



基于混合策略 ISSA-XGBoost 的高速公路工程造价预测研究

李珏*, 刘洋

长沙理工大学 交通运输工程学院, 长沙 410114

摘要: 高速公路造价预测是对高速公路建设项目前期阶段进行造价控制的重要手段。本文针对工程实践中小样本数据和工程造价特征指标之间高维、非线性关系的特点, 融合正余弦算法和 Lévy 飞行改进的麻雀算法来优化 XGBoost 超参数, 对高速公路项目进行造价预测, 同时与该改进麻雀算法优化的 RF、SVM 模型比较, 结果表明 ISSA-XGBoost 模型具有更好的泛化性和可解释性, 可为高速公路项目的投资决策提供可靠依据。

关键词: 高速公路; 造价预测; Lévy 飞行; 改进麻雀算法; ISSA-XGBoost

中图分类号: TU723.3

文献标识码: A

文章编号: 1674-4969(2026)01-0070-15

引言

高速公路造价高昂, 施工周期长, 技术复杂, 工程成本投入大, 资本运行风险高^[1], 投资管控难度大。高速公路项目前期决策阶段的投资估算准确与否不仅影响到可行性研究工作的质量和经济评价结果, 而且也直接关系到下一阶段设计概算和施工图预算的编制, 对建设项目资金筹措方案也有直接的影响^[2]。在工程项目前期可获得的造价信息有限, 传统估算方

法无法满足投资决策的精度及效率要求, 也不能满足高速公路建设造价全过程管理精细化发展的实际需要^[3]。对于工程特征和市场环境的变化, 传统方法具有一定的局限性、滞后性和偏差性, 因此探索更为普适化、精确化的估算模型显得尤为必要。

随着人工智能的快速发展, 机器学习算法也逐渐被应用到工程建设各个领域的投资估算中, 对成本估算的发展产生了深远的影响^[4]。一些研究学者^[5-9]基于支持向量机(support vector machine, SVM)及其融

收稿日期: 2024-05-21; **修回日期:** 2024-06-13

基金项目: 湖南省教育厅科研项目: 基于“信息融合+价值链”的全过程工程咨询模式研究(22A0202)

作者简介: *李珏(1978—), 男, 博士研究生, 副教授, 研究方向为公路工程项目管理、安全事故分析、文本挖掘。E-mail: Lij93@csust.edu.cn (通讯作者)

刘洋(2000—), 男, 硕士研究生, 研究方向为文本挖掘、安全事故分析。E-mail: 22101040132@stu.csust.edu.cn

引用格式: 李珏, 刘洋. 基于混合策略 ISSA-XGBoost 的高速公路工程造价预测研究[J]. 工程研究——跨学科视野中的工程, 2026, 18(1): 70-84. DOI: 10.3724/j.issn.1674-4969.20240064. CSTR: 32282.14.JES.20240064

LI J, LIU Y. Research on Expressway Engineering Cost Prediction Based on Hybrid ISSA-XGBoost[J]. Journal of Engineering Studies, 2026, 18(1): 70-84. DOI: 10.3724/j.issn.1674-4969.20240064. CSTR: 32282.14.JES.20240064

合模型对隧道工程、水利工程、住宅工程及电力工程造价进行估算与控制,表明该算法具有良好的泛化能力和预测精度。对于公路领域,许多学者^[10-15]以人工神经网络(artificial neural network, ANN)为基础,结合各类优化增强算法开发了许多公路成本估算模型,展示了人工神经网络在预测公路建设成本方面的优越性。极端梯度上升(extreme gradient boosting, XGBoost)是由Chen等^[16]提出的基于决策树的集成算法,在许多数据竞赛中大放异彩,很多学者也开始尝试将XGBoost算法应用于建设项目的初期成本估算^[17-19]。郭婧娟和刘曜玮^[20]基于XGBoost算法模型与造价指标体系对高速公路工程概算进行了预测模型的创建,分析了XGBoost算法模型与指标体系的优势。工程建设项目造价数据具有高维、非线性、大样本等特点,传统方法^[21-22]对于这些特点无能为力,同时传统方法具有应用上的局限性及计算上的偏差性,无法处理海量样本且学习新的样本,具有明显的滞后性。而机器学习算法能自动从这些高维数据中提取有用信息,解决复杂的系统问题,并且模型具有很好的解释性,这些优点对于投资估算应用来说具有很强的适用性。

本研究以建立起面向我国高速公路领域的造价预测模型为目标,使用训练效率高、预测性能好的XGBoost算法来构建模型。同时XGBoost超参数众多,模型训练效果好坏取决于超参数的选择。本文使用薛建凯^[23]于2019年提出的麻雀搜索算法(sparrow search algorithm, SSA)对XGBoost超参数进行寻优,考虑到基本麻雀算法容易陷入局部最优,故采用一种混合正余弦算法和Lévy飞行的策略^[24]来改进基本麻雀算法,以加强全局搜索能力,旨在使模型具有更快的收敛速度、更强的泛化能力以及更高的预测精度,为高速公路项目前期的投资决策提供依据。

1 ISSA-XGBoost模型

1.1 XGBoost算法

XGBoost算法是基于boosting思想,将许多弱学习器集成在一起形成的一个强学习器。其核心思想是利用贪心方法学习每棵基树。通过不断形成新的决策

树来拟合之前预测的残差,不断减小预测值与真实值之间的残差误差,从而提高预测精度。优化目标可由式(1)、式(2)定义:

$$\text{Obj}^{(t)} = \arg \min \sum_i L(y_i, \hat{y}_i^{(t)}) + \Omega(f_t) \quad (1)$$

$$\Omega(f_t) = \gamma N + \frac{1}{2} \lambda \sum_{j=1}^N w_j^2 \quad (2)$$

式中, L 为损失函数; $\Omega(f_t)$ 为正则化项; γ 、 λ 均为超参数,可以控制惩罚的力度; N 为该棵树叶子结点的个数; w_j 为叶子结点的值。

对损失函数进行二阶泰勒展开并回代至目标函数就可计算出每个叶子结点的值 w_j 以及最小损失值 L 。因为每棵树的叶节点值都是拟合上一轮的残差,将每棵树的计算结果相加就可得到最终的预测值,即加法模型式(3):

$$\hat{y}^{(T)} = \sum_{j=1}^T f_j(x) \quad (3)$$

式中, T 为树的棵数。

1.2 基本麻雀搜索算法(SSA)

麻雀搜索算法是通过模仿麻雀的觅食行为和反捕食行为来实现模型参数寻优的。它可以有效减少搜索空间,提高搜索效率,避免搜索空间中的局部最优解。在该算法中麻雀种群由发现者、加入者和预警者三部分组成。

用式(4)来更新发现者的位置信息:

$$X_{ij}^{t+1} = \begin{cases} X_{ij}^t \exp\left(\frac{-i}{\alpha \times \text{iter}_{\max}}\right) & \text{if } R_2 < \text{ST} \\ X_{ij}^t + QL & \text{if } R_2 \geq \text{ST} \end{cases} \quad (4)$$

式中, t 代表当前迭代数, $i=1,2,\dots,n$, $j=1,2,\dots,d$; X_{ij} 为第 i 只麻雀在第 j 维中的位置信息; $\alpha \in (0,1]$ 是一个随机数; $\text{iter}_{\max}$ 为最大迭代次数; $R_2 \in [0,1]$ 和 $\text{ST} \in [0.5,1]$ 分别表示预警值和安全值; Q 是服从正态分布的随机数; L 表示 $1 \times d$ 的全 1 矩阵。

