

付伟,祝玉青,司红君,等.基于决策树的水稻病虫害发生程度预测模型——以芜湖市为例[J].气象研究与应用,2024,45(1):102-107.

FU Wei, ZHU Yuqing, SI Hongjun, et al. A prediction model for the occurrence degree of rice diseases and pests based on decision tree algorithm: A case study in Wuhu[J]. Journal of Meteorological Research and Application, 2024, 45(1): 102-107.

基于决策树的水稻病虫害发生程度预测模型

——以芜湖市为例

付伟¹, 祝玉青¹, 司红君¹, 邹莹瑾²

(1. 芜湖市气象局, 安徽 芜湖 241000; 2. 无为市气象局, 安徽 无为 238300)

摘要: 利用芜湖市1988—2022年7种水稻病虫害发生及种植面积数据, 国家气候中心逐月大气环流和海温指数, 基于C5.0决策树算法, 构建一种以大气环流和海温指数作为预报因子的水稻病虫害发生程度的长期预测模型。结果表明, 该模型可较好地预测未来一年芜湖市各种水稻病虫害发生程度, 2022年7种病虫害发生程度预测准确率平均为85.7%, 为水稻病虫害发生程度的预测提供一种有效的实用方法。

关键词: C5.0决策树算法; 水稻病虫害发生程度; 预测模型; 大气环流和海温指数

中图分类号: S431

文献标识码: A

doi: 10.19849/j.cnki.CN45-1356/P.2024.1.17

水稻是我国第一大粮食作物, 水稻生产是我国粮食安全的重要组成部分^[1], 而病虫害是我国最主要的农业生物灾害之一, 每年造成包括水稻在内的大量粮食作物损失^[2]。近年来, 水稻病虫害的防治取得大量的成果, 通过优化种植制度, 在播种时进行灭菌处理等方式, 恶化病虫害的发生和生存环境, 部分病虫害如白叶枯和三化螟, 在一些地区已很少或者基本不发生, 主要水稻产区的虫害发生面积呈逐年减少趋势, 实际产量损失也呈下降趋势^[3-4], 但水稻病虫害的防控工作依然不容放松。

在防控研究之外, 水稻病虫害的预测也是重要的研究方向。在地理环境、水稻品种、种植制度相对稳定的情况下, 气象因子已成为影响病虫害发生年际变化的主控因子^[5]。因此许多学者开展以气象条件作为影响因子的水稻病虫害预测方法研究, 通常的研究方法是首先分析与病虫害显著相关、有生物学意义的气象要素作为影响因子, 如罗伟等^[6]发现自贡市冬季平均气温是水稻二化螟发生最为显著的影响因子。黄珍珠等^[7]指出广东稻瘟病致病日

的气象条件为日平均气温20~30℃、日平均相对湿度≥90%、日照时数≤2 h、日降水量≥1 mm。彭荣南等^[8]发现影响水稻白叶枯病发生的关键气象因子分别是8月中旬至9月中旬降水量、9—10月降雨系数和8月下旬降水强度。此外, 大气环流指数和海温指数也可作为水稻病虫害长期预测的预报因子^[9-10]。确定预报因子后采用规律统计^[10]、回归分析^[6-9]、经验法则^[11]等建立病虫害的预测模型和方法。但各种统计方法都有自己的局限性, 如一些统计方法对已有的变量样本拟合较好甚至过拟合, 对新的样本预测准确率不够稳定^[12]。建立预测模型常见的回归方法可以定量预测病虫害^[6-9], 其预测的目标变量通常为连续变量, 预测结果是数值, 但实际农业气象服务过程中, 需要直接预测病虫害偏多偏少等偏离平均状态的程度, 预测结果为分类数据, 就需要借助其他类型的预测模型。

决策树模型, 也称规则推理模型, 通过对特征属性和目标变量构成的训练数据集进行循环分析, 建立分类规则, 实现对新样本的分类。它由一个根

收稿日期: 2024-01-04

基金项目: 国家重点研发计划项目(2018YFD0300905)、科技助力经济2020重点专项气象行业项目(KJZLJ202002)

