基于模型集群的马铃薯叶绿素检测光谱变量筛选讨论

刘 宁[1]，邢子正[1]，乔 浪[1]，李民赞[1]，孙 红[1*]，Qin Zhang[2]

1. 中国农业大学现代精细农业系统集成研究教育部重点实验室，北京 100083
2. Center for Precision & Automated Agricultural System，Washington State University，Pullman WA 99350，USA

摘 要 为了探究马铃薯作物叶绿素吸收特征，充分解析光谱特征波长变量，建立高精度叶绿素含量检测模型。在马铃薯发棵期(M1)、块茎形成期(M2)、块茎膨大期(M3)和淀粉积累期(M4)4个关键生长期，利用ASD便携式光谱仪采集80个样本区的314组作物冠层反射率数据，并同步采集叶片测定叶绿素含量。在光谱数据预处理之后，分析了马铃薯不同生长期的光谱反射率变化特征。利用基于模型集群思想的蒙特卡洛无信息变量消除(MC-UVE)、随机蛙跳(RF)、竞争自适应重加权采样(CARS)三种算法筛选叶绿素特征波长，建立叶绿素含量检测PLS模型。对4个生长期的314个样本，采用SPXY算法分别按照3∶1的比例划分，得到建模集240个样本、验证集74个样本。利用MC-UVE，RF，CARS三种算法筛选叶绿素特征波长，讨论迭代次数(N)和特征变量个数(LV)对MC-UVE和RF算法、迭代次数(N)对CARS算法筛选特征波长结果的影响，对迭代次数设置6个梯度，分别为N=50，100，500，1 000，5 000和10 000；对特征变量数设置4个梯度，分别为LV=15，20，25和30。以PLSR模型的验证集结果为评价指标，分析迭代次数(N)和特征变量数(LV)的最优参数组合。最后基于MC-UVE，RF和CARS算法筛选得到的最佳特征波长建立叶绿素检测PLSR模型，分别记为MC-UVE-PLSR，RF-PLSR，CARS-PLSR。结果表明，CARS，RF和MC-UVE三种算法的迭代次数(N)、特征变量数(LV)参数最佳组合分别为：(1)MC-UVE：迭代次数N=50 特征变量数LV=30；(2)RF：迭代次数N=500、特征变量数LV=30；(3)CARS：迭代次数N=100。对比在最佳特征波长建立的MC-UVE-PLSR，RF-PLSR，CARS-PLSR叶绿素含量检测，发现RF-PLSRRR模型的性能最优，R_v^2为0.786，RMSEV为3.415 mg·L^{-1}；MC-UVE-PLS模型性能次之，R_v^2为0.696，RMSEV为4.072 mg·L^{-1}；CARS-PLS模型的性能最差，R_v^2为0.689，RMSEV为4.183 mg·L^{-1}。以上结果说明：在筛选马铃薯叶绿素特征波长方面RF算法优于MC-UVE和CARS，得到的特征波长能够较全面地反映与马铃薯叶绿素相关的物质信息。

关键词 马铃薯；叶绿素检测；模型集群；光谱变量筛选；偏最小二乘(PLS)
中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)07-2259-08

引 言

叶绿素含量是评价马铃薯作物光合作用能力与营养水平的重要指标之一[1-2]。在可见光-近红外区域，分析含氢基团(O—H，N—H，C—H)振动合频和各级倍频的特性，是开展作物叶绿素、氮素、水分等参数光谱学检测的理论基础，取得了重要进展[3]。

作物叶绿素光谱学检测中，常通过筛选特征波长来达到解析光谱变量、剔除冗余信息、压缩计算量、提高诊断模型精度与鲁棒性等目的[4]。因为相关分析筛选变量存在高度自相关导致的多重共线性问题，在主成分分析的基础上，连续投影算法(successive projection algorithm，SPA)、无信息变量消除法(uninformative variables elimination，UVE)、间隔最小二乘波长选择方法(interval partial least square，iPLS)、变量投影重要程度系数法(variable importance in the projection，VIP)等算法被用于筛选特征波长并建立诊断模型[5-6]。

上述一次性建模筛选特征波长的方法，数据处理易受样

收稿日期：2019-06-13，**修订日期**：2019-10-24
基金项目：国家自然科学基金项目(31971785，31501219)，“海外名师”高端外国专家项目(MS2017ZGNY004)，中国农业大学研究生培养项目(ZYXW037，HJ2019029，JG2019004)和中央高校基本科研业务费专项资金项目(2020TC036)资助
作者简介：刘 宁，1995年生，中国农业大学信息与电气工程学院研究生 e-mail：ningliu@cau.edu.cn
*通讯联系人 e-mail：sunhong@cau.edu.cn

本个数的影响[7]。针对此问题 Li 等提出基于模型集群思想的蒙特卡洛无信息变量消除(Monte Carlo uninformative variables elimination，MC-UVE)[8]、随机蛙跳(random frog，RF)[9]、竞争自适应重加权采样(competitive adaptive reweighted sampling，CARS)[10]等变量筛选算法。有报道应用 CARS 算法设置迭代次数为 50，选取 10 个波长建立南瓜叶绿素检测模型，精度为 0.846。郑涛等[11]采用 MC-UVE 算法迭代次数为 500，选出 12 个马铃薯叶绿素特征波长。程萌等[12]基于 RF 算法筛选小麦叶绿素特征波长，迭代次数为 10 000，选出 8 个最优波长。

