改进的联合区间随机蛙跳算法的近红外光谱波长选择

程介虹[1]，陈争光[1,2*]

1. 黑龙江八一农垦大学电气与信息学院，黑龙江 大庆 163319
2. 黑龙江省水稻生态育秧装置及全程机械化工程技术中心，黑龙江 大庆 163319

摘 要 在近红外光谱的建模预测分析中，数据的冗余及共线性会严重影响模型的预测精度和稳健性。特征波长选择是提高定量分析预测精度的一种有效方法。随机蛙跳(RF)是一种依据不同的变量具有不同的被选择可能性的特征波长选择算法，近年来在特征波长提取方面展现良好的性能。该方法通过多次迭代，计算每个变量被选择的概率，以优选概率高的变量为特征波长。但由于其初始变量集 V_0 的产生是随机的，具有较大的不确定性，可能会包含无用或干扰信息，难以保证初始信息的有效性，使得迭代次数过大，运行时间过长。故而提出一种改进的联合区间随机蛙跳(Si-RF)特征波长选择算法，通过联合区间偏最小二乘法(Si-PLS)对全谱进行变量初选，此时得到的波长对目标变量变化最为敏感，将其作为RF的初始变量子集，以解决RF运行时间较长、效率较低的问题。另一方面，RF在选择特征波长时，选择被选概率值大于阈值的变量为特征波长，但对概率值阈值的设定无理论依据，易受人为因素影响。通过对变量按被选概率值降序排列后逐次增加一个波长建立多元线性回归(MLR)模型，以验证均方根误差(RMSEV)值最低时的变量子集为特征波长，以找到预测精度最高点所包含的波长，提高预测精度。针对上述两点进行改进，将其应用于一组土壤样本近红外光谱数据集，进行特征波长选择后，建立MLR模型，与RF-MLR及全谱-PLSR模型的预测精度进行比较。结果表明：RF经过10 000次迭代，优选出10个波长点，建立的MLR模型的预测均方根误差(RMSEP)为1.6276；而改进后Si-RF只需进行1 000次迭代，优选出17个波长点，其MLR模型的RMSEP减小到0.818 4，大大提升了预测精度，提高运行效率。相较于全谱，也极大的提高了预测精度，简化模型的复杂度，证明改进的Si-RF是一种有效的特征波长选择算法。

关键词 近红外光谱；特征波长选择；多元校正；随机蛙跳；联合区间偏最小二乘

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)11-3451-06

引 言

近红外光谱区(800～2 500 nm)的含氢基团的倍频和合频吸收峰组成的吸收强度较弱灵敏度较低，吸收带较宽且严重重叠。若采用全谱建模，不仅会存在某些光谱区域与待测组分相关性弱，而且相邻的波长高度相关，包含了大量的冗余信息，这都会影响模型的精度和稳健性。克服这些问题的有效途径是对所测得的光谱进行波长选择，减少建模所需的波长点和计算工作量，进而得到预测能力强、鲁棒性高的模型。在众多特征波长选择算法中，随机蛙跳(random frog，RF)[1]是近年来提出的一种新型特征波长选择算法。其依据不同的变量具有不同的被选择可能性，通过多次迭代，计算每个变量被选择的概率，选择概率高的变量为特征波长。

陈立旦等[2]通过RF选出特征波长后，建立最小二乘支持向量机(least squares support vector machine，LS-SVM)模型，对生物柴油的含水量进行预测，发现RF-LS-SVM模型的相关系数大于0.95，可以准确地预测生物柴油的含水量。胡孟晗等[3]通过RF对特征波长进行提取，建立LS-SVM模型预测蓝莓硬度和弹性模量，与全谱模型对比，RF算法可以有效地去除冗余信息，提升模型预测准确率。孙红等[4]采用相关系数法(correlation coefficient，CC)和RF算法筛选对叶绿素含量敏感的波长，建立偏最小二乘回归(partial least squares regression，PLSR)模型对马铃薯作物的叶绿素含量

