



# 基于集成学习和近红外光谱的玉米种子含水率预测方法研究

杨琳<sup>1</sup>, 张林<sup>1</sup>, 叶泽辉<sup>2</sup>

(1. 商洛学院 电子信息与电气工程学院, 陕西商洛 726000; 2. 陕西省商丹高新学校, 陕西商洛 726000)

**摘要** 针对传统化学方法测定玉米种子含水率存在工序复杂、周期长、成本高等问题, 提出一种基于集成学习算法和近红外光谱技术的快速、无损预测玉米种子含水率的方法。以‘陕科9号’等8个品种的320份玉米种子作为研究对象, 用近红外光谱仪(Antaris II型, 美国Nicolet公司)采集玉米种子的近红外漫反射光谱。统一采用偏最小二乘回归(Partial Least Squares Regression, PLS)方法对比分析SG平滑滤波(Savitzky-Golay, SG)结合4种光谱预处理方法对玉米种子近红外光谱的预处理效果, 发现Savitzky-Golay方法结合多元散射校正法去噪效果最优。采用竞争性自适应重加权算法(CARS)进行特征波长的提取, 前7个光谱特征的累计贡献率超92%以上。以GBDT(Gradient Boosting Decision Tree, 梯度提升决策树)、RF(Random Forest, 随机森林)、XGB(或XGBoost, Extreme Gradient Boosting 极端梯度上升)作为基础模型, 采用Stacking作为融合策略, 建立Stacking集成学习模型。预处理后的数据, 提取前7个主成分作为特征向量, 用直接干燥法得到这些种子的含水率作为标签, 分别训练4种玉米种子含水率预测模型, 对比分析该4种模型的性能指标, Stacking集成模型经过2163次训练后预测相关系数 $R^2=0.9391$ , 相对分析误差PRD=2.91。结果表明, Stacking集成模型融合了GBDT、RF、XGB3个基础模型的优势, 精度高, 收敛特性好, 泛化能力强, 为玉米种子含水率快速、无损的测定提供了新的思路。

**关键词** 玉米种子; 近红外光谱; 集成学习; Stacking; 含水率

**中图分类号** O657.3

**文献标志码** A

**文章编号** 1004-1389(2022)08-1025-09

玉米是全世界范围内主要的粮食作物之一, 富含氨基酸、维生素A、维生素E、卵磷脂、亚油酸等营养成分, 不但是水产养殖、养殖、畜牧等行业至关重要的饲料原料来源, 还是医疗卫生、轻工化工、战略物资储备, 以及食品餐饮等领域的必不可少的原料之一<sup>[1-2]</sup>。20世纪90年代随着美国率先推出玉米“单粒种子”播种技术, 迅速在世界种粮大国得到推广, 中国玉米种植模式也随之发生改变<sup>[3]</sup>。种子质量是决定玉米产量的关键因素, 但种子质量容易受到环境条件、机械脱粒、干燥等操作的影响<sup>[4]</sup>, 这些因素可能导致严重的种子损伤和种子外观的变化。机械揉捏压力造成的损害、病原体的侵袭、种子破损、种皮破裂、水分损害和发绿的种子可能会对种子活性产生一定的影响, 对玉米种子质量的监控显得非常重要<sup>[5-7]</sup>。

玉米种子的质量取决于水分、油、蛋白质和淀

粉几种主要化合物, 这些化合物赋予玉米种子营养价值, 尤其是水。水分含量太高, 会引发种子霉变, 水分含量太低, 种子活性会大幅度降低。因此, 在玉米种子处理工序和储存过程中, 应保持特定的水分含量限值, 以确保其质量的稳定性<sup>[8]</sup>。目前, 物质成分测定的常规物理和化学方法主要包括烘箱干燥法、分光光度法和高效液相色谱法(HPLC)<sup>[9-12]</sup>。虽然基于大型设备的技术提供了各种可靠的方案, 具有良好的准确性和灵敏度, 但它们通常存在预处理程序复杂、操作耗时、成本高以及需要专业操作人员等缺点<sup>[13]</sup>。

由于拥有快速、无损的优势, 近红外光谱技术近年来得到了广泛的关注。例如刘秀英等<sup>[14]</sup>将BP神经网络对光谱的分析, 应用于牡丹种子含水率的估算, 姜妍等<sup>[15]</sup>用近红外光谱检测了青饲大豆主要品质参数, 焦俊等<sup>[16]</sup>用近红外光谱完成了

收稿日期: 2022-05-19 修回日期: 2022-06-23

基金项目: 商洛学院科学研究项目(18SKY-FWDF001)。

第一作者: 杨琳, 女, 硕士, 讲师, 主要从事机器学习在农业的应用研究。E-mail: 1742198437@qq.com

通信作者: 张林, 男, 硕士, 讲师, 主要从事近红外光谱分析研究。E-mail: 759013520@qq.com

核桃的品种鉴别,余克强等<sup>[17]</sup>用近红外光谱检测了猕猴桃不同贮藏期果肉色泽。康文翠等<sup>[18]</sup>利用光谱对霉变小麦菌落数定量分析。近红外光谱技术的原理是近红外光照射到样品时,一部分光会被物质吸收,同时会发生漫反射,不同的成分对应物质内部的 C-H、N-H、O-H 和 S-H 等不同化学键,对近红外光谱的吸收波长不同,因此可以利用近红外光谱分析技术进行定性和定量分析研究<sup>[19-22]</sup>。Bai 等<sup>[23]</sup>基于机器学习算法,根据速溶茶的近红外光谱,对其水分、咖啡因等相关参数进行预测,表现出很高的精度和准确性。虽然近红外光谱已经得到了广泛的应用,但对玉米种子的各种成分进行无损定量分析的研究很少,且准确率还有待提高。在多元数据分析中,机器学习<sup>[24]</sup>的应用非常广泛,其中的集成学习算法具有可以融合多个基础模型的优势,从而达到更高的精度。针对上述问题,本试验拟结合近红外光谱技术与机器学习方法的优势,研究含水率预测的新方法,在提高速度的同时提高准确率。

