陈志浩, 王建华, 龙拥兵, 等. 基于 Spark 的 WOA-BP 水稻产量预测 [J]. 华南农业大学学报, 2023, 44(4): 613-618.
CHEN Zhihao, WANG Jianhua, LONG Yongbing, et al. WOA-BP rice yield prediction based on Spark[J]. Journal of South China Agricultural University, 2023, 44(4): 613-618.

基于 Spark 的 WOA-BP 水稻产量预测

陈志浩✉, 王建华✉, 龙拥兵, 兰玉彬, 刘军和, 熊弘依, 肖方军, 肖艺铭
(华南农业大学 电子工程学院 (人工智能学院)/国家精准农业航空施药技术国际联合研究中心/
岭南现代农业科学与技术广东省实验室, 广东 广州 510642)

摘要:【目的】随着大数据技术和人工智能的快速发展, 针对当前水稻产量预测模型精度低、预测区域范围过大、模型优化时间过长等问题, 本文提出一种基于 Spark 的鲸鱼优化算法−反向传播神经网络 (Whale optimization algorithm-backpropagation, WOA-BP) 水稻产量预测方法。【方法】本文以广东省西部地区的县/市/区水稻产量及气象数据作为研究对象, 采用 WOA 对 BP 网络的权值和偏置值进行优化, 并构建水稻产量预测模型, 提升预测精度; 此外, 在 Spark 框架下, 实现 WOA-BP 算法并行化, 减少算法时间开销。【结果】模型精度方面, 通过对预测结果进行反归一化后比较, 经 WOA 优化后的 BP 神经网络模型, 平均绝对百分比误差 (Mean absolute percentage error) 从 8.354% 降至 7.068%, 平均绝对误差 (Mean absolute error) 从 31.320 kg 降至 26.982 kg, 均方根误差 (Root mean square error) 从 41.008 kg 降至 33.546 kg; 运行时间方面, 3 节点 Spark 集群比非 Spark 模式减少了 11 742 s, 减少 44% 的时间开销。【结论】基于 Spark 的 WOA-BP 水稻产量预测方法, 能够较好地预测出广东西部县/市/区的水稻产量, 同时可以很好地反映气象因素对广东省西部地区水稻产量的影响情况, 对研究广东西部县/市/区乃至整个广东的水稻产量情况具有一定的参考价值。

关键词: 气象因素; 水稻产量; 反向传播神经网络; 鲸鱼优化算法; Spark
中图分类号: TP391; S512 **文献标志码:** A **文章编号:** 1001-411X(2023)04-0613-06

WOA-BP rice yield prediction based on Spark

CHEN Zhihao✉, WANG Jianhua✉, LONG Yongbing, LAN Yubin, LIU Junhe, XIONG Hongyi, XIAO Fangjun, XIAO Yiming
(College of Electronic Engineering (College of Artificial Intelligence), South China Agricultural University/National Center for International Collaboration Research on Precision Agricultural Aviation Pesticides Spraying Technology/
Guangdong Laboratory for Lingnan Modern Agriculture, Guangzhou 510642, China)

Abstract: 【Objective】With the rapid development of big data technology and artificial intelligence, aiming at the problems of low accuracy, too large prediction area, too long model optimization time of the current rice yield prediction model, etc., a whale optimization algorithm-backpropagation (WOA-BP) rice yield prediction method based on Spark was proposed.【Method】This paper took rice yield and weather data of counties/cities/districts in the western region of Guangdong Province as the research object, used WOA to

收稿日期: 2022–07–04 **网络首发时间:** 2023–05–11 09:32:34
首发网址: https://kns.cnki.net/kcms/detail/44.1110.S.20230510.1127.002.html
作者简介: 陈志浩, 硕士研究生, 主要从事农业人工智能与大数据处理研究, E-mail: 1217405445@qq.com; 通信作者: 王建华, 副教授, 博士, 主要从事农业人工智能与大数据处理、物联网与虚拟现实技术研究, E-mail: jhw655@scau.edu.cn
基金项目: 岭南现代农业实验室资助项目 (NT2021009); 广东省基础与应用基础研究基金 (2021A1515011514); 高等学校学科创新引智计划 (D18019); 中新国际联合研究院项目 (No.206-A021006); 广东省重点领域研发计划 (2019B020214003)

