DOI:10.11918/202501033

供水管线失效事件预测模型精度研究

侯本伟,周宝进,吴 珊

(北京工业大学 建筑工程学院,北京 100124)

摘 要:构建城市供水管线失效事件预测模型,可用于评估管线的失效可能性,是供水管网更新改造的重要依据。供水管线失效模型的建模方法包括分类和回归两类,现有失效模型研究往往采用其中1种方法进行案例分析,缺乏两种建模方法适用性和精度的比较。为此,基于某实例管网数据,采用随机森林(RF)、误差反向传播神经网络(BPNN)和支持向量机(SVM)3种机器学习算法,建立供水管线失效分类模型和回归模型。采用一致性指数(C-index)对比分类与回归模型的准确性,并使用分类指标与回归指标分别分析建模数据集划分方式与构成比例对供水管线失效模型的影响。结果表明:RF构建的失效模型均表现出最好的性能,分类模型的C-index比回归模型相应结果高5.4%~32.8%;与按照年份划分建模数据集的方式相比,随机划分建模数据集能够提升两类模型的预测精度;建模数据集构成比例对两类模型预测精度的影响存在差异,当未失效管线数据占比增大时,分类模型预测管线失效事件的准确度降低,而回归模型预测管线失效时间的误差减小。在实际构建供水管线失效模型时,需要根据对象数据集的特征,合理选择建模方法,并关注数据集的划分方式和构成比例对模型结果的影响。

关键词:供水管线;漏损事件;失效模型;分类模型;回归模型

中图分类号:TU911 **文献标志码**:A **文章编号**:0367-6234(2026)02-0012-10

Accuracy analysis of water supply pipeline failure prediction models

HOU Benwei, ZHOU Baojin, WU Shan

(College of Architecture and Civil Engineering, Beijing University of Technology, Beijing 100124, China)

Abstract: Constructing a predictive model for urban water supply pipeline failure events is crucial for assessing the likelihood of pipeline failures and serves as an important basis for the renovation and upgrading of water supply networks. The modeling methods for water supply pipeline failure models include classification and regression. Current research on failure models often employs only one of these methods for case analysis, lacking a comparison of the applicability and accuracy of both modeling methods. To address this gap, based on data from a specific instance of a water supply network, this paper establishes water supply pipeline failure classification and regression models using three machine learning algorithms: Random Forest (RF), Backpropagation Neural Network (BPNN), and Support Vector Machine (SVM). The concordance index (C-index) is used to compare the accuracy of the classification and regression models. Additionally, classification and regression indicators are employed to analyze the impact of modeling dataset division, as well as composition ratios of the dataset on the water supply pipeline failure models. The results show that the failure models constructed by RF exhibit the best performance, with the C-index of the classification models being 5.4% to 32.8% higher than that of the corresponding regression models. Compared to dividing the modeling dataset by year, randomly dividing the modeling dataset can enhance the predictive accuracy of both types of models. Furthermore, the impact of the modeling dataset composition ratio on the predictive accuracy of both types of models varies; as the proportion of non-failure pipeline data increases, the accuracy of the classification model in predicting pipeline failure events decreases, while the regression model shows reduced error in predicting pipeline failure times. Therefore, when constructing water supply pipeline failure models in practice, it is necessary to choose the modeling method appropriately based on the characteristics of the target dataset and pay attention to the impact of dataset division methods and composition ratios on the model results.

Keywords: water supply pipeline; leakage event; failure model; classification model; regression model

收稿日期:2025-01-14;**录用日期**:2025-02-22;**网络首发日期**:2025-05-09
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20250508.1706.002
基金项目:国家自然科学基金(52478486);北京工业大学城市更新科技创新基金(2024-4)
作者简介:侯本伟(1984—),男,博士,副教授
通信作者:吴 珊,wushan@bjut.edu.cn

供水管线是保障城市正常运行的重要生命线工程，在日常运维中由于管线自身和外部环境变化，导致管线漏损失效事件的发生，不仅造成大量水资源浪费，还会产生道路塌陷、区域生产生活中断等次生影响[1]。基于供水管线破损维修记录数据建立的供水管线失效模型，可用于评估管线失效事件发生的可能性。研究人员已经开发了多种依赖于历史数据的供水管线失效模型，这些数据包括管线破损记录、管线属性和水力模型[2]。采用更准确的供水管线失效模型可以带来显著的成本节省。供水管线失效模型在各种文献综述中已有详细描述[3-7]，这些文献综述中都有不同的分类方法来描述各种模型。Taiwo 等[7]综述了物理模型、基于统计的模型和基于机器学习的模型在预测管线失效概率中的应用，详细阐述了每种模型类型的优势和劣势，提供了一个全面的视角来审视管线失效模型。

物理模型试图分析作用于管线的荷载作用，从力学角度出发，计算管线内外环境与管线之间的相互作用，该方法需要严格的控制条件和现场测量数据，研究过程对数据要求较高，一般用于漏损成本较高的重要输水管线或关键供水管线[4]。统计模型运用统计学方法对管线失效记录历史数据进行分析，以识别各类影响因素（自变量）和管线失效（因变量）之间的关系。赵洪宾等[8]根据实际给水管网的漏失数据，采用时间序列分析方法建立了给水管网漏失预测的线性指数平滑模型和二次曲线指数平滑模型，发现管网漏损频率随时间非线性增加。Al-Ali 等[9]使用逻辑回归分析了 8 种类型管线的失效情况，从 43 个变量中识别出 16 个对回归方程有统计学意义的变量，模型预测准确率超过 70%。

