中文引用格式:马辉,贺鹰霞,陈杨杨. 基于 XGBoost 的城市污水管道缺陷发生概率预测[J]. 中国安全科学学报,2024,34(11):163-171.

英文引用格式:MA Hui,HE Yingxia,CHEN Yangyang. Prediction of urban sewage pipeline defect probability based on XGBoost [J]. China Safety Science Journal,2024,34(11):163-171.

基于 XGBoost 的城市污水管道缺陷发生概率预测*

马 辉 教授, 贺鹰霞**, 陈杨杨

(天津城建大学 经济与管理学院,天津 300384)

中图分类号:X928.03 **文献标志码**:A **DOI**: 10.16265/j.cnki.issn1003-3033.2024.11.0368

资助项目:教育部人文社科项目(22YJCZH022);天津市研究生科研创新项目(2022SKYZ330)。

【摘 要】 为提高城市污水管道缺陷检测效率,减少地毯式检测带来的资源浪费,降低环境安全风险,利用极致梯度提升(XGBoost)模型预测城市污水管道缺陷发生概率。首先,统计分析污水管道缺陷成因,筛选出能够表征管道缺陷状况的关键性指标,作为 XGBoost 模型的输入;其次,选择合适的目标函数和基学习器参数,利用网格搜索算法寻优基学习器的关键参数,完成模型训练和优化;最后,以广东省中山市某区域污水管网数据为例,验证 XGBoost 模型的有效性,根据模型输出寻找影响缺陷发生的主要因素和路径,并将区域内污水管网的缺陷发生概率划分出 4 个不同等级后进行可视化展示。结果表明:XGBoost 模型在 10 折交叉验证下的曲线下面积(AUC)均值达到 0.97,模型的预测准确率为 93%;管道埋深、坡度和长度 3 个特征对管道缺陷发生概率变化的影响程度最高;当管长增加,坡度越大、埋深越浅,污水管道发生缺陷的概率会随之增长。

【关键词】 极致梯度提升(XGBoost); 城市污水管道; 缺陷发生概率; 决策树; 预测模型

Prediction of urban sewage pipeline defect probability based on XGBoost

MA Hui, HE Yingxia, CHEN Yangyang

(School of Economics and Management, Tianjin Urban Construction University, Tianjin 300384, China)

Abstract: To improve the efficiency of urban sewage pipeline defect detection, reduce resource wastage resulting from indiscriminate inspection methods, and mitigate environmental safety risks, the XGBoost model was used to predict the probability of urban sewage pipeline defects. Firstly, the causes of sewage pipe defects were statistically analyzed to determine key indicators that can characterize the pipeline defects as the inputs of the XGBoost model. Secondly, appropriate objective functions and base learner parameters were selected. Then the model training and optimization were performed by a grid search algorithm to determine the key parameters of the base learner. Finally, the XGBoost model prediction performance was validated against an area of the sewage pipeline network in Zhongshan, Guangdong province. Moreover, the main factors and paths affecting defect probability were investigated based on the model output, and the defect probability of the sewage pipe network in the area was divided into 4 different levels for

* 文章编号:1003-3033(2024)11-0163-09; 收稿日期:2024-06-12; 修稿日期:2024-08-15

** 通信作者:贺鹰霞(1997—),女,四川达州人,硕士研究生,主要研究方向为土木水利工程项目管理。E-mail:872695175@qq.com。

visualization. The results indicated that the average area under curve (AUC) of the XGBoost model was 0.97 under 10-fold cross-validation with a prediction accuracy of 93%. Pipeline depth, slope, and length had the greatest impact on the probability of pipeline defect. As the pipe length increases, the sewage pipe defect probability will increase if the slope becomes greater and the buried depth becomes shallower.

Keywords: eXtreme Gradient Boosting (XGBoost); urban sewage pipelines; defect probability; decision tree; prediction model

0 引 言

城市污水管道所处环境复杂,易受多种因素影响,因而其状况难以持续保持完好,在运行负荷过重且管理粗放的情况下,管道易产生缺陷问题。这些潜在的缺陷逐渐积累,会引发加严重的环境问题和卫生事件。目前,我国多数城市仍采用传统的管理维护模式,对于污水管道的检测手段多为地毯式的盲目检测,或是等到问题发生再去被动抢修,不仅浪费人力物力,也严重影响居民的日常生活。因此,基于数据驱动和人工智能应用背景,构建一个污水管道缺陷发生概率预测模型,不仅可提前规划管道检测工作,避免紧急情况发生,还可减少管道全面检测的必要,降低成本,合理分配人力和物力资源。