optimize the weights and bias values of BP neural network, and constructed a rice yield prediction model to improve the prediction accuracy. In addition, the WOA-BP algorithm was parallelized in the Spark framework to reduce the algorithm time overhead.【Result】In terms of model accuracy, by comparing the prediction results after inverse normalization, the mean absolute percentage error of the BP neural network model optimized by WOA decreased from 8.354% to 7.068%, and the mean absolute error decreased from 31.320 kg to 26.982 kg, the root mean square error dropped from 41.008 kg to 33.546 kg. In terms of run time, 3-node Spark cluster reduced runtime by 11 742 s over non-Spark mode, reducing time overhead by 44%.【Conclusion】The WOA-BP rice yield prediction method based on Spark can better predict rice yield in western Guangdong counties/cities/districts, and at the same time can well reflect the influence of weather factors on rice yield in western Guangdong Province, which is a reference for studying the rice yield situation in western Guangdong counties/cities/districts and even the whole Guangdong.

Key words: Weather factor; Rice yield; Backpropagation neural network (BP); Whale optimization algorithm (WOA); Spark

水稻作为人类主要粮食之一，受众十分广泛，遍布亚、欧、非洲以及热带美洲，全球约一半的人口以稻米作为主食，因此水稻的产量问题一直备受关注，水稻产量预测也成为当前水稻生产中的一个重要研究方向。当前作物估产方面，主要有气象产量预测法、遥感技术和统计动力学模拟法[1]，通常使用多元线性回归、决策树、神经网络等构建模型；水稻产量受多种因素影响，如气候、病虫害、农药化肥使用量等，导致产量数据呈现非线性分布，预测效果整体较差。

现今我国的水稻主要有早、中、晚 3 种水稻，水稻的分布位置与气候条件密切相关，光照、温度、风向、水分等因素的变化会影响水稻的生长，进而影响水稻的产量。比如，气温变化会对水稻花器官分化、发育以及水稻同化物合成、累积、转运及分配过程产生影响；水稻在孕穗期和灌浆期对水分变化最为敏感，在这期间水稻缺水会阻碍分蘖穗的形成，并影响谷粒的灌浆充实程度；水稻从孕穗期到出穗期叶面积较大，蒸腾强度达到高峰，蒸发量过大会对水稻生长造成影响；水稻属喜阳短日照作物，光照强度直接影响水稻同化物的形成速率，进而影响产量[2-5]。当前，在气象估产方面，国内外学者已进行了相关的研究，比如，刘洪英等[6] 利用四川省南充市 1989—2018 年气象数据和水稻单产数据，采用线性回归方法建立了基于气象因子的水稻产量预报模型；高俊杰等[7] 利用 1982—2020 年广东省肇庆市高要区气象因子与早稻产量的数据，采用逐步线性回归方法建立了早稻产量预报模型；Chutia 等[8] 利用 1990—2012 年水稻作物产量数据和周天气数据，建立了阿萨姆邦 13 个地区的水稻产量预测模型；Kaeomuangmoon等[9] 通过研究泰国 77 个区域的气候数据变化，利用 Rice4cast 平台预测季节性 KDML 105 水稻的产量；Traore 等[10] 使用决策分析针对萨赫勒地区气候条件进行水稻估产；Jha 等[11] 通过作物动态模型根据每日气象数据对尼泊尔水稻产量进行估产；Dhekale 等[12] 针对印度克勒格布尔市日降雨量数据，采用 CERES-Rice(DSSATv4.5) 模型进行水稻产量预测；Nain 等[13] 针对印度哈里亚纳邦卡尔纳尔地区的气象及水稻产量数据，使用多元线性回归等不同统计方法对该地区的水稻产量进行预测；Guo等[14] 通过气象和水稻产量等农艺性状数据，分别使用反向传播神经网络和偏最小二乘法构建模型，预测华东地区的水稻产量；杨北萍等[15] 通过长春市 2 个地区的气象、水稻遥感及产量数据，使用随机森林算法对 2 个地区的水稻产量进行预估；徐强强等[16] 通过浙江省台州市椒江区的气象及水稻产量数据，使用指数平滑法对该地区早稻产量进行预测。其他作物方面，路智渊等[17] 通过气象因子结合固原市小麦产量进行回归分析，进行小麦产量预测；马凡[18] 基于气象数据及安徽省小麦产量，构建小麦产量预测模型。以上方法不同程度地存在模型精度低、预测区域级别过大、模型优化时间过长等缺陷，如模型的误差超过 10%，预测区域的级别为国家或省市，使用群智能算法等优化神经网络时间过长等。为了解决上述问题，本文提出一种基于 Spark 的鲸鱼优化算法−反向传播神经网络 (Whale optimization algorithm-backpropagation，WOA-BP) 水稻产量预测方法。首先，以县/市/区作为研究区域级别，避免研究区域范围过大和数据量太少的问题，可以很好地反映气象因素对县/市/区

