基于遗传算法组合 XGBoost 模型的开采沉陷预测

Prediction of Mining Subsidence Based on Genetic Algorithm Combined with XGBoost Model

韩连昌（贵州盘江精煤股份有限公司金佳矿，贵州 六盘水 553000）

摘　要：为了准确预测矿山开采引起的地表沉陷，引入遗传算法对 XGBoost 模型进行优化，并运用 python 程序语言开发了 GA-XGBoost 组合模型。首先通过随机初始化 XGBoost 的超参数向量，经过训练与测试得到模型的预测误差，通过 GA 对 XGBoost 模型进行优化，最终得到性能最佳的 XGBoost 模型，并对国内 78 例煤矿开采沉陷数据进行预测。预测结果表明：GA-XGBoost 模型预测结果的 R^2（决定系数）为 0.931 8，RMSE（均方根误差）为 0.398 9，MAE（平均绝对误差）为 0.298 9。与单一的 XGBoost、随机深林以及 Gradient Boost 等集成学习模型相比，GA-XGBoost 模型开采沉陷预测精度更高。

关键词：开采沉陷；XGBoost；遗传算法；集成学习

中图分类号：TD823　　**文献标志码**：A　　**文章编号**：1672-609X(2025)02-0009-06

Abstract: The XGBoost ensemble learning algorithm has excellent performance in solving complex nonlinear relationship problems. In order to accurately predict the surface subsidence caused by mining, the genetic algorithm was introduced to optimize the XGBoost model, and the GA-XGBoost combined model was developed using the python programming language. Firstly, the hyperparameter vector of XGBoost is randomly initialized, the prediction error of the model is obtained after training and testing, and the XGBoost model is optimized by GA, and finally the XGBoost model with the best performance is obtained, and 78 domestic coal mining subsidence data are predicted. The prediction results show that the R^2 (coefficient of determination) of the prediction results of the GA-XGBoost model is 0.931 8, the RMSE (root mean square error) is 0.398 9, and the MAE (mean absolute error) is 0.298 9. Compared with ensemble learning models such as single XGBoost, random deep forest, and Gradient Boost, the GA-XGBoost model has higher mining subsidence prediction accuracy.

Key words: mining subsidence; XGBoost; genetic algorithm; ensemble learning

1　前言

煤矿开采沉陷不仅容易引起地表损坏，建筑垮塌、公路、铁路损毁而且容易诱发滑坡等地质灾害以及地表水流失等一系列安全与环境问题，开采沉陷问题已经严重制约了矿山的可持续发展，为社会与环境带来了严重的影响[1-2]。通过提前预计地表下沉的基本数据信息，不仅可以指导矿区生产与灾害防治，还可以通过预先优化开采方案以及处理对策降低开采损害，缓和资源开发利用与环境保护以及社会公共安全之间的矛盾。矿山开采沉陷的研究一直是矿业科技人员的研究热点，近年来矿山开采沉陷常用的研究方法主要有相似模拟、数值模拟等方法，相似模拟可模拟开挖引起的围岩的应力分布、覆岩破断以及地表沉陷规律，在研究采矿等复杂问中具有直观、简便等特点，为工程应用与科学研究提供有价值的参考[3-5]。数值模拟能够模拟开采过程中复杂的动态变化过程，成本低、可快速获取结果，在采矿领域也得到了广泛的应用[6-7]。当前我国开采沉陷的主要预计方法是概率积分法，由于其理论丰富，计算方便等特点而应用广泛。然而由于开采条件、地质条件复杂，概率积分法预计存在一定的误差，大量的学者围绕概率积分法的预测精度做了很多有益的研究[8-9]。近年来，人工智能方法成功应用于岩体工程领域[10-11]，越来越多的矿业科技人员引入用人工智能技术解决复杂的开采沉陷预测问题，并取得了丰硕的研究成果[12-14]。矿山开采沉陷的研究方法较多，每种方法都具有一定的优势，然而在大多数情况下，决策者更关注的是模型预测结论与实际观测数据的误差，而不是某种单一的研究方

[作者简介] 韩连昌(1994—)，男，云南宣威人，研究生学历，工程师，主要从事采矿工程。

[引用格式] 韩连昌. 基于遗传算法组合 XGBoost 模型的开采沉陷预测[J]. 中国矿山工程,2025,54(2):9-14.