此类研究中尚有如下问题需要深入讨论，一方面应用不同算法选取变量是否存在差异，建立的模型是否最优且稳健；另一方面，MC-UVE，RF 和 CARS 等算法中初始参数迭代次数普遍采用固定值，修改迭代次数与其他约束是否对变量筛选结果有影响，需要开展比较和分析。

因而，在马铃薯作物叶绿素光谱学检测中，分别应用 MC-UVE，RF 和 CARS 算法，讨论迭代次数(number of iteration，N)参数和特征变量个数(latent variable，LV)对特征波长筛选结果的影响。通过建立 PLS 模型，阐明特征波长分布与叶绿素含量的解析能力，以模型验证集精度为评价标准，明确参数最优组合，以期为马铃薯叶绿素光谱降维与高鲁棒性诊断建模奠定基础，也为同类研究提供参考。

1 实验部分

1.1 材料

2018 年在北京市昌平区小汤山国家精准农业示范基地开展实验，马铃薯品种为“大西洋”。30 m×40 m 范围内设 80 个采样区，在发棵期(M1)、块茎形成期(M2)、块茎膨大期(M3)和淀粉积累期(M4)4 个生长期跟踪采集马铃薯冠层光谱并进行理化测试。

1.2 田间光谱数据采集与叶绿素含量测定

采用 ASD FieldSpec HandHeld2 便携式地物光谱仪测定 325～1 075 nm 内 751 个波长处作物冠层光谱反射率，采样间隔 1 nm，每点重复采集 3 次取平均值。同步随机采集叶片经浸提后，利用紫外分光光度计测定叶绿素含量，测定方法参考相关文献。每个生长期采集 80 组数据，其中 M1 因植被覆盖度较低导致无效数据，保留 74 组有效数据后，全生长期共获取 314 组数据。数据采集预处理总体流程如图 1 所示。其中，采用标准正态变量(standard normal variate，SNV)方法，对原始光谱曲线进行预处理来消除环境噪声的干扰。光谱与处理、特征波长筛选以及 PLSR 建模均在 matlab2014.a 环境中完成。

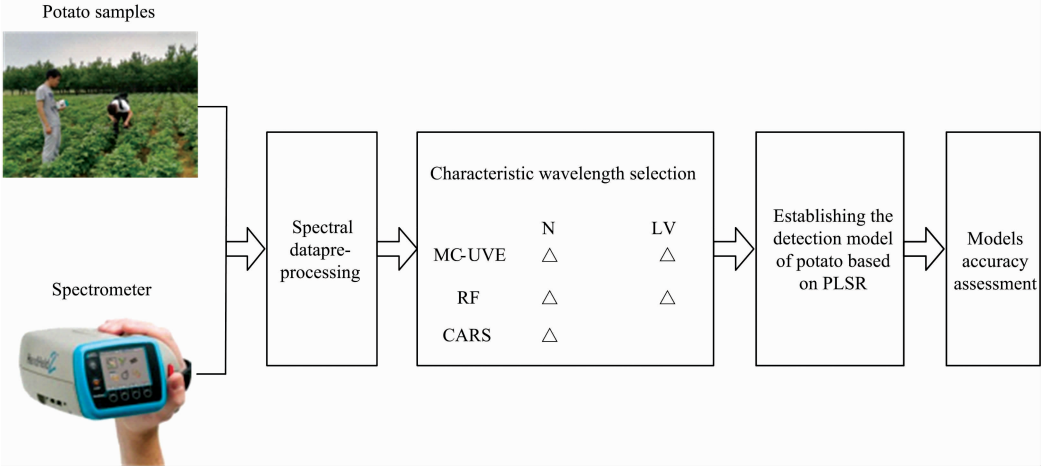

图 1 数据处理总体流程图

Fig. 1 Flow chart of data processing

1.3 光谱特征波长筛选方法

基于模型集群分析的思想，比较 MC-UVE，RF 和 CARS 3 种变量筛选算法，在 matlab2014.a libpls 软件中实现。

(1)MC-UVE 算法

MC-UVE 算法基于偏最小二乘回归(partial least squares regression，PLSR)提出，从训练集中取出一定数目(M 个)样本构建 PLS 子集，重复 M 次计算 PLS 回归系数矩阵，引入变量稳定指数为筛选标准，计算得到每个变量稳定指数值，并从高到低排序筛选变量[6]。其中，保留的 LV 数量决定着模型的预测能力和模型的稳定性。

(2)RF 算法

RF 算法类似于可逆跳转马尔可夫链蒙特卡洛。与 PLSR 相结合，通过 PLSR 结果模拟一条服从稳态分布的马尔可夫链来计算每个变量被选择的概率，从而进行重要变量的筛选[7]。

(3)CARS 算法

CARS 算法基于自适应重加权采样和指数衰减函数，选取在 PLSR 模型中回归系数绝对值大的变量，得到一系列波长变量子集；然后对每个波长子集采用交叉验证建模，从中挑选出模型均方根误差最小的子集[8]。因此 CARS 算法筛选得到的特征变量个数一定。

