第42卷，第6期 光 谱 学 与 光 谱 分 析 Vol 42，No. 6，pp1956-1964
2022年6月 Spectroscopy and Spectral Analysis June，2022

基于集成学习的水稻氮素营养及籽粒蛋白含量监测

张 杰[1,2]，徐 波[1]，冯海宽[1]，竞 霞[2]，王娇娇[1]，明世康[1]，傅友强[3]，宋晓宇[1*]

1. 北京市农林科学院信息技术研究中心，北京 100094
2. 西安科技大学测绘科学与技术学院，陕西 西安 710054
3. 广东省农业科学院水稻研究所，广东 广州 510640

摘 要 利用高光谱遥感技术在水稻收获前对籽粒品质相关的蛋白质含量进行监测，一方面可以及时调整栽培管理方式，指导合理追肥，另一方面，有助于提前掌握籽粒品质信息，明确市场定位。该研究以广东省典型优质籼稻为研究目标，基于2019年和2020年两年氮肥梯度实验，以水稻分化期和抽穗期冠层尺度高光谱数据、水稻氮素参数，包括叶片氮素含量(LNC)、叶片氮素积累量(LNA)、植株氮素含量(PNC)、植株氮素积累量(PNA)及籽粒蛋白含量数据为基础，利用四种个体机器学习算法 partial least square regression (PLSR)、K-nearest neighbor (KNN)、Bayesian ridge regression (BRR)、support vector regression (SVR)，三种集成学习算法 random forest (RF)、adaboost、bagging，针对水稻不同生育期氮素状况进行监测建模，在此基础上构建基于水稻冠层光谱信息、光谱信息结合水稻农学氮素参数的籽粒蛋白含量的监测模型，并对模型进行精度对比。研究结果表明，在水稻氮素营养监测方面，利用水稻冠层454～950 nm波段信息，采用RF及Adaboost算法，在水稻分化期、抽穗期及全生育期LNC、LNA、PNC及PNA模型 R^2 均达到0.90以上，同时也具有较低的RMSE和MAE。在水稻籽粒蛋白品质监测方面，采用全波段光谱信息进行籽粒蛋白含量监测时，RF具有最高的精确度与稳定性，两生育期的RF模型对籽粒蛋白含量的监测结果 R^2 分别为0.935和0.941，RMSE分别为0.235和0.226，MAE分别为0.189和0.152；两生育期以全波段光谱信息结合长势参数进行籽粒蛋白监测时，Adaboost模型具有最高的精确度和稳定性，其中分化期全波段光谱信息结合PNA作为输入参数，Adaboost模型 R^2 为0.960，RMSE为0.175，MAE为0.150，以抽穗期全波段光谱信息结合PNC作为输入参数，R^2 为0.963，RMSE为0.170，MAE为0.137。研究结果表明，与PLSR，KNN，BRR和SVR几种个体学习器算法相比，集成算法RF，Adaboost和Bagging具备良好的处理多重共线性的能力，适合用于高光谱数据的分析与处理，在作物氮素营养监测及水稻品质的早期遥感监测方面具有明显优势。

关键词 高光谱遥感；水稻品质；机器学习；集成算法；Adaboost；Random forest
中图分类号：TP79 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)06-1956-09

引 言

水稻是我国种植面积最大、覆盖范围最广的粮食作物，长期以来我国注重水稻产量发展，以解决人民生活温饱为目标，近年来，随着人民生活消费水平不断提高，对稻米的需求也从过去的“高产”向“品质-食味”转变，获取高产、高品质的水稻是我国实施精准农业的新要求[1]。稻米蛋白质含量是决定水稻营养品质的重要指标，氮素是水稻生长发育的关键因素，既影响水稻营养代谢、物质积累，也影响稻米最终营养及食味品质。高光谱遥感技术是实现对水稻氮素营养及品质进行绿色无损监测的重要途径，建立水稻氮素营养及籽粒蛋白含量监测模型，是精准制定田间管理措施、快速评估水稻品质的可靠依据[2]。一些学者已利用高光谱遥感对水稻氮素状况及籽粒蛋白含量等品质相关参数进行了深入研究[3-4]。张浩等采用主成分分析(principal component analysis，PCA)对200～1 100 nm之间的光谱波段进行降维，选用贡献率最高的两个主成分作为模型的输入变量，对水稻叶片氮素含量

收稿日期：2021-10-27，**修订日期**：2022-02-25
基金项目：广东省重点领域研发计划项目(2019B020214002)，国家自然科学基金项目(42171394)资助
作者简介：张 杰，1997年生，西安科技大学测绘科学与技术学院硕士研究生 e-mail：zhangj0216@163.com
*通讯作者 e-mail：songxy@nercita.org.cn

及籽粒蛋白含量进行预测，模型决定系数 R^2 达到 0.847 以上[5]。孙雪梅等探讨了不同氮素水平下的水稻光谱曲线特征，利用统计相关分析，研究了 9 个植被指数和 8 个微分参数与叶片叶绿素、全氮含量的相关关系，建立了叶绿素和全氮含量的监测模型，并通过叶绿素监测模型间接对水稻籽粒蛋白含量进行预测[6]。刘芸等分析了米粉光谱与蛋白质含量、直链淀粉含量的关系，通过提取敏感波段的特征参数，利用多元逐步回归构建的模型决定系数 R^2 达到 0.7 以上，检验精度也达到了 80%以上[7]。

