基于遗传算法的近红外光谱定性分析特征波长提取方法研究

李浩光[1,2]，于云华[1,2]，逄 燕[1]，沈学锋[1,2]

1. 山东石油化工学院机械与控制工程学院，山东 东营 257061
2. 中国石油大学(华东)新能源学院，山东 东营 257061

摘 要 近红外光谱分析技术虽在多领域获得广泛应用，但应用时仍以实验室仪器为主，目前光谱仪存在体积大、功耗高、价格贵等问题，有能力购买与使用此类仪器的主要是高校、科研院所、大型企业等，常用的基于傅里叶变换或光栅原理的光谱仪价格通常高达几十万元，超出中小企业、普通百姓的经济承受能力，因此近红外光谱仪的进一步推广应用仍有难度。降低仪器造价并实现微型化，是推广近红外光谱技术应用的一个重要方向，近红外光谱仪小型化的努力方向有CT正交型光栅技术以及微机电系统技术，但这两种技术方案对光谱仪体积缩小幅度有限，仍存在价格高、内部有移动部件等问题，难以做到真正微型化。据光谱仪的工作原理可知，其价格高低及微型化难度与仪器所能检测波段以及分辨率密切相关，以线性渐变滤光片与InGaAs探测器为例，分辨率越高，检测的波长点越多，其价格越高，制造难度越大。针对某一特定的定性分析任务，若能从大量波长点中挑选出少量特征波长点，并利用挑选得到的少量特征波长点完成对被测样本的定性分析任务，则可降低仪器制造成本，并降低光谱仪微型化难度，从而有利于近红外光谱分析技术的推广与应用。以玉米单倍体和多倍体籽粒作为研究对象，针对两类籽粒分类任务，分多天以漫透射方式采集被研究对象的近红外光谱，按时间顺序将所采数据分为5个数据集，对第1个数据集使用遗传算法提取出10个特征波长点，再将提取得到10个特征波长点，用于剩余4个数据集的单倍体、二倍体鉴别，以检验方法的有效性。实验结果表明使用10个特征波长点能够获得与全光谱基本一致的鉴别效果，说明使用少量特征波长点上的吸光度值也能够有效鉴别单倍体，可为其他领域某特定任务开发低成本便携式微型近红外光谱仪提供借鉴。

关键词 遗传算法；近红外光谱；特征波长；微型近红外光谱仪

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)08-2437-06

引 言

近红外光谱仪制造是我国仪器制造业的短板，目前国产近红外光谱仪多应用于实验室场合[1-2]，体积大、功耗高、价格昂贵、难以二次开发，极大地限制了近红外光谱分析技术的推广应用[3-4]。近红外光谱仪昂贵的价格及其较大体积制约了近红外光谱分析技术的大范围推广应用，究其原因还是近红外光谱仪本身价格昂贵且体积尚未做到便携化、微型化。

MicroNIR-1700光谱仪是美国VIAVI公司生产的一种微型便携式光谱仪，该型光谱仪将近红外光源、分光部件及近红外检测器集成于ϕ45×42 mm的体积之内，且内部不含移动部件，重量仅为60 g左右，是目前世界上体积最小的微型近红外光谱仪，该型光谱仪能实现微型化主要原因是使用了线性渐变滤光片技术及InGaAs探测器(128线元)，目前国内仪器厂商及科研院所研发或生产的相关产品尚不能达到该型光谱仪的技术性能；同时VIAVI公司出于技术保护的目的，不对外出售相关元件，因此国内尚无厂家有能力生产或仿制类似性能的微型光谱仪。据光谱仪的工作原理可知，其价格高低及微型化难度与光谱仪所能检测波段以及分辨率密切相关，以线性渐变滤光片与InGaAs探测器为例，分辨率越高，检测的波长点越多，其价格越高，制造难度越大。若能够借鉴MicroNIR-1700光谱仪的微型化设计思路，并在其基础上设计只需要采集少量波长点吸光度的光谱仪，则分光部件与InGaAs探测器元件成本与制造难度可大幅降低，

收稿日期：2021-01-19，**修订日期**：2021-05-05

基金项目：国家重大科学仪器设备开发专项(2014YQ470377)，山东省教育厅科技计划项目(J18KA329)，东营市科技发展基金项目(DJ2020032)资助