为了检测作物叶绿素含量，本研究以马铃薯作物为例，

对 CARS 算法的迭代次数(N)参数、RF 和 MC-UVE 算法的迭代次数(N)参数和特征变量数(LV)参数对叶绿素特征波长筛选结果的影响进行讨论。迭代次数设置 6 个梯度，分别为 N=50，100，500，1 000，5 000 和 10 000；特征变量数设置 4 个梯度，分别为 LV=15，20，25 和 30，分析迭代次数(N)和特征变量数(LV)两个参数的最优组合情况。

1.4　PLSR 模型建立与模型评价

利用偏最小二乘回归(PLSR)建模[13]，利用 SPXY(sample set partitioning based on joint X-Y distance)算法分别在 M1，M2，M3 和 M4 个生长期按照 3∶1 的比例划分样本集，采用留一交互验证法进行内部交互验证，以交叉验证均方差(root mean square error of cross validation，RMSECV)为标准选取 PLSR 模型最优特征变量数。特征波长筛选的结果以 PLSR 模型验证集模型决定系数(determination coefficients of validation set，R_v^2)以及验证集均方根误差(root mean square error of validation，RMSEV)为模型评价指标。其中 R_v^2 反映模型验证的稳定性，R_v^2 越接近于 1，说明模型的鲁棒性越高。RMSEV 用来反映模型的预测能力，RMSEV 越小，模型的预测能力越高。

2　结果与讨论

2.1　马铃薯作物生长期冠层反射光谱响应分析

SNV 校正后的各生长期的马铃薯冠层反射光谱曲线如图 2 所示，总体而言，在可见光波段，由于色素体对蓝、红光的强吸收存在 400～500 与 611～710 nm 低反射率区，并在 400 和 680 nm 附近出现吸收谷；520～610 nm 体现为色素体的强反射，550 nm 附近为绿色反射峰。受到叶肉内海绵组织结构内的空腔反射率增强影响，近红外 711～760 nm 快速攀升后进入 761～1 000 nm 高反射平台区，其中 970 nm 附近出现水分的微弱吸收谷。由 M1 至 M4 推进，在 400～500 和 740～880 nm 反射率降低；在 530～640 和 910～960 nm 反射率升高，且 M4 和 M1 分别呈现与其他生长期较大的差别。综上说明作物光谱响应是对植物生长过程中色素体、水分分子、结构等的综合表现，针对叶绿素指标，挖掘全谱中特征波长十分必要。

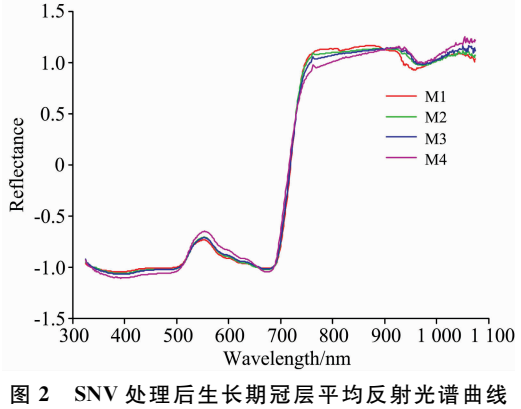

图 2　SNV 处理后生长期冠层平均反射光谱曲线

Fig. 2　Average spectrum of growth potato after SNV

2.2　相关性分析与样本集划分结果

分析光谱反射率与叶绿素含量的相关性，结果如图 3 所示。在 387～509，519～633 和 744～844 nm 波段，二者相关系数绝对值(|r|)均高于 0.6，在 678 nm 达正相关峰值 0.411；在 702 nm 存在负相关峰值−0.715。845～917 nm 正相关系数逐渐降低，917 nm 之后呈负相关。此结果与叶绿素吸收可见光蓝、红光，反射绿光的物理现象一致，但相关性曲线显示相邻波长之间的相关系数接近。若选取相关系数较高者为特征波长，会存在波长冗余与多重共线性问题。因此，利用 SPXY 算法划分样本集结果如表 1 所示，后续建模开展特征波长变量筛选方法讨论，用建模集筛选特征波长、建立回归模型，以验证集的结果评价特征波长筛选结果。

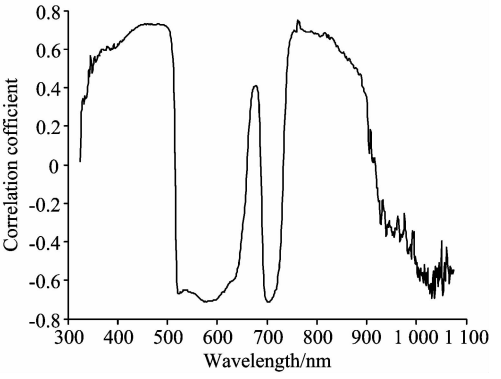

图 3　光谱反射率与叶绿素含量相关性曲线

Fig. 3　Correlation between reflectance and chlorophyll content

表 1　建模集与验证集划分统计

Table 1　Statistical results of calibration set and validation set

数据集	样本量	最大值	最小值	平均值	标准差
总样本	314	41.20	7.66	24.05	7.95
建模集	240	41.20	7.66	24.07	7.95
验证集	74	37.46	8.20	24.00	8.00