级别水稻产量的影响，在研究小区域水稻产量时更具有参考意义；此外，BP 神经网络具有优良的非线性映射能力，利用其构建水稻产量模型能够提升模型的预测效果，同时利用 WOA 对 BP 网络的权值和偏置值进行优化，改善 BP 神经网络收敛慢、易局部收敛等缺陷，能够进一步提升模型的效果，避免误差较大等问题；最后，将现有的大数据技术与农业和人工智能进行结合，利用大数据 Spark 框架，搭建 Spark 集群，将改造后的 WOA-BP算法在集群环境下实现并行化运算，减少算法优化过程的时间开销，充分发挥大数据技术的优势，实现对水稻产量与气象数据的快速建模以及县/市/区水稻产量的精准预测。

1　材料与方法

1.1　试验环境

模型的训练在 TensorFlow 框架下完成，优化算法在 Spark 集群下运行，其中 Spark 集群由 3 台相同配置的联想台式电脑组成，硬件环境：联想 3148 主板、AMD Ryzen5 3 600 6-Core 双线程 CPU、16 GB DDR4 3 000 MHz 内存、TP-LINK 路由器，软件环境：Ubuntu16.04 系统、TensorFlow2.8、Spark3.2.0、Python3.7，编程语言为 Python；通过路由器将 3 台电脑构成局域网，按照 1 主节点 2 子节点搭建 Spark 集群环境，Spark 集群模式为 Standalone 模式。

1.2　数据来源与处理

本文以广东省西部地区 4 座城市 (湛江、茂名、阳江、云浮)23 个区县 2000—2020 年水稻的单产 (每 667 m^2) 数据及该地区的气象因素作为研究对象，其中，该区域的水稻单产数据共计 482 条，数据来源于广东省统计局历年的《广东农村统计年鉴》；气象因素选取 2000—2020 年每年 3—10 月的每月气温 (最高、最低、平均)，土温 (最高、最低、平均)，露点温度 (最高、最低、平均)，积温，降水量，蒸发量和太阳辐射量，来源于欧洲中期天气预报中心 (ECMWF) 的气象数据。为了降低后期 BP 神经网络模型结构的复杂程度，通过主成分分析 (Principal component analysis，PCA) 对影响因素进行降维，降维后累积方差贡献度保持在 0.95 以上；此外为了加快 BP 神经网络的收敛，需对数据进行归一化处理，归一化公式如下：

$$X' = \mathrm{MIN} + \frac{X - X_{\min}}{X_{\max} - X_{\min}} (\mathrm{MAX} - \mathrm{MIN}), \qquad (1)$$

式中，X 为当前元素，X'为 X 归一化后的值，MIN、MAX 分别为 X 整体数据集中所有元素的最小值和最大值，$X_{\min}$与$X_{\max}$为当前所在列的最小值与最大值。

1.3　BP 神经网络和 WOA 算法

BP 神经网络是 1986 年由 Rumelhart 等[19] 提出，是一种按误差逆传播算法训练的多层前馈网络，是目前应用最广泛的神经网络之一。BP 神经网络学习过程分为输入信息正向传播和误差反向传播 2 个阶段[20]。

