

基于随机森林算法的打叶工艺参数优化

卢沛临¹, 田青¹, 李瑞东¹, 吴箭¹, 邓绍坤¹, 范茂青¹, 李强², 逢涛^{3*} (1. 云南烟叶复烤有限责任公司, 云南昆明 650000; 2. 云南烟叶复烤有限责任公司泸西复烤厂, 云南红河 652400; 3. 云南省烟草农业科学研究院, 云南昆明 650000)

摘要 通过建立随机森林回归模型, 筛选出在梗叶分离参数中影响片烟中片率的 9 个重要特征参数。通过模型优化和预测, 模型回归评分为 0.622, 预测值与真实值之间的平均相对误差小于 5%。模拟结果的相关性分析表明, 在多因素影响、非线性关系且关系复杂的情况下, 随机森林回归模型仍然可以较好地预测不同叶梗分离参数下的中片率。该方法可为打叶工艺参数的优化提供技术支持。

关键词 打叶复烤; 中片率; 参数优化; 随机森林

中图分类号 TS44⁺3 文献标识码 A

文章编号 0517-6611(2022)20-0162-04

doi:10.3969/j.issn.0517-6611.2022.20.042

开放科学(资源服务)标识码(OSID): 

Optimization of Threshing and Redrying Process Parameters Based on Random Forest Algorithm

LU Pei-lin, TIAN Qing, LI Rui-dong et al (Yunnan Tobacco Redrying Co., Ltd., Kunming, Yunnan 650000)

Abstract Through establishing random forest regression model, 9 important characteristic parameters that influenced the percentage of medium-sized strips in the stem-leaf separation parameters were screened out. Through model optimization and prediction, the model regression score was 0.622, and the average relative error between the predicted value and actual value was less than 5%. The correlation analysis of the simulation results showed that under the conditions of multi-factor influences, non-linear relationship and complex relationship, random forest regression model could predict the percentage of medium-sized strips under different leaf-stem separation parameters.

Key words Threshing and redrying; Percentage of medium-sized strips; Parameter optimization; Random forest

打叶复烤是卷烟工业企业进行原料初挑、模块配方、均匀性提升的重要环节^[1]。近年来, 随着各卷烟企业重点品牌对成品片烟核心质量指标均匀性水平要求的不断提高, 打叶复烤企业在参数设置和优化管控方面进行了大量研究。杨凯等^[2]通过烟碱值的组配模式, 探索出基于烟碱变异系数的均质化控制模式。皮亮等^[3]结合历史数据, 以过程工艺和参数标准化为切入点, 有效控制过程质量稳定性, 从而使最终产品质量指标稳定性显著提升。杨洋等^[4]通过多指标权重分析和正交试验对遵义复烤新线工艺参数进行优化。黄小艳等^[5]通过 Scikit_learn 中的决策树算法和正交试验, 根据加工历史数据, 使用机器学习方式对打叶复烤的水分参数进行了优化。通过以上研究使产品的化学成分指标、叶片结构指标的均匀性有了明显提升, 但在过程质量管控和数据价值的利用上仍有待加强。

2001 年, Breiman^[6]将决策树集成组合成随机森林。随机森林算法是一种非常具有代表性的机器学习 Bagging 集成算法^[7], 它以决策树作为基评估器, 多棵树随机组成的森林也叫随机森林, 包括随机森林分类器和随机森林回归器(random forest regressor)。成浩科等^[8]使用随机森林算法建立了河流总磷的预测模型, 分析了河流总磷的影响因素。苏志同等^[9]也使用随机森林算法对铝电解煅烧工艺参数进行了研究。由于决策树算法本身存在的局限性, 决策树模型预测效果比随机森林模型容易过拟合且受异常值的影响更大。张莉等^[10]使用随机森林和逻辑回归分类模型, 对各类烟叶样品的外观质量指标和感官质量指标进行关联分析, 使精选效率大幅度提升, 选出的烟叶工业适用性也明显提高。为适应