作者简介：李浩光，1981年生，山东石油化工学院副教授 e-mail：lihaoguang@upc.edu.cn

并进一步降低光谱仪整体造价。

针对某一特定的定性分析任务，若能够从大量波长点中挑选出少量的特征波长点，并利用挑选得到的少量特征波长点完成对被测样本的定性分析任务，则可以降低仪器制造成本，并降低光谱仪微型化的难度，从而有利于近红外光谱分析技术的大面积推广与应用。

以玉米单倍体和多倍体籽粒作为研究对象，针对二类籽粒分类任务，多天以漫透射方式采集被研究对象的近红外光谱，按时间顺序将所采数据分为 5 个数据集，对第 1 个数据集使用遗传算法提取出 10 个特征波长点，再将提取得到 10 个特征波长点，用于剩余 4 个数据集的单倍体二倍体鉴别，以检验方法的有效性。实验结果表明使用 10 个特征波长点能够获得与全光谱基本一致的鉴别效果，说明使用少量特征波长点上的吸光度值也能够有效鉴别单倍体，针对玉米单倍体鉴别这一特定任务，可以降低仪器制造的成本与微型化的难度，为加快近红外光谱定性分析技术在单倍体育种行业的应用提供技术基础。虽然研究对象为玉米籽粒，但是方法思路亦可推广至其他被测对象，可为其他领域某个特定任务开发低成本便携式微型近红外光谱仪提供借鉴。

1 基于遗传算法的特征波长点选择方法

1.1 算法原理

在近红外光谱定性分析[5-7]问题中，将遗传算法(genetic algorithms，GA)与某一分类算法结合搜寻最有利于分类任务的原始光谱特征波长点子集，即可构成基于遗传算法的近红外光谱特征波长点选择方法。

遗传算法模仿自然界中生物进化过程，在算法中包含了繁殖、交叉、变异等生物进化过程中的重要步骤。

在实际自然界生物进化过程中，生物染色体具有一定变异概率，上一代染色体在繁殖时，染色体会相互交叉并遗传给下一代，生物进化时总是选择并保留最能适应其生活环境的遗传基因，而遗传算法则是选择最有利于分类的特征子集，且在每一代分类时都选择一组当前最优的特征子集，进行循环的变异、繁殖、交叉、分类等步骤，直至满足算法设定条件。

遗传算法首先将需要特征筛选的向量编码成一条染色体，对于特征选择，其目标是从 D 个特征中挑选 d 个特征，首先需将全部特征表示成一个由 D 个二进制代码构成的字符串，二进制字符串中的 0 表示该维对应特征未被选择，1 则表示该维对应特征被选择，该字符串就代表遗传算法中的染色体，可将其用 m 来表示。若在 D 维特征中选择 d 维全是 1 的有效特征，则存在种组合。

对于分类任务，遗传算法最优目标就是选择最适合分类的特征子集，因此分类器的鉴别准确率就可作为遗传算法的适应度函数值，对于算法中每个迭代步骤中的若干条染色体，每一条染色体即对应一个适应度值，即分类器鉴别准确率。

若待挑选的特征波长点为 n 个，基于遗传算法的特征选择方法可由以下几个步骤实现：

(1)对所有特征是否被选择使用二进制编码：采用 0 和 1 对本节中数据中的全光谱所有波长点进行编码，每一个波长点对应染色体中的一个基因。若编码为 1 则表示该波长点被选中。若编码为 0 则表示该波长点未被选中，一种 0 和 1 编码组合即可当作一条染色体。

(2)初始化染色体种群：染色体种群规模设定为 N，采用随机初始化的方式，产生 N 个编码长度为 n 的染色体，设定迭代次数为 100。

(3)解码并以 SVM 分类器的分类准确率作为适应度函数：将染色体解码，并采用 SVM 方法进行鉴别。使用交叉验证的方法，计算每一个染色体对应的平均正确识别率与平均正确拒识率。