本研究采用竞争性自适应重加权算法(CARS)进行特征波长的提取,建立集成学习<sup>[25]</sup>算法模型,选择 RF<sup>[26]</sup>、GDBT<sup>[27]</sup>、XGB<sup>[28]</sup>作为基础模型,以 Stacking 作为融合规则,使用不同的模型进行玉米种子含水率预测的试验,以异质集成的方式,有效融合基算法,实现优势互补。结果表明,Stacking 集成学习算法较其他算法而言,精度高,有良好的泛化能力,显著提升玉米种子含水率预测的效果。本研究为近红外光谱结合集成学习算法分析玉米种子质量提供了新的思路,提高了准确率和应用性,对高效率播种,与提高现代化农业生产效率有非常重要的意义。

## 1 材料与方法

### 1.1 试验材料

本试验中玉米样本为未包裹种衣剂等药物的玉米种子,由种子公司提供,分别为‘陕科 9 号’‘颍玉 8 号’‘东单 159’‘潞玉 6 号’‘正大 999’‘正大 12’‘均隆 608’‘郑单 958’共计 8 个品种,每个品种选择 40 粒,共计 320 份玉米种子样本。首先对 320 个玉米种子进行编号,采用近红外光谱仪(Antaris II 型,美国 Nicolet 公司)采集每粒玉米种子的近红外漫反射光谱作为特征,然后采用直接干燥法测定每个玉米种子含水率作为标签。

为了增强所建立模型的鲁棒性,训练集和测试集的样本更具有代表性,按 3 : 1 的比例,采用 K-S 算法(Kennard-Stone),将所有数据集划分为训练集和测试集。K-S 算法是将所有的样本全部看作是训练集的候选样本,然后依次从其中挑选出测试集样本进行划分,其挑选的依据是依次选择欧氏距离最远的 2 个样本作为训练集,随后对其余的样本分别计算到训练集中每一个样本的欧氏距离,分别找到拥有最大欧氏距离的样本和最小欧氏距离的样本存放入训练集中,依次重复以上过程,直到训练集样本数量达到设置要求为止,剩余样本划为测试集。该算法最大的优势在于可以确保测试集中的样本最大限度按照空间距离均匀分布,且不受异常样本的干扰。样本间欧氏距离计算方法如公式(1)所示。

$$d_x(p, q) = \sqrt{\sum_{j=1}^m [x_p(j) - x_q(j)]^2} \quad (1)$$

式中:  $N$  表示样本点数,  $p, q \in [1, N]$ ,  $x_p$  和  $x_q$  表示 2 个不同的样本,  $m$  表示空间维度。图 1 为采用直接干燥法测得的玉米种子样本含水率

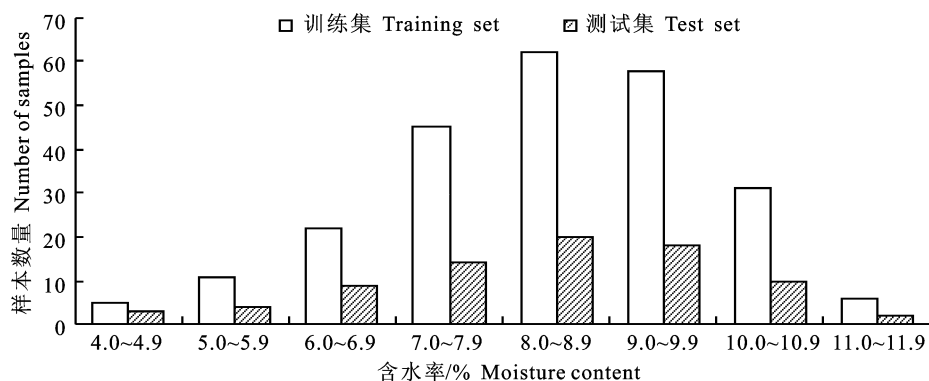


图 1 玉米种子样本含水率

Fig. 1 Distribution of moisture content maize seed samples

分布,从图中可知玉米种子样本含水率基本符合正态分布。

## 1.2 近红外光谱采集

本试验过程中采用近红外光谱仪(Antaris II 型,美国 Nicolet 公司)及其漫反射部件对玉米种子近红外光谱进行采集,在采集过程中为减少因玉米种子摆放角度引起的干扰,分别对玉米种子样本胚芽正面朝上和胚芽朝下采集 3 次光谱,对每个样本的共计 6 个光谱取平均作为该玉米种子的近红外光谱数据使用。仪器参数设定为:光谱采集波长范围 1 100~2 400 nm,扫描间隔为 2 nm,每个样品获得 700 个样本点,扫描次数为 32 次。玉米种子样本近红外漫反射光谱如图 2 所示。

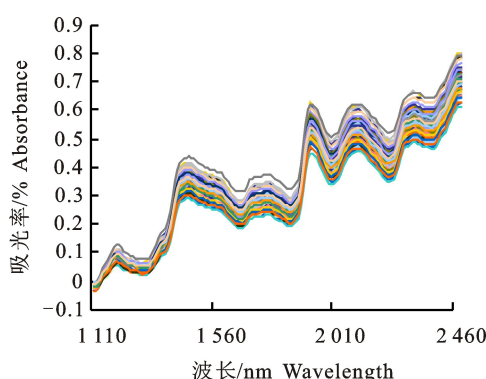


图 2 玉米种子样本近红外光谱图

Fig. 2 Near infrared spectra of maize seed samples

## 1.3 光谱预处理

由于不同的玉米种子颗粒及种子表面平整度的差异及种子形状和种皮之间均存在着较明显的差异,导致近红外光谱在采集过程中光的反射、散射影响程度不一,采集到的玉米种子原始近红外光谱中包含了大量的干扰信息,导致测量误差很难消除,影响最终的定量分析结果。因此为了减少外界环境和种子外观不同产生的影响,选择适当的预处理方法对光谱进行预处理是必要的。本次试验采用 SG 平滑滤波法,分别结合多元散射校正法(Multiplicative scatter correction, MSC)、归一化(Normalization, NOR)、均值中心化(Mean centering, MC)、标准正态变量变换(Standard normal variate transformation, SNV) 4 种预处理方法对玉米种子近红外光谱的处理。

## 1.4 提取特征波长

由于全波段的吸收光谱中包含有大量的干扰信息和冗余信息,为了减少计算量,简化模型,提高模型的运行效率,需要对预处理后的光谱信息

进行特征提取。采用竞争性自适应重加权算法(competitive adaptive reweighted sampling, CARS)来提取玉米种子近红外光谱的特征波长。图 3 为玉米种子近红外光谱主成分累计贡献率,由图中可知前 7 个主成分的累计贡献率达 92% 以上,包含了大部分光谱有效信息,因此建模时提取玉米种子光谱的前 7 个主成分作为特征信息进行建模。