卷烟工业企业对片烟中片率的要求, 进一步分析中片率与打叶复烤工艺参数之间的相关性, 笔者采用随机森林回归器对 2021 年叶梗分离工艺参数和片烟中片率进行回归建模, 以期找到打叶复烤打叶工艺参数优化调控的方向。

1 研究数据与方法

1.1 研究数据 选取 2021 年泸西复烤厂配方烟叶产地相近、产品质量指标要求相同的出口备货烟叶模块共计 43 个, 从 17 646 个数据中筛选出涉及叶梗分离段一打、二打、三打、四打、五打和一至十二风分共计 33 项工艺参数及中片率数据, 共 9 792 个数据。相关数据来自泸西复烤厂 2021 年中控系统的操作日志。将选取的数据集记为 $D = \{x_{ij}\} (i = 1, 2, \dots, n; j = 1, 2, \dots, m)$, 其中 n 和 m 分别表示样本数和指标数。

1.2 随机森林回归算法 随机森林是一种集成学习算法, 集成多个决策树算法对相同现象产生重复的预测结果, 利用 bootstrap 重抽样方法从原始样本中抽取多个样本, 对每个 bootstrap 样本构建决策树, 然后将所有决策树预测平均值作为最终预测结果。随机森林回归可以看成是由多个弱预测器(决策树)集成的强预测器, 抵消了部分随机误差, 对异常值和噪声具有很好的容忍度。

随机森林回归算法步骤可归纳如下:

设从独立分布的随机向量 (X, Y) 中抽取训练集, 输入向量为 X , 输出向量为 Y , 则预测输出 $h(X)$ 的均方泛化误差为 $E_{X,Y}[Y-h(X)]^2$ 。

设 θ 为随机参数向量, 则对应的决策树为 $T(\theta)$ 。

(1) 数据集的随机选取。利用 bootstrap 方法从原始数据集 D 中有放回地随机抽取 K 个样本子集, 记作 $\theta_i (i = 1, 2, \dots, k)$, 并由此生成 k 棵回归树 $\{T(x, \theta_i)\} (i = 1, 2, \dots, k)$, 每次 boot-strap 抽样未抽到的样本组成了 k 个袋外数据(out-of-bag, OOB), 未被抽中的概率为 $(1-\frac{1}{n})^n$, 其中 n 为原始数据

作者简介 卢沛临(1994—), 男, 云南曲靖人, 助理工程师, 从事打叶复烤工艺参数的控制研究。* 通信作者, 副研究员, 博士, 从事烟草化学研究。

收稿日期 2021-11-05

D 的样本量。当样本量 n 较大时, $(1-\frac{1}{n})^n$ 将收敛于 $\frac{1}{e}$, 约为 0.368, 表明原始样本集中约有 36.8% 的“袋外数据”可能不在子样本集中, 将其作为随机森林的测试样本。

(2) 特征的随机选取。在每棵回归树的每个节点处从 m 个特征中随机抽取 m_{train} 个特征 ($m_{\text{train}} < m$), 作为当前节点的分裂特征集, 并以这 m_{train} 个特征中最好的分裂方式对该节点进行分裂。

(3) 随机森林的生成。每棵回归树开始自顶向下的递归分枝, 直到满足分割终止条件。

(4) OOB 估计及模型评价。将生成的 k 棵回归树组成随机森林回归模型, 回归的效果评价采用袋外数据 (OOB) 预测的残差均方 (MSE) 和拟合系数 (R^2)。计算公式如下:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i^{\text{OOB}})^2$$
$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i^{\text{OOB}})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

式中, y_i 为袋外数据中因变量的真实值, $\hat{y}_i^{\text{OOB}}$ 为随机森林对袋外数据的预测值, $\bar{y}$ 表示样本平均值。

(5) 模型优化。通过多次循环生成随机森林回归模型, 求出每次循环的模型拟合系数和均方根误差值, 得到最优决策树大小参数、决策树最大深度参数、内部节点再划分所需的最小样本参数、叶子节点最小样本数、最大特征数。

