻优, 并引入自适应惯性权重和 Lévy 飞行两种策略改进蜉蝣优化算法在寻优过程容易陷入局部最优解的问题, 增强蜉蝣优化算法的全局搜索能力和稳健性. 实验结果表明: 改进后的蜉蝣优化算法能够更有效地寻找到支持向量机的最优参数组合, 提升鉴别模型的整体性能, 该模型对 4 种菜籽油的识别精度为 100%. 因此, 本研究为转基因菜籽油的类型鉴别提供了一种快速有效的新方法, 也为其他转基因物质的鉴别提供了有价值的参考.

关键词: 转基因菜籽油, 太赫兹光谱, 分类鉴别, 蜉蝣优化算法

PACS: 87.50.U, 87.64.-t, 07.57.-c

DOI: 10.7498/aps.73.20231569

1 引言

菜籽油是世界上第三大植物油品种, 其富含不饱和脂肪酸、维生素 E 和多种矿物质, 有助于心血管健康, 维持皮肤健康, 为人体提供重要的营养成分和能量来源. 据农业生物技术应用国际服务机构统计, 2019 年, 全球油菜中有 27% 是转基因作物^[1]. 转基因油菜是全球四大转基因作物之一, 其主要用途是生产菜籽油. 虽然转基因菜籽油已成为生活中常见的食用油, 但截至目前还没有任何研究能够彻底否认其潜在危害^[2]. 在消费市场上, 不注明转基因标示或将转基因产品标识为非转基因的情况屡见不鲜. 因此, 基于对公众食品安全的考虑, 对转基因菜籽油的鉴别具有重要的现实意义. 目前常见

的转基因产品检测方法有两种: 一种是基于脱氧核糖核酸 (deoxyribonucleic acid, DNA) 的方法^[3], 另一种是基于蛋白质的检测技术^[4]. 由于转基因菜籽油中 DNA 和蛋白质含量极低, 采用上述两种方法均存在提取过程繁琐、耗时较长、会损坏原有物质和非专业人员难以胜任等问题. 因此, 寻找一种快速无损和操作便捷的转基因菜籽油检测方法显得尤为重要.

太赫兹 (terahertz, THz) 波是指频率在 0.1—10 THz 范围的一段电磁波, 是宏观电子学和微观光子学的交叉研究领域, 具有很大的应用价值和学术价值^[5,6]. 理论研究表明, 许多生物分子 (如 DNA、蛋白质和脂肪等) 的振动和转动能级正好处于 THz 频带范围内^[7,8]. 因此, 应用太赫兹时域光谱 (terahertz time-domain spectroscopy, THz-TDS) 技术

* 国家自然科学基金 (批准号: 62261012, 61841502) 资助的课题.

[†] 通信作者. E-mail: tchen@guet.edu.cn

探测生物样品产生共振吸收峰, 并通过 THz 光谱来识别生物样品成为了可能^[9]. 目前, 利用 THz 光谱进行转基因食用油的检测识别已较多. 文献^[10]报道了 THz-TDS 在检测转基因大豆油上的应用, 文献^[11]报道了 THz-TDS 在检测转基因玉米油上的应用, 文献^[12]报道了 THz-TDS 在检测转基因山茶油上的应用.

然而, 通过对文献^[10–12]的分析可知, 同种转基因和非转基因植物油的 THz 光谱极为相似, 难以直接从光谱上对它们进行准确区分, 需要结合一些模式识别方法才能实现对它们的准确区分. 因此, 本文应用支持向量机 (support vector machine, SVM) 方法对转基因和非转基因菜籽油进行鉴别. 由于 SVM 对参数较为敏感, 选取合适的参数才可较好提升其性能^[13], 因此 SVM 常与优化算法结合使用. 蜉蝣优化算法 (mayfly optimization algorithm, MOA) 与其他传统优化算法相比, 有着较好的求解精度和较快的收敛速度, 但也由于较快的收敛速度, 其在寻优过程中容易陷入局部最优解, 全局搜索能力较弱^[14], 因此为了提升 MOA 的整体搜索性能和精度, 本文引入自适应惯性权重 (adaptive inertia weight, AIW) 以及 Lévy 飞行两种策略来改进 MOA (命名为 ALMOA). 本文将 ALMOA 应用于 SVM 重要参数的寻优过程中, 从而得到一种基于改进蜉蝣优化算法的支持向量机模型 (ALMOA-SVM), 来实现对转基因和非转基因菜籽油的快速

准确鉴别.

2 实验部分

2.1 实验设备

本文采用的实验设备为美国 Zomega 公司生产的 Z-3 THz-TDS 系统, 该系统主要由超快飞秒光纤激光器、THz 辐射产生装置、THz 辐射探测装置和延时控制装置四部分组成, 系统原理图如图 1 所示. 该系统激光的中心波长为 780 nm, 脉冲宽度低于 100 fs, 信噪比高于 70 dB. 整个实验在室温下进行, 为避免潮湿空气中水分对 THz 波吸收的影响, 实验前在样品实验舱中充满干燥的氮气, 使其内部密闭空间的相对湿度小于 2%, 以保证实验数据的准确性.

2.2 样品制备

实验选取的样品为在市面上容易获取的 4 种不同品牌的转基因和非转基因菜籽油, 样品信息如表 1 所示. 所有油样均为具有国家质量监督检验检疫认证的合格产品. 实验样品在实验前都在低温避光环境下储存以防止变质和氧化. 实验样品架选择窗片材料为聚四氟乙烯薄膜的可拆卸液体池, 由于聚四氟乙烯在 THz 波段具有较低的吸收特性, 所以不会对待测样品产生干扰. 可拆卸液体池的厚度为 0.5 mm, 中心为面积为 270 mm² 的椭圆孔.