基于机器学习的模型能够从一组数据中学习或识别特定模式，对新输入的数据进行未来预测，目前常用于供水管线失效模型。这些模型包括人工神经网络（ANN）、支持向量机（SVM）、极端梯度提升（XGBoost）和随机森林（RF）等，均可以用来解决回归问题和分类问题[10]。Fan 等[11]使用了 5 种机器学习算法（LightGBM、ANN、k-NN、SVM 和 LR）对管网中的每根管线进行分类，判断管线是否失效。Chen 等[12]通过结合供水管线历史失效记录数据，利用 RF、GBDT 和 XGBoost 3 种算法预测管线失效概率，其中，RF 算法显示出最高的预测准确性。侯本伟等[13]利用 BP 神经网络建立了管线破损数与天气指标、管径、管龄、管长的回归模型，预测未来 1 年不同时段的管线失效数。

基于机器学习的模型已被证明可以提高管线失效模型预测的准确性[14-15]。目前，供水管线失效模型的研究主要集中在利用或改进机器学习算法提升模型的精度。根据输出变量的类型，管线失效模型可分为分类和回归两种模型，由于模型应用案例管网特征和可用数据结构的差异，先前的研究中往往直接选择其中 1 种模型进行结果分析，没有分析两种建模方式的适用性。此外，失效模型构建过程中，对于建模数据集的处理方式以及管线失效数据与未失效管线在建模数据集中的比例等，均对失效模型的预测精度造成较大问题，这在先前的研究中也鲜有涉及。为此，基于北方某实例管网数据，采用 BPNN、SVM 和 RF 3 种机器学习算法构建管线失效分类模型和回归模型，利用一致性指数（C-index）等指标比较了两种模型的适用性特征。通过改变数据集划分方式和建模数据集构成比例，分析了模型数据集对两类失效模型预测结果的影响。

1　供水管线失效模型

1.1　分类模型与回归模型

在建立供水管线失效模型时，将收集到的管线信息和失效记录作为建模信息数据，模型会不断学习训练管线特征与失效状态之间的关联规律。失效模型构建所需要的数据集为 $\{S_1, S_2, \cdots, S_D\}$，$D$ 表示建模数据集数量。其中，$S_i = \{X_i; Y_i\} = \{x_{i1}, x_{i2}, \cdots, x_{in}; y_{i1}, y_{i2}, \cdots, y_{im}\}$，$X_i = \{x_{i1}, x_{i2}, \cdots, x_{in}\}$ 表示第 i 个数据的输入变量（自变量），$Y_i = \{y_{i1}, y_{i2}, \cdots, y_{im}\}$ 表示第 i 个数据输出变量（因变量），n 和 m 分别表示自变量和因变量的数量。供水管线失效模型一般为多输入单输出的问题。对于分类问题，其输出变量 y_i 为离散状态值集合 $\{C_1, C_2, \cdots, C_N\}$，$N$ 为类别数量；对于回归问题，其输出变量 y_i 为区间 $[y_L, y_U]$ 内的某个实数[16]。

在供水管线失效分类模型中，建模数据的类别标签为两类（$N=2$），类别标签“1”代表发生失效事件，类别标签“0”代表未发生失效事件，模型输出类别标签“1”的概率可视为管线失效概率。回归模型的输出变量一般包括管线失效率[8]、管线失效数量[13]、管线失效事件发生时间[17]等。在本文的研究中，以管线失效发生时间作为回归模型的输出变量。

1.2　建模数据集

在供水管线失效模型中，删失数据（censored data）是指在记录期内未发生失效事件的管线数据。删失数据分为左删失（left-censored）和右删失（right-censored）两种情况。右删失指在记录期内管线未发生失效事件。图 1i）中，记录期内有 1 根管线发生了失效事件（叉号表示），而另外两根管线在记录

期内未失效,这些管线的数据属于右删失数据。左删失指管线的敷设时间早于记录期的起始时间,且在记录期之前可能已发生失效事件,但未被记录。图 1ii)中的 3 根管线敷设时间均早于记录期,其失效情况在记录期之前未被记录。此外,图 1iii)中绿色标识的管线虽然发生了失效事件,但未被发现并记录,这种情况也属于左删失数据。

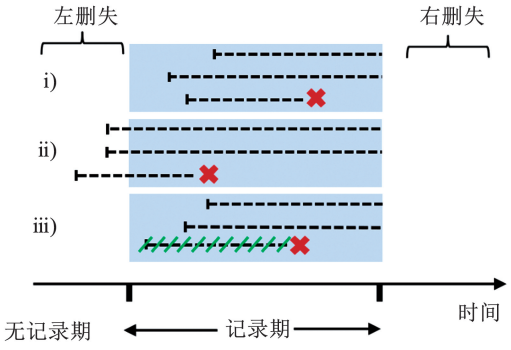


图 1 删失数据示意

Fig. 1 Censored data illustration

删失数据在供水管线失效模型建立过程中十分常见,在建立分类模型时,要抽取一定的删失数据与管线失效记录数据一起构成建模数据集,以平衡模型训练数据集,保证模型更好地学习管线失效数据的特征。回归模型的输出变量是管线失效时间,未发生失效事件的管线没有实际的失效时间,因此,在建模时会将删失数据排除在外,只采用管线失效数据进行建模。