(6) 特征重要性评价。特征重要性评价通常使用基尼指数 (Gini index) 或者袋外数据 (OOB) 错误率作为评估指标来衡量, 采用袋外数据 (OOB) 错误率作为评估指标。对于随机森林回归中的变量重要性评分 (variable importance measure, VIM), 使用基于 permutation 随机置换的残差均方减小量进行衡量。具体过程如下:

a. 每一个自助样本建立一个回归树模型, 同时使用该模型对相应的袋外数据 OOB 进行预测, 得到 k 个袋外数据的残差均方, 记为 $\text{MSE}_1, \text{MSE}_2, \dots, \text{MSE}_k$ 。

b. 变量 X_i 在 k 个 OOB 样本中随机置换, 形成新的 OOB 测试样本, 然后用已建立的随机森林对新的 OOB 进行预测, 与第一步的计算方法相同, 得到随机置换后的 OOB 残差均方, 得到以下矩阵:

$$\begin{bmatrix} \text{MSE}_{11} & \cdots & \text{MSE}_{1k} \\ \vdots & & \vdots \\ \text{MSE}_{q1} & \cdots & \text{MSE}_{qk} \end{bmatrix}$$

c. 用 $\text{MSE}_1, \text{MSE}_2, \dots, \text{MSE}_k$ 与 OOB 残差均方矩阵对应的第 i 行向量相减, 平均后再除以标准误差得到变量 X_i 的重要性评分, 即

$$\text{score}_i = \frac{1}{S_E} \left[\sum_{j=1}^k \frac{1}{k} (\text{MSE}_j - \text{MSE}_{ij}) \right]$$

2 结果与分析

2.1 随机森林规模参数及其他参数的优化 模型规模参数 ($n_{\text{estimators}}$) 代表随机森林中树木的数量, 即基评估器的数量。这个参数对随机森林模型精确度的影响是单向的, 模型规模参数越大, 模型的效果往往越好。同时, 任何模型都具有决策边界, 当模型规模参数达到一定数值后, 随机森林的精确度往往不再上升。为了获取最优的模型效果, 使用 300 次循环, 计算不断调试模型规模参数 (得到学习曲线), 求出每次循环的模型拟合系数和均方根误差, 并以可视化形式进行展现, 便于求得最优参数值。

经过循环计算得出的最优拟合系数和均方根误差见表 1。当模型规模参数为 93 时, 以最优拟合系数作为衡量标准时, 随机森林模型精确度最大; 当模型规模参数为 17 时, 均方根误差最小。

表 1 随机森林规模参数
Table 1 The scale parameter of the random forest

模型评估指标 Evaluation indices of the model	$n_{\text{estimators}}$	评估值 Evaluated value
R^2	93	0.59
MSE	17	-6.78
	93	-7.21

通过观察学习曲线 (图 1) 发现, 当模型规模参数为 93 时, 曲线均处于平稳状态, 此时均方根误差为 7.21, 均方根误差稍有增加。为了保障模型精确且稳定可靠, 选择 93 作为随机森林规模参数 (Scikit_learn 中均方根误差被认定为一种误差, 因此显示为负值, 真实的均方根误差为其显示的绝对值)。