(4)计算适应度函数的值：正确识别率与正确拒识率的均值越高，则适应度越高。考虑到收敛速度，将适应度函数设为正确识别率与正确拒识率的均值。

(5)使用选择、交叉、变异操作繁殖下一代染色体：采用“轮盘赌”选择法，按设定交叉概率对染色体进行交叉，采用精英主义策略，只留下最优值，并按设定变异概率进行变异。

(6)将步骤(5)中的新一代染色体代入步骤(3)，重复步骤(3)—(5)，直到满足收敛条件。

1.2 算法设计

图 1 是采用基于遗传算法的选择波长点方法示意图。首先采用遗传算法与分类算法结合从原始光谱中提取最有利于分类的少量特征波长点，再利用少量特征波长点吸光度对待鉴别光谱类别进行鉴定。为对提出的基于遗传算法的特征波长点方法进行优化设计，选择如下近红外光谱数据集作为实验数据集：

以中国农业大学国家玉米改良中心提供的某品种玉米单倍体和二倍体籽粒作为研究对象，分 5 日连续采集其近红外光谱，使用自制近红外光谱采集装置，并以漫透射采集方式交替采集单倍体、二倍体单籽粒近红外光谱各 100 条，共 5 组数据，5 个实验数据集按时间顺序依次编号为 T1—T5。

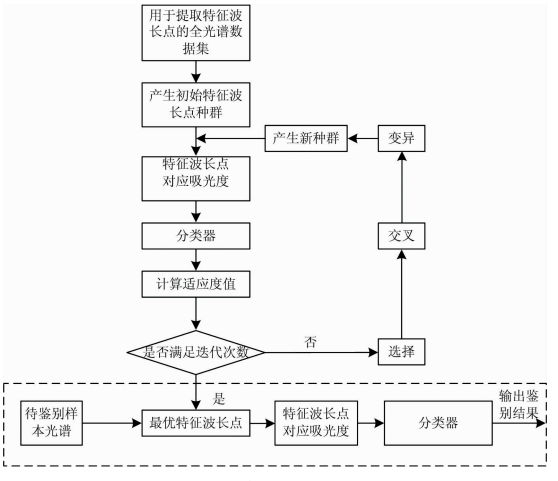

图 1 基于遗传算法的选择波长点方法

Fig. 1 Selection of wavelength points based on genetic algorithms

遗传算法中的适应度函数值使用 SVM 分类器所得的被测光谱正确识别率与正确拒识率[9-12]的均值来衡量。

SVM 使用 LIBSVM 工具箱，设置 SVM 分类器[8]类型为二分类类型，以最优识别率为标准，在高斯核参数 σ 及正则化参数 C 指数增长的过程中以网格的方式搜索最优高斯核参数 σ 及正则化参数 C，高斯核参数 $\sigma=3.2$，正则化参数 $C=0.56$。

基于遗传算法的特征选择方法中种群规模、交叉率、变异率三个参数对识别性能、收敛速度具有明显影响。为确定适合本任务的种群规模、交叉率、变异率，以单倍体二倍体籽粒鉴别任务为例对三个参数进行如下分析研究。

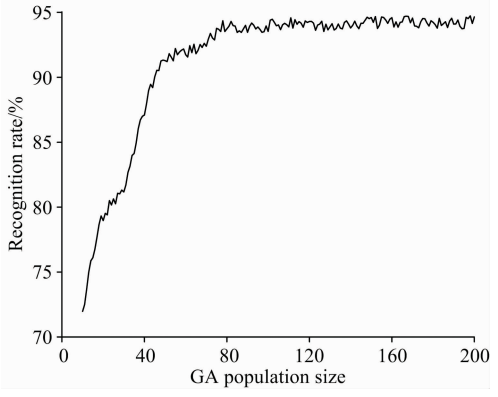

图 2　识别率随种群规模变化曲线图

Fig. 2　Recognition rate as population size increasing

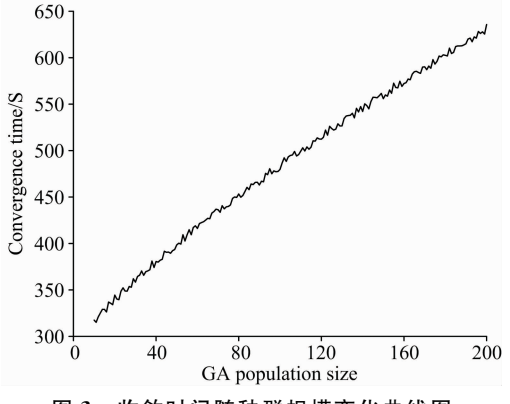

图 3　收敛时间随种群规模变化曲线图

Fig. 3　Curve of convergence time as population size increasing

由图 2 与图 3 可知，随着遗传算法中群体规模增大，迭代次数及迭代时间显著变化，分类准确率随种群规模增大首先出现上升趋势，当种群规模达到 80 时分类准确率趋于平缓，而程序运行时间始终是直线上升趋势。

