# 改进特征波段选取和混合集成建模的东北粳稻叶绿素含量估算

刘 潭[1,2]，许童羽[1,2*]，于丰华[1,2]，袁青云[1,2]，郭忠辉[1]，徐 博[1]

1. 沈阳农业大学信息与电气工程学院，辽宁 沈阳 110161
2. 沈阳农业大学辽宁省农业信息化工程技术中心，辽宁 沈阳 110161

**摘 要** 利用光谱信息快速、无损和准确的检测水稻冠层叶片叶绿素含量，对水稻的长势评估、精准施肥、科学管理都具有非常重要的现实意义。以东北粳稻为研究对象，以小区试验为基础，获取关键生长期的水稻冠层高光谱数据。首先采用标准正态变量校正法(SNV)对光谱数据进行预处理，针对处理后光谱数据，以随机蛙跳(RF)算法为基础，结合相关系数分析法(CC)和续投影算法(SPA)，提出一种融合两种初选波段的改进型随机蛙跳算法(fpb-RF)筛选叶绿素含量的特征波段，并分别与标准RF，CC和SPA方法进行对比。以提取的特征波段作为输入，结合线性模型和非线性模型各自优势，提出一种高斯过程回归(GPR)补偿偏最小二乘(PLSR)的叶绿素含量混合预测模型(GPR-P)：利用PLSR法对水稻叶绿素含量初步预测，得到叶绿素含量的线性趋势，然后利用具有较好非线性逼近能力的GPR对PLSR模型偏差进行预测，两者叠加得到最终预测值。为了验证所提方法优越性，以不同方法提取的特征波段作为输入，分别建立PLSR、最小二乘支持向量机(LSSVM)、BP神经网络预测模型。结果表明：相同预测模型条件下，改进fpb-RF算法提取特征波段作为输入可较好的降低模型复杂性、提高模型预测性能，各模型测试集的决定系数($R_P^2$)和训练集的决定系数($R_C^2$)均高于0.704 7。另外，在各算法提取特征波段进行建模时，GPR-P模型的$R_C^2$和$R_P^2$均高于0.755 3，其中，采用fpb-RF方法提取的特征波段作为输入建立的GPR-P模型预测精度最高，$R_C^2$和$R_P^2$分别为0.781 5和0.779 6，RMSEC和RMSEP分别为0.904 1和0.928 3 mg·L$^{-1}$，可为东北粳稻叶绿素含量的检测与评估提供有价值的参考和借鉴作用。

**关键词** 水稻；叶绿素含量；光谱分析；特征波段提取；fpb-RF算法；混合预测模型

**中图分类号**：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)08-2556-09

## 引 言

水稻是中国主要粮食作物之一，叶绿素含量与其营养元素、产量等有着极为密切的关系，不仅在光合过程中起着重要作用，同时也是评价水稻生长、诊断病理的重要生理生化参数[1]。水稻叶绿素含量的精准估测对其生长监测、施肥施药调控及田间的精准管理都具有重要的实际意义。

近年来，随着高光谱技术的快速发展，国内外学者采用高光谱技术对田间作物的冠层叶片叶绿素含量检测做了大量的研究，并取得了一定的成果。以往研究在对水稻等作物叶绿素估测中，多是利用光谱波段组合植被指数的方法，大致可分为单波段植被指数和多波段植被指数，还有一些研究采用作物叶片的全光谱信息。若采用单一敏感波段建模，会导致有效信息无法充分利用，模型精度降低。而当采用全波段进行建模时，由于光谱中存在大量的冗余信息，同样会使模型精度受到一定的影响。因此，通过分析水稻叶绿素的光谱特征，确定其敏感波段，是提升模型运行效率、简化模型结构、增强模型稳定性的首要条件[2]。

常见的特征波段提取主要包括2类，一类是以数理统计为基础，如连续投影法(successive projections algorithm，SPA)和相关系数法(correlation coefficient，CC)等。虽然一

**收稿日期**：2020-06-04，**修订日期**：2020-10-09

**基金项目**：国家“十三五”重点研发计划项目(2016YFD0200600)，国家自然科学基金项目(32001415)，中国博士后科学基金项目(2018M631820)，辽宁省博士科研启动基金项目(2019-BS-207)，辽宁省自然基金指导计划项目(2019-ZD-0720)资助

**作者简介**：刘 潭，1985年生，沈阳农业大学信息与电气工程学院讲师 e-mail：liutan_0822@126.com

*通讯作者 e-mail：xutongyu@syau.edu.cn

定程度可以剔除包含冗余信息的变量，但仍存在如保留变量多和筛选结果中存在较低信噪比变量等不足[3]。另一类是较为新颖的基于智能优化算法的特征波段寻优方法，如遗传算法(GA)、蚁群算法(ACA)，随机蛙跳算法(random frog，RF)。毛博慧等[4]利用 GA 算法对冬小麦苗期冠层叶绿素含量的敏感波长进行优选，并在此基础上建立其预测模型。为实时检测作物叶绿素含量。孙红等[5]采用 RF 算法筛选叶绿素含量的敏感波段，并建立偏最小二乘(partial least squares regression，PLSR)模型。此类方法可全局搜索有效信息变量，可较好地实现高光谱特征波段的选择，但也有运算时间长、模型参数复杂等问题。

在水稻等作物叶绿素含量反演建模方面，主要分为数据建模和机理建模两类。在机理建模方面，曾毓燕等[6]应用 PROSPECT 结合 DART 模型估算作物冠层尺度和叶片尺度的叶绿素含量。Sun 等[7]通过叶片辐射传输模型研究了 HSL 系统估算叶绿素含量的可能性，并取得了较好的预测结果。虽然机理建模物理意义较为明确，且反演过程较稳定，但通常模型参数较多且确定较为复杂，另外，地表环境系统包含许多不确定性因素，都会对模型精度产生较大影响。因此，结构简单、分析方便的数据驱动建模方法被广泛应用于叶绿素反演建模中，主要包括线性数据模型(如偏最小二乘、多元线性回归)及非线性数据模型(如神经网络、支持向量机)[8]。与线性模型相比，非线性模型精度有所提高，但也有一些不足。同时，叶绿素反演建模中，由于输入与输出变量间的关系既包含线性成分又包含非线性成分，一些研究也表明非线性预测模型对模型中的线性成分预测精度有限，所以较为理想的办法是对模型中的线性成分采用线性模型预测，而非线性成分则采用非线性模型预测。