WOA 算法是由澳大利亚学者 Mirjalili 等[21] 于 2016 年提出的新型群智能寻优算法，该算法主要分为座头鲸识别并包围猎物、螺旋泡网攻击、鲸鱼根据同类位置随机搜索捕食 3 个阶段；WOA 算法已经被运用到复杂函数优化、路径规划、图像分割和光伏模型等领域，并取得显著效果[22]。

1.4　基于 Spark 的 WOA-BP 算法并行化

Spark 是一个基于内存运算的大数据计算框架，在时间性能上优于 MapReduce，Spark 为了能够实现高并发和高吞吐率的数据处理过程，封装了弹性分布式数据集 (Resilient distributed datasets，RDD)、累加器、广播变量 3 大数据结构，应对不同场景下数据处理，其中，RDD 是 Spark 中最基本的数据单元，同时也是 1 个不可变、可分区且支持并行计算的数据集合。RDD 用于支持 Spark 框架的并行计算，而累加器以及广播变量则用于数据同步，其中累加器是 1 个只写变量，变量一旦被修改，该变量在所有节点的值将同步更新；广播变量是 1 个只读变量，当变量被广播后，成为该节点的局部变量，节点修改该变量不会影响其他节点[23-24]。

由于 WOA 算法优化 BP 神经网络时，存在大量迭代计算，除鲸鱼自身的信息不一样外，每头鲸鱼在寻找自身最优解以及更新自身位置信息的过程中，所有更新逻辑均相同，因此，结合 Spark 并行计算框架，实现基于 Spark 的 WOA-BP 算法并行化，减少算法的时间开销。图 1 为基于 Spark 的 WOA-BP 算法的并行化流程图，算法的具体步骤如下。

