

doi: 10.3969/j.issn.1002-0268.2023.03.008

基于 XGBoost 的高速公路工程概算 预测方法研究

郭婧娟¹, 刘曜玮^{*2}

(1. 北京交通大学 经济管理学院, 北京 100044; 2. 中交投资有限公司, 北京 100101)

摘要:我国长期以来,高速公路工程概算的编制是基于定额体系进行编制,而定额体系的确定具有滞后性、偏差性的缺点,导致所编制的高速公路工程概算存在偏差,而越来越大的投资规模和承包模式,对于投资资金的管控也提出了更高的要求。为解决该问题,经过文献分析与方法研究,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。以大量筛选指标和项目概算信息为样本库,使用网格搜索和交叉验证的方式优化了 XGBoost 算法模型,调整了模型相关的参数组合,优化了 XGBoost 算法模型的参数设置。以 $RMSE$, MAE , R^2 这 3 个常用预测精度评价指标组合评价作为评价体系,对模型的预测从精准性、稳定性和偏差性上进行了评价。最终提出了造价预测的新方案,并对所创建的 XGBoost 算法预测模型进行了抽样验证,展现该算法模型在高速公路工程概算预测过程中的预测能力,对推广造价指标体系、提高高速公路工程造价预测的精度具有一定的促进作用。同时也对造价指标的影响因素进行了详细分析,以特征重要性为指标进行分析与探究,探究对于我国高速公路造价指标的重要影响因素,探究影响因素与造价指标之间的潜在关系,为投资方和设计单位提供概算编制的新的辅助参考指标和研究方向。

关键词:道路工程; 预测模型; XGBoost 算法; 工程概算; 造价指标

中图分类号: U412; TP181; TU723.3

文献标识码: A

文章编号: 1002-0268 (2023) 03-0058-11

Study on Prediction Method of Expressway Project Estimate Based on XGBoost

GUO Jing-juan¹, LIU Yao-wei^{*2}

(1. School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China;

2. CCCC Investment Co., Ltd., Beijing 100101, China)

Abstract: For a long time in China, the compilation of expressway project estimates is based on a quota system, and the determination of the quota system has the disadvantages of hysteresis and deviation, leading to deviations in the compiled expressway project estimates. The increasing investment scale and contracting mode also put forward higher requirements for the management and control of investment funds. In order to solve the above problems, through literature analysis and method research, the prediction model of expressway project estimate is established based on XGBoost algorithm model and cost indicator system, and their advantages are analyzed. Using a large number of screened indicators and project estimate information as the sample library, the XGBoost algorithm model is optimized by grid search and cross validation, the relevant parameter combination of the model is adjusted, and the parameter settings of the XGBoost algorithm model are optimized. Using 3 commonly used prediction accuracy evaluation indicators ($RMSE$, MAE , R^2) as the evaluation system, the prediction accuracy, stability and deviation of the model are evaluated. Finally,

收稿日期: 2021-06-15

基金项目: 住建部软科学研究项目 (2021-R-089)

作者简介: 郭婧娟 (1972-), 女, 甘肃平凉人, 教授. (jjguo@bjtu.edu.cn)

* 通讯作者: 刘曜玮 (1993-), 男, 河北唐山人, 硕士研究生. (302920110@qq.com)

a new cost prediction scheme is proposed, and the established XGBoost algorithm prediction model is sampled and verified to show the prediction ability of the algorithm model in the process of expressway project estimate prediction. It has a certain promotion effect on popularizing the cost indicator system and improving the accuracy of expressway project cost prediction. Meanwhile, a detailed analysis of the influencing factors of the cost indicators is conducted, the importance of characteristics as the indicators are analyzed and explored, i. e., the important influencing factors of the expressway cost indicators in China are explored, the potential relationship between the influencing factors and the cost indicators are explored, which provides the investors and designers with new auxiliary reference indicators and research directions for estimate compilation.

Key words: road engineering; prediction model; XGBoost algorithm; project estimate; cost indicator

0 引言

在“交通强国”发展战略指引下,截至2022年底,我国公路总里程达到535万km,稳居世界第一。“十三五”规划期间,我国公路总投资规模将达到7.8万亿元人民币。在公路建设规模稳步扩大、投资需求持续增长的客观条件下,公路建设也迎来了高质量发展的转型时期。受到市场因素影响,公路工程成本投入波动大,资本运行风险高^[1],投资管控压力大。目前,对公路建设的投资管控,采用设计概算作为控制指标,但设计概算的编制主要依据政府定额体系对量、价、费等^[2]的规定,具有一定的局限性、滞后性和偏差性,对于工程特征和市场环境的变化不能及时应对和快速调整,也缺乏系统的公路工程造价指标体系作为衡量概算合理性的参考,导致了设计概算与实际投资偏差超出合理范围。随着投资主体多元化^[3],PPP和EPC等新型承包模式的广泛应用,从公路项目的规划、立项、投资、建设及后期的运营,建设主体都需要对资金的投入使用和保值增值给予预测、预警和监控,探索更为市场化、智能化的管控手段必要而迫切。

借助大数据和机器学习技术在预测算法上的强大优势,回归预测分析已被充分证实并广泛应用于房价、股市及天气预测等各个方面。本研究应用全国各省市高速公路工程概算原始数据,基于XGBoost算法,建立高速公路工程概算与工程特征、宏观经济因素之间的关联关系,开发高速公路工程概算预测模型,用于投资审批和监控过程中预测高速公路造价,衡量设计概算编制的合理性,同时也为投资者的管理与决策提供参考依据。

1 文献综述

高速公路工程具有项目周期长^[4]、资金投入大、工程技术复杂、项目地域分布分散等特征,其造价

工作也受到各方面因素影响,由于施工时间长、地质环境复杂^[5],还涉及到区域经济因素、政府相关法律等因素的影响,设计总概算不能有效地成为建设方进行造价管理的工具。高速公路工程概算的编制是基于定额体系以实物量法为基础进行的,定额是设计单位及咨询单位进行工、料、机计算消耗量的重要依据^[6]。根据公路工程建设市场上现行定额标准,劳动定额、材料消耗定额和机械台班定额是公路定额中的主要内容^[7]。在传统的定额体系中,虽然定额的发布呈周期性、大数据性的特点不断地在完善,但还存在着管理途径有限,缺乏现代理念的缺点。对于公路工程成本的管理和组织依然缺乏时效性和系统性,因此需要新造价预测方法来辅助投资建设单位进行资金的管控和管理层决策。

