

基于粒子群-随机森林算法和气象数据的三七叶面积生长预测模型

范升旭¹, 杨春曦^{2*}, 杨启良³, 韩世昌²

1. 昆明理工大学化学工程学院, 云南 昆明 650500

2. 昆明理工大学机电工程学院, 云南 昆明 650500

3. 昆明理工大学农业与食品学院, 云南 昆明 650500

摘要: 目的 基于数据挖掘技术, 建立三七叶面积生长预测模型, 对于三七整个生长期的精准管理与决策提供参考。方法 基于粒子群-随机森林算法, 采用 2018、2019 年 4~10 月云南省红河自治州泸西县三七种植基地棚内气象因子数据以及三七叶面积生长数据作为训练集和测试集构建生长预测模型。结果 通过特征工程中皮尔森系数分析可知, 三七叶生长与土壤温度、上方水蒸气压和下方水蒸气压等气象因子呈正相关, 其中土壤温度正相关程度最大, 其皮尔森相关系数在 0.75~0.90; 下方土壤热通量与三七叶生长呈负相关, 其皮尔森相关系数为-0.4~-0.3; 通过粒子群优化随机森林算法训练的生长预测模型, 其均方根误差 (root mean square error, RMSE) 收敛时值为 0.021 82, 模型优化后的三七叶生长预测模型决定系数 R^2 达到 0.999 97。结论 通过多种算法对比实验结果表明, 粒子群-随机森林算法构建的三七叶面积生长预测模型具有较高的预测精度。该方法为三七叶的生长预测提供了新的研究思路。

关键词: 预测模型; 三七叶; 气象因子; 特征工程; 随机森林; 粒子群优化

中图分类号: R282 **文献标志码:** A **文章编号:** 0253-2670(2022)10-3103-08

DOI: 10.7501/j.issn.0253-2670.2022.10.021

Prediction model of *Panax notoginseng* leaf area growth based on particle swarm-optimization random forest algorithm and meteorological data

FAN Sheng-xu¹, YANG Chun-xi², YANG Qi-liang³, HAN Shi-chang²

1. Faculty of Chemical Engineering, Kunming University of Science and Technology, Kunming 650500, China

2. Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming 650500, China

3. Faculty of Agriculture and Food, Kunming University of Science and Technology, Kunming 650500, China

Abstract: Objective Based on data mining technology, the growth prediction model of Sanqi (*Panax notoginseng*) leaf area was established to provide reference for accurate management and decision-making of *P. notoginseng* during the whole growth period. **Methods** Based on the particle swarm-random forest algorithm, the meteorological factor data in the shed of *P. notoginseng* planting, Luxi County, Honghe Autonomous Prefecture, Yunnan Province from April to October 2018 and 2019 and leaf area growth data of *P. notoginseng* were used as the training set and test set of machine learning methods to build a growth prediction model. **Results** After doing the Pearson coefficient analysis of the characteristic engineering, the simulation results showed that the leaf growth of *P. notoginseng* was positively correlated with meteorological factors such as soil temperature, upper water vapor pressure and lower water vapor pressure. The positive correlation degree of soil temperature was the largest one with 0.75—0.90 Pearson correlation coefficient. On the contrary, the soil heat flux below was negative correlated with the leaf growth of *P. notoginseng*, and the Pearson correlation coefficient was -0.4—-0.3. For the prediction model trained by the proposed particle swarm-random forest algorithm, the convergence value of the root mean square error (RMSE) was 0.021 82, and the coefficient of determination R^2 of *P. notoginseng* leaf growth prediction model reaches 0.999 97 after model optimization. **Conclusion** The comparative results among different algorithms showed that the prediction model of *P. notoginseng* leaf area growth constructed by particle swarm optimization random forest algorithm has high prediction accuracy. Meanwhile, the combined algorithm proposed in this paper provides a new idea for the growth prediction research of stems and leaves of *P. notoginseng*.

收稿日期: 2022-02-20

基金项目: 国家自然科学基金项目 (62063011, 51979134, 51779113); 云南省科技厅科技计划项目 (202001AU070032)

作者简介: 范升旭, 硕士研究生, 研究方向为机器学习算法与大数据技术研究。E-mail: 2507197269@qq.com

*通信作者: 杨春曦, 教授, 硕士研究生导师, 主要从事于无线传感器网络以及大数据技术研究工作。E-mail: ycx@kmust.edu.cn

Key words: predication model; leaf of *Panax notoginseng* (Burk.) F. H. Chen; meteorological factors; characteristic engineering; random forest; particle swarm optimization

三七叶系五加科人参属植物三七 *Panax notoginseng* (Burk.) F. H. Chen 的干燥叶, 性温、味辛, 具有止血、消肿、定痛, 治吐血、外伤出血、痈肿毒疮等作用, 主要生长区域为云南和广西等地, 经过初步统计, 每年可采收三七叶约 2500 t^[1]。研究表明, 三七叶可药用也可食用, 且毒性低、不良反应小^[2-3]。因此, 研究三七叶生长期的气象数据特征, 获得三七叶生长与气象因子之间的关系, 对指导三七叶生长栽培以及整个三七生长期的精准管理与决策具有重要意义。