加入者的位置信息更新见式(5):

$$X_{ij}^{t+1} = \begin{cases} Q \exp\left(\frac{X_{\text{worst}}^t - X_{ij}^t}{i^2}\right) & \text{if } i > n/2 \\ X_p^{t+1} + |X_{ij}^t - X_p^{t+1}| A^+ L & \text{otherwise} \end{cases} \quad (5)$$

式中, X_p 是发现者占据的最优位置; X_{worst} 表示当前的全局最差位置; A 表示一个 $1 \times d$ 的矩阵, 每个元素随机分配 1 或 -1, $A^+ = A^T(AA^T)^{-1}$ 。

预警者的位置更新表达式如式(6):

$$X_{i,j}^{t+1} = \begin{cases} X_{\text{best}}^t + \beta |X_{i,j}^t - X_{\text{best}}^t| & \text{if } f_i > f_g \\ X_{i,j}^t + K \left(\frac{X_{i,j}^t - X_{\text{worst}}^t}{(f_i - f_w) + \varepsilon} \right) & \text{if } f_i = f_g \end{cases} \quad (6)$$

式中, X_{best} 为当前全局的最优位置; β 是服从 $N(0,1)$ 的随机数, 作为步长控制参数; K 为 $[-1,1]$ 的随机数, 也是步长控制系数; f_i 为当前麻雀个体的适应度值; f_g 和 f_w 分别为当前全局最佳和最差适应度值; ε 为大于 0 的常数, 以避免分母为 0。

1.3 混合正余弦算法和 Lévy 飞行的改进麻雀算法 (improved sparrow search algorithm, ISSA)

基本麻雀算法虽然相对于其他群智能优化算法而言具有收敛速度快、求解精度高等优点, 但后期同样容易陷入局部最优。为了解决这一问题, 提高算法全局搜索能力, 本文引入正余弦算法平衡局部和全局搜索能力, 并加快收敛速度, 同时加入 Lévy 飞行策略对当前最优解进行扰动变异, 加强局部逃逸能力。在发现者位置更新计算式中引入非线性正弦学习因子, 改进后如式(7)、式(8):

$$\omega = \omega_{\min} + (\omega_{\max} - \omega_{\min}) \sin(\pi / \text{iter}_{\max}) \quad (7)$$

$$X_{i,j}^{t+1} = \begin{cases} (1 - \omega)X_{i,j}^t + \omega \sin(r_1) |r_2 X_{\text{best}} - X_{i,j}^t|, & R_2 < \text{ST} \\ (1 - \omega)X_{i,j}^t + \omega \cos(r_1) |r_2 X_{\text{best}} - X_{i,j}^t|, & R_2 \geq \text{ST} \end{cases} \quad (8)$$

式中, ω 为惯性权重, $\omega_{\max}$ 和 $\omega_{\min}$ 分别是惯性因子上、下限; r_1 是 $[0, 2\pi]$ 内的随机数; r_2 是 $[0, 2]$ 内的随机数。

在加入者位置更新计算式中引入 Lévy 飞行策略, 改进后如式(9)~式(11):

$$X_{i,j}^{t+1} = \begin{cases} Q \exp\left(\frac{X_{\text{worst}}^t - X_{i,j}^t}{t^2}\right) & \text{if } i > n/2 \\ X_p^{t+1} + X_p^{t+1} \otimes \text{Lévy}(d) & \text{otherwise} \end{cases} \quad (9)$$

$$\text{Lévy}(x) = 0.01 \frac{r_3 \sigma}{|r_4|^{(1/\xi)}} \quad (10)$$

$$\sigma = \left\{ \frac{\Gamma(1+\xi) \sin(\pi\xi/2)}{\Gamma[(1+\xi)/2] \xi 2^{(\xi-1)/2}} \right\}^{(1/\xi)} \quad (11)$$

式中, r_3 、 r_4 均为 $[0,1]$ 范围内的随机数; ξ 可取 1.5; $\Gamma(x) = (x-1)!$ 。

1.4 建模流程

在高速公路造价预测模型构建中, 选择模型最优参数是保证模型回归结果精确的关键, 本文在预测之前首先运用 ISSA 对 XGBoost 中的 learning_rate(学习率)、n_estimators(树数量)、max_depth(树最大深度)、subsample(行抽样)、colsample_bytree(列抽样)、reg_alpha(正则化参数)这六个参数进行择优选取。计算流程如图 1 所示。同时通过 ISSA、SSA、粒子群优化算法(particle swarm optimization, PSO)之间迭代收敛曲线的比较, 以及该改进麻雀算法优化的 XGBoost、随机森林(random forest, RF)和 SVM 模型的比较来验证本文提出的 ISSA-XGBoost 模型的优越性, 全文总体研究框架如图 2 所示。