1) 设置相关参数：设置种群规模，如鲸鱼数量 n，参数维度 d 等，同时设置 Spark 广播变量。

2) 初始化种群：创建含 d 个元素的一维零数组，通过该数组构建 RDD，之后通过 map 算子进行种群的初始化，实现并行化初始化操作，减少时间开销。

3) 更新鲸鱼位置和适应度：更新每头鲸鱼的位置信息，并进行越界检查，之后计算该鲸鱼的适应度 (Fitness)，以样本的均方根误差 (Root mean

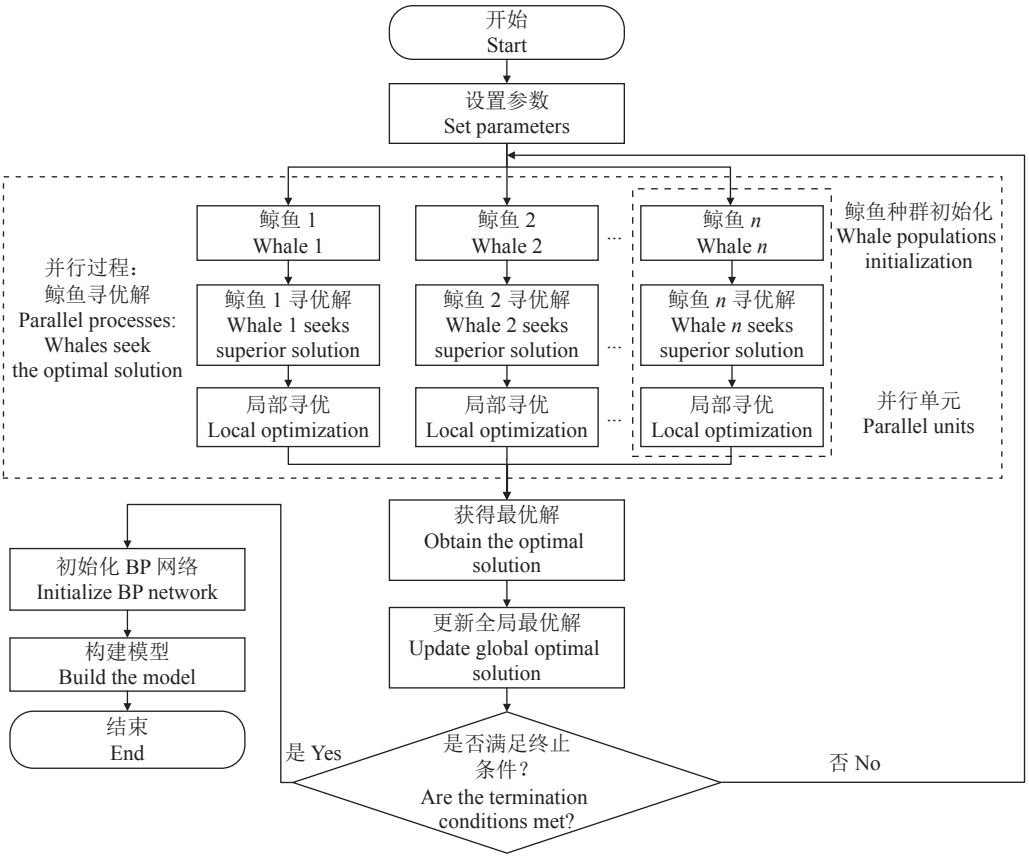

图 1　WOA-BP 算法的并行化流程

Fig. 1　The parallelization process of the WOA-BP algorithm

square error，RMSE) 作为适应度，计算公式如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}-y'_{ij}\right)^2}, \tag{2}$$

式中，n 为样本数，m 为网络输出层输出个数，y_{ij}为样本的实际值，y'_{ij}为网络的实际输出值。

4) 更新全局最优解和最小适应度：通过 sortBy 算子获取最小适应度以及该适应度对应的鲸鱼位置信息，更新全局最优解。

5) 终止条件判断：若不满足终止条件，则程序继续执行，否则，通过 collect 算子收集各个分区的数据，完成算法的优化阶段，得到全局最优解。

6) 构建 BP 神经网络：利用全局最优解对网络的权值和偏置值进行初始化，构建模型。

2　结果与分析

2.1　基于 WOA-BP 水稻产量预测

本文以广东省西部地区 2000—2020 年县/市/区水稻单产及气象数据为基础，按照 3∶1∶1 进行数据集划分：2000—2012 年数据作为训练集，2013—2020 年数据作为验证集 (50%) 和测试集 (50%)，通过 BP 神经网络建模，分别使用粒子群优化算法 (Particle swarm optimization，PSO) 和 WOA 对 BP 神经网络进行优化，得到 BP、PSO-BP、WOA-BP 3 种产量预测模型，之后对模型的预测结果进行反归一化。图 2 是 3 种模型预测值与真实值的绝对误差对比，由图 2 可以清晰看出，WOA-BP 模型的曲线整体上更加贴近横坐标，即测试集样本的整体绝对误差小于另外 2 种模型的。表 1 为 3 种模型的预测精度对比，可以明显看出，与传统 BP 模型相比，经 WOA 优化后的产量预测模型的平均绝对百分比误差 (Mean absolute percentage error，MAPE) 减少了 1.286 个百分点，平均绝对误差 (Mean absolute error，MAE) 减少了 4.338 kg，RMSE 减少了 7.462 kg。虽然 PSO-BP 模型相较传统 BP 模型在精度上有一定提升，但效果明显不如 WOA-BP。此外，试验过程中发现，相同种群规模下，WOA 与 PSO 2 种算法的优化时间相差较大，其中 WOA 为 26 637 s，PSO 为 48 518 s，WOA 比 PSO 少了约 45% 的时间开销，显然 WOA 的时间性能更优。因此，WOA 在算法优化的时间开销以及模型效果上均优于 PSO，故本文采用 WOA-BP 对广东省西部地区县/市/区的水稻产量进行最终建模。

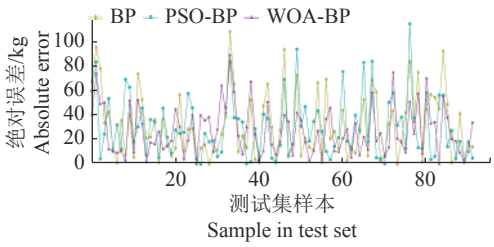

图 2　3 种模型的绝对误差

Fig. 2　Absolute error of the three models

表 1　3 种模型精度对比

Table 1　Precision comparison of the three models

模型 Model	平均绝对百分比误差/% MAPE	平均绝对误差/kg MAE	均方根误差/kg RMSE
BP	8.354	31.320	41.008
PSO-BP	7.890	29.999	38.786
WOA-BP	7.068	26.982	33.546