罗群等^[4]指出三七在不同生长时期所需的光照、温度、湿度、土壤等环境条件不同, 其中, 温度持续过高或者过低都会对三七生长造成危害。熊凯等^[5]提出影响三七生长的主要气象因子有太阳净辐射、棚内温湿度、土壤热通量、饱和水蒸气压、土壤温度、三七冠层上方温湿度、饱和水蒸气压。目前, 对影响三七生长因素的研究已经取得一定成果。罗美佳等^[6]提出光质对三七生长、光合特性及有效成分均有显著影响, 其中红光有利于三七的株高生长, 青、黄、紫、蓝光均有利于三七地下部分生物量的积累, 蓝、黄光有利于三七光合作用。张子龙等^[7]指出连作土壤对三七种子的萌发及幼苗的生长均表现明显的障碍效应, 化感(自毒)作用可能是造成三七连作障碍的原因之一。金航等^[8]发现云南文山所产三七无论在质量还是产量上均优于广西靖西的主要原因是云南文山具有全年日照充足、温度适宜、降雨适中、时间变化合理等有利的气象条件, 有利于三七的生长以及有效成分和干物质的积累。唐建楷等^[9]指出不同灌水频率和施肥量对三七形态指标、净光合速率、蒸腾作用、叶水势、水分利用效率、有效成分积累及发病率有明显影响。王朝梁等^[10]认为不同比例的自配肥可能增加三七的茎粗、叶面积和株高, 改善其生物学性状; 不同比例自配肥处理的三七出苗率均达到差异显著或极显著水平, 并且肥料中 N:P₂O₅:K₂O 比例以 1:1:2 处理时三七出苗率最好。上述研究主要侧重于探讨影响三七生长的因素, 鲜有综合考虑多个气象因子与三七叶生长之间的耦合关系, 也未将气象因子与三七叶生长的关系进行定量分析。

近年来, 随着大数据时代的快速发展, 基于数

据训练得到预测模型的机器学习方法被广泛应用于各个学科的研究中。其中, 随机森林^[11]是一种在土地利用^[12-13]、作物产量预测^[14-15]、山体滑坡空间预测^[16]、植株生长估测^[17-18]等不同领域都取得较好应用效果的机器学习算法。Kennedy 等^[19]设计了粒子群优化算法, 通过迭代遵循适应度函数规则找到最优解, 目前已被广泛应用于算法以及函数优化等领域。龙泉等^[20]提出基于粒子群优化 back propagation 神经网络处理风电机齿轮箱故障诊断, 提高了神经网络的训练效率, 加快了网络的收敛速度。穆朝絮等^[21]提出基于粒子群优化的非线性系统最小二乘支持向量机预测控制方法, 为非线性系统在数学模型未知的情况下设计出有效的预测控制器, 且具有良好的自适应能力和鲁棒性。

本研究通过机器学习方法分析三七叶生长期的气象因子的数据特征, 进而获得主要影响三七叶生长的气象因子特征, 气象因子是具有连续性、非线性的变量, 随机森林算法在处理这类数据上效果突出, 性能稳定且具有良好的抗干扰能力^[22-25]。利用粒子群算法优化随机森林算法^[26-28]构建三七叶生长期的主要气象因子数据与三七叶面积生长数据的生长预测模型, 为提高三七叶产量的环境调控和气象预警提供理论依据和技术支持。通过集成多学科知识与技术交叉融合, 揭示三七生长习性规律, 监控三七生长与环境变化全过程, 为中药资源新兴交叉学科的建立提供可能性方向, 也为三七生长环境精细化管理奠定良好的基础^[29]。

1 材料与方法

1.1 研究区域概况

试验于 2018 年 3 月至 2019 年 11 月在云南省红河哈尼族彝族自治州泸西县昆明理工大学三七控水减排提质增效关键技术与示范基地(103°57'E, 24°26'N)进行。试验基地平均海拔 1842 m, 属北亚热带低纬高原山地季风气候。试验期间, 基地最高和最低气温分别为 31 ℃和 1 ℃, 平均相对湿度 58%。供试土壤为微酸性红壤土, pH 值 6.0 左右。试验基地为塑料大棚, 采用塑料膜挡雨, 顶部设置遮阳网, 透光率为 8.3%。三七田挖沟起垄, 垄沟深 0.40 m, 每垄间距为 0.50 m, 每垄地长为 18 m, 宽为 1.90 m, 垄面均匀覆盖 5 mm 干燥的松针, 两侧

埋设有黑色不透水薄膜。供试作物为 2 年生三七，种植间距为 15 cm，种植密度为 4.40×10^5 株/hm²。

1.2 数据来源

三七生长高峰期在 4~10 月，因此选用 2018 年和 2019 年的 4~10 月三七叶面积的生长数据、气象因子数据作为总样本。气象因子样本数据每小时采集 1 次，样本数据来源于三七种植基地无线气象因子远程监测系统，其组成如图 1 所示。