建模数据集划分方式如图 2 所示,图 2(a)按照年份进行划分,使用 a_1 年 ~ a_{n-1} 年管线失效数据和部分删失数据 c_{n1}(删失数据量与失效数据量相同)作为训练数据,模型测试数据为第 a_n 年的管线失效数据以及删失数据 c_{n2}(与训练数据相互独立)。另一种将第 a_1 年 ~ 第 a_n 年所有失效数据进行混合(图 2(b)),然后随机划分模型训练数据集和测试数据集,为便于比较分析,测试数据集大小与第 a_n 年管线失效数据保持一致,删失数据保持不变。

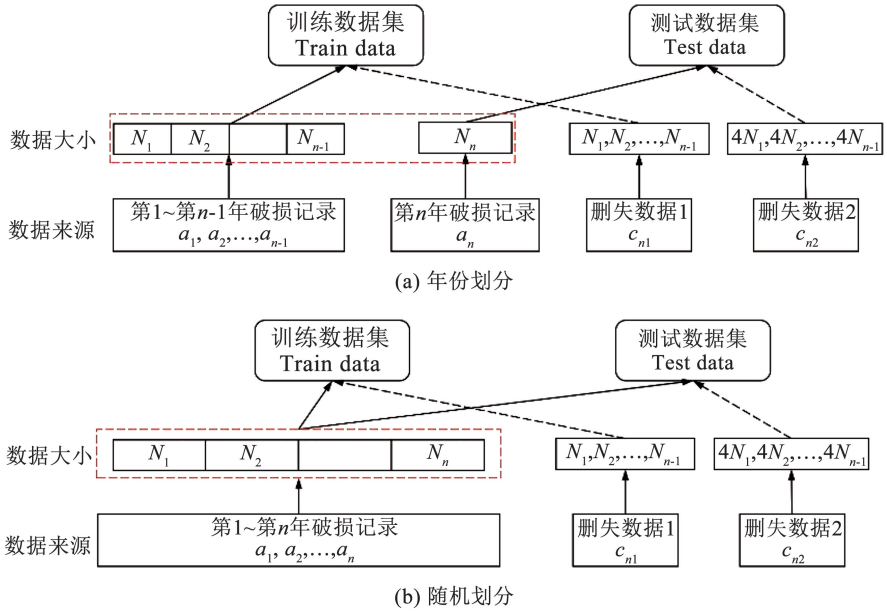


图 2 数据划分示意

Fig. 2 Data division illustration

1.3 建立失效模型的算法

主要采用 BPNN、SVM 和 RF 3 种常用的机器学习算法建立管线失效模型,这些方法在先前的研究中多次被使用[11-13,17]。

误差反向传播神经网络(BPNN)是一种多层前馈神经网络,通过反向传播算法调整网络权重和偏置,以最小化误差[18]。BPNN 的主要特点是信号向前传递和误差反向传播,输入信号从输入层经隐含层逐层处理,直至输出层,每一层的神经元状态只影响下一层神经元状态;如果输出层得不到期望输出,则转入反向传播,根据预测误差调整网络权值和阈值,从而使 BPNN 预测输出不断逼近期望输出。

支持向量机(Support Vector Machine,SVM)属于机器学习中的监督学习范畴,其核心思想是通过构造一个超平面最大化不同类别之间的间隔,从而实现对数据的有效分类[19]。该算法具有良好的泛化能力,尤其在高维特征空间中表现突出。通过使用核函数,SVM 能够处理非线性可分的数据,将其映射到更高维的特征空间,从而实现线性可分的分类效果。

随机森林(RF)是一种集成机器学习方法，它通过使用多个基础决策树提高模型的准确性和鲁棒性[20]。这种方法采用装袋策略(bagging strategy)，通过随机选择一组特征增强预测能力，旨在通过减少每棵树之间的相关性减轻过拟合问题。

1.4　模型评估指标

为了评估分类模型与回归模型的预测精度和有效性，采用以下 3 类评估指标，即分类指标、回归指标、排序指标。

分类模型的评估指标为混淆矩阵，混淆矩阵是评估分类模型最基本的工具，其可以将真实的管线失效情况与模型预测的管线失效情况进行比对，计算出真正类(TP，P_T)、假正类(FP，P_F)、真负类(TN，N_T)、假负类(FN，N_F)。通过混淆矩阵可以得到准确度(Accuracy，a)、召回率(Recall，r)、假正类率(FPR，R_{FP})、精确度(Precision，p)等指标用于评估分类模型性能。上述指标计算公式见式(1)～(4)。

分类模型还可以通过 ROC 曲线下的面积(AUC)表征模型分类性能优劣，AUC 值介于 0～1，越接近 1，表示分类性能越好。

$$a = \frac{P_T + N_T}{P_T + N_F + P_F + N_T} \tag{1}$$

$$r = \frac{P_T}{P_T + N_F} \tag{2}$$

$$R_{FP} = \frac{P_F}{P_F + N_T} \tag{3}$$

$$p = \frac{P_T}{P_T + P_F} \tag{4}$$

回归模型的评估指标采用决定系数(R^2)、均方根误差(E_{RMS})、平均绝对误差(E_{MA})以及平均偏差(E_{MB})，公式如下：

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{5}$$

$$E_{RMS} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{6}$$

$$E_{MA} = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \tag{7}$$

$$E_{MB} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i) \tag{8}$$

式中：y_i与 $\hat{y}_i$ 分别为在第 i 个管线的实际失效时间和预测失效时间，$\bar{y}$ 为所有管线实际失效时间的平均值，n 为管线数量。

R^2反映回归模型对因变量变异的解释比例，取值范围为 0～1，越接近 1，拟合效果越好。E_{RMS}是衡量预测值与实际值的差异标准度量，对异常值敏感，能识别极端情况下的表现；E_{MA}表示预测值与实际值的平均绝对差异，对异常值不敏感；E_{MB}衡量预测值与实际值的平均差异，可直观反映模型偏差。