综上分析，为进一步提高水稻叶绿素含量预测的精确性和稳定性，以东北粳稻为研究对象，采用无人机高光谱成像技术，首先对水稻叶绿素相关特征波段的提取方法展开研究。基于智能优化算法的优势，在标准 RF 算法基础上，结合相关系数分析法和连续投影法，提出一种融合两种初选波段的随机蛙跳算法选取特征波段，然后以提取的特征波段为输入，结合线性模型和非线性模型各自优势，首先利用 PLSR 法对水稻叶绿素含量进行预测，得到叶绿素含量的线性趋势，然后利用高斯过程回归(gaussian process regression，GPR)模型对 PLSR 模型的偏差进行预测，最后将两个模型的输出叠加得到水稻叶绿素含量的最终预测值。

## 1　实验部分

### 1.1　试验设计

试验区在辽宁省沈阳市沈北新区清水镇，位于 123°63′E，42°01′N(试验地点如图 1)，试验时间为 2018 年 6 月—9 月，粳稻品种为“秋光”。试验田设有 16 个随机分布的试验小区。试验小区设 4 个氮素水平：0，50，100 和 150 kg·hm$^{-2}$，分别记为 N0，N1，N2 和 N3，每个水平设置 4 个重复。在试验期间对水稻的返青、分蘖、拔节和抽穗等几个关键生长期分别进行采样，并对测得的叶绿素含量数据按鲁棒 3$\sigma$ 原则检测异常值，如果样本偏差大于 3 倍标准差，视为显著离群点，进行相应的剔除，同时采用 Monte Carlo 方法剔除样本中的异常光谱数据，最终得到 102 个样本。

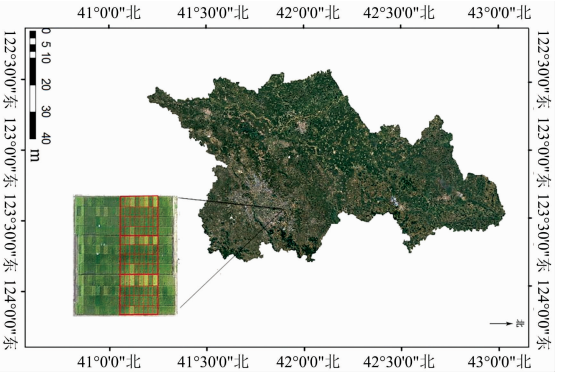


**图 1　试验地点区位图**

**Fig. 1　Location map of test site**

### 1.2　高光谱数据获取

水稻冠层光谱数据用 M600 智能六旋翼无人机搭载 GaiaSky-min 成像光谱仪测定。光谱范围 400～1 000 nm，分辨率为 2.0 nm。每次试验均选择在天气晴朗及无风的气象条件下进行，且于 10:00—14:00 时间段对光谱反射率测量，并保持与地面样本采集日期相同。光谱测量时，无人机飞行高度设置为 50 m，地面分辨率设置为 2 cm。使用 ENVI5.3 软件获得各试验小区感兴趣区的平均光谱反射率，作为该试验小区的水稻冠层光谱反射率。

作者所在团队对 PROSAIL 辐射传输模型敏感性分析的前期研究发现水稻叶绿素含量主要与 400～800 nm 区间波段的冠层光谱反射率存在较大相关性，800 nm 之后波段的光谱反射率与水稻叶绿素含量相关性不大[9]。因此，选取 400～800 nm 之间的波段进行相关研究。

### 1.3　水稻叶绿素含量测定

水稻叶绿素含量用 Spectrum752 型号-紫外可见分光光度计测定。在各试验小区中，获取冠层光谱数据的同时，选取 4 株具有代表性水稻，带回实验室。将无水乙醇和蒸馏水以 9∶2 比例制成 50 mL 的混合溶液，从带回的每个样本上选取 10 片叶片，用蒸馏水清洗干净，去除中脉剪碎后，称取 0.4 g，置于配制好的混合溶液中，避光静置 24 h 至样品完全发白。利用分光光度计测 649 和 665 nm 处的吸光度，并取 3 次测量的平均值，根据该光度值可计算水稻叶绿素含量，如式(1)—式(3)。

$$c_a = 13.95 \times D_{665} - 6.88 \times D_{649} \quad (1)$$

$$c_b = 24.96 \times D_{649} - 7.32 \times D_{665} \quad (2)$$

$$c_{ab} = c_a + c_b \quad (3)$$

式中，$c_a$ 和 $c_b$ 分别为水稻叶绿素 a 和 b 的含量(mg·L$^{-1}$)；$c_{ab}$ 为总的叶绿素含量(mg·L$^{-1}$)；$D_{649}$ 和 $D_{665}$ 为 649 nm 波段和 665 nm 波段处的吸光度值(%)。

### 1.4　水稻叶绿素光谱特征波段提取

为了减少水稻叶片结构背景噪声及样本表面纹理等因素的影响，在建模前，采用标准正态变量校正(standard normal

variate，SNV)方法对水稻冠层原始光谱处理，如图 2 所示。处理后，采用改进的优化算法选取水稻叶绿素含量特征波段。

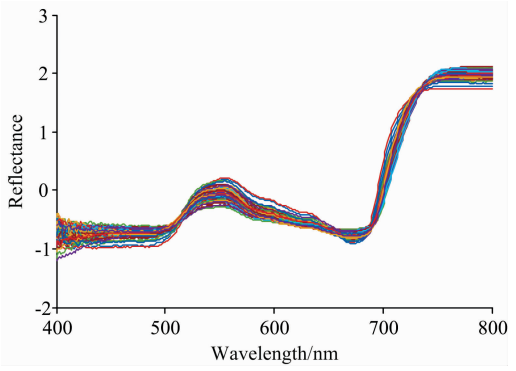


**图 2 SNV 预处理后的光谱曲线**

**Fig. 2 Spectral curves after SNV pretreatment**

1.4.1 随机蛙跳算法

随机蛙跳算法(RF)是一类解决组合优化问题的后启发式群体进化算法，通过特征变量被选择的概率来确定其重要程度[13]。该算法融合了粒子群算法(PSO)和模因演化算法(memetic algorithm，MA)各自的优点，具有调节参数少、收敛速率快、全局寻优能力强等特点，相对于差分进化、遗传算法等优化算法具有较强的竞争力。RF 的主要运算步骤为：

(1)初始化，随机产生一个包含 $Q$ 个变量(光谱波段)的初始变量集 $V_0$；

(2)基于初始变量集，选出候选变量集 $V^*$，包含 $Q^*$ 个变量；选择 $V^*$ 作为 $V_1$ 来代替 $V_0$，直到达到最大迭代次数 $G_{max}$；

