

基于 Group-Lasso 天麻品质形成关键因子的分析

王红洁¹, 王 科², 余水祥², 马云桐^{3*}

1. 桂林理工大学理学院, 广西 桂林 541000

2. 成都工业学院大数据与人工智能学院, 四川 成都 611730

3. 成都中医药大学药学院, 四川 成都 610075

摘要: 目的 为提高人工种植天麻的质量, 基于 Group-Lasso 变量筛选构建随机森林回归模型分析影响天麻品质形成的关键因子。方法 基于 Group-Lasso 法, 对 2007—2022 年天麻质量研究文献中天麻素含量及产地环境变量等数据进行变量筛选, 并在筛选出的变量基础上建立随机森林回归模型及计算变量重要性得分。结果 最终选择了产区、生长状况、种质类型、产地气候类型、产地土壤类型、最热月均温、产地年降水量、产地年日照时数和无霜期 9 个变量, 基于被选变量与天麻素含量建立随机森林回归模型, 模型的均方误差 (mean square error, MSE) 和平均绝对百分误差 (mean absolute percentage error, MAPE) 分别为 0.103 2 和 14.08%, 特征重要性排序显示天麻素含量的最大影响因素是产地年降水量, 其次是产地土壤类型、无霜期和产地年日照时数。结论 随机森林回归模型有相对较低的误差和较高的预估精度, 更适合用于对天麻种植环境的分析和天麻素含量的估算, 为人工种植天麻提供参考。

关键词: 天麻; 天麻素; Group-Lasso; 变量筛选; 随机森林回归; 变量重要性评分

中图分类号: R282.2 文献标志码: A 文章编号: 0253-2670(2023)13-4278-08

DOI: 10.7501/j.issn.0253-2670.2023.13.020

Analysis of key factors in *Gastrodia elata* quality formation based on Group-Lasso

WANG Hong-jie¹, WANG Ke², YU Shui-xiang², MA Yun-tong³

1. College of Science, Guilin University of Technology, Guilin 541000, China

2. School of Big Data and Artificial Intelligence, Chengdu Technological University, Chengdu 611730, China

3. College of Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 610075, China

Abstract: **Objective** In order to improve the quality of artificially planted Tianma (*Gastrodia elata*), a random forest regression model based on Group-Lasso variable screening was constructed to analyze the key factors affecting the quality of *G. elata*. **Methods** Based on the Group-Lasso method, the data of gastrodin content and environmental variables of origin in the literature of *G. elata* quality research from 2007 to 2022 were screened, and the random forest regression model was then established on the selected variables, and importance score of the variables was calculated. **Results** Finally, nine variables including production area, growth status, species, production area climate type, production area soil type, average temperature in the hottest month, annual precipitation in the production area, annual sunshine hours in the production area, and frost-free period were selected. A random forest regression model was established based on the selected variables and gastrodin content. The mean square error (MSE) and mean absolute percentage error (MAPE) were 0.103 2 and 14.08%, respectively. The ranking of feature importance showed that the biggest influencing factor of gastrodin content was the annual precipitation in the production area, followed by the production area soil type, frost-free period, and annual sunshine hours in the production area. **Conclusion** The random forest regression model had relatively low error and high prediction accuracy, and was more suitable for the analysis of *G. elata* planting environment and the estimation of gastrodin content. **Key words:** *Gastrodia elata* Bl.; gastrodin; Group-Lasso; variable screening; random forest regression; variable importance measures

收稿日期: 2023-02-23

基金项目: 四川省科技厅重点研发项目: 川产地道药材大品种精深加工关键技术及产品开发生态 (2020YFN0152); 川产地道药材品质评价关键技术装备研究 (2021YFS0045)

作者简介: 王红洁, 女, 在读硕士, 研究方向为机器学习和分布式优化。E-mail: 1103623812@qq.com

*通信作者: 马云桐, 男, 教授, 从事中药资源领域相关研究。E-mail: mayuntong06@163.com

天麻 *Gastrodia elata* Bl. 属于兰科非自养型的植物, 没有根和绿叶, 种子结构简单, 没有胚乳及其他营养贮备, 仅由胚及种皮构成, 其在种子萌发后主要通过分解侵入其体内的蜜环菌菌丝获得营养以进行生长发育^[1]。天麻主要生产于我国的部分省区, 国外的研究报道甚少。研究发现天麻具有健脑、抗肿瘤及增强免疫力等作用^[2], 其入药已有悠久的历史, 在现代临床上可以有效对抗头痛、眩晕、肢体麻木、冠心病和高血压等病症^[3-4]。天麻具有重要价值, 而早期只有野生天麻, 但经过长期研究后, 成功实现了天麻的人工种植。随着技术的深入, 天麻的栽培方式也逐渐进步, 其产量和质量得到提高。由于天麻的药用价值, 市场对天麻的需求量与日俱增、质量要求更高^[5], 但影响天麻品质的原因不止是栽培和加工方式, 其生长的环境也有重要影响, 天麻的生长过程主要有避光、向气、向湿 3 个的特性, 适合在偏酸性的生态环境中生长^[6]。天麻中活性成分最高的成分是天麻素, 天麻素含量的高低是衡量天麻质量优劣的首要指标, 被《中国药典》2020 年版定为天麻的质量标准控制成分^[7-8]。探讨天麻素含量与环境因子、土壤类型、产区和生长状况之间的关系, 能够减少因人为培育不当而导致天麻数量和质量下降的问题, 对促进该产业的健康发展具有现实意义。