目前对高光谱数据的处理大多通过主成分分析等方法实现对数据的特征降维，以较少的特征参数参与模型的构建，在降低模型构建难度的同时可能会丢失部分有效信息，而机器学习模型具备强大的处理高维数据与冗余数据的能力，基于更高效益的数学方法与数据处理方式实现对数据中有效信息的提取[8-9]。Bao 等在小麦品种快速分类模型的构建中，使用连续投影算法(successive projections algorithm，SPA)、主成分分析算法(PCA)和随机蛙跳(random frog，RF)三种特征提取方法，从数百个光谱波段中筛选可用于建立分类模型的光谱变量，使用线性判别分析(linear discriminant analysis，LDA)支持向量机(support vector machine，SVM)、极限学习机(extreme learning machine，ELM)三种机器学习算法分别以全波段和经过特征筛选的波段作为输入变量进行小麦品种分类模型构建，以全波段作为输入变量的 ELM 算法分类精度最优[10]。上述研究大多使用个体学习器的机器学习方法进行建模，相比于个体学习器，集成学习往往具备更好的稳定性和更高的精度。Chan 等使用 RF 和 Adaboost 对航空高光谱图像进行了生态区分类研究，并评估了分类精度，表明 RF 和 Adaboost 的精度均优于神经网络分类器[11]。Gislason 等对 RF，Boosting 和 Bagging 三种集成方法进行了土地覆盖分类的精度对比研究，表明 RF 在与其他集成学习方法精度相当的同时，拥有更快的训练速度并且不会过拟合[12]。Pham 等在边坡稳定性问题上应用集成学习建立了分类模型，并与八种传统机器学习模型进行了对比，模型的平均 F1 分数、准确度与 ROC 曲线下面积(AUC)分别提高 2.17%，1.66%和 6.27%，认为集成学习中的极端梯度增强集成分类器(XGB-CM)更适用于滑坡风险评估问题[13]。

基于遥感的植被理化参数的监测有助于适时、精准地获取作物氮素营养信息，是作物肥水诊断及管理决策的基础。而水稻品质形成与其生长过程中氮素营养代谢的合成与转运直接相关，结合作物光谱响应机理与碳氮代谢转运机制进行作物蛋白质含量预测具有可行性[14]。本文在前人研究的基础上，选择水稻关键生育期冠层全波段光谱数据，分别基于四种个体学习器算法(PLSR，KNN，BRR，SVR)及三种集成学习算法(RF，Adaboost，Bagging)，开展：(1)水稻氮素营养遥感监测；(2)水稻籽粒蛋白含量遥感监测，将集成学习应用于水稻生长参数及品质的遥感监测，通过研究，探索其适用性，分析不同算法在水稻氮素营养监测中的优劣，同时筛选水稻籽粒蛋白含量遥感建模的最优算法及最优变量因子，为水稻品质的监测应用提供依据。

1　实验部分

1.1　试验设计与数据采集

2019 年—2020 年在广东省广州市白云区钟落潭试验基地(23°23′24″N—23°23′59″N，113°25′48″E—113°26′24″E)开展水稻变量施肥的小区实验。试验基地内，2019 年的试验品种为美香占 2 号(V1)，插秧时间 2019 年 8 月 8 日，插秧密度为 20 cm×20 cm，试验共设计 15 个小区采样测试；每个小区插秧规格为 16 穴×16 穴。2020 年的试验品种为美香占 2 号(V1)和五丰优 615(V2)，插秧时间为 2020 年 8 月 8 日，共 30 个小区，插秧密度为 20 cm×20 cm，根据插秧规格，每个小区 16 穴×20 穴。

2019 年及 2020 年试验共设计 5 个氮素水平(N0，N1，N2，N3，N4)，分别为 0，60，120，180，240 kg N·ha^{-1}，设 3 次重复；其中基肥、分蘖肥、穗肥的施用比例为 5∶2∶3；磷、钾肥用量分别为 54 和 144 kg·ha^{-1}。分化期(2019.9.13，2020.9.10)和抽穗期(2019.10.11，2020.10.9)进行田间植株取样，获取水稻叶片及植株氮素含量数据，水稻品质数据分区于 2019 年及 2020 年水稻成熟期获取，其中 2019 年每时期获取 15 个样本，2020 年每时期获取 30 个样本，数据获取情况见表 1。

表 1　试验数据获取情况

Table 1　Test data acquisition

参数	2019 年		2020 年	
	分化期	抽穗期	分化期	抽穗期
冠层光谱	√	√	√	√
长势参数(LNC，LNA，PNC，PNA)	√	√	√	√
水稻品质数据	√	√	√	√

1.2　光谱测定与预处理

水稻冠层高光谱数据采集使用的是美国 ASD Filed Spec Pro 2500 背挂式野外光谱仪，仪器采集的光谱范围为 350～2 500 nm。根据前人的研究，作物光谱的可见光与近红外范围已能够反映作物的生长状况，因此本次实验采用 454～950 nm 的冠层光谱数据，重采样后间隔为 4 nm[15-16]。测量时间为北京时间 10:00—14:00，期间天气晴朗，在每一个采样点测量前后均用标准白板对冠层辐亮度数据进行校正。冠层光谱数据采集时距离冠层高度约 1 m，探头垂直向下，探头角度为 25°，每个采样点取 10 次测量平均值作为该样方的冠层辐亮度值。同一年的试验中，记录采样点的位置，保证不同生育期同一小区在相同位置采集数据。对测定的冠层辐亮度和白板辐亮度利用式(1)计算目标的光谱反射率。

$$R = \frac{L_{\text{target}}}{L_{\text{board}}} \times R_{\text{board}} \tag{1}$$

式(1)中，R 为冠层反射率，L_{target} 为冠层辐亮度(μW·cm^{-2}·nm^{-1}·sr^{-1})，L_{board} 为白板辐亮度(μW·cm^{-2}·nm^{-1}·sr^{-1})，R_{board} 为白板反射率。

1.3　植株氮素参量

光谱测量完成后，在小区内随机选择 6 穴水稻植株样

本，去根并逐丛计数茎蘖数，分化期茎叶分离，抽穗期将茎叶和穗分离，分开放入 105 ℃烘箱杀青 30 min，并在 80 ℃下干燥至恒重，分别称重并记录。然后使用凯氏定氮仪分别测量茎、叶、穗的氮含量。计算公式如式(2)

$$NC = (V \times 0.05 \times 14 \times 1\,000)/(1\,000 \times M) \quad (2)$$

式(2)中，NC 为氮含量(%)，V 为盐酸体积(mL)，M 为样品质量(g)。

根据采样点种植密度和水稻样本的干重计算单位面积叶片和植株的生物量和氮积累量，计算公式如式(3)

$$LNA = (LAGB \times LNC)/1\,000 \quad (3)$$

$$PNA = (LAGB \times LNC + SAGB \times SNC + EAGB \times ENC)/100 \quad (4)$$

式(3)和式(4)中，LNA 为叶片氮积累量(kg·m^{-2})；PNA 为植株氮积累量(kg·m^{-2})；LAGB，SAGB 和 EAGB 分别为测试样本中叶片、茎、穗的生物量(g·m^{-2})；LNC，SNC 和 ENC 分别为叶片、茎、穗的氮浓度(%)。在分化期时，由于水稻的穗还未发育，只计算叶片和茎的相关参量即可。