法或模型,因为每种方法和模型都有各自的优缺点。因此,一些理论原理相对简单,模型开发方便,具有一定通用性和可靠性的分析方法和模型往往具有更好的辅助决策功能。鉴于以上分析,本文选取了经典的 XGBoost 集成算法用于处理开采沉陷预测的高度非线性问题,同时针对 XGBoost 超参数多,模型复杂等问题引入了遗传算法(GA)对 XGBoost 的超参数进行优化,建立了预计精度自适应调整的 GA-XGBoost 模型,提高了单一 XGBoost 模型的泛化能力与预测精度。本文提出的组合模型方法能帮助科技人员使用人工智能方法预测开采沉陷问题,为开采计划的定制和治理方案的优化提供参考。

2 基于 GA 优化的 XGBoost 模型

2.1 XGBoost 基本原理

XGBoost 是一类典型的集成学习算法,其基本思路是组合若干个弱评估器,并且通过不断拟合单一弱评估器的预测残差,进而形成一个强评估器用于预测分析[15]。常见 XGBoost 集成的弱学习器是树模型,在建立目标函数时同时考虑了传统的损失函数与模型复杂度,并且增加了正则项用于防止过拟合,因此 XGBoost 具备高效的运算性能和预测精度[16]。由于 XGBoost 在处理非线性问题上表现出优秀的性能与泛化能力,目前已经广泛应用于机器学习领域,用于分类、回归等任务,其目标函数为[17]

$$Obj^{(r)} = \sum_{i=1}^{m} L(y_i, \hat{y}_i^{(r)}) + \sum_{k=1}^{r} \Omega(g_r) \quad (1)$$

式中,i 表示数据集的第 i 个样本;m 表示导入第 k 棵树的导入数据总量;r 表示建立的所有树;y_i 表示真实值;$\hat{y}_i^{(r)}$ 表示第 r 次预测的预测值;g_r 表示树模型的结构项;$L(y_i, \hat{y}_i^{(r)})$ 表示模型损失函数;$\Omega(g_r)$ 表示模型的复杂度。

模型的复杂度公式为

$$\Omega(g_r) = \gamma T + \frac{1}{2}\lambda \sum_{j}^{T} w^2 \quad (2)$$

式中,T 表示树的叶子节点数;w 表示叶子权重;γ 为控制叶子数量的参数;λ 为正则向系数。XGBoost 在考虑传统误差函数的基础上,兼顾了模型复杂度影响,在数学原理上对模型的泛化误差进行了优化,在模型的方差 – 偏差困境中寻求平衡点,以求模型的泛化误差最小,运行速度最快。XGBoost 已经被认为是在分类和回归问题上拥有极高性能的评估器。但是由于 XGBoost 算法复杂,超参数多,模型的预测性能直接受到模型超参数影响,在工程应用中受到极大限制。然而,超参数的优化与分析人员对数据物理意义的理解和分析经验密切相关,通常情况下的人工调参,效率低下且往往难以快速优化模型,难以满足工程应用要求。