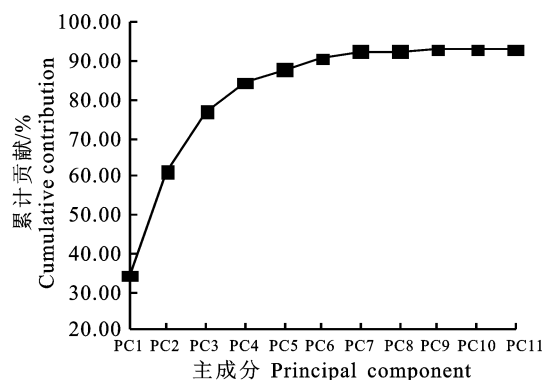


图 3 玉米种子样本近红外光谱主成分累计贡献率

Fig. 3 Cumulative contribution of principal components in near-infrared spectra of maize seed samples

## 2 算法原理

集成学习<sup>[25]</sup>(Ensemble learning)是一种融合多个基础模型的集成模型,通过某种融合策略,对几个可能含有错误的输出进行分析后,得到一个对性能提高具有积极影响的结果,提升模型学习的性能。

### 2.1 基础模型选择

本试验中,通过对 8 个玉米品种的光谱特征在 GBDT 等 10 个常用模型进行性能验证,其中 RF 算法对数据集的适应能力强,训练速度快,有一定的抗噪能力;GBDT 的基学习器之间存在串行关系,在相对较少的调参时间下,预测的准确度较高;XGB 能够自动地运用 CPU 的多线程进行并行计算,同时算法精度得以提高。因此选择以上 3 个模型作为集成模型的基础模型。

RF<sup>[26]</sup>(Random Forest, 随机森林),即用随机的方式将多个没有关联的决策树组合起来的模型。得到随机森林模型后,当新样本进入时,随机森林中的每一棵决策树分别进行判断, T 个基础模型得到的回归结果进行算术平均,作为最终模型输出。这种方式提高了预测精度,并对多元共

线性不具有敏感性,对缺失数据和非平衡数据的结果而言,相对稳健,对于几千个解释变量,有很好的预测表现。

GBDT<sup>[27]</sup> (Gradient Boosting Decision Tree, 梯度提升决策树)是一种迭代决策树算法,由多棵回归决策树组成。GBDT 的核心在于累加所有树的结果作为最终结果,该算法的泛化能力较强。GBDT 的基础学习器表示为式(2):

$$h(X; \{b_j, R_j\}^J) = \sum_{j=1}^J b_j I(x \in R_j) \quad (2)$$

回归树将输入变量和输出变量组成的空间划分成  $J$  个不相交的区域  $R_j$ ,  $j$  表示第  $j$  个区域。样本点落入区域  $R_j$  预测值为  $b_j$ , 否则为 0,  $\{b_j, R_j\}^J$  表示每个模型的参数。 $I(\cdot)$  为示性函数,当前迭代的结果表示为式(3):

$$f_b(x) = f_{b-1}(x) + \rho_b \sum_{j=1}^J b_{jb} I(x \in R_{jb}) \quad (3)$$

其中,  $f_{b-1}(x)$  表示上一次迭代结果,  $R_{jb}$  为第  $b$  次迭代基于伪响应变量  $\tilde{y}_j$  建立回归树获得的区域划分。 $b_{jb} = \text{ave}_{x \in R_{jb}} \tilde{y}_j$ , 即预测值为落入区间  $R_{jb}$  的样本观测点的均值,系数  $\rho_b$  为搜索所得。

XGB<sup>[28]</sup> (或 XGBoost, Extreme Gradient Boosting 极端梯度上升),以提升效果为目的,依次加入不同的树,来对损失函数 Loss 进行二阶的泰勒展开,同时还添加了正则项,以得到最优解,通过权衡模型的复杂程度,以及降低目标函数,达到避免过拟合的效果。

XGB 算法完整的目标函数为式(4):

$$Obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{t=1}^T \Omega(f_t) = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + const \quad (4)$$

其中,  $i$  表示数据集的第  $i$  个样本,  $n$  表示样本总量,  $t$  表示迭代次数,  $y_i$  表示第  $i$  个样本的真实值,  $\hat{y}_i^{(t)}$  表示第  $i$  个样本第  $t$  次迭代的预测值,  $\sum_{i=1}^n l(y_i, \hat{y}_i^{(t)})$  是自身损失函数,而  $\Omega(f_t)$  是正则化惩罚项,表示模型的复杂度,  $const$  是常数项。给目标函数对权重求偏导,得到一个能够使目标函数最小的权重,将这个权重代入到目标函数中,算得的结果就是目标函数求解后的最小值。

## 2.2 融合策略

集成学习一般有平均法、投票法和学习法几种融合策略,分为 Bagging、Boosting、Stacking 三大类,分别侧重于减少方差、减少偏差和提升预测结果。其中, Bagging 是并行同质集成方法,利用基础模型的独立性,通过平均能够较大地降低误差; Boosting 属于串行集成方法,利用基础模型之间的依赖,通过给错分样本一个较大的权重来提升性能; Stacking 是通过一个元分类器或者元回归器来整合多个分类模型或回归模型的集成学习技术。本研究使用 Stacking 策略融合 RF、GBDT 和 XGB 3 个基础模型算法,其中 RF 属于 Bagging, GBDT、XGB 属于 Boosting。建立一种新的 Stacking 算法,如图 4 所示,其流程分为 6 步。

