第42卷，第10期
2022年10月

光谱学与光谱分析
Spectroscopy and Spectral Analysis

Vol. 42, No. 10, pp3135-3142
October, 2022

耦合平均影响值-连续投影算法优化种子活力近红外检测模型

杨冬风[1]，李爱传[1]，刘金明[1]，陈争光[1]，时　闯[1]，胡　军[2*]

1. 黑龙江八一农垦大学信息与电气工程学院，黑龙江 大庆　163319
2. 黑龙江八一农垦大学工程学院，黑龙江 大庆　163319

摘　要　目前，近红外光谱(NIRS)可以实现种子活力的快速、无损检测，但区分的活力等级一般少于3级且精度不高。建立种子活力多等级、高精度的NIRS检测模型，解决活力等级增加与预测模型精度之间的矛盾是现阶段近红外种子活力检测的主要任务。以玉米种子为研究对象，采用人工老化的方法获得5种活力等级的种子样本并采集对应的光谱数据建立反向神经网络(BP)预测模型。为了提高模型的精度和稳健性，提出一种耦合平均影响值-连续投影特征波长提取算法(MIVopt-SPAsa)。该算法针对连续投影算法(SPA)耗时过长的问题，采用平均影响值算法(MIV)对其预降维。MIV方法实现了对波长影响值的排序，但缺乏选取波长影响阈值的指标，因此引入相对距离比对MIV算法进行优化(MIV_{opt})，实现特征波长范围的有效分割。针对SPA提取特征变量数目确定的问题，设定了特征波长数目范围并在此范围内优中选优，实现了自适应的SPA(SPA_{sa})特征提取。使用耦合MIV_{opt}-SPA_{sa}算法对具有1 845个波长的玉米种子近红外全谱数据进行特征提取，提取出特征波长37个，主要分布在玉米种子近红外光谱的7个主要吸收峰附近，表明该算法可以有效提取出与玉米种子生化物质近红外吸收特性一致的特征波长。为了测试该算法对模型性能的影响，建立了全谱BP模型、MIV-BP模型、SPA_{sa}-BP模型、MIV_{opt}-SPA_{sa}-BP模型和竞争自适应重加权CARS-BP模型对5个等级的玉米种子活力进行分级，MIV_{opt}-SPA_{sa}-BP模型的预测平均准确率可达99.1%，预测精度高于其他模型；其计算平均时间为14.382 s，低于MIV-BP模型的计算时间(24.523 s)、CARS-BP模型的计算时间(97.226 s)和SPA_{sa}-BP模型的计算时间(101.224 s)，但高于全谱模型的平均计算时间(0.253 1 s)；其最佳表现交叉熵为0.007 892，远远低于另外4个模型。实验结果表明：MIV_{opt}-SPA_{sa}算法可以有效地提高玉米种子活力近红外检测模型的精度，实现种子活力多等级、精确、无损检测，为种子活力检测模型的优化提供参考。

关键词　近红外光谱；种子活力；玉米；平均影响值算法；连续投影算法

中图分类号：O657.33　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2022)10-3135-08

引　言

种子活力是指种子的潜在发芽能力或者种胚所具有的生命力，是预期种子具有长成正常幼苗的潜在能力。种子活力水平测定在育种、种子生产、种子加工、种子收购、种子贮藏、种子检验及种子调运等环节中是不可缺少的重要方法[1]。国际种子检验协会规定的常规种子活力测定方法主要有标准发芽试验、四唑染色试验、离体胚测定法、电导率测定法等。上述方法不仅检测周期长、操作步骤复杂，而且都是有损检测。

近红外光谱(near infrared spectroscopy，NIRS)涵盖了有机分子的倍频与合频的吸收光谱，能够反映分子的结构、组成和状态信息。随着NIRS技术在农业领域研究的不断深入，NIRS技术也开始在种子活力无损检测中崭露头角。Maythem等[2]采用偏最小二乘(partial least squares，PLS)建立大豆种子活力的等级预测模型，对于两种等级(高、低)活力的预测准确率在85.7%～89.7%之间，对三种等级(高、中、低)活力预测时，不能正确区分高活力和中等活力种子。He等[3]采用极限学习机(extreme learning machine，ELM)

收稿日期：2021-07-29，**修订日期**：2021-10-25

基金项目：国家重点研发计划项目(2018YFE0206300)，黑龙江省自然科学基金项目(C2018050)，大庆市科技局科技项目(zd-2019-38)，黑龙江省省属高校基本科研业务费科研项目(ZRCPY201914)资助

作者简介：杨冬风，女，1977年生，黑龙江八一农垦大学信息与电气工程学院副教授　　e-mail：yangsansun@sina.com

*通讯作者　　e-mail：hj_1977@sohu.com

模型对三种不同年份的水稻种子进行鉴别，鉴别精度较高；采用连续投影算法(successive projection algorithm，SPA)结合支持向量机(support vector machine，SVM)从不同年份的可存活种子中鉴定出不可存活的种子，其分类准确率达94.38%。李武等[4]利用前向间隔偏最小二乘(forward interval partial least squares，FiPLS)、竞争自适应重加权(competitive adaptive reweighted sampling，CARS)、无信息变量消除(uninformative variable elimination，UVE)等变量筛选方法对甜玉米种子的近红外光谱进行特征波长区域选择，采用 PLS 建立发芽率、发芽指数和活力指数的预测模型，取得了较好的预测效果。金文玲等[5]利用主成分分析(principal component analysis，PCA)结合 PLS 建立带稃壳水稻种子的近红外超连续激光光谱的预测模型，对 3 种不同年份的水稻种子进行分类，训练集和预测集的准确率分别为 94.44%和 95.92%。