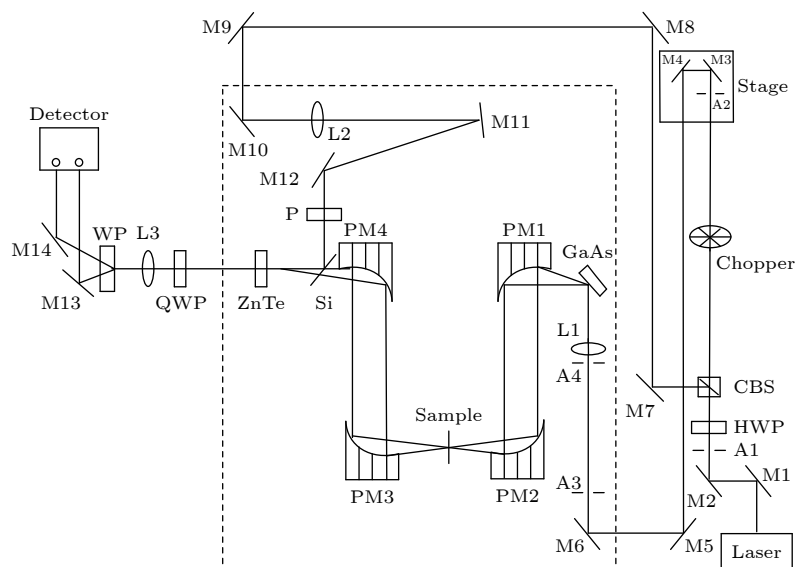


图 1 THz-TDS 系统原理图

Fig. 1. Schematic diagram of THz-TDS system.

在制样时, 采用 5 mL 的一次性医用注射器吸取约 2 mL 的油样, 沿液体池壁轻压注射器, 使油样缓慢注入液体池中, 以避免气泡的产生. 每种菜籽油制作 90 个样本, 共计 360 个, 其中每种菜籽油随机选取 70% 的样本作为训练集, 剩余的 30% 作为测试集.

表 1 实验样品信息
Table 1. The information of experimental sample.

标识符	品牌	类型	样本数	
			训练集	测试集
Non-GMO1	道道全	非转基因	63	27
Non-GMO2	鲁花	非转基因	63	27
GMO1	金龙鱼	转基因	63	27
GMO2	乡佬坎	转基因	63	27

2.3 数据处理方法与模型评价指标

在太赫兹时域光谱中, 获取的信息较为有限, 为进一步研究转基因和非转基因菜籽油在 THz 波段的吸收特性, 对实验测得的太赫兹时域参考信号和样品信号进行快速傅里叶变换, 得到各自的频域信号, 然后通过 (1) 式计算获得样品的吸光度, 以此来表征 4 种菜籽油对 THz 波的吸收程度.

$$A(\omega) = \log_{10} \left| \frac{E_{\text{ref}}(\omega)}{E_{\text{sam}}(\omega)} \right|^2, \quad (1)$$

其中, $E_{\text{ref}}(\omega)$ 为频域参考信号, $E_{\text{sam}}(\omega)$ 为频域样品信号, ω 为角频率.

为了更好地对分类鉴别模型的性能进行评估, 采用查准率 P 、查全率 R 和精度 A 作为模型评价指标, 计算公式如下:

$$P = \frac{TP}{TP+FP}, \quad (2)$$

$$R = \frac{TP}{TP+FN}, \quad (3)$$

$$A = \frac{TP+TN}{TP+FP+FN+TN}, \quad (4)$$

其中, TP 为真正类, 即模型正确地将某类物质 (设为正类) 预测为该类药物 (正类) 的个数; FP 为假正类, 即模型错误地将其他类药物 (设为负类) 预测为该类药物 (正类) 的个数; TN 为真负类, 即模型正确地将其他类药物 (负类) 预测为其他类药物 (负类) 的个数; FN 为假负类, 即模型错误地将该类药物 (正类) 预测为其他类药物 (负类) 的个数.

3 分类模型

3.1 支持向量机

SVM 是一种基于统计学习理论的有监督学习方法 [15,16]. 其核心原理在于将数据映射到高维空间, 以寻找一个能够最大化不同类别数据间边界距离的超平面, 从而实现对数据的有效分类. 通过引入核函数, SVM 可以处理非线性分类问题, 将其转化为在高维特征空间中的线性分类任务. 同时, SVM 以结构风险最小化为原则, 通过在特征空间中找到最优超平面来解决分类问题, 具有较强的泛化能力和对噪声的抵抗能力.

在实际的应用中, 合适的 SVM 参数选择将决定模型的泛化能力和分类性能优劣, 本文选择径向基函数 (radial basis functions, RBF) 作为 SVM 的核函数, 因此该模型的分类能力主要取决于正则化参数 c 和径向基函数 g 两个参数, 本文进一步采用蜉蝣优化算法 (MOA) 对 SVM 的参数进行寻优.