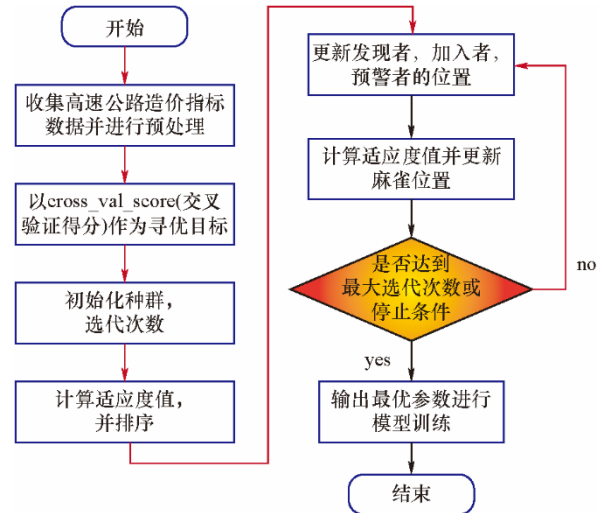


图1 ISSA 算法优化的计算流程

Figure 1 The calculation process of ISSA algorithm optimization

2 高速公路造价预测指标选取

2.1 数据来源

本文的造价数据整理自近十年来各省份高速公路

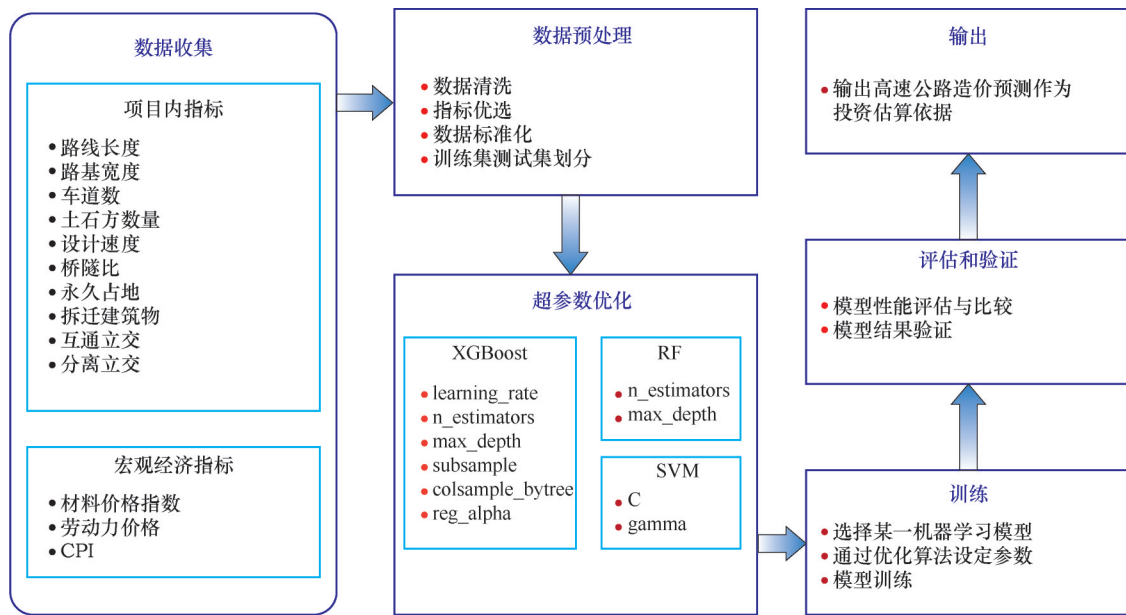


图2 研究框架

Figure 2 Research framework

表 1 部分原始造价数据

Table 1 Part of original cost data

项目	安徽东至至九江 高速公路安徽段	安徽阜阳至新蔡 高速公路		重庆丰都至石柱 高速公路	重庆江津至贵州习水 高速公路(重庆境)
编号	1	2		65	66
路线长度/km	10.802	68.048		53.929	70.458
路基宽度/m	26.0	26.0		24.5	24.5
车道数/个	4	4		4	4
土石方数量/ 10^4 m^3	271.35	1438.52		1384.21	2170.00
设计速度/(km/h)	100	100		80	80
桥隧比/%	8.00	2.59		50.73	33.15
永久占地/ hm^2	78.68	345.12		301.87	388.51
拆迁建筑物/ m^2	3258	64880		135598	134665
互通立交/个	0	2		5	7
分离立交/个	2	27		4	27
竣工时间/年	2015	2013		2013	2018
CPI	615.2	594.8		594.8	650.9
材料价格指数	358.9	385.9		385.9	389.4
劳动力价格/元	48886	42072		42072	60501
工程造价/亿元	5.53	24.44		52.30	82.11

竣工文件。对于宏观经济指标,许多学者的研究都涉及了劳动力价格^[25],因为在国内环境下土木行业仍然属于劳动密集型产业^[26],同时居民消费价格指数

(consumer price index, CPI)则是国外学者比较看重的对高速公路工程造价有较大影响的因素^[27, 28],而材料价格作为工程采购成本的直接体现具有研究意义^[29],因此本

文引入了三个宏观经济指标: CPI、劳动力价格(以建筑业城镇单位就业人员平均工资表示)、材料价格指数(以工业生产者出厂价格指数表示), 数据均来自国家统计局。在剔除了改扩建项目、其他等级公路项目后整理形成66条原始造价数据, 部分示例于表1, 其原始造价指标如表2所示。

表2 原始造价指标
Table 2 Original cost indicators

项目内指标	宏观经济指标
路线长度; 车道数; 设计速度; 永久占地; 互通立交; 路基宽度; 土石方数量; 桥隧比; 拆迁建筑物; 分离立交	CPI; 劳动力价格; 材料价格指数

2.2 指标优选

合理选取造价影响指标是模型构建的前提。本研究基于原始数据集的相关性分析以及RF特征重要性排序来进行指标选取。为了剔除路线长度对某些解释

变量的影响(如土石方数量), 采用单位指标(原始数据/路线长度)来代替原始造价指标: 土石方数量、占地面积、拆迁建筑物面积变为单位土石方数量以及单位占地面积和单位拆迁建筑物面积, 被解释变量变为单位造价。对原始数据集进行变量相关性分析, 采用Pearson相关系数法, 结果如图3所示。虽然基于决策树的XGBoost算法本质上并不受多重共线性的影响, 但即便如此, 当数据集中出现了与原始变量相关的变量时, 原始变量的重要性可能会降低, 而删除两个相关变量之一并不会对模型的准确性产生重大影响, 因为共线性不是增强问题, 决策树并不关心使用哪个变量, 而另一个变量包含相似的信息。无论算法模型如何, 从训练的数据集中删除冗余特征是一种很好的做法。由图3可以看出, 路基宽度与车道数、设计速度存在明显的正相关关系, CPI与劳动力价格相关性达到0.99, 可以看作是线性关系, 因此根据上述原则应首先删去车道数、设计速度、劳动力价格这三个特征。