国内外学者们针对城市污水管道缺陷状况评估预测的主要研究成果集中于 3 个方面:①基于系统工程的综合评价方法。HAWARI[1]建立了基于规则的模拟模型,利用模糊分析网络确定了影响值,从而判断管道的缺陷状况;ELHAM 等[2]针对暴雨天气下污水管道的溢流问题,提出基于层次分析法(Analytic Hierarchy Process,AHP)的状态评估模型,经验证后该模型成功用于评估管道的破损状况;罗同顺等[3]从污水管道的密封性、稳定性和功能性 3 方面出发,针对管道的结构缺陷,利用模糊综合评价法评估了中国南方某市的污水管道;徐得潜等[4]考虑了污水管道的爆炸、中毒、溢流污染和环境风险,使用 AHP 和灰色关联法评估了城市污水管道的风险;巴振宁[5]考虑管道失效风险因素之间的相互影响,提出基于改进模糊网络分析法的市政排水管网运行风险评估方法。②基于数理统计的评估方法。ALTARABSHEH 等[6]利用蒙特卡罗模拟法来评估管道故障风险,有效估计了实际工程中管道状况等级并确定管网中风险较高的管道;KABIR 等[7]在识别影响管道缺陷状况的显著变量后,使用贝叶斯优化 logistic 回归模型评估了污水管道的结构缺陷状况,获得显著效果;黄荣敏[8]提出一种基于风险指数的描述性统计方法,并探究了我国 W 市某排水区的排水管道属性与结构、功能以及健康状况的相关关系;③基于机器学习的评估方法。杨利伟等[9]建立遗传算法优化的反向传播(Genetic Algorithm Back Propagation,GA-BP)神经网络模型评估了污水管道的健康状态;郑茂辉等[10-11]使用遗传算法和粒子群算法优化神经网络模型,实现了管道结构性缺陷的分类诊断。

上述研究均有程度不同的局限性,综合评估方法主观性较强,多数情况下依赖于专家经验;数理统计方法很难对数据进行深度挖掘从而达到高精度与高适用性。鉴于此,笔者将选择机器学习法,并采用梯度提升树解决神经网络易产生的过拟合、梯度消失和爆炸问题。首先,基于缺陷成因分析,寻找管道缺陷状况的表征指标作为模型输入;然后,使用极致梯度提升(eXtreme Gradient Boosting,XGBoost)技术结合参数寻优构建污水管道缺陷发生概率预测模型;最后,以广东省中山市某区域污水管网为例,验证了模型有效性的同时,提取出影响缺陷发生概率的关键因素,并进行概率分布的可视化,为更加主动地维护和管理策略提供参考依据。

1 城市污水管道缺陷状况表征指标

1.1 城市污水管道缺陷成因

引起污水管道缺陷问题的原因多样且复杂,在许多内部因素和外部因素的相互作用、共同影响下造成了污水管道的缺陷问题,从而导致健康状态恶化。内部因素是指由管道自身存在的问题引起的缺陷因素,外部因素是指诸如温度、过负荷、外力干扰等管道运行过程中周围环境或运行环境给污水管道施加的运行压力导致管道出现缺陷的因素。参阅以往相关研究成果[12-14],管道缺陷成因大致可分为管道自身性状、外部环境因素、管道运行状况 3 大类,其中外部环境因素可分为地上环境因素及地下环境因素,见表 1。

1.2 城市污水管道缺陷状况表征指标选择

尽管每种缺陷类型都有其独特的影响因素,但

表 1　污水管道缺陷成因分析

Table 1　Sewage pipeline cause analysis

管道自身性状	外部环境因素		管道运行状况
	地上	地下	
管道年龄			
管道直径		埋深	负荷状况
管道材料	地面荷载	地下水位	水流速度
管道长度	人类活动	土壤质地	管内温度
接口形式	植被状况	土壤酸碱度	管内水质酸碱度
管道坡度	自然状况	环境温度	污染物浓度
安装质量		垫层材料	运维水平
管道自重			

通过聚焦于那些在多种缺陷类型中均能起到影响作用的因素,能够更全面地捕捉到污水管道缺陷状况的关键特征,解释多数情况下污水管道缺陷的变化情况。参考《城镇排水管道检测与评估技术规程》(CJJ 181—2012)、《城镇排水管道结构等级评定》(DB11/T 1492—2017)及《城镇排水管道功能等级评定》(DB11/T 1277—2015)等行业标准,选定管龄、管材、管径、管长、埋深、坡度、地面荷载、负荷状况、土壤质地作为后续管道缺陷状况判别工作基础。此外,由于直接计算污水管道可能受到的外部荷载难度较高,故以道路类型和用地类型间接表征路面交通荷载和负荷状况,最终筛选结果为:管龄、管材、管径、管长、埋深、坡度、道路类型、土壤质地、用地类型 9 个表征指标。

2　基于 XGBoost 的污水管道缺陷发生概率预测模型

2.1　XGBoost 模型原理介绍

XGBoost 本质上是一种梯度提升式的计算技术,它的基本思想是结合多个弱学习器(通常为决策树)的预测结果来构建一个强学习器,过程中每个弱学习器都是在尝试纠正前一个学习器的错误的基础之上构建的。更具体地,XGBoost 在训练过程中会不断优化一个目标函数,当其达到最小值时模型训练完成。

首先,给定数据样本 $\{x_i,y_i\}(i=1,2,\cdots,n)$,对每个样本构建一颗决策树,则会生成 k 棵树:

$$\hat{y}_i=\sum_{i=1}^{n}f_k(x_i),f_k\in F \tag{1}$$

式中:$\hat{y}_i$ 为预测值;n 为决策树的数量;x_i 为输入的第 i 个样本;$f_k(x_i)$ 为第 i 个变量的第 k 个决策树对应的目标函数;F 为所有可能的决策树的集合。

