杨明欣, 高鹏, 陈文彬, 等. 基于机器学习的油青菜心水分胁迫研究 [J]. 华南农业大学学报, 2021, 42(5): 117-126.
YANG Mingxin, GAO Peng, CHEN Wenbin, et al. Research of *Brassica chinensis* var. *parachinensis* under water stress based on machine learning[J]. Journal of South China Agricultural University, 2021, 42(5): 117-126.

基于机器学习的油青菜心水分胁迫研究

杨明欣[1]，高　鹏[1]，陈文彬[1]，周　平[1]，孙道宗[1,2]，谢家兴[1,2]，陆健强[1,2]，王卫星[1,2]
(1 华南农业大学 电子工程学院/人工智能学院，广东 广州 510642；
2 广东省农情信息监测工程技术研究中心，广东 广州 510642)

摘要:【目的】以油青菜心 *Brassica chinensis* var. *parachinensis* 为试验对象，基于冠层温度研究其生长过程中的水分胁迫变化规律，并利用机器学习方法，以水分胁迫指数 (Crop water stress index, CWSI) 和光合有效辐射预测光合作用速率。【方法】试验期间，在营养生长阶段 (V 期) 和生殖生长阶段 (R 期) 对油青菜心进行不同田间持水量处理，采集冠层温度、空气温湿度数据，建立无蒸腾作用基线 (上限方程)、无水分胁迫基线 (下限方程)，通过经验公式计算 CWSI。利用基于密度的空间聚类方法和空气温度研究油青菜心的冠气温差上限分布情况，选取固定值作为上限；以 CWSI 经验公式为基础，使用不同温度定值的无蒸腾作用基线计算 CWSI，验证聚类效果。为更简便获取光合作用速率，使用 4 种机器学习方法：最邻近节点算法 (*k*-Nearest neighbor，KNN)、支持向量回归 (Support vector regression，SVR)、极端梯度提升法 (Extreme gradient boosting，XGBoost)、随机森林 (Random forest，RF) 进行预测，并对比预测效果。【结果】在不同田间持水量处理下，CWSI 能较好地监测油青菜心水分胁迫状况。通过聚类分析，将 V 期和 R 期冠气温差上限分类到 2 个簇中，得到簇心分别为 3.4 和 4.2 ℃，与 CWSI 经验公式计算值显著相关，表明使用固定值作为油青菜心冠气温差上限值具有可行性。KNN、SVM、XGBoost 和 RF 预测模型均取得较好效果，相关系数分别为 0.873、0.877、0.887 和 0.863。【结论】机器学习方法可用于油青菜心光合作用速率的预测，可以避免使用大型笨重仪器，降低对油青菜心叶片的损伤，减少测量时间。

关键词: 油青菜心；水分胁迫指数；光合作用速率；机器学习
中图分类号: S27　　**文献标志码:** A　　**文章编号:** 1001-411X(2021)05-0117-10

Research of *Brassica chinensis* var. *parachinensis* under water stress based on machine learning

YANG Mingxin[1], GAO Peng[1], CHEN Wenbin[1], ZHOU Ping[1], SUN Daozong[1,2],
XIE Jiaxing[1,2], LU Jianqiang[1,2], WANG Weixing[1,2]
(1 College of Electronic Engineering/College of Artificial Intelligence, South China Agricultural University, Guangzhou 510642, China; 2 Guangdong Engineering Research Center for Monitoring Agricultural Information, Guangzhou 510642, China)

Abstract: 【Objective】*Brassica chinensis* var. *parachinensis* was used as the experimental object to study the change rule of water stress during growth process based on canopy temperature, and the machine learning method was used for predicting the photosynthetic rate based on crop water stress index (CWSI) and photosynthetically active radiation. 【Method】During the test, the experiment adopted different field capacities

收稿日期: 2021–01–12　　**网络首发时间:** 2021–06–18 15:04:58
网络首发地址: https://kns.cnki.net/kcms/detail/44.1110.s.20210618.0936.002.html
作者简介: 杨明欣 (1995—)，女，硕士研究生，E-mail: 736517614@qq.com；通信作者：王卫星 (1963—)，男，教授，博士，E-mail: weixing@scau.edu.cn
基金项目: 广东省重点领域研发计划 (2019B020214003)

for *B. chinensis* var. *parachinensis* at the vegetative growth stage (V stage) and reproductive growth stage (R stage), collected the canopy temperature, air temperature and humidity data, established non-transpiration-baseline (upper limit equation), non-water-stress-baseline (lower limit equation), and calculated CWSI by empirical formulation. Cluster method of density-based spatial clustering of application with noise and air temperature were used to study the upper limit distribution of canopy temperature minus air temperature of *B. chinensis* var. *parachinensis*, and the fixed values were selected as the upper limit. Based on the CWSI empirical formulation, CWSI was calculated using the non-transpiration-baselines with different temperature fixed values to verify the clustering effect. In order to obtain the photosynthetic rate more easily, four machine learning methods of *k*-nearest neighbor (KNN), support vector regression (SVR), extreme gradient boosting (XGBoost) and random forest (RF) were used for prediction, and the prediction effects were compared.【Result】Under different field capacities, CWSI could better monitor the water stress status of *B. chinensis* var. *parachinensis*. Through cluster analysis, the upper limit of canopy temperature minus air temperature at V stage and R stage was classified into two clusters, and the cluster centers were 3.4 and 4.2 ℃, respectively, which were significantly correlated with the values calculated by the empirical formula of CWSI, indicating that it was feasible to use a fixed value as the upper limit of canopy temperature minus air temperature in *B. chinensis* var. *parachinensis*. The prediction models of KNN, SVM, XGBoost and RF all achieved good results, and the correlation coefficients were 0.873, 0.877, 0.887 and 0.863, respectively.【Conclusion】Machine learning can be used for predicting the photosynthetic rate of *B. chinensis* var. *parachinensis*, avoid the use of large and cumbersome instruments, reduce the damage to the leaves of *B. chinensis* var. *parachinensis*, and reduce the measurement time.