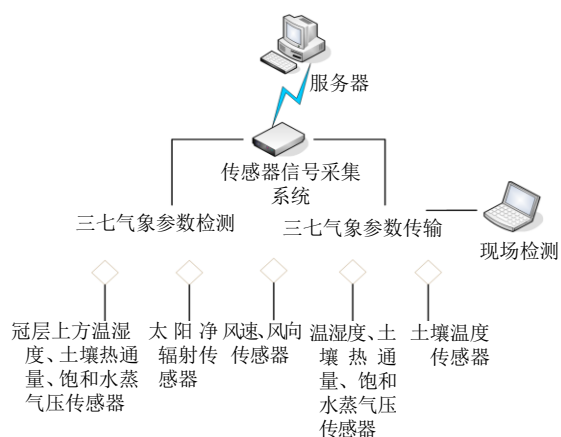
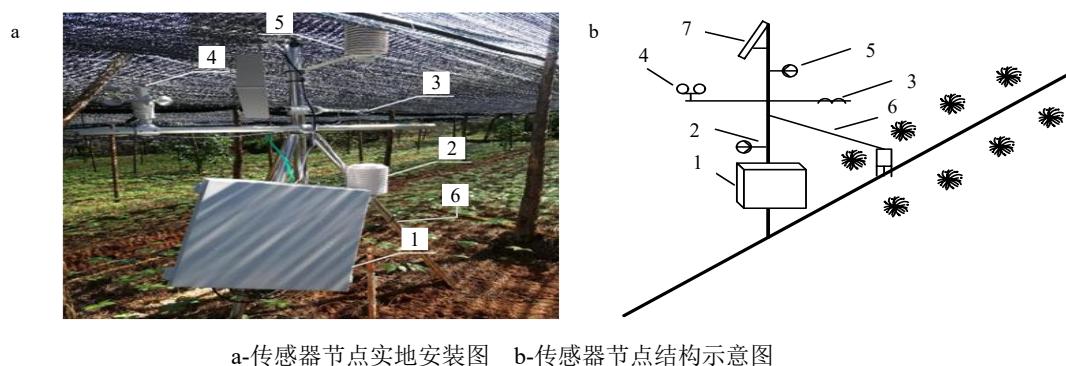


图 1 三七种植基地无线气象因子远程监测系统

Fig. 1 Remote monitoring system of wireless meteorological factors in *P. notoginseng* planting base

采集的气象因子数据包括太阳净辐射 (W/m^2)、棚内上下方温度 ($^{\circ}C$)、棚内上下方湿度 (%)、棚内上下方土壤热通量 (W/m^2)、棚内上下方饱和水蒸气压 (kPa)、土壤温度 ($^{\circ}C$)、三七冠层上下方温度 ($^{\circ}C$)、三七冠层上下方湿度 (%)、三七冠层上下方土壤热通量 (W/m^2) 和三七冠层上下方饱和水蒸气压 (kPa)。分别在施肥水平 F1: 60 kg/hm²、F2: 90 kg/hm²、F3: 120 kg/hm²、F4: 150 kg/hm²；灌水水平 W1: 100 m³/hm²、W2: 150 m³/hm²、W3: 200 m³/hm² 组合搭配后共计 12 个区域进行三七叶面积生长数据的采集。每月在各个区域贴有固定标签的三七植株里取 3 片长势相近的叶片进行叶面积均值计算统计。其中通过纸样称重法，对三七叶面积进行测量，新取的三七叶片平铺于标准网格纸上，在纸上描绘出叶片轮廓后剪下，称质量 (m_1)，面积为 A_1 (未知)；取标准网格纸，称质量 (m_2)，面积为 A_2 ，则叶片面积 $A_1 = m_1 \times A_2 / m_2$ ，其余三七叶面积按此方法计算，从每年 4 月 4 日开始，至 10 月 4 日测定结束。共计 18 个气象因子特征以及三七叶面积实测数据 14 915 个，样本组数为 785 组，随机分成 2 部分，75% ($n=588$) 作为训练样本建立模型，25% ($n=197$) 作为测试样本评价模型。传感器节点和气象因子采集系统布置图见图 2，其中，图 2-



a-installation drawing of sensor nodes b-schematic diagram of sensor nodes structure

图 2 传感器节点和气象因子采集系统布置图

Fig. 2 Working diagram of sensor nodes and meteorological factor acquisition system

a 为传感器节点实地安装图，图 2-b 为传感器节点结构示意图，节点 1 是信号采集系统；节点 2 是冠层上方温湿度、土壤热通量、饱和水蒸气压传感器；节点 3 是太阳净辐射传感器；节点 4 是风速、风向传感器；节点 5 是棚内温湿度、土壤热通量、饱和水蒸气压传感器；节点 6 是土壤温度传感器，埋于试验地地下；节点 7 是太阳能电源模块。

1.3 数据预处理

由于采集系统中传感器在长时间运行过程中会出现数据缺失及其他异常数据，为保证特征在时间维度上一致，需要进行数据清洗。同时，为了保证模型能够更好地反映三七叶生长高峰期与气象因子的关系，本研究需要将每年 4~10 月份采集的气象因子数据提取出来进行模型训练。

不同气象因子往往具有不同的量纲和量纲单位, 为了避免研究中不同气象因子对三七叶生长的影响权重不同, 从而导致模型评估不精确, 需对数据进行标准化处理, 原始数据经过数据标准化处理后, 各指标处于同一数量级, 可以进行综合对比评价。Min-max 标准化也称为离差标准化, 是对原始数据的线性变换, 使结果映射到 $[0, 1]$, 转换如公式 (1) 所示。

$$A^* = \frac{A - \min}{\max - \min} \quad (1)$$

max 为特征数据中最大值, min 为特征数据中最小值, A 为归一化前数据, A^* 为归一化后数据

1.4 模型构建流程

1.4.1 皮尔森相关系数分析 皮尔森相关系数是一种线性相关系数, 也是最常用的 1 种相关系数^[30]。首先通过皮尔森系数判断各气象因子与三七叶生长之间的相关性。皮尔森相关系数 $r(x, y)$ 的取值范围为 $[-1, 1]$, 若取值在 $(0, 1]$, 取值越大表示正相关性越强; 若取值在 $[-1, 0)$, 取值越小表示负相关性越大; 取 0 值表示无相关性。皮尔森系数评价指标如公式 (2) 所示。