第一作者简介: 付伟(1981—), 男, 高级工程师, 从事生态气象研究工作。E-mail: fw075@126.com

节点、一系列分支及叶子节点组成,并通过叶子节点实现分类。有别于一般统计分类模型,决策树的分类是基于逻辑的,具有优秀的学习效率和出色的解释能力^[13],分类逻辑清晰易懂,已广泛用于灾害天气识别^[14-15]、农产品产量预测^[16]、商业分析^[17]、林业信息^[18]和植被类型判别^[19]等领域,但鲜有在病虫害发生程度预测方面的应用研究。

司红君等^[9]分析验证芜湖市主要病虫害均受气象要素显著影响,大气环流和海温指数也可以作为病虫害长期预测的预报因子,因此本文将在此基础上,继续以芜湖市为例,使用决策树算法构建水稻主要病虫害偏多偏少等发生程度的长期预测模型,以期丰富病虫害预测产品和预测方法,为植保部门开展病虫害防治提供科学依据。

1 资料与方法

1.1 资料来源

本文使用的水稻相关资料包括 1988—2022 年芜湖市逐年水稻种植和病虫害发生面积。水稻种植面积数据摘自安徽省统计年鉴,病虫害发生面积数据来自芜湖市植保站,包括稻瘟病、纹枯病和稻曲病 3 种病害,二化螟、稻纵卷叶螟、稻蝗和褐飞虱 4 种虫害,白叶枯和三化螟近年来发生面积很少,因此本文不做讨论。

气候数据包括 1987—2021 年的逐月大气环流和海温指数,来自国家气候中心网站(<http://cmdp.ncc-cma.net/cn/monitoring.htm>),包括 88 项大气环流指数和 26 项海温指数,后文统称为气候指数,各项指数的名称和含义可在上述网页查询。为方便描述,本文使用 $AtmNN(mm)$ 和 $OceNN(mm)$ 代表各月气候指数变量,其中 Atm 表示大气环流指数, Oce 表示海温指数, NN 为该气候指数的编号, mm 表示该指数的月份,如 $Atm40(11)$ 表示 11 月第 40 项大气环流指数, $Oce5(9)$ 表示 9 月第 5 项海温指数。

1.2 方法

本文使用逐年病虫害发生面积偏离平均状态的程度(距平百分比)代表病虫害发生程度。由于 2011 年安徽省撤销地级巢湖市,其所辖无为县并入芜湖市,成为芜湖市面积最大、人口最多的县级行政区,因此芜湖市 2012 年起水稻种植总面积较之前大幅增加,病虫害发生面积相应出现较大变化,为保证历史数据可比性,定义病虫害发生面积比=病虫害发生面积/水稻种植面积(单位:%)^[9],代表病虫

害的发生量。

本文使用 C5.0 决策树算法进行预测模型的训练和构建。C5.0 算法由 ID3、C4.5 算法不断改进演变而来^[20-21]。相较于 ID3 算法, C5.0 改进为采用信息增益率来选择决策树节点的分裂属性,并且可以通过剪枝简化决策树规模,防止出现过拟合^[22]。相较于 C4.5 算法, C5.0 引入以 Boosting 提升法为代表的新技术,决策树构建速度更快,规模更小,适用性更强。

C5.0 决策树算法的大概训练过程如下:设训练样本集 S 共有 n 个样本,构建决策树模型的训练次数(迭代次数)为 T ,用 C^t 表示由第 t 次训练产生的模型,最终构建的模型记为 C^* 。用 ω_i^t 表示第 i 个样本在第 t 次模型构建中的权重($i=1, 2, \dots, n; t=1, 2, \dots, T$),然后进行如下模型训练过程^[23-24]:

(1) 设定初始权重值 $\omega_i^1=1/n$, 令 $t=1$ 开始第一次训练。

(2) 计算 ω_i^t 的归一化因子 p_i^t 。

(3) 为每个样本赋予归一化权重值 p_i^t , 在此基础上构建模型 C^t 。

(4) 计算第 t 次训练的分类错误率 ε^t 。

(5) 如果 $\varepsilon^t > 0.5$, 令 $T=T-1$, 返回步骤(1); 如 $\varepsilon^t = 0$, 模型训练过程结束, 令 $t=T$; 如果 $0 < \varepsilon^t \leq 0.5$ 进行步骤(6)。

(6) 计算权重值的调整因子 $\beta^t = \varepsilon^t / (1 - \varepsilon^t)$ 。

(7) 更新样本的权重值, 如果样本分类正确 $\omega_i^t + 1 = \omega_i^t \beta^t$; 如果样本分类错误 $\omega_i^t + 1 = \omega_i^t$ 。