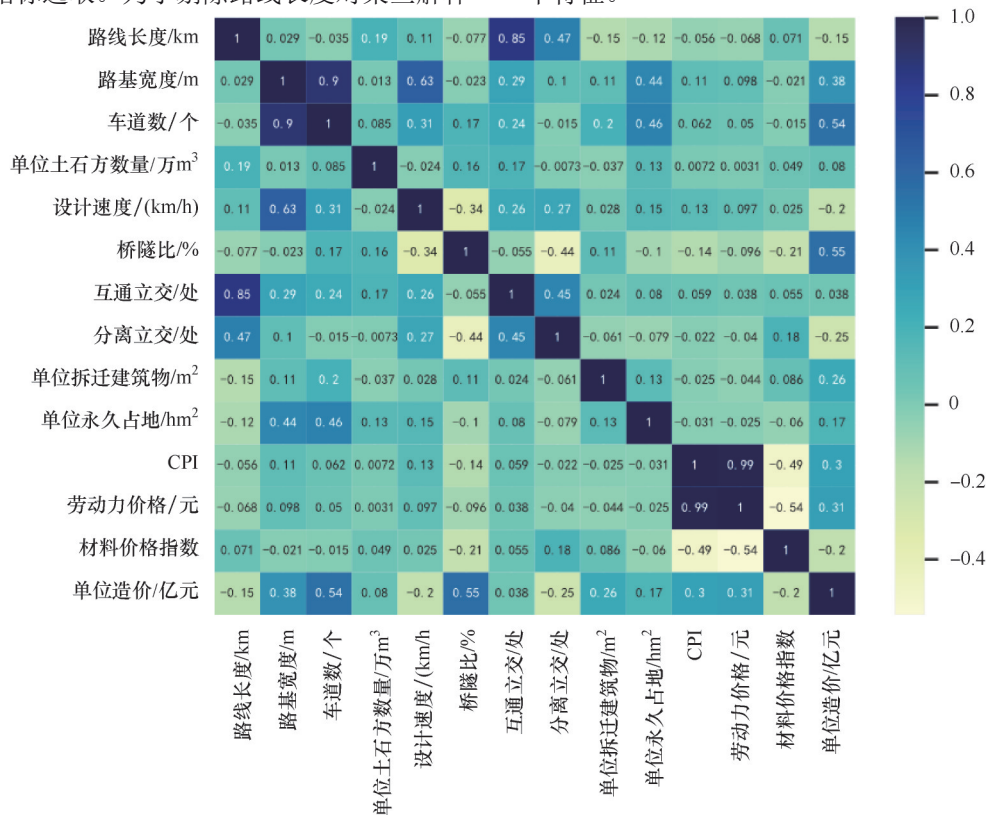


图3 特征间相关系数图

Figure 3 Correlation coefficient graph between features

表3 输入样本集

Table 3 Input sample set

编号	路线长度	路基宽度/m	单位土石方数量/ 10^4 m^3	桥隧比/%	互通立交/处	单位永久占地/ hm^2	CPI	材料价格指数	单位造价/亿元
1	178.135	28.0	11.15	3.38	4	6.61	637.5	376.2	0.282
2	68.048	26.0	21.14	2.59	2	5.07	594.8	385.9	0.359
3	39.583	26.0	19.77	9.85	4	6.79	615.2	358.9	0.378
4	88.633	28.0	21.65	4.76	8	7.40	579.7	393.4	0.396
5	32.746	29.0	10.44	2.55	3	6.90	565.0	400.2	0.420
6	139.060	27.0	11.21	8.52	11	5.29	606.7	378.6	0.449
7	139.140	28.0	35.61	7.52	9	6.95	579.7	393.4	0.470
8	40.26	24.5	27.87	10.81	5	8.06	615.2	358.9	0.486
9	23.347	28.0	14.67	6.33	2	7.08	594.8	385.9	0.488
10	50.498	26.0	20.10	5.95	5	7.44	579.7	393.4	0.497
11	49.42	24.5	12.70	5.17	1	6.78	615.2	358.9	0.508
12	10.802	26.0	25.12	8	0	7.28	615.2	358.9	0.512
13	80.613	26.0	27.62	8.58	7	7.46	579.7	393.4	0.523
14	78.76	26.0	19.61	14.05	4	5.43	594.8	385.9	0.529
15	69	26.0	26.04	18.53	4	6.19	579.7	393.4	0.539
16	116.899	26.0	33.99	15.53	10	6.70	627.5	353.9	0.557
17	88.45	28.0	31.07	11.96	7	6.84	615.2	358.9	0.576
18	34.78	21.5	16.79	14.8	3	6.94	606.7	378.6	0.581
19	43.149	28.0	16.62	11.17	4	6.80	615.2	358.9	0.600
20	39.388	26.0	20.23	22.94	5	6.90	615.2	358.9	0.601
21	30.82	21.5	35.22	25.12	3	7.74	579.7	393.4	0.607
22	56.815	28.0	9.73	12.776	5	5.57	627.5	353.9	0.609
23	145.187	24.5	16.14	26.33	6	5.17	594.8	385.9	0.620
24	13.44	24.5	12.37	16.43	1	5.20	594.8	385.9	0.624
25	77.268	26.0	12.42	20.17	5	6.26	594.8	385.9	0.624
26	36.66	28.0	17.08	10.01	4	4.33	615.2	358.9	0.634
27	3.3	26.0	22.97	12.18	2	9.40	627.5	353.9	0.639
28	58.672	26.0	12.32	15.48	4	6.15	615.2	358.9	0.641
29	154.58	26.0	42.10	17.01	12	5.85	637.5	376.2	0.645
30	53.401	28.0	27.31	11.51	4	8.91	594.8	385.9	0.647
31	37.5	34.5	22.85	12.66	6	6.66	594.8	385.9	0.647
32	121.805	24.5	15.17	33.15	9	7.60	536.1	377.5	0.650
33	47.718	24.5	20.16	29.64	3	6.99	606.7	378.6	0.688
34	38.824	24.5	28.90	30.9	2	6.83	615.2	358.9	0.688
35	54.764	26.0	41.31	24.23	5	7.96	627.5	353.9	0.695
36	24.08	34.5	39.71	11.35	3	9.15	565	400.2	0.713
37	40.202	26.0	25.62	39	4	5.48	615.2	358.9	0.728
38	12.545	21.5	26.45	41.89	1	6.31	579.7	393.4	0.729

续表

编号	路线长度	路基宽度/m	单位土石方 数量/ 10^4m^3	桥隧比/%	互通立 交/处	单位永久 占地/ hm^2	CPI	材料价格 指数	单位造 价/亿元
39	123.52	26.0	29.85	31.1	7	9.47	565	400.2	0.737
40	203.03	21.5	33.04	29.78	13	5.23	579.7	393.4	0.749
41	105.248	26.0	27.04	19.27	10	6.99	594.8	385.9	0.750
42	33.796	26.0	25.04	19.52	4	9.66	606.7	378.6	0.760
43	99.196	26.0	31.38	16.77	9	6.89	579.7	393.4	0.766
44	95.713	26.0	43.31	20.03	10	7.78	579.7	393.4	0.814
45	98.544	26.0	14.22	16.15	12	7.73	615.2	358.9	0.815
46	23.323	24.5	25.38	46.85	1	9.00	536.1	377.5	0.836
47	83.64	26.0	52.81	27.6	6	6.11	579.7	393.4	0.841
48	73.027	26.0	22.05	20.11	5	6.36	579.7	393.4	0.847
49	46.07	24.5	40.78	52.02	5	7.05	637.5	376.2	0.858
50	59.55	24.5	14.79	73.54	3	8.04	519	357.8	0.877
51	56.078	24.5	32.45	44.27	5	5.71	579.7	393.4	0.946
52	53.929	24.5	25.67	50.73	5	5.60	594.8	385.9	0.970
53	48.4	24.5	37.48	34.1	6	6.60	627.5	353.9	0.975
54	35.924	24.5	14.06	63.12	2	5.44	615.2	358.9	0.998
55	21.6	24.5	23.79	42.28	2	7.20	627.5	353.9	1.014
56	138.21	34.5	45.46	37.49	14	7.82	606.7	378.6	1.051
57	18.99	24.5	39.86	41.45	0	5.72	594.8	385.9	1.083
58	21.623	24.5	34.82	43.67	1	4.97	637.5	376.2	1.138
59	67.179	24.5	34.53	52.06	3	4.94	615.2	358.9	1.152
60	20.417	24.5	16.75	60.64	2	4.60	594.8	385.9	1.261
61	46.369	24.5	15.77	76.87	2	5.24	594.8	385.9	1.285
62	25.688	33.5	24.04	68.81	3	7.77	627.5	353.9	1.327
63	63.200	33.5	37.66	49.67	10	7.52	594.8	385.9	1.360
64	91.086	24.5	11.77	44.13	6	5.55	627.5	353.9	1.460
65	63.577	34.5	19.08	59.85	12	7.64	606.7	378.6	1.529
66	19.490	34.5	20.46	46.65	5	8.90	627.5	353.9	1.594

本文不仅将RF算法作为对比模型,而且利用其变量重要性评价进行指标选取,来作为一种特征选择方法。在剔除了车道数、设计速度、劳动力价格这三个特征后,将剩余特征进行RF回归(参数设置为默认),特征重要性可视化如图4所示。通过分析,互通立交特征重要性略高于分离立交,由于互通立交造价以及施工难度、周期明显高于分离立交,两者包含相似的信息,而互通立交这个指标更能代表高速公路整体造价,故剔除分离立交特征。同时,拆迁建筑物

面积并没有一个统一的评价及补偿标准,所包含的高速公路整体造价信息并不充分,故删去。本研究最终保留八个特征指标作为模型输入指标:路线长度、路基宽度、单位土石方数量、桥隧比、互通立交、单位永久占地、CPI、材料价格。

XGBoost算法是基于决策树的集成算法,而树模型对于数据没有什么要求,不需要对数据进行太多的预处理。本文数据集的八个特征均为连续型变量,无须对其作标准化或归一化的处理,只需对类别型变量

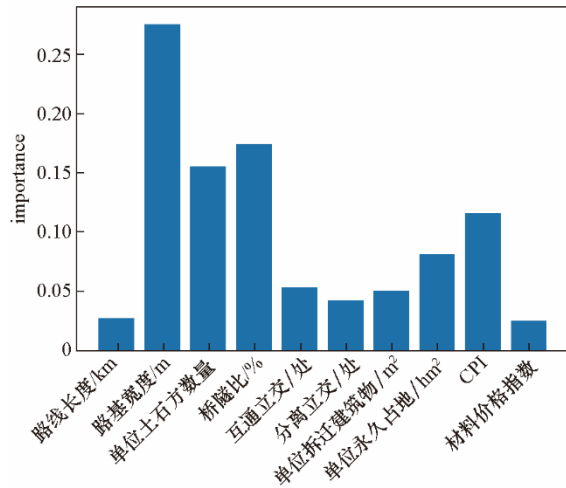


图4 RF特征重要性可视化

Figure 4 Visualization of RF feature importance

用 one-hot-encoding 等方法进行编码。对保留指标进行整理后形成的最终输入样本集如表3所示。

3 ISSA-XGBoost造价预测模型的建立

3.1 ISSA 参数择优

本文通过 Python 实现 ISSA 优化算法以及构建 ISSA-XGBoost 高速公路造价预测模型, 以 $\text{neg_mean_squared_error}$ 负均方误差为评分标准的 cross_val_score 交叉验证得分作为种群适应度。ISSA 种群初始化参数以及 XGBoost 参数寻优范围如表4所示。

