doi:10.6041/j.issn.1000-1298.2021.S0.019

基于集成学习的农业生产技术效率评价方法

冯建英　苏允汇　龚劭齐　王　智　穆维松
(中国农业大学信息与电气工程学院, 北京 100083)

摘要: 提高农业生产技术效率是推动农业高质量发展的重要内容,但传统的基于前沿面的技术效率评价模型在实际应用中存在模型运算速度慢和灵活性低等问题,难以对大量新增样本的效率进行快速评价。基于此,本研究将基于前沿面的 DEA 技术效率测算模型与集成学习模型相结合,提出一种农业生产技术效率的评估预测方法,并利用葡萄生产技术效率数据集验证了模型的效果。研究结果显示,Stacking 融合模型的准确率和 AUC 分别达到了 94.8% 和 0.984,均优于其他对比模型,表明基于 Stacking 集成学习模型具有较高的预测准确性,能够实现更加高效、快速的技术效率评价。

关键词: 农业生产; 技术效率; 效率层级分析; 数据包络分析; 多模型融合; Stacking 集成学习
中图分类号: TP391.4　　**文献标识码**: A　　**文章编号**: 1000-1298(2021)S0-0148-08

Evaluation Method of Agricultural Production Technical Efficiency Based on Ensemble Learning

FENG Jianying　SU Yunhui　GONG Shaoqi　WANG Zhi　MU Weisong
(*College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*)

Abstract: Improving the technical efficiency of agricultural production is an important part to promote the high-quality development of agriculture. However, in practical application, there exist some flaws in the traditional technical efficiency evaluation model based on the frontier, such as slow computing speed and low flexibility, which make it difficult to evaluate the efficiency of a large number of new samples. For the above reasons, a method for evaluating and predicting the technical efficiency of agricultural production was proposed, which combined the DEA technical efficiency measurement model based on the frontier with the ensemble learning model, and the grape production technical efficiency dataset was used to verify the effect of the model. Experiments showed that the Stacking fusion model reached the accuracy and AUC of 94.8% and 0.984 respectively, with promising result that surpassed the other comparison models, indicating that the Stacking ensemble learning model had high accuracy, robustness and generalization ability, and can achieve more efficient, fast and stable technical efficiency evaluation.

Key words: agricultural production; technical efficiency; efficiency tier analysis; DEA; multi-model fusion; Stacking ensemble learning

0　引言

推进农业高质量发展,需要转变农业发展方式,加快推进农业实现质量变革、动力变革和效率变革[1],提高农业生产技术效率是推动农业高质量发展的重要内容和必经路径。

在农业生产效率测算和评价方面,目前研究一般均是基于构造前沿面的参数方法或非参数方法模型及其变形[2-5]。其中,以数据包络分析(Data envelopment analysis,DEA)方法为代表的非参数模型因其不需要确定具体的生产函数形式、能避免生产函数设定不当引起的测算误差,被国内外众多学者用以测算农业生产技术效率[6-9]。但基于 DEA 方法的效率测度模型在实际应用中存在运算效率较

收稿日期: 2021-07-13　修回日期: 2021-09-07
基金项目: 财政部和农业农村部:国家现代农业产业技术体系项目(CARS-29)
作者简介: 冯建英(1982—),女,副教授,博士,主要从事农业大数据分析与智能决策研究,E-mail: fjying@cau.edu.cn
通信作者: 穆维松(1967—),女,教授,博士,主要从事农业大数据分析与智能决策研究,E-mail: wsmu@cau.edu.cn

慢、模型灵活性不高等问题[10]。DEA 模型的效率测算原理依赖于其中若干样本构成的技术前沿面，在实践中当需要评价一个新增的生产单元的技术效率时，必须将其加入整体样本集，通过线性规划求解新的前沿面，从而获得各个生产单元的效率[10-11]。

机器学习中的神经网络、支持向量机、随机森林等模型在拟合输入输出关系中具有优异表现，显示了该类算法模型在发现数据内在非线性关系和规律方面的优势[12-14]。集成学习（Ensemble learning）是机器学习的一个分支。作为一类组合优化的学习方法，集成学习不仅能通过组合多个简单模型以获得一个性能更优的组合模型，而且允许研究者针对具体的机器学习问题设计组合方案以得到更为强大的解决方案[15]。近年来，许多学者致力于将集成学习理论应用于农业领域的分类、识别、预测建模等研究中，如谢文涌等[16]提出了多模型融合的 Stacking 集成学习方法对 6 种不同品系的金线莲叶片进行分类识别；袁培森等[17]利用堆叠式两阶段集成学习的分类器组合模型，建立了水稻表型组学实体分类模型；海兵帅[18]将集成学习机制应用于光谱与品质相关性分析，建立了稳健精简的苹果品质定量分析模型；谢元澄等[19]利用基于特征选择集成学习模型，实现了对果蝇求偶行为的自动识别。这些研究表明，集成学习在分类问题的应用中具有泛化能力强、稳定性高的优势。