收稿日期：2019-10-15，**修订日期**：2020-02-06

基金项目：国家重点研发计划项目(2016YFD0701300)，黑龙江省农垦总局重点科研计划项目(HKKYZD190804)，黑龙江八一农垦大学科研团队计划项目(TDJH201807)资助

作者简介：程介虹，1997年生，黑龙江八一农垦大学硕士研究生 e-mail：1024212535@qq.com

*通讯联系人 e-mail：ruzee@sina.com

进行预测，结果表明 RF-PLSR 模型预测精度优于 CC-PLSR，可实现马铃薯不同叶位叶绿素含量的无损检测。此外，Yu 等[5]采用 RF 和 PLSR 建立校准模型，发现通过 380～1 030 nm 区域的波长可实现辣椒植物的总氮含量的预测。Zhao 等[6]通过 RF 算法选择特征波长，建立 RF-PLSR 和 RF-LS-SVM 模型预测桑葚果实的总可溶性固体值含量，两个模型皆具有良好的性能。以上结果表明 RF 算法在数据降维方面是有效的。

尽管 RF 算法在特征波长选择方面具有一定优势，但存在两方面的不足：其一是，初始变量集 V_0 的产生是随机的，难以保证初始信息的有效性；算法为保证运行过程中遍历整个数据集，要求迭代次数 N 需足够大，从而导致算法的运行时间长、收敛速度慢。其二是，RF 在选择特征波长时，选择被选概率值大于阈值的变量为特征波长，但对阈值的设定无理论依据，易受人为因素影响。

针对上述两点，对 RF 算法进行了改进，提出一种联合区间随机蛙跳(synergy interval-random frog，Si-RF)算法，以一组公开的土壤样本近红外光谱数据为例，分别利用 RF 和改进的 Si-RF 进行特征波长选择，建立多元线性回归(multiple linear regression，MLR)模型，比较预测精度，并与全谱的 PLSR 模型进行对比，以证明改进的 Si-RF 算法的有效性。

1 实验部分

1.1 样本数据

所用数据为一组土壤样本近红外光谱数据，来自于网站 Quality & Technology。该数据集包含 108 个土壤样本。样本光谱的波长范围为 400～2 500 nm，采样间隔为 2 nm，共计 1 050个波长点。本文以土壤有机质(soil organic matter，SOM)的含量作为因变量进行波长选择及近红外光谱数据建模预测分析。

1.2 随机蛙跳算法

1.2.1 算法步骤

RF 是 Li[1]提出的一种类似于可逆跳跃马尔可夫链蒙特卡罗(reversible jump Markov Chain Monte Carlo，RJMCMC)的算法，它以迭代的方式进行，计算每个变量在每次迭代中被选择的概率，概率越高变量重要性越大，优选概率高的变量为特征变量。

随机蛙跳的主要步骤包括以下三步[1]：

(1)初始化：参数设置，随机选择一个包含 Q 个变量的变量子集 V_0；

(2)概率引导模型搜索：基于 V_0，选择包含 Q^*(随机产生)个变量的候选变量子集 V^*，以一定概率接受 V^* 作为 V_1，并用 V_1 代替 V_0，循环此步骤直至 N 次迭代完成；

(3)变量评估：计算每个变量被选择的概率，概率越高变量重要性越大。

其中概率引导模型搜索和变量评估具体方法如下。

1.2.2 概率引导模型搜索

首先，从均值为 Q、方差为 $0.3Q$ 的正态分布中随机选择一个整数 Q^*，之后通过以下三种方式之一产生一个包含 Q^* 个变量的候选变量子集 V^*：

(1)如果 $Q^*=Q$，则令 $V^*=V_0$。

(2)如果 $Q^*<Q$，则首先对 V_0 建立 PLS 模型，记录并比较模型中每个变量的回归系数的值，将 $Q-Q^*$ 个最小回归系数的相关变量从 V_0 中移除。剩余的 Q^* 个变量为候选子集 V^*。