从以上研究中可以看出，NIRS 技术在种子活力检测方面是有效的，可以对 3 种活力等级的种子进行较为准确的区分。采用的建模方法主要是适合于线性预测的 PLS 和适合于小样本分类的 SVM，而适合于非线性大样本建模的 BP 神经网络用的不多。从目前的文献来看，种子活力检测所采用的特征波长优选方法只有少数的几种常规算法，譬如 SPA，CARS 以及 UVE 等，或几种方法的简单组合。以上两点使得目前的种子活力检测陷于等级少(3 个等级以下)、检测精度不够高的状况。

在 NIRS 分析领域，特征波长优选和预测方法的确定始终是决定模型优劣的关键。针对不同研究对象，学者们采用不同的波长选择方法[6-8]、多种选择方法组合[9-11]、改进的选择方法[12-13]以及不同的预测方法[14-17]以增强模型的鲁棒性和准确性。为了构建多等级、高精度的种子活力检测模型，首先对 SPA 算法加以改进得到自适应 SPA(SPA_{sa})，然后对在 BP 神经网络中评价输入变量对结果影响较为有效的指标—平均影响值(mean impact value，MIV)加以优化得到 MIV_{opt} 方法，将 MIV_{opt} 与 SPA_{sa} 算法进行耦合，建立既适合线性模型又适合非线性模型的特征波长提取方法 MIV_{opt}-SPA_{sa}。然后建立全谱、MIV、SPA_{sa}、MIV_{opt}-SPA_{sa} 和 CARS 的 BP 预测模型并比较模型的预测精度和效率，以证明 MIV_{opt}-SPA_{sa} 算法优化近红外种子活力检测模型的有效性。

1 MIV_{opt}-SPA_{sa} 算法原理

1.1 SPA 算法

SPA 算法[18]是一种基于变量信息的变量降维技术，它利用向量的投影分析来寻找含有最低冗余信息的变量组合，能够有效地消除光谱波长共线性、奇异性和不稳定性的影响，使向量间的共线性达到最小，减少建模所用变量的个数，降低模型复杂度。对于光谱矩阵 $\boldsymbol{X}_{n\times m}$($n$ 为样本数，m 为光谱变量数)，首先设定待选特征变量个数 H，然后执行以下步骤：

(1)初始迭代 $t=1$ 时，在光谱矩阵中任选一列向量 x_j，记为 $x_{k(0)}$，$k(0)$为所选变量的最初位置($j=k(0)$，$1\leqslant j\leqslant m$)，则其他剩余变量位置的集合定义为 s

$$s=\{j,1\leqslant j\leqslant m,j\notin\{k(0),\cdots,k(H-1)\}\} \quad (1)$$

(2)计算剩余列向量 x_j($j\in s$)在所选向量 $x_{k(t-1)}$ 构成的正交向量空间中的投影

$$P=\boldsymbol{I}-\frac{x_{k(t-1)}(x_{k(t-1)})^{\mathrm{T}}}{(x_{k(t-1)})^{\mathrm{T}}x_{k(t-1)}} \quad (2)$$

$$x_j=Px_j \quad (3)$$

其中，$\boldsymbol{I}$ 为单位矩阵；P 为投影算子。

(3)提取最大投影值变量 $\arg[\max(\|Px_j\|)]$，$j\in s$，加入所选变量集。

(4) $t=t+1$，如果 $t<H$，则返回步骤(2)循环计算。

当循环终止时，得到的变量集合$\{x_{k(0)},x_{k(1)},\cdots,x_{k(H)}\}$就是选取的特征波长集合。因为迭代的第一个变量 $x_{k(0)}$ 是随机选取的，因此令光谱中的每个波长都作一次初始变量，进行上述迭代，每次迭代选取 H 个变量，即可得到 $n\times H$ 维矩阵 $\boldsymbol{X}=\{X_1,X_2,\cdots,X_n\}^{\mathrm{T}}$，此矩阵为基于 n 个初始变量的迭代所选取的 n 个候选变量集。然后对每个变量集进行 PLS 交叉验证，得到交叉验证均方根误差 $RMSECV_j$($1\leqslant j\leqslant n$)，取最小的 RMSECV 所对应的 $k(0)$和所选出的变量组合，即为最终筛选出的最优组合。

1.2 改进的 SPA 算法(SPA_{sa})

SPA 方法在特征波长选择方面具有一定的优势，但存在两方面的不足：一是候选变量个数的确定没有标准，H 过小会造成所选变量不能包含光谱中的大部分关键信息，而由于变量之间的共线性，H 又不能设置过大以至于超过独立变量的个数，因此需要寻找取得最优 H 值的方法；二是将每一个光谱变量都作为初始变量进行迭代，得到备选变量组合，之后要对每组变量进行 PLS 交叉验证，当光谱变量较多时，算法的效率不高，需要寻找合适的降维方法，在 SPA 之前对光谱数据进行预降维。针对 H 确定的问题，采取的改进方法如下：

(1)令 H 在某个合理的范围$[M,N]$内变化，得到 $N-M$ 组最小 $RMSECV_i$($1\leqslant i\leqslant N-M$)和对应的变量组合。

(2)从中选取最小 $RMSECV_i$ 所对应的 H 以及变量组合。

这样获取的 H 值具有区域范围的最优性，称之为自适应(Self Adaption)SPA，简称 SPA_{sa}。

1.3 优化的 MIV 方法(MIV_{opt})