2.3　基于模型集群分析的马铃薯叶绿素特征波长筛选

2.3.1　MC-UVE 算法

由于 MC-UVE 算法对于同一批光谱数据，设置同样的迭代次数，运行多次计算变量的稳定指数不一致，因此分别讨论迭代次数(N)和特征波长数量(LV)的影响。

首先设置 N 分别为 50，100，500，1 000，5 000 和 10 000次 6 个迭代梯度，运行 5 次计算各个波长变量所对应的稳定指数平均值。以 N=500 为例，运行结果如图 4 所示，稳定指数越高代表此波长变量越具有信息价值。然后对 6 个迭代梯度改变 LV，按照稳定指数从高到低选择 LV 为 15，20，25 和 30 个建立马铃薯叶绿素检测 PLSR 模型。共得到 24 种模型，结果如表 2 所示。当 N 值增加并未有效提升检测模型精度，但是 LV 增加时，建模特征变量增多可以提升建模型精度。其中 N=50 时其特征波长的位置分布如图 5 所示，精度最优的模型为 LV=30 时，预测集精度 R_v^2 为 0.696，验证集均方根误差 RMSEV 为 4.072 mg·L^{-1}。

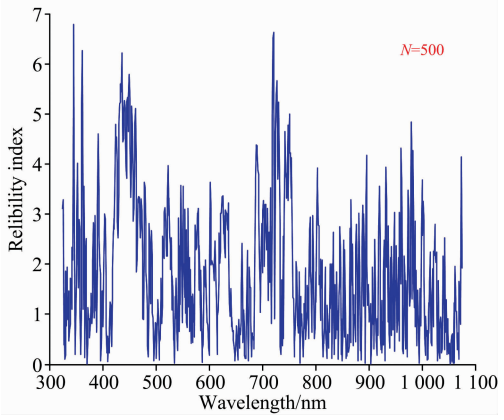

图 4 MC-UVE 算法在迭代次数为 500 时的运行结果

Fig. 4 Run results of MC-UVE at $N=500$

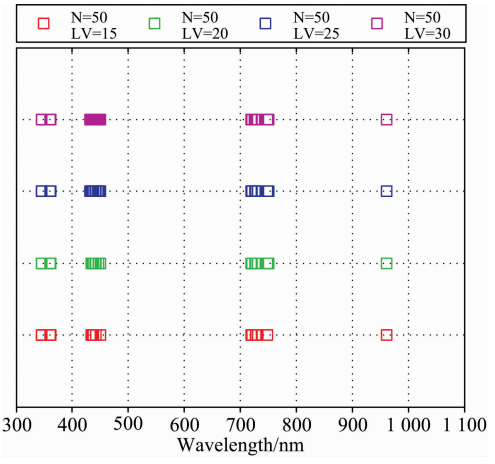

图 5 MC-UVE 在 LV 梯度下最佳迭代次数时特征波长位置

Fig. 5 Wavelengths selected by MC-UVE at LV gradients

由图 5 对比相关性分析结果可知，$N=50$ 时 LV 从 15 增至 30 过程中，被选取的波长数在 400～500 和 720～800 nm 范围增加。该区间相关系数绝对值$|r|>0.6$，但是并未体现红光吸收特征，使得该类马铃薯叶片叶绿素含量诊断模型验证集 R_v^2 约为 0.60～0.70 之间。

表 2 基于 MC-UVE 的叶绿素含量检测 PLSR 模型验证集结果(RMSEV：mg·L^{-1})

Table 2 PLSR validation results on the chlorophyll content detection using MC-UVE (RMSEV：mg·L^{-1})

迭代次数 (N)	LV=15		LV=20		LV=25		LV=30	
	R_v^2	RMSEV	R_v^2	RMSEV	R_v^2	RMSEV	R_v^2	RMSEV
50	**0.656**	**4.327**	**0.670**	**4.243**	**0.687**	**4.133**	**0.696**	**4.072**
100	0.619	4.566	0.655	4.345	0.665	4.278	0.68	4.183
500	0.599	4.681	0.642	4.429	0.640	4.438	0.672	4.236
1 000	0.608	4.632	0.624	4.536	0.650	4.374	0.649	4.383
5 000	0.611	4.611	0.637	4.456	0.649	4.380	0.645	4.408
10 000	0.626	4.521	0.647	4.398	0.649	4.383	0.636	4.460

2.3.2 RF 算法

RF 算法与 MC-UVE 算法类似，首先讨论迭代次数 N 的影响，分别设置 N 为 50，100，500，1 000，5 000 和10 000 次 6 个梯度，运行 5 次取平均值。以 $N=10\ 000$ 为例的运行结果如图 6 所示，纵坐标为每个波长的被选择概率(selection probability)，被选择概率越高说明波长越重要。其次讨论波长个数 LV 的影响，按照选择概率从大到小设置 LV 分别为 15，20，25 和 30 建立马铃薯叶绿素检测 PLS 模型，共得到 24 种模型。

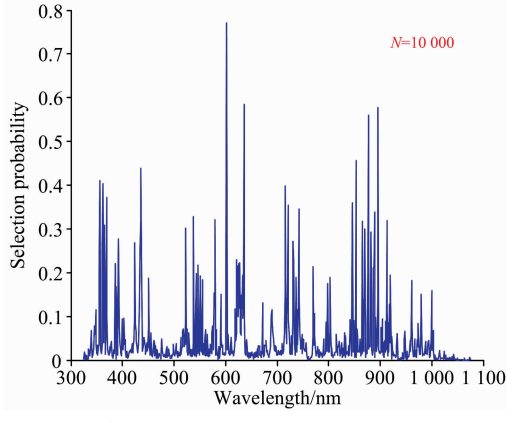

图 6 RF 算法在迭代次数为 10 000 时的运行结果

Fig. 6 Results of RF algorithmat $N=10\ 000$

结果如表 3 所示，整体而言 N 值增加使模型精度提升有限，LV 增加模型精度上升趋势明显。较优的组合选取特征波长结果如图7所示，包括：$N=500$时LV为15或30组合，与