1.4 水稻籽粒蛋白含量测定

于成熟期逐小区实收 125 丛稻株(5 m^2)，水稻植株脱粒，籽粒晒干 3 个月后，脱壳碾磨成精米，然后磨细成粉，采用半微量凯氏定氮法测定籽粒氮素含量，籽粒蛋白质含量(%)=籽粒氮素含量×5.95。

1.5 回归技术

机器学习中根据算法的构建形式，可以将其分为个体学习器与集成学习器，其中个体学习器各自遵循独立的学习策略对目标进行预测，而集成学习则是将多个已有的个体学习器通过某种策略结合起来，建立一个新的学习器，最终以多个基学习器的预测结果的均值或加权均值作为最终预测结果。本文选择了四种基于不同理论的个体机器学习算法，三种基于不同构建思想的集成学习算法，研究两类算法在水稻氮素营养与籽粒蛋白品质监测上的优缺点。

个体学习器算法包括 PLSR，KNN，BRR 和 SVR，其中 PLSR 通过最小化误差的平方和，寻找一组新的潜在变量来解释自变量 X 与因变量 Y 之间存在的统计关系，是一种常见的对数据进行降维处理、解决数据多重共线性问题、简化建模过程的方式[17]。KNN 是通过 k 个最接近的邻居计算与预测因子之间空间相似性关系进行预测，常被用来分类问题，后来逐渐应用于参数估计[18]。BRR 是基于贝叶斯方法与最小二乘法的改进而提出的，通过对线性贝叶斯回归模型加入 L2 正则化，结合相关参数的先验信息形成先验分布并给出预估数值[19]。SVR 的基本思想是通过寻找最优划分超平面，忽略小于偏差 ε 的的样本，对其他样本进行回归，偏差 ε 的引入是 SVR 区别于传统回归模型的地方，即以预测 y 值为中心，与真实 y 值之间存在一个宽度为 2ε 的区域，在此区域内，预测 y 值与真实 y 值的差别认为是 0。其回归模型为 $f(x)=w^{\mathrm{T}}x+b$，w 和 b 为模型待确定参数[20]。

集成学习算法包括 Bagging，RF 和 Adaboost，其中 Bagging 的个体学习器的训练集通过自助采样得到，每个个体学习器采用的训练集不同，但个体学习器权重相同，在每个个体训练完之后进行平均，从而得到更高的准确性；RF 是 Bagging 以决策树为基学习器的拓展变体，并进一步引入属性的随机选择，抗噪性能和泛化性能有所提高，RF 和 Bagging 均属于并行化集成；Adaboost 是每个个体使用相同的训练集，但每轮训练中样本权重不同，并且后一个学习器的运行依赖前一个学习器的结果，运行过程中不断优化和提升，最终将一族弱学习器提升为损失函数极小的强学习器，属于序列化集成[21,23]。

1.6 模型验证方法

利用 PLSR，KNN，BRR，SVR，RF，Adaboost 和 Bagging 七种算法构建水稻氮素参数及籽粒蛋白含量预测模型时，采用 k-fold 交叉验证方法($k=5$)进行建模。采用决定系数(R^2)，均方根误差(root mean square error，RMSE)，平均绝对误差(mean absolute error，MAE)三个指标联合验证模型预测精度，R^2 越大，代表模型拟合度越高，RMSE 和 MAE 越小，模型稳定性越好。

2 结果与讨论

2.1 籽粒蛋白与不同生育期氮素参数的相关性分析

2019 年与 2020 年不同生育期水稻氮素参量与籽粒蛋白含量进行相关性分析，分化期 LNC，LNA，PNC，PNA 与籽粒蛋白含量的相关性系数分别为 0.452，0.794，0.499 和 0.804，抽穗期 LNC，LNA，PNC，PNA 与籽粒蛋白含量的相关性系数分别为 0.787，0.774，0.824 和 0.756，LNC 与 PNC 在分化期与籽粒蛋白含量的相关性低于抽穗期，PNC 较分化期提高 LNA 与 PNA 在分化期与抽穗期均具有较好的相关性，且除分化期的 LNC 外，其余氮素参量均达到了 0.001 水平显著。同时，四种氮素参量之间相关性为 0.7 左右，LNC 和 LNA 代表叶片尺度的氮素含量与积累情况，PNC 和 PNA 表示植株地上部分整体的氮素含量与积累情况，由于水稻籽粒蛋白的形成是一个动态的生物学过程，同一时期，不同部位与不同形式的氮素参量可能对籽粒蛋白的形成具备不同的转运与作用机理，故建模过程中分别加入 LNC，LNA，PNC 和 PNA 四个参量探究其对水稻籽粒蛋白形成的影响。

2.2 水稻氮素参数遥感模型构建与分析

2.2.1 冠层光谱与水稻氮素参数的相关性分析

图 1 是两年试验不同生育期水稻冠层光谱与氮素相关参数的相关系数图。整体来看，所有长势参数在两生育期相关系数有所不同，但均具有相似的变化趋势，近红外部分均保持在某一值持平，整体变化幅度很小且以正相关为主，可见光部分则以负相关为主。所有长势参数在 550 nm 附近出现相关性“低谷”，相关系数低于其他可见光部分。在可见光区域与近红外区域的交界处，光谱反射率受叶片内细胞间隙折射率不同的影响，反射率急剧增加，相关系数迅速由负转正，有明显的降低后再抬升的趋势。

LNC 随着生育期的推进，与冠层反射率相关性在全波段均有所提高，分化期最大相关系数为 0.450(950 nm)，抽穗期最大相关系数达到 0.585(942 nm)。LNA 在可见光部分与冠层光谱的相关性分化期高于抽穗期，而在近红外部分则

相反，两生育期最大相关系数分别为－0.602(662 nm)，0.662(950 nm)。PNC 与 LNC 具有相似的趋势，在分化期与冠层反射率相关性普遍较低，最大相关系数仅为 0.431(950 nm)，抽穗期 PNC 与冠层反射率的相关性在可见光部分与近红外部分均有较大提升，特别是近红外部分，在 750～950 nm 区间，最大相关系数达到 0.710(940 nm)。分化期 PNA 在可见光部分的 658 nm 附近，与冠层反射率存在较好的相关性，最大相关系数为－0.641(666 nm)，随着生育期的推进，在抽穗期的近红外部分，PNA 与冠层反射率的相关性优于分化期，最大相关系数为 0.663(922 nm)。