2.2 Genetic Algorithm(GA)基本原理

遗传算法(GA)是一种启发式的群体搜索算法,其主要思路是参考大自然中生物的“优胜劣汰,适者生存”的法则,借助于计算机程序模拟生物种群染色体的选择、复制、交叉、变异等操作,以群体的方式进行淘汰搜索,最终在解空间范围内搜索到最优解[18]。GA 首先将潜在的可能解当做独立的个体,将个体进行编码形成初始的种群,然后通过逐代进化,以产生近似最优解,每一代的遗传操作中根据个体适应度的值进行选择操作,并借鉴遗传学的交叉和变异操作产生新一代的种群。整个过程模拟自然界的进化过程,后代种群更加适应环境,逐渐优于前代,最后将后代最优个体经过解码进而得到问题的最优解[19]。GA 通用性强、操作简单,已经被广泛应用于众多领域的优化问题[20],GA 的基本运行流程如图 1 所示。

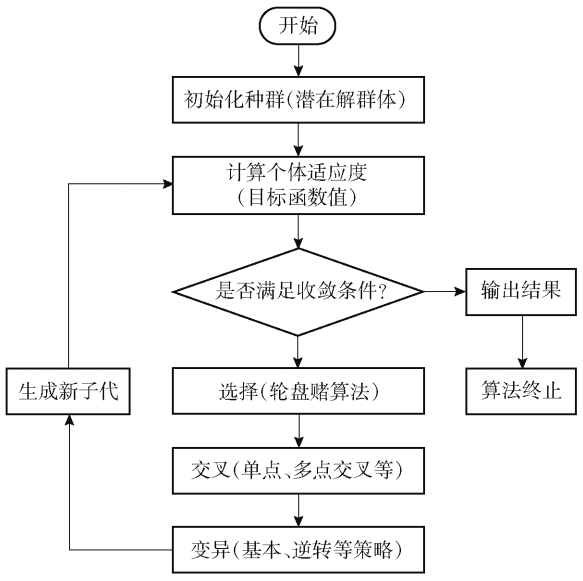

图 1 遗传算法基本流程

2.3 GA-XGBoost 模型

鉴于上述存在的问题,本文针对开采沉陷的非线性回归预测问题,以 XGBoost 算法为基础,同时引用了 GA 对 XGBoost 的超参数进行优化,建立了开采沉陷预测的 GA-XGBoost 预测模型。首先将数据集分割为训练集与测试集,其中训练集用于训练模

型,学习数据集中特征与标签之间的非线性关系,测试集用于评估模型的预测性能。然后随机初始化 XGBoost 的超参数向量作为个体并进行编码,并以 XGBoost 的预测误差作为目标函数,通过遗传算法对超参数向量进行进化迭代,搜索解空间范围内的最佳超参数组合,GA-XGBoost 模型的工作流程图如图 2 所示。

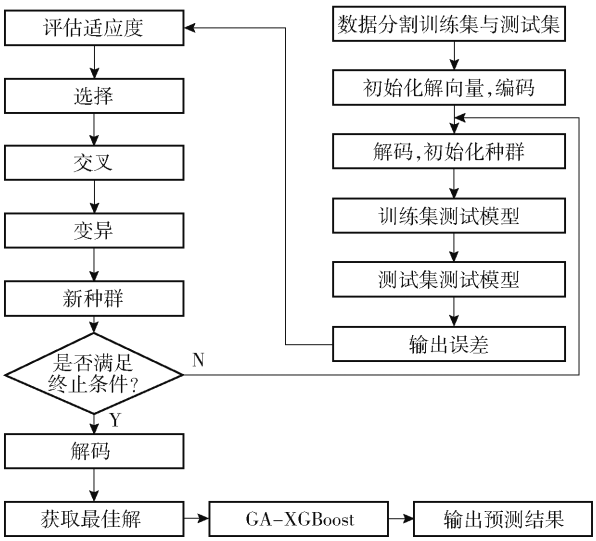

图 2　GA-XGBoost 优化流程图

在开展遗传优化前,设定 GA 基本参数如种群数、迭代次数、变异概率等,同时定义优化目标解的维数、各分量范围,以及相关的目标函数。采用遗传算法优化 XGBoost 模型的关键步骤在于通过 GA 选择最优 XGBoost 超参数的解向量,GA 在 XGBoost 训练过程中,对 XGBoost 的训练实现了自优化,提高了学习效率,同时通过其全局优化及隐含并行性的优点提高了 XGBoost 超参数的优化速度。