(8) 如果 $t=T$, 模型训练过程结束, 否则, 令 $t=t+1$, 再跳至步骤(2)进行下一次模型训练。直至求得最终的决策树模型 C^* 。

C^* 由 C^1, C^2, \dots, C^T 加权求和得到, 每个权重值均为 $\log(1/\beta^t)$ 。即当利用决策树模型对训练样本进行分类时, 先用 $C^t (1 \leq t \leq T)$ 分别对样本进行分类, 一共得到 T 个分类结果, 然后根据每个 C^t 的权重值统计每种类别的得数, 以得数最高的类别作为整个模型的分类结果。

构建 C5.0 决策树模型时, 使用病虫害发生程度的分类数据作为目标变量(因变量), 气候指数作为自变量。构建模型前, 首先确定变量类型, 对变量的缺失值进行审核, 过滤掉含有缺失值的数据记录, 然后确定目标变量(因变量)和自变量。构建模型时使用交叉验证, 折叠次数 10, 自检模型中可解释样本的比例作为模型准确率。

2 模型变量的预处理

首先,为达到预测病虫害发生程度的目的,计算1988—2022年芜湖市病虫害发生面积比的距平百分比时间序列。计算距平时,为保证后期科研工作时历史距平值具有可比性,参考国家气候中心气候标准期的概念,将1991—2020年病虫害发生面积比的平均值设为其多年平均的标准值,具体平均值计算结果如下:稻瘟病28.85%、纹枯病80.03%、稻曲病19.05%、二化螟56.04%、稻纵卷叶螟67.61%、稻蝗9.19%和褐飞虱65.18%。在此基础上计算1988—2022年芜湖市病虫害发生面积比的距平百分比时间序列(图1、图2)。

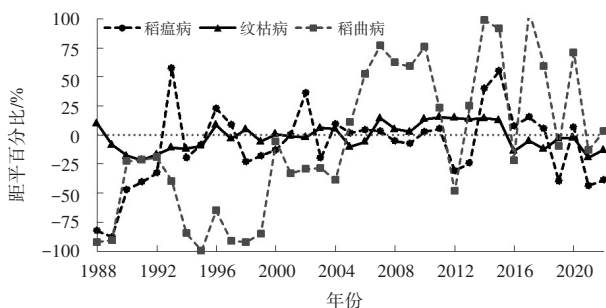


图1 芜湖市1988—2022年稻瘟病、纹枯病和稻曲病发生面积比的距平百分比变化

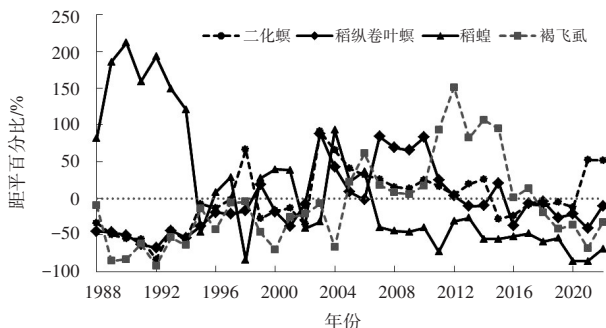


图2 芜湖市1988—2022年二化螟、稻纵卷叶螟、稻蝗和褐飞虱发生面积比的距平百分比变化

其次,根据建模的需要,进一步将目标变量(因变量)转换为分类数据。各病虫害发生数据仅有34个样本(34 a),不宜在一次模型构建中分类过多,造成每类数据样本过少,影响模型的准确性。因此,按照距平百分值,每次将一种病虫害的发生面积比数据转换为一组两类分类数据,分多次放入模型中建模,每次仅放入同一组的两类分类数据。按照上述方法,将纹枯病数据转换为偏多和偏少、偏多一

成以上和以下、偏少一成以上和以下三组分类数据,并分三次放入模型中建模,其中未来一年纹枯病发生面积比是否偏多、偏少、偏多一成以上、偏少一成以上为预报项目,偏多一成以下和偏少一成以下用于辅助同组分类数据构建模型。其他病虫害数据转换为偏多和偏少、偏多二成以上和以下、偏少二成以上和以下三组分类数据,分别放入模型中建模,其中预报项目为未来一年各病虫害发生面积比是否偏多、偏少、偏多二成以上和偏少二成以上,偏多二成以下和偏少二成以下用于辅助同组分类数据构建模型。