(3)计算 $G_{max}$ 次迭代后各变量被选择的概率，以此作为选择变量的标准，越高越好。

(4)按被选概率降序顺序，依次联合各变量组合实施交叉验证并分别计算，若其均方根误差(root mean square error，RMSE)最小，则此时的变量组合即为选取的最佳特征波段组合。

由于标准 RF 算法的初始变量集是随机产生的，无法保证初始波段变量的有效性，很可能降低算法的收敛速度和精度。因此，为进一步提高标准 RF 算法的收敛性能，提出一种融合两种初选波段的改进型随机蛙跳算法(fusion of two primary bands-random frog，fpb-RF)，首先采用相关系数分析法和连续投影法提取特征波段，然后融合两种方法初选的特征波段作为 RF 算法的初始变量集，在此基础上利用 RF 算法进一步寻优，减少无用的迭代次数，进而改善算法的收敛性。

1.4.2 相关系数分析法

相关系数法(CC)是一种以偏最小二乘回归模型为基础进行分析的方法，主要是将各波段对应的光谱数据和水稻叶绿素含量进行相关性计算与分析，从而筛选出相关性较大的特征波段组合[11]。

1.4.3 连续投影算法

连续投影法(SPA)是一种新型的特征波段选取方法，采用投影策略筛选出的特征波段可有效消除光谱矩阵中的冗余信息，从而大大减少波段数量以简化模型、提高模型运行效率，该算法步骤主要包括 3 个阶段[12]。

第一阶段，筛选出共线性相对较小的若干组候选波段变量子集。

第二阶段，利用各子集中的变量建立多元线性回归模型，并筛选出使 RMSE 最小的变量子集。

第三阶段，对筛选出的变量子集逐步回归建模，以在保证模型精度的同时获得具有较少波段数目的集合，即为所选的特征波段。

### 1.5 叶绿素含量反演建模方法

为了提高模型的预测精度，避免对样本集的选择陷入局部最优，利用 Kennard-Stone 算法将样本数据划分为训练集和测试集，以用于建模和模型预测，具体如表 1 所示。

**表 1 水稻叶绿素含量统计表**

**Table 1 Statistics of chlorophyll content in rice**

| 样本集 | 样本数 | 最大值/(mg·L$^{-1}$) | 最小值/(mg·L$^{-1}$) | 平均值/(mg·L$^{-1}$) | 标准差/(mg·L$^{-1}$) |
|---|---|---|---|---|---|
| 总样本 | 102 | 12.811 6 | 1.702 2 | 6.772 9 | 2.338 6 |
| 训练集 | 77 | 12.811 6 | 1.702 2 | 6.700 6 | 2.416 8 |
| 测试集 | 25 | 11.686 3 | 1.934 1 | 6.989 7 | 2.117 8 |

提出一种 GPR 补偿 PLSR 的混合集成建模方法，同时采用 PLSR、BP 神经网络、最小二乘支持向量机(least square support vector machine，LSSVM)等 3 种不同方法分别建立水稻叶绿素含量的预测模型，并进行仿真对比，得到最优的预测模型。选用测试集的决定系数($R_P^2$)和均方根误差(RMSEP)对模型的预测精度及可靠性进行评价，同时融入训练集的决定系数($R_C^2$)和均方根误差(RMSEC)作为模型的辅助评价指标，$R_P^2$ 越大、RMSEP 越小，表明模型的预测效果越好。

1.5.1 偏最小二乘回归方法

偏最小二乘回归(PLSR)作为一种结合了多元线性回归分析、典型相关分析和主成分分析技术的新型多元回归方法，可以较好的解决多重共线性问题，其表达见式(4)

$$Y_i = \beta_0 + \sum_{k=1}^{r} \beta_k T_{ik} + e_i \quad (i = 1, 2, \cdots, n)$$

$$T_{ik} = \sum_{j=1}^{m} \beta C_{kj} x_{ij} \quad (k = 1, 2, \cdots, r) \qquad (4)$$

式(4)中，$x_{ij}$ 为模型输入变量，如光谱特征波段；$Y_i$ 为输出变量，即水稻叶绿素含量；$m$ 为输入变量的维数，即筛选的特征波段个数；$n$ 为样本数；$\beta_k$ 为回归系数；$e_i$ 为偏差；$T_{ik}$ 和 $C_{kj}$ 分别为第 $k$ 个潜在变量及相应系数。

1.5.2 高斯过程回归方法

高斯过程回归(GPR)是在贝叶斯和统计学习理论基础上发展起来的一种机器学习方法，在处理高维、小样本数据及非线性等问题时表现出较好的适用性[13]。

给定训练样本集 $D=\{(x_i, y_i) | i=1, 2, \cdots, n\}=(X,$

$y$)，GPR 模型如式(5)所示

$$y = f(x) + \varepsilon, \varepsilon \sim N(0, \sigma_n^2) \quad (5)$$

其中，$x_i \in R^m$ 为 $D$ 中的第 $i$ 个 $m$ 维输入变量，$y_i$ 为对应的输出变量，$\varepsilon$ 表示均值为 0，方差为 $\sigma_n^2$ 的高斯白噪声。若 $f$ 为一个高斯过程，则可由其均值和协方差函数来确定，即

$$f(x) \sim GP(0, k(x, x')) \quad (6)$$

其中 $k(x, x')$为协方差函数。

1.5.3　水稻叶绿素含量混合预测模型

建立基于 GPR 补偿 PLSR 的混合预测模型(GPR-P)充分利用了线性与非线性模型的优势，从而可实现水稻叶绿素含量的精准估测，且提高模型的稳定性，模型整体结构如图 3 所示。

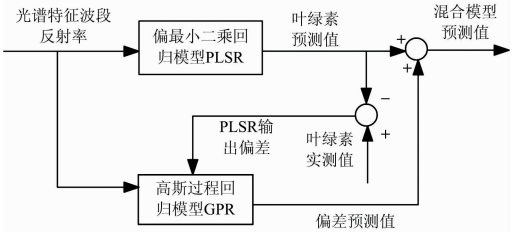


**图 3　GPR-P 整体结构图**

**Fig. 3　Overall structure diagram of GPR-P model**

具体的建模步骤为：

步骤 1：模型训练，利用训练集构建水稻叶绿素含量的 PLSR 预测模型，并通过 PLSR 的叶绿素含量预测值与实测值之间的偏差建立 GPR 预测模型，同时采用共轭梯度法优选 GPR 模型的超参数。