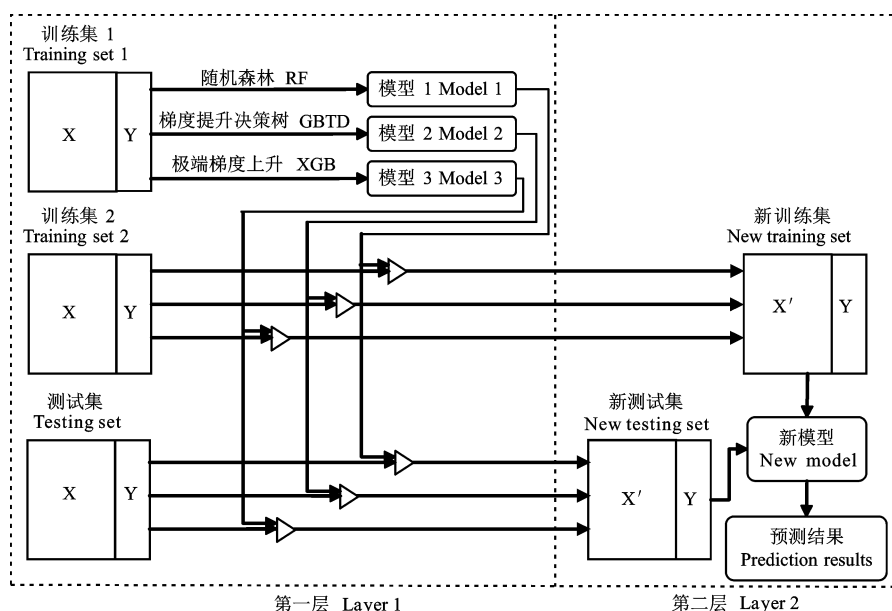


图 4 Stacking 集成学习算法框图

Fig. 4 Block diagram of Stacking ensemble learning algorithm

(1)将数据划分为 2 个训练集(Training set 1, Training set 2)和 1 个测试集(Testing set);

(2)创建第一层模型,利用 Training set 1 来训练 GDBT、RF、XGB 3 个基学习器的模型;

(3)每个训练集划分为 5 个大小相同的集合,取其中一份作为测试集,其余的为训练集。每个模型都要进行 5 折交叉验证,如图 5 所示,分别得到一个结果,合并构成新的训练集;

(4) Training set 2 分别通过 GDBT、RF、

XGB 3 个基模型,生成的预测结果作为新的特征,并拼接起来作为第二层的训练集 New training set;

(5)测试集(Testing set)也经过 3 个模型后,取输出的平均值,形成第二层的测试集 New testing set;

(6)以 New training set 作为训练集,以 New testing set 作为测试集,进行第二层模型的训练与预测,得到最终的结果。

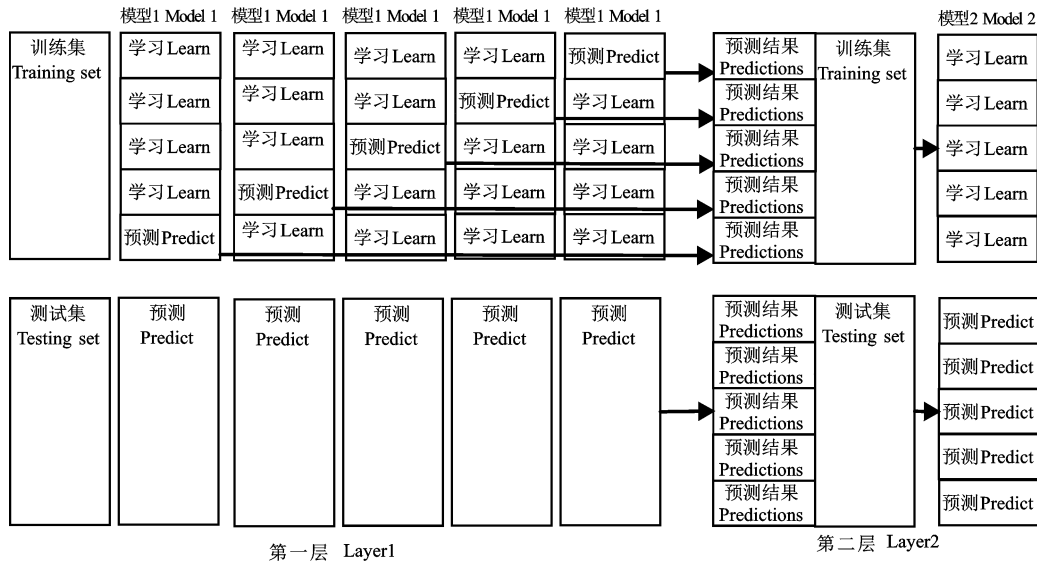


图 5 Stacking 集成学习算法数据流图

Fig. 5 Data flow diagram of Stacking ensemble learning algorithm

### 2.3 模型的评价

在本次研究过程中采用决定系数( $R^2$ )、校正均方根误差(Root mean square error of cross validation,  $RMSECV$ )、预测均方根误差(Root mean square error of prediction,  $RMSEP$ )和相对分析误差(relative percent deviation,  $RPD$ )对模型进行评价。其计算方法如式(5)、式(6)和式(7)所示:

$$RMSECV \text{ or } RMSEP = \sqrt{\frac{\sum_{j=1}^m (y_j - \hat{y}_j)^2}{m}} \quad (5)$$

$$R^2 = 1 - \frac{\sum_{j=1}^m (y_j - \hat{y}_j)^2}{\sum_{j=1}^m (y_j - \bar{y}_j)^2} \quad (6)$$

$$RPD = \frac{1}{\sqrt{1 - R^2}} \quad (7)$$

式中: $m$  表示样本数量, $y_j$  表示第  $j$  个样本的真实值, $\hat{y}_j$  表示第  $j$  个样本的预测值, $\bar{y}_j$  表示

所有  $m$  个训练样本真实值的平均值。

通常模型最终得到的  $RMSECV$  和  $RMSEP$  的值越小, $R^2$  的值越接近于 1,则表面该模型的预测精度越高。如果  $RMSECV$  和  $RMSEP$  的值相差过大,则表明训练集样本和测试集样本不具有代表性,发生了过拟合或者欠拟合的情况,需要对样本重新划分训练集和测试集。

$RPD$  的值小于 1.4,表示模型不可靠; $RPD$  的值在 1.4~2 之间表示模型有一定的可靠性; $RPD$  的值在 2~2.5 之间表示模型较可靠,能够用于模型分析; $RPD$  的值大于 2.5,表示模型可靠性较高<sup>[29]</sup>。

## 3 结果与分析

### 3.1 光谱预处理

采用 SG 平滑滤波法(Savitzky-Golay smoothing, SG),分别结合多元散射校正法(Multiplicative scatter correction, MSC)、归一化(Normal-

ization, NOR)、均值中心化(Mean centering, MC)、标准正态变量变换(Standard normal variate transformation, SNV) 4 种预处理方法对玉米种子原始近红外光谱的处理后,选用偏最小二