2.2　基于 Spark 的 WOA-BP 算法时间性能

由表 1 和图 2 的结果可知，经 WOA-BP 算法得到的预测模型效果最佳，但算法的优化时间仍旧较长，故在此基础之上，结合 Spark 并行计算框架，减少优化过程的时间开销。因此，使用 3 台台式主机按照 1 主节点 2 子节点的形式搭建 Spark 集群，同时改造 WOA-BP 算法实现并行化，并按照 2 倍物理核数的规则对 RDD 进行分区，提升集群整体的并行度，充分利用 CPU 性能。表 2 为不同节点性能对比及配置信息，图 3 为不同节点算法运行时间对比，由表 2 和图 3 可以清晰看出，随着节点数量的增加，算法的优化时间随之减少，其中 3 节点比 2 节点和 1 节点分别减少了 21.4% 和 39.3% 的时间开销，大幅度缩短算法的优化时间。同时与“2.1”中非 Spark 的 WOA 的优化时间相比，减少了 44% 的时间开销，充分体现算法与 Spark 框架结合后的优势，真正实现对水稻产量与气象数据的快速建模。

表 2　不同节点数量性能对比及配置信息

Table 2　Performance comparison and configuration information under different node number

节点数量 Node number	总内存/G Total memory	总物理核数 Total physical nuclei number	分区数量 Partition number	*t*/s
1	16	12	24	24 534
2	32	24	48	18 955
3	48	36	72	14 895

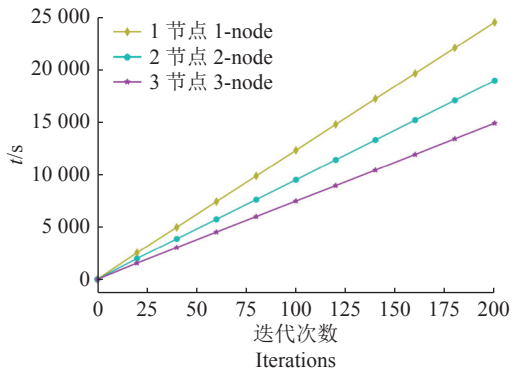

图 3　不同节点数时间开销对比

Fig. 3　Time overhead comparison under different node number

3　结论

本文以广东省西部地区所有县/市/区作为研究区域，针对气象因素对水稻单产的影响，提出一种基于 Spark 框架的 WOA-BP 水稻县/市/区级别的单产预测方法。首先通过 WOA 对 BP 神经网络进行优化，避免 BP 神经网络收敛慢、易局部收敛等缺陷，提升 BP 模型的整体预测精度；其次，结合 Spark 并行计算框架，实现 WOA-BP 算法并行化，加快 WOA-BP 算法的运算速度，减少算法的时间开销；最后通过 WOA-BP 算法得到的最优解对网络进行初始化并构建网络模型，之后进行水稻单产的预测。测试集的预测结果表明，该模型的预测精度较高，预测结果较精确，论证了该方法的可行性及有效性；同时，该模型可以很好地反映气象因素对广东省西部地区县/市/区水稻单产的影响情况，对研究广东西部县/市/区乃至整个广东的水稻单产具有一定的借鉴意义。

参考文献：

[1] 李晔, 白雪. 基于新维无偏灰色马尔可夫模型的小麦产量预测[J]. 江苏农业科学, 2021, 49(15): 181-186.

[2] 韩芳玉, 张俊飚, 程琳琳, 等. 气候变化对中国水稻产量及其区域差异性的影响[J]. 生态与农村环境学报, 2019, 35(3): 283-289.

[3] 闫蓉, 李凤霞, 赵维忠, 等. 气象条件对水稻蒸腾速率的影响[J]. 宁夏农林科技, 2005(2): 7-8.