为弥补 DEA 模型的局限性，并发挥机器学习算法非线性关系拟合的优势，本文提出一种将基于前沿面的 DEA 技术效率测算模型与基于机器学习的智能算法相结合的农业生产技术效率评价方法，将集成学习模型应用于农业生产技术效率的评估预测。

1　算法设计

1.1　预测评价整体思路

技术效率反映了生产投入要素的组合与产出之间的关系，基于集成学习的农业生产技术效率评价和预测方法的核心思想是：首先对没有技术效率标签的样本，基于 DEA 模型计算各生产单元的技术效率，然后基于机器学习中的分类算法构建拟合要素投入产出和技术效率标签的个体学习器，通过不断调整模型的参数使学习器的性能得以优化，并通过 Stacking 融合策略进行模型集成融合，最终构建具有良好预测准确率和较高预测效率的集成学习模型。则对于新增样本，可以直接利用该样本的投入、产出数据和集成学习模型预测其技术效率，而不用对包含新、老样本的数据集重新进行数学规划求解。该方法一方面可以发挥机器学习模型的运算速度优势，另一方面可以避免传统技术效率评价方法无法迅速、灵活处理新增评价单元、必须重新进行数学规划运算而导致的模型灵活性差、效率低等问题。

1.2　DEA－BCC 模型

DEA－CCR 模型衡量的是相对效率，即需要通过比较不同决策单元（Decision making unit，DMU）的投入产出来计算各决策单元的技术效率[20]。在本研究中，一个生产单元（农户）是一个决策单元。选取的每个生产单元用 DMU_j 表示（$j=1,2,\cdots,n$），将投入指标数量设为 m，而产出指标数量设为 s，则生产单元的投入可用向量表示为 $\boldsymbol{X}_j=(x_{1j},x_{2j},\cdots,x_{mj})$，产出则用向量 $\boldsymbol{Y}_j=(y_{1j},y_{2j},\cdots,y_{sj})$ 来表示。用 v_i 表示投入的权重（$i=1,2,\cdots,m$），u_r 表示各产出项的权重（$r=1,2,\cdots,s$）。将 DMU_j 的产出项与产出项权重相乘并求和，除以投入项和投入项权重相乘后的和，就得到了所有生产单元的相对效率 h_j，h_j 满足

$$h_j=\max\left(\frac{\sum_{r=1}^{s}u_r y_{rj}}{\sum_{i=1}^{m}v_i x_{ij}}\right)\leqslant 1 \tag{1}$$

将不等式转化为对偶规划模型，并加入分别关于投入、产出的对偶变量 S_i^- 和 S_r^+，并转换为如下规划问题

$$\begin{cases}\min\theta\\ \text{s.t.}\begin{cases}\sum_{j=1}^{n}\boldsymbol{X}_j\lambda_j+S_i^-=\theta\boldsymbol{X}_0\\ \sum_{j=1}^{n}\boldsymbol{Y}_j\lambda_j-S_r^+=\boldsymbol{Y}_0\end{cases}\quad(\lambda_j\geqslant 0,S_i^-\geqslant 0,S_r^+\geqslant 0)\end{cases} \tag{2}$$

式中　θ——被评价决策单元的纯技术效率

$\boldsymbol{X}_0$——被评价决策单元的投入

$\boldsymbol{Y}_0$——被评价决策单元的产出

λ_j——第 j 个决策单元样本权重系数

基于规模报酬变化的 DEA－BCC[20] 模型与 CCR 模型不同之处在于前者将 $\sum_{j=1}^{n}\lambda_j=1$ 作为求解目标的一个凸约束条件加入到约束条件集合当中，即

$$\begin{cases}\min\theta\\ \text{s.t.}\begin{cases}\sum_{j=1}^{n}\boldsymbol{X}_j\lambda_j+S_i^-=\theta\boldsymbol{X}_0\\ \sum_{j=1}^{n}\boldsymbol{Y}_j\lambda_j-S_r^+=\boldsymbol{Y}_0\\ \sum_{j=1}^{n}\lambda_j=1\end{cases}\quad(\lambda_j\geqslant 0,S_i^-\geqslant 0,S_r^+\geqslant 0)\end{cases} \tag{3}$$

该模型中 θ 表示纯技术效率,当 $\theta=1$ 时,说明该生产单元满足技术有效,当其小于 1 时说明存在技术效率损失。

1.3 效率层级分析

根据已有研究结果和经验,基于 DEA - BCC 模型测算的农业生产技术效率,一般只有极少数生产单元位于前沿面,而绝大多数生产单元不位于前沿面[22-23],这将导致数据集存在严重的数据不平衡问题,对算法的学习过程产生极大的干扰,因此,将生产单元是否位于前沿面作为机器学习的分类标签是不合适的。此外,评价农业生产单元的技术效率,根本目的是为了帮助生产者最大化发挥技术优势、优化生产过程中的投入和产出,促进生产节能降耗和提质增效。但是,不同农业生产单元的自然、经济、技术等条件可能存在巨大差异,很多决策单元能够改进投入产出状况,但并不一定达到前沿面,如果仅将极少数位于生产前沿面的生产单元作为样本集的标准,大多数决策单元依然难以达到要求,将使模型的实际价值下降。