3.2 蜉蝣优化算法

MOA 是 2020 年由 Konstantinos 等 [17] 根据蜉蝣的飞行和繁衍行为提出的启发式算法, 用于解决复杂的函数优化问题. 算法的工作原理如下: 最初, 随机生成两组蜉蝣, 分别代表雄性和雌性种群. 将每个蜉蝣随机放置在问题空间中, 作为由 d 维向量 $\mathbf{x} = (x_1, x_2, x_3, \dots, x_d)$ 表示的候选解, 并在预先定义的适应度函数 $f(\mathbf{x})$ 上评估其性能. 蜉蝣的速度 $\mathbf{v} = (v_1, v_2, v_3, \dots, v_d)$ 定义为其位置的变化, 每只蜉蝣的飞行方向是个体和社会飞行经验动态交互作用. 雄性通过全局最优位置和自身历史最优位置移动, 雌性则是向优于自己的配偶移动, 若配偶弱于自己则自行局部搜索, 移动结束后, 雌性和雄性蜉蝣进行交配并产生后代, 子代有较小的概率产生变异, 最后淘汰子代和亲代中适应度较差的个体, 维持种群整体数量不变, 重复上述过程.

3.3 蜉蝣优化算法的改进

3.3.1 引入自适应惯性权重

惯性权重对解的搜索精度和收敛次数有着良好的指导性作用, 较大的惯性权重有利于全局搜索, 较小的惯性权重则有利于局部搜索. 由于 MOA

采用的是线性的惯性权重, 其全局和局部搜索能力一般, 为了更好地发挥算法的全局搜索以及局部搜索能力, 本文采用一种自适应非线性惯性权重^[18,19], 使之在迭代初期缓慢减小, 主要发挥算法的全局搜索能力, 从而达到圈定最优解范围的目的, 在迭代后期, 惯性权重减小加快, 从而快速增强算法的局部搜索能力, 精准锁定最优解位置. 这里, 定义自

适应非线性惯性权重 w 如 (5) 式所示:

$$w = w_{\max} - (t/t_{\max})^3 (w_{\max} - w_{\min}), \quad (5)$$

其中, w_{\max} 和 w_{\min} 分别为最大和最小惯性权重, 分别取值 0.8 和 0.4; t_{\max} 为最大迭代次数; t 为当前迭代次数.

将惯性权重 w 引入 MOA 中, 雄性蜉蝣个体的速度更新为

$$v_{ij}^{t+1} = \begin{cases} wv_{ij}^t + a_1 e^{-\beta r_p^2} (p_{\text{best}ij} - x_{ij}^t) + a_2 e^{-\beta r_g^2} (g_{\text{best}j} - x_{ij}^t), & f(x_i) > f(g_{\text{best}}), \\ wv_{ij}^t + dr, & f(x_i) \leq f(g_{\text{best}}), \end{cases} \quad (6)$$

其中, x_i^t 为在第 t 次迭代时雄性蜉蝣 i 在搜索空间中的当前位置; v_{ij}^{t+1} 为在第 $t+1$ 次迭代时蜉蝣 i 的速度; v_{ij}^t 为第 t 次迭代时蜉蝣 i 在 j 维上的速度; x_{ij}^t 为在第 t 次迭代时雄性蜉蝣 i 在 j 维上的位置; g_{best} 为全局最优位置; p_{best} 为自身历史最优位置; a_1 和 a_2 为蜉蝣游动行为的吸引系数; β 为能见度系数, 用于控制蜉蝣的能见范围; r_p 为当前位置 x_i^t 与 p_{best} 的距离; r_g 为当前位置 x_i^t 与 g_{best} 的距离; d 为舞蹈系数; $r \in [-1, 1]$, 是一个随机值.

雌性蜉蝣个体的速度更新如 (7) 式所示:

$$v_{ij}^{t+1} = \begin{cases} wv_{ij}^t + a_2 e^{-\beta r_{\text{mf}}^2} (x_{ij}^t - y_{ij}^t), & f(y_i) > f(x_i), \\ wv_{ij}^t + c_{\text{fl}} r, & f(y_i) \leq f(x_i), \end{cases} \quad (7)$$

其中, y_i^t 为在第 t 次迭代时雌性蜉蝣 i 在搜索空间中的当前位置; r_{mf} 为雌雄蜉蝣之间的笛卡尔距离; c_{fl} 为随机游走系数.

3.3.2 融合 Lévy 飞行策略

针对 MOA 容易陷入局部最优的问题, 利用 Lévy 飞行的跳跃能力来增强其跳出局部最优的能力^[20]. Lévy 飞行策略模拟自然界中动物的随机觅食行走, 假设种群中的蜉蝣均存在一定的概率不直接沿着最优路径移动, 而是根据 Lévy 飞行策略在最优路径附近进行随机游走, 从而达到跳出当前局部最优位置, 扩大全局搜索能力的目的. 同时为了避免在迭代后期, 蜉蝣一直在全局最优位置周围游走, 而不收敛于全局最优位置, 为 Lévy 飞行增加步长调整参数 δ ^[21]:

$$\delta = \delta_{\max} - \left[\frac{\delta_{\max} - \delta_{\min}}{\arctan(b)} \right] \times \arctan \left[b \left(\frac{t}{t_{\max}} \right)^a \right], \quad (8)$$

其中, δ_{\max} 和 δ_{\min} 分别为最大和最小步长调整参数, 分别取值 1 和 0; a, b 为常数, 分别取值 4 和 20.

通过上述参数的取值, 此时 $\delta \in [0, 1]$, 在迭代前期, δ 从 1 开始缓慢减小, 发挥 Lévy 飞行的全局游走优势, 增强算法的全局搜索能力, 在迭代中期 δ 开始迅速减小, 并至迭代后期逐渐趋于零, 目的是为了保证算法在迭代后期主要进行局部搜索, 从而快速收敛于全局最优位置.