Key words: *Brassica chinensis* var. *parachinensis*; crop water stress index; photosynthetic rate; machine learning

油青菜心 *Brassica chinensis* var. *parachinensis* 是一种重要的绿色蔬菜，广泛种植于广东省及全国多个城市，其叶片为油绿色，形状尖细，风味甜爽，菜苔油绿有光泽[1-2]。油青菜心对土壤水肥具有较高的要求，在生长过程中需要保证足够的水分供应，以满足其正常生长需求[3]。在持续缺水状态下油青菜心的多项生理活动受到影响，如叶绿素含量下降、叶片含水量下降等[4]。通过监测油青菜心的水分状况，采取高效的水分灌溉策略可保证油青菜心的品质，同时达到节约水资源、提高水资源利用效率的效果。

当作物缺水时冠层温度升高，基于冠气温差的水分胁迫指数能实时反映作物的水分亏缺状态[5-9]。水分胁迫指数 (Crop water stress index，CWSI) 是反映作物水分状况的无量纲因子，范围在 [0，1] 之间，0 表示作物无水分胁迫或者灌溉充分，1 表示作物无蒸腾作用或严重胁迫。国内外研究在使用 CWSI 经验模型时，上下限方程确定方法并不统一[10]。Jones 等[11] 使用涂抹凡士林和水的干、湿参考面温度计算葡萄的 CWSI；但是使用干、湿参考面需考虑位置对计算模型的影响[12]。王卫星等[13] 采用水汽压差为 0 和 6 时对应的温度研究柳叶菜心的水分状况；但是水汽压差随着地域和季节变化，意味着基准线也会变化[14]。张立元等[15]、Agam 等[16] 使用空气温度+5 ℃ 作为上限研究玉米、橄榄树等作物的水分变化规律，Kumar 等[17]、Khorsandi 等[18] 使用空气温度+2、+3 ℃ 作为上限研究芥菜、芝麻的 CWSI，避免了建立上限方程；虽然能构建 CWSI，但是不少研究直接套用经验值，没有结合作物本身及环境情况进行分析。屈振江等[19]研究苹果冠层温度的变化特征时发现，冠层温度峰值出现的时间、变化趋势与空气温度一致。针对相似的数据，具有噪声的基于密度的空间聚类 (Density-based spatial clustering of applications with noise，DBSCAN) 算法可将其划分到集合中，以簇中心代表集合的数据[20]。因此，本研究拟结合 DBSCAN 聚类算法，在给定的空气温度范围内探讨油青菜心冠气温差上限的固定值。

谢慧婷[21]、Matese 等[22] 研究生菜、葡萄藤的水分状况时发现，在缺水状态下光合作用速率呈下降趋势，而 CWSI 与光合作用速率呈负相关的关系。利用光合作用测量系统可直接获取作物的光合作用速率[23-24]，通过叶室夹住被测叶片形成固定被测空间并取样实现数据的自动采集，但叶室对叶片有

一定程度的损坏。通过其他容易获取的数据模拟作物的光合作用速率，能有效解决损坏叶片的问题，刘煦[25] 使用有效辐射、相对湿度等环境数据，基于机器学习方法预测了林下参的光合作用速率；陈硕博[26]、陈俊英等[27] 利用棉花冠层的光谱数据，以线性回归、机器学习多种方法反演了光合参数。机器学习方法已被广泛应用于农业领域，宋飞扬等[28]、Botula 等[29] 使用最邻近节点 (k-Nearest Neighbor，KNN) 算法筛选特征、模拟土壤持水量，提高了模型的预测精度；Mohammadi 等[30] 使用支持向量回归 (Support vector regression，SVR) 算法模拟每日参考蒸发量，结果表明预测值能较好地拟合真实值；Wang 等[31] 利用土壤回声结合极端梯度提升 (Extreme gradient boosting，XGBoost) 算法预测土壤 pH，达到预测精度高、误差低的效果；白婷等[32] 应用光谱数据和随机森林 (Random forest，RF) 算法获得较高精度的土壤有机质估测值。本研究拟建立 CWSI 与光合作用速率的关系，以 CWSI 预测光合作用速率，探究不同胁迫程度下油青菜心光合作用速率的变化规律。在油青菜心从四叶一心到植株现蕾的营养生长阶段 (Vegetative growth stage，V 期) 和从植株现蕾到菜苔高度与苔叶先端齐平的生殖生长阶段 (Reproductive growth stage，R 期) 进行不同水分处理，计算 CWSI，运用 DBSCAN 聚类分析方法选取作为冠气温差上限的固定值，采用 KNN、SVR、XGBoost、RF 这 4 种机器学习方法预测光合作用速率。

1　材料与方法

1.1　试验材料

以油青菜心为试验材料，试验于 2020 年 11—12 月在华南农业大学工程学院进行。11 月初将种子播种于花盆中，花盆高度 15 cm，上口径 22 cm，每盆装适量的土壤。待菜心出芽长出第 1 片真叶后，将相同长势的菜心移到相同规格的花盆中，每盆 2 株，间距为 10 cm，待幼苗长到四叶一心时期开始进行不同水分处理，如图 1 所示。

图 1　油青菜心盆栽图

Fig. 1　The pot cultivation picture of *Brassica chinensis*

试验前每个花盆灌溉充足水分，静置 1 h，每个花盆取表层 0~10 cm 的土壤，称质量 (m)，放入烘箱在 105 ℃ 条件下烘干 6 h，待土壤冷却后再称质量 (m_d)，按公式 (1) 计算该花盆土壤的田间持水量 (w)[33]：

$$w=\left(\frac{m}{m_d}-1\right)\times 100\%, \tag{1}$$

以平均值作为试验采用的土壤田间持水量，计算结果为 32.2%。

1.2　水分处理及数据测定方法

根据本试验土壤的田间持水量 (32.2%)，将试验对象油青菜心的水分胁迫梯度设置为 5 组，分别用 T1~T5 表示。T1 的田间持水量为 32.2%，以 T1 作为对照，T2~T5 的田间持水量分别为 85%T1、70%T1、55%T1 和 40%T1。

数据采集方法：选取菜心冠层顶部完全展开的、并能获取充足的阳光的叶片作为测量目标。数据采集时间为每天 10:00—16:00，每 30 min 采集 1 次。使用手持式红外测温仪 (型号 Raytek ST18，雷泰公司产品) 进行冠层温度测量，该设备的光学分辨率为 12∶1，发射率为 0.95，光谱响应范围 8~14 μm，测温范围–20~500 ℃。测量时，保持红外测温仪与测量目标的距离在叶片大小的 12 倍以内。每个目标点测量 3 次，取平均值。使用手持式温湿度计 (型号 Aicevoos W8，艾沃斯公司产品) 进行空气温度、相对湿度的测量，测量时保持设备与测量目标的距离为 10 cm，1 min 内测量 3 次后取平均值。使用土壤水分传感器 (型号 RS485，鹰都公司产品) 进行土壤含水量测量，测量时将设备埋在 2 个目标叶片中间，与土壤表面的距离为 10 cm。使用光合作用测定仪 (型号 SYS-GH30D，塞亚斯公司产品) 进行光合作用速率、光合有效辐射的测量，该仪器基于快速准确的红外线 CO_2 气体分析仪法，测定量程为 0~3 000 mg/kg，精度为 3 mg/kg，测量时使用叶室夹住叶片，保持 1 min，测量 5 次后取平均值作为该目标点的实际光合参数。