(3)如果 $Q^*>Q$，则从 $V-V_0$(V 代表包含全部 p 个变量的集合)中随机抽取 $\omega(Q^*-Q)$ 个变量，ω 默认值为 3，生成一个变量子集 T，通过 V_0 和 T 的组合建立 PLS 模型，保留模型中回归系数最大的 Q^* 个变量，并将其设为候选子集 V^*。

简而言之，利用所提出的正态分布控制变量数，实现变量的增、删操作。在得到候选变量子集 V^* 后，下一步是确定 V^* 是否可以被接受。分别对 V_0 和 V^* 建立 PLS 模型，计算交叉验证均方根误差(cross-validation root mean square error，RMSECV)，得到 RMSECV 和 RMSECV*。如果 RMSECV* ≤RMSECV，接受 V^* 为 V_1，否则接受 V^* 为 V_1 概率为 0.1RMSECV/RMSECV*。最后，使用 V_1 中的变量更新 V_0，并重复 N 次迭代，直至循环结束。

1.2.3 变量评估

N 次迭代之后，总共获得 N 个变量子集。对于每个变量，可以使用式(1)计算其被选择的概率。

$$\mathrm{Prob}_j=\frac{N_j}{N},\quad j=1,2,\cdots,p \tag{1}$$

式(1)中，N_j 为第 j 个变量在 N 次迭代中被选择的次数，变量越重要，被这 N 个变量子集选择的机会就越多。因此，该选择概率可以用作变量重要性的度量，可以用作变量选择的标准。

1.3 对 RF 算法的改进

1.3.1 V_0 子集的初选

在 RF 算法中，初始变量集 V_0 的产生是随机的，具有较大的不确定性，可能会产生无信息变量或干扰信息，从而导致算法的迭代次数大，运行时间长。为了提高初始集 V_0 变量的有效性，减少迭代次数，对 V_0 子集的产生进行改进。

联合区间偏最小二乘法(synergy interval partial least squares，SiPLS)是 Norgaard 提出的一种波长选择算法。该方法将光谱划分为等宽的 n 个子区间，对其中 m 个子区间任意组合为联合区间。基于联合区间建立 PLS 模型，比较各 PLS 模型的 RMSECV 的值，将最小 RMSECV 值所对应的联合区间的波长设为初始变量集 V_0，开始迭代，可以消除 V_0 的随机性，避免无信息变量及噪声的干扰，从而减少迭代次数。

1.3.2 建模波长的优选

在 RF 算法中，一般选择概率值较大的前 10 或 15 个变量，或者通过人为设定概率的阈值，取概率值大于阈值的变量来选择符合要求的特征波长，建模波长数量选择存在不确定性。

本文的改进是：对排序后的变量从第一个波长开始，每次增加一个波长，建立光谱数据和有机质含量数据之间的

MLR 模型。计算每个模型的验证均方根误差(root mean square error of validation，RMSEV)值，其中最小 RMSEV 值所对应的变量子集即为特征波长。RMSEV 可以使用式(2)计算

$$\mathrm{RMSEV}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \tag{2}$$

式(2)中，y_i 为样本实测浓度值，$\hat{y}_i$ 为模型计算出来的浓度值，n 为验证集的样本数量。

这样可以找到预测精度最优所包含的波长数，提高预测精度。

1.4　建模方法

现有研究大多对 RF 所选特征波长建立 PLSR 模型。而 MLR 是一种常规的校正方法，直观简单，且具有良好的统计特性，应用非常普遍，其优点是产生的模型比主成分回归(principal components regression，PCR)和 PLSR 模型更简单，更易于解释。