基于大数据理论,创建机器学习模型进行预测已在各个领域得到了广泛的应用。在国外的研究中,英国专家^[8]第1次将回归模型应用于造价预测。也有专家^[9]就回归预测方法进行进一步改进,使其更加完善和准确。Liu等^[10]针对计算方法进行深层次的研究,提高了模型的精度,提出了分阶段成本控制系数法。Shash^[11]研发了公路造价预测软件TRACER,用于公路工程造价的预测,但需要进行实时的维护与更新数据,需要大数据的支持。Mahfouz^[12]通过选取11个影响洛杉矶地区全部工程造价的工程量影响因子,构建了神经网络预测模型。随着“十四五”规划的到来,大数据中心建设与5G、工业互联网一同被提及,成为衡量大国信息化与现代化实力的一个重要指标。我国学者也对造价预测方面展开了诸多研究,周仁强^[13]通过典型工程法,利用已知的造价信息建立造价测算模型,得出各单项工程造价指数、综合指数及平均数指数的测算编制方法。任宏等^[14]为了进一步改善BP神经网络的性能,将学习自适应调整法和动量法引入模型。陈源等^[15]基于实例推理的估算方法利用灰色关联理

论对公路工程造价指标的影响因素进行预测, 得出了公路工程造价投资估算预测值案例推理的方法。

2016年, 陈天琦等^[16]针对多变量数据集提出了学习能力强、稳定性好、运算速度快的 XGBoost (eXtreme Gradient Boosting) 模型, 即基于决策树的提升算法, 在近年的建模竞赛上得到了很多的应用, 且已经被广泛应用于医学领域、房地产领域和金融领域。龚军等^[17]使用“单因素分析+logistic”筛选独立危险因素, 用于儿童脓毒性休克的危险因素筛查。张家棋等^[18]使用 XGBoost 算法建立预测模型进行房价预测, XGBoost 的学习优势已经在多领域得到了验证。

综上所述, 针对定额体系的缺点及工程概算在造价管理中的重要性, 本研究拟提出一种基于 XGBoost 算法的高速公路工程概算预测模型, 利用回归分析法, 基于多种造价指标的选取, 探索高速公路设计概算预测的新模式。

2 研究方法

2.1 XGBoost 模型

XGBoost 算法模型是采用加法模型与前向分步算法结合的算法模型。该模型基于 Boosting 方法, 通过将许多树模型结合的方式来提高计算的准确率要求, 通过不断地进行迭代, 生成很多新的树模型, 名为梯度提升树。前向分步算法指在叠加新的基础模型上进行同步优化, 即每次叠加的模型对上次模型拟合后的残差进行拟合, 从算法模型的角度解释即为决策树的加法模型:

$$f_M(x) = \sum_{m=1}^M T(x, \theta_m), \quad (1)$$

式中, M 为决策树的数量; $T(x, \theta_m)$ 为其中的某个决策树; θ_m 为对应决策树的所有参数合集, 其中, 假设 $f_0(x)=0$, 则相应的第 m 步的模型是:

$$f_m(x) = f_{m-1}(x) + T(x, \theta_m). \quad (2)$$

为求解对应的参数 θ_m , 需要最小化相应损失函数 θ'_m 来确定, 具体公式为:

$$\theta'_m = \underset{\theta_m}{\operatorname{argmin}} \sum_{i=1}^n L[y_i, f_{m-1}(x_i) + T(x_i; \theta_m)], \quad (3)$$

式中, x_i 为第 i 个特征值; y_i 为第 i 个真实值; L 为损失函数。

前向分步算法得到 M 棵决策树 $T(x, \theta_m)$ 后, 再进行叠加, 便可以得到提升树模型 $f_M(x)$, 参数值的求解依赖于损失函数的求解情况。

假设在提升树的模型中, 加入不停迭代到 k 棵

形成 f_k 树函数公式, 预测值 $\hat{y}_i$ 的函数关系表达式为:

$$\hat{y}_i = \varphi(x_i) = \sum_{k=1}^K f_k(x_i), f_k \in F, \quad (4)$$

式中, $F = \{f(x) = w_{q(x)}\} (q: R^m \rightarrow T, w \in R^T)$, 为所有回归树模型构建的函数空间; $q(x)$ 为 x 映射为决策树叶子节点索引的函数; R^m 为输入空间划分为 m 个区域; R^T 为 T 个叶子节点的输入空间区域; T 为叶子节点的数量; w 为叶子节点对应的权重。因此每次迭代生成的树, 即每次前一棵树 $k-1$ 所迭代出的预测值与实际值都会有一定差额, 这个差额即为损失量, 将损失量表示为损失函数 $L(\phi)$, 即为:

$$L(\phi) = \sum_{i=1}^n l(\hat{y}_i, y_i) + \sum_{k=1}^K \Omega(f_k), \quad (5)$$

式中, $l(\hat{y}_i, y_i)$ 为预测值与真实值的差异; $\Omega(f_k)$ 是一个正则项, 是使用叶子节点数与叶子节点对应值向量控制整个训练模型的复杂程度^[19]。假设迭代训练到第 t 棵树, 首先定义 $f_0(x_i) = \hat{y}_i^{(0)} = 0$, 则第 t 次迭代得到的模型 $\hat{y}_i^{(t)}$ 则为:

$$\begin{cases} \hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\ \hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \end{cases} \quad (6)$$

目标函数 Obj 经过 t 次迭代后进行累加, 即可表示为:

$$Obj^{(t)} = \sum_{k=1}^t l[y_i, \hat{y}_i^{(t-1)} + f_t(x_i)] + \Omega(f_i) + const. \quad (7)$$

对误差项进行二阶泰勒展开:

$$\begin{cases} Obj^{(t)} \approx \sum_{i=1}^n [l(y_i, \hat{y}_i^{(t)}) + g_i f_t(x_i) + 0.5 h_i f_t^2(x_i)] + \Omega(f_i) + const \\ g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \\ h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)}) \end{cases} \quad (8)$$