一致性指数(简称 C-index，C_{index})，是一种用于评估模型性能的统计指标，在生存分析和临床预测模型中广泛应用，大于 0.7 表明模型较为准确[21]。C-index 的计算涉及将所有研究对象随机地两两配对。对于每一对个体，如果生存时间较长的个体预测生存时间也较长，或者预测生存概率较高的个体实际生存时间也较长，则认为预测结果与实际结果一致。C-index 即所有这样的一致配对数占所有可能配对的比例。

$$C_{index} = \frac{1}{m}\sum_{i;d_i=1}\sum_{j;y_i<y_j} I[\hat{y}_i < \hat{y}_j] \tag{9}$$

式中：m 为所有可能配对的数量，$d_i = 1$ 表示管线发生过失效事件，y_i与 y_j分别表示管线 i 和 j 实际的值，$I(\cdot)$为指示函数，$\hat{y}_i$ 与 $\hat{y}_j$ 为管线 i 和 j 预测的值。

2　案例分析

选择北方某城市的供水管网数据作为研究对象，该管网管线长度超过 6 000 km，其中约 20% 的管线服务年限超过 30 a。管线失效记录发生在 2009—2019 年，共有超过 10 000 条失效记录数据。研究聚焦于 4 种主要的管线类型，即普通铸铁(CI)管、球墨铸铁(DI)管、镀锌(GS)钢管和钢塑复合(SP)管，这些管线长度占据了管网总长度的 90% 以上。CI 和 GS 管线铺设始于 1950 年，DI 和 SP 管线则是在 1990 年之后才开始铺设。CI 和 DI 管线通常用于直径超过 50 mm 的配水干线，GS 和 SP 管线则主要用于直径在 15～50 mm 的配水支线。CI 和 DI 管线分布在道路区域和非道路区域，道路区域占比约 60%；GS 和 SP 管线主要分布在非道路区域(小区内)，占比约 80%。管线平均压强水头范围为 20～45 m，大多数管线的压强水头范围集中于 25～35 m。

图 3 显示了按管线年龄分布的管线失效统计。4 种管材对应的管线失效频率峰值年龄有所不同，CI 管线约为 21 a，DI 管线约为 9 a，GS 管线约为 18 a，SP 管线约为 13 a。CI 管和 GS 管开始敷设时间较早，两种管线失效随管龄变化规律类似，符合 Weibull 分布，DI 管线与 SP 管线变化规律类似。

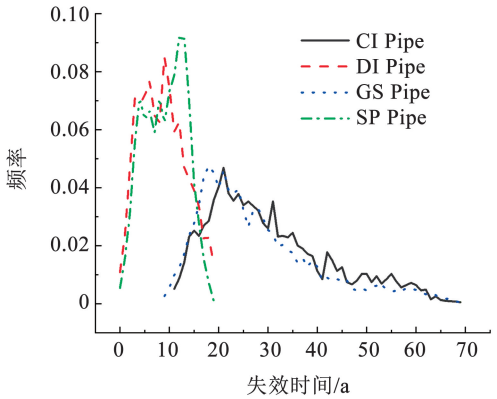


图 3 不同类型管线失效时间情况统计

Fig. 3 Statistic of failure time for different types of pipeline

2.1 模型方法比较

2.1.1 建模方法的比较

在对分类模型和回归模型进行具体分析之前，先选择一个合适的机器学习算法。选用 RF、BPNN、SVM 3 种方法建立管线失效分类模型和回归模型。以年份划分数据集的方式，将 2009—2018 年的所有管线失效数据作为训练数据，将 2019 年的失效数据作为测试数据得到模型预测结果。在图 4 中，RF 分类模型的准确度(Accuracy)为 0.825，AUC 为 0.78，均大于 BPNN 分类模型和 SVM 分类模型。RF 回归模型的 R^2 和 C-index 分别为 0.59 和 0.82，均大于 BPNN 回归模型和 SVM 回归模型。分类模型和回归模型中 RF 的表现均优于其他两种方法，在后续分析中均使用 RF 算法。

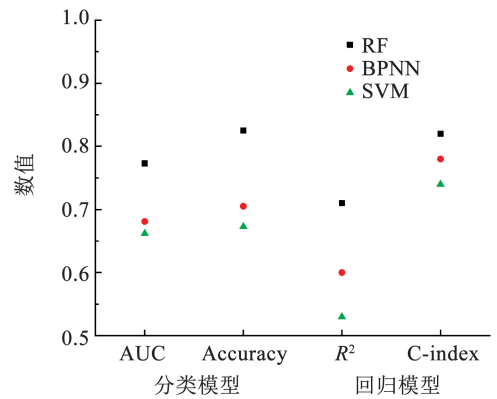


图 4 3 种机器学习方法性能评估结果

Fig. 4 Performance evaluation results of three machine learning (ML) methods

2.1.2 输入变量重要度分析

基于管网管线数据和失效记录，分类模型和回归模型的输入变量包括管径(D)、管材(M)、管龄(A)、管长(L)、压力(P)、道路类型(R)和埋深(H)，其中，回归模型中管龄为因变量，未作为输入变量分析。从表 1 可以看出，基于分类模型和回归模型两种不同建模方式所得到的输入变量重要性有所区别。管龄、管长在分类模型中相对更重要，而管材在回归模型中最为重要，结果差异源于两种建模方式的目标和数据处理方式不同。