为解决上述问题,引入 HONG 等[11]提出的效率层级分析(Efficiency tier analysis)概念,在 DEA 技术效率模型的基础上扩大对“高效率”这一概念的内涵,以期训练出更有实际指导意义的机器学习模型。

效率层级分析是一个不断迭代、将位于前沿面和不在前沿面的生产单元不断分离的过程。在每一轮迭代中,都会利用 DEA 模型进行运算,得出位于当前层级的前沿面的样本,效率层级流程图如图 1 所示。

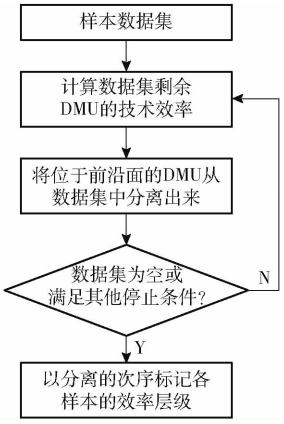

图 1 效率层级分析流程图

Fig. 1 Flow chart of efficiency tier analysis

经效率层级迭代,测算出农业生产单元对应的分层技术效率,为每个生产单元找到技术效率标签(所属效率层级),得到相对平衡的技术效率数据,实现分层的多类别的效率度量。

1.4 集成学习算法

集成学习将多个个体学习器(或称组件学习器)结合起来,得到最终训练好的模型,最终的预测结果取决于多个学习器的整合。

按照个体学习器之间的依赖关系,可将集成学习分为 2 类:个体学习器之间没有依赖或依赖关系极弱,个体模型的训练和预测都独立进行,算法在工程上可以通过并行方式执行,提高模型程序的执行速度,此类算法的代表是 Bagging 算法;个体学习器之间的依赖关系较强,无论模型训练还是预测,每个学习器的构建都需要基于已经生成的其他学习器,因而不同的个体学习器只能通过串行方式依次生成,该类算法以 Boosting 系列算法为代表。

个体学习器的结合策略主要包括平均法、投票法和学习法。其中,学习法通过另一个学习器(称为次级学习器)来进行个体学习器(称为初级学习器)的结合,是一种更为强大的结合策略。

因此,本文选择 Bagging 算法中表现较好的随机森林(Random forest)算法、Boosting 算法中表现较好的 XGBoost 算法和“学习法”的典型代表 Stacking 算法,作为农业生产技术效率预测评价的主要集成学习算法。

1.4.1 随机森林算法

Bagging 算法通过设置相关参数,对原始训练集进行随机采样,得到每个弱学习器的训练集。即训练一个含有 T 个弱学习器的 Bagging 模型,一共需要经过 T 次随机采样的过程,每个弱学习器都有属于自己的一个采样集,基于这 T 个采样集,可以独立地构建出 T 个互不干扰的弱学习器,再采用特定的结合策略对 T 个学习器的模型结果进行处理,从而得到最终的强学习器。自助采样法(Bootstapped sampling)是最为广泛使用的随机采样策略。随机森林算法是在 Bagging 算法基础之上衍生而来,它的基分类器是决策树模型,在通过随机采样生成基模型的训练集时,同样会有基于 Bootstrap 的采样过程,同时,又加上了特征维度的随机选择。即每次采样都在原始数据集的特征空间中随机选择某几个特征作为弱学习器的特征,最后采用投票方式对随机采样生成的模型预测结果进行汇总,得到最终的分类结果。大量的实践证明,随机森林算法具有极高的预测精度,且能够有效地避免抗噪声和异常值的影响,也不易出现过拟合现象[24-25]。随机森林算法的训练可以通过并行处理来实现,有效地保证了算法的效率和可拓展性[26]。

1.4.2 XGBoost 算法

XGBoost 算法本质上是在循环迭代过程中不断生成决策树的过程[27],最终该样本对应的预测值只需要将每棵树对应的分数加起来便可得到[28-30],即

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i) \tag{4}$$

式中　$\hat{y}$——最终预测值

K——决策树总数

$f_k(x_i)$——第 k 棵树的拟合结果

在第 t 轮迭代中，当前的预测分数可以用前 $t-1$ 轮的迭代结果加上本轮的残差拟合值来表示，即

$$\hat{y}^{(t)} = \sum_{k=1}^{t} f_k(x) = \hat{y}^{(t-1)} + f_t(x) \tag{5}$$

式中　$\hat{y}^{(t)}$——第 t 轮迭代的预测结果

$f_t(x)$——第 t 轮迭代的残差

根据该模型的定义，对于 n 个样本、K 个个体学习器，XGBoost 算法的目标函数 O_{bj} 可以表示成误差函数和正则化项的和，即

$$O_{bj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

其中

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \sum_{j=1}^{T_k} w_j^2 \tag{7}$$