本工作建立三种模型：基于全谱的 PLSR 模型、基于 RF 波长选择的 MLR 模型和基于 Si-RF 改进的波长选择的 MLR 模型。通过三种模型预测能力的比较验证本法的有效性。模型的预测能力主要通过校正相关系数(R_c)、校正均方根误差(RMSEC)、预测相关系数(R_p)、预测均方根误差(RMSEP)指标来评价。其中，R 取值越接近 1，RMSEC 和 RMSEP 越接近 0，模型的拟合性越好，预测精度越高。

1.5　数据分析

软件采用 MATLAB R2015b 及 The Unscrambler X 10.3 (64-bit)，光谱数据的预处理、建模分析及预测在 Unscrambler 软件中实现，特征波长提取、图形的绘制在 MATLAB 中实现。计算机硬件的配置为 Intel(R)Core(TM)i5-3450CPU @3.50GHz 处理器，8GB 内存，操作系统为 windows10。

2　结果与讨论

2.1　光谱数据特征

土壤样本的原始近红外光谱图如图 1(a)所示。为校正光谱基线，消除其他背景的干扰，提高光谱分辨率，并且在一定程度上减少各变量间的线性相关性，利用 Savitzky-Golay 窗口宽度为 11 的一阶求导法对原始光谱数据进行预处理，预处理后的近红外光谱图如图 1(b)所示，可以发现通过预处理后的近红外光谱曲线，能更精确地确定吸收峰的位置。

将 108 个土壤样本通过 SPXY(sample set portioning based on joint x-y distance)算法分为 75%训练集和 25%预测集，建模集包含 81 个样本，预测集包含 27 个样本，土壤有机质含量统计数据结果如表 1 所示。划分后的建模集的 SOM 含量范围涵盖预测集的 SOM 含量，建模集具有代表性。

2.2　特征波长选取

2.2.1　RF 变量选择结果

如前所述，首先对 RF 进行初始化参数设置，N 设定为 10 000，Q 设定为 10，开始运行。每个变量被选择的概率结果如图 2 所示，选择概率大于 0.2 的变量为最终特征波长，得到满足条件的有 10 个波长点分别为 1 420，1 390，1 392，1 394，1 388，1 422，2 318，1 424，1 396 和 1 922 nm。

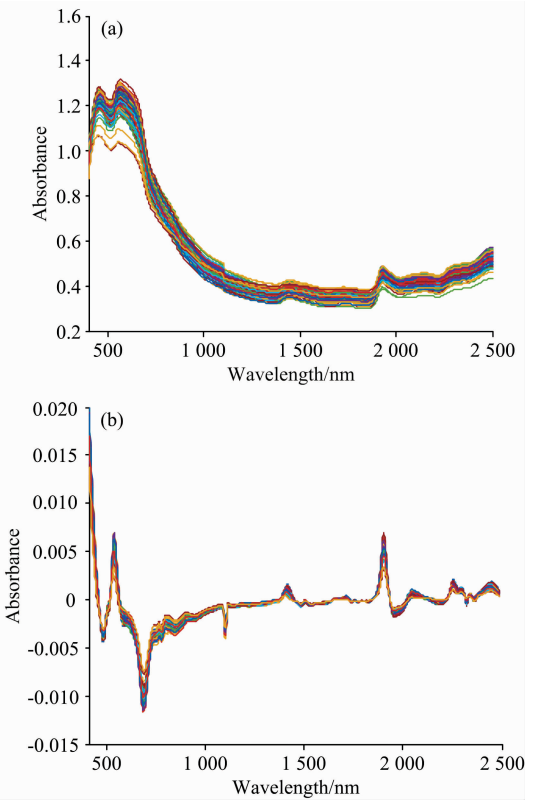

图 1　原始光谱图及预处理后的光谱图

(a)：原始光谱图；(b)：S-G 一阶导处理后的光谱图

Fig. 1　Original and pre-processed spectra

(a)：Original；(b)：S-G first derivative

表 1　土壤有机质含量统计数据结果

Table 1　Statistical data of soil organic matter content

样本类型	样本数	最小值	最大值	均值	标准差	变异系数/%
总体样本	108	42.91	95.85	85.43	10.82	12.7
建模集	81	42.91	95.85	83.73	11.79	14.0
预测集	27	74.99	93.46	90.52	4.30	4.8