受试验期间天气影响，在保证数据充足的前提下，本文剔除了阴雨天气 (12 月 13—20 日) 测量的数据。

1.3　水分胁迫指数（CWSI）计算

CWSI 的计算如以下公式所示[15]：

$$\mathrm{CWSI}=\frac{(\theta_c-\theta_a)-(\theta_c-\theta_a)_{ll}}{(\theta_c-\theta_a)_{ul}-(\theta_c-\theta_a)_{ll}}, \tag{2}$$

式中：θ_c 为作物冠层温度；θ_a 为空气温度；$(\theta_c-\theta_a)_{ll}$ 为下限方程或无水分胁迫基准线，是作物在无水分胁迫时或充分灌溉下的冠气温差；$(\theta_c-\theta_a)_{ul}$ 为上限方程或无蒸腾作用基准线，是作物在无蒸腾作用时、气孔关闭状态下的冠气温差；单位均为 ℃。

$(\theta_c-\theta_a)_{ul}$ 和 $(\theta_c-\theta_a)_{ll}$ 的计算如以下公式所示：

$$(\theta_c-\theta_a)_{ul}=A+B\times \mathrm{VPG}, \tag{3}$$

$$(\theta_c-\theta_a)_{ll}=A+B\times \mathrm{VPD}, \tag{4}$$

$$\mathrm{VPD}(\theta_a)=0.610\,8\frac{100-\mathrm{RH}}{100}\mathrm{e}^{\left(\frac{17.27\theta_a}{\theta_a+237.3}\right)}, \tag{5}$$

$$\mathrm{VPG}=\mathrm{VPD}(\theta_a)-\mathrm{VPD}(\theta_a+A), \tag{6}$$

式中：A 和 B 为回归系数；RH 为空气相对湿度；VPD 和 VPG 分别为饱和水汽压差和饱和水汽压差梯度。

根据不同的 CWSI 经验模型，分别使用公式 (3) 计算$(\theta_c-\theta_a)_{ul}$和将固定值作为$(\theta_c-\theta_a)_{ul}$，固定值由 DBSCAN 算法得到。

1.4 基于密度的聚类算法 (DBSCAN)

DBSCAN 算法能找到任意形状、集中分布的簇[34]。在给定的数据集中，DBSCAN 算法将所有对象标记为未访问，随机选择未访问的对象 p 并标记为已访问，通过检查 p 的 ε-(ε>0) 领域内包含的 MinPts 个对象判断其属性，其中 ε 为半径参数，MinPts 为领域密度阈值。若为核心点，将 p 的 ε-领域中的所有对象添加到集合 C 中，直到 C 不再扩展，迭代停止。

距离指标常用欧式距离和曼哈顿距离，特征向量$\boldsymbol{X}_i$、$\boldsymbol{X}_j$之间的欧式距离、曼哈顿距离计算如公式 (7)、(8)：

$$\text{欧式距离：}\ d_{ij}=\sqrt{\sum_{m=1}^{k}\left[x_i(m)-x_j(m)\right]^2}, \tag{7}$$

$$\text{曼哈顿距离：}\ d_{ij}=\sum_{m=1}^{k}\left|x_i(m)-x_j(m)\right|, \tag{8}$$

式中：d_{ij}表示第 i、j 个特征向量$\boldsymbol{X}_i$、$\boldsymbol{X}_j$的距离；k 表示特征向量的维度；$x_i(m)$、$x_j(m)$ 分别为$\boldsymbol{X}_i$、$\boldsymbol{X}_j$ 在第 m 维的值，m 的取值范围为 1, 2, …, k。

DBSCAN 不需要预先设定簇数目，由算法自动决定，但需要设定半径参数 ε 和领域密度阈值 MinPts。在本研究中，以欧式距离作为度量，ε 为 0.44，MinPts 为 4。

1.5 最邻近节点算法 (KNN)

KNN 是数据挖掘算法中最简单的一种，精度高且对异常值不敏感[35]。对于要预测的点 $(\boldsymbol{x}_i, y_i)$，KNN 在一系列样本坐标中选择 k 个离$\boldsymbol{x}_i$最近的样本坐标，对其 y 值求平均，结果为 KNN 模型的预测值。

预先设定的值包括 k 值和 k 个邻近点的权重，在本研究中，以曼哈顿距离作为度量，k 为 3，各邻近点的权重一致。

1.6 支持向量回归 (SVR)

SVR 被广泛应用于处理模式识别问题，泛化错误率低[36]。该算法结合了核函数和线性回归，让训练集中的每个点 $(\boldsymbol{x}_i, y_i)$ 尽量拟合到一个线性模型 $y_i=\boldsymbol{\omega x}+b$。将最优超平面记作$\boldsymbol{\omega x}+b=0$，样本$\boldsymbol{x}$到最优超平面的距离为 r，通过引入松弛变量和惩罚系数求最小距离，将最优超平面问题转化为最优化问题：

$$\text{约束条件：}\quad r=\frac{1}{2}\|\boldsymbol{\omega}\|^2_{\min}+C\sum_{i=1}^{n}(\varepsilon_1,\varepsilon_2), \tag{9}$$

$$\text{服从条件：}\quad y_i-(\boldsymbol{\omega x}_i)-b\leqslant\varepsilon_1+\varepsilon_2, \tag{10}$$

$$(\boldsymbol{\omega x}_i)+b-y_i\leqslant\varepsilon_1+\varepsilon_2, \tag{11}$$

$$\varepsilon_1,\varepsilon_2\geqslant 0, \tag{12}$$

最终支持向量回归模型如公式 (13) 所示：

$$f(\boldsymbol{x})=\sum_{i=1}^{m}\left(\widehat{\alpha_i}-\alpha_i\right)k\left(\boldsymbol{x}_i^T\boldsymbol{x}\right)+b, \tag{13}$$

式中：$\|\boldsymbol{\omega}\|$为欧几里得范数，即$\sqrt{\boldsymbol{\omega}\cdot\boldsymbol{\omega}}$；$\boldsymbol{\omega x}$表示向量$\boldsymbol{\omega}\in R^N$与向量$\boldsymbol{x}\in R^N$的内积；$\varepsilon_1,\varepsilon_2$为松弛变量；$C$ 为惩罚系数；$k\left(\boldsymbol{x}_i^T\boldsymbol{x}\right)$为核函数；$\widehat{\alpha_i}$、$\alpha_i$、b 由约束条件求得。在本研究中，惩罚系数 C 为 4，核函数为径向基函数核。

1.7 极端梯度提升法 (XGBoost)