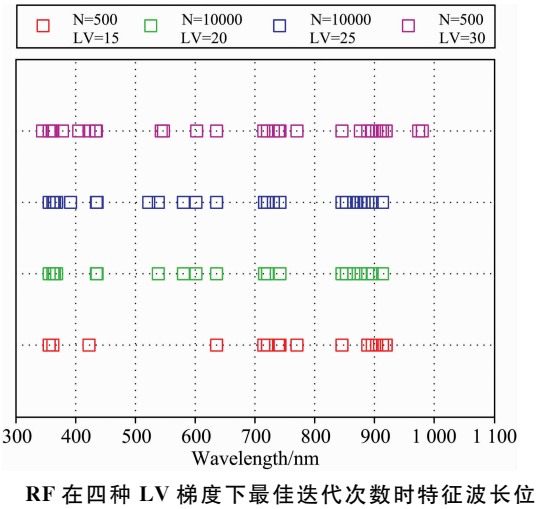

图 7 RF 在四种 LV 梯度下最佳迭代次数时特征波长位置

Fig. 7 Wavelengths selected by RF algorithm at LV gradients

表 3　基于 RF 在不同输入参数下的叶绿素含量检测 PLSR 模型验证集结果(RMSEV：mg·L^{-1})

Table 3　PLSR validation results on the chlorophyll content detection using RF with different setting(RMSEV：mg·L^{-1})

迭代次数	LV=15		LV=20		LV=25		LV=30	
(N)	R_v^2	RMSEV	R_v^2	RMSEV	R_v^2	RMSEV	R_v^2	RMSEV
50	0.682	4.162	0.688	4.116	0.713	3.957	0.708	3.990
100	0.694	4.090	0.716	3.942	0.709	3.991	0.739	3.781
500	**0.715**	**3.949**	0.740	3.770	0.766	3.574	**0.786**	**3.415**
1 000	0.696	4.044	0.730	3.835	0.746	3.719	0.760	3.620
5 000	0.690	4.117	0.727	3.867	0.775	3.512	0.779	3.480
10 000	0.673	4.223	**0.754**	**3.667**	**0.775**	**3.510**	0.773	3.520

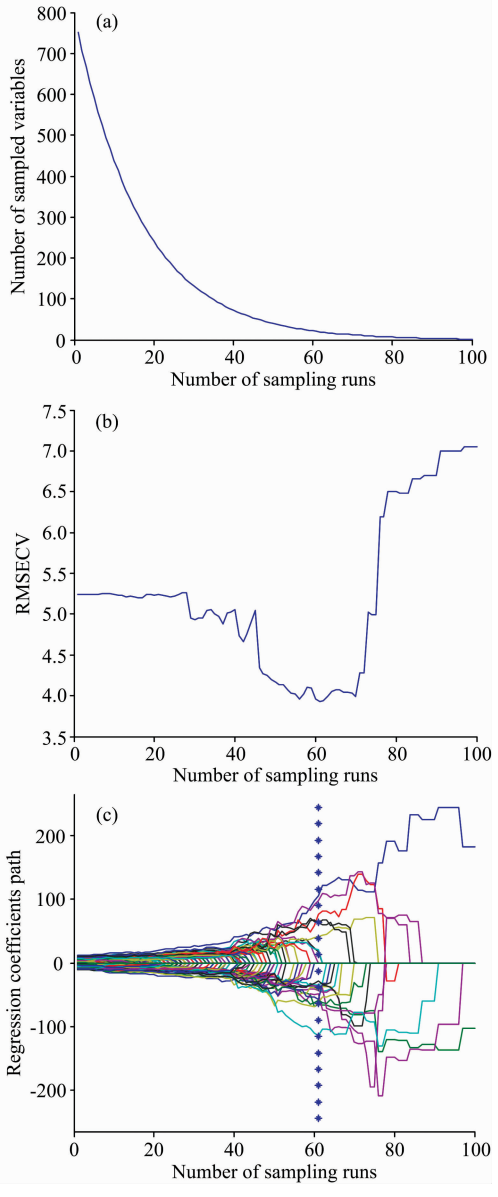

图 8　CARS 在迭代次数 N 为 100 时运行结果

(a)：选择变量个数变化趋势；(b)：RMSECV 变化趋势；(c)回归系数变化趋势

Fig. 8　Results of CARS algorithm at N=100

(a)：Trend of selected variable number；(b)：Trend of RMSECV；(c)：Trend of regression coefficient

N=10 000 时 LV 为 20 或 25 组合，马铃薯叶片叶绿素含量诊断模型验证集 R_v^2 约为 0.68～0.79 之间。其中，最优模型 N 为 500、LV 为 30 时验证集精度 R_v^2 为 0.786，验证集均方根误差 RMSEV 为 3.415 mg·L^{-1}。