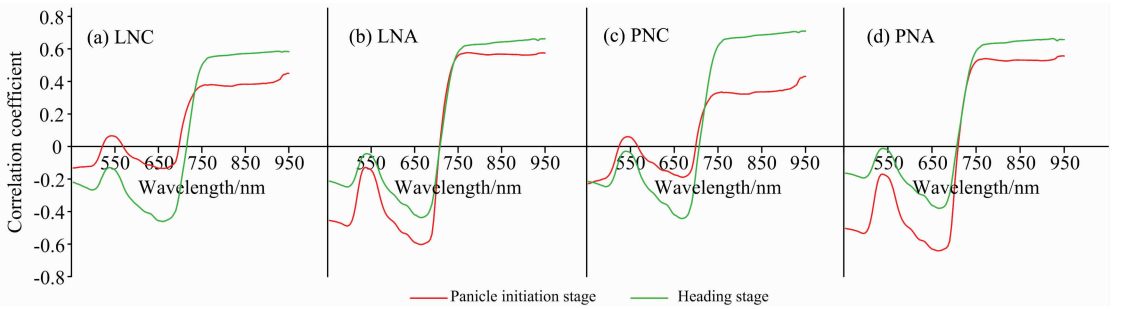

图 1　2019 年与 2020 年水稻不同生育期冠层光谱与氮素参数相关性(n=45)

Fig. 1　Correlation between canopy spectrum and nitrogen parameters at different growth stages of rice in 2019 and 2020

2.2.2　基于光谱信息的水稻氮素含量监测

基于 2019 年与 2020 年数据，以全波段光谱作为输入参数，分别采用 PLSR，KNN，BRR，SVR，RF，Adaboost 和 Bagging 七种不同算法构建分化期、抽穗期及全生育期水稻氮素参数 LNC，LNA，PNC 和 PNA 监测模型。结果如表 2、表 3、表 4 所示，七种算法均能利用全波段光谱信息实现对各个氮素参数不同程度的表达。其中，分化期、抽穗期及全生育期 LNC 的最优建模精度 R^2 分别为 0.927，0.954 和 0.922，RMSE 为 0.110，0.241 和 0.135，MAE 为 0.187，0.185 和 0.107，LNA 的最优建模精度 R^2 分别为 0.944，0.948 和 0.943，RMSE 为 0.272，0.591 和 0.440，MAE 为 0.230，0.436 和 0.325，PNC 的最优建模精度 R^2 分别为 0.930，0.951 和 0.925，RMSE 为 0.084，0.073 和 0.115，MAE 为 0.079，0.057 和 0.091，PNA 的最优建模精度 R^2 为 0.938，0.920 和 0.952，RMSE 为 0.399，1.346 和 0.978，MAE 为 0.353，1.042 和 0.716，LNC 在分化期、抽穗期的最优监测模型为 Adaboost，全生育期最优监测模型为 RF，在进行 LNA，PNC 和 PNA 监测时，均为 RF 模型表现最优，表明 RF 方法在氮素监测时具有良好的适应性与精度。这些参数和籽粒蛋白含量的相关性显著，故利用作物生长前期的光谱数据进行籽粒蛋白含量监测是可行的。

表 2　基于水稻分化期冠层光谱数据的氮素参数模型精度(n=45)

Table 2　Model accuracy of nitrogen parameters based on canopy spectral data of rice at Panicle Initiation stage

方法	LNC/%			LNA/(g·m^{-2})			PNC/%			PNA/(g·m^{-2})		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
PLSR	0.283	0.284	0.198	0.581	0.648	0.477	0.319	0.197	0.131	0.608	0.913	0.595
KNN	0.527	0.232	0.155	0.609	0.627	0.411	0.490	0.174	0.097	0.613	0.911	0.562
BRR	0.418	0.258	0.097	0.678	0.569	0.217	0.571	0.158	0.071	0.767	0.706	0.323
SVR	0.606	0.213	0.096	0.697	0.556	0.259	0.712	0.130	0.070	0.772	0.708	0.352
RF	0.912	0.120	0.110	0.944	0.272	0.230	0.930	0.084	0.079	0.938	0.399	0.353
Adaboost	0.927	0.110	0.187	0.932	0.304	0.504	0.920	0.082	0.144	0.935	0.402	0.734
Bagging	0.807	0.154	0.219	0.915	0.311	0.532	0.861	0.103	0.153	0.914	0.465	0.723

表 3　基于水稻抽穗期冠层光谱数据的氮素参数模型精度(n=45)

Table 3　Model accuracy of nitrogen parameters based on canopy spectral data of rice heading stage

方法	LNC/%			LNA/(g·m^{-2})			PNC/%			PNA/(g·m^{-2})		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
PLSR	0.699	0.202	0.156	0.645	1.081	0.834	0.705	0.110	0.080	0.619	2.256	1.727
KNN	0.630	0.199	0.141	0.677	1.084	0.740	0.679	0.098	0.083	0.612	2.354	1.696
BRR	0.738	0.112	0.084	0.658	0.504	0.385	0.779	0.061	0.049	0.652	1.221	0.943
SVR	0.749	0.104	0.089	0.682	0.506	0.410	0.827	0.063	0.055	0.626	1.216	1.071
RF	0.944	0.137	0.103	0.948	0.591	0.436	0.951	0.073	0.057	0.920	1.346	1.042
Adaboost	0.954	0.241	0.185	0.937	1.057	0.826	0.938	0.133	0.109	0.909	2.387	1.918
Bagging	0.894	0.216	0.159	0.905	1.100	0.833	0.911	0.127	0.093	0.888	2.358	1.792

表 4 基于水稻全生育期冠层光谱数据的氮素参数模型精度(n=90)

Table 4 Model accuracy of nitrogen parameters based on canopy spectral data of rice whole growth period