3　GA-XGBoost 模型预测开采沉陷

为了全面验证遗传算法(Genetic Algorithm, GA)优化极端梯度提升(Extreme Gradient Boosting, XGBoost)模型在矿山开采沉陷预测中的可行性和有效性,本文系统性地收集了国内众多学者在矿山开采沉陷领域的研究成果,精心筛选出 78 例具有代表性的实际案例作为研究数据集。这些案例涵盖了不同地质条件、开采方式及规模的矿山,确保了数据的多样性和实用性,为模型训练和验证提供了坚实的基础。

研究的核心在于探究开采下沉量与采矿活动参数及地质特征之间的复杂关系。具体而言,我们选择了四个关键特征作为输入变量:采深(H),它反映了开采活动对地下岩层的扰动深度;采高(m),直接影响开采后上覆岩层的垮落程度和地表沉陷的范围;采空区面积(s),是衡量开采规模的重要指标,与地表沉陷的广度和强度密切相关;表土松散层厚度(h),这一地质因素显著影响着沉陷传播至地表的效率和形态。作为因变量,我们关注的是最大下沉值(w),它是评估矿山开采对地表影响程度最直接且重要的指标之一。通过对这一指标的准确预测,可以为矿山安全评估、土地复垦规划及环境保护措施提供科学依据。

在模型构建与优化方面,本文创新性地引入了遗传算法(GA)来优化 XGBoost 模型的超参数配置。XGBoost 作为一种高效的梯度提升决策树算法,在分类和回归任务中展现出强大的性能,但其性能高度依赖于一系列超参数的设置,包括树棵树(num_round)、树深度(max_depth)、学习率(eta)、λ 值。

3.1　数据集

在数据处理和机器学习项目中,数据集分割是至关重要的一步。它不仅有助于确保模型在未见过的数据上具有良好的性能,还有助于避免过拟合和欠拟合等问题。通过将数据集分为训练集和测试集,我们可以分别评估模型在训练数据和测试数据上的表现,从而更全面地了解模型的性能。在本例中,我们选择将 80% 分割为训练集用于训练模型,并用剩余 20% 数据进行模型测试。局部数据集的格式见表 1。

3.2　预测结果

在本文的研究中,笔者探讨了遗传算法(Genetic Algorithm, GA)作为一种先进的优化策略,在提升极端梯度提升(eXtreme Gradient Boosting, XGBoost)模型性能方面的应用与效果。为了全面评估 GA 算法对 XGBoost 模型优化的效能,设计了实验,其中核心变量为遗传算法中的种群数(Population Size)。种群数作为 GA 算法中的一个关键参数,直接影响了搜索空间的覆盖广度及算法的探索与利用能力,进而对最终优化结果产生显著影响。

具体而言,我们设置了多个不同的种群数水平进行对照分析,包括 10、20、30、40、50、60、70 及 80,旨在探究不同种群规模下 GA 算法对 XGBoost 模型超参数优化的动态过程及效果差异。通过这一系列的实验设置,我们期望能够揭示种群数对 GA 算法搜索效率、全局最优解发现能力以及模型最终性能

表 1 模型采用数据集(局部)

序号	采深/m	采厚/m	采空区面积/m²	松散层厚度/m	最大下沉量/m
1	224	1. 8	102 300	8. 1	1. 16
2	318	1. 6	70 224	2. 7	1. 06
3	81	1. 5	17 220	3. 2	0. 92
4	60	2. 1	47 760	10	1. 20
5	150	1. 85	46 250	8	1. 00
6	227	2	39 750	32	0. 48
7	264	4. 9	107 250	14	1. 31
8	120	1. 6	37 960	24	0. 82
9	606	8. 05	176 000	8	4. 95
10	253	2	98 000	3	1. 21
11	123. 2	2. 2	14 186	5. 8	0. 93
12	225	3	5 550	7	0. 59
13	125	25	34 726	7	10. 40
14	158. 76	3. 78	61 944	7	2. 29
15	49. 92	0. 96	131 320	7	2. 29
16	33. 6	0. 96	52 020	20	0. 83
17	61. 92	1. 44	76 590	20	0. 99
18	94. 6	2. 15	86 010	20	1. 26