目标函数由损失函数和正则化项构成,损失函数用于拟合当前训练数据,正则化项用于表示学习器的复杂程度,目标函数公式如下:

$$L(\varnothing)=\sum_i l(\hat{y}_i,y_i)+\sum_k \Omega(f_k) \tag{2}$$

式中:$l(\hat{y}_i,y_i)$ 为第 i 个样本的损失函数;$\Omega(f_k)$ 为正则化项,越小抗过拟合能力越低。

为更快地优化目标函数,一般情况下会用二阶泰勒展开来近似原来的目标函数,第 t 次迭代的目标函数可用下式表述:

$$L^{(t)}\approx\sum_{i=1}^{n}l(y_i,\hat{y}^{(t-1)})+g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)+\Omega(f_t) \tag{3}$$

式中:$l(y_i,\hat{y}^{(t-1)})$ 为第 i 个样本前 $t-1$ 个弱学习器的损失值;g 和 h 分别为一阶偏导和二阶偏导,$g_i=\partial_{\hat{y}^{(t-1)}}l(y_i,\hat{y}^{(t-1)})$,$h_i=\partial^2_{\hat{y}^{(t-1)}}l(y_i,\hat{y}^{(t-1)})$。

其次,寻找最佳分割点,每棵决策树可写作:

$$f_t(x)=w_{q(x)},w\in R^T,q:R^d\to\{1,2,\cdots,T\} \tag{4}$$

式中:q 为每棵决策树的结构;T 为树中叶子节点数;w 为叶子节点的权重,每个 f_t 都通过训练产生各自的树结构 q 与叶子权重 w。树的复杂度函数与叶子节点数量和权重相关,其结构可表示为:

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda w^2 \tag{5}$$

式中:γ 为叶子节点数量惩罚正则项,用于限制决策树产生分支,起到剪枝效果;λ 为叶子节点权重惩罚正则项,起到抑制过拟合的作用。

为简化计算,所有被划分到第 j 个叶子节点的样本 x_i 可以组成一个样本集合 $I_j=\{i\mid q(x_i)=j\}$,根据式(4)和式(5)将式(3)进一步重构,得到:

$$L^{(t)}=\sum_{j=1}^{T}\left[\left(\sum_{i\in I_j}g_i\right)w_j+\frac{1}{2}\left(\sum_{i\in I_j}h_i+\lambda\right)w_j^2\right]+\gamma T \tag{6}$$

对于特定的树结构 $q(x)$,其目标函数的最小值可由式(6)对 w_j 求导得到,叶子节点 j 的最优权重 w_j^* 为:

$$w_j^*=-\frac{\sum_{i\in I_j}g_i}{\sum_{i\in I_j}(h_i+\lambda)} \tag{7}$$

由此得到树结构 $q(x)$ 目标函数最小值为:

$$L^{(t)}(q)=-\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j}g_i)^2}{\sum_{i\in I_j}(h_i+\lambda)}+\gamma T \tag{8}$$

式(8)可作为一个评价指标来衡量决策树的质量,值越小代表树结构越好。设 I_L 和 I_R 是节点分裂后左右节点的实例集,$I = I_L \cup I_R$,通过衡量拆分前后式(8)的差值,确定差值最大的分割点为最佳分割点:

$$L_s = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L}(h_i+\lambda)} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R}(h_i+\lambda)} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I}(h_i+\lambda)}\right] - \gamma \tag{9}$$

在最佳分割点进行分裂后,即完成了一次节点分裂,重复上述步骤,直至分裂带来的收益 L_s 小于设定的阈值或分裂至最大树身,即完成一棵决策树的训练。

2.2 缺陷发生概率预测模型评估指标

1) 混淆矩阵混淆矩阵是评估模型性能的一种重要工具,通过将模型的预测结果与实际情况相比,以此展示模型在各个类别上的表现。一个城市污水管道缺陷发生概率预测模型的混淆矩阵见表2。

表 2　二分类模型的混淆矩阵

Table 2　Confusion matrix for binary classification model

预测情况	真实情况	
	缺陷管道	正常管道
缺陷管道	TP	FP
正常管道	FN	TN

注:真阳性(True Positive,TP):XGBoost 模型正确地预测为缺陷管道的实际缺陷管道样本数。假阳性(False Positive,FP):XGBoost 模型错误地预测为缺陷管道的实际正常管道样本数。真阴性(True Negative,TN):XGBoost 模型正确地预测为正常管道的实际正常管道样本数。假阴性(False Negative,FN):XGBoost 模型错误地预测为正常管道的实际缺陷管道样本数。

2) 混淆矩阵仅统计了样本个数,仅用其来评估 XGBoost 模型的优劣是不充分的,基于上述内容,在评估模型分类性能时,延伸了以下指标:

真阳率(True Positive Rate,TPR):正确预测为缺陷管道的样本占实际缺陷管道样本的比例,计算公式如下:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

假阳率(False Positive Rate,FPR):错误预测为缺陷管道的正常管道样本占所有实际正常管道样本的比例,计算公式如下:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{11}$$