第三,为充分检验模型的准确性和业务适用性,在模型自检的基础上,对模型实际的预测能力进行检验,因此建模时因变量仅使用1988—2021年的病虫害发生数据,模型构建后,使用2022年的数据对模型的预测结果进行实况检验。

最后,从实际业务考虑,在年初预测新一年病虫害发生程度时,应使用前一年的气候指数进行预测,因此建模时使用1987—2020年的逐月气候指数作为自变量建模。另外,逐月114项气候指数共1368组变量,需先筛选相关性较好的变量,将1988—2021年的各病虫害发生面积比时间序列分别与前一年(1987—2020年)的月气候指数时间序列进行相关性分析,相关系数通过0.05显著性检验的气候指数时间序列方可作为自变量参与建模。经过计算,各病虫害预测模型参与建模的自变量数量为:稻瘟病127个,纹枯病133个,稻曲病359个,二化螟163个,稻纵卷叶螟200个,稻蝗229个,褐飞虱213个,也就是各水稻病虫害显著相关气候指数的数量。

3 模型构建结果

将处理完成的各病虫害年发生程度分类数据(因变量)和对应显著相关月气候指数数据(自变量)放入预测模型,基于C5.0决策树算法构建预测模型。7种病虫害,每种4个预测项目,共28个预测项目。所有预测项目的预测模型均构建成功,可输出各预测模型的决策树示意图和预测规则的判断公式,举例展示如下:

稻蝗年发生面积比偏多和偏少的预测决策树示意图显示(图3),模型中偏多和偏少的样本分别有15和19个。构建的决策树模型共有4层,6个节点,有3个用于判断分类的预测变量,其中位置越高

的预测变量重要性越高, 3 个预测变量从上到下分别在 2、3、4 层的分枝处用于分类的判断。该决策树中, $Oce16(6)$ (6 月西太平洋暖池强度指数) 是最重要的预测变量。当 $Oce16(6) > 29.347$ 时, 稻蝗年发生面积比 100% 偏少, 反之 28.57% 偏少, 71.43% 偏多。第二个预测变量是 $Oce25(9)$ (9 月热带印度洋海温偶极子指数), 在 $Oce16(6) \leq 29.347$ 的情况下, 当 $Oce25(9) > 0.447$ 时, 稻蝗年发生面积比 100% 偏少, 反之 11.77% 偏少, 88.24% 偏多。第三个预测变量是 $Atm58(7)$ (7 月北半球极涡中心强度指数), 在 $Oce16(6) \leq 29.347$, $Oce25(9) \leq 0.447$ 的情况下, 当 $Atm58(7) > 5\ 433.030$ 时, 稻蝗年发生面积比 100% 偏少, 反之 100% 偏多。即 6 个节点中, 节点 3 为偏多的预测规则, 4、5、6 为偏少的预测规则。

稻蝗年发生面积比偏多的预测规则用公式表示为: $Oce16(6) \leq 29.347$ 且 $Oce25(9) \leq 0.447$ 且 $Atm58(7) \leq 5\ 433.030$, 偏少的预测规则用公式表示为 $Oce16(6) > 29.347$ 或 $Oce16(6) \leq 29.347$ 且 $Oce25(9) > 0.447$ 或 $Oce16(6) \leq 29.347$ 且 $Oce25(9) \leq 0.447$ 且 $Atm58(7) > 5\ 433.030$ 。模型自检表明, 预测规则可以解释模型中的 33 个样本, 准确率为 97.06%。

当模型中的两类分类数据, 仅有一类为预报项目, 另一类用于辅助建模时, 如褐飞虱年发生面积比偏多二成以上的预测决策树示意图 (图 4), 模型中偏多二成以上的样本有 7 个, 其余 27 个样本为偏多二成以下, 用来辅助建模。构建的决策树模型共有 3 层, 4 个节点, 有 2 个用于判断分类的预测变量。该决策树中, $Atm72(1)$ (1 月东大西洋遥相关型指数) 是最重要的预测变量, 其次是 $Atm70(8)$ (8 月北大西洋涛动指数)。两个预测变量分别在 2、3 层的分枝处用于分类的判断。从上到下根据分枝处的分类规则依次判断, 判断出的 4 个节点中, 节点 1、3 为偏多二成以上的预测规则。