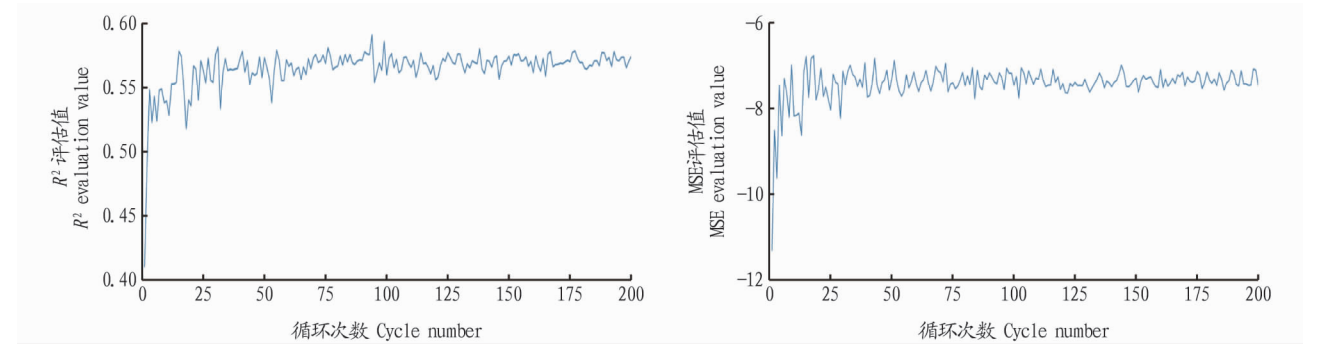


图 1 随机森林模型规模参数学习曲线
Fig.1 Random forest model scale parameter learning curve

随机森林模型在模型规模参数为 93 时,采用循环算法绘制出的树最大深度 (Max_depth) 学习曲线见图 2。表 2 为优化后模型拟合系数与均方根误差的变化。由于拟合系数和均方根误差的最优值不同,根据参数变化百分比选择对损失百分比最小的参数模型,并建立随机森林回归模型。

2.2 随机森林模型模拟结果 随机森林对中片率的模拟结

果见图 3。图 3 显示出坐标点距离直线越近,表明预测值与实际值越接近。对预测值和实际值进行相关性分析,相关系数为 0.782,预测值与实际值存在较强的非线性相关性。在多因素影响、非线性关系且关系复杂的情况下,随机森林仍然可以较好地预测不同叶梗分离参数下的中片率。

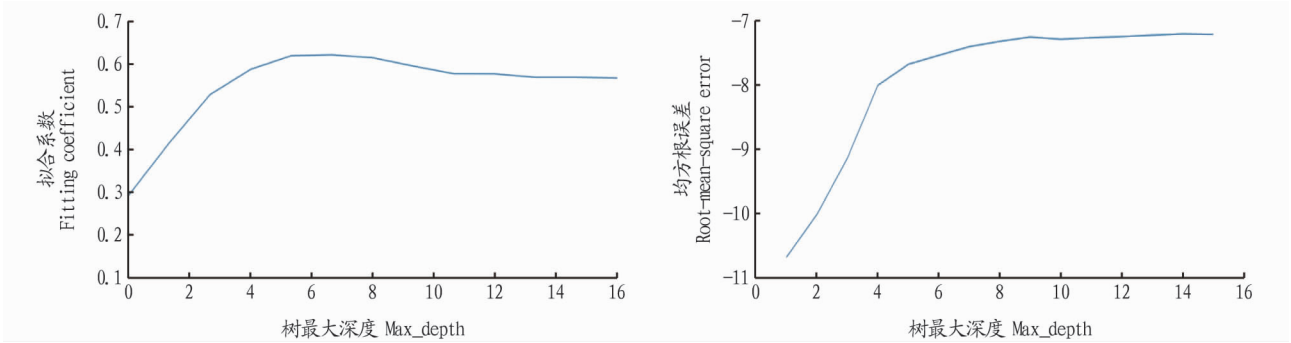


图 2 优化后的随机森林模型参数学习曲线

Fig.2 Random forest model parameter learning curve after optimization

2.3 模型的重要特征因素评价 在兼顾模型精确性和稳定性,实现综合性能最好时,将剩余的 70%测试集数据导入模型,选取决策影响度大于 5%的因素作为重要特征参数,得出影响模型决策的参数特征重要度排序,见表 3。从表 3 可以看出,针对此次建模所采用的参数指标,二打一联打辊转速、五打打辊转速、三打打辊转速和一打打辊转速是叶梗分离参数中影响最大的 4 个工艺参数指标,其中二打一联打辊转速和五打打辊转速对中片率的影响最大,分别占比 18.01%和 18.31%。

表 2 模型优化前后参数的变化

Table 2 The changes of parameters before and after the optimization of the model

模型参数 Model parameters	损失值 Loss value	损失百分比 Loss per- centage // %
拟合系数 Fitting coefficient	0.052 8	8.48
均方根误差 Root-meansquare error	0.312 8	4.14