针对 SPA 预降维的问题，提出一种改进的平均影响值(MIV)方法。MIV 方法[19]最初由 Dombi 提出，用于表示神经网络中权重矩阵的变化，且可用于评估输入变量对神经网络模型性能的影响，通常应用于质谱分析、生物医学中[20-21]。本研究中的最终建模方法采用 BP 神经网络，使用 MIV 方法可以精确给出波长变量对种子活力等级影响程度的排序，然后使用 SPA 方法从排序靠前的波长中进一步进行优选，达到提高算法效率的目的。MIV 的具体步骤如下，对于光谱矩阵 $\boldsymbol{X}_{n\times m}$($n$ 为样本数，m 为光谱变量数)有：

(1)首先采用全谱 BP 建模，训练出预测准确率超过90%的模型。

(2)将训练集 T 第 i 个样本的第 j 个波长的吸光度分别

增加和减少 10%，得到新的样本集 T_{j+}^{i} 和 T_{j-}^{i}。

(3)将 T_{j+}^{i} 和 T_{j-}^{i} 作为全谱模型的输入进行预测，得到的平均预测值代入式(4)，计算出样本变动对输出的影响变化值 IV_j^i(Impact Value)。

$$IV_j^i = \mathrm{BP}_{\mathrm{full}}(T_{j+}^{i}) - \mathrm{BP}_{\mathrm{full}}(T_{j-}^{i}) \tag{4}$$

(4)对各样本所得的 IV_j^i 求平均，即为各波长的平均影响值 MIV_j，如式(5)所示

$$MIV_j = \frac{1}{N}\sum_{I=1}^{n} | IV_j^i | \tag{5}$$

(5)对得到的平均影响值进行排序。

MIV 方法可以得到波长变量的平均影响值排序，但如何确定序列前多少个波长作为有效波长不同文献的方法不一。李大虎等[22]采用相对贡献率来表征某个因素的 MIV 值对于全部因素 MIV 总和的百分比。在其研究中，全部因素只有 9 个，以相对贡献率超过 10%作为特征筛选的依据可以有效地实现 BP 神经网络输入特征的筛选。但对于数目众多的近红外全谱数据来说，每个波长的相对贡献率很小且相互之间的差异不大，以相对贡献率作为筛选指标不够妥当。因此，提出相对距离比 δ_j 作为选择标准

$$\delta_j = \frac{2\times| MIV_j - \overline{MVI} |}{MIV_{\max} - MIV_{\min}} \tag{6}$$

当 δ_j 大于某个数值时的变量即选入备选变量集。其中，$\overline{MIV}$为全部波长的平均 MIV，$MIV_{\max}$ 为最大 MIV，$MIV_{\min}$ 为最小 MIV。

1.4　耦合 $\mathrm{MIV_{opt}}$-$\mathrm{SPA_{sa}}$ 算法

将优化的 MIV 算法 $\mathrm{MIV_{opt}}$ 对全谱 BP 模型进行平均影响值排序及选择，得到的光谱变量作为 $\mathrm{SPA_{sa}}$ 算法的输入，由此降低 SPA 算法的循环次数，提高波长选择的效率。耦合 $\mathrm{MIV_{opt}}$-$\mathrm{SPA_{sa}}$ 算法是一种基于变量信息的波长选择方法，在 $\mathrm{MIV_{opt}}$ 阶段，将非线性的 BP 模型预测用于平均影响值排序，剔除了非线性模型无关的波长变量；在 $\mathrm{SPA_{sa}}$ 阶段，使用相对全谱变量较少的波长变量，并采用线性的 PLS 作为交叉验证，剔除了与线性模型无关的波长变量，并进行自适应的筛选，得到最优的波长变量组合。

2　实验部分

2.1　材料与仪器

实验用种子样本购自大庆市萨尔图萨中种子公司，为黑龙江省农垦科学院作物所玉米育种研究室杂交培育的垦粘一号玉米品种。实验用仪器如图 1 所示。(a)是近红外光谱采集仪器，为德国 Bruker 公司 Tango 近红外光谱仪，采用积分球漫反射测量方式，分辨率为 8 $\mathrm{cm^{-1}}$，样品和背景的扫描时间均为 32 s，谱区范围 11 550～3 950 $\mathrm{cm^{-1}}$，每条光谱采集的数据点数为 1 845 个；(b)是科文 KW-TH 型种子老化(恒温恒湿)实验箱；(c)是上海菁华公司 JA2003N 高精度电子天平，精确到 1 mg；(d)是近红外光谱仪配套的 IN312-SHD0 型量杯。光谱分析及建模采用的软件主要使用挪威 CAMO 公司的 UnscrambX10.3 和美国 MathWorks 公司的 Matlab R2020。

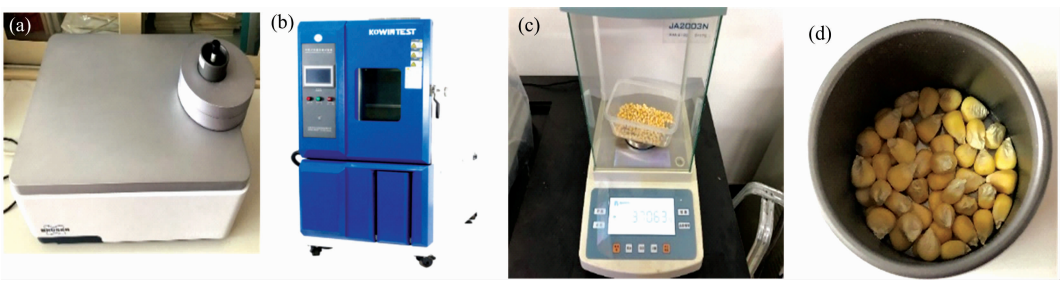

图 1　近红外光谱仪(a)，老化实验箱(b)，电子秤(c)和 IN312-SHD0 量杯(d)