表 1 两类模型的输入变量重要性

Tab. 1 Importance of input variables in two types of models

模型	D	M	A	L	P	R	H
分类	4.45	5.17	13.65	13.05	6.51	4.73	11.95
回归	3.54	15.78	—	3.57	5.54	1.49	2.10

2.1.3 分类与回归的比较

AUC 适用于分类模型，能够评估模型在不同阈值下的分类能力，而 C-index 适用于回归模型，能够评估模型对连续变量预测的一致性。AUC 和 C-index 均可反映模型在预测结果中的性能优劣，通过比较两种指标大小，可以直观地评估分类模型和回归模型在管线失效预测中的相对性能。从表 2 可以看出，RF 分类模型 4 种管材计算得到的 AUC 值相比 RF 回归模型得到的 C-index 高 5.4% ~32.8%。

表 2 两种模型的预测性能比较

Tab. 2 Comparison of predictive performance of two models

预测模型	RF 分类模型(AUC)	RF 回归模型(C-index)
CI Pipe	0.85	0.64
DI Pipe	0.78	0.61
GS Pipe	0.78	0.74
SP Pipe	0.73	0.62

进一步比较两类建模方式的结果差异，分析 CI Pipe和 GS Pipe 的预测结果，分类模型和回归模型分别按照预测失效概率和预测失效时间排序，统计一定预测管线失效数下对应的预测管线累积长度，如图 5 所示。CI Pipe 中，分类模型的预测管线累积长度始终小于回归模型，若预测出一半的管线失效数时，分类模型的累积管线长度为 46.7 km，比回归模型小 32.8%。但图 5(b)中 GS Pipe 表现出不同的结果，分类模型的预测管线累积长度大于回归模型，表 2 GS Pipe 的一致性指数大于 0.7，回归模型准确性较高，在进行排序时管线失效数据排名靠前。此外，CI Pipe 和 GS Pipe 两种管线的管材、管径、管长、失效记录存在差异，也会影响模型的结果。

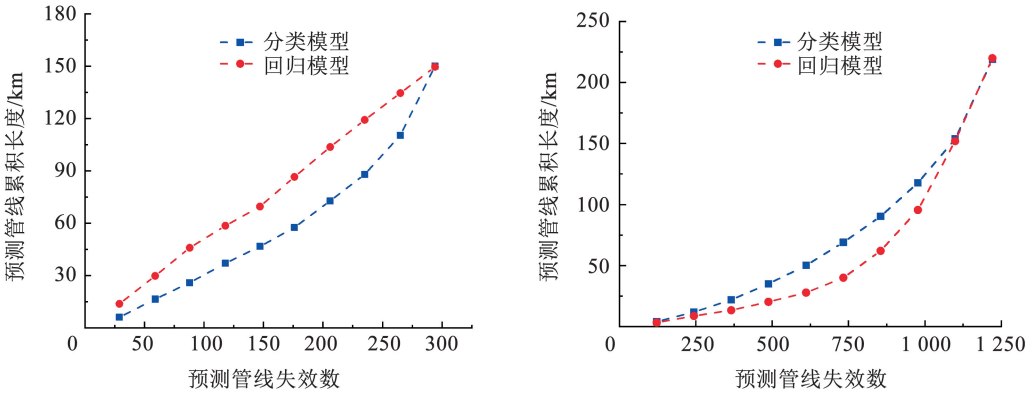


图 5　预测失效管线失效数与累积管线长度的关系

Fig. 5　Number of pipeline failures predicted by the model under different pipe lengths

2.2　分类模型结果分析

2.2.1　分类模型的结果分析

分类问题中利用混淆矩阵和 AUC 值表示模型预测性能，表 3 为 RF 分类模型预测结果的评价指标。模型的 AUC 值均大于 0.73，表现出了良好的分类效果。All Pipe 模型整体准确度为 0.825，召回率(Recall)为 0.537。4 种管线模型中，CI Pipe 的 Recall 值最大，为 0.762，对应的 AUC 也最大，为 0.853；而其余 3 种管线的 Recall 相对来说较小，不同管材的预测效果不一。4 种管线模型中，SP Pipe 的 FPR 值最小为 0.118，其在预测未失效管线的效果上更优。由于测试数据的不平衡性，数据集中约有 97% 的管线未发生失效事件，各个预测模型中的精确度(Precision)都比较低，4 种类型管线的模型精确度范围在 0.07 ~ 0.15 不等。

图 6 展示了 4 种失效管线和未失效管线的失效概率预测情况。可以看出，模型预测失效管线的失效概率集中在 0.5 之上，未失效的管线失效概率集中在 0.3 之下。CI Pipe 失效管线和未失效管线的失效概率均值分别为 0.63 和 0.27，DI Pipe 失效管线和未失效管线的失效概率均值分别为 0.57 和 0.28，均表现出良好的分类结果。而 GS Pipe 和 SP Pipe 的预测偏差较大，失效管线的失效概率均值分别为 0.51 和 0.46，主要原因是这两种管材均为小口径管线，其失效模式与大口径管线不一致。此外，GS Pipe 和 SP Pipe 的失效记录偏多，且很多记录中管龄数据较小，这使得模型在预测时可能低估失效概率。

表 3　分类评价指标(年份划分数据集)

Tab. 3　Classification evaluation metrics (dataset divided by year)