对于污水管道缺陷发生概率预测的问题,受试者工作特征曲线(Receiver Operating Characteristic Curve,ROC)是一系列分类阈值下,TPR 与 FPR 数值点的连线,提供了一种模型在所有可能阈值下性能的动态视图,其中,横轴为 FPR、纵轴为 TPR。一个理想的分类器模型的 ROC 曲线应尽可能地靠近左上角,即 TPR 越高越好、FPR 越小越好。

3) ROC 是一条蜿蜒的曲线,在实际应用中难以比较曲线的优劣,为避免这个局限性,采用 ROC 曲线下面积(Area Under Curve,AUC)[15] 作为模型泛化性能评估指标,它提供了一个将分类模型性能量化为单一数值的方法,范围为 0~1。当 AUC≤0.5,意味着模型的训练结果随机性较大,没有意义;当 AUC≤0.7,模型准确性较差;当 0.7<AUC≤0.9,表示模型准确性较好;当 AUC>0.9,此时模型的准确性非常高。

3 城市污水管道工程案例分析

3.1 研究数据准备与处理

以广东省中山市某区域的污水管网为例,实例验证所提方法的有效性。该区域内管道基础数据包含 2 部分:①污水管网地理信息系统(Geographic Information System,GIS)数据库,包含管材、管径、管长、管龄、埋深、坡度以及所在道路等;②该片区 2023 年的污水管网闭路电视(Closed Circuit Television,CCTV)监测数据。研究所用数据样本的管径范围为 DN100~N1500,管道长度总计 100.7 km。在污水管道 GIS 数据库中,管材、管径、管长、埋深、坡度等指标均有直接标注;部分重要的外部因素如道路类型、用地类型、土壤质地等缺乏详细记录,根据数据库中提供的位置信息在各类公开的大型地理数据集中检索查询,数据资料和来源见表3。最后,基于管段唯一性标识建立污水管网 GIS 数据和 CCTV 检测数据的对应关系,共提取有效样本数据 5 231 条,缺陷管段样本 4 135 条,正常管段样本 1 096 条,按 2 类样本的占比随机选取 3/4 的样本作为训练集,1/4 的样本作为测试集。

为避免各指标量纲和数量级不同造成的不平衡性,对连续型数据进行最大最小归一化,归一化公式如下:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{12}$$

式中:x' 为归一化后的数据;x 为输入的原始数据;

表 3　污水管道基础数据资料

Table 3　Basic data of sewage pipelines

数据来源	指标内容	指标参数形式
污水管网 GIS 数据库	管龄	连续型数据
	管径	连续型数据
	管材	分类型数据
	管长	连续型数据
	坡度	连续型数据
	埋深	连续型数据
	缺陷类型	分类型数据
OpenStreetMap 精准查询结果以及中山市兴趣点数据中有关道路类型的部分	道路类型	分类型数据
中国基本城市土地利用类型制图	用地类型	分类型数据
Google Earth Engine 土壤质地分类数据集以及中国科学院地理科学与资源研究所提供的中国土壤质地空间分布数据	土壤质地	分类型数据

x_{max} 为输入数据的最大值; x_{min} 为输入数据的最小值。

此外,以“是否存在缺陷”为样本标签,1 表示存在缺陷的管道,0 表示正常管道。对于分类型数据,采用数字编码的方式代替文本,以便后续训练模型时能够正常进行,以管材为例,编码内容见表 4。

表 4　管材数字编码

Table 4　Pipe material digital code

管道材质	混凝土管道	高密度聚乙烯管道	聚乙烯管道	硬聚氯乙烯管道
数字编码	1	2	3	4

3.2　XGBoost 模型参数选择与调节

选择正确的目标函数对于模型的性能至关重要,因为它直接影响着模型如何拟合训练数据。鉴于缺陷发生概率预测是一个二分类任务,即管道是否发生缺陷、概率有多大,故选择“binary:logistic”作为指定目标函数的参数值。基学习器(booster)参数通过控制 XGBoost 模型中决策树的生长过程、调节正则项中常数大小来提高模型的准确性,表 5 为基学习器参数选择及其作用。

在调节基学习器参数时,网格搜索法是一种非常彻底和系统的模型参数优化方法,它通过系统地遍历所有可能的参数组合来寻找最优解。模型训练过程中选择调用 Python 中 sklearn 库的 Stratified K-Fold 类和 GridSearchCV 类进行网格参数调节,寻优过程及结果见表 6。

表 5　基学习器参数名称及作用

Table 5　Booster parameters names and functions

参数名称	参数作用
learning_rate	该参数用于控制每一步迭代中模型参数的更新幅度
n_estimator	该参数表示要构建的树的数量,即最大迭代次数
max_depth	该参数定义了树的最大深度
min_child_weight	该参数用于控制树的生长
subsample	该参数指定了用于每次训练迭代的数据子集的比例
colsample_bytree	该参数用于控制每棵树在训练时随机采样的特征的比例
gamma	该参数用于控制树的生长,指定节点分裂所需的最小损失函数减少量
lambda	该参数用于 L2 正则化,防止过拟合
scale_pos_weight	该参数用于处理类别不平衡的问题