褐飞虱年发生面积比偏多二成以上的预测规则用公式表示为: $Atm72(1) \leq -0.404$ 或 $Atm72(1) > -0.404$ 且 $Atm70(8) \leq -0.522$ 。模型自检表明, 预测规则也可以解释模型中的 33 个样本, 准确率为 97.06%。

其余各病虫害预测项目的预测决策树示意图、公式表示的预测规则和自检模型准确率不再一一展示, 其中稻蝗年发生面积比偏少和偏少二成以上的训练样本基本相同, 因此构建出的预测规则也相同。自检模型准确率结果显示, 各模型的预测规

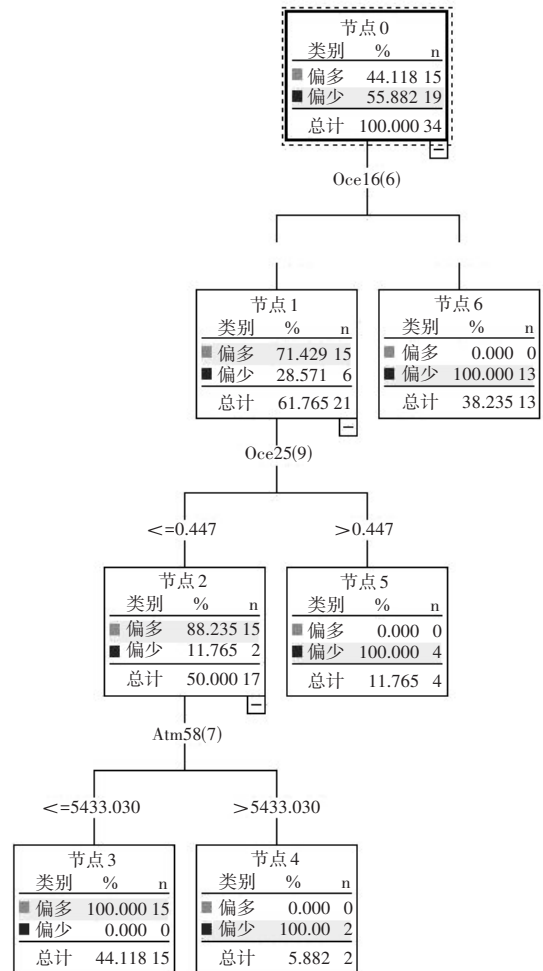


图3 稻蝗年发生面积比偏多和偏少的预测决策树

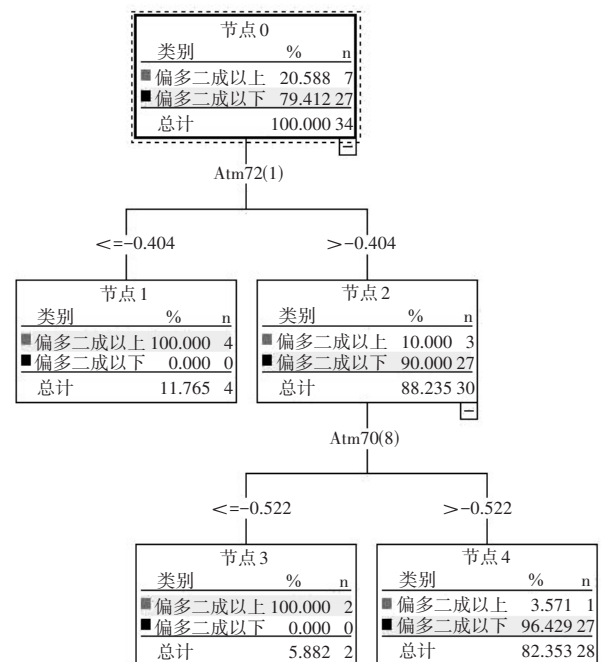


图4 褐飞虱年发生面积比是否偏多二成以上的预测决策树

则,可以较好地解释模型中的样本,准确率均在90%以上,其中6个模型的准确率达到100%(年发生面积比稻曲病偏多、偏少和偏多二成以上,稻纵卷叶螟偏少二成以上,稻蝗偏多二成以上,褐飞虱偏少二成以上)。