步骤 2：利用测试集的输入数据，通过建立的 PLSR 模型预测水稻叶绿素含量值。

步骤 3：利用输入数据，通过 GPR 模型得到叶绿素含量偏差的预测值。

步骤 4：将 PLSR 模型的输出值与 GPR 模型的偏差输出值叠加，得到混合模型预测值。

## 2　结果与讨论

### 2.1　特征波段的选择

2.1.1　基于 CC 算法的特征波段选取

将水稻叶绿素含量与对应的冠层光谱反射率进行相关性分析，如图 4 所示。图中显示了各光谱波段对应的相关系数，按相关系数绝对值的大小进行降序排列，以筛选出绝对值较大的 10 个波段，分别为 702，701，703，704，700，705，699，706，571 和 572 nm，作为特征波段。

2.1.2　基于 SPA 的特征波段选取

利用 SPA 算法对冠层光谱特征波段进行提取，最佳光谱波段数由内部交叉验证的 RMSE 值来确定。从图 5(a)可以看出，RMSE 值随着特征波段数目增加逐渐下降，当 RMSE 最小，代表运算结果最优，此时选取的特征波段数为 6。由图 5(b)看出 SPA 的征波段选取分布情况，分别为 459，481，533，648，702 和 798 nm，可将这些波段的光谱反射率作为预测模型的输入。

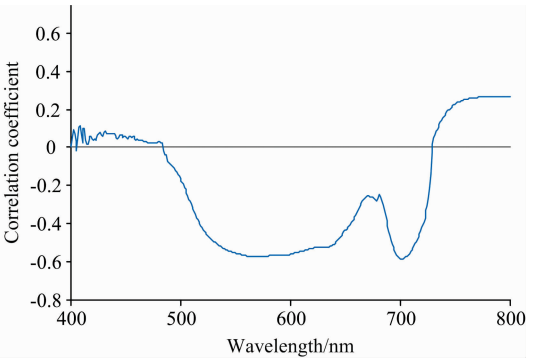


**图 4　相关系数分析法结果**

**Fig. 4　Results of correlation coefficient analysis**

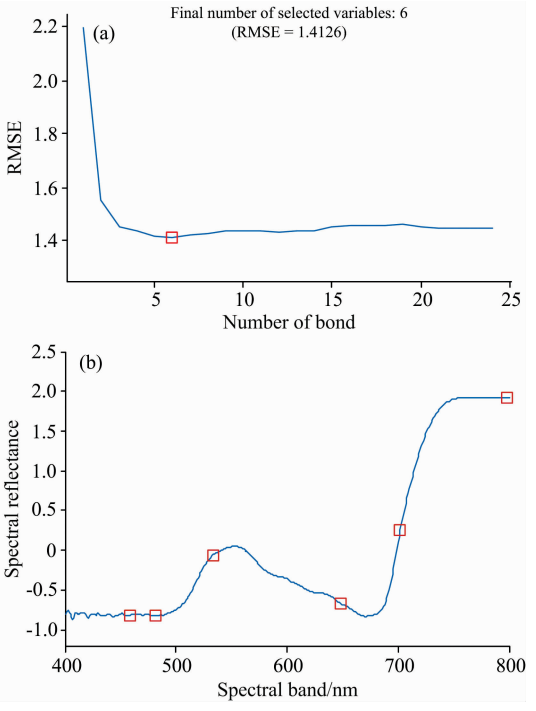


**图 5　基于 SPA 算法的特征波段选取结果**

(a)：样本模型中最佳特征波段个数；(b)：提取的特征波段分布

**Fig. 5　Feature band selection results based on SPA algorithm**

(a)：Number of the best feature bands in the sample model;
(b)：Distribution of extracted feature bands

2.1.3　基于 RF 和 fpb-RF 算法的特征波段选取

采用 Matlab R2016b 软件分析，RF 算法的参数设置：最大迭代次数为 5 000，最大主成分数为 10。模型中初始变量个数为 15，在改进算法中，经过 CC 和 SPA 算法共同选取产

**表 2　RF 和 fpb-RF 算法寻优结果**

**Table 2　Optimization results of RF and fpb-RF algorithms**

| 方法 | 迭代次数 | 收敛时间/min | 均方根误差 |
|---|---|---|---|
| RF | 5 000 | 15.2 | 1.314 0 |
| fpb-RF | 800 | 2.1 | 1.232 8 |

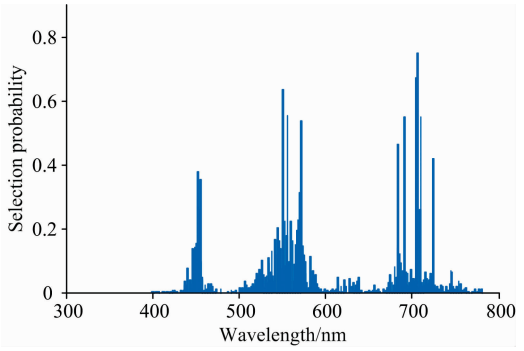


**图 6 基于 fpb-RF 算法的各波段选择概率**

**Fig. 6 Each band selection probability based on fpb-RF algorithm**

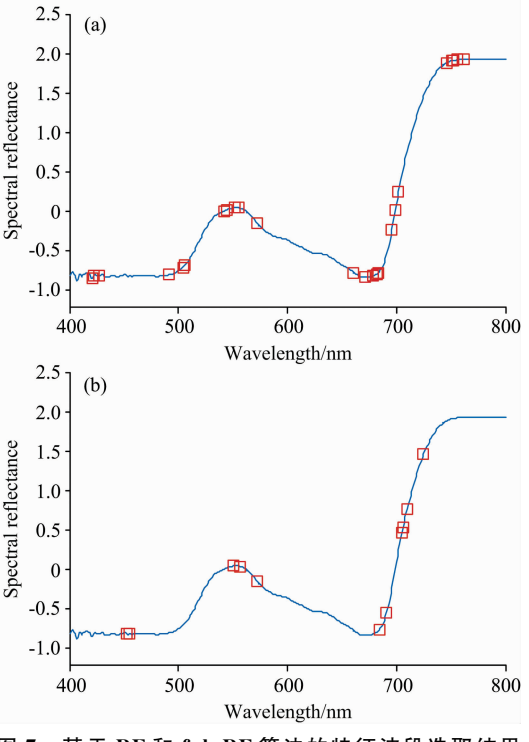


**图 7 基于 RF 和 fpb-RF 算法的特征波段选取结果**

(a)：RF 算法选取的特征波段分布；

(b)：fpb-RF 算法选取的特征波段分布

**Fig. 7 Feature band selection results based on RF and fpb-RF algorithms**

(a)：RF algorithm；(b)：fpb-RF algorithm

生的 16 个初始特征波段作为 fpb-RF 算法的初始变量集，迭代次数设为 800，分别采用两种算法提取水稻叶绿素含量特征波段，并进行对比分析，如表 2 所示。可见，标准 RF 算法迭代 5 000 次收敛，运行时间为 15.2 min，而 fpb-RF 算法迭代 800 收敛仅需 2.1 min，寻优速度极大地提高，且均方根误差较小，收敛性能相对较好。