模型	Accuracy	Recall	FPR	Precision	AUC
All Pipe	0.825	0.537	0.165	0.097	0.773
CI Pipe	0.793	0.762	0.206	0.070	0.853
DI Pipe	0.782	0.603	0.212	0.084	0.776
GS Pipe	0.833	0.527	0.158	0.089	0.778
SP Pipe	0.860	0.419	0.118	0.153	0.730

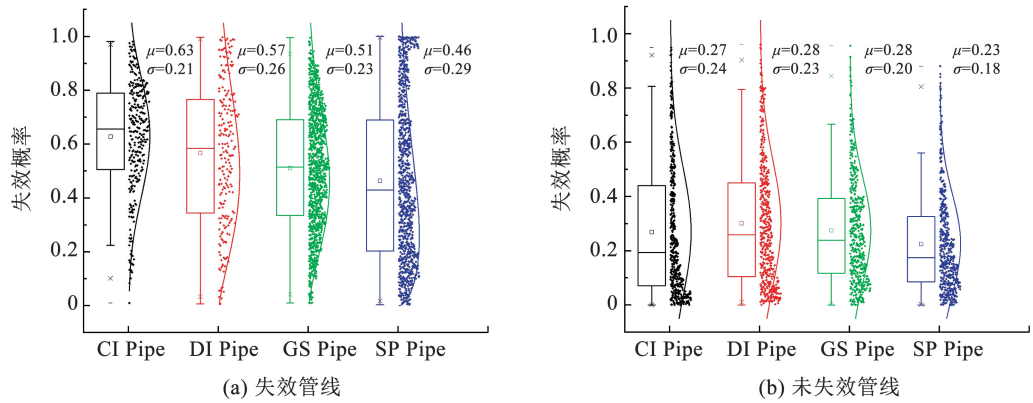


图 6　测试集中失效管线与未失效管线失效概率预测分布

Fig. 6　Predicted failure probability distribution for failed and non-failed pipelines in the test data

2.2.2 建模数据集划分方式的影响

采用随机划分建模数据集的方式进行建模,模型评估结果如表 4 所示。对比表 4 和表 3 可以看出,按照随机划分的方式,失效模型预测结果的召回率有明显的提升,模型的 AUC 值大于 0.85 以上。相比按照年份划分建模数据集的方式,CI Pipe 和 GS Pipe 模型的 AUC 值分别提升了 4.6% 和 12.7%,All Pipe 的 AUC 值提升了 14.2%。现有研究大多数采用随机划分的方式[11-12],分类效果较好。然而在实际应用时,通常采用 2.2.1 节中按照年份划分数据集的方式进行建模,来预测在役管线未来几年的失效概率情况。

2.2.3 建模数据集构成比例的影响

按照年份划分数据集的方式,分析建模数据集中未失效管线和失效管线数据比例对模型评价结果的影响。由图 7 可知,建模管线数据的不平衡比例对各类指标和不同管材的影响规律均存在差异。未失效的管线数据占比增大,管线数据的召回率(Recall)显著下降,表明实际失效且被预测为失效的比例下降,对于大部分实际失效的管线,模型预测失效概率低于分类阈值 0.5。预测模型的精确度(Precision)和准确度(Accuracy)均有所上升,在精确度和召回率的交点之后,准确度上升幅度缓慢。

表 4 分类评价指标(随机划分数据集)

Tab. 4 Classification evaluation metrics (randomly divided dataset)

模型	Accuracy	Recall	FPR	Precision	AUC
All Pipe	0.819	0.780	0.180	0.125	0.883
CI Pipe	0.787	0.823	0.213	0.073	0.892
DI Pipe	0.780	0.766	0.219	0.100	0.843
GS Pipe	0.827	0.762	0.171	0.115	0.877
SP Pipe	0.859	0.788	0.137	0.226	0.898

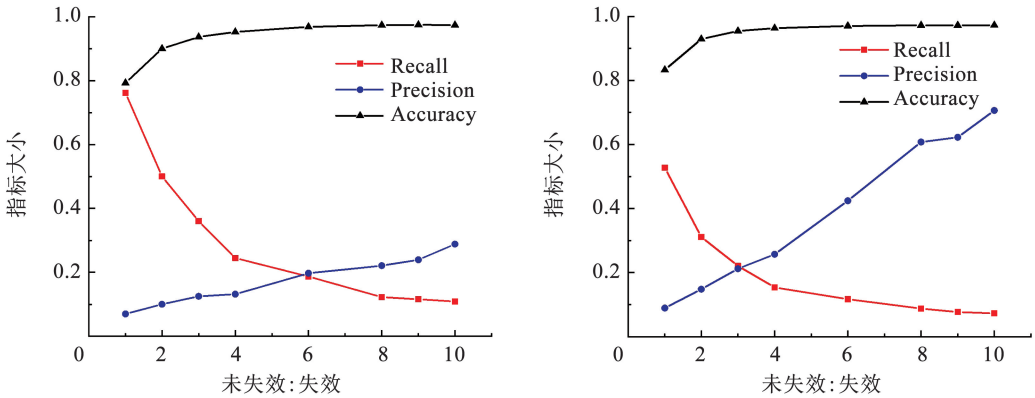


图 7 建模管线数据不平衡比例下的模型结果

Fig. 7 Model results under imbalanced pipeline data ratios

当建模数据集为平衡数据集(未失效:失效 = 1:1)对应的召回率(1:1)较大时,精确度指标和召回率指标的交点所对应的不平衡比例也偏大。如 CI Pipe 召回率(1:1)为 0.76,所对应的交点数据比例为 6:1;GS Pipe 召回率(1:1)为 0.53,所对应的交点比例为 3:1。在进行管线失效模型构建时,可以根据实际数据的情况,通过调整模型阈值、采用欠采样或过采样来平衡数据集,以提高模型的准确性和可靠性。