由图 7 对比相关性分析结果可知，在 LV 从 15 增至 30 过程中分布愈加广泛，反映的信息愈加全面。在 LV=15 时，在绿光区域没有筛选到特征波长，而在 LV=20，25 和 30 时，筛选到的特征波长在蓝、绿、红区域均有分布。LV=30 时，970 nm 附近反映水分弱吸收的波长被选中，说明该方法筛选波长对含氢基团具有较好的选择性。

2.3.3　CARS 算法

CARS 算法与 RF 和 MC-UVE 不同，对于同一批数据，在相同的迭代次数(N)下变量筛选结果唯一，所以仅考虑设置 N 为 50，100，500，1 000，5 000 和 10 000 次 6 个梯度。N=100 时的运行结果如图 8 所示，图 8(a)为筛选过程中变量数随着迭代次数 N 的变化曲线，筛选的波长数(LV)随运行次数的增加而减少；图 8(b)为 RMSECV 随着迭代次数的变化曲线，在前 30 次时 RMSECV 保持不变，30 次后下降，在迭代 61 次时 RMSECV 的值最小为 3.928，之后逐步攀升；图 8(c)为各光谱波长的回归系数的变化趋势，其中“* *”列表示 RMSECV 最小时所对应的迭代运行次数。运行后得到的波长变量集采用交叉验证，根据 RMSECV 的值来确定最优波长变量子集为 21 个特征波长。

CARS 算法筛选得到的特征波长位置如图 9 所示，分别

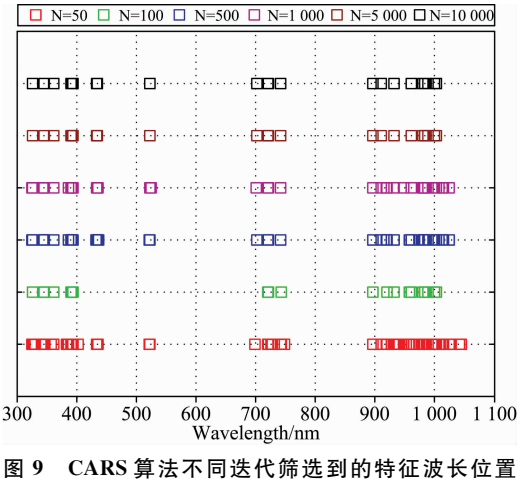

图 9　CARS 算法不同迭代筛选到的特征波长位置

Fig. 9　Location of wavelengths selected by CARS

建立叶绿素诊断 PLSR 模型结果如表 4 所示，当 N 值增加时 CARS 筛选得到的最优变量个数呈先上升后下降趋势；参与建模波长数增加并不能提升诊断模型精度，马铃薯叶片叶绿素含量诊断模型 R_v^2 约为 0.63～0.69。

表 4 基于 CARS 算法不同迭代次数的叶绿素含量检测 PLSR 模型验证集结果(RMSEV：mg·L^{-1})

Table 4 PLSR validation results on the chlorophyll content detection with iteration of CARS(RMSEV：mg·L^{-1})

迭代次数(N)	最佳迭代次数	特征波长个数(LV)	R_v^2	RMSEV
50	21	67	0.645	4.408
100	**61**	**21**	**0.689**	**4.183**
500	249	39	0.636	4.460
1 000	502	38	0.649	4.379
5 000	2960	22	0.680	4.294
10 000	5918	22	0.672	4.301

由图 9 对比相关性分析结果可知，在 N=50 时，筛选得到的特征波长存在显著“边缘指纹效应”，即在 325～500 和 900～1 100 nm 区域信息冗余与噪声降低了模型精度和鲁棒性，所以 N=50 时模型的验证集精度最低。与其他迭代次数相比，N=100 时在 900～1 000 nm 范围波长数精简，冗余信息较少；N=5 000 和 N=10 000 时，筛选得到的特征波长基本一致。其中，N=100 时选取 21 个特征波长建立的模型最优，预测集精度 R_v^2 为 0.689，验证集均方根误差 RMSEV 为 4.183 mg·L^{-1}。

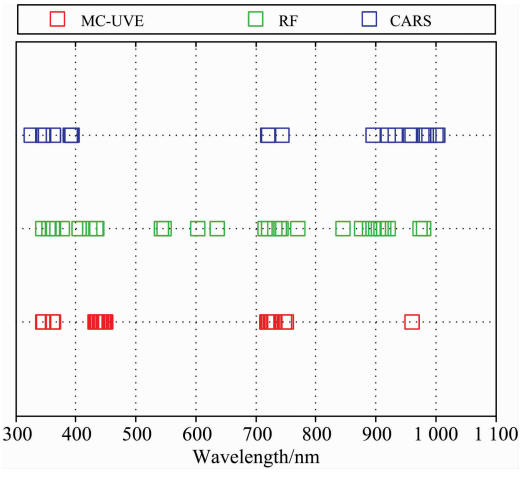

图 10 MC-UVE，RF 和 CARS 最优特征波长位置

Fig. 10 Wavelengths selected by optimal MC-UVE，RF，CARS

2.4 三种波长筛选方法建模比较与讨论

综合上述三种方法，当 N 值或 LV 个数增加时模型验证精度得到一定提升，其中 LV 对 MC-UVE 算法影响较为显著，马铃薯叶片叶绿素含量诊断验证集 R_v^2 约为 0.60～0.70 之间；N 和 LV 对 RF 算法均有显著影响，R_v^2 约为 0.68～0.79 之间；CARS 算法主要考虑 N 值设置，R_v^2 约为 0.63～0.69 之间。将基于 MC-UVE，RF 和 CARS 算法在最佳 N 和 LV 组合下筛选特征波长，分别建立叶绿素含量检测模型记为 MC-UVE-PLSR，RF-PLS 和 CARS-PLSR，结果如表 5 所示。其中，RF-PLSR 模型最优，验证集精度 R_v^2 为 0.786，验证集均方根误差为 3.415，MC-UVE-PLSR 模型次之，CARS-PLSR 模型最差。