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \quad (2)$$

$\text{Cov}(X, Y)$ 表示 X 与 Y 的协方差, $\text{Var}(X)$ 为 X 的方差, $\text{Var}(Y)$ 为 Y 的方差

其次, 去除掉气象因子与三七叶生长相关性较弱的特征, 提取与三七叶生长有相关特征的样本导入到预测模型训练, 达到降低维度, 加快模型训练速度。

1.4.2 随机森林回归预测算法 随机森林是由 Breiman^[11]提出的一种分类算法, 其通过自助法 (bootstrap) 重采样技术, 从原始训练样本中有放回地重复随机抽取 n 个样本, 生成新的训练样本集合训练决策树, 然后按以上步骤生成 m 棵决策树组成随机森林, 新数据的分类结果按树的投票数量决定。随机森林算法的实质是对决策树算法的一种改进, 将多个决策树合并在一起, 每棵决策树的建立依赖于独立抽取的样本, 分类能力较弱。但产生大量的决策树后, 待测试的样本数据可以通过每一棵树的分类结果, 经统计后选择最可能的分类, 由于采用多棵决策树并行的训练模型, 因此每棵决策树都可以选择部分样本及特征, 能够在一定程度上避免过拟合; 同时, 每棵决策树随机选择样本以及特征, 使得随机森林算法具有很好的抗噪能力, 性能稳定。

1.4.3 粒子群优化随机森林回归预测算法 粒子群

优化算法^[16]是通过无质量的粒子来模拟鸟群里的鸟捕食行为设计的。其中每个粒子都具有 2 个属性, 即速度和位置。通过粒子的不断搜索, 得到最优解, 其中个体搜索最优解为 $Pbest$, 群体最优解为 $Gbest$, 粒子在迭代寻优过程中, 通过 $Pbest$ 和 $Gbest$ 不断地更新自身的速度和位置, 迭代过程如公式 (3) 所示。

$$V_{\lambda q}^k = \alpha V_{\lambda q}^{k-1} + c_1 r_1 (Pbest_{\lambda q} - X_{\lambda q}^{k-1}) + c_2 r_2 (Gbest_q - X_{\lambda q}^{k-1})$$

$$X_{\lambda q}^k = X_{\lambda q}^{k-1} + V_{\lambda q}^{k-1} \quad (3)$$

其中, $V_{\lambda q}^k$ 为第 k 次迭代下粒子 λ 飞行速度矢量的第 q 维分量, $X_{\lambda q}^k$ 为第 k 次迭代下粒子 λ 位置矢量的第 q 维分量, α 为粒子位置惯性权重, q 为粒子维数, c_1 、 c_2 为学习因子, 取值为 $c_1 = c_2 \in [0, 4]$, r_1 、 r_2 为 2 个取值在 $[0, 1]$ 的随机数。为了增加搜索随机性和有效性, 一般粒子位置取值在 $[X_{\max}, X_{\min}]$, 速度取值在 $[V_{\max}, V_{\min}]$, 其中 $V_{\min}$ 为粒子最小搜寻速度, $V_{\max}$ 为粒子最大搜寻速度, $X_{\min}$ 为粒子最小搜寻位置, $X_{\max}$ 为粒子最大搜寻位置, 保证粒子不会盲目进行搜索。

通过已完成初始化的粒子群算法对随机森林算法中初始参数进行最优值搜寻, 可以避免算法训练模型中的初始参数采用人工经验值设置导致的参数不确定性。粒子群算法智能选择合适的决策树数目 ($n_estimators$) 和决策树最大深度 (max_depth), 可以保证在有效增强模型的预测能力的情况下, 增强训练模型训练的泛化能力; 再进一步, 选择决策树数目和决策树最大深度 2 个参数作为粒子维数, 可以有效减少粒子群算法搜索运行时间, 以及一定程度上增强搜寻效果。由于构建的算法搜索区域为二维空间, 在确保能够搜索到能优化训练模型所需参数值的同时, 又缩短了算法搜索运行时间, 从而保证算法效率。同时, 选取随机森林算法训练模型的均方根误差 (root mean square error, RMSE) 作为粒子群算法的适应度函数, RMSE 越小, 则表示模型的预测性能越精确。利用粒子群算法对适应度函数 RMSE 进行最小值寻优, 便可以确定随机森林回归模型达到搜寻条件时最佳性能所对应的参数值。模型构建流程图见图 3。

本模型构建的实验平台单机处理器为英特尔酷睿 i5-4590, 频率为 CPU 3.3 GHz, 运行内存 12 GB, 操作系统为 64 位的 Windows 7 旗舰版, 程序编译语言为 Python 3.7, 编译器为 PyCharm。