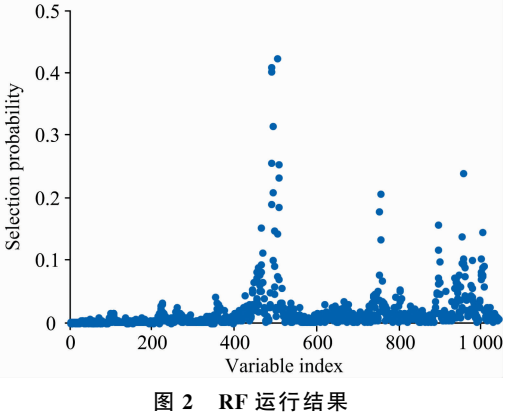

图 2　RF 运行结果

Fig. 2　The result of random frog

2.2.2 Si-RF 变量选择结果

首先利用 SiPLS 对全谱数据进行特征波长选择，将全谱数据依次等分成 20，25，30 和 35 个区间，由于联合区间的组合会有 C_n^m 种情况，为避免计算量过大，组合数分别设置为 2 和 3，得到结果如表 2 所示。

表 2 SiPLS 子区间优选结果

Table 2 Sub-interval optimization results of SiPLS

区间总数	组合数	所选区间	RMSECV
20	2	[10 18]	1.678
	3	[10 15 19]	1.595
25	2	[12 23]	1.787
	3	[12 19 24]	1.575
30	2	[15 28]	1.556
	3	[12 15 28]	1.466
35	2	[15 17]	1.703
	3	[17 18 32]	1.583

由表 2 可以发现，将全谱等分为 30 个区间，组合数设置为 3 时，RMSECV 最小，此时所选的特征波长点为 104 个，将这三个波段 1 182～1 250，1 392～1 460 和 2 288～2 354 nm，共计 104 个波长点作为初始变量子集 V_0，RF 算法的迭代次数分别设置为 500，1 000，1 500 和 2 000 次，得到结果如表 3 所示。

表 3 不同迭代次数的优选结果

Table 3 Optimal results of different iteration times

迭代次数	变量数	$RMSEV_{min}$
500	13	0.988 4
1 000	17	0.818 4
1 500	24	0.876 5
2 000	17	0.981 9

由表 3 可知，当 N 设置为 1 000 次时，RMSEV 值最小。该情况下 Si-RF 运行结果如图 3 所示，每个变量被选择的概率结果如图 3(a)所示。将每个变量被选择的概率值进行降序排列，从第一个波长开始，逐次增加一个波长建立 MLR 模型。各模型的 RMSEV 值如图 3(b)所示，正方形标记所示为最低 RMSEV 值，为 0.818 4，此时选择的特征波长数为 17 个，分别为 1 392，1 394，1 420，2 332，2 330，1 418，1 440，1 348，1 920，1 402，2 000，1 424，2 312，1 442，1 426，1 444和 2 364 nm。

2.3 模型建立与比较

将全谱、RF 以及 Si-RF 选择的特征波长，建立回归模型比较预测能力，得到模型的校正、预测相关系数和校正、预测均方根误差的值如表 4 所示。

从表 4 可以看出，RF 和 Si-RF 模型的各项参数均优于全谱，改进的 Si-RF 算法模型的各项参数均优于 RF。基于 RF 所选特征波长的 MLR 模型的 R_p 为 0.9354，RMSEP 为 1.627 6，而改进后 Si-RF 选择的特征波长 MLR 模型的 R_p 为 0.984 8，RMSEP 减小到 0.818 4，大大提升了预测精度。