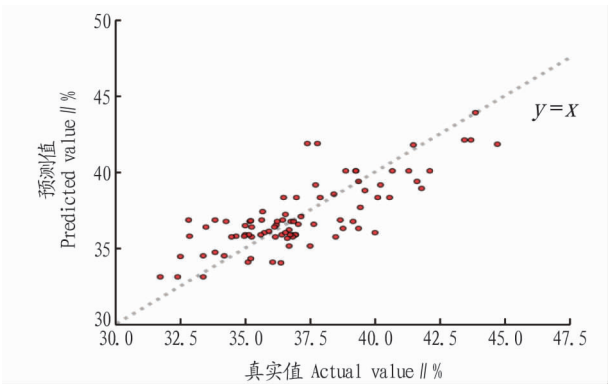


图 3 随机森林模型中片率模拟结果

Fig.3 The simulation results of the percentage of medium-sized strips in random forest model

2.4 基于重要特征因素和回归关系的试验验证 利用 Values 语法读取出测试集数据,选取测试集中的 9 个重要特征

因素进行预测,结果如表 4 所示。然后,再将中片率的预测值与真实值进行对比,结果见图 4。

表 3 模型中片率与梗叶分离工艺参数特征重要度排序

Table 3 The characteristic importance ranking of the percentage of medium-sized strips and stem-leaf separation process parameters in the model

因素 Factor	特征重要度排序 Characteristic importance ranking	特征因素 Characteristic factors	重要度 Importance %
A	1	打辊转速(五打)	18.31
B	2	一联打辊转速(二打)	18.01
C	3	打辊转速(三打)	14.90
D	4	四联打辊转速(一打)	13.90
E	5	打辊转速(四打)	8.05
F	6	二联打辊转速(二打)	7.53
G	7	二联打辊转速(一打)	7.11
H	8	一联打辊转速(一打)	6.29
I	9	三联打辊转速(一打)	5.91

使用 Scikit_learn 回归模型中的 3 个重要评判指标对预测效果进行评价,模型回归评分为 0.622(评分越接近 1,则模型预测准确度越好;当评分为负数时,表示模型预测效果较差)。该回归模型 MSE 为 1.824,预测值与真实值相比存在 1.428 的平均误差。在相对误差方面,平均相对误差为 3.93%,说明随机森林回归模型准确度能够满足现场参数调控需求。

3 结论与讨论

目前打叶复烤企业的工艺参数选择往往是依靠现场操作人员的经验来确定,热风润叶和叶梗分离参数的调控对片烟的叶片结构具有较为明显的调节作用,但在实际生产中负责不同工段的操作人员很可能不同,依靠经验进行调控的方式难以形成固有的规律。针对加工出口烟叶梗叶分离段的工艺参数,建立随机森林回归模型,初步筛选出 9 个影响中

表 4 测试集数据及中片率预测值
Table 4 The testing set data and predicted value of the percentage of medium-sized strips

序号 No.	I	II	III	IV	V	VI	VII	IX	X	中片率预测值 Percentage of mediumsized strips predicted//%
1	32	32	32	32	30	30	34	35	35	40.21
2	34	34	34	34	31	31	34	34	35	40.57
3	32	32	32	32	30	30	32	35	35	41.61
4	31	31	31	31	30	30	29	33	35	35.12
5	29	29	29	29	29	29	29	32	35	38.50
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
84	29	29	29	29	28	28	28	31	35	34.47
85	31	31	31	31	30	30	30	34	39	36.87
86	35	35	35	35	31	31	34	34	35	37.80
87	31	31	31	31	32	32	28	33	35	35.65

注:I、II、III、IV、V、VI、VII、IX、X分别为一联打辊转速(一打)、二联打辊转速(一打)、三联打辊转速(一打)、四联打辊转速(一打)、一联打辊转速(二打)、二联打辊转速(二打)、打辊转速(三打)、打辊转速(四打)、打辊转速(五打)
Note:I,II,III,IV,V,VI,VII,IX,X were the rotational speed of one-roll threshing drum(threshing 1),the rotational speed of two-roll threshing drum(threshing 1),the rotational speed of three-roll threshing drum(threshing 1),the rotational speed of four-roll threshing drum(threshing 1),the rotational speed of one-roll threshing drum(threshing 2),the rotational speed of two-roll threshing drum(threshing 2),the rotational speed of threshing drum(threshing 3),the rotational speed of threshing drum(threshing 4),and the rotational speed of threshing drum(threshing 5) respectively