式中　$L(y_i, \hat{y}_i)$——误差函数，用于衡量拟合值和预测值之间的误差

$\Omega(f_k)$——正则化项，用于防止模型过拟合

T_k——第 k 棵树的叶子节点个数

w_j——叶子节点上各类别的预测分数

γ、λ——人为定义的系数，分别用于控制叶子节点个数和防止叶子节点的分数值过高

将式(7)代入式(6)，得到经过 t 轮迭代之后的优化目标函数为

$$O_{bj}^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{8}$$

将式(8)中的误差函数项对 $\hat{y}^{(t-1)}$ 做二阶泰勒展开，即可得到目标函数的近似值为

$$O_{bj}^{(t)} \approx \sum_{i=1}^{n} \left(L(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right) + \Omega(f_t) \tag{9}$$

其中

$$g_i = \partial_{\hat{y}^{(t-1)}} L(y_i, \hat{y}^{(t-1)})$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} L(y_i, \hat{y}^{(t-1)})$$

将式(9)作变换，可以推导出目标函数为

$$O_{bj}^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \tag{10}$$

为方便表示，令 $G_j = \sum_{i \in I_j} g_i$，$H_j = \sum_{i \in I_j} h_i$。再将 $O_{bj}^{(t)}$ 对 w_j 求导，并令其导函数为 0，此时的等式即为目标函数得到最小值时的等价条件，可以计算出此时每个叶子节点的值 w_j^*，此时 w_j^* 和目标函数表达式为

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{11}$$

$$O_{bj}^{(t)} = \frac{1}{2} \sum_{j=1}^{T} -\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{12}$$

此时目标函数即评判该轮迭代生成的决策树性能的指标。另外，XGBoost 算法引入了一种类似于 CART 回归树的贪婪算法机制来防止模型过拟合。

1.4.3　Stacking 融合模型

与一般的集成学习算法不同，模型融合方法通过将不同算法结果进行融合，充分吸收不同算法的优势，同时避免某些算法对特定数据集的局限性，从而组合成一个更为强大的预测模型。本文采用堆叠法(Stacking)融合策略。Stacking 是将几个分类或回归模型的结果通过元分类器或元回归器进一步拟合输出最终结果的集成学习技术[31]。Stacking 有 2 层学习框架，第 1 层由若干个个体学习器组成，个体学习器基于整个数据集进行训练，第 2 层模型则将第 1 层基模型的输出结果作为特征来进行训练。Stacking 框架示意图如图 2 所示。

1.5　机器学习模型性能评价指标

选取准确率(Accuracy)[16-17]和受试者工作特征曲线下面积(Area under the ROC curve，AUC)[32]来评价模型的性能。准确率是测试集上被正确分类的样本数占总体样本的比例。

分类算法可以输出某个样本属于各个预测标签的概率。将模型在测试集上预测的样本属于正例的概率按照从大到小排列，然后依次将上述概率作为输出正或负例标签的阈值。高于阈值的样本作为正例样本，反之为负例样本，那么在每个阈值下都可以计算出此时模型在测试集上的真正例率(TPR)和假正例率(FPR)，以 FPR 为横坐标，TPR 为纵坐标，将各个阈值下的 FPR、TPR 在坐标系中对应的坐标点连成线，就构成了受试者工作特征曲线(ROC 曲线)。曲线下方面积即 AUC，根据定义，AUC 值越大，则分类器的性能表现越好。