提升的潜在影响机制。

随着 GA 算法的迭代推进,我们观察到 XGBoost 模型的误差呈现出一种系统性的下降趋势。这一现象充分展示了 GA 算法在超参数空间内进行有效搜索并逐步逼近最优配置的能力。每一次 GA 迭代都代表了一次在给定种群内基于适应度函数(在此为 XGBoost 模型的误差)的选择、交叉和变异操作,这些操作共同驱动着种群向更优的超参数组合进化。

图 3 直观地展示了在不同种群数设置下,遗传迭代过程中模型误差的变化规律。从图 3 可以清晰地看到,随着迭代次数的增加,所有种群数配置下的模型误差均呈现逐步减小的趋势,这验证了 GA 算法在优化 XGBoost 模型超参数方面的有效性。同时,不同种群数下的误差下降曲线呈现出一定的差异性,具体而言,种群数较大的实验组(如 60、70、80)往往能在更早的迭代阶段实现更快的误差下降速度,这可能是因为较大的种群数提供了更多的遗传多样性,有助于算法在更广泛的搜索空间内快速定位到较优的解区域。然而,随着迭代深入,所有实验组之间的误差差异逐渐缩小,表明在足够的迭代次数后,GA 算法均能较好地逼近全局最优解,尽管种群数对其收敛速度有所影响,但对最终优化结果的影响趋于稳定。

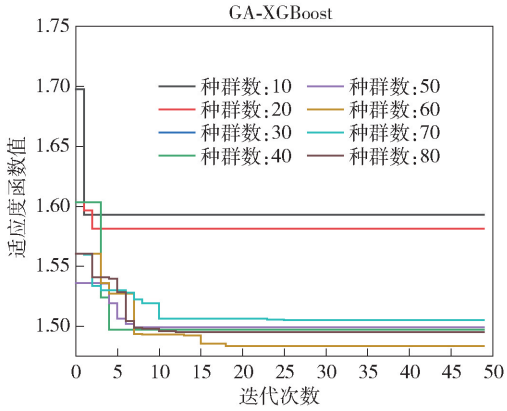

图 3 遗传迭代误差变化规律

通过遗传算法优化后,我们搜索到了最佳的 XGBoost 模型超参数的解向量为[num_round = 137, max_depth = 4, eta = 0. 420 945 01, λ = 5. 923 841 51]。同时我们分别采用单一的 XGBoost 模型,以及随机深林回归(RFR),Gradient Boost 等集成算法与真实数据(REF)进行对比分析,结果如图 4 所示。

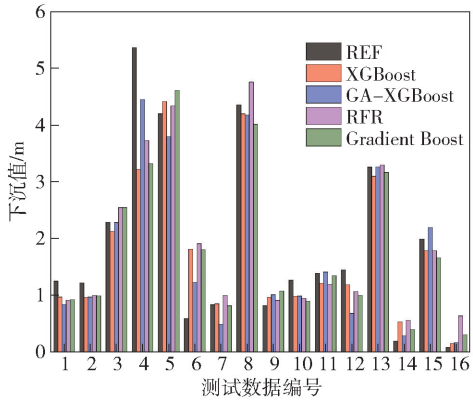

图 4 不同模型预测分析

在图 4 所展示的对比分析中,我们观察到四个不同的预测模型在预测开采过程中的最大下沉值时,均呈现出与真实值之间的偏差。特别是在下沉幅度较大或较小的极端情况下,这些预测误差变得尤为显著,这不仅反映了模型在极端条件下的局限性,也强调了提高预测精度的紧迫性和重要性。为了全面且精确地评估 GA-XGBoost 模型在这一特定应用背景下的预测性能,本文采用了三种广泛认可且各具特色的统计指标:决定系数(R^2)、均方根误差($RMSE$)以及平均绝对误差(MAE)。