乘回归(Partial Least Squares Regression, PLS)方法分别建立原始光谱、SG 平滑滤波后的光谱,以及 SG 平滑滤波结合四种预处理的光谱,建立含水率预测模型,其结果如表 1 所示。

表 1 玉米种子含水率预测模型不同预处理方法建模结果

Table 1 Modeling results of different preprocessing methods for prediction model of moisture content in maize seed

| 预处理方法<br>Preprocessing methods | 训练集 Training set                                     |                                                                                 | 测试集 Testing set                                      |                                                                       |                                                  |
|--------------------------------|------------------------------------------------------|---------------------------------------------------------------------------------|------------------------------------------------------|-----------------------------------------------------------------------|--------------------------------------------------|
|                                | 决定系数<br>( $R^2$ )<br>Coefficient of<br>determination | 校正均方根误差<br>(RMSECV)<br>The root mean<br>square error<br>for cross<br>validation | 决定系数<br>( $R^2$ )<br>Coefficient of<br>determination | 预测均方根误差<br>(RMSEP)<br>The root mean<br>square error<br>for prediction | 相对分析误差<br>(RPD)<br>Relative percent<br>deviation |
| 原始光谱 Original spectrum         | 0.798 2                                              | 0.257 4                                                                         | 0.770 9                                              | 0.265 4                                                               | 1.57                                             |
| 平滑滤波 SG                        | 0.867 5                                              | 0.204 3                                                                         | 0.856 9                                              | 0.216 2                                                               | 1.94                                             |
| 平滑滤波+均值中心化 SG+MC               | 0.870 2                                              | 0.194 2                                                                         | 0.868 9                                              | 0.203 5                                                               | 2.02                                             |
| 平滑滤波+归一化 SG+NOR                | 0.891 8                                              | 0.178 4                                                                         | 0.887 5                                              | 0.183 1                                                               | 2.17                                             |
| 平滑滤波+标准正态变量变换 SG+SNV           | 0.894 8                                              | 0.162 4                                                                         | 0.896 8                                              | 0.186 3                                                               | 2.26                                             |
| 平滑滤波+多元散射校正法 SG+MSC            | 0.903 2                                              | 0.159 4                                                                         | 0.902 3                                              | 0.159 7                                                               | 2.32                                             |

由表 1 可知,无论采用哪一种光谱预处理方法,分别建立玉米种子含水率的预测模型,训练集和测试集的决定系数均大于 0.77,测试集的相对分析误差也均大于 2,说明采用近红外光谱分析法对玉米种子含水率进行检测分析是可行的。对比分析建模参数可以发现采用 Savitzky-Golay 方法结合多元散射校正法预处理后的光谱建立的模型,达到了较好的效果,决定系数  $R^2=0.903\ 2$ ,预测均方根误差  $RMSEP=0.159\ 7$ ,且相对分析误差  $RPD=2.32$ ,既能够有效去除干扰信息,又能够最大限度保留有效信息。

### 3.2 建模结果分析

为检验 Stacking 集成算法的优劣,将其与 RF、GDBT、XGB 3 个算法进行比较,分别进行

3 000 次训练后的建模结果如表 2 所示,在同样玉米种子光谱数据对模型进行训练的情况下,本研究提出的 Stacking 集成算法预测效果更优,预测相关系数  $R_p^2=0.939\ 1$ ,相对分析误差  $PRD=2.91$ ,模型预测精度和稳定性较好。

为了进一步优化模型参数和收敛特性,对模型进行重复训练,训练结果如图 6 所示,由曲线对比可知,四种模型测试集的损失函数,均随着训练次数的增加而减小,在训练 1 051 次之前损失函数下降速度较快,模型优化速度明显,本研究提出的 Stacking 集成算法收敛速度最快,明显优于 Bagging、Boosting 策略,在训练 2 163 次以后模型性能提升有限,因此在实际应用中将模型的训练次数设定为 2 163 次对比 4 个模型的预测效果。

表 2 4 种模型的预测效果对比

Table 2 Comparison of prediction effect of four models

| 模型<br>Model  | 训练集 Training set                                     |                                                                                 | 测试集 Testing set                                      |                                                                       |                                                  |
|--------------|------------------------------------------------------|---------------------------------------------------------------------------------|------------------------------------------------------|-----------------------------------------------------------------------|--------------------------------------------------|
|              | 决定系数<br>( $R^2$ )<br>Coefficient of<br>determination | 校正均方根误差<br>(RMSECV)<br>The root mean<br>square error<br>for cross<br>validation | 决定系数<br>( $R^2$ )<br>Coefficient of<br>determination | 预测均方根误差<br>(RMSEP)<br>The root mean<br>square error<br>for prediction | 相对分析误差<br>(RPD)<br>Relative percent<br>deviation |
| 随机森林 RF      | 0.157 6                                              | 2.32                                                                            | 0.901 4                                              | 0.158 3                                                               | 2.31                                             |
| 梯度提升决策树 GDBT | 0.156 6                                              | 2.35                                                                            | 0.905 8                                              | 0.156 6                                                               | 2.36                                             |
| 极端梯度上升 XGB   | 0.153 3                                              | 2.38                                                                            | 0.911 4                                              | 0.153 2                                                               | 2.43                                             |
| 堆叠 Stacking  | 0.103 4                                              | 2.51                                                                            | 0.939 1                                              | 0.091 6                                                               | 2.91                                             |

4 个模型经过 2 163 次训练后,模型测试集真实值与预测值的散点拟合图如图 7 所示。可以看出采用 RF、GDBT、XGB 建立的预测模型的预测

值较为分散,而采用本试验提出的 Stacking 集成算法,数据点密集地集中在拟合直线两侧。因此,Stacking 模型预测玉米种子含水率效果明显优于

RF、GDBT、XGB 3 种基算法。

将 80 个测试集样本按其含水率从小到大进行排序,从中等间隔抽取 23 测试集样本进行重复试验,将抽取的 23 个重复试验样本重新按 1~23 进行编号,分别采用 Stacking 模型进行 5 次预测,得到 23 组数据分别计算其均值和方差,其结果如图 8 所示。由图 8 可以看出,选取的 23 个测试样本的 5 次预测平均值大多落在直线  $y=x$  上,说明模型的预测准确度较高。第 21 号样本的波动范围最大,其 5 次预测值的标准差为 0.403 73,其余样本的 5 次重复预测值的标准差均集中在 0.260 77 以内,说明该模型的预测稳定性较好,与前面通过模型参数预测相关系数  $R_p^2$  和相对分析误差 PRD 对模型的评价结果一致,