2　数据集建立与数据预处理

将算法应用于葡萄生产技术效率的预测评价，数据集采用国家葡萄产业技术体系 2017 年在全国葡萄主产区采集的葡萄园投入产出基础数据。

本研究选取 5 个变量作为技术效率评价模型的输入：

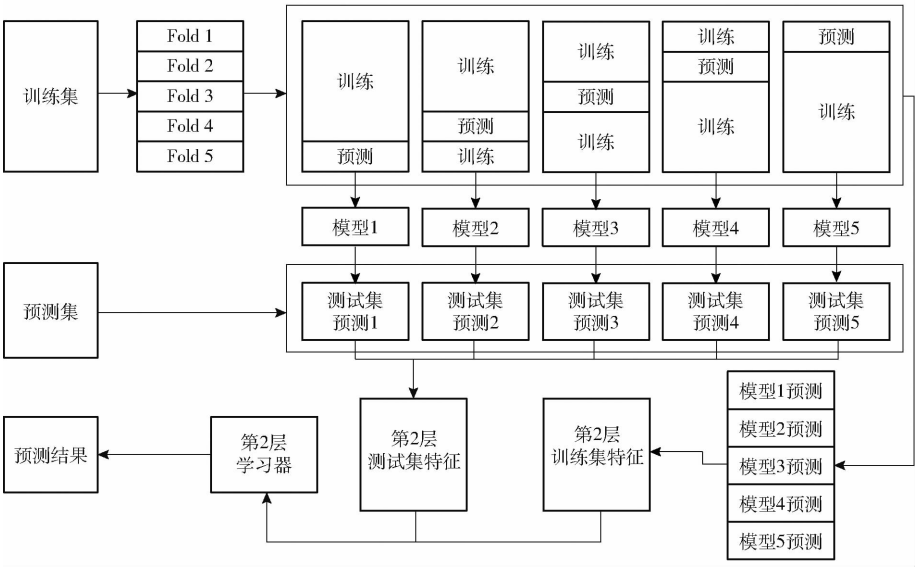

图 2 Stacking 框架示意图

Fig. 2 Frame diagram of Stacking

建园成本:将葡萄园建设所投入的资金折旧后分摊到每年每公顷土地的均值,加上每年每公顷额外的葡萄园修缮维护费,单位:元/hm^2。

土地成本:选取葡萄园占用土地所产生的租金、流转费用或自家农用地按照当地的土地租金和流转费用的平均值,单位:元/hm^2。

物质成本:包括种植过程中投入的肥料、农药、套袋、水、电、有机肥和生长调节剂等生产要素,以及使用农业机械等造成的折旧费用,单位:元/hm^2。

人工投入:包括在劳动生产过程中以资金流出形式产生的雇佣劳动力产生的费用以及不计资金形式的自家劳动者的劳动投入按当地的工价折算所产生的费用,单位:元/hm^2。

每公顷产值:每公顷土地收入,单位:元/hm^2。

初始数据集中的葡萄园分布在全国 23 个省市,覆盖了全国五大葡萄种植区[33],包含 783 份观测样本。经过筛选和数据清洗等预处理工作,去除了因格式混乱、缺失值过多而无法进行数据分析的样本记录,最终数据集包含 702 个观测样本,每个样本有 4 个投入特征和 1 个产出特征,通过 DEA 模型测算结果和效率层级分析结果指定每个观测样本的技术效率标签。

建模前,需要进行数据标准化(归一化)处理,消除指标之间的量纲影响[34]。通过数据标准化将训练集中样本特征的特征值 $x_i(i=1,2,\cdots,n)$ 转换成均值为 0、方差为 1 的标准值 x_i^*,即

$$x_i^* = \frac{x_i - \mu}{\sigma} \tag{13}$$

式中 μ、σ——样本均值、样本均方差

3 结果与讨论

3.1 基于 DEA 模型的技术效率测算结果

利用 MaxDEA 软件、DEA - BCC 模型测算出规模报酬可变假设下的葡萄园纯技术效率,其分布如表 1 所示。可以看出,葡萄园的技术效率集中分布在 0.2 ~ 0.399 和 0.4 ~ 0.599 这 2 个区间,占比分别为 34.62% 和 35.04%,而位于效率前沿面上的生产单元数量只有 48 个,仅占样本总数的 6.84%,如果以是否达到前沿面(技术效率为 1)为标准划分样本集,明显会导致前文所述数据极不平衡。这一结果也显示,大量葡萄园集中分布在中低水平的技术效率区间,没有达到生产前沿面,表明当前我国葡萄园的技术效率整体水平并不高,葡萄生产中存在着显著的技术效率损失,需要通过人员培训、组织方式优化、管理水平提升等途径提高生产中对技术效用的最大化利用。

表 1 葡萄园技术效率区间分布

Tab. 1 Interval distribution of vineyards technical efficiency

技术效率	样本数	占比/%
0 ~ 0.199	5	0.71
0.2 ~ 0.399	243	34.62
0.4 ~ 0.599	246	35.04
0.6 ~ 0.799	134	19.09
0.8 ~ 0.999	26	3.70
1	48	6.84

3.2 效率层级分析结果

利用效率层级分析对样本葡萄园的技术效率进行层级化。分层后的各层级频数分布结果见表 2,

表中效率等级数越小，表明决策单元位于越高层次的技术效率前沿面上。由表 2 可以看出，效率等级 1～8 的样本数变化较为缓和。

表 2　葡萄园技术效率等级分布

Tab. 2　Rank distribution of vineyards technical efficiency

效率等级	样本数	占比/%
1	48	6.84
2	56	7.98
3	82	11.68
4	93	13.25
5	85	12.11
6	83	11.82
7	70	9.97
8	55	7.83
≥9	130	18.52

考虑到数据集的特征（数据量不充分、样本中达到较高技术效率的观测样本占少数），并追求较高的分类准确率，构建机器学习模型时将技术效率位于前 3 个等级的样本作为高效率的样本（即正例样本），而将剩余样本作为低效率样本（即负例样本），以减弱数据不平衡可能带来的建模误差。