1.5 模型评价指标

采用 4 个指标作为评价模型拟合程度的优劣, 即决定系数 (coefficient of determination, R^2)、均方

误差 (mean square error, MSE)、平均绝对误差 (mean absolute error, MAE)、RMSE。

$$R^2(X, h) = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (h(x_i) - \bar{y})^2} \quad (4)$$

$$\text{RMSE}(X, h) = \sqrt{\frac{1}{N} \sum_{i=1}^N (h(x_i) - \hat{y}_i)^2} \quad (5)$$

$$\text{MSE}(X, h) = \frac{1}{N} \sum_{i=1}^N (h(x_i) - \hat{y}_i)^2 \quad (6)$$

$$\text{MAE}(X, h) = \frac{1}{N} \sum_{i=1}^N |h(x_i) - \hat{y}_i| \quad (7)$$

式中 $h(x_i)$ 表示第 i 个三七叶面积数据的预测值, $\hat{y}_i$ 表示第 i 个三七叶面积生长数据的真实值, $\bar{y}$ 表示三七叶面积生长数据均值, 样本个数为 N 个。其中决定系数 R^2 取值范围在 $[0, 1]$, 数值越接近 1 表示模型拟合效果越好; MSE 是测试集实际值与预测值之差的平方和, RMSE 为均方误差平方根, MAE 是预测值与真实值之差的绝对平均值。MSE、MAE 以及 RMSE 的数值越趋近于 0 表示模型预测精度越高。

2 结果与分析

2.1 气象因子数据与三七叶面积生长数据间的皮尔森相关系数

不同气象因子数据与三七叶面积生长数据之间的皮尔森相关系数见表 1。

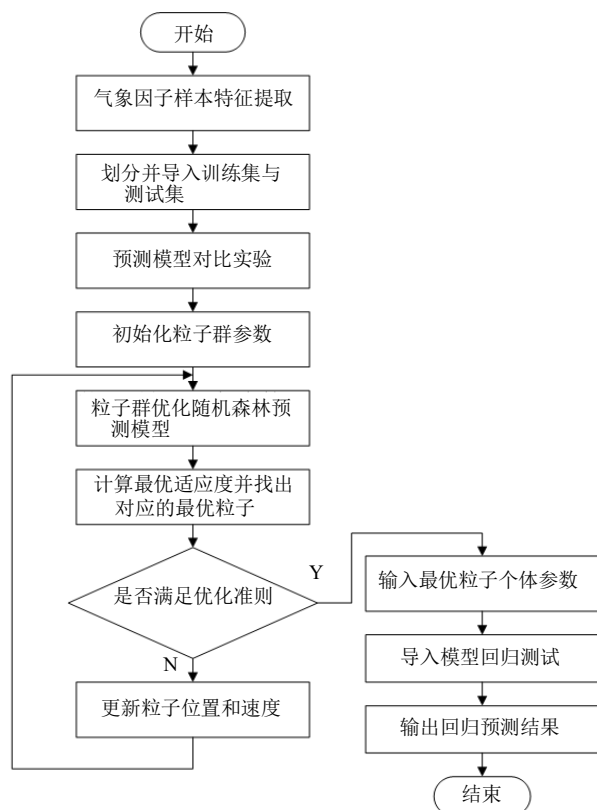


图 3 模型构建流程

Fig. 3 Flow chart of model construction

表 1 三七叶面积生长与各气象因子间皮尔森相关系数

Table 1 Pearson correlation coefficient between leaf area growth and meteorological factors of *P. notoginseng*

序号	特征变量	皮尔森相关系数
1	太阳净辐射 (net solar radiation, $\text{W} \cdot \text{m}^{-2}$)	-0.298 924 920 8
2	上方温度 (temperature above, $^{\circ}\text{C}$)	-0.185 395 671 1
3	上方湿度 (humidity above, %)	0.521 236 432 5
4	上方土壤热通量 (soil heat flux above, $\text{W} \cdot \text{m}^{-2}$)	-0.335 497 475 7
5	土壤温度 (soil temperature, $^{\circ}\text{C}$)	0.885 152 925 6
6	下方温度 (temperature below, $^{\circ}\text{C}$)	-0.192 979 257 9
7	下方湿度 (humidity below, %)	0.545 357 700 0
8	下方土壤热通量 (soil heat flux below, $\text{W} \cdot \text{m}^{-2}$)	-0.352 431 631 9
9	上方潜热通量 (latent heat flux above, $\text{W} \cdot \text{m}^{-2}$)	-0.123 115 322 8
10	下方潜热通量 (latent heat flux below, $\text{W} \cdot \text{m}^{-2}$)	-0.109 738 246 0
11	上方显热通量 (sensible heat flux above, $\text{W} \cdot \text{m}^{-2}$)	0.296 560 680 5
12	下方显热通量 (sensible heat flux below, $\text{W} \cdot \text{m}^{-2}$)	0.036 766 140 5
13	上层饱和水蒸气压 (upper saturated water vapor pressure, kPa)	-0.221 604 227 4
14	下层饱和水蒸气压 (lower saturated water vapor pressure, kPa)	-0.229 728 512 2
15	上方水蒸气压 (upper water vapor pressure, kPa)	0.799 377 189 7
16	下方水蒸气压 (lower water vapor pressure, kPa)	0.797 353 375 5

由表 1 可知, 土壤温度、上方水蒸气压、下方水蒸气压、下方土壤热通量 4 个气象因子, 对三七叶生长的影响较为显著。其中, 土壤温度、上方水蒸气压、下方水蒸气压与三七叶片生长呈正相关, 其皮尔森相关系数在 0.75~0.90; 下方土壤热通量与三七叶片生长呈负相关, 其皮尔森相关系数为 -0.4~-0.3。