设置数据集中的80%为训练集, 20%为测试集。本文以 PSO 算法为基准, 比较 SSA 与 ISSA 算法的迭

表4 ISSA-XGBoost 参数设置

Table 4 ISSA-XGBoost parameter settings

参数	取值	寻优范围
种群规模	100	$\text{learning_rate}[0.01, 0.1]$
最大迭代次数	50	$\text{n_estimators}[100, 500]$
安全值 ST	0.6	$\text{max_depth}[3, 10]$
发现者比例	0.7	$\text{subsample}[0.6, 1]$
意识到危险的麻雀比重	0.2	$\text{colsample_bytree}[0.6, 1]$
$\omega_{\max}$	1	$\text{reg_alpha}[0.001, 10]$
$\omega_{\min}$	0.4	—

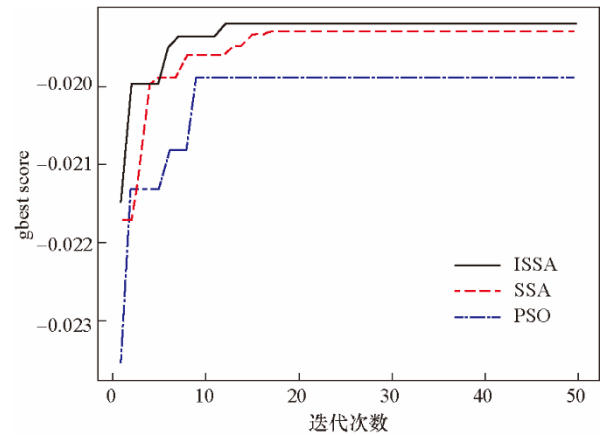


图5 ISSA、SSA、PSO算法迭代收敛曲线

Figure 5 Iterative convergence curves of ISSA, SSA, and PSO algorithms

代收敛曲线如图5所示, 可以看出 ISSA 算法无论在收敛速度还是收敛精度上都优于 SSA 与 PSO 算法, 验证了该改进麻雀算法的稳健性与有效性。ISSA、PSO、SSA 择优参数组合如表5所示。

表5 ISSA、SSA、PSO 择优参数组合

Table 5 ISSA, SSA, PSO optimization parameter combination

模型参数	ISSA	SSA	PSO
max_depth	3	7	7
colsample_bytree	1	0.8846	0.889
learning_rate	0.0847	0.05	0.0184
n_estimators	323	355	374
reg_alpha	0.001	0.001	0.0027
subsample	1	0.964	1

3.2 模型结果

将本研究样本数随机划分80%训练集, 20%测试集。分别设定模型参数, 通过真实值与预测值的比较, 可以对 ISSA 的优化效果进行定量评价, 具体结果见表6。ISSA-XGBoost 预测值与真实值对比结果如图6所示。表7则给出了 ISSA-XGBoost、PSO-XGBoost、SSA-XGBoost 模型的评价指标得分。

由表6可知, PSO-XGBoost 模型预测的最大误差率为56.03%, 最小误差率为-2.18%, 平均绝对误差率为13.2%; SSA-XGBoost 模型预测的最大误差率为56.03%, 最小误差率为-1.24%, 平均绝对误差率为

表6 PSO-XGBoost、SSA-XGBoost、ISSA-XGBoost 预测结果对比

Table 6 Comparison of PSO-XGBoost, SSA-XGBoost, and ISSA-XGBoost prediction results

样本 编号	实际值/ (亿/km)	PSO-XGBoost		SSA-XGBoost		ISSA-XGBoost	
		预测值/(亿/km)	相对误差/%	预测值/(亿/km)	相对误差/%	预测值/(亿/km)	相对误差/%
1	0.601	0.618	2.82	0.633	5.32	0.636	5.82
2	0.814	0.775	-4.79	0.787	-3.32	0.869	6.76
3	0.750	0.692	-7.73	0.671	-10.53	0.739	-1.47
4	0.647	0.696	7.57	0.683	5.56	0.633	-2.16
5	0.858	1.010	17.72	1.037	20.86	1.125	31.12
6	0.620	0.721	16.29	0.698	12.58	0.692	11.61
7	0.529	0.556	5.10	0.579	9.45	0.611	15.50
8	0.688	0.673	-2.18	0.675	-1.89	0.683	-0.73
9	1.152	1.110	-3.65	1.111	-3.56	1.139	-1.13
10	0.282	0.440	56.03	0.440	56.03	0.408	44.68
11	1.360	1.016	-25.29	1.072	-21.18	1.185	-12.87
12	0.576	0.645	11.98	0.638	10.76	0.656	13.89
13	0.645	0.743	15.19	0.732	13.49	0.735	13.95
14	1.050	0.961	-8.48	1.037	-1.24	1.043	-0.67

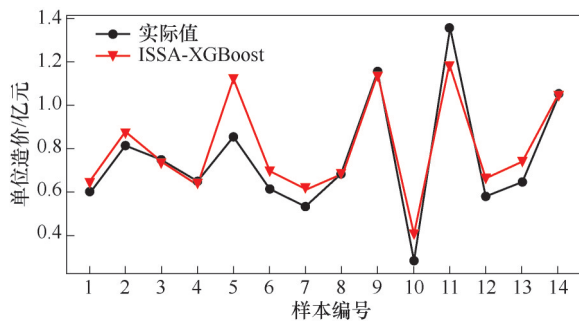


图6 ISSA-XGBoost 模型预测结果

Figure 6 ISSA-XGBoost model prediction results

表7 ISSA-XGBoost、PSO-XGBoost、SSA-XGBoost 模型评价指标得分

Table 7 ISSA-XGBoost, PSO-XGBoost, SSA-XGBoost model evaluation metrics

模型	决定 系数 R^2	调整后决定系数 $\text{Adjusted } R^2$	平均绝对 误差 MAE	均方误差 MSE
ISSA-XGBoost	0.852	0.831	0.074	0.011
PSO-XGBoost	0.791	0.762	0.090	0.015
SSA-XGBoost	0.830	0.806	0.081	0.012

12.56%；ISSA-XGBoost 模型预测的最大误差率为 44.68%，最小误差率为-0.67%，平均绝对误差率为 11.6%。无论是从误差率极差还是平均绝对误差率来看，改进后的麻雀算法都略优于其他两种算法。通过

上述分析可以看出 ISSA-XGBoost 模型的预测精度更稳定。同时，表 6 中出现了预测误差率很高的样本，是因为本研究样本集来自于全国各省份，非单一省份数据，高速公路造价数据有很大的区域差异，样本中特征数据范围跨度很大。本研究并不追求过拟合的效果，由表 7 可知 ISSA-XGBoost 的 $\text{Adjusted } R^2$ 为 0.831，说明该模型的预测结果在全国范围内具有一定的广泛性与普适性。对于投资估算而言，建设项目规划和项目建议书阶段的投资估算精度的要求为误差控制在 $\pm 30\%$ 以内；预可行性研究阶段的投资估算精度要求为误差控制在 $\pm 20\%$ 以内；可行性研究阶段的投资估算精度要求为误差控制在 $\pm 10\%$ 以内^[30]。通过上述分析可知本文提出的 ISSA-XGBoost 预测模型的平均绝对误差率为 11.6%，其预测值可作为投资估算阶段的前期依据。