进一步说明了该模型的预测能力和模型稳定性。

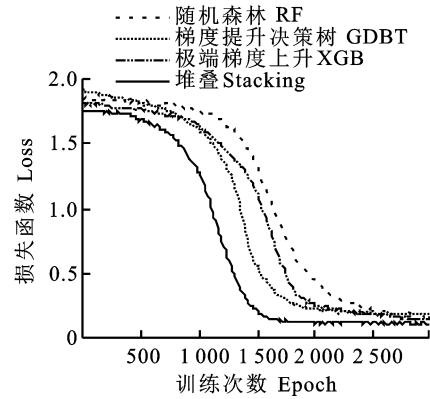
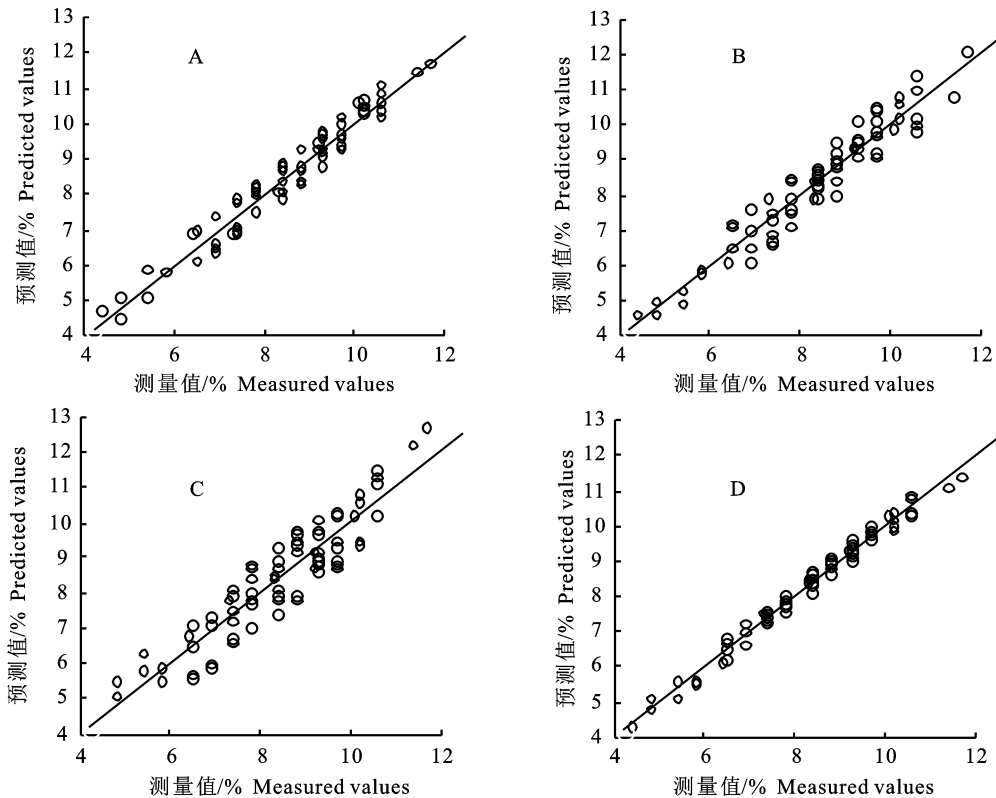


图 6 4 种模型测试集损失函数

Fig. 6 Loss function of the testing set of four models



A. 随机森林; B. 梯度提升决策树; C. 极端梯度上升; D. 堆叠

A. RF; B. GDBT; C. XGB; D. Stacking

图 7 4 种建模方法验证结果

Fig. 7 Verification results of four models

为了进一步研究训练集样本与测试集样本的代表性,对测试集样品实测值与预测值结果进行  $T$  检验,结果见表 3。由表 3 可知,测试集样品的实测值与预测值方差方程 Levene 检验结果  $P$  值为 0.961,不显著,测试集样本预测值与实测值均值方程  $T$  检验结果  $P$  值(双侧)为 0.901,不显著,表明测试集样本具有普适性,进一步说明所建

立的玉米种子含水率预测模型能够很好地对未知玉米种子样本含水率进行预测。

## 4 结 论

本研究提出一种基于集成学习的近红外光谱快速预测玉米种子含水率的方法。对于种子公司提供的,未包裹种衣剂等药物的 8 个品种 320 份

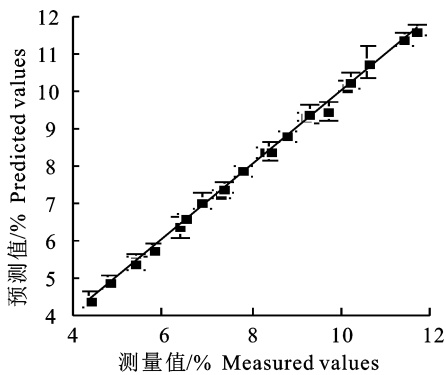


图 8 Stacking 模型重复 5 次测试结果  
Fig. 8 Test results five times by stacking model

表 3 测试集样品实测值与预测值结果进行 T 检验结果

Table 3 T test results of measured and predicted values of testing set sample

| 组别<br>Group         | 方差方程的 Levene 检验<br>Levene test of variance equation |       | 均值方程 T 检验<br>T test of mean equation |                           |                    |
|---------------------|-----------------------------------------------------|-------|--------------------------------------|---------------------------|--------------------|
|                     | F                                                   | P     | T                                    | 自由度<br>Degrees of freedom | P(双侧)<br>Bilateral |
| 实测值 Measured value  | 0.001                                               | 0.961 | 0.149                                | 69.4                      | 0.901              |
| 预测值 Predicted value |                                                     |       | 0.148                                | 68.8                      | 0.901              |

0.939 1, 相对分析误差  $PRD$  为 2.91, 5 次重复性试验的结果显示, 平均值集中在直线  $y=x$  上, 且其标准差均不超过 0.403 73。Levene 检验结果  $P$  值为 0.961,  $T$  检验结果  $P$  值(双侧)为 0.901, 表明测试样本具有普适性。

(4) 本研究所述的 Stacking 模型的预测精度和预测稳定性远高于 GDBT、RF、XGB 这些基础模型, 收敛特性好、泛化能力强, 有助于实现玉米种子含水率快速、无损、稳定地检测, 为农业生产过程管理和质量控制提供一种有效方法。

参考文献 Reference:

[1] 李少昆, 赵久然, 董树亭, 等. 中国玉米栽培研究进展与展望 [J]. 中国农业科学, 2017, 50(11): 1941-1959.  
LI SH K, ZHAO J R, DONG SH T, *et al.* Advances and prospects of maize cultivation in China [J]. *Scientia Agricultural Sinica*, 2017, 50(11): 1942-1959.