XGBoost 是 Chen 等[37] 开发的一个开源机器学习项目，处理速度快，高度灵活，能更好地控制过拟合。该算法在梯度提升决策树基础上对损失函数进行二阶泰勒展开并添加正则项，通过不断形成新决策树拟合之前预测的残差，从而减少预测值与真实值残差。其目标函数 L 由损失函数l和正则项Ω组成，如公式 (14)、(15)：

$$L(\varnothing)^t=\sum_{i=1}^{n}l\left[y_i,\hat{y}_i^{t-1}+f_t(\boldsymbol{x}_i)\right]+\Omega(f_k), \tag{14}$$

$$l\left(y_i,\hat{y}_i^{t-1}\right)=\left(y_i-\hat{y}_i^{t-1}\right)^2, \tag{15}$$

$$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\|\boldsymbol{\omega}\|^2, \tag{16}$$

式中：$L(\varnothing)^t$表示第 t 次迭代的目标函数；y_i为目标值；$\hat{y}_i^{t-1}$表示前 t−1 次迭代的预测值；l为目标值与预测值的平方差；$f_t(\boldsymbol{x}_i)$为第 t 次迭代产生的新模型；$\Omega(f_k)$表示第 t 次迭代的模型的正则项；γ 和 λ 表示正则项系数；T 表示该模型的叶结点个数。

对公式 (14) 使用泰勒展开，可得：

$$L(\varnothing)^t\cong\sum_{i=1}^{n}\left[g_if_t(\boldsymbol{x}_i)+\frac{1}{2}h_if_t^2(\boldsymbol{x}_i)\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2, \tag{17}$$

$$\omega_j^2=-\frac{\sum_{i\in I_j}g_i}{\sum_{i\in I_j}h_i+\lambda}, \tag{18}$$

式中：g_i表示样本$\boldsymbol{x}_i$的一阶导数；h_i表示样本$\boldsymbol{x}_i$的二阶

导数；ω_j 表示第 j 个叶子结点的输出值；I_j 表示第 j 个叶子结点值的样本子集。对 ω_j 求导并且令导函数等于 0，可求得使目标函数达到最小值的 ω_j。在本研究中，学习步长为 0.09，最大的树深度为 4，弱学习器数量为 70，正则化系数 γ 和 λ 为 0.001 和 1。

1.8　随机森林算法 (RF)

RF 最早由 Breiman[38] 于 2001 年提出，该算法基于分类树，是 Bagging 算法的优化，具有运算速度快、稳定性好等优势，在处理大数据集时预测精度高。RF 从总体数据集中采用多次、随机的方法取一部分样本构成样本簇，在每一个新生成的样本簇中训练回归决策树，组合每一棵决策树的输出结果加权平均。

对于给定的训练数据集，$D=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2),\cdots,(\boldsymbol{x}_i,\boldsymbol{y}_i)\}$，其中 $\boldsymbol{x}_i$ 为输入实例 (特征向量)，$\boldsymbol{y}_i$ 为目标向量，i=1，2，…，n，n 为样本容量。每次划分特征空间，逐一考察当前集合中所有特征的所有取值，根据平方误差最小化准则选择其中最优的一个作为切分点。对于训练集中第 j 个特征变量 $\boldsymbol{x}^j$ 和取值 s，作为切分变量和切分点，并定义 2 个区域 $d_{1(j,s)}=\{\boldsymbol{x}|\boldsymbol{x}^j\leqslant s\}$ 和 $d_{2(j,s)}=\{\boldsymbol{x}|\boldsymbol{x}^j>s\}$，为找出最优 j 和 s，对以下公式求解：

$$d=\left[\sum_{x_i\in d_{1(j,s)}}(\boldsymbol{y}_i-c_1)^2\right]_{\min}+\left[\sum_{x_i\in d_{2(j,s)}}(\boldsymbol{y}_i-c_2)^2\right]_{\min},\tag{19}$$

式中：c_1、c_2 为划分后 2 个区域内固定的输出值；利用选定划分区域内 $\boldsymbol{y}_i$ 求平均得到相应的输出值，直到生成完整的决策树。

1.9　模型评估

本研究运用 SPSS 软件，采用误差分析评估机器学习模型的预测效果，利用最小显著极差法分析水分处理间的显著水平。误差分析包括相关系数 (Correlation coefficient，R^2)、平均绝对误差 (Mean absolute error，MAE)、均方根误差 (Root mean square error，RMSE)，如公式 (20)~(22) 所示：

$$R^2=\frac{E(y_if_i)-E(y_i)E(f_i)}{\sqrt{E(y_i^2)-E^2(y_i)}\sqrt{E(f_i^2)-E^2(f_i)}},\tag{20}$$

$$\mathrm{MAE}=\frac{\sum|f_i-y_i|}{n},\tag{21}$$

$$\mathrm{RMSE}=\sqrt{\frac{\sum(f_i-y_i)^2}{n}},\tag{22}$$

式中：E 为数学期望；f 表示预测值；y 表示实际值；i 为第 i 个 ($i\leqslant n$) 数据，n 为数据量。

最小显著极差 (Least significant difference，LSD) 法，对任何 2 个 i、j 处理平均数间的均数 $(\bar{x}_i-\bar{x}_j)$，若其绝对值≥LSD_α，则为在 α 的水平上差异显著，LSD_α 的计算如公式 (23) 所示：

$$\mathrm{LSD}_\alpha=t_\alpha(\mathrm{df_e})\times\sqrt{\frac{2\mathrm{MS_e}}{n}},\tag{23}$$

式中：$\mathrm{df_e}$ 为误差自由度；t_α 为误差自由度下的临界值；$\mathrm{MS_e}$ 为误差均方；n 为各处理内的重复数。

2　结果与分析

2.1　充分灌溉下冠气温差与饱和水汽压差的关系分析

在无云晴天采集充分灌溉 T1 处理的冠层温度、空气温湿度数据，建立 CWSI 经验模型的冠气温差下限。图 2a、2b 分别为营养生长阶段 (V 期，11 月 27 日—12 月 12 日) 生殖生长阶段 (R 期，12 月 21 日—12 月 31 日) 的冠气温差与饱和水汽压差的散点图，通过回归方程拟合得到冠气温差的下限方程，方程如下所示：

V 期：$(\theta_c-\theta_a)_{ll}=-2.473\,6\times\mathrm{VPD}+2.291\,2$，($R^2$=0.847)；

R 期：$(\theta_c-\theta_a)_{ll}=-2.114\,0\times\mathrm{VPD}+3.080\,5$，($R^2$=0.785)。

由图 2 可以看出，冠气温差与饱和水汽压差的拟合结果较好，两者显著相关 (P<0.01)，均方根误差分别为 1.54(V 期) 和 1.07(R 期)。在 V 期和 R 期冠气温差的下限有一定差异，R 期的冠气温差下限比 V 期高，因此在建立油青菜心 CWSI 经验模型时应针对不同生长期建立相应的冠气温差下限。