雄性和雌性蜉蝣个体的位置更新为

$$x_i^{t+1} = x_i^t + v_i^{t+1} + \delta L(\alpha), \quad (9)$$

$$y_i^{t+1} = y_i^t + v_i^{t+1} + \delta L(\alpha), \quad (10)$$

其中, $L(\alpha)$ 符合 Lévy 分布, 稳定参数 $\alpha = 1$.

通过上述两种策略的改进, 相比于 MOA, ALMOA 在迭代前期具有更强的全局搜索能力, 在迭代后期具有更强的局部搜索能力. 由此构建得到的 ALMOA-SVM 模型, 解决了 MOA 在 SVM 参数寻优过程中容易陷入局部最优解的问题, 增强了 SVM 最优参数的搜索精度, 提升了模型的整体性能.

4 实验结果与分析

4.1 光谱分析

通过实验获取 4 种菜籽油共计 360 个样本的 THz 时域光谱如图 2 所示, 实验设置的扫描窗口长度为 30 ps, 光谱分辨率约为 33.3 GHz, 图中 Reference 表示参考信号, 为实验舱中样品架空载时的测量值. 由图 2 可见, 同种菜籽油不同样本的时域波形之间存在一定的差异, 不同菜籽油样本的时域波形之间存在一定的交叉重叠. 为了更清楚地

观测到转基因与非转基因菜籽油存在的差异, 对每种菜籽油 90 个样本的 THz 时域光谱数据求平均, 得到 4 种菜籽油的 THz 平均时域光谱如图 3 所示. 可以看出, 所有菜籽油的谱线相对于参考信号, 在幅值上均呈现一定程度的衰减, 在时间上均呈现一定的时延, 表明菜籽油对 THz 光谱具有一定的吸收特性. 其中, Non-GMO1 油样的相位延迟最长, GMO2 油样的振幅衰减最多. 总体上看, 转基因菜籽油样品相对于非转基因菜籽油样品, 在相位上延迟更少, 在幅值上衰减更大.

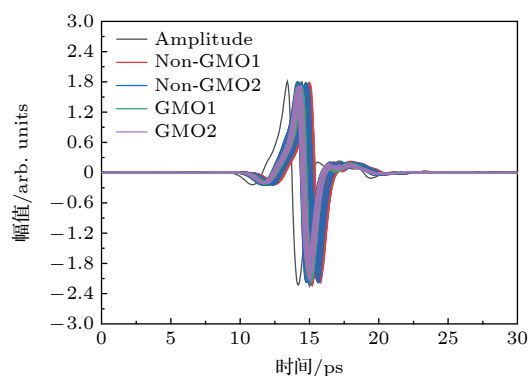


图 2 360 个菜籽油样本的 THz 时域光谱

Fig. 2. THz time-domain spectra of 360 rapeseed oil samples.

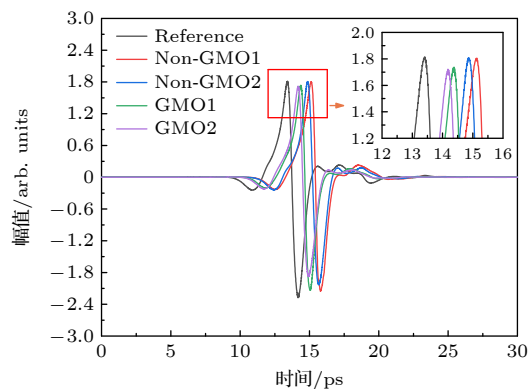


图 3 4 种菜籽油及参考信号的 THz 时域光谱

Fig. 3. THz time-domain spectra of four types of rapeseed oils and reference signal.

为了进一步研究转基因和非转基因菜籽油在 THz 波段内各频率的变化特性, 将平均时域光谱补零后进行快速傅里叶变换得到其平均频域谱, 如图 4 所示. 可见, 所有样品信号相对于参考信号, 在 0.3 THz 之后均开始出现一定程度的衰减, 同时在 1.8 THz 之后参考信号和样品信号均开始出现明显的振荡现象, 表明在 1.8 THz 之后信号受噪声

影响加剧. 从整体上看, 在 0.3—1.8 THz 波段, 转基因菜籽油样品相对于非转基因菜籽油样品, 在幅值上呈现出更大的衰减趋势. 通过上述分析可知, 转基因菜籽油样品相对于非转基因菜籽油样品, 在 THz 波段表现出更强的吸收特性.

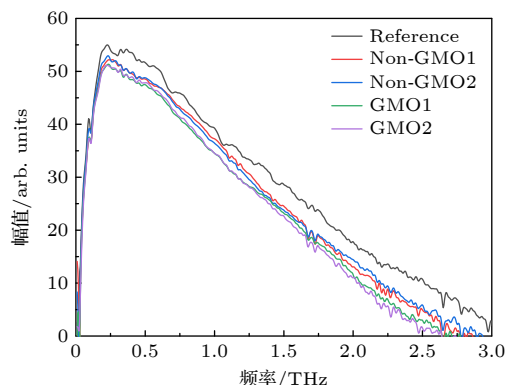


图 4 4 种菜籽油及参考信号的 THz 频域光谱

Fig. 4. THz frequency-domain spectra of four types of rapeseed oils and reference signal.