由此可见，识别率满足条件时，增大种群规模需花费较大计算代价，因此遗传算法种群规模设置不宜过大。综合考虑特征波长点的分类识别性能与算法收敛时间，利用遗传算法挑选特征波长点时，设置遗传算法种群规模为 80。

由图 4—图 7 分析可知，随交叉率与变异率增大，识别率到达一定值以后，其变化趋势趋于稳定，增长趋势并不明显，与此同时，收敛时间却呈现线性增长趋势，分析认为随

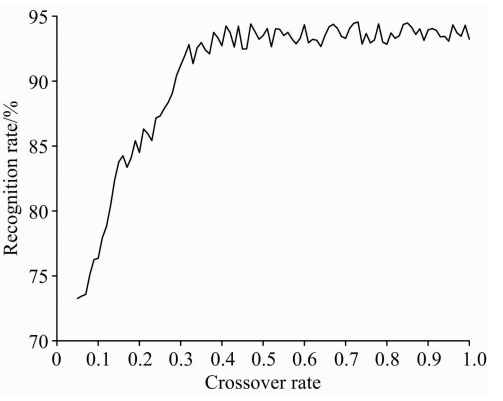

图 4　识别率随交叉率变化曲线图

Fig. 4　Curve of recognition rate as crossing rate changing

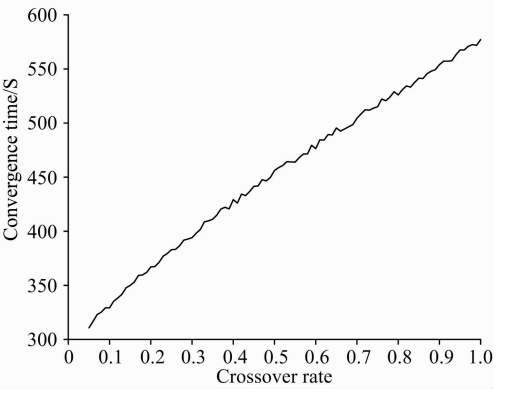

图 5　收敛时间随交叉率变化曲线图

Fig. 5　Curve of convergence time as crossover rate changing

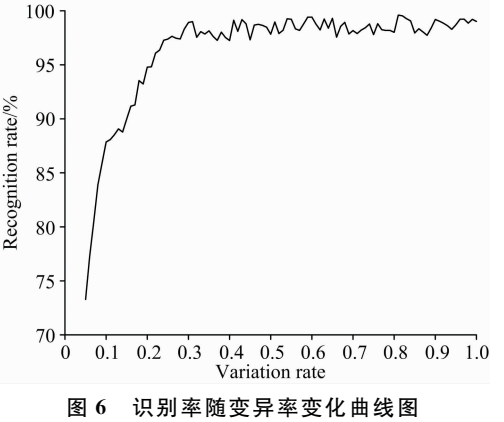

图 6　识别率随变异率变化曲线图

Fig. 6　Curve of recognition rate as variation rate changing

着遗传算法中交叉率越大，遗传种群中产生新模式的概率相应增大，在开始阶段有时能够扩展至整个编码空间，但原有模式被破坏的可能性也随之增大，而交叉率过小导致每一步搜索空间过小，导致算法难以收敛。

相比于交叉率，变异能够提高算法所得解的多样性，但变异率较大时，易导致遗传算法变为随机搜索，变异率设置过小，种群易出现早熟或易陷入局部最优解。综合考虑所选特征建立模型的识别效果与收敛时间，使用遗传算法方法进