本研究所采用的气象因子均是气温相关的统计量, 从已有的结果可以知道, 土壤温度对三七叶生长影响较为显著, 这与罗群等^[4]、崔秀明等^[31]研究的实验结果相吻合。土壤温度成为三七叶片生长过程的主要影响气象因子, 这可能归咎于土壤温度的变化会影响土壤呼吸, 土壤呼吸又会影响三七根系的生理作用和营养物质传输^[32], 进而影响三七叶片生长情况。

同时, 由于所采用的气象因子均是气温及气温基础上的统计量, 为了探讨其他气象因子与土壤温度之间的关系, 分别计算不同气象因子与土壤温度之间的皮尔森相关系数, 见表 2。由表 2 可知, 上方水蒸气压和下方水蒸气压 2 个气象因子与土壤温度呈正相关关系较为显著, 其皮尔森相关系数在 0.8~0.9。综合表 1、2 可知与土壤温度呈正相关的气象因子对三七叶片生长具有明显影响。

表 2 土壤温度与各气象因子间皮尔森相关系数

Table 2 Pearson correlation coefficient between soil temperature and meteorological factors

序号	特征变量	皮尔森相关系数
1	太阳净辐射	0.003 397 036 1
2	上方温度	0.304 849 554 2
3	上方湿度	0.275 722 660 2
4	上方土壤热通量	0.100 898 898 8
5	下方温度	0.321 069 998 4
6	下方湿度	0.288 651 785 5
7	下方土壤热通量	0.168 593 904 9
8	上方潜热通量	-0.004 345 964 8
9	下方潜热通量	-0.014 487 835 2
10	上方显热通量	-0.021 597 128 9
11	下方显热通量	0.004 651 436 2
12	上层饱和水蒸气压	0.208 791 311 7
13	下层饱和水蒸气压	0.222 562 127 2
14	上方水蒸气压	0.827 128 041 3
15	下方水蒸气压	0.831 953 445 5

2.2 三七叶生长预测模型构建

经过皮尔森系数分析后, 为了降低特征维度, 加快训练速度, 去除掉气象因子与三七叶片生长相关性较弱的特征变量, 如上方潜热通量、下方潜热通量、上方显热通量以及下方显热通量, 并提取与三七叶片生长有相关特征样本导入至预测模型进行训练, 预测模型对比实验中的参数设置如表 3 所示。表 3 是未进行智能算法寻优的支持向量机^[33]、k 近邻^[34]、随机森林^[11]模型初始参数, 均采用网格搜索确定。

2.3 三七叶生长预测模型评价

测试样本集测试结果表明, 不同模型预测三七叶面积生长数据的能力不同, 结果如图 4 所示。

表 3 预测模型参数设置

Table 3 Prediction model setting parameters

预测模型	参数
支持向量机	C=105, gamma=0.015
k 近邻	n_neighbors=5
随机森林	n_estimators=5, max_depth=3

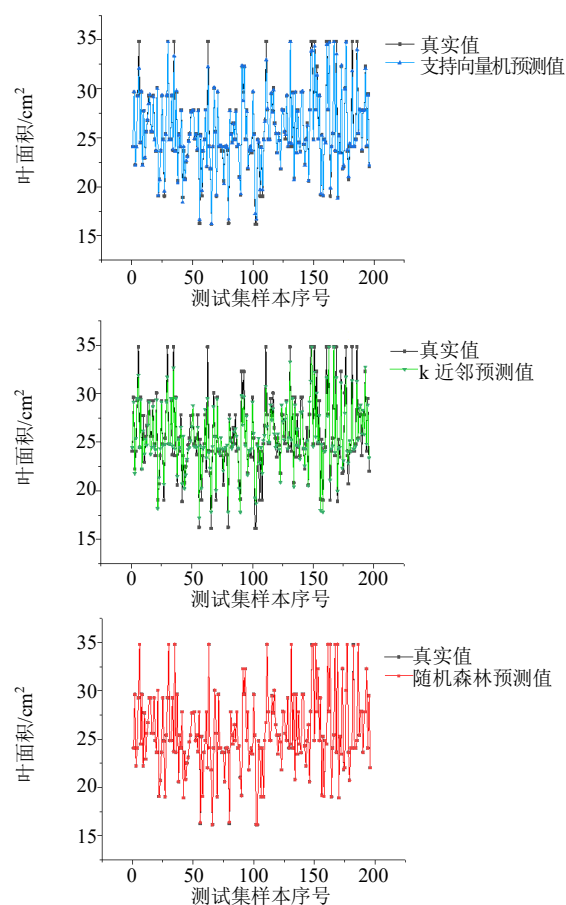


图 4 不同预测模型预测结果对比图

Fig. 4 Comparison of prediction results of different prediction models

由图 4 可知, k 近邻算法的预测值与测试实际数据相差较大, 支持向量机、随机森林预测模型的预测结果基本符合测试集数据真实分布变化趋势, 但模型之间都存在不同程度的偏差波动。如表 4 所示, 粒子群优化随机森林算法的 MSE、MAE、RMSE 值均小于其他预测模型, 且决定系