方法	LNC/%			LNA/(g·m^{-2})			PNC/%			PNA/(g·m^{-2})		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
PLSR	0.379	0.309	0.236	0.580	1.028	0.760	0.438	0.267	0.212	0.592	2.553	1.953
KNN	0.375	0.313	0.245	0.639	0.962	0.704	0.332	0.293	0.231	0.584	2.585	1.921
BRR	0.596	0.250	0.198	0.731	0.824	0.622	0.833	0.146	0.116	0.837	1.616	1.202
SVR	0.601	0.250	0.190	0.711	0.868	0.596	0.736	0.186	0.149	0.781	1.900	1.304
RF	0.922	0.135	0.107	0.943	0.440	0.325	0.925	0.115	0.091	0.952	0.978	0.716
Adaboost	0.778	0.199	0.181	0.876	0.596	0.509	0.848	0.154	0.133	0.919	1.426	1.213
Bagging	0.896	0.139	0.105	0.905	0.531	0.376	0.896	0.123	0.094	0.935	1.073	0.789

2.3 水稻籽粒蛋白含量监测模型构建与分析

利用 PLSR，KNN，BRR，SVR，RF，Adaboost 和 Bagging 七种不同算法，以水稻不同生育期冠层光谱数据，以及冠层光谱数据结合氮素参量为输入参数，构建水稻蛋白品质监测模型，并分析对比模型精度。

2.3.1 基于水稻冠层光谱信息的籽粒蛋白含量预测

图 2 是基于不同算法，利用水稻分化期与抽穗期冠层全波段光谱信息所建模型的精度对比结果。由于采用分化期与抽穗期全波段光谱信息进行建模，多重共线性可能是一个问题，而不同算法处理多重共线性的能力不同，各个算法在水稻籽粒蛋白含量的监测上表现差距较大。从图 2 中可以看出，分化期 KNN、PLSR 的预测结果 R^2 仅分别为 0.538 和 0.580，而 RF、Adaboost、Bagging 监测结果 R^2 则分别达到了 0.935，0.916 和 0.874，同时也具有更低的 RMSE 和 MAE，各个算法的监测能力依据 R^2 排名为 RF＞Adaboost＞Bagging＞SVR＞BRR＞PLSR＞KNN；利用抽穗期数据进行预测时，各个算法的监测精度均有提高，监测能力依据 R^2 及 RMSE 排名为 RF＞Adaboost＞Bagging＞SVR＞BRR＞KNN＞PLSR。三种集成算法(RF、Adaboost、Bagging)在处理多重共线性问题上表现出更为良好的性能。

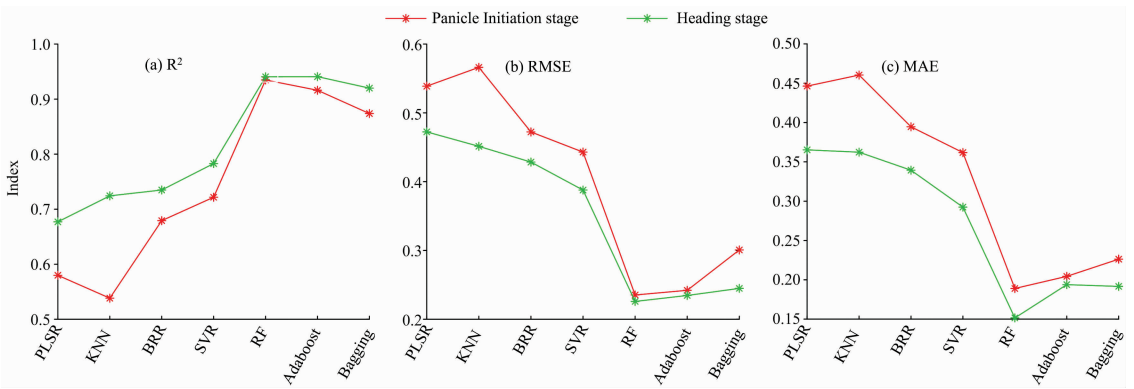

图 2 基于水稻不同生育期冠层光谱数据籽粒蛋白含量模型的 R^2，RMSE 和 MAE

Fig. 2 R^2, RMSE and MAE based on the canopy spectral data seed protein content model of rice at different fertility stages

2.3.2 基于水稻冠层光谱信息氮素参量的籽粒蛋白含量监测

利用 PLSR，KNN，BRR，SVR，RF，Adaboost 和 Bagging 七种不同算法，分别以水稻不同生育期冠层光谱数据与四个不同实测氮素参量为输入参数，构建水稻蛋白品质监测模型，分析对比了模型验证精度，图 3 显示了七种算法在两个不同生育期，采用不同参数组合所构建模型的 R^2，RMSE 和 MAE 变化的统计图，以光谱数据结合不同氮素参数作为输入参数的模型，与仅采用光谱信息所建立的模型相比，大部分算法的监测精度和稳定性均得到了提升，即在不同运行规则下的大部分算法认为氮素参数是监测籽粒蛋白含量的有效参数，其含量的高低受到植株氮素的影响。

通过综合对比，在分化期，光谱信息结合 PNA 作为输入参数时，各个算法的精度提升最明显，较以光谱信息作为输入参数的模型，各个算法 R^2 分别提高 0.131，0.182，0.013，0.041，0.020，0.044 和 0.063；在抽穗期，以光谱信息结合 PNC 作为输入参数时，各个算法的精度提升最明显，较以光谱信息作为输入参数的模型，各个算法 R^2 分别提高 0.073，0.054，0.028，0.043，0.013，0.022 和 0.019。在这两组输入参数下，两时期均为 Adaboost 表现最优，RF 和 Bagging 方法稍低于 Adaboost，但也表现极好，PLSR，KNN，BRR 和 SVR 在氮素参数影响下，模型精度提升更为明显，但仍未能超过 RF，Adaboost 和 Bagging。