根据所拟合的冠气温差下限方程和公式(3)计算 CWSI，并绘制 CWSI 每日变化曲线，如图 3 所示。随着田间持水量降低，水分胁迫程度加深，CWSI 明显增加。处理间的 CWSI 平均涨幅为 0.13，具有显著差异 (P<0.01)，表明 CWSI 能较好地反映油青菜心的水分状况。T1 处理为充分灌溉，其每日平均 CWSI 在 [0, 0.2] 波动，T2、T3 和 T4 处理的每日平均 CWSI 分别是 [0.2, 0.4]、[0.2, 0.5] 和 [0.3, 0.6]，受胁迫最多的 T5 处理的每日平均 CWSI 在 [0.4, 0.8] 波动，变化范围较大，表明水分胁迫程度越大，油青菜心越容易受到气候的影响。处理开始时变化曲线密集，随着胁迫时长增加，变化曲线无重叠或交点，表明水分的长期缺失对油青菜心的生理活动造成了一定影响，使得曲线波动范围缩小。

2.2　不同生长期的冠气温差上限分析

根据公式 (3) 获取的冠气温差上限及对应的空气温度，运用 DBSCAN 算法对油青菜心进行聚类，

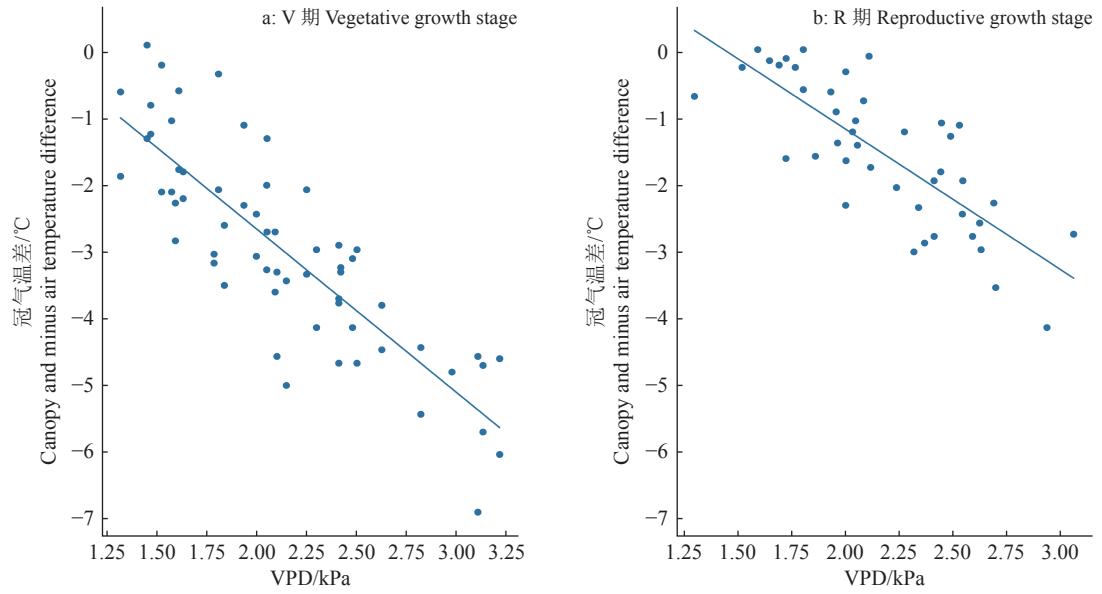
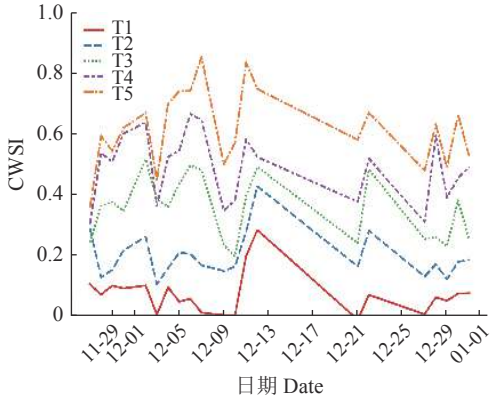


图 2 冠气温差与饱和水汽压差 (VPD) 的关系

Fig. 2 Relationship between canopy and air temperature difference and vapor pressure deficit (VPD)

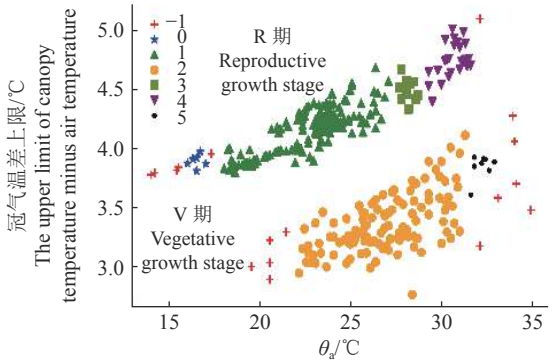


T1：田间持水量为 32.2%；T2~T5：田间持水量分别为 T1 的 85%、70%、55% 和 40%

T1: Field water holding capacity was 32.2%; T2~T5: Field water holding capacity was 85%, 70%, 55% and 40% of T1 respectively

图 3 5 种田间持水量处理的水分胁迫指数 (CWSI) 变化曲线

Fig. 3 Crop water stress index(CWSI) change curves under five field water holding capacities



−1 表示离群点，0~5 表示聚类形成的簇

−1 means outliers, 0-5 means clusters formed by clusters

图 4 基于 DBSCAN 的冠气温差上限聚类

Fig. 4 Clusters of the upper limit of canopy temperature minus air temperature by DBSCAN

结果如图 4 所示。由图 4 可以看出，随着空气温度上升，2 个时期的油青菜心冠气温差上限均呈现上升的趋势，其中 R 期的冠气温差上限比 V 期高，主要分布在 [4, 5] ℃，V 期的冠气温差主要分布在 [3, 4] ℃。由于试验期间空气温度变化较大，算法自动聚类的簇共有 6 个，且 R 期的空气温度变化比 V 期更大，因此 R 期有 4 个簇，V 期有 2 个簇，表明空气温度对 DBSCAN 的聚类效果影响较大。跨度越大则簇越多，增加了选取固定值的难度。数据集主要分布在 [20, 30] ℃，随着数据的稠密程度上升，簇内包含的数据增多，其中 2 个大簇的质心分别为 3.4 ℃ (类别为 2，116) 和 4.2 ℃(类别为 1，164)，分别对应 V 期和 R 期。此外，聚类结果显示有小部分噪声点 (类别为−1，是离群点)，大多分布在空气温度区间边缘，剔除该部分数据并不影响聚类的效果。