对比分析 MC-UVE，RF 和 CARS 筛选的最优特征波长，位置如图 10 所示，从特征波长分布角度，在可见光范围(400～710 nm)，RF 算法筛选波长分布均匀；MC-UVE 算法对 550 nm 附近绿光区域不敏感，而在 450 nm 附近蓝光区域“波长聚集”现象显著；CARS 算法对该区域筛选变量较少。在近红外区域(711～1 100 nm)，RF 算法得到的特征波长分布仍然较为均匀；MC-UVE 在 800～1 000 nm 只筛选到一个特征波长；CARS 筛选到的特征波长均聚集在 900～1 000 nm 内。综上说明 RF 算法在可见光和近红外区筛选得到的特征波长对叶绿素光谱吸收和反射等特征具有较为全面的代表性。

从相关性的角度考虑，RF 算法筛选得到的特征波长在叶绿素高相关范围(387～509，519～633，744～844 和 845～917 nm)和相关性峰值(702 nm)均有分布。而 MC-UVE 算法筛选变量只在 387～509 和 744～844 nm 两个范围，CARS 算法筛选变量则只有 391，392，393，394 和 896 nm 五个波长落入高相关性范围内，且前四个为相邻波长而存在波长信息冗余。上述结果在 PLSR 模型中也得到了验证，RF-PLSR 模型的精度最优，MC-UVE-PLSR 模型次之，CARS-PLSR 模型最差。

表 5 MC-UVE-PLSR，RF-PLSR 和 CARS-PLSR 验证集结果

Table 5 Results of validation set of MC-UVE-PLS，RF-PLS and CARS-PLS

模型	N	LV	R_v^2	RMSEV
MC-UVE-PLSR	50	30	0.696	4.072
RF-PLSR	**500**	**30**	**0.786**	**3.415**
CARS-PLSR	100	22	0.689	4.183

综上表明，当合理选择 N 和 LV 参数时，RF 算法对马铃薯叶绿素特征波长筛选能力优于 MC-UVE 和 CARS 两种算法，同时也避免了高相关性区间筛选相邻波长存在的高度自相关导致的多重共线性问题。所建立的 RF-PLSR 模型可为马铃薯叶绿素含量诊断提供支持，而研究讨论的变量筛选方法与参数分析过程，可为其他同类光谱学检测提供参考。

3 结 论

为了高精度地检测马铃薯作物叶绿素含量，利用基于模型集群思想的 CARS，RF 和 MC-UVE 三种算法筛选叶绿素特征波长，建立叶绿素含量检测 PLS 模型。以 PLS 模型验证集结果为评价指标，讨论三种算法的迭代次数(N)和特征变量个数(LV)参数对模型结果的影响，确定三种算法的最佳输入参数组合，对比分析 MC-UVE，RF 和 CARS 筛选的最优特征波长，结论如下：

对叶绿素含量和光谱数据做相关性分析，发现在 387～509，519～633 和 744～844 nm 三个波段内，叶绿素含量与光谱反射率的相关系数较高，其相关系数绝对值均高于 0.6；在 678 和 702 nm 处存在相关性极值，相关系数分别为 0.411 和－0.715。

当 N 值或 LV 个数增加时各算法模型验证精度得到提升，其中 LV 对 MC-UVE 算法影响较为显著，马铃薯叶片叶绿素含量诊断验证集 R_v^2 约为 0.60～0.70 之间；N 和 LV 对 RF 算法均有显著影响，R_v^2 约为 0.68～0.79 之间；CARS 算法主要考虑 N 值设置，R_v^2 约为 0.63～0.69 之间。

对比分析基于三种算法选取最优特征波长建立的叶绿素 PLS 检测模型可知：迭代次数 N＝500、特征变量数 LV＝30 时，RF-PLSR 模型性能最优，R_v^2 为 0.786，RMSEV 为 3.415 mg·L^{-1}；迭代次数 N＝50、特征变量数 LV＝30 时，MC-UVE-PLSR 模型 R_v^2 为 0.696，RMSEV 为 4.072 mg·L^{-1}；迭代次数 N＝100 时 CARS-PLSR 模型 R_v^2 为 0.689，RMSEV 为 4.183 mg·L^{-1}。说明 RF 算法筛选得到的特征波长能够更加全面的反映与马铃薯叶绿素相关的物质信息，所以基于 RF 建立的模型精度优于 MC-UVE 和 CARS 两种算法。