当训练到第 t 棵树时, g_i 和 h_i 为已知的量, 去掉公式中的常数项, 便是统一的第 t 次迭代的目标函数。

XGBoost 算法的优势在于对于样本数据集进行正则化后对过拟合修正, 擅长解决高速公路工程造价这类具有高维因子和大量数据特征的样本预测, 通过二阶导数的运算, 使损失更加精确, 降低最小损失函数, 而其算法的并行化也使运算效率得到更大

的提升。

2.2 建模流程

XGBoost 是基于梯度提升的一种目标函数计算方式,提高了模型的精度,并将目标函数的求解方式转化成二次函数求解最小值问题,将正则项转化为树模型的复杂度加入目标函数中,提升了模型的运算能力和泛化能力,建模过程如下:

(1) 数据预处理阶段,高速公路工程造价影响因子较多,影响机制复杂,使用 SPSS 软件和 Python 软件将选取的概算预测指标进行多重共线性、标准化等处理,剔除异常值,减小异常值对模型预测准确度的影响。

(2) 数据预处理后,将数据分为训练集和测试集,为模型的正常运行做好前期准备。

(3) 利用网格搜索和交叉验证的方法对 XGBoost 算法模型进行参数组合的优化和筛选,在训练集中训练出得分最高的最优预测模型参数组合并输出最优参数。

(4) 利用测试样本对步骤(3)所选取的模型参数进行检验,利用 R^2 , MAE , $RMSE$ 对测试集的结果进行评价分析,随后随机选取样本数据对 XGBoost 算法模型进行预测验证,检验模型预测的准确性和稳定性。

3 高速公路概算影响指标分析

3.1 指标的识别

影响高速公路造价的指标种类繁多,按来源主要包含项目内特征指标和宏观经济指标 2 类。本研究采用文献研究法和问卷调查法综合对相关特征指标进行识别和重要性排序。

(1) 基于文献研究的影响指标识别

筛选出有关高速公路工程造价影响因素分析及造价预测模型的国内外学术期刊文献 45 篇(国内 33 篇,国外 12 篇),将文献中所涉及到的影响指标按项目内部和外部宏观影响因素进行归类,文献研究统计结果详见表 1。

如表 1 所示,从文献的统计结果可见,对公路工程造价影响较大的项目特征指标有桥隧比(11 篇)、地貌特征(11 篇)、路线长度(7 篇)、路基宽度(6 篇)、通道数量(6 篇)、施工方案(5 篇)、建安费(5 篇)、土石方数量(5 篇);宏观经济指标有劳动力价格(9 篇)、CPI 指数(8 篇)、征地费标准(7 篇)、材料价格(6 篇)、银行贷款利率(5 篇)。

表 1 高速公路造价影响指标文献分析

Tab. 1 Literature analysis of influencing indicators of expressway cost

影响指标	文献数量	影响指标	文献数量
连接线数量	4	施工方案	5
路基宽度	6	所在地区	2
路线长度	7	CPI	8
车道数	3	GDP	2
桥隧比	11	PPI	2
土石方数量	5	银行贷款利率	5
防护工程数量	3	劳动力价格	9
通道数量	6	人均收入水平	1
底面层结构厚度	3	政策处理	3
地貌特征	11	标准规范变化	2
建安费用	5	投资渠道	1
征地费标准	7	材料价格	6

(2) 基于问卷调查的影响指标识别

在文献研究基础上,设计公路工程造价影响指标的调查问卷,剔除文献数量较少的所在地区(2 篇)、人均收入水平(1 篇)和投资渠道(1 篇)。将同类型指标进行合并,例如,将标准规范变化和政策处理合并为当地政府政策。将辅助性指标并入主指标,例如,将通道数量并入桥隧比,将施工方案并入建安费用指标。剩余 18 项指标进行调查分析。

问卷采用五点式李克特量表,由被调查者对问卷中的高速公路工程造价影响指标的重要程度进行判断。共发放 199 份调查问卷,发放对象中施工单位和设计单位占比 79.9%,一线员工及部门主办占比 80.4%,95.48% 的被调查者参与过高速公路工程建设工作。收回有效问卷 190 份,统计得到 18 项高速公路造价影响指标的均值和方差情况见表 2。

如表 2 所示,通过调查,共 7 项指标在“重要”分值以上,分别是路基宽度(3.49)、路线长度(3.46)、地形地貌(3.38)、桥隧比(3.35)、劳动力价格(3.34)、材料价格(3.29)、征地费指标(3.1)。

(3) 综合指标筛选

综合文献研究和问卷调查 2 种方法研究结果,对交叉重复部分进行合并。 CPI 指标在文献研究中所占比例较大,研究发现 CPI 指标属于国家宏观经济指标,一线工作人员对其影响作用并不明确,所涉及文献大多为国外学者研究文献,研究意义较大,因此纳入指标筛选范围中。

表 2 高速公路造价影响指标重要性排序

Tab. 2 importance ranking of influencing indicators of expressway cost			
序号	影响指标	均值	标准差
1	路基宽度	3.49	0.48
2	路线长度	3.46	0.48
3	地形地貌	3.38	0.48
4	桥隧比	3.35	0.46
5	劳动力价格	3.34	0.46
6	材料价格	3.29	0.45
7	征地费标准	3.1	0.41
8	车道数量	2.91	0.42
9	土石方数量	2.86	0.4
10	建筑安装费用指标	2.85	0.37
11	连接线数量	2.85	0.42
12	CPI (消费者物价指数)	2.84	0.42
13	防护工程数量	2.83	0.44
14	底面层结构厚度	2.8	0.39
15	银行贷款利率	2.78	0.39
16	GDP (国内生产总值)	2.77	0.44
17	PPI (生产价格指数)	2.75	0.4
18	当地政府政策	2.66	0.39

3.2 指标的选取

通过文献综述和问卷调查,选取 8 个指标作为影响高速公路工程造价的影响指标,其中路基宽度、路线长度、地形地貌和桥隧比属于项目内影响指标,劳动力价格、材料价格、征地费用指标和 CPI 指标属于宏观经济指标。为了检验所选指标的合理性,针对部分公路工程概算数据进行分析。从交通运输部路网监测与应急处置中心对各省公路项目造价指标的调研报告中,根据涉及到的路线长度、路基宽度、地貌特征等众多因素对比需求考虑,选取 2019—2020 年度广西省和河南省新建高速公路批复概算数据为样本,对指标进行验证分析,概算样本数据如表 3 所示。