决定系数(R^2),作为衡量模型预测值与观测值

之间拟合优度的一个关键指标，其值越接近于 1，表明模型的预测性能越好，即模型能够更准确地捕捉数据中的变化趋势。然而，值得注意的是，即便 R^2 值较高，也可能存在模型在特定区间（如极大或极小下沉值）内预测误差较大的情况，因此需结合其他指标综合判断。

均方根误差（*RMSE*）则通过计算预测值与真实值之间差异的平方的平均值后取平方根，提供了预测误差的量化表示。该指标对极端误差值较为敏感，能够直观地反映出模型在整体预测上的准确性。*RMSE* 值越小，意味着模型的预测结果与实际观测更为接近，预测精度更高。

平均绝对误差（*MAE*）则是预测误差绝对值的平均值，它同样反映了模型预测的准确性，但与 *RMSE* 相比，*MAE* 对误差的惩罚是线性的，因此对极端误差的敏感度较低。这一特性使得 *MAE* 在评估模型预测稳定性的方面具有独特优势，尤其是在处理含有异常值的数据集时。

综上，通过综合运用 R^2、*RMSE* 和 *MAE* 这三种评估指标，从多个维度剖析 GA-XGBoost 模型在预测开采最大下沉值方面的性能表现。

3.3　模型分析

GA-XGBoost 模型预测精度采用决定系数 R^2，均方根误差 *RMSE*，以及平均绝对误差 *MAE* 进行评价，各个指标的数学公式定义如下：

$$R^2 = 1 - \sum_{i=1}^{N} (y_i - y_i')^2 \Big/ \sum_{i=1}^{N} (y_i - \bar{y}_i)^2 \quad (3)$$

$$RMSE = \sqrt{\sum_{i=1}^{N} (y'_i - y_i)^2 / N} \quad (4)$$

$$MAE = \sum_{i=1}^{N} |y - y_i'| / N \quad (5)$$

其中 y_i 是代表真实值，y_i'是模型预测值，$\bar{y}_i$ 是真实值均值，N 是测试样本数量。

图 5 所示是单一的 XGBoost、GA-XGBoost、RFR，Gradient Boost 等模型的预测情况，从图 5 可以看出，由于文中所选模型均为集成算法模型，因此 XGBoost 与 RFR 以及 Gradient Boost 预测性能近似，相关系数 R^2均在 0.8 左右，*RMSE* 在 0.6 左右，*MAE* 在 0.4 左右，作为强大的集成算法，其性能表现是欠佳的。通过 GA 算法优化以后，XGBoost 模型性能得到显著的提升，文中构建的 GA-XGBoost 模型的 R^2 提升到 0.931 8，*RMSE* 降低至 0.398 9，*MAE* 为 0.298 9。GA-XGBoost 模型表现是最佳的，可见利用