机器学习在特征选择和对模型的拟合上有着比浅层模型突出的优点。回归是用于分析变量间关系的一种技术^[9]。多元线性回归是自变量有 2 个及以上的回归分析, 若自变量间存在多重共线性, 那么普通线性回归将失效^[10]。共线性是指 2 个或多个自变量之间出现了相关关系, 其会对回归分析造成很大的影响^[11]。Lasso 算法在变量选择、解决多重共线性问题上具有重要作用, 它通过在回归优化函数中增加 1 个偏置项, 用绝对值偏差作为正则化项, 以减少共线性的影响, 从而减少模型误差^[12]。但实际中常常出现分类型变量的情况, Lasso 方法只能选择单个哑变量, 所以 Lasso 方法对于存在分类变量的情形通常不能得到满意的结果。而对于含有多个分类变量的哑变量, Group-Lasso 方法^[13]很好地解决了这个问题。另外, 随机森林是多棵决策树集成在一起的算法, 多棵决策树共同预测能够提高模型的准确度^[14]。随机森林回归可以有效避免模型过拟合, 提高模型精确度。

本研究从改善天麻的人工种植环境入手, 通过

Group-Lasso 算法筛选变量, 并建立天麻素含量与被选变量间的随机森林回归模型, 开展天麻种植环境的系统研究, 实现对天麻品质形成关键因子的分析, 为天麻药材种植环境的选择提供参考, 同时为人工种植天麻奠定理论基础。

1 模型与方法

1.1 Group-Lasso 方法

Lasso 回归模型属于线性回归模型的一种, 其在模型拟合过程中增加了回归系数的绝对值之和, 以残差平方和及回归系数的绝对值之和达到最小为目标来求解, 这样可以让一些系数被压缩为 0, 达到子集收缩的效果。Lasso 估计定义^[15]如下。

$$\hat{\beta}_{\lambda} = \arg \min_{\beta} (\|Y - X\beta\|^2 + \lambda \sum_{j=1}^d |\beta_j|) \quad (1)$$

$Y = (y_1, y_2, \dots, y_n)^T$ 为因变量, $X = (X^{(1)}, X^{(2)}, \dots, X^{(d)})$ 为自变量, 回归系数为 $\beta = (\beta_1, \beta_2, \dots, \beta_d)^T$, λ 为调节参数

在线性回归模型中, 当自变量除连续变量还含有分类变量时, Lasso 法通常不再适用。Lasso 方法只能选择单个的哑变量, 不能将分类变量整个选择, Group-Lasso 能在 Lasso 法的基础上解决这个问题。其估计定义^[16]如下。

$$\hat{\beta}_{\lambda} = \arg \min_{\beta} (\|Y - X\beta\|^2 + \lambda \sum_{g=1}^G \|\beta_{I_g}\|) \quad (2)$$

I_g 指第 g 组变量的下标集, β_{I_g} 指第 g 组变量的系数向量

Group-Lasso 方法的惩罚项能够看成是 L1 惩罚和 L2 惩罚的中间状态^[17], Group-Lasso 方法是在组的水平上选择变量, 也就是成组地选择变量。例如, 有 K 个水平的分类变量, 在建模过程中, 该分类变量被转化为 $K-1$ 个 0-1 变量, 并被看作是 1 个组。Group-Lasso 可以对 $K-1$ 个哑变量同时进行选择, 但 Lasso 法只能筛选出这 $K-1$ 个哑变量中的一部分, 没有实际意义。

1.2 随机森林回归

随机森林是在 bagging^[18]和决策树二者之上进行了提升。单棵决策树在回归预测时会有一定的精度, 通过增加树的数量来提高预测精度, 整个森林中的每棵树均会参与决策, 这就是随机森林的基本思想。随机森林算法提供了检验特征交互的方法, 且具有较强的泛化能力^[19]。最优的是, 由于各个决策树是独立的, 所以可以并行处理对随机森林的训练, 进一步提高成模型的效率, 有效抑制过拟合发生。

随机森林回归算法步骤^[20]: (1) 用 bootstrap 方法从 n 个原始样本中有放回地抽取 m 个样本集, 用

来构建 m 棵回归树, 未被抽到的样本组成了 m 个袋外数据集。(2) 在每个节点处, 从所有 p 个解释变量中随机抽取 k 个分割变量 ($k < p$), 根据分枝优度准则选取最优分枝。(3) 每棵回归树开始自顶向下的递归分枝, 直到满足分割的终止条件。