(1) 征地费标准。所有样本项目的征地费用占概算造价总额的平均比例为 10.95%,说明征地费对高速公路工程造价具有较强的影响作用。

(2) 车道数与路基宽度。车道数与路基宽度相关性较高,进行统一分析,沙井至吴圩高速公路设计为六车道,广西省内其余高速公路均为四车道。由表中可见,在不考虑其他因素的情况下,该公路的每公里平均造价高于其他大部分四车道平均造价。

(3) 桥隧比。选取桥隧比 74.4%的南丹—天峨下老公路作为主要研究对象,不考虑其他因素的影响,其每公里平均造价最高,甚至高于六车道的沙井—吴圩高速公路。

表 3 概算样本数据

Tab. 3 Estimate sample data										
省份	项目名称	平原	山岭	设计时速/ (km·h ⁻¹)	路基宽 度/m	车道 数	路线长 度/km	桥隧 比/%	土石方数量/ (万 m ³)	防护及排水圬工 数量/(万 m ³)
广 西	1 田林—西林(滇桂界)高速公路	0	100	80	25.5	4	191.21	42.10	5 389.55	135.98
	2 南宁南过境线(吴圩机场— 隆安段)	0	100	120	28	4	44.88	11.86	765.37	16.47
	3 沙井—吴圩高速公路	0	100	120	34/43	6	25.60	12.89	808.11	21.64
	4 象州—来宾段初步设计	0	100	120	26.5	4	45.81	18.07	892.59	26.88
	5 蒙山—象州二期高速公路	0	100	100	26	4	56.67	19.31	1 285.44	32.01
	6 信都—梧州二期高速公路	0	100	120	26.5	4	52.86	30.58	1 503.66	22.18
	7 隆安—硕龙南宁段	0	100	120	27	4	43.17	30.73	1 594.28	11.45
	8 来宾—都安高速公路	0	100	100	26	4	134.04	48.83	1 818.49	39.27
	9 融安—从江高速公路	0	100	100	26	4	43.04	50.90	715.00	14.25
	10 南宁六景—宾阳高速公路	0	100	100	26.5	4	43.32	36.80	1138.79	28.57
	11 南丹—天峨下老公路	0	100	100	26	4	104.65	74.40	863.52	69.07
	12 连山—贺州高速公路	0	100	100	26	4	47.48	46.80	666.36	18.76
河 南	13 林州(豫冀省界)—桐柏高速公 路—北段	0	100	80	25.5	4	16.84	0.39	443.00	15.00
	14 林州(豫冀省界)—桐柏高速公 路—南段	100	0	100	26	4	42.10	0.12	1 009.00	43.00

续表 3

省份	项目名称	互通立交 数量/处	连接线 数量/ km	初步设计概算总价		初步设计概算建安费		征地拆迁补偿 费用/万元
				总额/ 万元	每公里造价/ (万元·km ⁻¹)	总额/ 万元	每公里造价/ (万元·km ⁻¹)	
1	田林—西林(滇桂界)高速公路	10.00	1.30	2 806 036.91	14 675.23	2 101 807.89	10 992.20	208 688.01
2	南宁南过境线(吴圩机场—隆安段)	4.00	7.63	510 196.42	11 367.51	338 395.74	7 539.68	830 93.18
3	沙井—吴圩高速公路	3.00	—	531 538.46	20 763.22	383 841.61	14 993.81	69 078.65
4	象州—来宾段初步设计	4.00	1.92	567 003.91	12 377.84	394 090.41	8 603.09	79 329.80
5	蒙山—象州二期高速公路	5.00	3.65	754 453.07	13 313.80	531 730.66	9 383.43	105 565.55
6	信都—梧州二期高速公路	4.00	10.56	758 273.35	14 345.48	563 221.89	10 655.38	80 896.75
7	隆安—硕龙南宁段	2.00	4.54	564 637.41	13 078.79	411 148.82	9 523.51	59 726.64
8	来宾—都安高速公路	5.00	12.50	2 125 477.77	15 857.40	1 642 909.79	12 257.14	165 247.61
9	融安—从江高速公路	5.00	13.04	808 600.13	18 789.36	641 136.53	14 898.03	51 079.56
10	南宁六景—宾阳高速公路	3.00	2.30	736 307.86	16 997.73	552 484.43	12 754.15	69 427.02
11	南丹—天峨下老公路	7.00	3.78	2 083 383.41	19 907.35	1 663 574.90	15 895.95	85 515.23
12	连山—贺州高速公路	6.00	6.00	804 976.22	16 955.08	617 451.45	13 005.28	57 018.65
13	林州(豫冀省界)—桐柏高速公路-北段	2.00	0.00	263 854.00	15 666.00	183 123.00	10 872.00	42 982.00
14	林州(豫冀省界)—桐柏高速公路-南段	5.00	1.32	465 610.00	11 061.00	289 750.00	6 883.00	110 094.00

数据来源: 交通运输部路网监测与应急处置中心各省公路项目造价指标调研报告。

(4) 地形地貌。地形地貌特征以平原微丘和山岭重丘的比例表示, 表中所示河南省林州(豫冀省界)—桐柏高速公路林州段为山岭重丘区新建高速公路, 在不考虑其他因素影响下, 每公里平均造价高于平原微丘区的项目。

通过对高速公路工程概算影响指标的识别和分析, 可以确定所筛选出的 8 项影响指标对造价的影响是相对较强的, 因此选用该 8 项指标作为 XGBoost 算法模型构建的自变量指标, 以此为基础对工程概算的预测研究进行分析与探究。

预测模型选取指标如表 4 所示。

表 4 预测模型指标

Tab. 4 Forecast model indicators

项目内指标	宏观经济指标
路基宽度	劳动力价格
路线长度	材料价格
地形地貌	占地拆迁费用指标
桥隧比	CPI

4 XGBoost 算法模型的构建与预测实现

4.1 模型构建与参数优化

从 sklearn 库中导入 XGBoost 模块, 导入所需样本指标, 进行多重共线性检验、标准化处理、异常

值处理、相关性分析等数据预处理, 随后设定参数调整范围: 包含总共迭代次数、学习速率(以避免在更新过程中的过拟合)、每棵树的最大深度和取值范围(默认值是 6), 树越深, 越容易过拟合、训练的实例样本占整体实例样本的比例(取值范围是 (0, 1], 默认为 1, 防止过拟合)。随后对样本指标以 7 : 3 的比例随机划分为训练集与测试集, 进入下一步训练阶段。