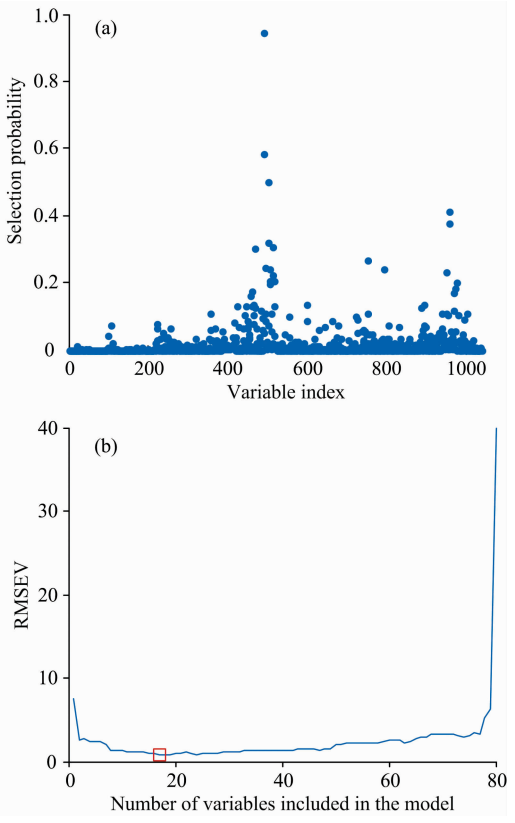

图 3 Si-RF 运行结果

(a)：各变量被选概率；(b)：各模型 RMSEV 值

Fig. 3 The result of Si-RF

(a): Selection probability of each variable;
(b): RMSEV values of each model

表 4 不同波长选择方法下模型的结果

Table 4 Results of model with different wavelength selection methods

模型	变量数	R_c	RMSEC	R_p	RMSEP
FULL-PLS	1 050	0.963 3	3.143 1	0.909 3	1.954 3
RF-MLR	10	0.982 7	2.172 0	0.935 4	1.627 6
Si-RF-MLR	17	0.993 6	1.327 0	0.984 8	0.818 4

图 4 分别为对建模集、预测集样本的全谱-PLS、RF-MLR 和 Si-RF-MLR 模型的 SOM 的实测值和预测值相关图。从图中可以更加直观的看出，基于 Si-RF 波长选择算法的 MLR 模型优于全谱模型及 RF 算法的 MLR 模型。

由于 RF 算法对初始变量集的产生是随机的，有较大的不确定性，可能会包含无信息变量或干扰信息，从而导致算法的迭代次数大、运行时间长。而通过 SiPLS 特征波长初选，得到的波长对于目标变量变化最为敏感，同时避免了其他光谱无信息变量与噪声的影响。所以首先对全谱通过 Si-PLS 进行特征波长初选，将其初选结果作为 RF 的初始变量子集 V_0，这样可以改善RF收敛速度慢的问题，减少RF算