为验证聚类的效果，根据无蒸腾作用基线获取的上限，分别采用空气温度+2.0、+5.0 ℃ 与聚类结果计算 T1、T5 处理的水分胁迫指数，其误差分析如表 1 所示。由表 1 可以看出，在 2 个阶段、2 个处理中，空气温度+2.0 ℃ 的 CWSI 平均值偏高，空气温度+5.0 ℃ 的 CWSI 平均值偏低，在充分灌溉 T1 处理的条件下与无蒸腾作用基线的 CWSI 均无显著差异，当田间持水量为 40%T1(T5 处理) 时与无蒸腾作用基线的 CWSI 均有显著差异 (P<0.05)。空气温度+3.4、+4.2 ℃ 的 CWSI 均位于空气温度+2.0、

+5.0 ℃ 之间，与无蒸腾作用基线的 CWSI 显著相关 (R^2=0.99)。不同生长期、同温度、同处理的 CWSI 存在显著差异 (P<0.05)，当 V 期和 R 期采用同一个温度上限时，4.2 ℃ 将造成 V 期的 CWSI 偏小，3.4 ℃ 将造成 R 期的 CWSI 偏大，因此应结合作物的生长期选取不同的固定值。与充分灌溉 T1 处理相比，采用固定温度计算 T5 处理的 CWSI 与无蒸腾作用基线的平均值差值差异较大，V 期 2.0、3.4 和 5.0 ℃ 的涨幅分别为 0.152、0.003 和 0.104，R 期 2.0、4.2 和 5.0 ℃ 的涨幅分别为 0.350、0.019 和 0.044。拟合结果表明，3.4 和 4.2 ℃ 可分别作为油青菜心在 V 期和 R 期的冠气温差上限，本研究将以 2 个温度计算的 CWSI 作为机器学习模型的输入向量之一。

表 1　不同冠气温差上限的水分胁迫指数 (CWSI) 误差分析

Table 1　Error statistics of crop water stress index (CWSI) under different upper limits of canopy and air temperature difference

生长期 Stage	处理[1)] Treatment	固定温度/℃ Fixed temperature	CWSI		
			平均值 Mean	与无蒸腾基线的平均值差值[2)] Difference with average of non-transpiration-baseline	均方根误差 RMSE
V	T1	2.0	0.129	−0.019	0.042
		3.4	0.103	−0.002	0.008
		5.0	0.084	0.011	0.027
	T5	2.0	0.801	−0.171*	0.179
		3.4	0.635	−0.005*	0.029
		5.0	0.515	0.115*	0.125
R	T1	2.0	0.066	−0.025	0.066
		4.2	0.043	−0.002	0.007
		5.0	0.038	0.003	0.010
	T5	2.0	0.908	−0.375*	0.402
		4.2	0.555	−0.021*	0.033
		5.0	0.487	0.047*	0.058

1) T1：田间持水量为32.2%，T5：田间持水量为T1的40%；2) “*”表示T5处理下各固定温度的CWSI平均值与无蒸腾基线的CWSI平均值差值达0.05的显著水平(LSD法)

1) T1: Field water holding capacity was 32.2%, T5: Field water holding capacity was 40% of T1; 2) “*” indicates the significant difference at 0.05 level between CWSI average values of each fixed temperature and non-transpiration-baseline in T5 treatment (Least significant difference method)

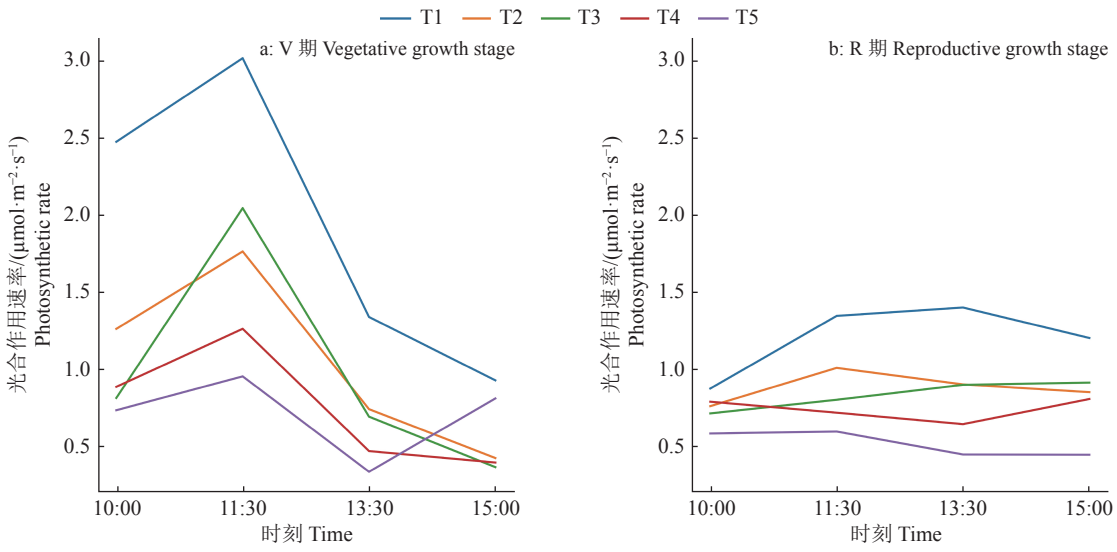
2.3　不同水分处理下的光合作用速率日变化分析

为研究油青菜心 CWSI 与光合作用速率的关系，绘制了光合作用速率的日变化曲线 (图 5)。由图 5 可以看出，光合作用速率呈现先增加而后减少的变化趋势，这是由于 11:30 时左右的太阳辐射较强、温度较高，有利于作物进行光合作用。在 V 期 11:30 时以后光合作用速率的下降趋势较明显，通过基于 LSD 法的多重比较分析，各个时刻的光合速率平均值差异均达 0.05 的显著水平，在 R 期光合作用速率的下降趋势较平缓。在 V 期、R 期中，CWSI 与光合作用速率是负相关的关系，充分灌溉处理 T1 的光合作用速率最高，当水分胁迫程度加深时，菜心叶面温度上升，光合作用速率下降，不同水分胁迫处理的光合作用速率在 0.05 水平具有显著差异 (LSD 法)。在 V 期，11:30 时 5 组水分胁迫处理的光合作用速率差距最大，在 15:00 时差距最小。

2.4　基于机器学习的光合作用速率预测分析

本试验运用 KNN、SVR、XGBoost、RF 这 4 种模型，以 CWSI 和光合有效辐射作为输入向量预测油青菜心的光合作用速率。光合作用速率测量值与预测值的散点图、拟合方程如图 6，数据经过标准化处理，预测模型的误差分析如表 2。