2.3 回归模型结果分析

2.3.1 回归模型的结果分析

回归模型选用一致性指数(C-index)和回归评估指标表示模型预测性能,评估结果如表 5 所示。模型的 C-index 介于 0.61 ~ 0.82,其中,All Pipe 的 C-index 最大。其涵盖了所有管线的数据集,在训练时能够学习更多的样本,预测一致性大于单一管线模型。同样,All Pipe 的 R^2 为 0.71,也大于 4 种单一管线模型。所有模型的平均偏差(E_{MB})均小于 0,说明模型预测管线失效时间偏小,即模型预测的失效时间小于实际管线失效时间。出现这种情况的原因是建模输入的数据仅包含了失效管线的数据,没有考虑在记录期内未发生失效事件的管线数据(即删失数据)。

表 5 回归模型评估指标

Tab. 5 Regression evaluation metrics

模型	R^2	E_{RMS}	E_{MA}	E_{MB}	C-index
All Pipe	0.71	8.16	5.87	-3.99	0.82
CI Pipe	0.16	13.22	9.38	-6.49	0.64
DI Pipe	0.12	6.57	5.55	-3.45	0.61
GS Pipe	0.43	8.87	6.31	-3.73	0.74
SP Pipe	0.11	5.04	4.31	-3.37	0.62

图 8 给出了随着管龄变化的预测管线失效时间平均误差分布(预测时间 - 实际时间),可以看出,

误差较大的管线数据其实际失效时间往往也较大。这是因为模型在训练数据集中无法考虑删失数据，且在建模数据集中具有较大失效时间的管线样本数据较少，不能学习到大龄管线的失效时间规律。

图 9 绘制了 CI Pipe 和 GS Pipe 的管龄分布情况，GS Pipe 和 SP Pipe 在 2000 年之后不再敷设，训练数据管线管龄大于 9 a，测试数据管线管龄均大于 19 a。训练数据管线平均管龄为 31.2 a，但测试数据管线平均管龄为 36.3 a。建模管线管龄数据的分布对模型预测结果的影响不容忽视。针对大龄管线数据较少的问题，可以对数据进行过采样生成更多大龄管线的样本，有效减少模型对小管龄管线的偏向，以此改善预测结果。