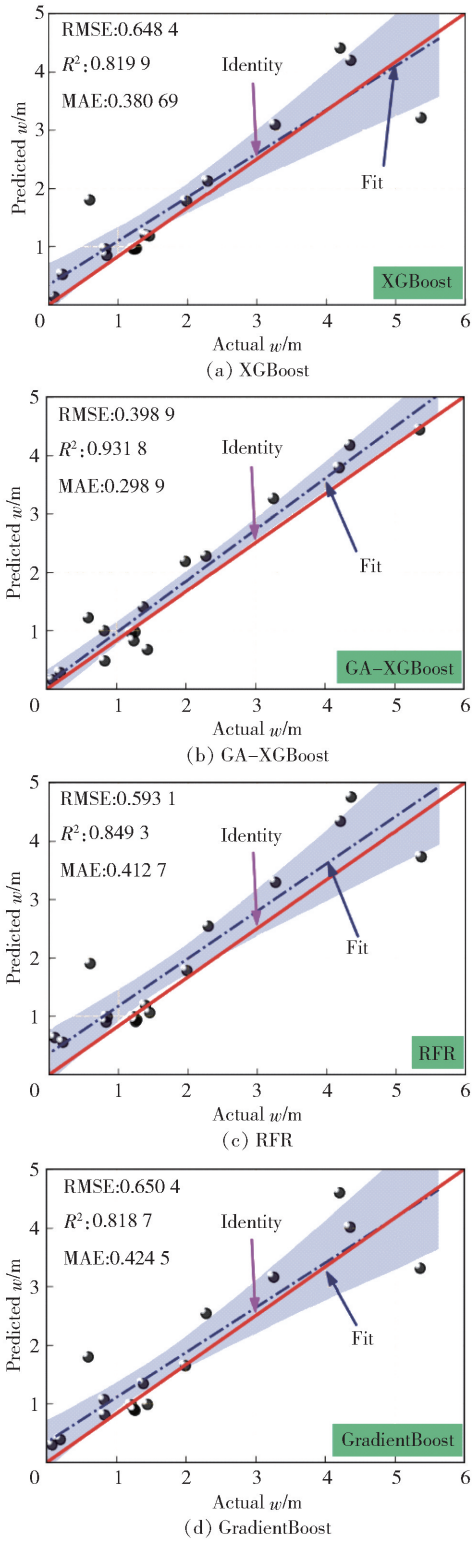

图 5　不同模型预测精度对比

GA 算法优化 XGBoost 模型提高开采沉陷的预测精度是可行的。

同时为了准确评价 4 个模型的优劣情况，我们采用了泰勒图描述模型的性能，具体如图 6 所示。

图中 XGBoost、RFR 以及 Gradient Boost 距离实际值 REF 较远,GA-XGBoost 距离 REF 更近,说明 GA-XGBoost 模型相比于其他单一模型更优秀。

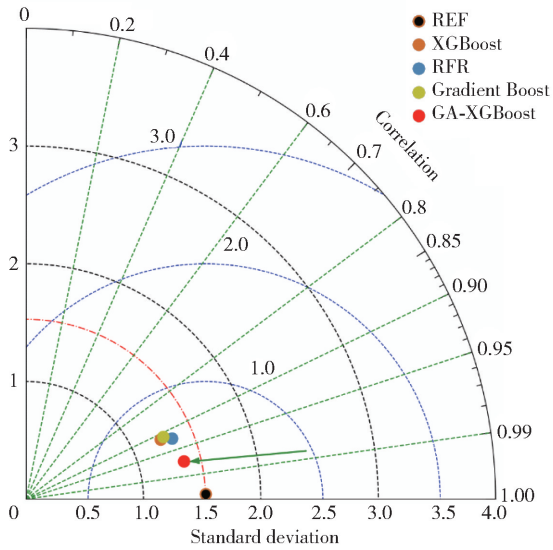

图 6 模型性能的泰勒图

4 结论

(1) GA-XGBoost 模型的预测精度比与单一的 XGBoost、RFR,Gradient Boost 等模型更高,说明利用 GA 算法优化 XGBoost 的超参数提高模型的预测性能是切实可行的。经过 GA 优化后的 XGBoost 模型,决定是系数 R^2 为 0.931 8,*RMSE* 为 0.398 9,*MAE* 为 0.298 9,可见 GA-XGBoost 模型适用于预测矿山开采沉陷,而且预测精度较高。