图 6 为利用 fpb-RF 算法寻优后各波段被选择的概率，当波段数为 11 时，RMSE 有最小值 1.232 8。因此，这 11 个波段为 fpb-RF 算法选取的特征波段组合。图 7 显示了 RF 和 fpb-RF 算法选取征波段对比情况，可见，两种算法提取的特征波段所在区间大概一致，为水稻叶绿素含量的敏感区间。采用标准 RF 算法选取与水稻冠层叶片叶绿素含量相关的特征波段共有 25 个，而 fpb-RF 算法选取的特征波段为 11 个。因此，采用改进 fpb-RF 算法可在提取有用信息的同时降低输入变量维数，进而大大简化模型的复杂性。

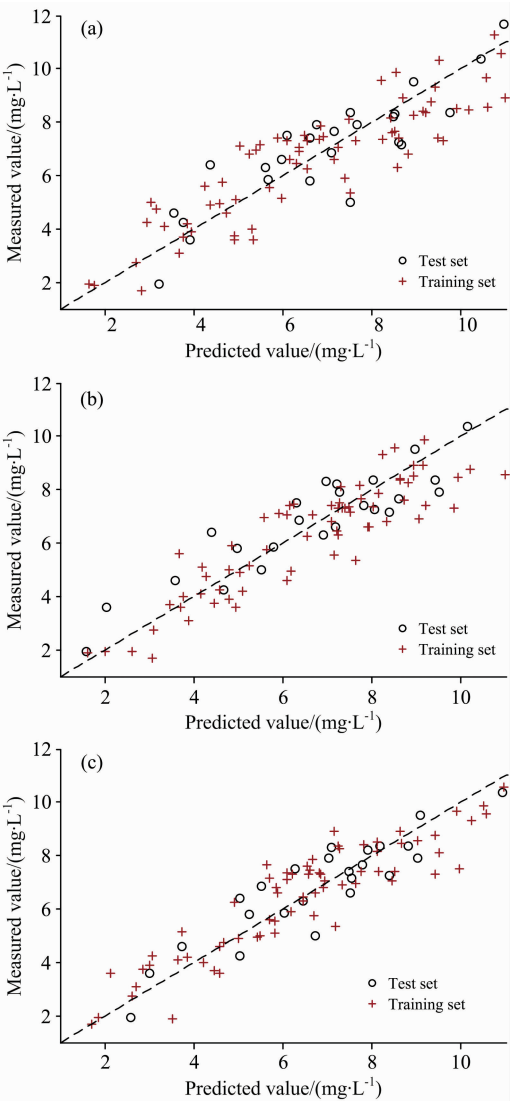


**图 8 不同特征波段提取方法的 GPR-P 模型预测结果**

(a)：CC-SPA 算法；(b)：RF 算法；(c)：fpb-RF 算法

**Fig. 8 GPR-P model prediction results of different feature band extraction methods**

(a)：CC-SPA algorithm；(b)：RF algorithm；(c)：fpb-RF algorithm

## 2.2 模型检验分析

### 2.2.1 GPR 补偿 PLSR 的混合预测模型建立

将采集并经过处理后的水稻叶绿素含量光谱反射率分别采用 CC 与 SPA 结合、标准 RF 和改进 fpb-RF 方法进行特征波段提取，将 3 种方法得到的结果作为预测模型的输入变量，水稻叶绿素含量实测值作为输出变量，建立基于 GPR 补

偿 PLSR 的混合预测模型，预测结果如图 8 所示，表 3 中给出模型的预测评价指标。

由图 8 可知，以不同方法提取的特征波段为 GPR-P 模型的输入，均具有较好的预测效果，其预测结果沿直线 $y=x$ 的分布状态表现较为优异，有较少的预测值偏离直线 $y=x$。由表 3 中可以看出 GPR-P 模型的 $R_C^2$ 和 $R_P^2$ 均高于 0.755 3，RMSEC 和 RMSEP 均低于 1.090 2 mg·L$^{-1}$。其中，采用 fpb-RF 算法提取的特征波段作为输入构建的 GPR-P 叶绿素含量模型预测精度最高，$R_C^2$ 和 $R_P^2$ 分别为 0.781 5 和 0.779 6，RMSEC 和 RMSEP 分别为 0.904 1 和 0.928 3 mg·L$^{-1}$，且优于采用标准 RF 和 CC-SPA 算法提取特征波段为输入建立的 GPR-P 模型。由此可见，在相同预测模型条件下，改进fpb-RF算法提取特征波段作为输入可以较大程度的降低模型复杂性、提高模型预测性能。