(1) 样本输入

使用交通运输部路网监测与应急处置中心发布的《2019 年高速公路造价指数指标分析报告》中的数据输入, 制订如下筛选标准。

①筛选高速公路工程项目。

②未标注路基宽度所占总路线长度比例和桥隧比情况的, 为保证预测模型的准确性, 进行剔除。

③仅保留单一数据类型的数据。

④仅保留新建高速公路工程项目。

最终整合出的数据共计 322 组, 作为初试输入样本。

(2) 样本预处理

①多重共线性检测

将数据进行多重共线性诊断, 得到结果如表 5 所示。

表 5 共线性诊断
Tab. 5 Collinearity diagnosis

模型	维	特征值	条件 指标	方差比例									
				常量	路线长度	山岭	路基宽度	桥隧比	征地指标	价格指数	钢筋单价	水泥单价	劳动力价格
1	1	8.428	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	0.56	3.878	0.00	0.07	0.08	0.00	0.00	0.47	0.00	0.00	0.00	0.00
	3	0.456	4.299	0.00	0.16	0.29	0.00	0.10	0.00	0.00	0.00	0.00	0.00
	4	0.268	5.6	0.00	0.69	0.05	0.00	0.01	0.20	0.00	0.01	0.00	0.01
	5	0.149	7.513	0.00	0.02	0.48	0.00	0.81	0.18	0.00	0.00	0.00	0.01
	6	0.067	11.231	0.00	0.00	0.02	0.01	0.07	0.01	0.00	0.02	0.01	0.98
	7	0.041	14.266	0.00	0.03	0.01	0.12	0.01	0.00	0.00	0.18	0.16	0.00
	8	0.018	21.571	0.00	0.00	0.01	0.01	0.01	0.11	0.00	0.72	0.68	0.00
	9	0.011	27.216	0.00	0.02	0.07	0.86	0.00	0.30	0.00	0.06	0.00	0.00
	10	0.000 018 15	681.344	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.01	0.14	0.07

根据 SPSS 软件中的结果输出,条件索引>10 或方差比例<0.5 时,自变量之间会出现严重的多重共线性,由表 5 可知,该样本中存在严重多重共线性问题,因此不能使用简单的多元线性回归模型,而使用 XGBoost 算法模型可以解决该问题。

②标准化及异常值处理

由于数据类型和统计量纲都不尽相同,在进行下一步数据分析之前,通常需要对数据进行标准化处理^[20]。随后使用标准化后的数据输出正态分布图和 QQ 图,观察其是否满足进一步数据预测的要求,直观的 QQ 图检验是常用的定性方法,在实践中应用广泛^[21]。QQ 图有 2 个作用:第一,检验 1 组数据是否服从某一分布。第二,检验 2 个分布是否服从同一分布。QQ 图全称是 quantile-quantile plot,从名称中可了解到是和分位数相关的图。正态分布图和 QQ 图如图 1 所示。

将造价指标较大的数值进行筛选,以 50 000 万元为分界点,对造价指标的样本进行了分割,筛选出 6 个异常值样本,如表 6 所示。

由表 6 可知,共计挖掘出 6 个异常值,异常值样本的共有特征为桥隧比均大于 90%,核对项目概况为桥梁或隧道导致造价指标异常。可见桥隧比极大的项目,由于单一原因导致的造价指标突高影响造价整体性结构,异常值的出现导致分布偏度非常大,需要对异常值进行处理。由于异常值只有 6 个,基本对样本容量无影响,遂采取删除处理。

(3) Grid SearchCV 参数优化

Grid SearchCV 即交叉验证+网格搜索的训练模式。本研究采用 5 折交叉验证的方式首先对训练组

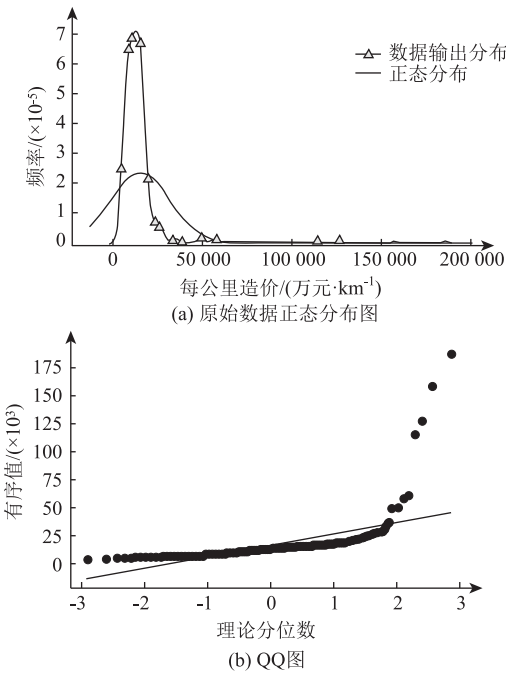


图 1 原始数据正态分布图与 QQ 图

Fig. 1 Original data normal distribution graph and QQ graph

再次进行划分,将训练组随机均分成 5 块,设定其中 1 块为预测集,其余 4 块为训练集,以 1 : 4 的方式进行交叉验证,随后选定另 1 块为预测集,其余 4 块为训练集再次训练,直到每块都被指定过预测集后,输出参数得分应用于测试集中进行预测分析。

本研究中需要选取出最优参数组合,因此在模型构建阶段便设定了每个参数的浮动范围,网格搜索的原理简单可以理解为“穷举法”,所有参数同时变化,进行交叉验证,这样每变动 1 次便又进行了 1 次 5 折交叉验证,通过将所有的参数范围都运算过后,所得出的得分最高的参数,即为最优参数。

表 6 异常值处理
Tab. 6 Outlier handling

序号	平原	山岭	路基宽度/m	路线长度/km	桥隧比/%	耕地补偿指标/(元·m ⁻²)	CPI	钢筋单价/(元·t ⁻¹)	水泥单价/(元·t ⁻¹)	劳动力价格/(元·工日 ⁻¹)	每公里造价/(万元·km ⁻¹)
22	100	0	32.00	1.69	100	86.40	101.70	2 920.00	290.00	87.00	57 107.10
35	100	0	33.50	4.93	100	270.00	103.10	4 080.00	576.30	102.00	126 721.00
36	100	0	26.00	10.03	90	270.00	103.10	4 080.00	576.30	102.00	157 674.27
58	100	0	33.50	7.91	90.35	202.50	101.90	3 412.00	380.00	71.00	114 726.07
153	0	100	35.50	32.40	90.07	547.50	103.40	4 360.00	534.57	220.00	59 851.45
158	100	0	41.00	24.00	100	525.00	102.30	3 910.00	445.00	200.00	186 390.00