表 6　参数设置及寻优结果

Table 6　Parameter setting and optimization results

参数名称	范围	寻优步长	最优取值
learning_rate	(0,1]	0.1	0.23
n_estimator	(0,1 000]	10	180
max_depth	(0,10]	1	7
min_child_weight	(0,10]	1	1
subsample	(0,1]	0.1	1
colsample_bytree	(0,1]	0.1	1
gamma	(0,10]	0.1	0
lambda	(0,10]	0.1	1
scale_pos_weight	(0,30]	1	15

3.3　XGBoost 模型效果评估

K 折交叉验证常用于评估机器学习模型的泛化性能,避免模型出现过拟合问题。为保证模型的效果,模型采用 10 折交叉验证下的平均 AUC 值评估污水管道缺陷发生概率预测模型的整体性能,如图 1 所示。10 折交叉验证每次训练均可绘制出一条 ROC 曲线,每条曲线求得对应的 AUC 值,将 10 次训练在同一阈值下的 TPR 与 FPR 求均值,可得到 10 折交叉验证的平均 ROC 曲线,图中以粗实线表示。虚线为 ROC 曲线的临界线,ROC 曲线越接近该线则表示模型越差,越远离该线则表示模型精度越良好。由图 1 可知:10 折交叉验证每次训练的 AUC 均值达到 0.97。

XGBoost 模型赋予每个管道样本一个 0~1 之间的概率值,表示污水管道存在缺陷的可能性。为进一步验证模型对污水管道缺陷发生缺陷的预测能力,设置 0.5 作为决策阈值,即当模型输出的概率≥

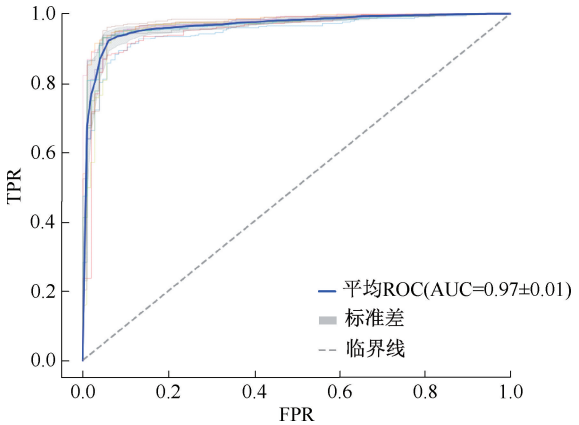

图 1　基于 10 折交叉验证的模型效果评估

Fig. 1　10 fold cross-validation model validation

0.5 时,判断管道存在缺陷;<0.5 时,判断管道为正常。表 7 以混淆矩阵形式给出 XGBoost 模型在测试集上的缺陷发生概率诊断识别效果。其中,模型正确识别正常管道 193 条样本,查准率为 95.54%,查全率为 70.44%;模型正确识别缺陷管道 1025 条样本,查准率为 92.62%,查全率为 99.13%,模型的整体准确率为 93%。图 2 给出测试集上各样本实际值与模型输出值的对照关系,结果表明:通过网格参数寻优的方法搭建起来的 XGBoost 模型能够较好地诊断识别污水管道缺陷发生概率,且识别精度能够满足实际应用要求。

表 7　混淆矩阵

Table 7　Confusion matrix

预测标签	真实标签	
	缺陷管道样本	正常管道样本
缺陷管道样本	1 025	81
正常管道样本	9	193

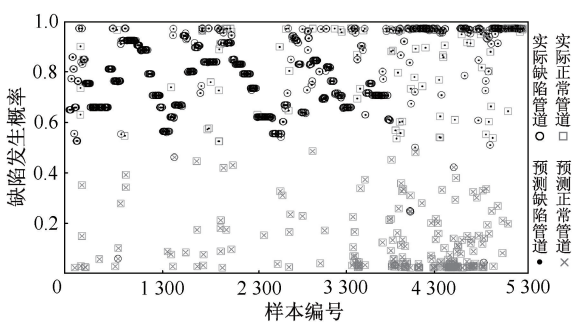

图 2　XGBoost 模型预测效果

Fig. 2　XGBoost prediction performance

4　污水管道缺陷发生概率预测分析

4.1　预测过程特征重要性分析

XGBoost 模型的一个显著优势在于能够对回归与分类目标中的独立变量按其重要性排序。在城市污水管道缺陷发生概率的模型训练过程中,模型会生成多棵决策树,在每棵树中选择最佳的分裂点将数据分成 2 个子集,以此方式递归构建整棵树,直到满足停止条件。一个变量被选为分裂点的次数越多,意味着它在预测缺陷发生概率中起到了更重要的作用。因此,通过统计一个变量在所有树中作为分裂点被使用的总次数,可得到一个权重,这个权重反映了该变量在模型中的重要性。例如:如果管龄在决策树 1、5、8 中分别分裂 1、5 和 4 次,那么管龄的权重将是 1+5+4=10。图 3 为 XGBoost 模型输出的特征重要性,埋深、坡度和管长是识别中山市某区域污水管网中管道缺陷发生概率的最重要变量。其他变量,如管材、用地类型和土壤质地,在 XGBoost 模型中的识别能力相对较低。