2 天麻品质形成关键因子分析

2.1 数据采集

根据数据已在期刊中公开发表, 发表时间为 2007—2022 年, 以天麻质量为研究主题的原则在中国知网数据库中采集研究数据, 环境因子来源于人地系统主题数据库 (<http://www.data.ac.cn/index.asp>)、中国科学院资源环境科学数据中心 (<http://www.resdc.cn>) 和 Wheat A 农业气象大数据 (<http://www.wheata.cn/>)^[21]。最终数据项包括天麻素、省区、产地、生长状况、种质类型、产地气候类型、产地土壤类型及环境因子 (年均温、最冷月均温、最热月均温、年降水量、年日照时长、无霜期)。共收集 810 条数据, 存在 3 条不完整数据, 做删除处理, 有效数据 807 条。

2.2 数据处理与分析

2.2.1 分类变量分析 天麻的产地共包含四川、云南、贵州、陕西、重庆、湖北、安徽、湖南、吉林、西藏、河南、甘肃、辽宁、黑龙江、河北和广西 16 个省份, 将省份按照 7 大地区划分, 即四川、云南、贵州、重庆和西藏为西南地区, 陕西和甘肃为西北地区, 湖北、湖南和河南为华中地区, 安徽为华东

地区, 吉林、辽宁和黑龙江为东北地区, 河北为华北地区, 广西为华南地区。天麻的产地气候类型包括亚热带湿润季风气候、大陆性季风气候、高原山地气候、低纬暖温带高原山地气候、亚热带向暖温带过渡区等数十类气候类型, 为简化模型复杂度, 将其分为亚热带气候、温带气候和高原气候 3 大类。

分析生长状况、产区、种质类型、产地气候类型和产地土壤类型 5 个分类变量与天麻素含量的关系, 如图 1 所示。本研究数据中天麻以野生为主, 红天麻的数量最多, 约占 74%, 目前红天麻和乌天麻的产量最高, 但红天麻仍是乌天麻产量的 4~5 倍, 红天麻产自黄河流域和长江流域诸省, 其种子发芽率和产量均高, 适应性和耐旱性强^[22]。天麻主要分布在西南地区, 约占 53%, 我国是野生天麻的主要分布国家之一, 主要分布在西南地区的云南、贵州和四川^[23], 云南昭通小草坝和四川平武产的野生天麻最为出名。同时, 黄壤和黄棕壤生长的天麻最多, 共占 61%, 黄壤和黄棕壤质地轻砾, 土层厚, 土壤成弱酸性, 有机质含量高及天麻所需的微量元素含量高, 是最有利于天麻生长的土壤类型^[24]。亚热带气候具有四季分明、季节分配均匀、降水量充足、热量资源丰富等特点, 更适合天麻的生长。

2.2.2 数值变量分析 图 2 展示了天麻素含量与年均温、最冷月均温、最热月均温、年降水量、年日照时长和无霜期的分布情况。天麻素含量和无霜期近似右偏分布; 产地年均温、最冷月均温、产地年