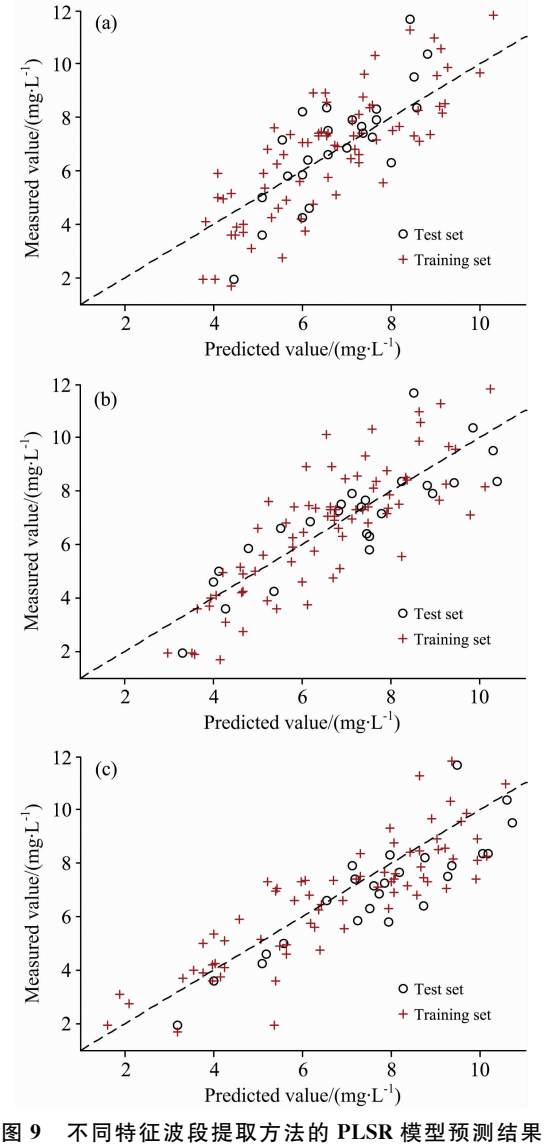


**图 9　不同特征波段提取方法的 PLSR 模型预测结果**

(a)：CC-SPA 算法；(b)：RF 算法；(c)：fpb-RF 算法

**Fig. 9　PLSR model prediction results of different feature band extraction methods**

(a)：CC-SPA algorithm；(b)：RF algorithm；(c)：fpb-RF algorithm

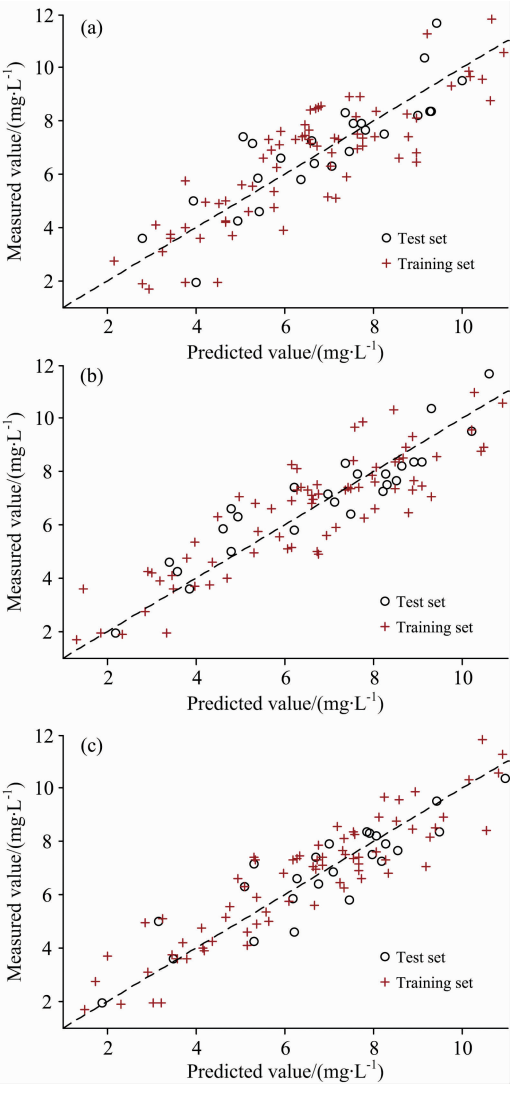


**图 10　不同特征波段提取方法的 LSSVM 模型预测结果**

(a)：CC-SPA 算法；(b)：RF 算法；(c)：fpb-RF 算法

**Fig. 10　LSSVM model prediction results of different feature band extraction methods**

(a)：CC-SPA algorithm；(b)：RF algorithm；(c)：fpb-RF algorithm

2.2.2　其他预测模型的建立及对比分析

将本文提出的 GPR-P 叶绿素含量预测模型同时与 PLSR 线性模型和应用较为广泛的 BP 神经网络、LSSVM 非线性模型进行比较，调整各模型参数至最佳状态，预测结果如图 9—图 11 所示。

由图 9—图 11 可知，对比不同特征波段提取方法建立的 PLSR，LSSVM 和 BP 神经网络叶绿素含量预测模型，利用 CC-SPA 算法提取特征波段建立的预测模型精度相对较低，且 PLSR 模型沿直线 $y=x$ 整体上较为分散，表明相应模型预测结果与实测值间存在较大的偏差。由表 3 中可以看出，基于 CC-SPA-PLSR 模型的拟合精度最低，相应的 RMSE-C 和 RMSE-P 分别为 1.377 9 和 1.398 2 mg·L$^{-1}$。而采用 fpb-RF 方法提取特征波段建立的各预测模型精度相对较高，

进一步显示了 fpb-RF 算法在提取特征波段时存在的一定优势，其中 fpb-RF-PLSR 模型的 $R_C^2$ 和 $R_P^2$ 分别为 0.717 9 和 0.704 7，RMSEC 和 RMSEP 分别为 1.232 8 和 1.275 5 mg·L$^{-1}$。fpb-RF-LSSVM 模型的 $R_C^2$ 和 $R_P^2$ 分别为 0.775 2 和 0.767 5，RMSEC 和 RMSEP 分别为 1.067 6 和 1.025 6 mg·L$^{-1}$。fpb-RF-BP 模型的 $R_C^2$ 和 $R_P^2$ 分别为 0.770 4 和 0.766 2，RMSEC 和 RMSEP 分别为 1.162 8 和 1.022 3 mg·L$^{-1}$。