Fig. 1　Near infrared spectrometer (a), Aging test box (b), Electronic scale (c) and IN312-SHD0 measuring cup (d)

2.2　玉米种子老化实验

种子在自然条件下的贮藏时间越长，种子的活力和生活力下降的越快。研究表明，人工加速老化与自然老化对种子内部物质含量及结构的影响差异不大，且发芽情况相近。将种子置于干燥(湿度低于 10%)、低温(温度 10～20℃)的环境中保存备用，实验前对种子进行筛选，清除干瘪、瘦小、损伤以及腐坏的种子，选出健康、饱满的种子总计 5 000 g。Tango 采集颗粒状样本时，要求样本的容量要达到量杯容量的 2/3 以上，以此确定每个样本的种子质量为(37.0±0.3) g，用高精度电子秤量出。将种子共分为 5 组(D0，D2，D4，D6，D8)，D0 组样本 15 个，不进行老化处理；其余各组每组样本 13 个，进行不同程度的老化处理，将样本装入尼龙袋中并编号。根据《国际种子检验规程》中对玉米种子人工加速老化测定的规定，将样本放入高温高湿老化箱中，薄层平铺于老化箱的网架上进行老化，温度设为 41 ℃，相对湿度设为 99%，5 组样本的老化时间分别为 0，48，96，144 和 192 h。

2.3　光谱数据采集

使用积分球漫反射测量方式采集光谱数据，为了扩大样本数目，将每个样本重复装样 3 次(每次装样都要将样本翻动摇匀)测 3 条光谱取平均。所有样本光谱采集的环境条件相同：温度 22 ℃，相对湿度 30%。采样点数为 1 845 个，开始波数为 11 542.16 $\mathrm{cm^{-1}}$，结束波数为 3 926.249 $\mathrm{cm^{-1}}$，采样间距为 4.119 $\mathrm{cm^{-1}}$。

3 结果与讨论

3.1 光谱数据预处理

采集共得到 402 条光谱数据，如图 2 所示。可以看出，不同老化时间的样本光谱的整体趋势、波峰位置高度相似，属于高相似度样本分类问题。

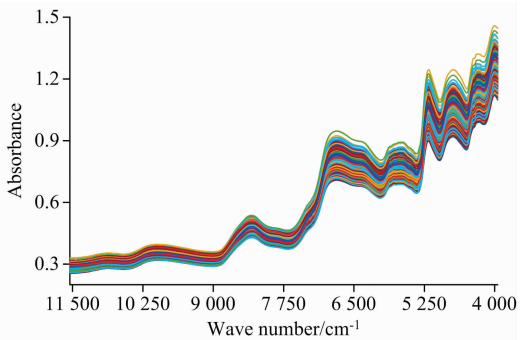

图 2 402 个玉米种子样本光谱图

Fig. 2 Spectra of 402 corn seed samples

测量的样品光谱中除了包含样品的真实信息还包括与仪器响应、测试条件和光的散射等有关的背景信息[23]，这些信息导致了光谱噪声和基线漂移。因此，在建立种子活力检测模型之前，进行光谱预处理以削弱各种背景信息对真实光谱的影响、降低模型的复杂度并提高模型的稳健性是十分必要的。在进行预处理方法选择时，首先使用高斯滤波(guassian filter，GS)、卷积平滑(Savitzky-Golay，SG)平滑、多元散射校正(multiplicative scatter correction，MSC)、标准正态变量变换(standard normal variate，SNV)方法及其组合对原始光谱进行预处理，然后使用 Kennard-Stone(KS)法将预处理后的光谱按 3∶1 划分为训练集和预测集，最后建立全谱 BP 预测模型并根据模型性能确定种子活力分级所采用的预处理方法。

表 1 不同预处理方法的 BP 全谱模型建模结果

Table 1 Results of BP full spectrum prediction model with different pretreatment methods

预处理方法	训练集		预测集	
	准确率/%	交叉熵	准确率/%	交叉熵
GS	82.14	0.026 3	76.25	0.031 4
SG	89.22	0.018 7	83.34	0.024 8
MSC	83.52	0.025 1	79.56	0.029 1
SNV	87.69	0.018 9	83.47	0.023 8
SG-SNV	92.53	0.013 4	90.18	0.016 2

由表 1 可知，在光谱平滑处理时，SG 平滑消除随机噪声的效果较 GS 好，SNV 消除样品颗粒大小和表面散射光的影响优于 MSC，组合预处理方法 SG-SNV 的模型表现最优，训练集的准确率达到 92.53%，预测集的准确率达到 90.18%。组合预处理之后的光谱如图 3 所示，与原始光谱相比减弱了噪声干扰和光谱散射问题，吸收峰的位置更加清晰。在波长 8 210，6 846，5 182，4 737 和 4 366 cm^{-1}处有 5 个显著的吸收峰。吸收峰在光谱的低频部分更为频繁，吸光度随着波数的减少而增加。