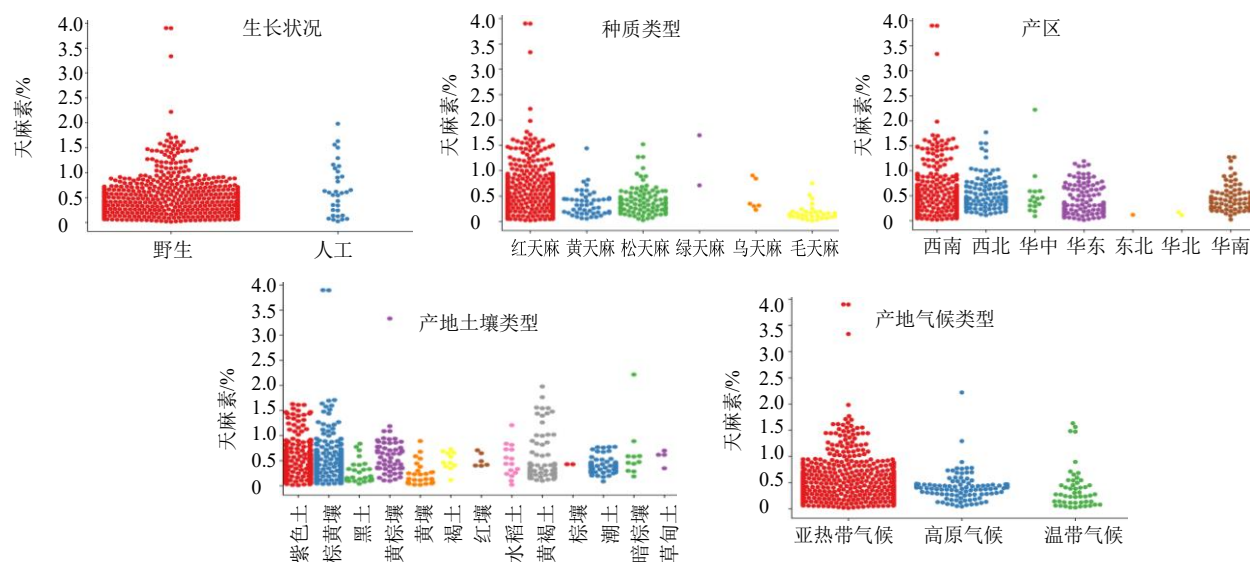


图 1 天麻素含量在分类变量中的分布

Fig. 1 Distribution of gastrodin content in categorical variables

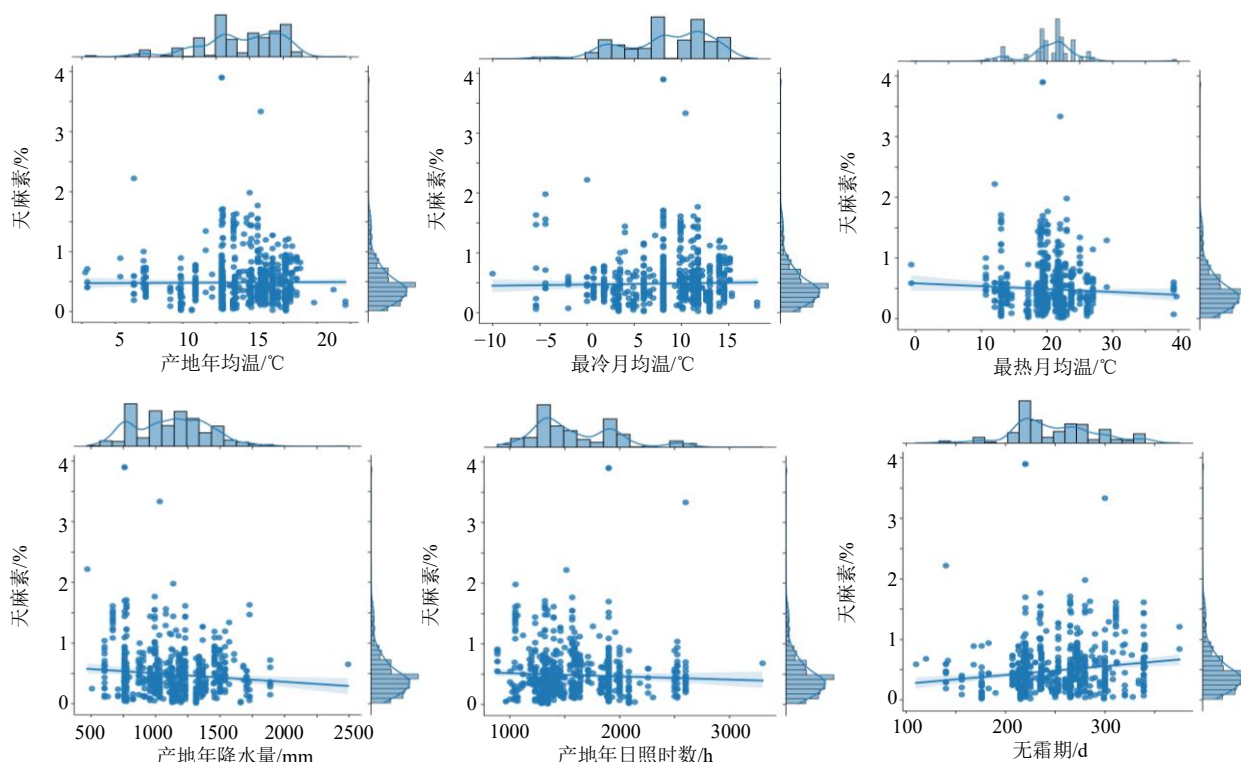


图 2 生态因子与天麻素含量关系

Fig. 2 Relationship between ecological factors and gastrodin content

降水量近似双峰分布；产地年日照时数和最热月均温的分布直方图近似双峰型与孤岛型的结合，主要由于产地不同导致。此外，图 2 还展示了天麻素含量与年均温等数值变量的关系，天麻素含量集中在产地年均温 12.5~18℃、最冷月均温在 0~15℃、最热月均温在 11~27℃、产地年降水量在 750~1500 mm、产地年日照时数在 1000~2000 h、无霜期在 210~340 d。其中，天麻素含量与无霜期的线性变化趋势最明显，随着无霜期增加天麻素含量呈上升趋势。天麻生长的区域年降水量一般在 900~1200 mm，李梁等^[25]研究指出天麻的块茎一般在 3~4 月能够萌动发芽，此时温度为 14℃左右，5~9 月适合天麻生长，此时地温在 20~25℃，天麻生长时温度不能超过 30℃。云南昭通所产天麻年平均气温为 11~13.0℃，最冷月平均气温 2.8~4.0℃，最热月平均气温 (18.0±0.4)℃，年均降雨量在 900 mm 左右^[26]。四川省野生天麻产区年温度变化较平缓，较少出现极冷、极热现象，年降水量比较充足，为 667~2033 mm，年均日照时数为 947~2079 h，无霜期 177~320 d。