3.3　基于集成学习的技术效率建模结果与分析

3.3.1　基模型选择及参数调优

机器学习和集成学习模型均在 Python 语言环境下借助 Jupyter Notebook 实现。其中，数据导入由 pandas 库实现，数据的标准化处理、机器学习模型的构建、模型的参数调优、模型评价由 sklearn 库和 XGBoost 库实现。将样本数据集按 7∶3随机划分为训练集和测试集。同时，采用 5 折交叉验证法，选取准确率作为评判指标，以确定各模型关键参数的最优取值。实验选择 Logistic 回归模型、支持向量机算法、决策树算法作为个体学习器，选择随机森林模型、XGBoost 模型作为集成学习模型，选择 Stacking 融合方法作为模型融合策略。

为了取得更好的模型预测效果，需要对模型进行参数调整和优化。参数调优过程中根据参数的重要程度和对模型结构的影响程度对参数的优化顺序进行指定，通过网格搜索方式实现遍历，确定各模型关键参数的最优取值，各模型待调优的参数及调优结果见表 3。

3.3.2　模型性能比较分析

以样本的投入和产出特征值作为输入，以基于效率层级分析的正负样本标签作为输出，选取前 3 等级样本为正例、剩余样本作为负例，分别利用 6 种模型进行输入输出关系的拟合，各模型的性能表现见表 4。

表 3　模型参数调整结果

Tab. 3　Adjustment results of model parameters

模型	参数	调优结果
Logistic 回归	正则化系数	0
支持向量机	惩罚系数	1 000
	Gamma	0.3
决策树	max_depth	30
	min_samples_leaf	15
	min_samples_split	20
随机森林	n_estimators	500
	max_depth	11
	min_samples_leaf	5
	min_samples_split	5
XGBoost	n_estimators	100
	eta	0.1
	gamma	0.1
	max_depth	3
	min_child_weight	1
	subsample	0.7
	colsample_bytree	0.8
	alpha	0.1
	lambda	0.1

表 4　基于集成学习的葡萄园技术效率评价模型性能比较结果

Tab. 4　Performance comparison of vineyard technical efficiency evaluation models based on ensemble learning

模型	准确率/%	AUC
Logistic 回归	83.4	0.837
支持向量机	90.0	0.935
决策树	94.3	0.936
随机森林	91.0	0.969
XGBoost	94.3	0.980
Stacking	94.8	0.984

由表 4 可得，基于 DEA 和集成学习的葡萄园技术效率评价模型表现优异，整体性能表现最佳的 Stacking 融合模型的准确率和 AUC 分别为 94.8% 和 0.984。因此，在实际的农业生产技术效率评价中，可以使用更快速和灵活的集成学习模型代替单一的 DEA 模型，实现决策单元技术效率的预测评价。

4　结论

（1）针对 DEA 模型在实际应用场景中的局限性，并发挥机器学习算法非线性关系拟合的优势，提出了一种基于 DEA 模型和集成学习结合的农业生产技术效率预测评价方法，将 DEA 模型中的投入产出变量作为集成学习模型的输入特征变量，将层级化的技术效率标签作为集成学习的预测输出，在 Stacking 融合模型框架下设计了以支持向量机等模

型为个体学习器、以随机森林算法和 XGBoost 算法为元学习器的生产技术效率预测模型。

(2)对葡萄生产技术效率评价数据集的处理结果显示,Stacking 模型表现最优,准确率和 AUC 分别为 94.8% 和 0.984,高于其他对比模型。表明基于 Stacking 的集成学习能够满足农业生产技术效率的预测评价需求,且具有较高的准确性,能实现及时高效的预测评价。

参 考 文 献

[1] 孙江超. 我国农业高质量发展导向及政策建议[J]. 管理学刊, 2019, 32(6): 28 - 35
SUN Jiangchao. The orientation and policy suggestions for the agricultural high-quality development[J]. Journal of Management, 2019, 32(6): 28 - 35. (in Chinese)

[2] 康海军. 基于 C - D 生产函数的福建省林业产业投入与产出分析[J]. 内蒙古财经大学学报, 2018, 16(2): 23 - 26
KANG Haijun. Analysis of the input and output of forestry industrial in Fujian Province based on C - D production function[J]. Journal of Inner Mongolia University of Finance and Economics, 2018, 16(2): 23 - 26. (in Chinese)

[3] 谭玉莲,漆雁斌,邓鑫. 农业生产效率的真实测度及其影响因素分析——基于四川省 411 户稻农的经验研究[J]. 粮食科技与经济, 2018, 43(1): 33 - 38
TAN Yulian, QI Yanbin, DENG Xin. The real measure of agricultural production efficiency and its influencing factors—based on the experience of 411 grain farmers in Sichuan Province[J]. Grain Science and Technology and Economy, 2018, 43(1): 33 - 38. (in Chinese)

[4] 游和远,吴次芳,林宁,等. 基于数据包络分析的土地利用生态效率评价[J]. 农业工程学报, 2011, 27(3): 309 - 315
YOU Heyuan, WU Cifang, LIN Ning, et al. Assessment of eco efficiency of land use based on DEA[J]. Transactions of the CSAE, 2011, 27(3): 309 - 315. (in Chinese)