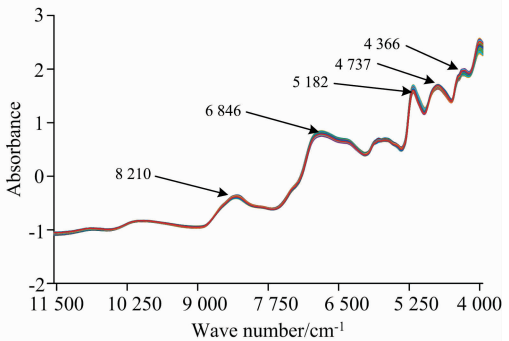

图 3 SG-SNV 预处理之后的光谱曲线

Fig. 3 Spectra after SG-SNV pretreatment

3.2 SPA 特征波长选择

首先利用经典 SPA 对全谱数据进行特征选择，设置选择变量数 20 个，采用 PLS 交叉验证择优，其中校正集样本 282 个，预测集样本 120 个。

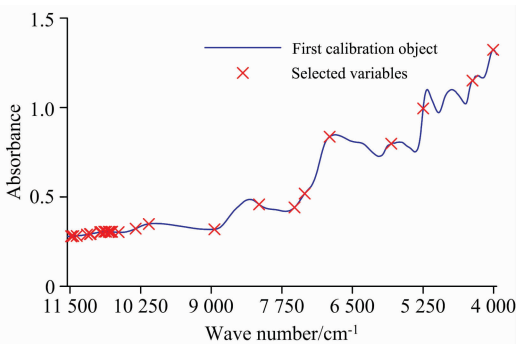

图 4 设置 20 个选择变量的 SPA 所选择的变量

Fig. 4 Selected variables of SPA with assigning value 20 to the number of variables selected

图 4 显示了 SPA 根据交叉验证均方根误差(RMSECV)从 1 845 个变量中选出的 20 个波长变量的位置。从图中可以看出，选中的波长集中在光谱第一个波峰附近，只有几个波长分布在其他几个吸收峰附近，这些特征波长所携带的信息量明显缺失，此时的交叉验证误差 RMSECV 为 0.714 7，SPA 波长选择的时间为 12.125 3 s。当设置选择变量数为 26 时的交叉验证误差 RMSECV 为 0.685 0，SPA 波长选择的时间为 14.417 5 s；当设置选择变量数为 35 时的交叉验证误差 RMSECV 为 0.669 1，SPA 波长选择的时间为 16.824 3 s。由此可见，设置不同的变量数得到的最小 RMSECV 是变化的，总的趋势是随着选择变量数的增加而减小，但什么时候 RMSECV 达到最小，很难通过不系统的抽样式的 SPA 选择来确定。

3.3 SPA_{sa} 波长变量选择

下面采用 SPA_{sa} 进行波长变量选择，设定 H 的变化范围为[1，80]，校正集和预测集样本的设定保持不变，变量的选择结果如图 5(a)所示，此时选中波长在各个吸收峰的附近都

有分布，提取的信息比较均衡。在 H 变化过程中，以选择的变量数为横坐标，以 RMSECV 为纵坐标，绘制 RMSECV 随变量数变化的趋势图，如图 5(b)所示。当选择变量数增加时，RMSECV 最小值是逐渐减小的，当变量数达到 47 时，RMSECV 值达到最小 0.6217；当变量数继续增加时，RMSECV 趋向于稳定。当变量数增加到接近 80，继续增加会引入与预测值无关的波长变量或具有较大噪声的变量，此时 RMSECV 会急剧增加。

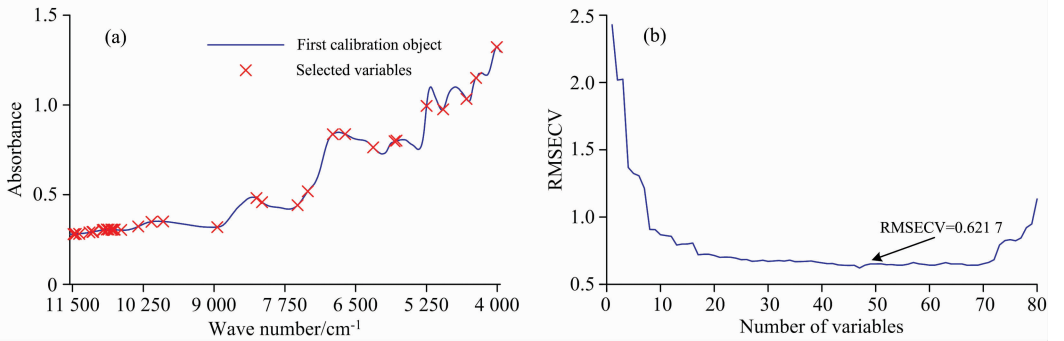

图 5　SPA_{sa} 特征波长选择结果(a)和 SPA_{sa} 波长选择中 RMSECV 随设定变量的个数的变化(b)

Fig. 5　Selected variables of SPA_{sa} (a) and the change of RMSECV with the setting number of selected variables in SPA_{sa} (b)

3.4　耦合 MIV_{opt}-SPA_{sa} 特征波长优选

由于 SPA_{sa} 需要在 H 的一定范围内反复进行 SPA 操作，因此其运算时间相当于多次 SPA 的时间累加。当光谱数据量较大时，算法的运行时间较长，因此对其进行预降维十分必要。下面首先对 11 542.16～3 926.249 cm^{-1} 范围的光谱数据进行 MIV 平均影响值计算，MIV 影响值随波长的分布如图 6 所示。

由图 6 可见，不同波长变量的 MIV 数值差异较大，为了去除与种子活力信息无关或相关性较小的波长，根据式(6)计算各个波长对应的相对距离比，选取的波长变量作为 SPA_{sa} 的备选光谱数据。为了找到最佳的 D，建立选取的波长数、BP 模型预测的准确率随 D 变化的关系，如图 7 所示。