由图 6 和表 2 可以看出，XGBoost 模型的预测值与测量值的相关系数最高，其次是 SVR、KNN 和 RF 模型，4 种预测模型的相关系数均大于 0.85，拟合方程的截距及斜率均没有显著差异，表明这 4 种机器学习均可以预测油青菜心的光合作用速率。由于光合作用速率的整个数据集分布并不均匀，随机划分的测试集在各区间的分布与整个数据集基本一致，主要集中在 [0.5,1.5]，当光合速率>2 时预测



T1：田间持水量为 32.2%；T2~T5：田间持水量分别为 T1 的 85%、70%、55% 和 40%

T1：Field water holding capacity was 32.2%；T2~T5: Field water holding capacity was 85%, 70%, 55% and 40% of T1 respectively

图 5　不同水分胁迫处理的油青菜心光合作用速率日变化

Fig. 5　Diurnal variations of photosynthetic rate of *Brassica chinensis* in different water stress treatments

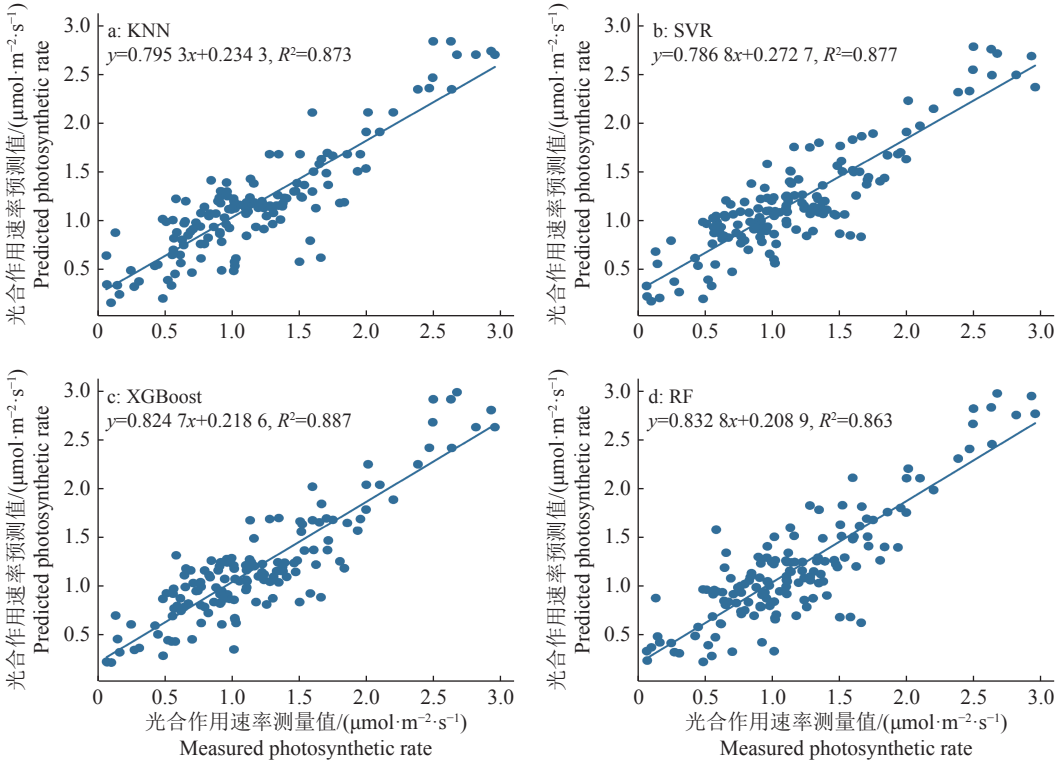


图 6　4 种模型的光合作用速率测量值与预测值的散点图

Fig. 6　The scatter plots of predicted and measured photosynthetic rates based on four models

表 2　4 种预测模型的误差分析

Table 2　Error statistics of four predicted models

模型 Model	决定系数 R^2	平均绝对误差 MAE	均方根误差 RMSE
KNN	0.873	0.226	0.294
SVM	0.877	0.234	0.289
XGBoost	0.887	0.227	0.279
RF	0.863	0.239	0.310

值均偏大，表明原始数据的分布对这 4 种模型的预测效果有一定影响。集成学习方法 XGBoost、RF 在建模时往往需要规模大的数据集，本试验用于预测的数据集共有 498 组输入向量，XGBoost、RF 的预测结果容易出现欠拟合的情况，无法体现提升树的优势。当数据集为原来的三分之一时，预测精度均下降，4 种模型建模的相关系数分别为 0.836(KNN)、

0.868(SVR)、0.871(XGBoost)、0.837(RF)，下降幅度最大的是 KNN 和 RF，表明 SVR 在小样本回归中比 RF、KNN 表现好。当数据集较小时，SVR 和 XGBoost 均可通过调整参数提高预测效果，如 SVR 减少惩罚系数和松弛变量，XGBoost 提高学习速率、降低最大生成树的数目，但 KNN 和 RF 并没有显著提高。XGBoost 和 RF 都属于集成学习，但 XGBoost 使用了一阶导数和二阶导数，使得预测值与实测值的拟合残差减少，又借鉴了 RF 列抽样的做法，降低过拟合和减少计算，大大提升效率，因此预测效果最好。

3　结论

本研究通过 5 组水分处理的对比分析揭示了油青菜心在营养生长阶段和生殖生长阶段的水分胁迫变化规律，运用 DBSCAN 算法对油青菜心冠气温差进行聚类，探讨了使用固定值作为上限的可行性，采用 KNN、SVR、XGBoost 和 RF 算法预测了油青菜心的光合作用速率。结果表明，CWSI 随田间持水量减少而升高，说明 CWSI 可以很好地识别油青菜心的水分胁迫；聚类结果显示，3.4 和 4.2 ℃ 与经验公式计算的 CWSI 具有显著的相关性，说明采用固定值计算油青菜心 CWSI 具有可行性，且 2 个阶段的温度存在差异，表明应结合作物的生长期进行聚类分析；4 种机器学习方法均取得较好的预测效果，XGBoost 模型的预测值与实测值拟合效果最好，其次是 SVR、KNN 和 RF，说明机器学习方法预测光合作用速率具有可行性，且实现了光合作用速率的快速检测。本研究只建立了油青菜心在 11—12 月的水分胁迫指数模型，由于经验方程受天气、季节变化的影响，下一步试验将采集全年不同月份的数据，并考虑太阳辐射、风速的影响。在预测光合作用效率时，数据尚不够充分，作物种类单一，下一步研究将扩充数据集，并期望能在多个作物上应用。

参考文献：

[1] 卢宇鹏, 夏岩石, 温少波, 等. 不同熟性菜心品质性状的多样性分析[J]. 广东农业科学, 2020, 47(5): 29-36.

[2] 叶红霞. 菜心的栽培季节和栽培方式[J]. 新农村, 2020(10): 26-27.

[3] 杨树涛, 黄永文, 刘泳涛, 等. 普宁市菜心生产特色农业气象指标研究[J]. 河南农业, 2018(11): 17-20.