行特征波长点选择时，为使所获特征波长点具有最优分类性能，交叉率设置为 0.5，变异率设置为 0.3。

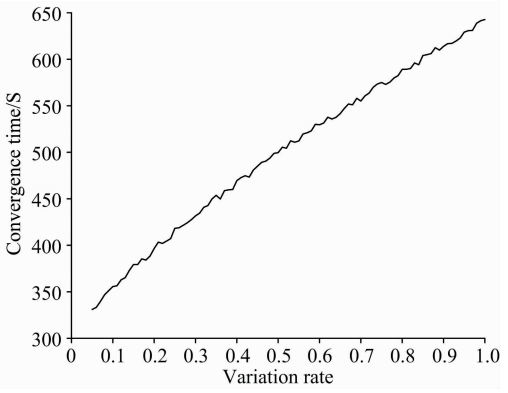

图 7 收敛时间随变异率变化曲线

Fig. 7 Curve of convergence time as variation rate changing

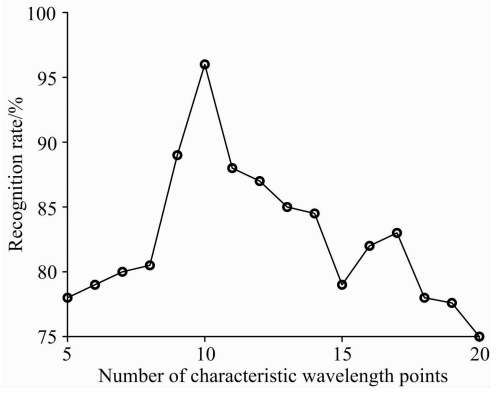

图 8 识别率随特征波长点个数变化曲线

Fig. 8 Curve of recognition rate when characteristic wavelength points increasing

图 8 是单倍体二倍体籽粒的识别率随特征波长点个数变化曲线，由该曲线可以看出，在特征波长点数目增长过程中，识别率表现出先升后降的趋势，特征波长点增加到 10 时，识别率达到最高，其后随着特征波长点数增加，识别率又出现下降趋势，说明选择 10 个特征波长点能够获得最优分类效果。特征波长点过少，模型出现欠拟合的情况，而特征波长点过多时，一方面会增加模型训练时间，另一方面易导致模型过拟合，因此以下实验选择挑选 10 个特征波长点进行实验。

以 T1—T5 作为实验数据集，利用 SVM 分类器的准确率作为 GA 算法适应度函数，得到 10 个特征波长点如表 1 所示。

表 1 特征波长点列表(nm)

Table 1 Characteristic wavelength point list (nm)

1	2	3	4	5	6	7	8	9	10
970	1 100	1 199	1 217.8	1 279.8	1 304.5	1 329.3	1 335.5	1 515.1	1 639

表 1 中为使用遗传算法对 T1 数据集进行特征选择得到的 10 个特征波长点。由此可知，针对数据集 T1 中的单倍体二倍体光谱数据，表 1 中的 10 个特征波长点最能够反映被测品种玉米籽粒单倍体与二倍体之间的差异信息。

2 结果与讨论

为验证所挑选的 10 个特征波长点用于近红外光谱定性分析的可行性，进行如下实验：

2.1 实验 1

利用挑选得到 10 个特征波长点，分别从 T1—T5 等 5 个实验集单倍体、二倍体数据中各随机抽取 50 条光谱建模，将剩余光谱作为测试集进行测试，共实验 20 次，识别率取平均，如表 2 和表 3 所示。

全光谱方式采用平滑(平滑窗口 9)、一阶导(9)、归一化、PLS(11)及 LDA(4)降维后，再使用 SVM 进行分类。

表 2 各数据集识别结果表

Table 2 Recognition results (Characteristic wavelength point)

测试集	T1	T2	T3	T4	T5
正确识别率/%	94.8	94.4	95.6	91.2	91.8
正确拒识率/%	94.6	93.2	98.0	93.2	93.2
平均识别率/%	94.7	93.8	96.8	92.2	92.5

表 3 全光谱各数据集识别结果表

Table 3 Recognition results (whole spectra)

测试集	T1	T2	T3	T4	T5
正确识别率/%	95.0	95.0	95.0	90.0	100.0
正确拒识率/%	95.0	95.0	100.0	95.0	90.0
平均识别率/%	95.0	95.0	97.5	92.5	95.0