数 R^2 值大于其他预测模型, 说明该算法具有更高预测精度。

采取随机森林算法训练模型的 RMSE 作为粒子群算法的适应度函数, RMSE 越小, 则表示模型的预测性能越强; 迭代 N 次 ($N=150$) 至 RMSE 收敛到最小, 迭代过程如图 5 所示。

表 4 不同模型预测能力评价指标表

Table 4 Prediction ability evaluation index of different models

预测模型	均方误差	平均绝对误差	均方根误差	R^2
支持向量机	0.218 42	0.207 86	0.467 35	0.987 89
k 近邻	1.928 72	1.018 74	1.388 78	0.893 08
随机森林	0.124 13	0.285 17	0.352 32	0.993 12
粒子群优化随机森林	0.000 48	0.007 07	0.021 82	0.999 97

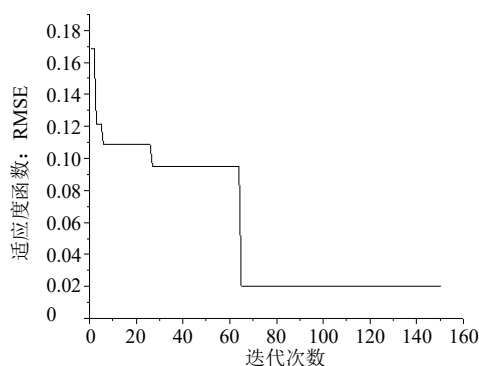


图 5 适应度函数与迭代次数间的关系

Fig. 5 Relationship between fitness function and number of iterations

皮尔 (R. Pearl) 生长曲线^[35]是描述生物生长与时间关系的模型, 补充以皮尔生长曲线在施肥水平 F1 (60 kg/hm^2) 和灌水水平 W1 ($100 \text{ m}^3/\text{hm}^2$) 拟合填充三七叶面积数据, 并以构建的预测随机森林模型进行预测, 结果如图 6 所示。随机森林模型对皮尔生长曲线模型处理下的三七叶面积进行预测, RMSE 为 0.182 71, 其中拟合曲线结果出现区域折线式上升, 是预测模型训练中选取的三七叶面积数据采样周期间隔导致, 拟合结果说明该模型同样符合常规植株生长预测趋势。

3 讨论

通过皮尔森系数筛选主要影响三七叶生长的气象因子特征, 筛选后的气象因子数据以及对在不同灌水以及施肥水平下的三七叶面积生长数据导入到粒子群-随机森林参数算法预测模型, 该模型的决定系数 R^2 、MSE、MAE 和 RMSE 均低于支持向量机

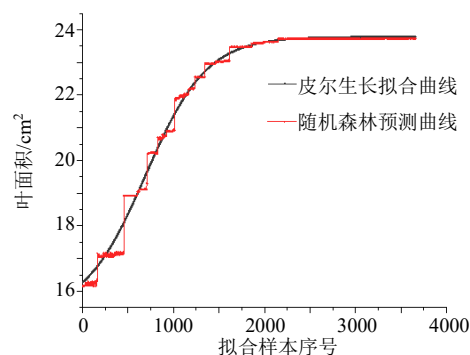


图 6 皮尔生长曲线与随机森林预测图

Fig. 6 Peel growth curve and random forest prediction diagram

以及 k 近邻预测模型训练结果, 得出粒子群-随机森林模型具有更高预测精度。

三七叶片生长与土壤温度、上方水蒸气压和下方水蒸气压等气象因子密切相关, 其中土壤温度正相关程度最大。分析与土壤温度强相关性的气象因子, 得到上方水蒸气压和下方水蒸气与三七叶片生长呈强正相关性, 而下方土壤热通量与土壤温度呈强负相关性。

本研究所提出的方法能在未来使种植人员得到较准确的三七叶片长势预测, 为后期三七整个生长期的精准管理与种植决策提供技术支撑, 也为中药资源新兴交叉学科的建立提供了可能性方向。以地区经济药用植物三七为研究对象, 揭示其生长习性规律, 服务于三七生产全过程, 推动三七规范化生产与品质提升、促进中药资源全产业链的提质增效和绿色发展, 从而降低三七种植风险, 提升三七产业经济效益。