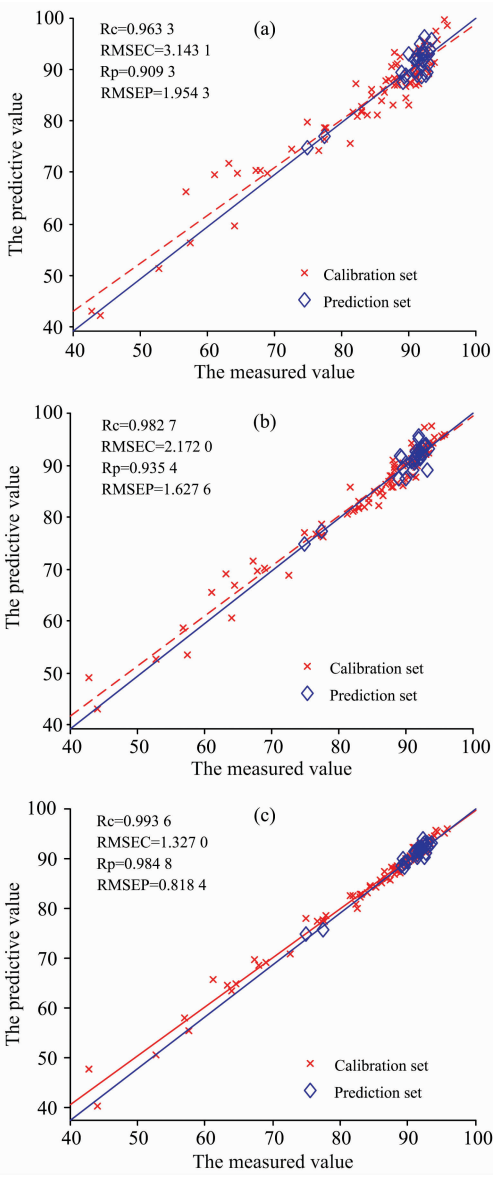

图 4　不同模型下土壤有机质的实测值和预测值相关图

(a)：全谱-PLS；(b)：RF-MLR；(c)：Si-RF-MLR

Fig. 4　Correlation between measured and predicted values of SOM obtained from different models

(a)：Full spectrum PLS；(b)：RF-MLR；(c)：Si-RF-MLR

法的迭代次数，大大节省运行时间，并且由于初始变量子集是针对于有效信息的波长，有利于 RF 每次迭代中 V^* 所包含的波长的选择，可以提高预测精度。在运行中，迭代次数也由 10 000 次减少至 1 000 次，提高运行效率。

通过 Si-RF 选出的特征波长点的范围在 1 348～1 444，1 920～2 364 nm 之间，这与许多前人研究所选波长点范围基本一致。如：白婷等[7]针对艾比湖 60 个表层土样，基于 CARS 算法提取的 SOM 特征波段主要集中在 1 970 和 2 340 nm 附近；朱亚星等[8]通过 UVE-CARS 优选出 84 个变量做为预测 SOM 含量的特征波长，分布于 561～721 和 1 920～2 280 nm 波段；于雷等[9]通过 CARS-SPA 优选出的 37 个特征波长，集中在近红外区域 1 800～2 400 nm，而且基于波长选择建立的 SOM 含量的 PLSR 模型预测精度最优。本工作 Si-RF 优选出的波段与图 2B 近红外光谱曲线吸收峰的位置也基本一致，符合高志海等[10]的论点，即光谱曲线上的凸起区可能对提取土壤有机质信息有实际意义。

对比 RF 及 Si-RF 所选波长点范围，RF 的范围在 1 388～1 424 和 1 922～2 318 nm 之间，Si-RF 的范围在 1 348～1 444和1 920～2 364 nm 之间，可以发现 Si-RF 已经基本涵盖 RF 所选波长的大部分，这也在一定程度上说明可以减少算法迭代次数。

3　结　论

提出了一种近红外光谱分析中特征波长选择的 Si-RF 算法，该方法通过对全谱进行 SiPLS 特征波长初选，将所得的波长做为初始变量子集，使得初始变量子集涵盖有效信息，以解决 RF 中迭代次数过多，运行效率较低的问题。将 RF 和改进的 Si-RF 应用于一组土壤样本近红外光谱数据集，将由 RF 选择的特征波长和改进的 Si-RF 选择的特征波长提取出来，建立 MLR 模型，发现 Si-RF-MLR 模型的预测精度优于 RF-MLR，并且在运行时间上也大大降低，提高运行效率；相较于全谱的 PLSR 模型，也极大的提高了预测精度，简化模型的复杂度。证明改进的 Si-RF 是一种有效的特征波长选择算法。