对比表 2 和表 3 发现，在各个独立的测试数据集中，特征波长点与全光谱两种方式下单倍体及二倍体的平均识别率基本接近，具体分析如下：

(1)10 个特征波长点方式：在 T1—T5 数据集上，特征波长点方法所得平均识别率在 92.2%～96.8%之间。

(2)全光谱方式：在 T1—T5 数据集上，平均识别率在 92.5%～97.5%之间。

由此可见，利用 10 个特征波长点的方式时，单倍体识别率相对于全光谱方式只略下降。说明采用特征波长点方式时，利用当天光谱数据进行训练，当天数据作为测试集进行测试，实验结果与全谱区识别性能相差不大，证明了使用特征波长点方法能够在当日数据集上取得较高的识别效果。

2.2 实验 2

利用挑选出的 10 个特征波长点，以 T1 作为训练集，利用 T2—T5 等 4 个实验数据集进行测试，检验特征波长点方式在多个测试集中的泛化能力。

由表 4 和表 5 可知，利用 T1 实验集的数据进行训练，T2—T4 数据集作为测试集测试时，10 个特征波长点方式所得的识别率与全谱区方式所得的识别率也非常接近，具体分

表 4　各数据集识别结果表(特征波长点)

Table 4　Recognition results (Characteristic wavelength point)

测试集	T2	T3	T4	T5
正确识别率/%	96.0	94.0	96.0	93.0
正确拒识率/%	96.0	94.0	91.0	92.0
平均识别率/%	96.0	94.0	94.0	92.5

表 5　各数据集识别结果表(全光谱)

Table 5　Recognition results (whole spectra)

测试集	T2	T3	T4	T5
正确识别率/%	95.0	95.0	96.0	93.0
正确拒识率/%	98.0	94.0	92.0	94.0
平均识别率/%	96.5	94.5	94.0	93.5

析如下：

(1)以 10 个特征波长点方式在 T2—T5 四个数据集上进行测试，所得的平均识别率在 92.5%～96.0%之间。

(2)以全光谱方式在 T2—T5 数据集上测试，所得的平均识别率在 93.5%～96.5%之间。

由此可见，特征波长点方式所建立的定性分析模型与全谱区方式所建立的定性分析模型性能基本接近，在不同数据集上都具有较强泛化能力。

以玉米单倍体和二倍体籽粒作为研究对象，针对两类籽粒分类任务，分多天以漫透射方式采集研究对象的近红外光谱，按时间顺序将所采数据分为 5 个数据集，对第 1 个数据集使用遗传算法提取出少量特征波长点，再将提取得到少量特征波长点，用于剩余 4 个数据集的单倍体二倍体鉴别，以检验方法的有效性。实验结果表明使用少量特征波长点能够获得与全光谱基本一致的鉴别效果，说明使用少量特征波长点上的吸光度值也能够有效鉴别玉米单倍体。

3　结　论

针对某一特定的定性分析任务，若能够从大量波长点中挑选出少量的特征波长点，并利用挑选得到的少量特征波长点完成对被测样本的定性分析任务，则可以降低仪器制造成本，并降低光谱仪微型化的难度，从而有利于近红外光谱分析技术的大面积推广与应用。本文研究了基于遗传算法的特征波长点选择方法，采用遗传算法与分类算法结合的特征波长点选择方法，以玉米籽粒为研究对象，从原始光谱中提取最有利于单倍体二倍体分类的十个特征波长点，再利用十个特征波长点的吸光度对多个测试集中的单倍体与二倍体籽粒进行分类，并将特征波长点方法与全光谱进行了对比实验，本研究可以为针对某一特定应用场景的近红外光谱仪小型化和简单化提供理论依据，虽然所研究对象为玉米籽粒，但是方法思路亦可推广至其他被检测对象，可以为其他领域某个特定任务开发低成本便携式微型近红外光谱仪提供借鉴。

References

[1] Chu X L, Shi Y Y, Chen B, et al. Journal of Instrumental Analysis, 2019, 38(5): 603.

[2] Yu F, Wen Q, Lei H J, et al. Laser & Optoelectronics Progress, 2018, 55(10): 30.

[3] Wang S H, Zhang X, Zhang G W, et al. Infrared Technology, 2020, 42(7): 688.