2.2.3 相关性分析 分析年均温、最冷月均温、最热月均温、年降水量、年日照时长和无霜期间的相关

性，计算相关系数。如表 1 所示，无霜期与最冷月均温、最热月均温之间，年降水量与最热月均温之间，有相对较高的正相关关系；产地年日照时数与产地年降水量、无霜期之间，有相对较高的负相关关系。说明变量间存在多重共线性问题，故采用 Group-Lasso 方法对变量进行筛选。

2.3 特征筛选

2.3.1 Group-Lasso 变量筛选 以天麻素含量作为被解释变量，产区、生长状况、种质类型、产地气候类型、产地土壤类型、产地年均温、最冷月均温、最热月均温、产地年降水量、产地年日照时数和无霜期作为解释变量，将分类自变量进行 One-Hot 编码，自变量由 11 维增至 37 维。

赤池信息量准则 (Akaike information criterion, AIC)^[27]是用来衡量统计模型拟合优良性的标准，建立在熵的概念基础上，可以权衡所估计模型的复杂度和模型拟合数据的优良性。利用 R 软件的 `grpreg` 函数实现 Group-Lasso 变量筛选，以 AIC 作为筛选准则，变量选择过程如图 3 所示。结果显示，最优 AIC 值下的产地年均温和最冷月均温 2 个变量的系数为 0，即保留产区等 9 个自变量。

2.3.2 交叉验证 利用交叉验证对 Group-Lasso 筛

表 1 相关系数

Table 1 Correlation coefficients

变量	产地年均温	最冷月均温	最热月均温	产地年降水量	产地年日照时数	无霜期
产地年均温	1.000 0	-0.017 4	0.096 5	0.057 3	-0.057 2	-0.009 6
最冷月均温	-0.017 4	1.000 0	0.013 0	-0.038 1	0.069 7	0.311 8
最热月均温	0.096 5	0.013 0	1.000 0	0.448 3	-0.085 5	0.229 0
产地年降水量	0.057 3	-0.038 1	0.448 3	1.000 0	-0.249 7	0.126 0
产地年日照时数	-0.057 2	0.069 7	-0.085 5	-0.249 7	1.000 0	-0.202 9
无霜期	-0.009 6	0.311 8	0.229 0	0.126 0	-0.202 9	1.000 0

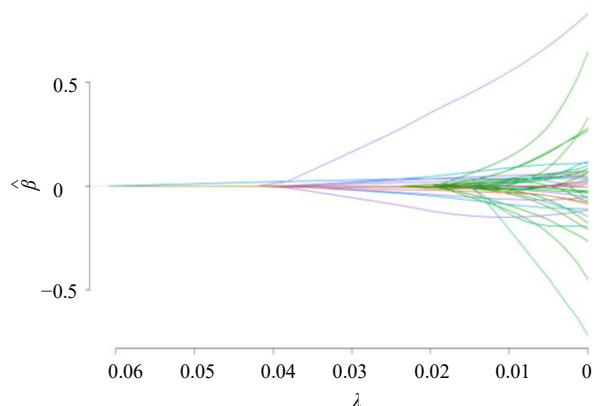


图 3 变量筛选过程

Fig. 3 Variable screening process

选结果进行验证, 结果如图 4 所示。得到最优的 lambda 参数值为 0.004 500 724, 交叉验证结果与上述筛选结果一致。

2.4 构建随机森林回归模型

2.4.1 模型建立 本研究共收集到天麻种植环境数据 807 条。以随机的方式将数据集按照 70% 的训练集和 30% 的测试集划分为 2 组, 564 条天麻数据作为训练集进行参数寻优和建立模型, 243 条天麻数据作为测试集对模型进行检验。模型自变量为 Group-

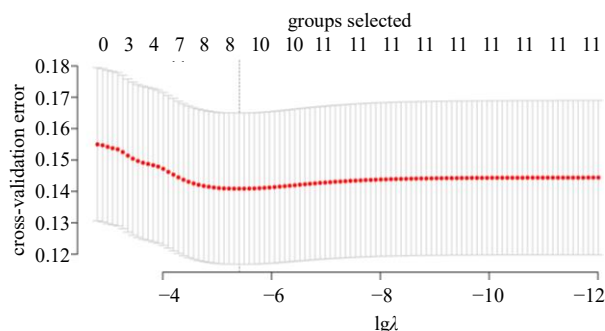


图 4 交叉验证结果

Fig. 4 Result of cross validation

Lasso 筛选出来的产区、生长状况、种质类型、产地气候类型、产地土壤类型、最热月均温、产地年降水量、产地年日照时数和无霜期 9 个变量。然后进行回归树数量的选择, 误差与回归树的关系见图 5。由图 5 可知, 回归树增大到一定量后, 误差就无太大变化, 随机森林回归树的数量达 2000 以后趋于稳定。通过多次试验, 在综合考虑均方误差大小和运算速率的情况下使用 2000 作为回归树数量。