[5] 张丽娜,陈志,杨敏丽,等. 我国玉米生产效率时空特征分析[J]. 农业机械学报, 2018, 49(1): 183 - 193.
ZHANG Li'na, CHEN Zhi, YANG Minli, et al. Spatio-temporal feature of maize production efficiency in main producing provinces of China[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(1): 183 - 193. (in Chinese)

[6] GENG Q, REN Q, NOLAN R H, et al. Assessing China's agricultural water use efficiency in a green-blue water perspective: a study based on data envelopment analysis[J]. Ecological Indicators, 2019, 96(1): 329 - 335.

[7] LI Nan, JIANG Yuqing, MU Hailin, et al. Efficiency evaluation and improvement potential for the Chinese agricultural sector at the provincial level based on data envelopment analysis (DEA)[J]. Energy, 2018, 164(45): 1145 - 1160.

[8] WANG Y, SHI L, ZHANG H, et al. A data envelopment analysis of agricultural technical efficiency of Northwest Arid Areas in China[J]. Frontiers Agricultural Science and Engineering, 2017, 4(2): 195 - 207.

[9] 孙翔,黄如晖,朱婧霖,等. 基于 DEA 模型的农村生活垃圾处理工程环境及经济效益评估[J]. 农业工程学报, 2018, 34(16): 190 - 197
SUN Xiang, HUANG Ruhui, ZHU Jinglin, et al. Environment and economy efficiency valuation for rural domestic solid waste treatment demonstration projects using DEA method[J]. Transactions of the CSAE, 2018, 34(16): 190 - 197. (in Chinese)

[10] CHENG Yijun, PENG Jun, ZHOU Zhuofu, et al. A hybrid DEA-Adaboost model in supplier selection for fuzzy variable and multiple objectives[J]. IFAC-PapersOnLine, 2017, 50(1): 12255 - 12260.

[11] HONG H K, HA S H, SHIN C K, et al. Evaluating the efficiency of system integration projects using data envelopment analysis (DEA) and machine learning[J]. Expert Systems with Applications, 1999, 16(3): 283 - 296.

[12] 金浏,赵瑞,杜修力. 混凝土抗压强度尺寸效应的神经网络预测模型[J]. 北京工业大学学报, 2021, 47(3): 260 - 268.
JIN Liu, ZHAO Rui, DU Xiuli. Neural network prediction model of concrete compressive strength size effect[J]. Journal of Beijing University of Technology, 2021, 47(3): 260 - 268. (in Chinese)

[13] 魏勤,陈仕军,黄炜斌,等. 利用随机森林回归的现货市场出清价格预测方法[J]. 中国电机工程学报, 2021, 41(4): 1360 - 1367.
WEI Qin, CHEN Shijun, HUANG Weibin, et al. Forecasting method of clearing price in spot market by random forest regression[J]. Proceedings of the CSEE, 2021, 41(4): 1360 - 1367. (in Chinese)

[14] 邹旺,江伟,冯俊杰,等. 基于 ANN 和 SVM 的轴承剩余使用寿命预测[J]. 组合机床与自动化加工技术, 2021(1): 32 - 35.
ZOU Wang, JIANG Wei, FENG Junjie, et al. Bearing remaining useful life prediction based on artificial neural network and support vector machine[J]. Modular Machine Tool & Automatic Manufacturing Technique, 2021(1): 32 - 35. (in Chinese)

[15] 徐继伟,杨云. 集成学习方法:研究综述[J]. 云南大学学报(自然科学版), 2018, 40(6): 1082 - 1092.
XU Jiwei, YANG Yun. A survey of ensemble learning approaches[J]. Journal of Yunnan University (Natural Sciences Edition), 2018, 40(6): 1082 - 1092. (in Chinese)

[16] 谢文涌,柴琴琴,甘勇辉,等. 基于多特征提取和 Stacking 集成学习的金线莲品系分类[J]. 农业工程学报, 2020, 36(14): 203 - 210
XIE Wenyong, CHAI Qinqin, GAN Yonghui, et al. Strains classification of *Anoectochilus roxburghii* using multi-feature extraction and Stacking ensemble learning[J]. Transactions of the CSAE, 2020, 36(14): 203 - 210. (in Chinese)