图 4 给出了基于不同机器学习算法以分化期水稻冠层全波段光谱信息及水稻植株氮素累积量 PNA 作为输入参数的水稻蛋白品质模型预测值与实测值的散点图。

图 5 给出了抽穗期以全波段光谱信息结合水稻 PNC 作为输入参数的水稻蛋白品质模型预测值与实测值的散点图。

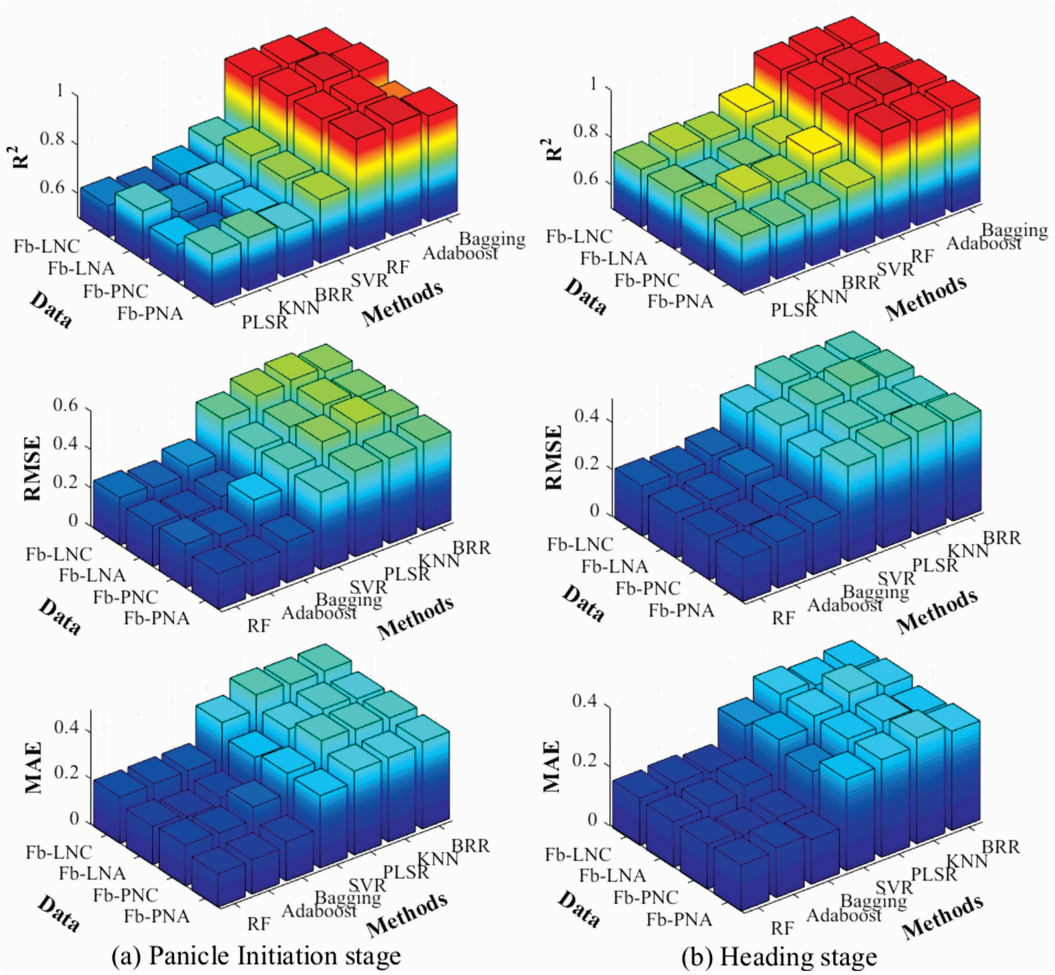

图 3　分化期和抽穗期不同参数组合下 7 种算法的 R^2、RMSE 和 MAE

Fig. 3　R^2, RMSE and MAE of seven algorithms under different parameter combinations at Panicle Initiation and Heading stage

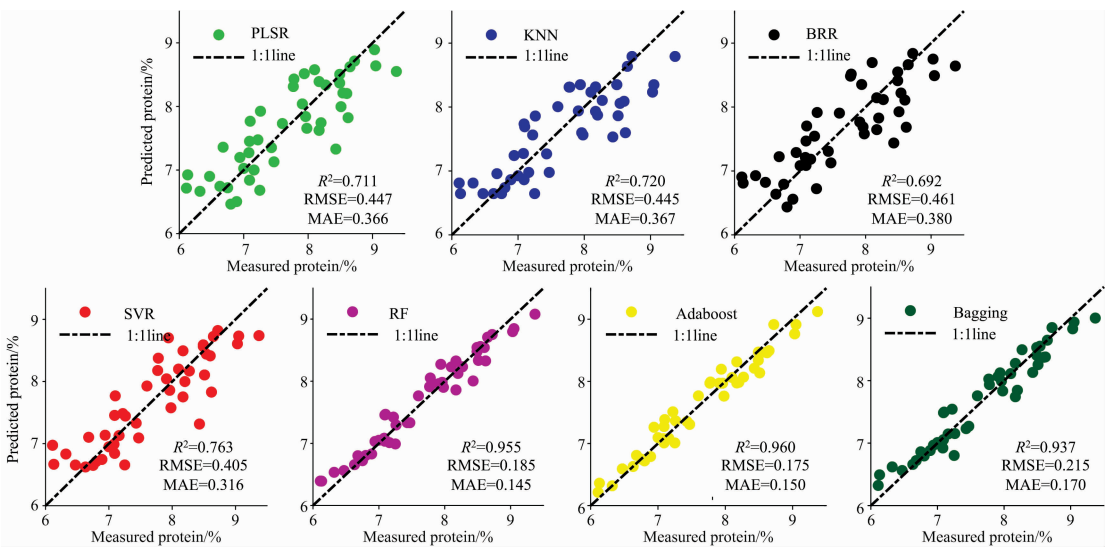

图 4　分化期以全波段光谱信息和 PNA 为输入的七种算法的 R^2，RMSE 和 MAE

Fig. 4　R^2, RMSE and MAE of seven algorithms with full band spectral information and PNA as input in Panicle Initiation stage