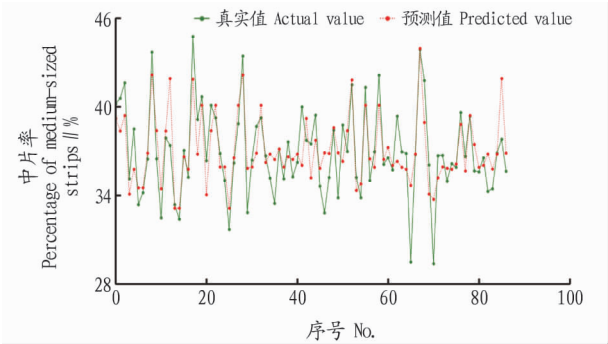


图 4 中片率预测值与实际值的比较

Fig.4 The predicted value and actual value comparison of the percentage of medium-sized strips

片率的重要特征参数,模型回归评分达到 0.622,回归模型预测效果较好。该方法可为打叶复烤打叶工艺参数的调控优化提供技术支持。

随着烟草行业数字化转型的逐步推进,在未来还需要进一步打通工商业相关业务的数据链条,获取更多高质量的业务数据,通过包括随机森林算法在内的大数据分析和机器学习

习算法,建立更加准确、适用的预测模型,实现烟叶复烤的数字化和智能化,支撑行业高质量发展。

参考文献

[1] 谢剑平.形势与未来:烟草科技发展展望[J].中国烟草学报,2017,23(3):1-7.
[2] 杨凯,陈清,徐其敏,等.打叶复烤配方均匀性控制模式的研究[C]//中国烟草学会.中国烟草 2013 年学术年会论文集.北京:中国烟草学会,2013:7-12.
[3] 皮亮,刘威,万明宇,等.基于参数标准化的片烟叶片结构管控[J].浙江农业科学,2021,62(6):1188-1191,1217.
[4] 杨洋,杨雨波,吴昊,等.烟草加工中打叶复烤工艺参数优化[J].农业工程,2018,8(8):83-85.
[5] 黄小艳,赵广仁,黄晶,等.基于决策树和正交试验设计的打叶复烤工艺参数优化[J].天津农业科学,2020,26(9):69-72.
[6] BREIMAN L.Random forests[J].Machine learning,2001,45(1):5-32.
[7] GENUER R,POGGI J M,TULEAU-MALOT C,et al.Random forests for big data[J].Big data research,2017,9:28-46.
[8] 成浩科,沈菲.基于随机森林的河流总磷预测模型及影响因素分析[J].环境保护科学,2021,47(3):62-67,117.
[9] 苏志同,汪武琚.基于随机森林的煅烧工艺参数的研究和分析[J].软件,2018,39(4):148-150.
[10] 张莉,纪铭阳,胡宗玉,等.基于随机森林和逻辑回归分类模型的烟叶精选品控指标筛选[J].江苏农业科学,2020,48(3):214-217.

(上接第 161 页)

[21] 田风霞,惠振,王国坤,等.植物体内的叶绿素降解与滞绿突变体[J].植物生理学通讯,2010,46(5):505-511.
[22] 樊昶昶,熊新星,邹亚红,等.热处理对菠菜和上海青类囊体膜及叶绿素稳定性的影响[J].中国食物与营养,2017,23(4):30-34.

[23] 曹晓倩,孙涛,帕尔哈提,等.不同处理条件对叶菜类蔬菜亚硝酸盐含量的影响[J].中国食物与营养,2018,24(2):33-36.
[24] HAO J X,LI H Y,WAN Y F,et al.Reduction of nitrite levels in fresh lettuces with aqueous chlorine dioxide treatment[J].Int J Food Sci Technol,2015,50(3):652-659.