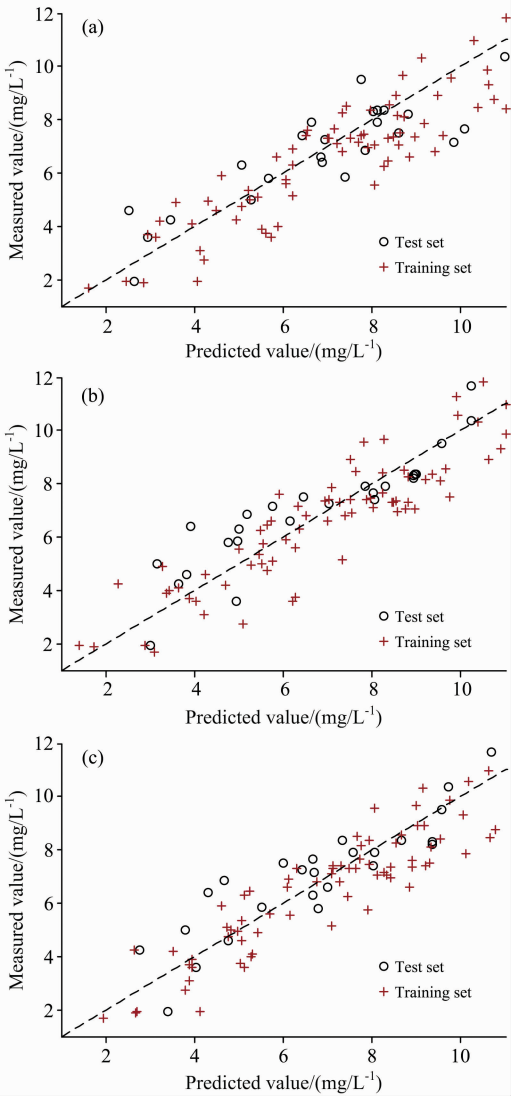


**图 11 不同特征波段提取方法的 BP 模型预测结果**

(a)：CC-SPA 算法；(b)：RF 算法；(c)：fpb-RF 算法

**Fig. 11 BP model prediction results of different feature band extraction methods**

(a)：CC-SPA algorithm；(b)：RF algorithm；(c)：fpb-RF algorithm

从整体角度看，非线性模型的预测效果优于 PLSR 线性模型，其中 fpb-RF－LSSVM 模型的预测精度较高，但与 GPR-P 模型相比，仍存在一定的差距。由此表明，在以相同特征波段为输入进行建模时，混合集成建模方法经过模型偏差补偿后，可进一步改善模型的预测性能，提高模型的精确性和稳定性。

**表 3 预测模型的评价指标**

**Table 3 Evaluation index of prediction model**

| 特征波段选取方法 | 模型 | $R^2$ | | RMSE | |
|---|---|---|---|---|---|
| | | 建模集 | 测试集 | 建模集 | 测试集 |
| CC-SPA | CC-SPA-PLSR | 0.670 7 | 0.592 5 | 1.377 9 | 1.398 2 |
| RF | RF-PLSR | 0.690 6 | 0.701 2 | 1.314 0 | 1.330 4 |
| fpb-RF | fpb-RF-PLSR | 0.717 9 | 0.704 7 | 1.232 8 | 1.275 5 |
| CC-SPA | CC-SPA-LSSVM | 0.736 3 | 0.739 2 | 1.215 1 | 1.156 9 |
| RF | RF-LSSVM | 0.760 1 | 0.757 9 | 1.184 5 | 1.071 7 |
| fpb-RF | fpb-RF-LSSVM | 0.775 2 | 0.767 5 | 1.067 6 | 1.025 6 |
| CC-SPA | CC-SPA-BP | 0.740 5 | 0.746 4 | 1.207 2 | 1.066 0 |
| RF | RF-BP | 0.763 9 | 0.765 0 | 1.184 1 | 1.048 9 |
| fpb-RF | fpb-RF-PB | 0.770 4 | 0.766 2 | 1.162 8 | 1.022 3 |
| CC-SPA | CC-SPA-GPR-P | 0.755 3 | 0.760 1 | 1.090 2 | 1.035 7 |
| RF | RF-GPR-P | 0.776 8 | 0.774 5 | 1.012 2 | 0.959 7 |
| fpb-RF | fpb-RF-GPR-P | 0.781 5 | 0.779 6 | 0.904 1 | 0.928 3 |

本工作以东北粳稻为研究对象，利用高光谱技术对水稻冠层叶片的叶绿素含量进行估测，提出了基于 fpb-RF 的特征波段提取方法，并通过与标准 RF 算法及 CC-SPA 算法比较，表明了 fpb-RF 算法在降低模型复杂性和提高模型精确性上具有一定优势。

利用 CC-SPA 算法提取的特征波段作为输入构建的叶绿素含量模型预测精度相对较低，主要原因可能在于该算法选取的特征波段相对较少且较为集中，多样性不足，进而导致部分有用信息缺失，影响模型的预测精确性。RF 算法由于具有较强的全局搜索能力，选取特征波段时，能一定程度上改善光谱信息缺失问题，提高特征波段的多样性。因此，所建模型的精度得到了明显提高。但标准 RF 算法存在收敛速度慢、提取特征波段较多等不足。本工作融合了 CC 和 SPA 方法提取特征波段，作为 RF 算法的初始变量集，在此基础上进一步寻优，可减少无用的迭代次数，改善算法收敛性，在尽可能降低模型复杂性的同时进一步提高预测模型的精确性。

采用不同特征波段提取方法建立的 LS-SVM 和 BP 神经网络非线性模型预测效果优于 PLSR 线性模型，这说明选择的特征波段和水稻冠层叶片叶绿素含量之间不仅存在线性关系，还存在非线性关系。利用非线性建模方法可进一步挖掘光谱数据与水稻叶绿素含量间隐藏的有效信息，提高模型预测精度。在此基础上，提出了基于 GPR 补偿 PLSR 的叶绿素含量混合预测模型，使得预测模型精度和稳定性进一步提高。原因在于该建模方法结合了线性与非线性模型各自的优势，其中光谱特征波段与叶绿素含量间存在的线性部分采用线性模型预测，而非线性部分通过非线性建模方法进一步补偿预测模型存在的偏差以提高模型预测性能。

从整体来看，基于 fpb-RF 选取的特征波段作为输入建立的 fpb-RF-GPR-P 模型最稳定，预测准确性最高，表明了基于高光谱的水稻叶绿素含量反演建模，改进蛙跳算法是一种较为有效的特征波段提取方法，且基于线性与非线性建模

相结合的建模方法可以进一步提高模型预测精度。本研究中还存在一定的不足之处，如在对水稻叶绿素含量预测时，目前采用是近 1～2 年数据来构建水稻叶绿素含量预测模型，但可为下一步开展长时间序列的水稻叶绿素含量预测奠定基础；此外，诸如水稻等作物叶绿素含量预测目前还没有统一的标准模型，虽然本工作构建了水稻在一些关键生育期内的整体混合预测模型，但最佳的叶绿素估测模型也会因各生育期、水稻品种及无人机的飞行高度等不同而存在一定差异。未来仍需广泛采集样本，进一步完善预测模型，以便更好的推广应用。本研究是基于典型的东北粳稻为例进行建模分析，由于受到天气、设备、技术、时间等多方面因素影响，获得样本数据有限。在后续工作中，拟进一步加强这方面的分析研究，针对无人机的不同飞行高度、多个水稻品种、不同年份水稻样本上继续测试，积累更多的试验数据，以得到更具普适性的水稻叶绿素含量的精确、稳定预测模型。