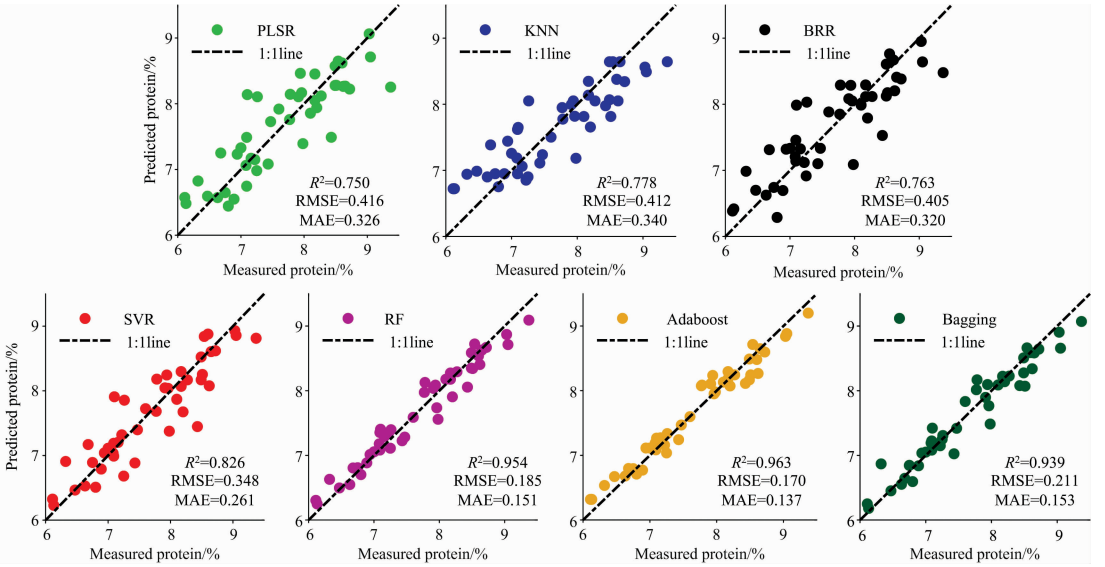

图 5 抽穗期以全波段光谱信息和 PNC 为输入的七种算法的 R^2，RMSE 和 MAE

Fig. 5 R^2, RMSE and MAE of seven algorithms with full band spectral information and PNC as input at Heading stage

从图 4 和图 5 中可以看出，三种集成算法更有利于在相同输入参数下获取更高精度，PLSR，KNN，BRR 和 SVR 算法预测实测值散点图围绕 1∶1 线仍然存在一定离散点，相比于单学习器的机器学习回归算法，三种基于集成学习器的算法 RF，Adaboost 和 Bagging 的准确度及稳定性均明显提高，且不同算法对输入参数具备不同的“适应性”，例如，RF 和 Bagging 算法获得最优精度时是抽穗期以光谱信息和 PNA 作为输入参数的，R^2 分别为 0.958 和 0.943，Adaboost 算法在水稻分化期、抽穗期的其中结合作物冠层光谱数据及氮素信息(PNC 及 PNA)蛋白品质模型 R^2 均达到 0.96 以上。

3 结 论

水稻品质形成与其生长过程中氮素营养代谢的合成与转运直接相关，本研究获取水稻不同生育期冠层光谱及水稻氮素营养参数，采用多种机器学习算法，进行作物氮素营养监测及水稻籽粒蛋白质含量监测研究。研究表明：

(1)基于水稻冠层光谱信息的水稻氮素参量遥感监测表明：冠层光谱信息能够较好的表达不同生育期水稻的氮素营养状况，其中基于不同算法的监测结果差异明显，传统回归方法和部分机器学习方法并不能够较好利用光谱信息对水稻氮素营养状况的监测，而 RF 和 Adaboost 对两时期的氮素参数监测结果 R^2 能够达到 0.90 以上，表明冠层光谱中与植株氮素营养相关的信息被较好的利用。由于水稻关键生育期的氮素含量与其籽粒蛋白含量相关性显著，因此利用分化期与抽穗期冠层光谱信息监测水稻籽粒蛋白品质是可行的。

(2)基于冠层光谱数据的水稻籽粒蛋白含量监测结果表明：在利用作物冠层全波段信息可能存在强共线性的情况下，七种算法的准确度与稳定性不同，三种集成算法在水稻蛋白品质监测上具有明显优势，其中 RF 算法两个不同生育期蛋白品质监测模型的 R^2 分别为 0.935，0.941，Adaboost 算法两个不同生育期蛋白品质监测模型的 R^2 分别为 0.916，0.941；Bagging 算法两个不同生育期蛋白品质监测模型的 R^2 分别 0.874，0.920，RF，Adaboost 和 Bagging 三种集成算法几乎不受多重共线性的影响，对比单学习器的机器学习算法，利用多个基学习器进行训练的算法能够解读更多籽粒蛋白含量与各种参数间尚未明确的关系，为最终的监测目标挖掘更多的决策信息。

(3)将氮素参量及水稻冠层光谱信息作为模型输入因子，进行水稻籽粒蛋白含量监测预测时，模型精度得到进一步改善，表明氮素参量对籽粒蛋白含量存在一定程度的影响，其中在分化期以光谱信息和 PNA 作为输入参数时，模型精度提升更明显(R^2 从 0.935 提高到 0.960)，在抽穗期以光谱信息和 PNC 作为输入参数时，模型精度提升更明显(R^2 从 0.941 提高到 0.963)。利用抽穗期数据进行监测时，获取的监测精度比分化期更高，可能是抽穗期水稻株体发育趋于完善，籽粒蛋白质的产生与植株氮素的关系更加明确，故利用抽穗期数据进行水稻籽粒蛋白含量能够获取更高的精度。

实验结果表明，全波段光谱携带的信息能够较好的对水稻氮素营养参数进行监测，也表明利用分化期与抽穗期冠层光谱监测当季水稻籽粒蛋白含量是可行的，七种方法中，仅输入全波段光谱的时候，RF 最优，最大 R^2 为 0.941，加入氮素参数后，Adaboost 表现最优，最大 R^2 为 0.963，Bagging 也取得了较好的监测结果。集成学习方法能够解决数据存在的多重共线性问题，k-fold 交叉验证方法也在一定程度上避免了模型的过拟合，利用所有信息进行回归建模，较大程度的保留了与水稻氮素营养和籽粒蛋白含量相关的信息，且简化了数据处理流程，在实际农业监测中更有利于推广与应用。论文在将不同参数组合作为自变量进行输入时只对各类参数简单拼接，未考虑不同参数的权重分布，不同参数与

籽粒蛋白含量是否存在最优映射的关系有待下一步研究。

References

[1] ZHANG Han, ZHAO Xiao-min, GUO Xi, et al(张　晗, 赵小敏, 郭　熙, 等). Jiangsu Agricultural Science(江苏农业科学), 2018, 46(12): 1.

[2] JIANG Huan-yu, YING Yi-bin, XIE Li-juan(蒋焕煜, 应义斌, 谢丽娟). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2008, 28(6): 1300.