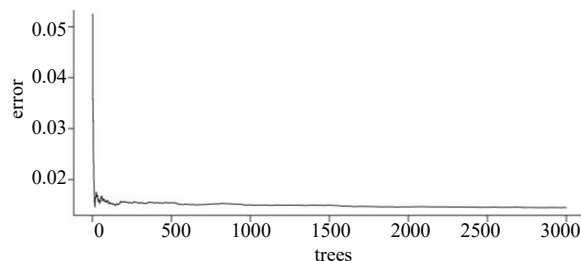


图 5 误差与回归树关系

Fig. 5 Relationship between error and regression tree

2.4.2 模型评价 为了验证随机森林算法结果的准确性, 本研究以均方误差 (mean square error, MSE) 和平均绝对百分误差 (mean absolute percentage error, MAPE) [28] 作为模型评价指标。MSE 是用来衡量实际观测值与模型预估值之间差异的指标, MAPE 是用来衡量预测模型有无偏性, 其可以准确反映实际预测误差的大小。二者计算公式如下。

$$MSE = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3)$$

$$MAPE = \frac{\sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right|}{n} \times 100, i=1, 2, \dots, n \quad (4)$$

$\hat{y}_i$ 代表模型预估值, 即天麻素含量的预测值, y_i 代表实际观测值, 即天麻素含量的实测值, n 为测试集样本数

MSE 越小, 则模型效果越好, 其预测的准确率

就越高。MAPE 值小于 10%，表示模型预测精度较高；大于 50%，则预测错误。

2.4.3 模型检验 本研究通过计算总体相对误差 (R_s)、绝对平均相对误差 (E) 2 个统计量来检验模型。

$$R_s = \frac{\sum y_i - \sum \hat{y}_i}{\sum \hat{y}_i} \quad (5)$$

$$E = \frac{1}{n} \sum \left| \frac{y_i - \hat{y}_i}{\hat{y}_i} \right| \quad (6)$$

2.4.4 特征重要性排序 随机森林可以在训练时输出变量的重要性，即哪个特征更有用。实现的方法有两种：Gini 法和置换法。置换法是改变特征观察对结果的影响，如果特征重要，那么结果应对其敏感。特征重要性在预测建模中起着重要作用，可以帮助了解数据集和模型，进行特征选择，甚至改进预测模型，从而提高预测模型的效率和有效性。回归问题的特征重要性公式^[29]如下。

$$v = \frac{\sum_{i \in oob} \exp(-(y_i - \hat{y}_i)^2) - \sum_{i \in oob} \exp(-(y_i - \hat{y}_{i,\pi})^2)}{|oob|} \quad (7)$$

y^* 为置换前的样本预测值， y 为实际值， y_{π}^* 为置换后的预测值， $|oob|$ 为样本数， m 为所有训练样本中标签值绝对值的最大值

公式 (7) 计算的是各变量在单棵决策树上的重要程度，计算出每棵树对各个变量的重要程度后，需要取平均值才能得到随机森林对各个变量的重要性，即 $\frac{1}{T} \sum_{i=1}^T v_i$ 。其中， v_i 为该变量对第 i 棵决策树的重要性， T 为决策树的数量。计算出每个变量的重要性后，将它们归一化得到最终的重要性值。

本研究将计算出各特征的相对得分，进而分析产区、生长状况、种质类型、产地气候类型、产地土壤类型、产地年降水量、产地年日照时数和无霜期对天麻素含量的影响程度，得出影响天麻有效成分的最关键因子。

2.5 模型评价结果

基于 Group-Lasso 筛选变量后，选取了多元线性回归和随机森林回归 2 种模型进行对比实验。通过计算模型 MSE 和 MAPE 来对模型的拟合效果进行评价，2 种模型的评价结果见表 2。

表 2 回归模型的预测评价

Table 2 Prediction and evaluation of regression model

模型	MSE	MAPE/%
多元线性回归	0.131 7	50.21
随机森林回归	0.103 2	14.08

从 2 种误差结果可以看出，随机森林回归误差明显小于多元线性回归误差，即随机森林的拟合效果较好，MSE 为 0.103 2，MAPE 约在 14%；多元线性回归拟合效果较差，MSE 为 0.131 7，MAPE 约在 50%。多元线性回归模型简单易操作，但其非线性学习能力不好。随机森林是多棵决策树并行的集成模型，一般预测效果更好，且能够处理非线性问题。

2.6 模型检验结果

2.6.1 比较偏差统计量 2 种模型的 R_s 、 E 结果见表 3。检验结果显示，随机森林回归的 R_s 为 0.02， E 为 37.93%，二者均小于多元线性回归。

2.6.2 评价模型预测能力 可视化多元回归和随机森林回归模型的实际观测值和预测估计值之间的差异，结果如图 6 所示。由图 6 可知，随机森林回归模型的预估精度优于多元回归模型。