4 模型预测效果检验

查询2021年逐月各气候指数,将对应的数值代入各病虫害预测规则中计算预测结论,并将预测结论与2022年各病虫害发生面积比的实际距平百分比值进行对比,检验模型的预测效果。

上一小节两个例子中,稻蝗年发生面积比偏多和偏少的预测变量为 *Oce16*(6)、*Oce25*(9)和 *Atm58*(7),其2021年的值分别为33.144、0.033和5403.323,将3个值代入预测规则中,计算出偏多的结论为“否”,偏少的结论为“是”,而2022年稻蝗年发生面积比的实况值为-69.0%,因此两个预测项目的检验结果均为“正确”。

褐飞虱年发生面积比偏多二成以上的预测变量为 *Atm72*(1)和 *Atm70*(8),其2021年的值分为-0.130和0.337,将2个值代入预测规则中,计算出偏多二成以上的结论为“否”,而2022年褐飞虱年发生面积比的实况值为-32.6%,因此预测项目的检验结果均为“正确”。

检验结果显示,在28个预测规则中,24个预测正确,2个预测错误(稻曲病和稻纵卷叶螟年发生面积比偏多二成以上),2个因 *Atm24*(2)(2月北非副高脊线位置指数)2021年数据缺测无法判断(二化螟年发生面积比偏多和偏少二成以上),正确率达到85.7%,可判断的预测规则正确率92.3%,各预测项目的预测变量值、预测结论、2022年病虫害发生程度实况值和检验结果不再一一展示。错误的2个预测项目用于训练的样本数量分别为11、8个,相对较少,可能造成模型训练不充分,影响预测效果。检验结果说明,总体上基于C5.0决策树算法所构建的病虫害发生程度的预测模型具有较好的预测能力和一定的实际应用价值。

5 结论和讨论

前期研究工作表明,芜湖市各水稻病虫害均受气象因子显著影响,可以使用气候指数作为预报因子对其进行长期定量预测。在此基础上,本文继续以芜湖市为例,基于C5.0决策树算法,构建一种以

大气环流和海温指数作为预报因子的病虫害发生程度长期预测模型,取得主要结论如下:

(1)对病虫害发生数据进行面积比和距平百分比处理,再转换为发生程度的分类数据作为目标变量(因变量),遴选前一年与因变量显著相关的月大气环流指数和海温指数作为自变量,成功构建7种常见病虫害发生程度的预测模型,得到具体的预测规则,可实现对未来一年各病虫害偏多、偏少和偏多二成以上、以下等多个发生程度的预测。

(2)各模型的自检结果表明,模型可以解释绝大部分的样本,准确率均在90%以上,其中6个模型的准确率达到100%。将2021年大气环流、海温指数和2022年各病虫害发生程度的实况值代入各预测规则,对预测结果进行检验,准确率达到85.7%,具有较好的预测能力。

本研究中对于各病虫害发生偏离平均状态程度的分类可能有不妥之处,实际业务中可根据植保部门的需求进行个性化划分,以适应实际业务。另外决策树算法也可用于目标变量(因变量)影响因子的筛选,如本研究可分析出某种病虫害发生程度预测中最重要的几个气候指数,后续的研究可在此基础上进一步分析这些气候指数对研究区域天气气候和病虫害发生影响的内在物理机制。

参考文献:

- [1] 程式华,胡培松.中国水稻科技发展战略[J].中国水稻科学,2008,22(3):223-226.
- [2] 霍治国,李茂松,王丽,等.气候变暖对中国农作物病虫害的影响[J].中国农业科学,2012,45(10):1926-1934.
- [3] 沈颖,周宇杰,骆琴,等.水稻白叶枯病“2+3”处理策略探讨[J].浙江农业科学,2023,64(12):2923-2925.
- [4] 车琳,蒋沁宏,王也,等.我国水稻五大产区虫害发生及防控情况差异的比较分析[J].植物保护,2022,48(3):233-241.
- [5] 廖雪萍,黄梅丽,雍阳阳,等.气候变化对广西农业影响的研究进展与展望[J].气象研究与应用,2020,41(4):72-80.
- [6] 罗伟,段修荣,钟莉,等.自贡市水稻二化螟发生气象等级预报研究[J].湖北农业科学,2020,59(9):20-23.
- [7] 黄珍珠,杜尧东,黄德超,等.基于气象因子的广东早稻穗颈瘟发生面积的预测[J].广东气象,2020,42(2):44-48.
- [8] 彭荣南,陈冰,陈观浩,等.水稻白叶枯病发生流行与气象条件关系及预测模型研究[J].农学学报,2020,10(2):29-33.