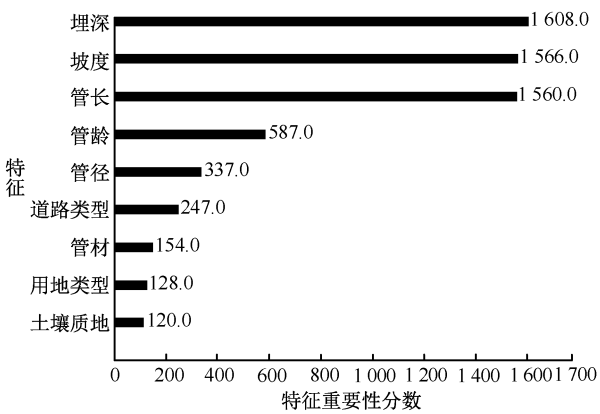

图 3　管道缺陷发生概率预测过程特征重要性排序

Fig. 3　Importance ranking of pipeline defect probability prediction process feature

此外,该模型可根据数据集中独立变量和目标变量来绘制出决策树,展示出决策树的不同层次以及独立变量的分割决策。决策树的分支和叶子是构成决策树的基本元素,揭示了如何基于独立变量的不同取值来进行决策分割,从而预测污水管道的缺陷发生概率。由于训练过程中生成的决策树较多,这里仅选取其中某一棵树展示,如图 4 所示。该树的第一个分割点展示了管龄对管道缺陷发生概率的影响。由此出发,污水管道被分为 2 组,一组是管龄小于 14 年的管道,另一组是管龄大于 14 年的管道。树的第 2 层将管道长度作为影响变量,在左侧节点中,分割点是 1.09 m,右侧节点则为 24.28 m。在第 3 层中,管道坡度作为第 3 个影响因素出现,此外,在该层中的最右侧还有一个叶节点,代表管道出现缺陷的概率。第 3 层右侧的叶节点显示,管龄超过

14 年且管长大于 24.28 m 的污水管道有 51%的可能性处于缺陷状态。除了管龄和管长外,第 4 层还涉及管道坡度和管径,右侧的叶节点显示,在管龄超过 14 年且管长大于 24.28 m 的基础上,当管道坡度小于 0.033 02 时,管道出现缺陷的概率达到 89%。在左侧,管龄小于 14 年且管长较短时,管道的状态较好。当管长增加,管道缺陷发生概率同时受到坡度、管径和埋深的影响,在多数情况下,埋深小于 2 m 时缺陷发生概率达到 70%以上。而较大尺寸的管道,在管龄较小时缺陷发生概率处于低位。