将输入与预处理后的模型进行导入, 设定参数调节范围。XGBoost 算法模型中所涉及到的参数较多, 因此仅选用对模型预测精度影响较大的 4 个参数进行调节并设置调节范围, 在此基础上进行交叉验证和网格搜索, 对设置的参数进行优化选择。优化的指标为 XGBoost 模型中的评分系统, 网格搜索自动寻找评分最高的参数组合并进行输出, 获得在训练集中得分最高的参数组合并进行预测集的预测, 调节参数如表 7 所示。

表 7 XGBoost 参数设定范围
Tab. 7 XGBoost parameter setting range

参数	参数调节范围
总迭代次数	[500, 1 000]
学习速率	[0.1, 0.3]
最大深度	[3, 4, 5, 6]
实例样本占整体比例	[0.9, 1]

4.2 评价体系与重要性排序

模型预测结果是否符合要求, 参数得分并不能作为评价模型预测准确度和稳定性的唯一指标, 因此, 拟选用均方根误差 $RMSE$ 、拟合优度 R^2 、平均绝对误差 MAE 这 3 个评价指标同时对预测模型的预测能力进行评价。

$RMSE$ 亦称标准误差, 均方根误差是预测值与真实值偏差的平方与观测次数 n 比值的平方根。回归模型预测效果的验证中, 将验证数据样本代入回归模型, 计算获得参数估计值 (计算值), $RMSE$ 值越小, 证明预测的效果越好。

R^2 是指回归直线对观测值的拟合程度。测量拟合优度的统计量是可确定系数 (也称为确定系数) R^2 , R^2 最大值为 1。 R^2 的值越接近 1, 说明回归拟合的曲线与实际值的拟合程度越好。

MAE 的观测值和实际值之间误差绝对值的平均值, 它实际上是误差平均的一种更一般的形式。

$RMSE$ 相当于 $L2$ 范数, MAE 相当于 $L1$ 范数。次数越高, 计算结果就越与较大的值有关, 而忽略较小的值, 这就是为什么 $RMSE$ 针对异常值更敏感的原因, 即有 1 个预测值与真实值相差很大, 那么 $RMSE$ 就会很大。 $RMSE$ 受异常值的影响更大, 而 MAE 受到的影响小很多。同样, MAE 的值越小, 则预测模型的准确性和有效性越好。

(1) 预测值与真实值对比

应用最优参数下的高速公路工程预测模型对测试集的数据进行预测, 随后使用选取的评价指标体系对预测结果进行评价, 预测值与真实值对比如图 2 所示。该预测模型的预测能力和学习能力由图 2 可见一斑。

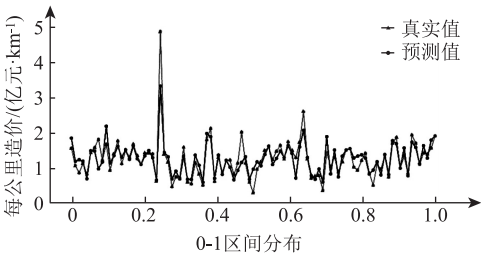


图 2 预测值与真实值的对比
Fig. 2 Comparison of predicted value and true value

(2) 预测模型评价体系

输出 $RMSE$, R^2 , MAE 指标, 如表 8 所示。

$RMSE$ 指标与 MAE 指标的得分越小, 预测效果越好, R^2 指标越接近 1, 预测模型的效果越好。由图可知, XGBoost 算法模型在 3 个评价指标中都有着出色的表现。

表 8 XGBoost 预测模型评价指标体系
Tab. 8 XGBoost prediction model evaluation indicator system

模型类别	$RMSE$	R^2	MAE
XGBoost	0.166 1	0.846	0.348 2

(3) 特征重要性排序

特征重要性 (Feature Importance) 衡量了特征在模型中的提升决策树构建中价值。一个特征越多地被用来在模型中构建决策树, 它的重要性就相对越高, 也可理解为该指标在整个预测模型中对于目标函数的影响权重。一个已训练的 XGboost 模型能够自动计算特征重要性, 这些重要性得分可以通过成员变量 feature_importances 得到。

本次高速公路工程造价 XGBoost 预测模型的特征重要性指标输出如图 3 所示。

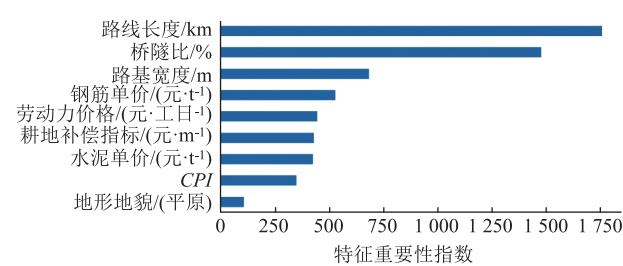


图 3 特征重要性排序
Fig. 3 Feature importance ranking

可以看出, 路线长度占据了特征指标重要性的首位, 则可认定在决策树迭代的过程中, 路线长度参与的树模型最多, 经过机器学习后认为其占每公里造价指标的权重最大。其次是桥隧比和路基宽度 2 个指标特征在 XGBoost 模型决策树迭代过程中也较多地被用来在模型中构建决策树, 其余特征值占较为近似, 地形特征值因为变化不明显, 所以处于权重的最后位置。