ISSA-XGBoost 模型的可解释性体现在其决策树的可视化(图 7)以及特征重要性排序(图 8)等方面。简言之，如果某一个特征在所有树中被用作分裂节点的次数越多，如在图 7 中桥隧比作为分裂节点出现了 4 次，同时该特征作为分裂节点所处理的样本数越多，带来的总增益越大，就说明该特征越重要。在图 8 中



图7 ISSA-XGBoost模型决策树

Figure 7 ISSA-XGBoost model decision tree

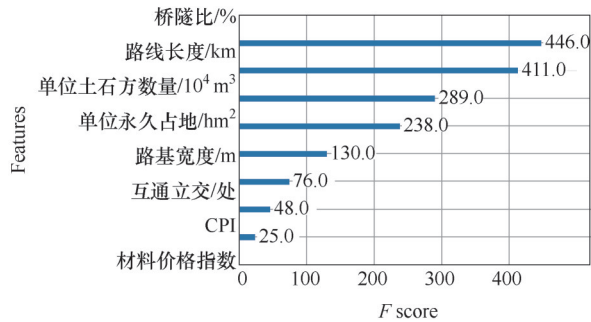


图8 ISSA-XGBoost模型特征重要性排序

Figure 8 ISSA-XGBoost model feature importance ranking

可以看出, 桥隧比、路线长度、单位土石方数量这三个特征对造价的影响分列前三, 而互通立交、CPI、材料价格指数这些特征对造价的影响相对而言没有那么大。

3.3 模型对比

为验证本研究提出的 ISSA-XGBoost 模型具有更好的可解释性和广泛性, 选取该改进麻雀算法优化随机森林(ISSA-RF)及支持向量机(ISSA-SVM)与所提出的 ISSA-XGBoost 模型进行预测性能对比。对比结果如图9所示。

为进一步对比模型预测结果的稳定性程度, 选用绝对误差率最大值 (maximum absolute error rate, MAER)、平均绝对误差 (mean absolute error, MAE)、均方根误差 (root mean square error, RMSE) 三项指标来进行评价。这三项指标的值越小就说明模型预测值与实际值的波动范围小, 该模型的预测稳定性就越高。

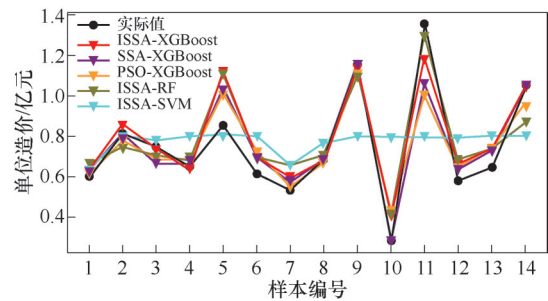


图9 模型对比结果

Figure 9 model comparison

对比结果如图10所示。

根据图9和图10可以看出, ISSA-XGBoost 模型在 MAER、MAE 与 RMSE 这三项指标上与其他四个模型相比表现都是最优的, 说明 ISSA-XGBoost 模型在处理过拟合问题及泛化回归预测方面要明显优于其他模型, 同时可以看出 SVM 的结果不是很理想, 或许是因为划分超平面的思想不足以应付这类数据集, 而 XGBoost 和 RF 这种集成基学习器的模型更适用。

3.4 模型验证

为了验证该模型在实际工程中的适用性, 本研究选取了3个在输入样本集之外的工程样本进行预测分析, 其具体数据如表8所示, 模型预测结果如表9所示。由表9可知 ISSA-XGBoost 模型对此3个样本的预测误差率分别为-8.87%、-13.44%、-7.94%, 满足前期投资估算的要求, 这也验证了该模型在高速公路工程造价预测方面的适用性。

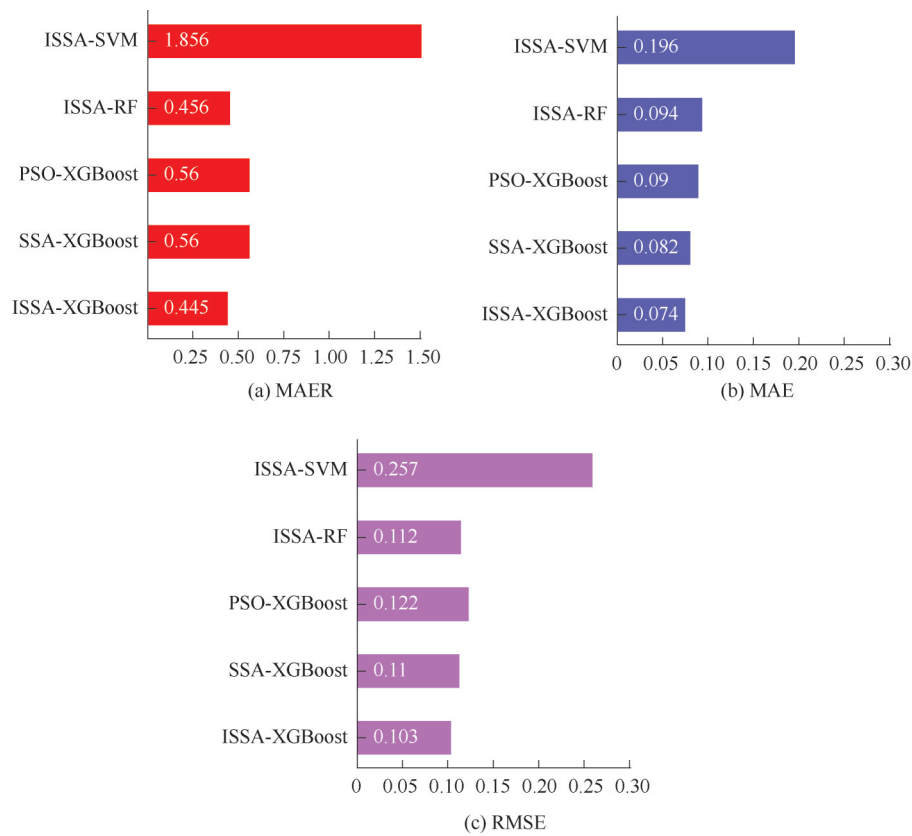


图10 模型评价指标对比
Figure 10 Comparison of Model Evaluation Indicators

表 8 高速公路验证样本
Table 8 Verification samples for expressways

名称	路线长度/km	路基 宽度/m	单位土石方 数量/10 ⁴ m ³	桥隧 比/%	互通立 交/处	单位永久 占地/hm ²	CPI	材料价格 指数
江苏阜兴泰高速公路兴化至泰州段	35.662	28	15.60	18.02	4	4.49	637.5	376.2
河南郑州至卢氏高速公路洛阳至洛宁段项目	69	26	26.04	18.53	4	6.19	579.7	393.4
上海至西安国家高速公路江都至六合段工程	76.1	34.5	18.85	17.51	11	7.75	579.7	393.4

表 9 模型预测结果
Table 9 Comparison of verification results

名称	单位实际造价/亿元	单位预测造价/亿元	相对误差率/%
江苏阜兴泰高速公路兴化至 泰州段	0.725	0.661	-8.87
河南郑州至卢氏高速公路 洛阳至洛宁段项目	0.626	0.542	-13.44
上海至西安国家高速公路 江都至六合段工程	0.896	0.825	-7.94

4 结论

本研究通过文献分析和统计方法,并结合工程实例,充分考虑项目特征和宏观经济指标对于高速公路造价的影响,最终确定了路线长度、路基宽度、单位土石方数量、桥隧比、互通立交、单位永久占地、CPI、材料价格指数八个造价影响指标作为输入特征,应用 ISSA-XGBoost 算法建立了高速公路工程造价预测模型,结论如下。