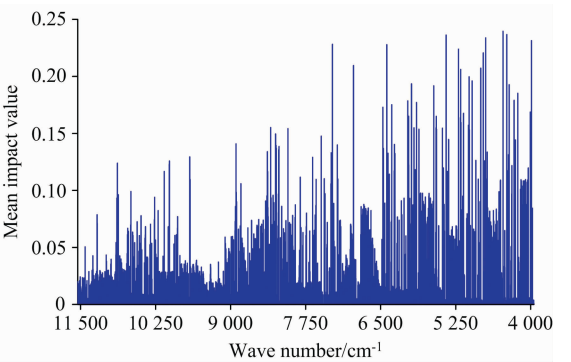

图 6　全谱数据的 MIV 值分布

Fig. 6　Distribution of MIV values of full spectrum data

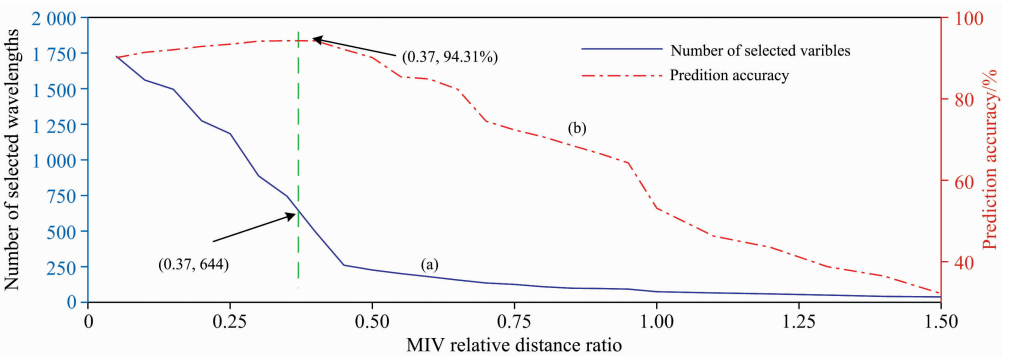

图 7　MIV_{opt} 预降维

(a)：选择的波长数随相对距离比的变化；(b)：预测准确率随相对距离比的变化

Fig. 7　MIV_{opt} pre-dimensionality reduction

(a)：Number of selected wavelengths varies with MIV relative distance ratio;

(b)：Prediction accuracy varies with MIV relative distance ratio

从图 7 可以看出，D 值的范围从 0.05 开始逐渐增加，随着 D 值增加，选择的波长数逐渐减少，BP 模型的预测准确率开始时逐渐增大，而后逐步降低。当 D 值在 0.35 附近时，预测准确率达到 94.31%，当 D 值取 0.40 时，预测准确率变为 94.28%。让 D 值在[0.35，0.40]范围内以步长 0.01 变化，求得对应的预测准确率。当 D 为 0.37 时达到预测准确率最大值 94.79%，此时选择的变量数目为 644 个。以这 644 个变量作为 SPA_{sa} 的备选光谱数据，设定 SPA_{sa} 的优选变量

个数小于 75 个，校正集和预测集样本数目保持 282 和 120 个不变，经过 SPA_{sa} 共筛选出变量 37 个，此时 PLS 模型的 RMSECV 为 0.504 9，运算时间为 14.357 s。

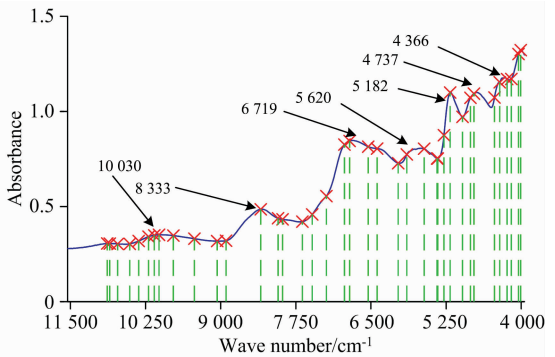

图 8 MIV_{opt}-SPA_{sa} 筛选出的特征波长分布

Fig. 8 The distribution of wavelengths selected by MIV_{opt}-SPA_{sa}

如图 8 所示，筛选出的特征波长主要集中在 7 个波峰附近。在 4 000～4 500 cm^{-1} 波段，以 4 235 为中心，左右分布着 4 165 和 4 358 等几个特征波长，此波段为玉米脂肪 C—H 基团的吸收峰。在 4 600～5 500 cm^{-1} 波段，以 5 000 为中心，两侧各有一个波峰，分布着 4 787，4 844，4 976，5 182 和 5 285等特征波长，此波段为玉米蛋白质 N—H 基团及淀粉 O—H 基团的合频吸收峰区。在 5 500～7 500 cm^{-1} 波段，有两个波峰，一个以 5 620 为中心，两边分布着 5 615 和5 903 等特征波长；一个以 6 719 为中心，两侧分布着 6 397，6 546，6 854 和 6 941 等特征波长，此吸收峰是水份的倍频吸收区。在 8 000～9 000 cm^{-1} 波段，以 8 333 为中心，分布着7 971，8 045 和 8 910 等特征波长，此波段为玉米淀粉甲基 C—H 基团二级倍频的吸收谱带。在 9 000～11 000 cm^{-1} 波段有一个波峰，以 10 030 为中心，两侧分布着 9 437，9 787，10 200，10 360 和 10 510 等特征波长，此波段是淀粉甲基 C—H 基团三级倍频及组合频的吸收谱带。

由上述分析可知，MIV_{opt}-SPA_{sa} 优选得到的特征波长分布与玉米种子生化物质构成有着高度的一致性，具有明显的物理意义，可以体现玉米老化过程中种子内部物质组成的变化；实现光谱数据的大幅度降维，是一种有效的基于变量信息的特征提取方法。