4.3 XGBoost 预测模型验证

最后, 随机选取样本库中的 2 组样本进行最优参数的预测分析, 分别选取江西抚吉高速和河南林桐高速的概算数据, 基础数据如表 9 所示。

将数据按指标类别输入模型中, 进行标准化处理, 标准化后数据集如表 10 所示。

将数据集标准化后, 代入选定的最优参数组进行模型预测, 预测值与实际值对比如表 11 所示。

由随机样本的预测值与实际值对比表可知, 误差在 4.9%~7%之间。而由图 3 预测值与实际值对比可知, 少部分样本的预测值与实际值差距较大, 差

表 9 高速公路相关基础数据

Tab. 9 Relevant basic data of expressway

序号	项目名称	路基宽度/ m	路线长度/ km	桥隧比/%	耕地补偿指标/ (元·m ⁻²)	CPI	钢筋单价/ (元·t ⁻¹)	水泥单价/ (元·t ⁻¹)	劳动力价格/ (元·工日 ⁻¹)	平原	山岭
1	抚州—吉安高速公路	25.50	16.80	38.50	75.00	103.00	3 987.00	440.00	106.00	0	100
2	林州—桐柏高速公路	26.00	179.00	12.00	30.00	105.20	4 950.00	480.00	89.00	100	0

表 10 数据集标准化

Tab. 10 Data set standardization

序号	路基宽度/ m	路线长度/ km	桥隧比/%	耕地补偿指标/ (元·m ⁻²)	CPI	钢筋单价/ (元·t ⁻¹)	水泥单价/ (元·t ⁻¹)	劳动力价格/ (元·工日 ⁻¹)	平原	山岭
1	1.324	-0.032	1.267	-0.675	-0.953	-0.793	-0.823	-0.292	-0.829	0.829
2	-0.319	-0.657	-0.581	-0.671	-0.531	-0.900	-1.072	-0.488	1.221	-1.221

表 11 预测值与真实值对比 (单位: 万元/km)

Tab. 11 Comparison of predicted value and real value
(unit: ×10⁴ yuan/km)

序号	省份	名称	实际值	预测值	误差
1	江西	抚州—吉安高速公路	5 276.9	5 239	37
2	河南	林州—桐柏高速公路	15 666	15 588	77

距较大的样本造价指标在 50 000 万附近, 原因可能为桥隧比极大而造成指标突高, 预测模型针对此类数据的预测能力较弱, 但大部分样本数据预测较为准确。这表明该参数选择下的 XGBoost 算法模型的优化效果较好, 预测误差较小, 预测效果比较符合预期, 可以实现投资方概算编制阶段的辅助预测、

辅助决策和辅助投资管控的功能。

5 结论

本研究综合考虑我国高速公路项目内特征指标和宏观经济指标对于工程造价的影响, 通过多种调研统计途径进行影响指标的筛选和分析, 共计筛选出路线长度、路基宽度、桥隧比、地形地貌、CPI、材料价格、劳动力价格和征地费用指标 8 个造价影响指标, 应用 XGBoost 算法综合影响指标建立高速公路工程造价预测模型, 得出以下结论。

(1) 依据特征重要性分析, 路线长度、桥隧比、路基宽度这 3 个指标对高速公路工程造价影响较大,

材料价格、劳动力价格、征地费用指标这4个指标对造价具有一定程度的影响,而地形地貌对造价具有较小的影响。

(2) 应用XGBoost算法建立我国高速公路造价预测模型是可以实现的,相较于其他回归预测算法,XGBoost算法在处理高维存在多重相关性的大样本数据具有显著优势,并行算法和增加正则项可以有效提高学习速率降低预测误差。

(3) 通过RMSE, MAE和 R^2 评价体系的检验,在322组数据的训练和预测过程中,XGBoost算法预测模型的表现较好,通过网格搜索和交叉验证进行算法优化,可以得到更加优化的模型参数组合,效果较好,可以推广到同类型其他高维复杂数据的预测模型中。

本研究提出了一种基于XGBoost算法的我国高速公路工程概算预测模型,由于受到样本容量限制,为保证预测精度和稳定性,选用全国各省数据综合建模,未能针对各省市不同的经济环境和项目特征进行模型细化,无法个性化表达各省市的造价影响因素,在以后的研究中,随着新基建和信息化的不断发展,可针对各省市的不同造价特征进行建模研究,分析造价指标影响因素。

参考文献:

References:

- [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, 15 (11): 274-276.
LIAO Zhen-hua. Key Points of Cost Control and Management of Highway Reconstruction and Expansion Project [J]. Construction & Design for Project, 2020, 15 (11): 274-276.
- [2] 周晓航, 褚春超, 周健. 经营性公路合理回报测算方法研究[J]. 公路交通科技, 2009, 26 (2): 145-149.
ZHOU Xiao-hang, CHU Chun-chao, ZHOU Jian. Study on Measuring Method of Reasonable Return of Commercial Highway [J]. Journal of Highway and Transportation Research and Development, 2009, 26 (2): 145-149.
- [3] 高娜. 定额计价在工程造价管理中的应用分析[J]. 建筑技术开发, 2020, 47 (23): 109-110.
GAO Na. Analysis on Application of Quota Valuation in Engineering Cost Management [J]. Building Technique Development, 2020, 47 (23): 109-110.
- [4] 董留群. 论高速公路工程项目全生命周期风险的识别[J]. 项目管理技术, 2008 (8): 37-41.
DONG Liu-qun. On Identification of Life Cycle Risk of Expressway Engineering Project [J]. Project Management Technology, 2008 (8): 37-41.
- [5] 黄嘉明. 地形复杂地段高速公路的造价控制分析[J]. 工程技术研究, 2020, 5 (12): 163-164.
HUANG Jia-ming. Cost Control Analysis of Expressway in Complex Terrain [J]. Engineering and Technological Research, 2020, 5 (12): 163-164.
- [6] 伊力哈木·艾尔肯. 工程造价计价模式和造价控制的研究[J]. 广西质量监督导报, 2020 (11): 33-34.
ILHAMU E. Research on Project Cost Pricing Mode and Cost Control [J]. Guangxi Quality Supervision Guide Periodical, 2020 (11): 33-34.
- [7] 文上勇, 张继钢. 工程造价市场化改革下定额的再认识与建议[J]. 建筑经济, 2021, 42 (1): 74-76.
WEN Shang-yong, ZHANG Ji-gang. Re-recognize and Suggestions about the Quota Based on the Market Reform of Construction Cost [J]. Construction Economy, 2021, 42 (1): 74-76.
- [8] TEICHOLZ P. Forecasting Final Cost and Budget of Construction Projects [J]. Journal of Computing in Civil Engineering, 1993, 7 (4): 511-529.
- [9] LOWE D J, EMSLEY M W, HARDING A, et al. Predicting Construction Cost Using Multiple Regression Techniques [J]. Journal of Construction Engineering and Management, 2006, 132 (7): 750-758.
- [10] LIU L, ZHU K, et al. Improving Cost Estimates of Construction Projects Using Phased Cost Factors [J]. Journal of Construction Engineering and Management, 2007, 133 (1): 91-95.
- [11] SHASH A A. Factors Considered in Tendering Decisions by Top UK Contractors [J]. Construction for Management and Economics, 1993, 11 (2): 111-118.
- [12] MAHFOUZ T. Application of Latent Semantic Analysis for Conceptual Cost Estimates Assessment in the construction Industry [C] // International Workshop on Computing in Civil Engineering 2011. [S. l.]: [s. n.], 2011.
- [13] 周仁强. 工程造价指数的信息采集和编制研究[D]. 重庆: 重庆大学, 2005.
ZHOU ren-qiang. Research on Information Collection and Compilation of Project Cost Index [D]. Chongqing: Chongqing University, 2005.
- [14] 任宏, 周其明. 神经网络在工程造价和主要工程量快速估算中的应用研究[J]. 土木工程学报, 2005 38 (8): 135-138.
REN Hong, ZHOU Qi-ming. Application of Neural Network for Quick Estimation of Engineering Construction Cost and Main Quantities [J]. Journal of Civil Engineering, 2005, 38 (8): 135-138.
- [15] 陈源, 崔文浩, 赖应良. 基于案例推理的山区高速公路