[4] 杨从党, 朱德峰, 周玉萍, 等. 不同生态条件下水稻产量及其构成因子分析[J]. 西南农业学报, 2004(S1): 35-39.

[5] 焦江华. 不同土壤有机碳含量下气象因子主导的水稻产量模拟及模型改进[D]. 北京: 中国农业科学院, 2020.

[6] 刘洪英, 鲜铁军, 李睿, 等. 基于气象因子的水稻产量预报模型[J]. 陕西气象, 2020(5): 45-47.

[7] 高俊杰, 袁业溶, 梁应. 高要区早稻产量预测模型的建立[J]. 广东气象, 2022, 44(2): 50-52.

[8] CHUTIA S, DEKA R L, GOSWAMI J, et al. Forecast-

ing rice yield through modified Hendrick and Scholl technique in the Brahmaputra valley of Assam[J]. Journal of Agrometeorology, 2021, 23(1): 106-112.

[9] KAEOMUANGMOON T, JINTRAWET A, CHOTAMONSAK C, et al. Estimating seasonal fragrant rice production in Thailand using a spatial crop modelling and weather forecasting approach[J]. Journal of Agricultural Science, 2020, 157(7/8): 566-577.

[10] TRAORE S, ZHANG L, GUVEN A, et al. Rice yield response forecasting tool (YIELDCAST) for supporting climate change adaptation decision in Sahel[J]. Agricultural Water Management, 2020, 239: 106242. doi: 10.1016/j.agwat.2020.106242.

[11] JHA P K, ATHANASIADIS P, GUALDI S, et al. Using daily data from seasonal forecasts in dynamic crop models for yield prediction: A case study for rice in Nepal's Terai[J]. Agricultural and Forest Meteorology, 2018, 265: 349-358.

[12] DHEKALE B S, NAGESWARARAO M M, NAIR A, et al. Prediction of kharif rice yield at Kharagpur using disaggregated extended range rainfall forecasts[J]. Theoretical and Applied Climatology, 2018, 133(3/4): 1075-1091.

[13] NAIN G, BHARDWAJ N, JASLAM P K M, et al. Rice yield forecasting using agro-meteorological variables: A multivariate approach[J]. Journal of Agrometeorology, 2021, 23(1): 100-105.

[14] GUO Y, XIANG H, LI Z, et al. Prediction of rice yield in East China based on climate and agronomic traits data using artificial neural networks and partial least squares regression[J]. Agronomy, 2021, 11(2): 282. doi: 10.3390/agronomy11020282.

[15] 杨北萍, 陈圣波, 于海洋, 等. 基于随机森林回归方法的水稻产量遥感估算[J]. 中国农业大学学报, 2020, 25(6): 26-34.

[16] 徐强强, 王旭辉. 指数平滑法在椒江区早稻产量预测中的应用研究[J]. 上海农业科技, 2021(4): 22-24.

[17] 路智渊, 顾娟, 龚小丽, 等. 固原市冬小麦产量预报与气象条件分析[J]. 现代农业, 2021(5): 111-112.

[18] 马凡. 基于气象数据的安徽省冬小麦产量预测模型研究[D]. 合肥: 安徽农业大学, 2020.

[19] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back propagating Errors[J]. Nature, 1986, 323(6088): 533-536.

[20] 苏博, 刘鲁, 杨方廷. GM(1, N) 灰色系统与 BP 神经网络方法的粮食产量预测比较研究[J]. 中国农业大学学报, 2006(4): 99-104.

[21] MIRJALILI S, LEWIS A. The Whale Optimization Algorithm[J]. Advances in Engineering Software, 2016, 95: 51-57.

[22] 高岳林, 杨钦文, 王晓峰, 等. 新型群体智能优化算法综述[J]. 郑州大学学报 (工学版), 2022, 43(3): 21-30.

[23] 翟光明, 李国和, 吴卫江, 等. 基于 Spark 的人工蜂群改进算法[J]. 计算机应用, 2017, 37(7): 1906-1910.

[24] 王诏远, 王宏杰, 邢焕来, 等. 基于 Spark 的蚁群优化算法[J]. 计算机应用, 2015, 35(10): 2777-2780.

【责任编辑 李庆玲】