[4] 徐燕. 土壤水分胁迫对菜心生理生化指标及气孔发育的影响[D]. 广州: 暨南大学, 2010.

[5] IDSO S B, JACKSON R D, PINTER P J, et al. Normalizing the stress-degree-day parameter for environmental variability[J]. Agricultural Meteorology, 1981, 24(1): 45-55.

[6] JACKSON R D, IDSO S B, REGINATO R J, et al. Canopy temperature as a crop water stress indicator[J]. Water Resources Research, 1981, 17(4): 1133-1138.

[7] JONE H G. Use of infrared thermometry for estimation of stomatal conductance as a possible aid to irrigation scheduling[J]. Agricultural and Forest Meteorology, 1999, 95(3): 139-149.

[8] RUD R, COHEN Y, ALCHANATIS V, et al. Crop water stress index derived from multi-year ground and aerial thermal images as an indicator of potato water status[J]. Precision Agriculture, 2014, 15(3): 273-289.

[9] 崔晓, 许利霞, 袁国富, 等. 基于冠层温度的夏玉米水分胁迫指数模型的试验研究[J]. 农业工程学报, 2005, 25(8): 22-24.

[10] 赵福年, 王瑞君, 张虹, 等. 基于冠气温差的作物水分胁迫指数经验模型研究进展[J]. 干旱气象, 2012, 30(4): 522-528.

[11] JONES H G, STOLL M, SANTOS T, et al. Use of infrared thermography for monitoring stomatal closure in the field: Application to grapevine[J]. Journal of Experimental Botany, 2002, 53(378): 2249-2260.

[12] BIAN J, ZHANG Z T, CHEN J Y, et al. Simplified evaluation of cotton water stress using high resolution unmanned aerial vehicle thermal imagery[J]. Remote Sensing, 2019, 11(3): 267. doi: 10.3390/rs11030267.

[13] 王卫星, 罗锡文, 区颖刚, 等. 基于冠层温度的菜心缺水指数模型初步试验研究 (英文)[J]. 农业工程学报, 2003, 19(5): 47-50.

[14] 孙道宗, 王卫星, 唐劲驰, 等. 茶树水分胁迫建模及试验[J]. 排灌机械工程学报, 2017, 35(1): 65-70.

[15] 张立元, 牛亚晓, 韩文霆, 等. 大田玉米水分胁迫指数经验模型建立方法[J]. 农业机械学报, 2018, 49(5): 233-239.

[16] AGAM N, COHEN Y, ALCHANATIS V, et al. How sensitive is the CWSI to changes in solar radiation?[J]. International Journal of Remote Sensing, 2013, 34(17): 6109-6120.

[17] KUMAR N, ADELOYE A J, SHANKAR V, et al. Neural computing modelling of the crop water stress index[J/OL]. Agricultural Water Management, 2020, 239: 1-10. [2021-01-05]. https://doi.org/10.1016/j.agwat.2020.106259.

[18] KHORSANDI A, HEMMAT A, MIREEI S A, et al. Plant temperature-based indices using infrared thermography for detecting water status in sesame under greenhouse conditions[J]. Agricultural Water Management, 2018, 204: 222-233.

[19] 屈振江, 郑小华, 王景红, 等. 渭北旱塬苹果园内外温度变化特征研究[J]. 干旱区地理, 2016, 39(2): 301-308.

[20] 吐尔逊·买买提, 谢建华. 基于 DBSCAN 的农机作业轨迹聚类研究[J]. 农机化研究, 2017, 39(4): 7-11.

[21] 谢慧婷. 基于红外热成像技术的生菜缺水指标的研究[D]. 福州: 福建农林大学, 2016.

[22] MATESE A, BARALDI R, BERTON A, et al. Estimation of water stress in grapevines using proximal and remote sensing methods[J]. Remote Sensing, 2018, 10(1): 114. doi: 10.3390/rs10010114.

[23] ZHANG L Y, NIUY X, ZHANG H H, et al. Maize canopy temperature extracted from UAV thermal and RGB imagery and its application in water stress monitoring[J]. Frontiers in Plant Science, 2019, 10: 1270. doi: 10.3389/fpls.2019.01270.

[24] SONG X Y, ZHOU G S, HE Q J, et al. Stomatal limitations to photosynthesis and their critical water conditions in different growth stages of maize under water stress[J/OL]. Agricultural Water Management, 2020, 241: 1-12. [2021-01-05]. https://doi.org/10.1016/j.agwat.2020.106330.

[25] 刘煦. 林下参种植光环境的动态预测与评价研究[D]. 长春: 吉林大学, 2014.

[26] 陈硕博. 无人机多光谱遥感反演棉花光合参数与水分的模型研究[D]. 杨凌: 西北农林科技大学, 2019.

[27] 陈俊英, 陈硕博, 张智韬, 等. 无人机多光谱遥感反演花蕾期棉花光合参数研究[J]. 农业机械学报, 2018, 49(10): 230-239.

[28] 宋飞扬, 铁治欣, 黄泽华, 等. 基于 KNN-LSTM 的 PM_(2.5) 浓度预测模型[J]. 计算机系统应用, 2020, 29(7): 193-198.

[29] BOTULA Y D, NEMES A, MAFUKA P, et al. Prediction of water retention of soils from the humid tropics by the nonparametric k-nearest neighbor approach[J]. Vadose Zone Journal, 2013, 12(2). doi: 10.2136/vzj2012.0123.

[30] MOHAMMADI B, MEHDIZADEH S. Modeling daily reference evapotranspiration via a novel approach based on support vector regression coupled with whale optimization algorithm[J/OL]. Agricultural Water Management, 2020, 237: 1-13. [2020-01-03]. https://doi.org/10.1016/j.agwat.2020.106145.

[31] WANG T T, YANG C H, LIANG J. Soil pH value prediction using UWB radar echoes based on XGBoost [C]//CSPS. International Conference in Communications, Signal Processing, and Systems. Urumqi: CSPS, 2019: 1941-1947.

[32] 白婷, 丁建丽, 王敬哲. 基于机器学习算法的土壤有机质质量比估算[J]. 排灌机械工程学报, 2020, 38(8): 829-834.

[33] 罗清元, 杨丹, 刘丽娜, 等. 基于不同环境下的河南省典型区域土壤田间持水量研究[J]. 节水灌溉, 2019, 4929(6): 35-38.

[34] MARTIN E, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//AAAI. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). California: AAAI, 1996: 226-231.

[35] ALTMAN N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992, 46(3): 175-185.

[36] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[37] CHEN T Q, GUESTRIN C. XGBoost: A scalable tree boosting system[C]//ACM. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. California: ACM, 2016: 785-794.

[38] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

【责任编辑 李晓卉】