[2] 张 乐, 吴静珠, 李江波, 等. 基于随机森林的单粒玉米种子水分近红外快速定量检测 [J]. 中国粮油学报, 2021, 36(12): 114-119.  
ZHANG L, WU J ZH, LI J B, *et al.* Near infrared fast quantitative detection of single maize seed moisture based on random forest [J]. *Journal of the Chinese Cereals and Oils*, 2021, 36(12): 114-119.

[3] 祖祯祯. 玉米单粒播种子质量标准即将实施 [N]. 农民日报, 2021-10-26(007).  
ZU Y Y. The quality standard of maizesingle seed will be

玉米种子进行了近红外漫反射光谱特征的 4 种集成学习算法分析, 得出如下结论:

(1) 采用竞争性自适应重加权算法 (CARS) 进行特征波长的提取, 发现前 7 个光谱特征波长包含了玉米种子含水率大部分特征信息, 所以选择前 7 个特征光谱作为样本的特征向量。

(2) 通过对比 4 种预处理方法, 发现 Savitzky-Golay 方法结合多元散射校正法对玉米种子近红外光谱的去噪效果最优。

(3) 建立玉米种子含水率预测的 Stacking 模型, 融合了 GDBT、RF、XGB 3 种算法的优势, 2 163 次训练后模型的预测相关系数  $R_p^2$  达到

implemented soon [N]. *Farmer Daily*, 2021-10-26(007).

[4] AGELET L E, GOWEN A A, JR C, *et al.* Feasibility of conventional and roundup ready soybeans discrimination by different near infrared reflectance technologies [J]. *Food Chemistry*, 2012, 134(2): 1165-1172.

[5] 王涪州, 张万旭, 王克如, 等. 新疆玉米机械收获籽粒含水率与相关性状的关系 [J]. 西北农业学报, 2019, 28(9): 1419-1427.  
WANG W ZH, ZHANG W X, WANG K R, XIAO CH H, *et al.* Relationship between grain moisture content and related characters in Xinjiang maize mechanical harvest [J]. *Acta Agriculturae Boreali-occidentalis Sinica*, 2019, 28(9): 1419-1427.

[6] 王铁固, 张怀胜, 马 娟, 等. 玉米种子活力相关性状的杂种优势 [J]. 西北农业学报, 2012, 21(5): 92-95.  
WANG T G, ZHANG H SH, MA J, *et al.* Heterosis of maize seed vigor related traits [J]. *Acta Agriculturae Boreali-occidentalis Sinica*, 2012, 21(5): 92-95.

[7] LIN P, LI X L, DU L, JIANG SH CH, *et al.* Rapidly and exactly determining postharvest dry soybean seed quality based on machine vision technology [J]. *Scientific Reports*, 2019, 9(11): 17143-17153.

[8] 周 雷. 玉米种子活力测定方法的研究 [J]. 西北农业学报, 1996, 5(3): 57-61.  
ZHOU L. Study on the determination method of maize seed vigor [J]. *Acta Agriculturae Boreali-occidentalis Sinica*, 1996, 5(3): 57-61.

[9] WEI Y, WU F, XU J, *et al.* Visual detection of the moisture content of tea leaves with hyperspectral imaging technology