- [9] 司红君, 付伟, 徐阳, 等. 皖江地区水稻病虫害长期定量预测方法研究——以芜湖为例[J]. 江西农业学报, 2022, 34(4): 58–63.
- [10] 王纯枝, 张蕾, 郭安红, 等. 基于大气环流的稻纵卷叶螟气象预测模型[J]. 应用气象学报, 2019, 30(5): 565–576.
- [11] 陈冰, 林汉龙, 宋祖钦, 等. 基于经验法则的水稻细菌性条斑病气候年型分析[J]. 气象研究与应用, 2020, 41(1): 26–30.
- [12] 方敏, 牛文科, 张晓松. 分类回归树多吸引子细胞自动机分类方法及过拟合研究[J]. 计算机研究与发展, 2012, 49(8): 1747–1752.
- [13] CHAN P K, WEI F, PRODROMIDIS A L, et al. Distributed data mining in credit card fraud detection[J]. IEEE Intelligent Systems, 1999, 14(6): 67–74.
- [14] 袁凯, 李武阶, 庞晶. 基于决策树算法的鄂东地区冰雹识别技术[J]. 应用气象学报, 2023, 34(2): 234–245.
- [15] 史达伟, 张静, 曹庆, 等. 基于决策树算法的海州湾地区海雾预测[J]. 气象科学, 2022, 42(1): 136–142.
- [16] 张帆, 王耐寒, 兰鹏, 等. 苹果产量集成预测模型研究[J]. 江苏农业科学, 2022, 50(4): 181–186.
- [17] 张宇, 张之明. 一种基于 C5.0 决策树的客户流失预测模型研究[J]. 统计与信息论坛, 2015, 30(1): 89–94.
- [18] 陈丽萍, 孙玉军. 基于不同决策树的面向对象林区遥感影像分类比较[J]. 应用生态学报, 2018, 29(12): 3995–4003.
- [19] 莫建飞, 陈诚, 陈燕丽, 等. 2000–2019 年广西喀斯特地区植被分布格局及时空变化分析[J]. 气象研究与应用, 2022, 43(4): 13–19.
- [20] TURE M, TOKATLI F, KURT I. Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients[J]. Expert Systems with Applications, 2009, 36(2): 2017–2026.
- [21] QUINLAN J R. C4.5: Programs for machine learning[M]. Boston: Kluwer Academic Publishers, 1994.
- [22] 杨已煜. 基于 C5.0 决策树算法的开放数据的效用预测研究[J]. 统计理论与管理, 2019(10): 17–22.
- [23] 庞素琳, 巩吉璋. C5.0 分类算法及在银行个人信用评级中的应用[J]. 系统工程理论与实践, 2009, 29(12): 94–104.
- [24] 帅青红, 李成林, 蒋佳霖. 商业银行客户细分与金融资产投资偏好——基于交易行为的新视角[J]. 财经科学, 2020(4): 25–36.

A prediction model for the occurrence degree of rice diseases and pests based on decision tree algorithm: A case study in Wuhu

FU Wei¹, ZHU Yuqing¹, SI Hongjun¹, ZOU Yingjin²

(1. Wuhu Meteorological Bureau, Anhui Wuhu 241000, China;

2. Wuwei Meteorological Bureau, Anhui Wuwei 238300, China)

Abstract: Based on the C5.0 decision tree algorithm, a long-term prediction model for the occurrence of rice diseases and pests using atmospheric circulation and sea surface temperature (SST) index as predictive factors was constructed using the occurrence of seven types of rice pests and diseases and planting area data of Wuhu City from 1988 to 2022, and the monthly atmospheric circulation and SST index of National Climate Centre (NCC). The results show that these models can effectively predict the occurrence degree of various rice diseases and pests in Wuhu City in the next year. The average predicting accuracy of the occurrence degree of seven diseases and pests in 2022 is 85.7%, which provides an effective and practical method for predicting the occurrence degree of rice diseases and pests.

Key words: C5.0 decision tree algorithm; occurrence degree of rice pests and diseases; prediction model; atmospheric circulation and sea surface temperature index