通过 (1) 式计算 4 种菜籽油在 0.3—1.8 THz 频段内的太赫兹吸光度, 获得 360 个菜籽油样本的太赫兹吸光度谱如图 5 所示. 可见, 所有菜籽油样本在 0.3—1.8 THz 波段呈现出相似的波形和相近的幅值, 无显著差异. 通过对每种菜籽油 90 个样本的吸光度取平均, 计算得到 4 种菜籽油样品的平均吸光度谱如图 6 所示. 可以看出转基因菜籽油样品相对于非转基因菜籽油样品, 在 THz 波段的吸光度更高, 说明转基因菜籽油样品在 THz 波段具有更强的吸收特性^[10,11], 与频域谱中观测到的结果相一致, 这可能是由于转基因油菜中引入了外源基因, 如高油酸基因、亚麻酸合成基因等, 改变了菜

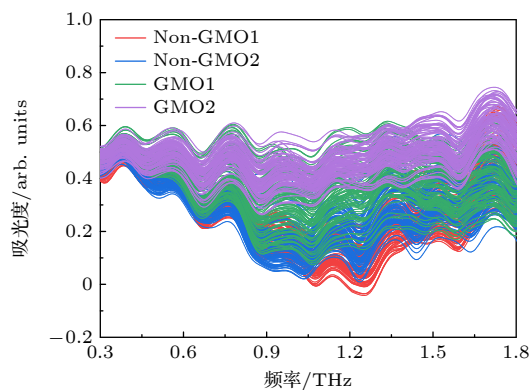


图 5 360 个菜籽油样本在 0.3—1.8 THz 波段内的吸光度谱

Fig. 5. Absorption spectra of 360 rapeseed oil samples in the 0.3—1.8 THz range.

籽油的脂肪酸组成含量,从而使转基因菜籽油在太赫兹波段具有更强的吸收特性^[22,23].同时可以清楚地发现转基因和非转基因菜籽油样品的波形极为相似,吸收峰所处频率位置也基本一致,这可能是由于转基因和非转基因菜籽油的成分极为相似所致,而波形存在差异的原因之一可能是由于不同来源菜籽油中相似成分的含量存在差异,从而导致它们与太赫兹共振吸收峰在光谱上呈现出一定的差异,因此,采用直接观察的方式很难对它们进行准确的鉴别.

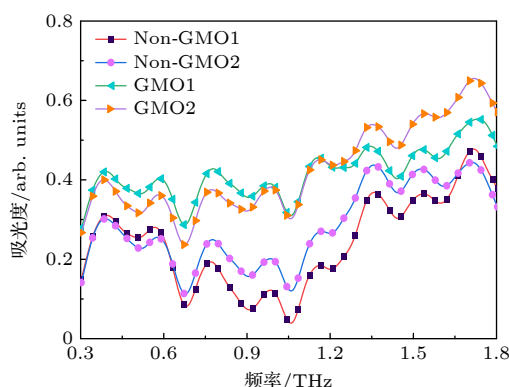


图6 4种菜籽油在0.3—1.8 THz波段内的平均吸光度谱
Fig. 6. Average absorption spectra of four types of rapeseed oils in the 0.3–1.8 THz range.

4.2 主成分分析

由于菜籽油样品的吸光度数据维数过高,若将其直接输入到鉴别模型中,计算量较大且十分耗时,这将会对模型性能产生负面影响.因此,为了减少光谱数据的冗余,提高建模效率,采用主成分分析 (principal component analysis, PCA) 对菜籽油吸光度谱中 0.3—1.8 THz 波段的原始数据 (330 维) 进行降维,得到各主成分的方差贡献率变化条形图如图 7 所示.可以看出,前 3 个主成分占据了原始数据的绝大部分信息,其累计方差贡献率达到了 98.27%,图 8 给出了前 3 个主成分的三维 (3D) 散点图,从图 8 可以看出,4 种菜籽油的主成分在三维空间中呈现出了不同的聚集区域,但也存在一些交叉重叠的地方,如 Non-GMO1 的主成分分布较为分散,与其他 3 种油样的主成分均有部分区域重叠;而 Non-GMO2, GMO1 和 GMO2 的主成分则分布则较为集中,但它们聚集区域的边缘位置也存在部分区域相互重叠.因此仅通过 PCA 不足以对样本进行完全正确的分类,但也说明了 PCA

能够有效提取不同菜籽油吸光度谱中的特征信息.从图 7 可以看出,前 9 个主成分的累积方差贡献率超过了 99.8%,可以近似解释所有原变量,因此采用这 9 个新变量代替原始光谱数据来进行后续建模处理.

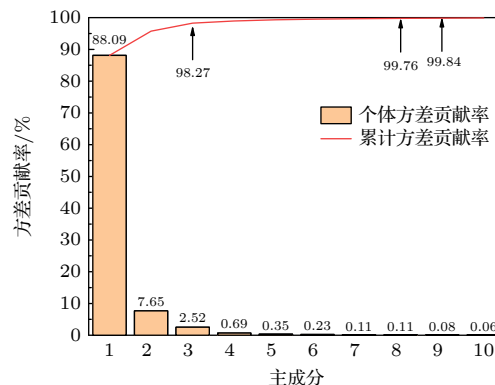


图7 吸光度的主成分方差贡献率变化条形图

Fig. 7. Bar chart of variance contribution rates for absorbance's principal components.

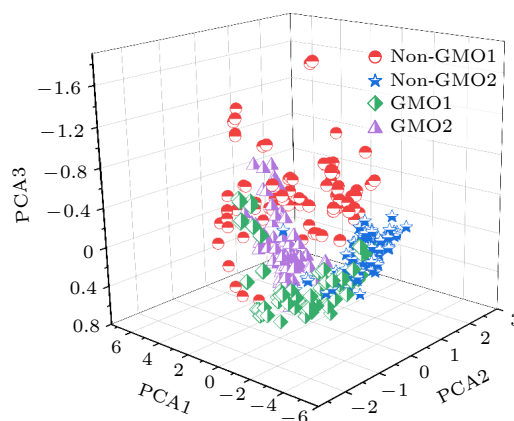


图8 吸光度前3个主成分的3D散点图

Fig. 8. 3D scatter plot of the first three principal components of absorbance.