[3] TANG Yan-lin, WANG Ji-hua, HUANG Jing-feng, et al(唐延林, 王纪华, 黄敬峰, 等). Acta Agronomica Sinica(作物学报), 2004, (8): 780.

[4] Liu X D, Sun Q H. International Journal of Pest Management, 2016, 62(3): 205.

[5] ZHANG Hao, HU Hao, CHEN Yi, et al(张　浩, 胡　昊, 陈　义, 等). Journal of Nuclear Agricultural Sciences(核农学报), 2012, 26(1): 135.

[6] SUN Xue-mei, ZHOU Qi-fa, HE Qiu-xia(孙雪梅, 周启发, 何秋霞). Acta Agronomica Sinica(作物学报), 2005, (7): 844.

[7] LIU Yun, TANG Yan-lin, HUANG Jing-feng, et al(刘　芸, 唐延林, 黄敬峰, 等). Scientia Agricultura Sinica(中国农业科学), 2008, (9): 2617.

[8] Bennett K P, Parrado-Hernández E. The Journal of Machine Learning Research, 2006, 7: 1265.

[9] Liakos K G, Busato P, Moshou D, et al. Sensors, 2018, 18(8): 2674.

[10] Bao Y, Mi C, Wu N, et al. Applied Sciences, 2019, 9(19): 4119.

[11] Chan J C W, Paelinckx D. Remote Sensing of Environment, 2008, 112(6): 2999.

[12] Gislason P O, Benediktsson J A, Sveinsson J R. Pattern Recognition Letters, 2006, 27(4): 294.

[13] Pham K, Kim D, Park S, et al. Catena, 2021, 196: 104886.

[14] ZHANG Fu-suo, CUI Zhen-ling, WANG Ji-qing, et al(张福锁, 崔振岭, 王激清, 等). Chinese Bulletin Botany(植物学通报), 2007, (6): 687.

[15] Fitzgerald M A, McCouch S R, Hall R D. Trends in Plant Science, 2009, 14(3): 133.

[16] ZHAO Xiao-min, SUN Xiao-xiang, WANG Fang-dong, et al(赵小敏, 孙小香, 王芳东, 等). Acta Agriculturae Universitatis Jiangxiensis(江西农业大学学报), 2019, 41(1): 1.

[17] Davari M, Karimi S A, Bahrami H A, et al. Catena, 2021, 197: 104987.

[18] Souza D V, Nievola J C, Santos J X, et al. Journal of Sustainable Forestry, 2019, 38(8): 755.

[19] Assaf A G, Tsionas M, Tasiopoulos A. Tourism Management, 2019, 71: 1.

[20] Chen Y W, Lin C J. Feature Extraction, 2006, 207: 315.

[21] Belgiu M, Drăguţ L. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 114: 24.

[22] Wu S, Nagahashi H. IEEE Signal Processing Letters, 2014, 21(6): 687.

[23] Wu X, Kumar V, Quinlan J R, et al. Knowledge and Information Systems, 2008, 14(1): 1.

Monitoring Nitrogen Nutrition and Grain Protein Content of Rice Based on Ensemble Learning

ZHANG Jie[1, 2], XU Bo[1], FENG Hai-kuan[1], JING Xia[2], WANG Jiao-jiao[1], MING Shi-kang[1], FU You-qiang[3], SONG Xiao-yu[1*]

1. Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100094, China
2. School of Surveying and Mapping Science and Technology, Xi'an University of Science and Technology, Xi'an 710054, China
3. Rice Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China

Abstract The use of hyperspectral remote sensing technology to monitor the protein content related to grain quality before rice matures is important. It can promptly adjust cultivation management methods and guide reasonable fertilization and help to grasp rice grain quality information in advance and clarify market positioning. This study took typical high-quality Indica rice in Guangdong Province as the research goal. Two-year nitrogen gradient experiments were carried on in 2019 and 2020. The canopy level hyperspectral data and rice nitrogen parameters, including leaf nitrogen content (LNC), leaf nitrogen accumulation (LNA), plant nitrogen content (PNC), and plant nitrogen accumulation (PNA), were collected at the rice panicle initiation stage and heading stage. Then, four individual machine learning algorithms, Partial Least Square Regression (PLSR), K-Nearest Neighbor (KNN), Bayesian Ridge Regression (BRR), Support Vector Regression (SVR), and three ensemble learning algorithms, Random forest (RF), Adaboost, Bagging were used for monitoring and modeling the nitrogen status of rice at different growth stages. After that, the rice grain protein content estimation models based on rice canopy spectral information, and spectral information combined with rice nitrogen parameterswere constructed by different machine learning algorithms. The rice nitrogen and grain protein content estimation models' accuracy were evaluated and compared. The study results showed that for rice nitrogen nutrition monitoring, using the rice canopy spectral information from 454～950 nm, the R^2 of LNC, LNA, PNC and PNA estimation models based on RF and Adaboost algorithms achieved above 0.90 at the rice, heading stage, with low RMSE and MAE. Panicle initiation stage When using full-band spectral information to estimaterice grain protein content, RF had the highest accuracy and stability, with R^2 of 0.935 and 0.941, RMSE of 0.235 and 0.226, and MAE of 0.189 and 0.152 at rice panicle initiation and heading stage, respectively. Adaboost model has the highest accuracy and stability for seed protein monitoring with full-band spectral information combined with growth parameters at both fertility stages, at the panicle initiation stage, the Adaboost algorithm with full-band spectral and PNA data can reach the bestfor rice grain proteinestimation, the R^2, RMSE and MAE was 0.960, 0.175, and 0.150. While at heading stage, the R^2, RMSE and MAE was 0.963, 0.170, 0.137, when using Adaboost algorithm with the full-band spectral and PNC data as input parameters. The results showed that the ensemble algorithms RF, Adaboost and Bagging have good ability to deal with multiple covariance compared with several individual learner algorithms PLSR, KNN, BRR and SVR. And they are suitable for the analysis and processing of hyperspectral data, which have obvious advantages in crop nitrogen nutrition monitoring and rice quality early monitoring through remote sensing.

Keywords Hyperspectral remote sensing; Rice grain protein; Machine Learning; Ensemble algorithms; Adaboost; Random forest

(Received Oct. 27, 2021; accepted Feb. 25, 2022)

* Corresponding author