(2) GA-XGBoost 的本质是利用 GA 的搜索能力,实现 XGBoost 模型的自适应与自优化,从而提高预测性能。随着矿山数据集的不断丰富与积累,模型的应用场景将会更加广泛,不仅可以补充开采沉陷领域的预测方法和理论,在其他矿业领域复杂高度非线性的问题分析中也有应用价值。在未来,可以通过研究更多矿山现场案例,收集更多高质量的沉陷数据和采用其他的优化策略,提高模型的预计精度和性能,为矿山开采沉陷治理提供可靠的理论支撑。

[参考文献]

[1] 芦家欣, 汤伏全, 赵军仪,等. 黄土矿区开采沉陷与地表损害研究述评[J]. 西安科技大学学报, 2019, 39(5):8.

[2] 李昱昊,安士凯,周大伟,等. 基于 UAV 摄影测量技术的开采沉陷全盆地建模和求参[J]. 煤矿安全, 2022, 53(2):179 – 186.

[3] 代张音,唐建新,王艳磊,等. 顺层岩质斜坡开采沉陷预测模型研究[J]. 岩石力学与工程学报, 2017, 36(12):3012 – 3020.

[4] 张会军. 黄河流域浅埋深煤层开采冒落沉陷特征及治理复垦技术研究[J]. 煤炭工程, 2021, 53(8):140 – 144.

[5] 朱晓峻. 带状充填开采岩层移动机理研究[D]. 徐州:中国矿业大学,2016.

[6] 邓伟男. 基于 FISH 语言的开采沉陷模拟数据处理方法[J]. 煤矿开采,2016,21(4):15 – 17 +9. .

[7] 李新岭,郭文兵,赵高博. 巨厚松散层土体压缩特性对开采沉陷影响研究[J]. 中国安全科学学报,2018, 28(7):135 – 141.

[8] 牛亚超,徐良骥,张坤,等. 基于 GA-BP 神经网络的概率积分法预计参数研究[J]. 金属矿山,2019(10): 93 – 100.

[9] 徐静宇. 基于卷积神经网络的开采沉陷预计参数计算模型研究[D]. 阜新:辽宁工程技术大学,2021.

[10] 丰土根,王超然,张箭. 基于 ABC-BP 模型的基坑地表沉降预测[J]. 河北工程大学学报(自然科学版), 2020,37(4):7 – 12.

[11] 许凯文. 基于遗传算法优化 BP 神经网络的深基坑地连墙变形预测[J]. 粉煤灰综合利用,2021,35(5): 6 – 11.

[12] 潘红宇,赵云红,张卫东,等. 基于 Adaboost 的改进 BP 神经网络地表沉陷预测[J]. 煤炭科学技术,2019, 47(2):161 – 167.

[13] 邢垒,原喜屯,张沛. 基于 Adaboost-PSO-BP 模型的开采沉陷预测研究[J]. 煤炭工程, 2020, 52(12): 141 – 144.

[14] 王雪英. 基于 BP 神经网络的山区开采沉陷预计[D]. 太原:太原理工大学,2010.

[15] 孙嘉琪. 基于改进 XGBoost 算法的企业信用评级预测方案设计[D]. 上海:上海师范大学, 2021.

[16] 杜雷. 基于 XGBOOST 算法的某铁路局旅客发送量预测分析[D]. 成都:西南交通大学, 2020.

[17] 武梦婷, 陈秋松,齐冲冲. 基于机器学习的边坡安全稳定性评价及防护措施[J]. 工程科学学报, 2022, 44(2):180 – 188.

[18] 刘微笑. 基于遗传算法的 A 企业车间布局优化研究[D]. 徐州:中国矿业大学, 2021.

[19] 谷晓琳. 基于改进遗传算法的柔性作业车间调度问题的应用研究[D]. 大连:大连交通大学,2020.

[20] 罗楚航. 基于并行化改进遗传算法的配电网电压优化方法研究[D]. 哈尔滨:哈尔滨工业大学, 2021.