4.3 参数寻优及模型鉴别

在训练集中分别用 MOA 和 ALMOA 对 SVM 进行参数寻优,寻找最佳的正则化参数 c 和径向基函数 g 参数,寻优过程如图 9 所示,寻优结果如表 2 所示.从图 9(a) 可以看出,MOA 的收敛速度很快,在迭代前期便快速取得了最佳适应度 97.22% (最佳参数 $(c, g) = (12.42, 0.79)$),同时平均适应度也几乎同步增长至最佳适应度附近,但在迭代中期和迭代后期,最佳适应度一直稳定不变,平均适应度也仅在最佳适应度下略微起伏,这说明 MOA 在迭代前期快速取得较高的局部最佳适应度后,迭代

中期至迭代后期一直在局部最佳适应度附近进行寻优, 未能跳出局部最优解扩大全局搜索范围. 经多次实验发现, MOA 常常在参数寻优的迭代前期便陷入了不同的局部最优解, 说明 MOA 较为依赖雌雄蜉蝣初始的随机位置, 全局搜索能力较差. 从图 9(b) 可以看出, ALMOA 在迭代前期也快速取得了局部最佳适应度 97.62%, 但由于该算法在迭代前期具有较强的全局搜索能力, 在图中具体表现为其平均适应度在迭代前期有较大的波动, 因此其顺利跳出了当前的局部最优解, 并在迭代中期再次跳出了局部最优解, 最终取得了全局最佳适应度 98.41% (最佳参数 $(c, g)=(84.62, 0.12)$). 同时, 从图 9(b) 中的平均适应度曲线变化可以发现, 其

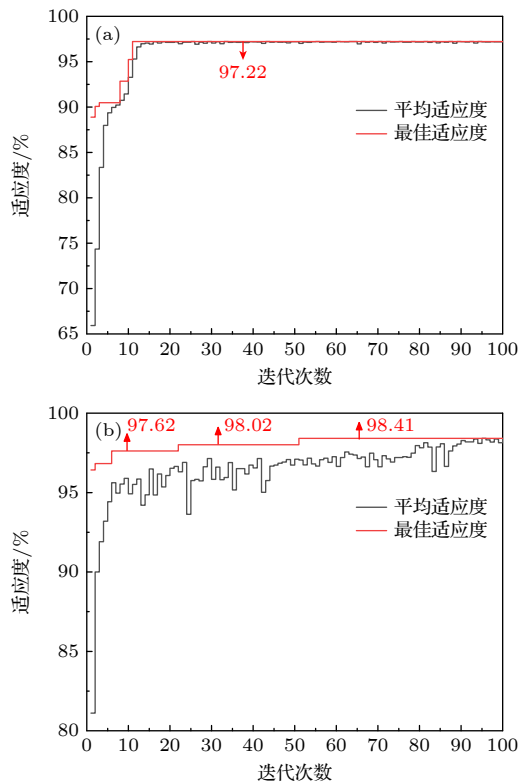


图 9 两种算法下 SVM 参数寻优过程中的适应度变化曲线 (a) MOA; (b) ALMOA

Fig. 9. Fitness evolution curves during SVM parameter optimization process for two algorithms: (a) MOA; (b) ALMOA.

表 2 两种算法的 SVM 参数寻优结果

Table 2. Results of SVM parameter optimization under two algorithms.

优化算法	最佳适应度/%	参数	
		c	g
MOA	97.22	12.42	0.79
ALMOA	98.41	84.62	0.12

波动幅度大致随着迭代次数增加而缓慢较小, 且曲线整体上呈现上升趋势, 并在迭代后期收敛于全局最佳适应度曲线附近, 说明 ALMOA 在迭代前期发挥了较强的全局搜索能力, 在迭代后期发挥了较强的局部搜索能力, 达到了预期的优化效果.

将 MOA 和 ALMOA 的最佳参数寻优结果分别代入 SVM 中, 并对测试集进行识别, 最终得到 MOA-SVM 模型和 ALMOA-SVM 模型的结果混淆矩阵如图 10 所示, 模型的性能评价如表 3 所示. 可见, 采用 MOA-SVM 模型的识别精度为 98.15%, 其预测结果中存在两个误判, 分别将两个 Non-GMO2 样品, 一个误判为 Non-GMO1 样品, 另一个误判为 GMO1 样品, 所得 Non-GMO2 的查全率为 92.59%, Non-GMO1 的查准率为 96.43%, GMO1 的查准率为 96.43%. 采用 ALMOA-SVM 模型的识别精度为 100%, 所有菜籽油样品均被正确识别. 由此可见, ALMOA 有效避免了参数寻优过