[17] 袁培森,杨承林,宋玉红,等. 基于 Stacking 集成学习的水稻表型组学实体分类研究[J]. 农业机械学报, 2019,

50(11)：144－152.
YUAN Peisen, YANG Chenglin, SONG Yuhong, et al. Classification of rice phenomics entities based on Stacking ensemble learning[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(11)：144－152. (in Chinese)
[18] 海兵帅. 基于集成学习的苹果品质高光谱检测方法研究[D]. 南京：南京农业大学, 2017.
[19] 谢元澄,梁敬东,王书平,等. 基于特征选择集成学习的果蝇求偶行为识别[J]. 南京农业大学学报, 2011, 34(6)：82－88.
XIE Yuancheng, LIANG Jingdong, WANG Shuping, et al. Recognition of *Drosophila courtship* behavior based on feature selection ensemble leaning[J]. Journal of Nanjing Agricultural University, 2011, 34(6)：82－88. (in Chinese)
[20] CHARNES A, COOPER W W, RHODES E. Measuring the efficiency of decision making units[J]. European Journal of Operational Research, 1978, 2(6)：429－444.
[21] BANKER R D, CHARNES A, COOPER W W. Some models for estimating technical and scale inefficiencies in data envelopment analysis[J]. Management Science, 1984, 30(9)：1078－1092.
[22] 曹盟盟. 基于 DEA－Tobit 模型的大泽山葡萄特色产业投入产出效率研究[D]. 淄博：山东理工大学, 2016.
[23] 梁流涛,曲福田,王春华. 基于 DEA 方法的耕地利用效率分析[J]. 长江流域资源与环境, 2008, 17(2)：242－246.
LIANG Liutao, QU Futian, WANG Chunhua. Analysis on cultivated land use efficiency based on DEA[J]. Resources and Environment in the Yangtze Basin, 2008, 17(2)：242－246. (in Chinese)
[24] 马玥,姜琦刚,孟治国,等. 基于随机森林算法的农耕区土地利用分类研究[J]. 农业机械学报, 2016, 47(1)：297－303.
MA Yue, JIANG Qigang, MENG Zhiguo, et al. Classification of land use in farming area based on random forest algorithm [J]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(1)：297－303. (in Chinese)
[25] 赵峦啸,刘金水,姚云霞,等. 基于随机森林算法的陆相沉积烃源岩定量地震刻画:以东海盆地长江坳陷为例[J]. 地球物理学报, 2021, 64(2)：700－715
ZHAO Luanxiao, LIU Jinshui, YAO Yunxia, et al. Quantitative seismic characterization of source rocks in lacustrine depositional setting using the random forest method:an example from the Changjiang sag in East China Sea basin[J]. Chinese Journal of Geophysics, 2021, 64(2)：700－715. (in Chinese)
[26] 王奕森,夏树涛. 集成学习之随机森林算法综述[J]. 信息通信技术, 2018, 12(1)：49－55.
WANG Yisen, XIA Shutao. A survey of random forests algorithms[J]. Information and Communications Technologies, 2018, 12(1)：49－55. (in Chinese)
[27] LIN Jianping, QI Chengwei, WAN Hailang, et al. Prediction of cross-tension strength of self-piercing riveted joints using finite element simulation and XGBoost algorithm[J]. Chinese Journal of Mechanical Engineering, 2021, 34(1)：1－11.
[28] PAN Bingyue. Application of XGBoost algorithm in hourly PM2.5 concentration prediction[J]. IOP Conference Series：Earth and Environmental Science, 2018, 113(1)：012127.
[29] GUO Junqi, YANG Lan, BIE Rongfang, et al. An XGBoost-based physical fitness evaluation model using advanced feature selection and Bayesian hyper-parameter optimization for wearable running monitoring[J]. Computer Networks, 2019, 151：166－180.
[30] 崔晓晖,师栋瑜,陈志泊,等. 基于 Spark 框架 XGBoost 的林业文本并行分类方法研究[J]. 农业机械学报, 2019, 50(6)：280－287.
CUI Xiaohui, SHI Dongyu, CHEN Zhibo, et al. Parallel forestry text classification technology based on XGBoost in Spark framework[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(6)：280－287. (in Chinese)
[31] 陈涵,张超,余树全. 基于 Stacking 模型集成算法的莲都区南方红豆杉潜在分布区[J]. 浙江农林大学学报, 2019, 36(3)：494－500.
CHEN Han, ZHANG Chao, YU Shuquan. Potential distribution area of *Taxus chinensis* var. mairei in Liandu District based on a Stacking algorithm[J]. Journal of Zhejiang A&F University, 2019, 36(3)：494－500. (in Chinese)
[32] 汪云云,陈松灿. 基于 AUC 的分类器评价和设计综述[J]. 模式识别与人工智能, 2011, 24(1)：64－71.
WANG Yunyun, CHEN Songcan. A survey of evaluation and design for AUC based classifier[J]. Pattern Recognition and Artificial Intelligence, 2011, 24(1)：64－71. (in Chinese)
[33] 穆维松,冯建英,田东. 中国葡萄产业经济研究[M]. 北京：中国农业出版社, 2016.
[34] 梁超. 基于 Stacking 模型融合的工程机械核心部件寿命预测研究[J]. 软件工程, 2019, 22(12)：1－4.
LIANG Chao. Life prediction of construction machinery core components based on Stacking model fusion[J]. Software Engineer, 2019, 22(12)：1－4. (in Chinese)