- 路工程造价估算研究 [J]. 价值工程, 2016, 35 (33): 37-38.
- CHEN Yuan, CUI Wen-hao, LAI Ying-liang. Study on the Estimation of Mountainous Expressway Project Cost Based on Case Reasoning [J]. Value Engineering, 2016, 35 (33): 37-38.
- [16] CHEN T, GUESTRIN C. XGBoost: A Scalable Tree Boosting System [C] // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S. l.]: [s. n.], 2016: 785-794.
- [17] 龚军, 钟小钢, 谈军涛, 等. “网格搜索+XGBoost”算法建立儿童脓毒性休克预测模型 [J]. 解放军医学杂志, 2020, 45 (12): 1270-1276.
- GONG Jun, ZHONG Xiao-gang, TAN Jun-tao, et al. Establishment of Children's Septic Shock Prediction Model with “Grid Search + XGBoost” Algorithm [J]. Medical Journal of Chinese People's Liberation Army, 2020, 45 (12): 1270-1276.
- [18] 张家棋, 杜金. 基于XGBoost与多种机器学习方法的房价预测模型 [J]. 现代信息科技, 2020, 4 (10): 15-18.
- ZHANG Jia-qi, DU Jin. House Price Prediction Model Based on XGBoost and Multiple Machine Learning Methods [J]. Modern Information Technology, 2020, 4 (10): 15-18.
- [19] DONG W, HUANG Y M, BARRY L, et al. XGBoost Algorithm-based Prediction of Concrete Electrical Resistivity for Structural Health Monitoring [J]. Automation in Construction, 2020, 114: 103155.
- [20] 李鹏飞, 何桢. 多元质量控制中数据处理的几个问题 [J]. 河北工业大学学报, 2005, 34 (4): 44-47.
- LI Peng-fei, HE Zhen. Some Problems of Data Manipulation in Multicontrol [J]. Journal of Hebei University of Technology, 2005, 34 (4): 44-47.
- [21] 魏艳华, 王丙参, 张艺馨. 利用蒙特卡罗方法对QQ图检验的改进与比较 [J]. 统计与决策, 2020, 36 (16): 13-17.
- WEI Yan-hua, WANG Bing-shen, ZHANG Yi-xin. Improvement and Comparison of QQ Chart Test by Using Monte Carlo Method [J]. Statistics and Decision, 2020, 36 (16): 13-17.
-
- (上接第 41 页)
- LI Fei-fei, XIE Kang-he, DENG Yue-bao. Analytical Solution for Consolidation by Vertical Drains with Exponential Flow under Vacuum Preloading [J]. Journal of Central South University (Science and Technology), 2015, 46 (3): 1075-1081.
- [17] 许波, 雷国辉, 郑强, 等. 考虑涂抹区压缩及渗透系数变化的堆载预压固结解 [J]. 岩土力学, 2014, 35 (6): 1607-1616.
- XU Bo, LEI Guo-hui, ZHENG Qiang, et al. Solution of Consolidation by Surcharge Preloading Considering Compressibility and Varying Hydraulic Conductivities in Smear Zone [J]. Rock and Soil Mechanics, 2014, 35 (6): 1607-1616.
- [18] 武崇福, 曹海莹, 刘光秀. 改进的砂井地基等效建模法及工程优化 [J]. 公路交通科技, 2012, 29 (8): 36-43.
- WU Chong-fu, CAO Hai-ying, LIU Guang-xiu. Improved Equivalent Modeling Method and Engineering Optimization for Sand Drain Foundation [J]. Journal of Highway and Transportation Research and Development, 2012, 29 (8): 36-43.
- [19] 董志良. 堆载及真空预压砂井地基固结解析理论 [J]. 水运工程, 1992 (9): 1-7.
- DONG Zhi-liang. Consolidation Theory on Heaped Load & Vacuum Preloading of Sand Drain Foundation [J]. Port and Waterway Engineering, 1992 (9): 1-7.
- [20] 郭彪, 龚晓南, 卢萌盟, 等. 真空联合堆载预压下竖井地基固结解析解 [J]. 岩土工程学报, 2013, 35 (6): 1045-1054.
- GUO Biao, GONG Xiao-nan, LU Meng-meng, et al. Analytical Solution for Consolidation of Vertical Drains by Vacuum-surcharge Preloading [J]. Chinese Journal of Geotechnical Engineering, 2013, 35 (6): 1045-1054.
- [21] HANSBO S. Consolidation of Fine-grained Soils by Prefabricated Drains [C] // Proceedings of the 10th International Conference on Soil Mechanics and Foundation Engineering. Publ Rotterdam: A. A. Balkema, 1981: 677-682.
- [22] 高长胜, 张凌, 汪肇京, 等. 塑料排水板的等效直径 [J]. 水利水运工程学报, 2002 (4): 28-32.
- GAO Chang-sheng, ZHANG Ling, WANG Zhao-jing, et al. Equivalent Diameter of Prefabricated Drains [J]. Hydro-Science and Engineering, 2002 (4): 28-32.