3.5 模型对比

BP 神经网络能学习和存贮大量的输入-输出模式映射关系，而无需揭示描述这种映射关系的数学方程。其学习规则为梯度下降法，通过反向传播不断调整网络的权值和阈值，使网络的误差平方和最小[24]。本研究采用三层的 BP 网络，隐层采用 sigmoid 激活函数，输出层采用 softmax 损失函数，训练算法采用比例共轭梯度反向传播算法。样本集划分方法采用 KS 法，训练集和预测集的比例为 3∶1，将 402 个样本数据划分为训练集(282)、预测集(120)，最大迭代次数设为 1 000。

为了对比全谱、MIV、SPA_{sa}、耦合 MIV_{opt}-SPA_{sa} 以及目前为许多学者青睐的 CARS 波长提取方法对模型性能的影响，建立了 5 个 BP 模型，Full-BP、MIV-BP、SPA_{sa}-BP 和 MIV_{opt}-SPA_{sa}-BP 以及 CARS-BP。待输入的光谱数据均经过 SG-SNV 预处理，模型的最佳隐层节点数根据经验公式和数据实测综合确定。所依据的经验公式为

$$m = \sqrt{n+l} + \alpha \tag{7}$$

式(7)中，m 为隐层节点数；n 为输入层节点数；l 为输出层节点数，α 为 1～10 之间的常数。

表征模型效率的评价指标是运算时间；表征模型精度的指标有准确率和交叉熵(cross-entropy，CE)，每个模型都经过 50 次训练，各个指标取平均来表征模型最终性能，5 种模型的性能对比如表 2 所示。

表 2 5 种模型的性能对比

Table 2 Performance comparison of 5 models

模型	输入节点数	隐层节点数	输出层节点数	运算时间/s	训练准确率/%	预测准确率/%	最佳交叉熵
Full-BP	1 845	49	5	0.253 1	90.3	89.4	0.081 241
MIV-BP	644	30	5	24.523	96.7	95.3	0.052 381
SPA_{sa}-BP	46	14	5	101.224	94.4	93.6	0.094 531
MIV_{opt}-SPA_{sa}-BP	37	11	5	14.382	99.2	99.1	0.007 892
CARS-BP	173	18	5	97.226	97.5	97.3	0.013 425

从表 2 可以看出，全谱模型的总的运算时间最少，其次是 MIV_{opt}-SPA_{sa}-BP，效率最低的是 SPA_{sa}-BP。这是因为具有相同输入变量的 SPA_{sa} 算法与 BP 的运算时间相比要大得多，对于具有 1 845 个输入的 BP 来说，其运算时间一般是几十毫秒，而 SPA_{sa}-BP 一般要达到 100 s 左右，而具有 1 845 个输入的 MIV_{opt}-SPA_{sa}-BP 的运算时间一般是十几秒。从模型的准确性和稳健性来看，MIV_{opt}-SPA_{sa}-BP 模型的准确率可达 99%以上，最佳交叉熵为 0.007 892 远远小于另外 4 个模型。

4 结 论

以提高玉米种子活力等级预测模型性能为目标，从优化特征波长提取的角度改进 BP 模型，提出了耦合 MIV_{opt}-SPA_{sa} 特征波长提取算法。该算法综合了 MIV 算法和 SPA 算法的优点，在 MIV 算法中引入相对距离比这个评价指标为数据降维提供了有效的衡量标准；在 SPA 算法中设定提取波长数量的范围，在此范围内优中选优，有效地解决了

SPA 算法特征波长数量确定的问题。由于在本质上 SPA 算法是一种基于偏最小二乘模型的特征提取方法，而 MIV 算法是一种评估输入变量对 BP 模型影响的算法，因此，耦合的 MIV_{opt}-SPA_{sa}算法融入了线性和非线性预测模型的内核，该算法对适合于线性模型和非线性模型预测的基于信息的特征波长提取兼收并蓄，提取出与玉米种子生化物质 NIRS 吸收特性一致的特征波长分布，极大地提高了 BP 预测模型的精度和稳健性，为基于信息的光谱数据特征波长提取提供了新思路。该算法需要进一步改进的地方是 SPA_{sa} 的运算效率不够高，在建立算法数据结构以及存取数据时进一步优化代码量并降低运算次数是解决该问题的关键。

References

[1] HU Jin, LI Yong-ping, HU Wei-min(胡　晋，李永平，胡伟民). Principle and Method of Seed Viability Determination(种子生活力测定原理和方法). Beijing: China Agricultural Publishing House(北京：中国农业出版社), 2009, 5.

[2] Maythem A A, Robert L G, Mauricio F S, et al. Seed Science Research, 2018, 28(3): 245.

[3] He X T, Feng X P, Sun D W, et al. Molecules, 2019, 24(12): 2227.

[4] LI Wu, LI Yan, LI Gao-ke, et al(李　武，李　妍，李高科，等). Journal of Nuclear Agricultural Sciences(核农学报), 2018, 32(8): 1611.

[5] JIN Wen-ling, CAO Nai-liang, ZHU Ming-dong, et al(金文玲，曹乃亮，朱明东，等). Chinese Optics(中国光学), 2020, 13(5): 1032.

[6] Rato Tiago J, Reis Marco S. Chemometrics and Intelligent Laboratory Systems, 2019, 186(6): 41.

[7] Zareef M, Chen Q S, Ouyang Q, et al. Analytical Methods, 2018, 10(25): 3023.