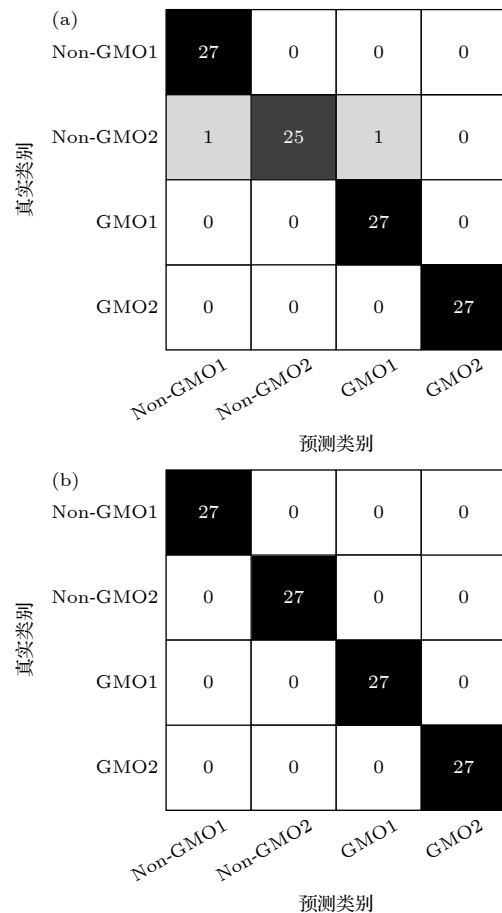


图 10 两种模型的分类结果混淆矩阵 (a) MOA-SVM 模型; (b) ALMOA-SVM 模型

Fig. 10. Confusion matrices of the classification results for the two models: (a) MOA-SVM model; (b) ALMOA-SVM model.

程中陷入局部最优解的情况, 增强了其全局搜索能力, 从而使鉴别模型的分类性能得到了较好提升.

表 3 MOA-SVM 模型与 ALMOA-SVM 模型的性能评价

Table 3. Performance evaluation of the MOA-SVM model and ALMOA-SVM model.

模型	样品	查全率/%	查准率/%	精度/%
MOA-SVM	Non-GMO1	100	96.43	98.15
	Non-GMO2	92.59	100	
	GMO1	100	96.43	
	GMO2	100	100	
ALMOA-SVM	Non-GMO1	100	100	100
	Non-GMO2	100	100	
	GMO1	100	100	
	GMO2	100	100	

5 结 论

本文采用 THz-TDS 技术研究了两种转基因和两种非转基因菜籽油的 THz 光谱, 发现转基因菜籽油相对于非转基因菜籽油在 THz 波段具有更强的吸收特性. 通过对 0.3—1.8 THz 范围内的菜籽油吸光度谱进行主成分分析, 选取累积方差贡献率超过 99.8% 的前 9 个主成分替代原始光谱数据, 降低了数据维度, 提升了后续建模效率. 在 SVM 参数寻优过程中, 针对 MOA 容易陷入局部最优解的问题, 引入自适应惯性权重和 Lévy 飞行两种改进策略, 提出了 ALMOA. 结果表明, 相比于 MOA, ALMOA 在迭代前期具备更强的全局搜索能力, 在迭代后期也具有较为出色的局部搜索能力, 对 SVM 参数的搜索精度更高; 基于本文实验获取的菜籽油吸光度数据集, ALMOA-SVM 模型对 4 种菜籽油的识别精度为 100%, 优于 MOA-SVM 模型获得的 98.15% 的识别精度. 因此, THz-TDS 技术结合 ALMOA-SVM 模型为转基因菜籽油的分类鉴别提

供了一种快速有效的新方法, 同时也为其他转基因物质的检测提供了方法参考.

参考文献

- [1] ISAAA 2021 *China Biotechnol.* **41** 114 (in Chinese) [国际农业生物技术应用服务组织 2021 中国生物工程杂志 **41** 114]
- [2] Kumar K, Gambhir G, Dass A, Tripathi A K, Singh A, Jha A K, Yadava P, Choudhary M, Rakshit S 2020 *Planta* **251** 91
- [3] Demeke T, Dobnik D 2018 *Anal. Bioanal. Chem.* **410** 4039
- [4] Gampala S S, Wulfkuhle B, Richey K A 2019 *Transgenic Plants* **1864** 411
- [5] Peng X Y, Zhou H 2021 *Acta Phys. Sin.* **70** 240701 (in Chinese) [彭晓昱, 周欢 2021 物理学报 **70** 240701]
- [6] Mittleman D M 2017 *J. Appl. Phys.* **122** 230901
- [7] Sun L, Zhao L, Peng R Y 2021 *Mil. Med. Res.* **8** 28
- [8] Hu Y, Wang X H, Guo L T, Zhang C L, Liu H B, Zhang X C 2005 *Acta Phys. Sin.* **54** 4124 (in Chinese) [胡颖, 王晓红, 郭澜涛, 张存林, 刘海波, 张希成 2005 物理学报 **54** 4124]
- [9] Chen T 2016 *Chin. J. Quantum Electron.* **33** 392 (in Chinese) [陈涛 2016 量子电子学报 **33** 392]
- [10] Zhang W T, Li Y W, Zhan P P, Xiong X M 2017 *Infrared Laser Eng.* **46** 1125004 (in Chinese) [张文涛, 李跃文, 占平平, 熊显名 2017 红外与激光工程 **46** 1125004]
- [11] Liu J J 2017 *Microw. Opt. Technol. Lett.* **59** 654
- [12] Liu J J, Fan L L, Liu Y M, Mao L L, Kan J Q 2019 *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **206** 165
- [13] Gu Q H, Chang Y X, Li X H, Chang Z Z, Feng Z D 2021 *Expert Syst. Appl.* **165** 113713
- [14] Guo L, Xu C, Yu T H, Tuerxun W 2022 *IEEE Access* **10** 36335
- [15] Cortes C, Vapnik V 1995 *Mach. Learn.* **20** 273
- [16] Tuerxun W, Xu C, Guo H Y, Jin Z J, Zhou H J 2021 *IEEE Access* **9** 69307
- [17] Zervoudakis K, Tsafarakis S 2020 *Comput. Ind. Eng.* **145** 106559
- [18] Ding Y H, You W B 2020 *IEEE Access* **8** 207089
- [19] Nickabadi A, Ebadzadeh M M, Safabakhsh R 2011 *Appl. Soft Comput.* **11** 3658
- [20] Syama S, Ramprabhakar J, Anand R, Guerrero J M 2023 *Results Eng.* **19** 101274
- [21] Liu N, Luo F, Ding W C 2019 2019 *IEEE Symposium Series on Computational Intelligence (SSCI)* Xiamen, China, December 6–9, 2019 p3104
- [22] Pan P Y, Xing Y H, Zhang D W, Wang J, Liu C L, Wu D, Wang X Y 2023 *J. Food Sci.* **88** 3189
- [23] Elahi N, Duncan R W, Stasolla C 2016 *Plant Physiol. Biochem.* **100** 52