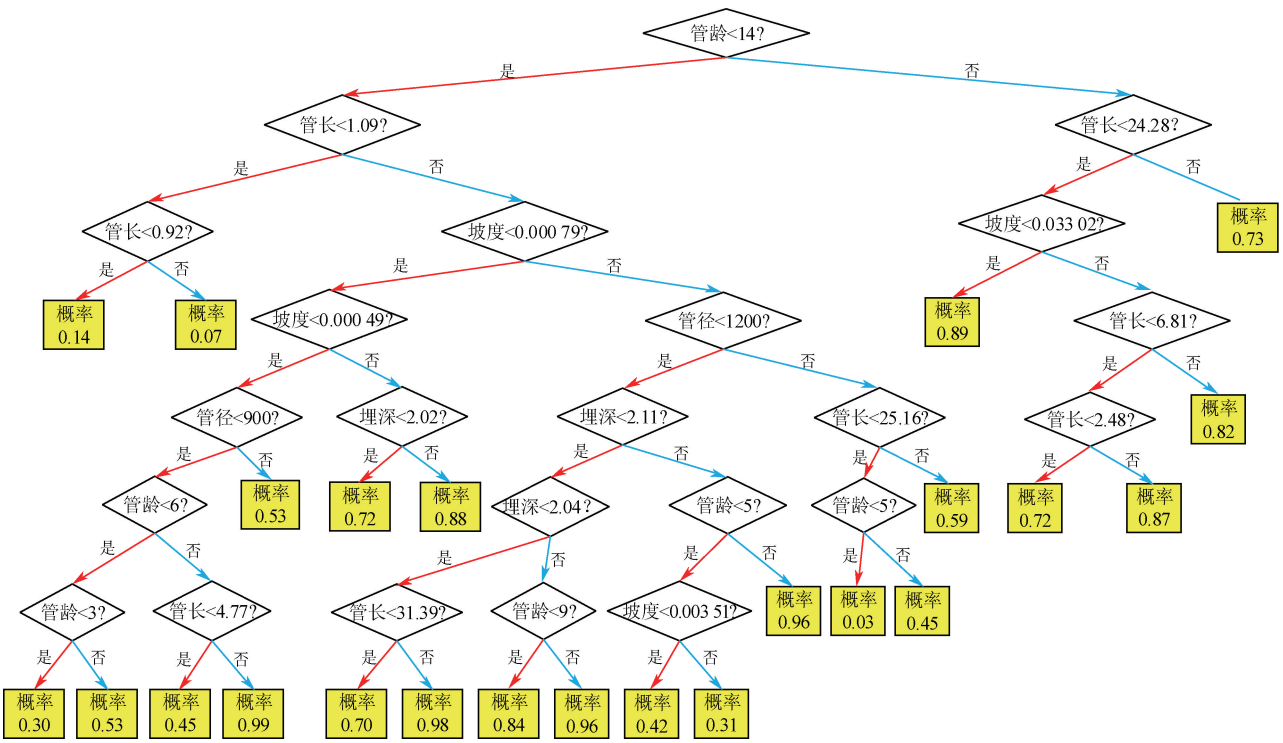

图 4　决策树生长过程

Fig. 4　Decision tree growth process

4.2　污水管道缺陷发生概率空间分布可视化

为更加直观地观察到该区域污水管网缺陷发生概率的空间分布情况,分级处理缺陷发生概率模型的识别结果,并在 ArcMap 中标注出来,实现污水管网缺陷发生概率识别结果的可视化。将污水管道缺陷发生概率以等间隔的方式划分为 4 类,并为每一类等级赋予不同的等级描述,代表不同的缺陷发生概率,划分结果见表 8。将该区域的污水管网缺陷发生概率识别结果按照表 8 进行概率等级分类,并利用 ArcMap 符号系统根据缺陷发生概率区间分级,将不同概率等级的污水管道由颜色深浅变化表示出来,可视化结果如图 5 所示。线段颜色越深表示发生缺陷的概率较高,污水管道正处于高风险状态,应引起检测单位的重视,及时安排管道巡检工作;反之,则表示污水管道处于相对平稳的状态,依据情况选择加强检修维护的频率或按照计划保证定期的巡检维护;颜色最浅处表示发生缺陷的概率极低,污水管道运行状况最平稳,对该类管道可适当减少巡检维护的频率,合理分配人力、物力和财力。

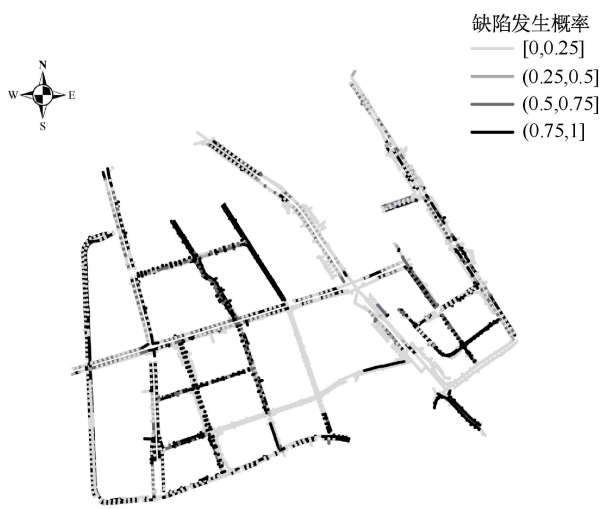

图 5　中山市某区域污水管网缺陷发生概率预测

Fig. 5　Probability prediction of sewage pipe network defects in a certain area of Zhongshan city

表 8 管道缺陷发生概率等级划分

Table 8 Classification of pipeline defect probability

缺陷发生概率等级	低概率	中低概率	中概率	高概率
缺陷发生概率区间	[0,0.25]	(0.25,5]	(5,0.75]	(0.75,1]

5 结 论

1) 提出基于 XGBoost 的污水管道缺陷发生概率预测方法,使用网格搜索优化参数设置,利用广东省中山市某区域污水管道数据仿真测试模型,结果表明:该方法能够有效提高污水管道缺陷发生概率预测的精确度,并具备良好的泛化性能。

2) 对污水管道特征变量进行重要度排序,结果显示,埋深、坡度、管长、管龄对缺陷是否发生的影响最大,当管长增加,坡度越大、埋深越浅,污水管道发生缺陷的概率会随之增长。

3) 将城市污水管道缺陷状况表征指标作为输入,利用决策树生成过程可以看到影响管道发生缺陷的关键因素,了解管道缺陷生成的主要路径,将缺陷发生概率划分为不同等级,帮助管理者安排更加具备针对性的管道检修计划,提高污水管网巡检效率。

4) 在后续研究中,可以加入管内污水流量、污水水质、管道内部压力等动态指标,全面反映污水管道缺陷状况的动态变化和实时情况。

参 考 文 献

[1] HAWARI A, ALKADOUR, FIRAS E. Condition assessment model for sewer pipelines using fuzzy-based evidential reasoning[J]. Australian Journal of Civil Engineering, 2018,16(1):45-67.

[2] ENNAOURI I, FUAMBA M. New integrated condition-assessment model for combined storm-sewer systems[J]. Journal of Water Resources Planning & Management, 2013,139(1):53-64.

[3] 罗同顺,左剑恶,干里里,等. 基于模糊综合评判模型的污水管道缺陷定量化评价方法[J]. 环境科学学报,2011,31(10):2 204-2 209.
LUO Tongshun, ZUO Jian´e, GAN Lili, et al. A quantitative evaluation method for sewage pipe defects based on fuzzy comprehensive evaluation model [J]. Acta Scientiae Circumstantiae, 2011,31(10):2 204-2 209.