(1) 基于容易陷入局部最优的基本麻雀算法,本文开发了一种融合正余弦算法和 Lévy 飞行的改进麻雀算法,大大提高了麻雀算法寻优的效率和精度。

(2) 相较于 PSO-XGBoost、ISSA-RF、SSA-XGBoost 和 ISSA-SVM 模型,ISSA-XGBoost 模型表现出更好的抗过拟合性能。同时根据各评价指标结果,ISSA-XGBoost 模型预测精度满足投资估算的要求,并且具有良好的适应性与可解释性。同时改进麻雀算法提高了传统 XGBoost 模型的训练效率与预测精

度。根据模型特征重要性分析结果,桥隧比、路线长度和单位土石方数量这三个特征对高速公路单位造价的影响最为显著,而两个宏观经济指标——CPI 和材料价格指数对单位造价也有一定的影响。

(3) 通过随机样本验证,本文提出的 ISSA-XGBoost 模型能够有效对全国各地高速公路造价作出预测,具有广泛性与普适性,并且该预测模型可以迁移至其他领域,为构建高效且完善的高速公路造价管理体系提供了一定的依据。

虽然本文建立的 ISSA-XGBoost 高速公路造价预测模型具有较高的预测精度和效率,但仍存在一些不足,如特征指标的筛选有待进一步研究和完善等,同时机器学习需要基于大量的样本数据,而本研究所使用的样本数其实是偏少的。在未来的研究中,可扩大样本量及特征类型,采用一些新的统计融合方法(如 PCA 主成分分析法等),不断强化统计学习模型的预测性能。

参考文献

- [1] 廖镇华. 改扩建高速公路工程造价控制及管理要点[J]. 工程建设与设计, 2020(11): 274-276.
Liao Z H. Key points of cost control and management of highway reconstruction and expansion project[J]. Construction & Design for Engineering, 2020(11): 274-276.
- [2] 陈晓莉. 投资项目评估[M]. 重庆: 重庆大学出版社, 2017.
Chen X L. Investment Project Evaluation[M]. Chongqing: Chongqing University Press, 2017.
- [3] 董娜, 卢泗化, 熊峰. 大数据背景下基于 ABC-SVM 的建筑工程造价预测[J]. 技术经济, 2021, 40(8): 25-32.
Dong N, Lu S H, Xiong F. Cost prediction in construction project based on ABC? SVM under the background of big data[J]. Journal of Technology Economics, 2021, 40(8): 25-32.
- [4] Tayefeh Hashemi S, Ebadati O M, Kaur H. Cost estimation and prediction in construction projects: A systematic review on machine learning techniques[J]. SN Applied Sciences, 2020, 2(10): 1703.
- [5] 冯瀚, 刘冰旂, 张玉鸿, 等. 基于 FCM 和 PSO-SVM 的电力工程造价预测模型研究[J]. 华东电力, 2014, 42(12): 2713-2716.
Feng H, Liu B Y, Zhang Y H, et al. Predicting model of power engineering cost based on the FCM and PSO-SVM[J]. East China Electric Power, 2014, 42(12): 2713-2716.
- [6] 刘少非, 侯大山. 基于支持向量机的铁路隧道洞身工程造价预测[J]. 铁道工程学报, 2022, 39(5): 108-113.
Liu S F, Hou D S. Construction cost prediction of main tunnel in railway tunnel based on support vector machine[J]. Journal of Railway Engineering Society, 2022, 39(5): 108-113.
- [7] 刘运鹄. 基于 IABC-LS-SVM 的水利工程项目建设成本估算与控制方法[D]. 大连: 大连理工大学, 2021.
Liu Y H. Construction cost estimation and control method of water conservancy project based on IABC-LS-SVM[D]. Dalian: Dalian University of Technology, 2021.
- [8] 彭军龙, 胡珂, 王梦瑶, 等. 基于 SSA-LSSVM 的住宅工程造价预测研究[J]. 长沙理工大学学报(自然科学版), 2023, 20(3): 137-145.
Peng J L, Hu K, Wang M Y, et al. Research on residential project cost prediction based on SSA-LSSVM[J]. Journal of Changsha University of Science & Technology (Natural Science), 2023, 20(3): 137-145.

- [9] 肖秋明, 刘昕尧. 基于 PSO-SVM 的山区营运高速公路边坡防治费用预测[J]. 长沙理工大学学报(自然科学版), 2022, 19(2): 120-128.
Xiao Q M, Liu X R. Prediction of slope prevention cost for mountainous operating expressway based on PSO-SVM[J]. Journal of Changsha University of Science & Technology (Natural Science), 2022, 19(2): 120-128.
- [10] Barros L B, Marcy M, Carvalho M T M. Construction cost estimation of Brazilian highways using artificial neural networks[J]. International Journal of Structural and Civil Engineering Research, 2018: 283-289.
- [11] AL-Zwainy F M S, Aidan I A A, Iraq A U C. Forecasting the cost of structure of infrastructure projects utilizing artificial neural network model (highway projects as case study)[J]. Indian Journal of Science and Technology, 2017, 10(20): 1-12.
- [12] Zhang Y X, Minchin R E Jr, Flood I, et al. Preliminary cost estimation of highway projects using statistical learning methods[J]. Journal of Construction Engineering and Management, 2023, 149(5): 04023026.
- [13] 赵伟佳, 罗德才, 陈方, 等. 基于 PCA-BP 神经网络的既有建筑改造成本预测[J]. 土木工程与管理学报, 2024, 41(2): 89-97.
Zhao W J, Luo D C, Chen F, et al. Cost prediction of existing building renovation based on PCA-BP neural network[J]. Journal of Civil Engineering and Management, 2024, 41(2): 89-97.
- [14] 刘仁杰. 基于模糊 BP 神经网络的建筑工程造价模型[D]. 济南: 山东建筑大学, 2023.
Liu R J. Construction project cost model based on fuzzy BP neural network[D]. Jinan: Shandong Jianzhu University, 2023.
- [15] 袁剑波, 殷婵. 基于卷积神经网络的高速公路建筑安装工程估价估算研究[J]. 工程研究——跨学科视野中的工程, 2023, 15(5): 446-455.
Yuan J B, Yin C. Cost estimation of expressway construction and installation engineering based on convolutional neural networks[J]. Journal of Engineering Studies, 2023, 15(5): 446-455.
- [16] Chen T Q. Scalable and intelligent learning systems[D]. Washington: University of Washington, 2019.
- [17] Zhang J S, Yuan J F, Mahmoudi A, et al. A data-driven framework for conceptual cost estimation of infrastructure projects using XGBoost and Bayesian optimization[J]. Journal of Asian Architecture and Building Engineering, 2024: 1-24.
- [18] Yan H Y, He Z, Gao C, et al. Investment estimation of prefabricated concrete buildings based on XGBoost machine learning algorithm[J]. Advanced Engineering Informatics, 2022, 54: 101789.
- [19] 王启正. 基于机器学习算法的住宅楼项目建筑成本预测方法研究[D]. 长春: 长春工程学院, 2022.
Wang Q Z. Research on building cost prediction method of residential building project based on machine learning algorithm[D]. Changchun: Changchun Institute of Technology, 2022.
- [20] 郭婧娟, 刘曜玮. 基于 XGBoost 的高速公路工程概算预测方法研究[J]. 公路交通科技, 2023, 40(3): 58-68.
- [21] Bayram, Savas, Al-Jibouri, et al. Efficacy of estimation methods in forecasting building projects' costs[J]. Journal of Construction Engineering and Management, 2016, 142(11): 5016012.1.
- [22] Abu Hammad A A, Ali S M A, Sweis G J, et al. Statistical analysis on the cost and duration of public building projects[J]. Journal of Management in Engineering, 2010, 26(2): 105-112.
- [23] 薛建凯. 一种新型的群智能优化技术的研究与应用[D]. 上海: 东华大学, 2020.
Xue J K. Research and application of a new swarm intelligence optimization technology[D]. Shanghai: Donghua University, 2020.
- [24] 毛清华, 张强, 毛承成, 等. 混合正弦余弦算法和 Lévy 飞行的麻雀算法[J]. 山西大学学报(自然科学版), 2021, 44(6): 1086-1091.
Mao Q H, Zhang Q, Mao C C, et al. Mixing sine and cosine algorithm with Lévy flying chaotic sparrow algorithm[J]. Journal of Shanxi University (Natural Science Edition), 2021, 44(6): 1086-1091.
- [25] 颜星星. 建筑工程造价中人工成本造价的影响以及应对措施[J]. 陶瓷, 2022(6): 189-191.
Yan X X. The influence of labor cost in construction engineering cost and its countermeasures[J]. Ceramics, 2022(6): 189-191.
- [26] 孙文凯. 中国劳动力流动问题研究[M]. 北京: 中国人民大学出版社, 2016.
Sun W K. Research on the Issue of Labor Mobility in China[M]. Beijing: China Renmin University Press, 2016.
- [27] Duzkale K A, Lucko G. Exposing uncertainty in bid preparation of steel construction cost estimating: I . Conceptual framework and qualitative C-I-V-I-L classification[J]. Journal of Construction Engineering and Management, 2016, 142 (10): 04016049.
- [28] Shahandashti M.S, Ashuri, et al. Highway construction cost forecasting using vector error correction models[J]. Journal of Management in Engineering, 2016, 32(2): 4015040.1.
- [29] Tong B, Guo J, Fang S. Predicting budgetary estimate of highway construction projects in China based on GRA-LASSO[J]. Journal of Management in Engineering, 2021(3): 37.
- [30] 石振香, 柯洪, 王卫波. 建设工程计价[M]. 北京: 中国计划出版社, 2022.
Shi Z X, Ke H, Wang W B. Construction Project Valuation[M]. Beijing: China Planning Press, 2022.