2.7 特征重要性

量化随机森林中各个变量对模型的贡献，各变量的重要性大小见表 4。结果显示，对天麻素含量

表 3 回归模型的检验结果

Table 3 Test results of regression model

模型	R_s	$E/\%$
随机森林回归	0.02	37.93
多元线性回归	0.06	48.79

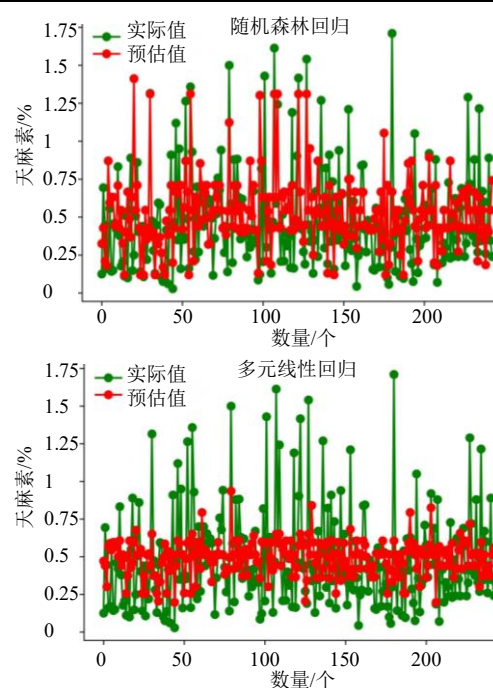


图 6 实际值与预估值的差异

Fig. 6 Difference between actual value and estimated value

表 4 特征重要性
Table 4 Feature importance value

变量	重要性
产地年降水量	41.594 0
产地土壤类型	37.386 3
无霜期	36.080 5
产地年日照时数	34.845 4
种质类型	32.535 6
最热月均温	31.087 6
产地气候类型	19.879 0
产区	18.982 3

影响最大的因素是产地年降水量,影响较大的因素有产地土壤类型,其次是无霜期和产地年日照时数,而生长状况对天麻素含量的影响相对较小。

3 讨论

中药的品质与环境具有密切的相关性,如虎杖在生长发育的过程中与生长环境密切相关^[30]。天麻在现代临床上具有重要的医用价值,需求量日益增加,天麻的人工种植技术也越来越受到人们的关注。早期关于天麻质量与种植环境之间的研究主要是通过生态背景调查的方法,如对天麻质量较好的产地云南昭通、四川平武等的生态环境进行调查,而本研究是利用机器学习方法探讨该问题。从天麻的种植环境出发,为了提高人工栽培天麻的质量,对天麻素含量与产区、生长状况、种质类型、气候类型、土壤类型、年均温、最冷月均温、最热月均温、年降水量、年日照时数和无霜期之间的关系进行探究。由于存在分类变量,且变量间具有多重共线性问题,故利用 Group-Lasso 算法筛选变量,为了找出影响天麻品质形成的关键因子,构建随机森林回归模型计算特征重要性。

利用 Group-Lasso 算法对产区、生长状况、种质类型、产地气候类型、产地土壤类型、产地年均温、最冷月均温、最热月均温、产地年降水量、产地年日照时数和无霜期 11 个变量进行筛选,最后保留了产区、生长状况、种质类型、产地气候类型、产地土壤类型、最热月均温、产地年降水量、产地年日照时数和无霜期 9 个变量,能够降低模型的复杂程度,提高效率。

以天麻素含量作为被解释变量,产区、生长状况、种质类型、产地气候类型、产地土壤类型、最热月均温、产地年降水量、产地年日照时数和无霜

期作为解释变量,分别建立多元线性回归和随机森林回归 2 种模型,通过结果的对比可知,随机森林回归模型有相对较低的误差和较高的预估精度,更适合用于对天麻种植环境的分析和天麻素含量的估算,同时随机森林回归模型具有易于实现、计算开销小,可以有效防止训练模型过拟合的优势。

根据特征重要性计算结果表明,对天麻素含量影响最大的环境因子是产地年降水量,天麻的生长需要产区的年降水量在 1000 mm 左右,不同时期水分需求也有所不同。4 月上旬块茎开始萌发,此时雨水满足土壤处于潮湿状态即可;7~9 月是天麻块茎生长的旺盛阶段,此时土壤需要充足的雨水供给水分。其次是产地土壤类型、无霜期和产地年日照时数,3 者特征重要性值接近。天麻和蜜环菌适合在比较疏松的沙质土壤中生长,若土壤黏重,则易积水,影响透气性,容易导致块茎死亡;若沙性过大,水分易流失,土壤缺水不利于天麻和蜜环菌的生长。无霜期主要与霜冻的时间有关,霜冻对光照产生影响。天麻具有避光性,块茎的生长在地下进行,但天麻的花茎具有趋光性,地上茎出土后,若遇到强烈的直射阳光会发生日灼病,进而导致植株死亡。此外,光照也间接影响地温和土壤水分,对蜜环菌和天麻的生长有一定影响。