[4] 徐得潜,张倩. 基于 AHP-GRA 的合流制污水管道风险评估[J]. 安全与环境学报,2019,19(4):1 149-1 154.
XU Deqian, ZHANG Qian. Risk assessment of combined sewage pipe based on AHP-GRA [J]. Journal of Safety and Environment, 2019,19(4):1 149-1 154.

[5] 巴振宁,王鸣铄,梁建文. 基于改进 F-ANP 方法的市政排水管网运行安全风险评估[J]. 安全与环境工程,2020,27(6):208-216.
BA Zhenning, WANG Mingshuo, LIANG Jianwen. Operational safety risk assessment of municipal drainage network based on improved F-ANP method [J]. Safety and Environmental Engineering, 2020,27(6):208-216.

[6] ALTARABSHEH A, MARIO V, AMR K. New approach for critical pipe prioritization in wastewater asset management Planning[J]. American Society of Civil Engineers, 2018,32(5): DOI:10.1061/(ASCE)CP.1943-5487.0000784.

[7] KABIR G, BALEK N B C, TESFAMARIAM S. Sewer structural condition prediction integrating bayesian model averaging with Logistic regression[J]. Journal of Performance of Constructed Facilities,2018, 32 (3):21-24.

[8] 黄荣敏,杜预,张浩,等. 基于风险指数法的排水管道健康状况影响因素研究[J]. 中国给水排水,2023,39(9):65-71.
HUANG Rongmin, DU Yu, ZHANG Hao, et al. Influencing factors of drainage pipeline health based on risk index method [J]. China Water & Wastewater, 2023,39(9):65-71.

[9] 杨利伟,邢雯雯,张莉平,等. 基于 GA 优化 BP 神经网络模型的污水管道系统健康状况评估[J]. 给水排水,2021,57(9):123-131.
YANG Liwei, XING Wenwen, ZHANG Liping, et al. Health assessment of sewage pipe system based on GA-optimized BP neural network model[J]. Water & Wastewater Engineering,2021,57(9):123-131.

[10] 郑茂辉,刘少非,柳娅楠,等. 基于粒子群优化极限学习机的排水管结构状况评价[J]. 同济大学学报:自然科学版,2020,48 (4): 513-516,551.
ZHENG Maohui, LIU Shaofei, LIU Ya´nan, et al. Evaluation of drainage pipe structure based on particle swarm

optimization extreme learning machine[J]. Journal of Tongji University:Natural Science,2020, 48 (4): 513-516,551.

[11] 郑茂辉,刘少非. GA 优化 ELM 神经网络的排水管道缺陷诊断[J]. 哈尔滨工业大学学报,2021,53(5):59-64.
ZHENG Maohui, LIU Shaofei. Ga-optimized ELM neural network for drainage pipe defect diagnosis[J]. Journal of Harbin Institute of Technology,2021,53(5):59-64.

[12] 王颖,王圃,王梓璇,等. 山地城市供水管网水质安全风险评价方法[J]. 中国安全科学学报,2023,33(8):205-211.
WANG Ying, WANG Pu, WANG Zixuan, et al. Water quality safety risk assesment method for water supply network inmountainous cities[J]. China Safety Science Journal,2023,33(8):205-211.

[13] 陈少博,杨宇轩,王浩,等. 南方滨海城市排水管网典型缺陷类型普查及成因机制分析[J]. 给水排水,2022,58(增 1):464-470.
CHEN Shaobo, YANG Yuxuan, WANG Hao,et al. Survey of typical defect types and analysis of cause mechanism of drainage pipe network in coastal cities of southern China[J]. Water & Wastewater Engineering, 2022,58(S1):464-470.

[14] 李若晗. 城市污水管道检测、评价与影响因素研究[D]. 北京:清华大学,2016.
LI Ruohan. Research on detection, evaluation and influencing factors of urban sewage pipeline[D]. Beijing:Tsinghua University,2016.

[15] 朱彤,秦丹,魏雯,等. 基于机器学习的公交驾驶员事故风险识别及影响因素研究[J]. 中国安全科学学报,2023,33(2):23-30.
ZHU Tong, QIN Dan, WEI Wen,et al. Research on accident nisk identification and infuencing factors of bus drivers based on machine learning[J]. China Safety Science Journal,2023,33 (2):23-30.

作者简介: 马 辉 (1979—),女,山西太原人,博士,教授,主要从事工程项目可持续建设管理、绿色建筑运营与管理等方面的研究。E-mail:tdmahui@ tju. edu. cn。

《中国安全科学学报》再次被收录为“中国科技核心期刊”

经过多项学术指标综合评定及同行专家评议推荐,《学报》再次被收录为“中国科技核心期刊”。根据《中国科技期刊引证报告(核心版)》(2024 年版),2023 年《学报》影响因子为 1. 702(2 165 种自然科学领域科技期刊的影响因子平均值为 1. 068),总被引频次为 3 946 次(2 165 种核心期刊均值为 1 673 次)。