## 3　结　论

采用高光谱技术实现对水稻冠层叶片叶绿素含量的预测，提出融合两种初选波段的随机蛙跳算法 fpb-RF 选取与叶绿素含量相关的特征波段，并与标准 RF 及 CC 结合 SPA 的特征波段筛选方法进行比较，在对比分析中提出一种基于 GPR 补偿 PLSR 的叶绿素含量混合预测模型，同时构建了 PLSR，BP 和 LS-SVM 等线性与非线性预测模型。通过不同特征波段选取方法获得的输入对不同预测模型的精度进行分析，得到如下结论：

(1)在相同预测模型条件下，fpb-RF 算法提取特征波段作为输入可较好的降低模型复杂性、提高模型预测性能，使模型具有较大的 $R^2$，较小的 RMSE，如 fpb-RF-PLSR 和 CC-SPA-PLSR 相比，$R_C^2$ 和 $R_P^2$ 分别提高了 0.047 2 和 0.112 2，RMSEC 和 RMSEP 分别减小了 0.145 1 和 0.122 7 mg·$L^{-1}$。证明了 fpb-RF 算法提取特征波段建立水稻叶绿素预测模型的优越性。

(2)在相同特征波段输入进行模型建立时，本文提出的 GPR-P 模型经过偏差补偿后，可以较大程度改善模型预测性能，进一步提高模型的预测精确性和稳定性，采用不同算法提取特征波段时，建立的 GPR-P 模型 $R_C^2$ 和 $R_P^2$ 均高于 0.755 3，RMSEC 和 RMSEP 均低于 1.090 2 mg·$L^{-1}$。其中，采用 fpb-RF 方法提取的特征波段建立的 GPR-P 模型得到了最高预测精度，与其他模型相比，显示出了一定的优势。因此，本方法可为准确检测水稻叶绿素含量提供新的研究方法和思路，在精准农业领域具有良好的应用前景。

## References


[1] ZHANG Zhi-yong, MA Xu-ying, LONG Yao-wei, et al(张智勇，马旭颖，龙耀威，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2019, 50(B07): 115.

[2] Barman B, Patra S. Knowledge-Based Systems, 2020, 193: 1.

[3] Xu S X, Zhao Y C, Wang M Y, et al. Catena, 2017, 157: 12.

[4] MAO Bo-hui, LI Min-zan, SUN Hong, et al(毛博慧，李民赞，孙　红，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2017, 33(1): 164.

[5] 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.

[6] ZENG Yu-yan, SHI Run-he, LIU Pu-dong, et al(曾毓燕，施润和，刘浦东，等). Remote Sensing Technology and Application(遥感技术和应用), 2017, 32 (4): 667.

[7] Sun J, Shi S, Yang J, et al. Remote Sensing of Environment, 2018, 212: 1.

[8] CHEN Peng, FENG Hai-kuan, LI Chang-chun, et al(陈　鹏，冯海宽，李长春，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2019, 35(11): 63.

[9] Yu F H, Xu T Y, Du W, et al. International Journal of Agricultural and Biological Engineering, 2017, 12(13): 110.

[10] Yun Y H, Li H D, Wood L R E, et al. Spectrochimica Acta Part A: Molecular & Biomolecular Spectroscopy, 2013, 111(7): 31.

[11] YANG Bao-hua, CHEN Jian-lin, CHEN Lin-hai, et al(杨宝华，陈建林，陈林海，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2015, 31(22): 176.

[12] XIE Ya-ping, CHEN Feng-nong, ZHANG Jing-cheng, et al(谢亚平，陈丰农，张竞成，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(7): 247.

[13] García-Nieto P J, García-Gonzalo E, Puig-Bargués J, et al. Biosystems Engineering, 2020, 195: 198.

# Chlorophyll Content Estimation of Northeast Japonica Rice Based on Improved Feature Band Selection and Hybrid Integrated Modeling

LIU Tan[1,2], XU Tong-yu[1,2*], YU Feng-hua[1,2], YUAN Qing-yun[1,2], GUO Zhong-hui[1], XU Bo[1]

1. College of Information and Electrical Engineering, Shenyang Agricultural University, Shenyang 110161, China

2. Liaoning Agricultural Information Technology Center, Shenyang Agricultural University, Shenyang 110161, China

**Abstract** Using spectral information to detect chlorophyll content in rice canopy leaves quickly, non-destructively and accurately has a great practical significance for rice growth evaluation, precise fertilization and scientific management. In this paper, japonica rice in northeast China is taken as the research object, and rice canopy hyperspectral data of key growth stages are obtained through plot experiments. Firstly, the standard normal variate (SNV) is used to preprocess the spectral data, based on the processed spectral data and the random frog (RF) algorithm, by combining a correlation coefficient analysis method (CC) and the successive projections algorithm (SPA), an improved random frog algorithm (fpb-RF) is proposed, which combines two primary bands to select the feature bands of chlorophyll content, It is compared with the standard RF, CC and SPA methods, respectively. A hybrid prediction model (GPR-P) with gaussian process regression (GPR) compensation partial least squares regression (PLSR) is proposed: PLSR method is used to preliminarily predict the chlorophyll content in rice to obtain the linear trend of chlorophyll content, and then the GPR with good nonlinear approximation ability is used to predict the deviation of PLSR model, then the final prediction value is obtained by superposition of two outputs. To verify the superiority of the proposed method, with the feature bands by different extraction methods as inputs, PLSR, Least Square Support Vector Machine (LSSVM) and BP neural network prediction models are respectively established. The results show that under the same prediction model conditions, the improved fpb-RF algorithm can better reduce the complexity and improve the model's prediction performance by extracting feature bands as input. Both the determination coefficient ($R_P^2$) of the test set and the determination coefficient ($R_C^2$) of each model's training set are higher than 0.704 7. In addition, the $R_C^2$ and $R_P^2$ of the proposed GPR-P model are both higher than 0.755 3 when each algorithm extracts feature bands. Among them, the GPR-P model with the input of the feature band extracted by the fpb-RF method has the highest prediction accuracy, $R_C^2$ and $R_P^2$ are 0.781 5 and 0.779 6 respectively, RMSE-C and RMSE-P are 0.904 1 and 0.928 3 mg·L$^{-1}$ respectively, which provides a valuable reference for the detection and evaluation of chlorophyll content in northeast japonica rice.

**Keywords** Rice; Chlorophyll content; Spectral analysis; Feature band selection; The fpb-RF algorithm; Hybrid prediction model

(Received Jun. 4, 2020; accepted Oct. 9, 2020)

* Corresponding author