环境因子是影响天麻药材产量和质量的重要因素。在人工种植天麻时,应注意环境因子的作用,充分考虑影响天麻品质形成的关键因子的特性,注意降水量、日照时数等环境因子的变化,同时选择具有合适土壤类型的区域作为天麻发展区进行人工种植。

利益冲突 所有作者均声明不存在利益冲突

参考文献

- [1] 杨世林, 兰进, 徐锦堂. 天麻的研究进展 [J]. 中草药, 2000, 31(1): 66-69.
- [2] 许廷生, 陆龙存, 黄子冬. 天麻有效成分的药理作用分析与临床应用研究进展 [J]. 中医临床研究, 2020, 12(21): 133-135.
- [3] 鞠桂春. 天麻及其制剂的药理作用和临床应用研究进展 [J]. 中国药业, 2008, 17(1): 64-66.
- [4] 于涵, 张俊, 陈碧清, 等. 天麻化学成分分类及其药理作用研究进展 [J]. 中草药, 2022, 53(17): 5553-5564.
- [5] 张伟, 宋启示. 贵州大方林下栽培天麻的化学成分研究 [J]. 中草药, 2010, 41(11): 1782-1785.
- [6] 王秋颖, 郭顺星. 天麻生长特性及其在栽培中的应用

- [J]. 中国中药杂志, 2001, 26(5): 353.
- [7] 胡国云, 姜月华, 赵海霞. 天麻素的含量测定、结构表征和药理活性研究进展 [J]. 当代化工, 2022, 51(1): 169-173.
- [8] 徐顶巧. 汉中地区天麻品质分析 [D]. 汉中: 陕西理工学院, 2014.
- [9] 陈希孺王松桂. 近代回归分析: 原理方法及应用 [M]. 合肥: 安徽教育出版社, 1987.
- [10] 何秀丽. 多元线性模型与岭回归分析 [D]. 武汉: 华中科技大学, 2005.
- [11] 赵东波. 线性回归模型中多重共线性问题的研究 [D]. 锦州: 渤海大学, 2017.
- [12] 方申奥. Lasso 类变量选择方法综述 [D]. 武汉: 武汉大学, 2018.
- [13] Friedman J, Hastie T, Tibshirani R. A note on the group lasso and a sparse group lasso [EB/OL]. arXiv [2010-01-05]. <https://arxiv.org/abs/1001.0736>.
- [14] Breiman L. Random forests [J]. *Mach Learn*, 2001, 45(1): 5-32.
- [15] Tibshirani R. Regression shrinkage and selection via the lasso [J]. *J Royal Statistical Society: Series B (Methodological)*, 1996, 58(1): 267-288.
- [16] Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression [J]. *J R Stat Soc Ser B*, 2008, 70(1): 53-71.
- [17] Zou H, Hastie T, Tibshirani R. On the “degrees of freedom” of the lasso [J]. *OALib J*, 2007, doi: 10.1214/009053607000000127.
- [18] Breiman L. Bagging predictors [J]. *Mach Lang*, 1996, 24(2): 123-140.
- [19] 董师师, 黄哲学. 随机森林理论浅析 [J]. 集成技术, 2013, 2(1): 1-7.
- [20] 曹正凤. 随机森林算法优化研究 [D]. 北京: 首都经济贸易大学, 2014.
- [21] 周碧乾. 天麻产地加工与品质特征的研究 [D]. 成都: 成都中医药大学, 2018.
- [22] 李慧, 钱润, 田娜, 等. 红天麻、乌天麻及其杂交天麻的 PCR 鉴别 [J]. 中国中药杂志, 2020, 45(15): 3666-3671.
- [23] 徐博, 吴翠, 李卓俊, 等. 天麻的资源分布及采后现状调研 [J]. 中国中医药信息杂志, 2021, 28(7): 11-16.
- [24] 李虎杰, 崔春龙, 李洪益. 四川平武野生天麻产区的土壤条件 [J]. 中国土壤与肥料, 2006(3): 59-61.
- [25] 李梁, 张艺, 成群芝. 中药天麻产区生态环境分析与评价 [J]. 中药研究与信息, 2004, 6(6): 14-16.
- [26] 石子为, 马聪吉, 康传志, 等. 基于空间分析的昭通天麻生态适宜性区划研究 [J]. 中国中药杂志, 2016, 41(17): 3155-3163.
- [27] Akaike H. A new look at the statistical model identification [J]. *IEEE Trans Autom Contr*, 1974, 19(6): 716-723.
- [28] 张倩. 基于随机森林回归模型的住房租金预测模型的研究 [D]. 长春: 东北师范大学, 2019.
- [29] 雷明. 机器学习: 原理、算法与应用 [J]. 自动化博览, 2020(3): 7.
- [30] 马云桐. 虎杖的资源、品质与药效的相关性研究 [D]. 成都: 成都中医药大学, 2006.

[责任编辑 潘明佳]