利益冲突 所有作者均声明不存在利益冲突

参考文献

- [1] 周家明, 崔秀明, 曾鸿超, 等. 三七茎叶的综合开发利用 [J]. 现代中药研究与实践, 2009, 23(3): 32-34.
- [2] 雷伟亚, 史栓桃, 余思畅, 等. 三七叶总皂甙的毒性研究 [J]. 云南医药, 1984, 5(4): 241-244.
- [3] 秦光和, 景箫, 王伟, 等. 三七茎叶毒理学安全性评价 [J]. 毒理学杂志, 2020, 34(2): 179-181.
- [4] 罗群, 游春梅, 官会林. 环境因素对三七生长影响的分析 [J]. 中国西部科技, 2010, 9(9): 7-8.
- [5] 熊凯, 杨启良, 杨春曦, 等. 基于病害高发期气象因子的三七病害发生率预测 [J]. 农业工程学报, 2020, 36(24): 170-176.
- [6] 罗美佳, 夏鹏国, 齐志鸿, 等. 光质对三七生长、光合特性及有效成分积累的影响 [J]. 中国中药杂志, 2014, 39(4): 610-613.
- [7] 张子龙, 王文全, 杨建忠, 等. 三七连作土壤对其种子萌发及幼苗生长的影响 [J]. 土壤, 2010, 42(6): 1009-1014.
- [8] 金航, 崔秀明, 朱艳, 等. 气象条件对三七药材道地性的影响 [J]. 西南农业学报, 2005, 18(6): 825-828.
- [9] 唐建楷, 韩焕豪, 刘冰, 等. 灌水频率和施肥量对三七有效成分积累与发病率的影响 [J]. 农业工程学报, 2020, 36(24): 55-63.
- [10] 王朝梁, 陈中坚, 孙玉琴, 等. 不同氮磷钾配比施肥对三七生长及产量的影响 [J]. 现代中药研究与实践, 2007, 21(1): 5-7.
- [11] Breiman L. Random forest [J]. *Machine Learning*, 2001, 45: 5-32.
- [12] 马玥, 姜琦刚, 孟治国, 等. 基于随机森林算法的农耕地土地利用分类研究 [J]. 农业机械学报, 2016, 47(1): 297-303.
- [13] 刘舒, 姜琦刚, 马玥, 等. 基于多目标遗传随机森林特征选择的面向对象湿地分类 [J]. 农业机械学报, 2017, 48(1): 119-127.
- [14] 王鹏新, 齐璇, 李俐, 等. 基于随机森林回归的玉米单产估测 [J]. 农业机械学报, 2019, 50(7): 237-245.
- [15] 程千, 徐洪刚, 曹引波, 等. 基于无人机多时相植被指数的冬小麦产量估测 [J]. 农业机械学报, 2021, 52(3): 160-167.
- [16] 余坤勇, 姚雄, 邱祈荣, 等. 基于随机森林模型的山体滑坡空间预测研究 [J]. 农业机械学报, 2016, 47(10): 338-345.
- [17] 李德, 陈文涛, 乐章燕, 等. 基于随机森林算法和气象因子的砀山酥梨始花期预报 [J]. 农业工程学报, 2020, 36(12): 143-151.
- [18] 冯海宽, 杨福芹, 杨贵军, 等. 基于特征光谱参数的苹果叶片叶绿素含量估算 [J]. 农业工程学报, 2018, 34(6): 182-188.
- [19] Kennedy J, Eberhart R. Particle swarm optimization [A] // Proceedings of ICNN'95-International Conference on Neural Networks [C]. Perth: IEEE, 1995: 1942-1948.
- [20] 龙泉, 刘永前, 杨勇平. 基于粒子群优化 BP 神经网络的风电机组齿轮箱故障诊断方法 [J]. 太阳能学报, 2012, 33(1): 120-125.
- [21] 穆朝絮, 张瑞民, 孙长银. 基于粒子群优化的非线性系统最小二乘支持向量机预测控制方法 [J]. 控制理论与应用, 2010, 27(2): 164-168.
- [22] Scornet E. On the asymptotics of random forests [J]. *J Multivar Anal*, 2016, 146: 72-83.
- [23] Delgado M F, Cernadas E, Barro S, et al. Do we need hundreds of classifiers to solve real world classification problems? [J]. *J Mach Learn Res*, 2014, 15(1): 3133-3181.
- [24] Cutler D R, Edwards T C Jr, Beard K H, et al. Random forests for classification in ecology [J]. *Ecology*, 2007, 88(11): 2783-2792.
- [25] 韩敏, 朱新荣. 不平衡数据分类的混合算法 [J]. 控制理论与应用, 2011, 28(10): 1485-1489.
- [26] 王杰, 程学新, 彭金柱. 一种基于粒子群算法优化的加权随机森林模型 [J]. 郑州大学学报: 理学版, 2018, 50(1): 72-76.
- [27] 王小杨, 罗多, 孙韵琳, 等. 基于 ABC-SVM 和 PSO-RF 的光伏微电网日发电功率组合预测方法研究 [J]. 太阳能学报, 2020, 41(3): 177-183.
- [28] 王东风, 孟丽. 粒子群优化算法的性能分析和参数选择 [J]. 自动化学报, 2016, 42(10): 1552-1561.
- [29] 段金彪, 宿树兰, 严辉, 等. 2016—2020 年我国中药资源学学科建设及科学研究进展与展望 [J]. 中草药, 2021, 52(17): 5151-5165.
- [30] 马瑞, 周谢, 彭舟, 等. 考虑气温因素的负荷特性统计指标关联特征数据挖掘 [J]. 中国电机工程学报, 2015, 35(1): 43-51.
- [31] 崔秀明. 三七 GAP 研究与实践 [M]. 昆明: 云南科技出版社, 2003: 32-35.
- [32] 李玉强, 赵哈林, 赵学勇, 等. 土壤温度和水分对不同类型沙丘土壤呼吸的影响 [J]. 干旱区资源与环境, 2006, 20(3): 154-158.
- [33] Bouboulis P, Theodoridis S, Mavroforakis C, et al. Complex support vector machines for regression and quaternary classification [J]. *IEEE Trans Neural Netw Learn Syst*, 2015, 26(6): 1260-1274.
- [34] Weinberger K, Saul L. Distance metric learning for large margin nearest neighbor classification [J]. *J Mach Learn Res*, 2009, 10: 207-244.
- [35] 吴强, PENG Yuanying, 马恒运, 等. 森林生态系统服务价值及其补偿校准: 以马尾松林为例 [J]. 生态学报, 2019, 39(1): 117-130.

[责任编辑 潘明佳]