Application of terahertz spectroscopy in identification of transgenic rapeseed oils: A support vector machine model based on modified mayfly optimization algorithm^{*}

Chen Tao[†] Li Xin

(School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin 541004, China)

(Received 27 September 2023; revised manuscript received 22 November 2023)

Abstract

To achieve rapid and accurate identification of genetically modified (GM) and non-GM rapeseed oils, a support vector machine (SVM) model based on an improved mayfly optimization algorithm and coupled with the terahertz time-domain spectroscopy, is proposed. Two types of GM rapeseed oils and two types of non-GM rapeseed oils are selected as research subjects. Their spectral information is acquired by using the terahertz time-domain spectroscopy. The observations show that GM rapeseed oils exhibit stronger terahertz absorption characteristics than non-GM rapeseed oils. However, their absorption spectra are highly similar, making direct differentiation difficult through visual inspection alone. Therefore, SVM is used for spectral recognition. Considering that the classification performance of SVM is significantly affected by its parameters, the mayfly optimization algorithm is combined to optimize these parameters. Furthermore, adaptive inertia weight and Lévy flight strategies are introduced to enhance the global search capability and robustness of the mayfly optimization algorithm, thus addressing the issue of easily becoming trapped in local optima in the optimization process. Moreover, principal component analysis is used to reduce the dimensionality of the absorbance data in a 0.3–1.8 THz range, aiming to extract critical features, thereby enhancing modeling efficiency and reducing redundancy in spectral data. Experimental results demonstrate that the improved mayfly optimization algorithm effectively identifies the optimal parameter combination for SVM, thereby enhancing the overall performance of the identification model. The proposed SVM model, in which the improved mayfly optimization algorithm is used, can achieve a recognition accuracy of 100% for the four types of rapeseed oils, surpassing the 98.15% accuracy achieved by the SVM model with the original mayfly optimization algorithm. Thus, this study presents a rapid and effective new approach for identifying GM rapeseed oils and offers a valuable reference for identifying other genetically modified substances.

Keywords: transgenic rapeseed oil, terahertz spectroscopy, classification discrimination, mayfly optimization algorithm

PACS: 87.50.U, 87.64.–t, 07.57.–c

DOI: [10.7498/aps.73.20231569](https://doi.org/10.7498/aps.73.20231569)

^{*} Project supported by the National Natural Science Foundation of China (Grant Nos. 62261012, 61841502).

[†] Corresponding author. E-mail: tchen@guet.edu.cn



太赫兹光谱在转基因菜籽油鉴别中的应用：基于改进蜉蝣算法的支持向量机模型

陈涛 李欣

Application of terahertz spectroscopy in identification of transgenic rapeseed oils: A support vector machine model based on modified mayfly optimization algorithm

Chen Tao Li Xin

引用信息 Citation: *Acta Physica Sinica*, 73, 058701 (2024) DOI: 10.7498/aps.73.20231569

在线阅读 View online: <https://doi.org/10.7498/aps.73.20231569>

当期内容 View table of contents: <http://wulixb.iphy.ac.cn>

您可能感兴趣的其他文章

Articles you may be interested in

准二维范德瓦耳斯本征铁磁半导体 CrGeTe_3 的THz光谱

Quasi-two-dimensional van der Waals ferromagnetic semiconductor CrGeTe_3 studied by THz spectroscopy

物理学报. 2022, 71(23): 237303 <https://doi.org/10.7498/aps.71.20221586>

激光诱导击穿光谱技术结合神经网络和支持向量机算法的人参产地快速识别研究

Rapid identification of ginseng origin by laser induced breakdown spectroscopy combined with neural network and support vector machine algorithm

物理学报. 2021, 70(4): 040201 <https://doi.org/10.7498/aps.70.20201520>

准二维范德瓦耳斯磁性半导体 CrSiTe_3 的THz光谱

Quasi-two-dimensional van der Waals semiconducting magnet CrSiTe_3 studied by using THz spectroscopy

物理学报. 2020, 69(20): 207302 <https://doi.org/10.7498/aps.69.20200682>

太赫兹时域光谱中脉冲太赫兹波全息探测

Holographic detection of pulsed terahertz waves in terahertz time-domain spectroscopy

物理学报. 2022, 71(18): 188704 <https://doi.org/10.7498/aps.71.20220983>

太赫兹雷达散射截面的仿真与时域光谱测量

Simulations and time-domain spectroscopy measurements for terahertz radar-cross section

物理学报. 2019, 68(16): 168701 <https://doi.org/10.7498/aps.68.20190552>

太赫兹实时近场光谱成像研究

Research on terahertz real-time near-field spectral imaging

物理学报. 2022, 71(16): 164201 <https://doi.org/10.7498/aps.71.20220131>