References

[1] Dong L, Xue W, Hengbiao Z, et al. Plant Methods, 2018, 14(1): 76.
[2] Jin Jia, Wang Quan. IEEE Transaction on Geoscience and Remote Sensing, 2019, 57(5): 3064.
[3] Roosjen P P J, Brede B, Suomalainen J M, et al. International Journal of Applied Earth Observation and Geoinformation, 2018, 66: 14.
[4] Di W, Chen X, Zhu X, et al. Analytical Methods, 2011, 3(8): 1790.
[5] Liu Y F, Chen X, Zheng B, et al. Advanced Materials Research, 2013, 726-731: 4337.
[6] Farrés M, Platikanov S, Tsakovski S, et al. Journal of Chemometrics, 2015, 29(10): 528.
[7] Li H D, Liang Y Z, Xu Q S, et al. Journal of Chemometrics, 2010, 24(7-8): 418.
[8] Cai W, Li Y, Shao X. Chemometrics and Intelligent Laboratory Systems, 2008, 90(2): 188.
[9] Li H D, Xu Q S, Liang Y Z. Analytica Chimica Acta, 2012, 740: 20.
[10] Li H D, Xu Q S, Liang Y Z. Chemometrics and Intelligent Laboratory Systems, 2018, 176: 34.
[11] ZHENG Tao, LIU Ning, SUN Hong, et al(郑　涛，刘　宁，孙　红，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2017, S1(48): 153.
[12] CHENG Meng, ZHANG Jun-yi, LI Min-zan, et al(程　萌，张俊逸，李民赞，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2017, S1(33): 157.
[13] Elsayed S, Rischbeck P, Schmidhalter U. Field Crops Research, 2015, 177: 148.

Discussion on Spectral Variables Selection of Potato Chlorophyll Using Model Population Analysis

LIU Ning[1], XING Zi-zheng[1], QIAO Lang[1], LI Min-zan[1], SUN Hong[1*], Qin Zhang[2]

1. Key Laboratory of Modern Precision Agriculture System Integration Research, China Agricultural University, Beijing 100083, China
2. Center for Precision & Automated Agricultural System, Washington State University, Pullman WA 99350, USA

Abstract The paper was aimed to explore the chlorophyll spectral absorption characteristics of potato crops, fully analyze the spectral characteristic wavelength variables, and establish a high－-precision chlorophyll content detection model. The 314 reflectance samples were collected using an ASD portable spectrometer at the seedling stage (M1), tuber formation stage (M2), tuber expansion stage (M3) and starch accumulation stage (M4). The chlorophyll content was determined by the simultaneous collection of leaves. After spectral data pre－-treatment, the spectral reflectance changes of different growth stages of potato were analyzed. The algorithms based on model population analysis were used to select chlorophyll characteristiccharacteristic chlorophyll wavelengths, including Monte Carlo uninformative variables elimination (MC－-UVE), random frog (RF) and competitive adaptive reweighted sampling (CARS) algorithm. The partial least square regression (PLSR) was used to establish the chlorophyll content detection model. The sample set was divided by a ratio of 3∶1 in each growth stage using the sample set partitioning based on joint X-Y distance algorithm (SPXY) with the 240 calibration samples and 74 validation samples. The different algorithms (MC-UVE, RF, CARS) were used to select chlorophyll characteristic wavelengths. The influence of the number of iteration (N) and the number of the latent variables (LV) on the results of characteristic wavelength selection of MC-UVE

and RF algorithms were discussed, and the influences of N on that of CARS algorithm were discussed. Six gradients were set for the number of iterations (N), which were N=50, 100, 500, 1 000, 5 000 and 10 000, respectively. Four gradients were set for the number of latent variables (LV), which were LV=15, 20, 25 and 30 respectively. Taking the validation set result of PLS model as the evaluation index, the optimal parameter combination of N and LV was analyzed. Based on the optimal characteristic wavelengths selected by the three algorithms, the chlorophyll detection PLSR models were established and denoted as RF-PLSR, MC-UVE-PLSR, and CARS-PLSR, respectively. The research results showed that the chlorophyll characteristic wavelengths selection results were optimal when N=50 and LV=30 of MC-UVE, N=500 and LV=30 of RF, N=100 of CARS. By comparing the RF-PLSR, MC-UVE-PLSR, and CARS-PLSR models, it was indicated that the performance of the RF-PLSR model was best, the determination coefficient of validation (R_v^2) was 0.786, the root means square error of validation (RMSEV) was 3.415 mg·L^{-1}; MC-UVE-PLSR was second, the R_v^2 was 0.696, the RMSEV was 4.072; and the CARS-PLSR was the worst, the R_v^2 was 0.689, the RMSEV was 4.183. Above results showed that the RF algorithm was superior to MC-UVE and CARS in selecting the characteristic chlorophyll wavelength of potato.

Keywords Potato; Chlorophyll detection; Model population analysis; Band selection; Partial least square (PLS)

(Received Jun. 13, 2019; accepted Oct. 24, 2019)

* Corresponding author

敬告读者——《光谱学与光谱分析》已全文上网

从 2008 年第 7 期开始在《光谱学与光谱分析》网站(www.gpxygpfx.com)“在线期刊”栏内发布《光谱学与光谱分析》期刊全文，读者可方便地免费下载摘要和 PDF 全文，欢迎浏览、检索本刊当期的全部内容；并陆续刊出自 2004 年以后出版的各期摘要和 PDF 全文内容。2009 年起《光谱学与光谱分析》每期出版日期改为每月 1 日。

《光谱学与光谱分析》期刊社