Research on Expressway Engineering Cost Prediction Based on Hybrid ISSA-XGBoost

LI Jue*, LIU Yang

School of Traffic Transportation of Engineering, Changsha University of Science and Technology, Changsha 410114, China

Highlights

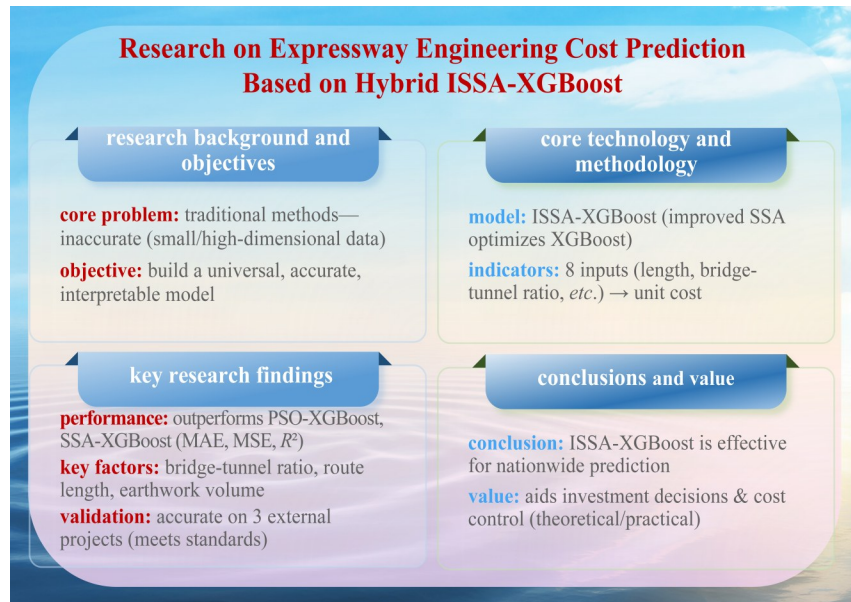
Proposes the ISSA-XGBoost hybrid model to solve prediction challenges with small, high-dimensional datasets.

The model's prediction accuracy and stability comprehensively outperform mainstream models like PSO-XGBoost.

Identifies bridge-tunnel ratio, route length, and earthwork volume as the three core cost-influencing factors.

Validated by real-world projects, the model is suitable for nationwide highway cost prediction and holds significant application value.

Graphical Abstract



Abstract: Expressway cost prediction is an important means of cost control in the preliminary stage of expressway construction projects. In this paper, for the characteristics of high dimensional and nonlinear relationship between the small sample data and the engineering cost feature indicators in engineering practice, the XGBoost algorithm with high training efficiency and good prediction performance is used to construct the prediction model, and at the same time, there are many hyperparameters of XGBoost, and the good or bad training result of the model depends on the selection of the hyperparameters, so this paper uses the Sparrow Search Algorithm (SSA) to search for the optimum of XGBoost hyperparameters.

Considering that SSA is easy to fall into the local optimum, a strategy that hybridises the sine-cosine algorithm and Lévy's flight is used to improve the basic SSA(ISSA), improve the global search ability of the algorithm and accelerate the convergence speed.

In the data collection part, this paper fully considers the influence of project indicators and macroeconomic indicators on the cost of expressways, and finally determines eight cost impact indicators as input features, namely, route length, subgrade width, quantity of earthworks per unit, the ratio of bridges and tunnels, number of interchanges, permanent land occupation per unit, CPI, and material price.

In the model establishing part, this paper firstly compares the iterative convergence curves of the three optimisation algorithms of PSO, SSA and ISSA, and it can be seen that the ISSA algorithm is better than SSA and PSO algorithm in both convergence speed and convergence accuracy, which verifies the robustness and effectiveness of the improved sparrow search algorithm.

At the same time, the improved sparrow search algorithm is used to optimise the three models of XGBoost, RF and SVM for comparison. The results show that compared with PSO-XGBoost, ISSA-RF, SSA-XGBoost and ISSA-SVM models, the ISSA-XGBoost model shows better antioverfitting performance. Meanwhile, according to the results of each evaluation metric, the ISSA-XGBoost model performs optimally compared with the other four models in the three metrics of MAER, MAE and RMSE, which indicates that the ISSA-XGBoost model is significantly better than the other models in dealing with the overfitting

problem and the generalised regression prediction, and that the model has a good adaptability and interpretability, and at the same time, the improved sparrow algorithm improves the training efficiency and prediction accuracy of the traditional XGBoost model.

According to the results of the importance of the model features, the three features of the ratio of bridges and tunnels, route length and quantity of earthworks per unit have the most significant influence on the cost of expressways, while the two macroeconomic indicators CPI and price index also have a certain impact on the cost.

In order to verify the applicability of the model in actual projects, this study selected three engineering samples outside the input sample set for prediction analysis, and the prediction error rates of the ISSA-XGBoost model for these three samples are -8.87%, -13.44%, and -7.94%, respectively, which meet the requirements of the pre-investment estimation.

Through the random sample verification, the ISSA-XGBoost model proposed in this paper can effectively predict the cost of expressways in various regions of the country, which is broad and universal, and the prediction model can be migrated to other fields, providing a certain basis for the construction of efficient and perfect expressway cost management system.

Keywords: expressway; cost prediction; Lévy flight; improve the sparrow algorithm; ISSA-XGBoost