[4] Huang Y W, Li H, Wang R L. Cereals & Oils, 2017, 30(7): 1.

[5] Miao X X, Miao Y, Gong H R, et al. Food Science and Technology, 2019, 44(10): 335.

[6] LIU Shuang, YU Hai-ye, PIAO Zhao-jia, et al(刘　爽，于海业，朴兆佳). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(11): 3542.

[7] Chen T T, Wang K J, Han X Z, et al. Food and Machinery, 2020, 36(11): 46.

[8] YAN Yan-lu, CHEN Bin, ZHU Da-zhou(严衍禄，陈　斌，朱大洲). Near Infrared Spectroscopy Analytical-Principles, Technology and Application(近红外光谱分析原理、技术与应用). Beijing: China Light Industry Press(北京：中国轻工业出版社), 2007.

[9] Karacaglar N N Y, Bulat T, Boyaci I H, et al. Journal of Food and Drug Analysis, 2019, 27(1): 101.

[10] Qin H, Ma J Y, Chen S J, et al. Infrared Technology, 2015, 1(37): 78.

[11] LI Hao-guang, LI Wei-jun, QIN Hong, et al(李浩光，李卫军，覃　鸿，等). Transactions of the Chinese Society of Agricultural Machinery(农业机械学报), 2016, 47(6): 259.

[12] LI Hao-guang, LI Wei-jun, QIN Hong, et al(李浩光，李卫军，覃　鸿，等). Transactions of the Chinese Society of Agricultural Machinery(农业机械学报), 2017, (S1): 422.

Study on Characteristic Wavelength Extraction Method for Near Infrared Spectroscopy Identification Based on Genetic Algorithm

LI Hao-guang[1, 2], YU Yun-hua[1, 2], PANG Yan[1], SHEN Xue-feng[1, 2]

1. College of Mechanical and Control Engineering, Shandong Institute of Petrochemical and Chemical Technology, Dongying 257061, China
2. New Energy College, China University of Petroleum (East China), Dongying 257061, China

Abstract At present, although the near-infrared (NIR) spectroscopy analysis technology has been widely used in many fields, it is mainly used as laboratory instruments, and the spectrometer used in the laboratory has the problems of large volume, high power consumption and high price. The main units that can purchase and use the NIR spectrometer are universities, scientific research institutes, large and medium-sized enterprises, etc. The price of a spectrometer based on the Fourier or grating principle is usually as high as several hundred thousand Yuan, which is beyond the affordability of small and medium-sized enterprises and ordinary people. Therefore, the application of NIR spectrometer is far away from ordinary people's daily life. The high price and large volume of near-infrared spectrometers restrict the large-scale application of near-infrared spectroscopy analysis technology. The reason is that the near-infrared spectrometer itself is expensive and the volume has not yet been portable and miniaturized. Reducing the cost of the NIR spectrometer and miniaturizing the spectrometer is an important direction to promote NIR spectroscopy technology. The efforts of miniaturization of NIR spectrometer include CT orthogonal grating technology and micro electro mechanical system technology. However, the volume reduction of the spectrometer by these two technical solutions is limited, and there are still some problems, such as high price, internal moving parts and real hard miniaturization. For a specific qualitative analysis task, a small number of characteristic wavelength points are selected from full spectra and used to build models which can recognize testes samples. The method mentioned above can reduce the cost of instrument manufacturing and difficulty of spectrometer miniaturization, and it is also conducive to the large-scale promotion and application of NIR analysis technology. Near infrared spectra of Haploid and diploid maize seeds are collected by diffuse transmission method in several days. The collected data are divided into five data sets in chronological order. For the first data set, 10 characteristic wavelength points are extracted by genetic algorithm, and then 10 characteristic wavelength points are extracted for the remaining four data sets In order to test the validity of the method, the haploid and diploid identification was carried out. The experimental results show that using 10 characteristic wavelength points can obtain the identification effect, which is consistent with the full spectrum, indicating that using a small number of characteristic wavelength points can also effectively identify haploids, which can provide a reference for the development of low-cost portable NIR spectrometer for a specific task in other fields.

Keywords Genetic algorithm; Near infrared spectroscopy; Characteristic wavelength; Micro near infrared spectroscopy

(Received Jan. 19, 2021; accepted May 5, 2021)