References

[1] Li H D, Xu Q S, Liang Y Z. Analytica Chimica Acta, 2012, 740(none): 20.
[2] CHEN Li-dan, ZHAO Yan-ru(陈立旦，赵艳茹). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2014, 30(8): 168.
[3] HU Meng-han, DONG Qing-li, LIU Bao-lin(胡孟晗，董庆利，刘宝林). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(11): 3651.
[4] SUN Hong, ZHENG Tao, LIU Ning, et al(孙　红，郑　涛，刘　宁，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2018, 34(1): 149.
[5] Yu K Q, Zhao Y R, Li X L, et al. PLoS One, 2014, 9(12): e116205.
[6] Zhao Y R, Yu K Q, He Y. Journal of Analytical Methods in Chemistry, 2015, 2015(2): 343782.
[7] BAI Ting, DING Jian-li, WANG Jing-zhe(白　婷，丁建丽，王敬哲). Journal of Drainage and Irrigation Machinery Engineering(排灌机

械工程学报)，2020，38(8)：829.
[8] ZHU Ya-xing，YU Lei，HONG Yong-sheng，et al(朱亚星，于　雷，洪永胜，等). Scientia Agricultura Sinica(中国农业科学)，2017，50(22)：4325.
[9] YU Lei，HONG Yong-sheng，ZHOU Yong，et al(于　雷，洪永胜，周　勇，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报)，2016，32(13)：95.
[10] GAO Zhi-hai，BAI Li-na，WANG Beng-yu，et al(高志海，白黎娜，王琲瑜，等). Scientia Silvae Sinicae(林业科学)，2011，47(6)：9.

Wavelength Selection of Near-Infrared Spectra Based on Improved SiPLS-Random Frog Algorithm

CHENG Jie-hong[1]，CHEN Zheng-guang[1,2*]
1. College of Electrical and Information，Heilongjiang Bayi Agricultural University，Daqing　163319，China
2. Helongjiang Engineering Technology Research Center for Rice Ecological Seedlings Device and Whole Process Mechanization，Daqing　163319，China

Abstract　In the modeling and prediction analysis of near-infrared spectroscopy，the redundancy and collinearity of the data will seriously affect the prediction accuracy and robustness of the model. The feature wavelength selection is an effective method to improve the prediction accuracy of quantitative analysis. Random frog (RF) is a feature wavelength selection algorithm based on different variables with different probability of being selected. In recent years，it has shown good performance in feature wavelength selection. The method calculates the probability of each variable being selected by iteration，and takes the variable with high probability as the feature wavelength. However，the initial variable set V_0 of RF is random and uncertain. It may contain useless or disturbing information. Moreover，it is difficult to guarantee the validity of the initial information，which makes the number of iterations too large and the running time too long. In this paper，an improved Si-RF feature wavelength selection algorithm is proposed based on RF. SiPLS is used to select the variables of the full spectrum. At this time，the wavelength obtained is the most sensitive to the change of the target variable. It is used as the initial variable subset of RF to solve the problem of long running time and low efficiency. On the other hand，when RF selects the feature wavelength，it selects the variable whose probability value is larger than the threshold value as the feature wavelength. However，there is no theoretical basis for setting the threshold value，which is easily influenced by human factors. In this paper，the MLR model is established by adding one variable each time in the descending order according to the probability values of being selected of each variable. The subset of variables with the lowest RMSEV value is taken as the feature wavelength，so as to find the wavelength subset contained in the highest prediction accuracy and improve the prediction accuracy. In view of the above two points，Si-RF was applied to soil near-infrared spectroscopy data sets. MLR model is established after selecting the feature wavelength，and the prediction accuracy was compared with that of RF-MLR and Full-PLSR models. The results show that the RF after 10 000 iterations，10 wavelength points are selected，and the RMSEP of the MLR model is 1.627 6. The improved Si-RF only needs 1 000 iterations to select 17 wavelength points. The RMSEP of MLR model is reduced to 0.818 4，which greatly improves the prediction accuracy and the running efficiency. Compared with the full spectrum，it also greatly improves the prediction accuracy，simplifies the complexity of the model. It proves that improved Si-RF is an effective feature wavelength selection algorithm.

Keywords　Near-infrared spectroscopy；Feature wavelength selection；Multivariate calibration；Random frog；Synergy interval partial least squares

(Received Oct. 15，2019；accepted Feb. 6，2020)

* Corresponding author