[8] ZHOU Hua-mao, CHEN Tian-bing, LIU Mu-hua, et al(周华茂，陈添兵，刘木华，等). Chin. J. Anal. Chem. (分析化学), 2020, 48(6): 811.

[9] Dong Chunwang, Liu Zhongyuan, Yang Chongshan, et al. Infrared Physics & Technology, 2021, 119(11): 103934.

[10] Liu Jinming, Jin Shuo, Bao Changhao, et al. Bioresource Technology, 2021, 321(2): 124449.

[11] Yun Y H, Bin J, Liu D L, et al. Analytica Chimica Acta, 2019, 1058(2): 58.

[12] CHENG Jie-hong, CHEN Zheng-guang(程介虹，陈争光). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(11): 3451.

[13] Jiang H, Xu W, Chen Q. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019, 214(10): 366.

[14] Yang M, Xu D, Chen S, et al. Sensors, 2019, 19(2): 263.

[15] Nie Pengcheng, Zhang Jinnuo, Feng Xuping, et al. Sensors and Actualtors B: Chemical, 2019, 296(8): 126630.

[16] Wakholi C, Kandpal L M, Lee H, et al. Sensors and Actualtors B: Chemical, 2018, 255(3): 498.

[17] Song J, Li G, Yang X. Journal of the Science of Food and Agriculture, 2019, 99(11): 4898.

[18] Araùjo M U, Saldanha T, Galvão R K. Chemometrics and Intelligent Laboratory Systems, 2001, 57(2): 65.

[19] Dombi G W, Nandi P, Saxe J M. The Journal of Trauma: Injury, Infection, and Critical Care, 1995, 39(5): 915.

[20] Tan X, Ji Z, Zhang Y. Technology & Health Care, 2018, 26(6): 1.

[21] Zhang Z, Yang J. International Journal of Oil, Gas and Coal Technology, 2016, 11(3): 279.

[22] LI Da-hu, LI Qiu-ke, WANG Wen-cai, et al(李大虎，李秋科，王文才，等). Coal Engineering(煤炭工程), 2020, 52(11): 154.

[23] YAN Yan-lu, CHEN Bin, ZHU Da-zhou(严衍禄，陈　斌，朱大洲). Principle, Technology and Application of Near Infrared Spectroscopy(近红外光谱分析的原理、技术与应用). Beijing: China Light Industry Press(北京：中国轻工业出版社), 2013. 2.

[24] PAN Qing-xian, DONG Hong-bin, HAN Qi-long, et al(潘庆先，董红斌，韩启龙，等). Journal of University of Science and Technology of China(中国科学技术大学学报), 2017, 47(1): 18.

Optimization of Seed Vigor Near-Infrared Detection by Coupling Mean Impact Value With Successive Projection Algorithm

YANG Dong-feng[1], LI Ai-chuan[1], LIU Jin-ming[1], CHEN Zheng-guang[1], SHI Chuang[1], HU Jun[2*]

1. College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China

2. College of Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China

Abstract At present, near-infrared spectroscopy (NIRS) technology, can realize the rapid and non-destructive detection of seed vigor, but the vigor grade is generally less than 3, and the accuracy is not high. The contradiction between the increase of vigor level and model precision urgently needs to be solved in the near-infrared spectrum detection of seed vigor. Five kinds of seed samples were obtained by the artificial aging method, and the corresponding spectral data were collected to establish the BP prediction model. In order to improve the accuracy and robustness of the model, an algorithm of coupled Mean Impact Value-Successive Projection Algorithm (MIV_{opt}-SPA_{sa}) is presented. Aiming at the problem of determining the number of feature variables extracted by the Successive Projection Algorithm(SPA), the algorithm sets the number range of feature wavelengths and selects the best in this range to realize adaptive SPA(SPA_{sa}). Aiming at the problem that SPA algorithm takes a too long time, MIV algorithm is used to reduce the dimension of SPA algorithm. Although the MIV method can sort the wavelength influence values, it lacks the threshold value for selecting wavelength influence. Therefore, the relative distance ratio is introduced to optimize the MIV algorithm to effectively segment the characteristic wavelength range. The full spectrum with 1 845 wavelengths is extracted by the MIV_{opt}-SPA_{sa} algorithm, and 37 characteristic wavelengths are extracted, which are mainly distributed near the 7 main absorption peaks of near-infrared spectrum of maize seeds. The results show that the algorithm can effectively extract the characteristic wavelength, which is consistent with the NIR absorption characteristics of maize seed biochemical substances. In order to verify the effect of the algorithm on the performance of the model, the full spectrum BP model, SPA_{sa}-BP model, MIV-BP model, MIV_{opt}-SPA_{sa}-BP model and competitive adaptive reweighting CARS-BP model were established to classify the five grades of maize seed vigor. The average prediction accuracy of the MIV_{opt}-SPA_{sa}-BP model is 99.1%, which is higher than other models; the average prediction time is 14.382 s, which is lower than that of the MIV-BP model (24.523), CAR-BP (97.226) and SPA_{sa}-BP model (101.224 s), but higher than that of full-spectrum model (0.253 1); The best performance cross-entropy is 0.007 892, which is far lower than other 4 models. The experimental results show that the MIV_{opt}-SPA_{sa} algorithm can effectively improve the accuracy of the near-infrared detection model of maize seed vigor, realize multi-level, accurate and nondestructive detection of seed vigor, and provide a reference for optimizing the optimisation seed vigor detection model.

Keywords Near infrared spectroscopy; Seed vigor; Maize; Mean impact value; Successive projection algorithm

(Received Jul. 29, 2021; accepted Oct. 25, 2021)

* Corresponding author