- [J]. *Journal of Food Engineering*, 2019, 248(5): 89-96.
- [10] REN G X, XUE P, SUN X Y, ZHAO G. Determination of the volatile and polyphenol constituents and the antimicrobial, antioxidant, and tyrosinase inhibitory activities of the bioactive compounds from the by-product of *Rosa rugosa* Thunb. var. *plena* regal tea [J]. *BMC Complementary and Alternative Medicine*, 2018, 307(18): 1-9.
- [11] BAE I K, HAM H M, JEONG M H, *et al.* Simultaneous determination of 15 phenolic compounds and caffeine in teas and mate using RP-HPLC/UV detection: method development and optimization of extraction process [J]. *Food Chemistry*, 2015, 172(4): 469-475.
- [12] XI X, WEI X, WANG Y, *et al.* Determination of tea polysaccharides in *Camellia sinensis* by a modified phenol-sulfuric acid method [J]. *Archives of Biological Sciences*, 2010, 62(3): 671-678.
- [13] JL A, BF A, DH A, *et al.* Discrimination of Chinese teas according to major amino acid composition by a colorimetric IDA sensor-ScienceDirect [J]. *Sensors and Actuators B: Chemical*, 2017, 240(3): 770-778.
- [14] 刘秀英, 余俊茹, 王世华. 光谱特征变量和 BP 神经网络构建油用牡丹种子含水量估算模型 [J]. *农业工程学报*, 2020, 36(22): 308-315.
- LIU X Y, YU J R, WANG SH H. Spectral characteristic variables and BP neural network are used to construct the estimation model of water content of oil peony seeds [J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2020, 36(22): 308-315.
- [15] 姜 妍, 蒙 何, 赵熠荣, 等. 青饲大豆主要品质参数的近红外光谱快速分析 [J]. *光谱学与光谱分析*, 2022, 42(3): 919-923.
- JIANG Y, MENG H, ZHAO Y R, *et al.* Fast analysis of main quality parameters of green fed soybean by near infrared spectroscopy [J]. *Spectroscopy and Spectral Analysis*, 2022, 42(3): 919-923.
- [16] 焦 俊, 圣 阳, 王 标, 等. 光谱技术在核桃检测领域的研究进展 [J]. *激光与光电子学进展*, 2021, 58(22): 44-54.
- JIAO J, SH Y, WANG B, *et al.* Research progress of spectral technology in walnut detection [J]. *Laser & Optoelectronics Progress*, 2021, 58(22): 44-54.
- [17] 余克强, 孟 浩, 曹晓峰, 等. 近红外光谱对贮藏期猕猴桃不同深度果肉色泽的变化研究 [J]. *光谱学与光谱分析*, 2020, 40(7): 2240-2245.
- YU K Q, MENG H, CAO X F, *et al.* Study on changes of flesh color of kiwifruit at different depths during storage by near infrared spectroscopy [J]. *Spectroscopy and Spectral Analysis*, 2020, 40(7): 2240-2245.
- [18] 康文翠, 林 颖, 左 敏, 等. 纳米色敏传感-可见/近红外光谱的霉变小麦菌落数定量分析 [J]. *光谱学与光谱分析*, 2020, 40(5): 1569-1574.
- KANG W C, LIN H, ZUO M, *et al.* Quantitative analysis of moldy wheat colony number by nano color sensitive sensing visible / near infrared spectroscopy [J]. *Spectroscopy and Spectral Analysis*, 2020, 40(5): 1569-1574.
- [19] ALISHAHI A, FARAHMAND H, PRIETO N, *et al.* Identification of transgenic foods using NIR spectroscopy: a review [J]. *Spectrochimica Acta Part A Molecular & Biomolecular Spectroscopy*, 2010, 75(1): 1-7.
- [20] MISHRA P, NORDON A, TSCHANNERL J, *et al.* Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products [J]. *Journal of Food Engineering*, 2018, 238(12): 70-77.
- [21] 谈爱玲, 王晓斯, 楚振原, 等. 基于近红外光谱融合与深度学习的玉米成分定量建模方法 [J]. *食品与发酵工业*, 2020, 46(23): 213-219.
- TAN A L, WANG X S, CHU ZH Y, *et al.* Quantitative modeling method of maize components based on near infrared spectrum fusion and deep learning [J]. *Food and Fermentation Industries*, 2020, 46(23): 213-219.
- [22] ZHANG Y, GUO W. Moisture content detection of maize seed based on visible/near-infrared and near-infrared hyperspectral imaging technology [J]. *International Journal of Food Science & Technology*, 2020, 55(2): 631-640.
- [23] BAI X L, ZHANG L, KANG C, *et al.* Near-infrared spectroscopy and machine learning-based technique to predict quality-related parameters in instant tea [J]. *Scientific Reports*, 2022, 12(3): 3833-3840.
- [24] HU S, LI H, CHEN C, *et al.* Raman spectroscopy combined with machine learning algorithms to detect adulterated Suichang native honey [J]. *Scientific Reports*, 2022, 12(3): 3456-3465.
- [25] MOHAMMADI, S, NARIMANI, Z, ASHOURI, M, *et al.* Ensemble learning from ensemble docking: revisiting the optimum ensemble size problem [J]. *Scientific Reports*, 2022, 12(1): 410-424.
- [26] WANG Y L, *et al.* Random Bits Forest: a Strong Classifier/Regressor for Big Data [J]. *Scientific Reports*, 2016, 6(7): 30086-30093.
- [27] WU Y, ZHANG Y, ZOU X, *et al.* Estimated Date of Delivery with Electronic Medical Records by a Hybrid GB-DT-GRU Model [J]. *Scientific Reports*, 2022, 12(3): 4892-4902.
- [28] CHEN W, FU K, ZUO J, *et al.* Radar emitter classification for large data set based on weighted-XGBoost [J]. *IET Radar Sonar Navigation*, 2017, 11(8): 1203-1207.
- [29] NICOLA B M, BEULLENS K, BOBELYN E, *et al.* Non-destructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review [J]. *Postharvest Biology & Technology*, 2007, 46(2): 99-118.

## Content in Maize Seeds Based on Ensemble Learning and Near Infrared Spectroscopy

YANG Lin<sup>1</sup>, ZHANG Lin<sup>1</sup> and YE Zehui<sup>2</sup>

(1. College of Electronic Information and Electrical Engineering, Shangluo University, Shangluo Shaanxi 726000, China;

2. Shaanxi Shangdan Gaoxin School, Shangluo Shaanxi 726000, China)

**Abstract** Traditional chemical methods for measuring moisture content in maize seeds are complex in process, long in time and high in cost etc., a fast and nondestructive method for predicting moisture contents in maize seeds was proposed in this paper based on ensemble learning algorithm and near infrared spectroscopy. In this study, 320 maize seeds of 8 varieties such as ‘Shannke 9’ were collected by near-infrared spectroscopy (Antaris II, Nicolet, USA). Partial Least Squares Regression (PLS) method was used to compare and analyze the pretreatment effects of Savitzky-Golay (SG) and combination of four spectral pretreatment methods on NIR spectra of maize seeds. It is found that combination of Savitzky-Golay method with multiple scattering correction method has the best denoising effect. Competitive adaptive reweighting algorithm (CARS) was used to extract feature wavelengths, and the cumulative contribution rate of the first seven spectral features was over 92%. GBDT (Gradient Boosting Decision Tree), RF (Random Forest), XGB (or XGBoost, Extreme Gradient Boosting) was used as the basic model, and Stacking was used as a fusion strategy to build a stacking ensemble learning model. After pretreatment, the first 7 principal components were extracted as feature vectors, and the moisture content of these seeds was obtained by direct drying method as labels. Four prediction models for moisture content in maize seeds were trained respectively, and the performance indicators of the four models were compared and analyzed. After 2 163 times training, the prediction correlation coefficient  $R_p^2$  of the stacking ensemble model was 0.939 1, and the relative analysis error PRD was 2.91. The results showed that the stacking ensemble model ensemble the advantages of GDBT, RF and XGB models, with high precision, good convergence characteristics and strong generalization ability, thus providing a new idea for fast and nondestructive determination of moisture contents in maize seeds.

**Key words** Maize seeds; NIRs; Ensemble learning; Stacking; Moisture content

**Received** 2022-05-19

**Returned** 2022-06-23

**Foundation item** Scientific Research Project of Shangluo University (No. 18sky-fwd001).

**First author** YANG Lin, female, master, lecturer. Research area: the application research of machine learning in agriculture. E-mail: 1742198437@qq.com

**Corresponding author** ZHANG Lin, male, master, lecturer. Research area: near infrared spectroscopy. E-mail: 759013520@qq.com

(责任编辑: 成敏 Responsible editor: CHENG Min)