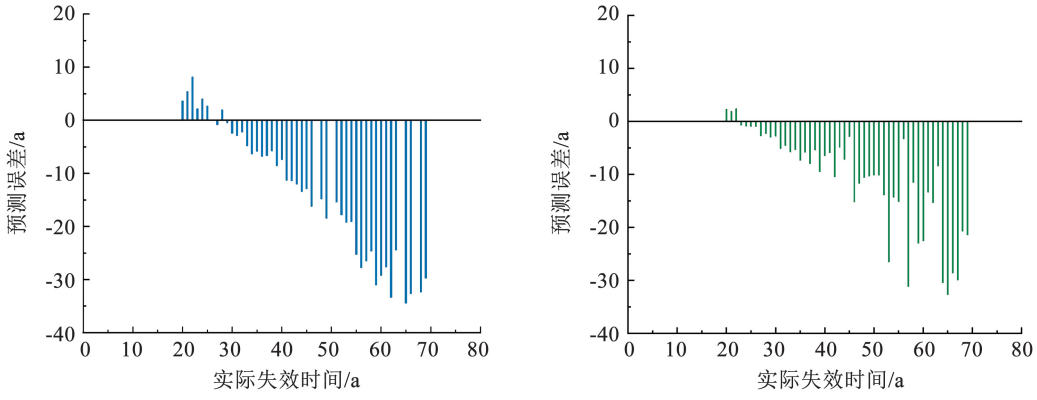


图 8　预测管线失效时间平均误差分布

Fig. 8　Distribution of average error in pipeline failure time prediction

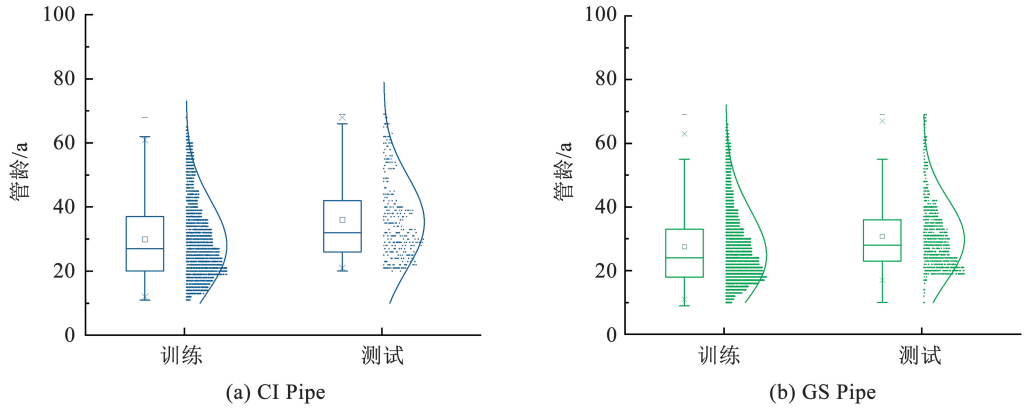


图 9　建模数据集管龄分布

Fig. 9　Distribution of pipe age in modeling dataset

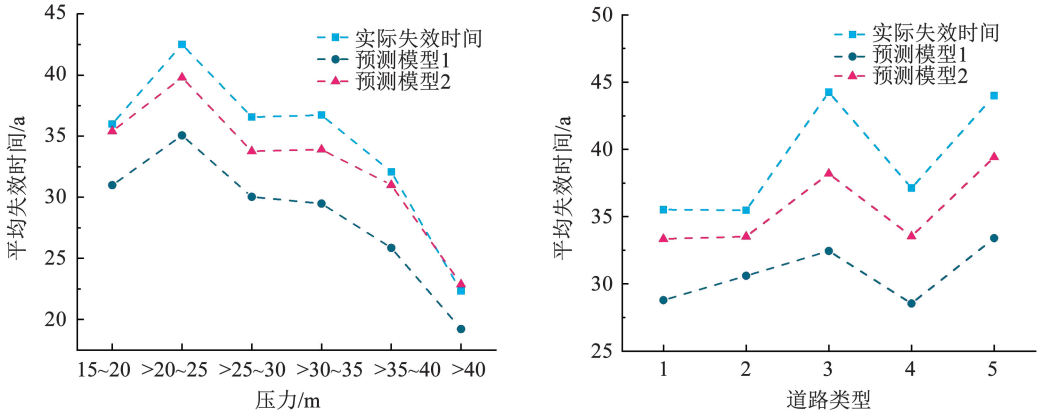
2.3.2　建模数据集划分方式的影响

在回归模型中，与分类模型类似，分析建模数据集划分方式对预测结果的影响。通过对 2009—2019 年全部的失效数据进行混合抽样，得到训练数据以及预测数据。按照随机划分建模数据得到的模型预测误差变小，CI Pipe 和 GS Pipe 的模型平均偏差（E_{MB}）为 −0.82 a 和 −0.94 a，而按照年份划分的模型平均偏差为 −6.49 a 和 −3.73 a。与按照年份划分数据集的方式相比，随机划分数据集使得训练数据集与测试数据集的差异变小，预测误差会降低。今后在进行回归模型的构建时，应该注意这一点。此外，回归模型的建模数据集均为失效管线数据集，有很多未发生失效事件的管线数据（即删失数据）被排除在模型之外，这种情况也会对模型的结果产生影响，2.3.3 中讨论了建模数据集中考虑删失数据的模型预测情况。

2.3.3　建模数据集构成比例的影响

选取普通铸铁（CI）管作为分析对象，统计不同参数（压力、道路类型）维度对应的管线平均失效时间预测结果，如图 10 所示。预测模型 1 的建模数据仅使用失效管线数据作为训练模型的数据，而预测模型 2 的建模数据集包含从未失效管线中抽样的数据（即删失数据），对该类删失数据，将记录期结束时的管线管龄视为管线失效时间。引入删失数据后，模型的预测精度有所提升。在普通铸铁管的分析中，预测模型 2 在不同参数（如压力、道路类型）

维度下的平均失效时间预测精度更高,不同压力区间下,预测时间精度提升 3.7 ~ 5.1 a;不同道路类型下,预测时间精度提升 2.9 ~ 6.0 a。在其余 3 种管线中也出现了类似的规律。删失数据的引入丰富了模型的训练样本,使得模型能够更好地学习管线失效时间的分布规律,有效缓解数据不平衡问题,提高模型对大龄管线失效时间的预测能力。后续可采用生存分析等专门用于处理删失数据的方法进行建模,以提升模型的预测性能。另外,图 10(a)对应不同压力区间管线的平均失效时间预测结果,管线运行压力增大容易造成管线发生失效事件。在评价模型的预测精度时,也要注意辨识实际建模数据集特征造成的精度差异。



(a) 不同压力区间 (b) 不同道路类型

图 10 不同参数统计得到的管线平均失效时间

Fig. 10 Average pipeline failure time based on different parameters

3 结 论

1)选择建模方法和方式均会对供水管线失效模型的预测结果产生重要影响。在 BPNN、SVM、RF 3 种机器学习算法中,RF 构建的失效分类模型和回归模型均表现出最好的性能。分类和回归两种建模方式得到的模型输入变量重要性有所区别,管龄、管长在分类模型中重要性显著,而管材在回归模型中更加重要。通过一致性指数评估两类模型的预测精度,分类模型比回归模型高 5.4% ~ 32.8%。

2)随机划分建模数据集的方式能够提高失效模型的预测精度。在分类模型中,相比按照年份划分建模数据集的方式,CI Pipe 和 GS Pipe 模型的 AUC 值分别提升了 4.6% 和 12.7%。在回归模型中,按照随机划分建模数据得到的 CI Pipe 和 GS Pipe 模型平均偏差(E_{MB})降低了 5.67 a 和 2.79 a。

3)建模数据集构成比例对分类模型与回归模型的预测结果影响有所区别。分类模型中,随着未失效管线与失效管线的数据比例增大,召回率显著下降,精确度和准确度均有所上升;回归模型中,将未发生失效记录的管线数据纳入建模数据集中,管线平均失效时间预测误差减小,证明模型预测精度有所提高。

4)除了建模方法影响模型预测精度,数据划分方式以及数据集构成比例均会影响失效模型的精度,因此,在构建供水管线失效模型时,通过优化数据预处理流程,可以显著提升模型的预测能力和实际应用价值,为供水管网的失效预测和风险管理提供有力的技术支持。

参考文献

[1] CLAIR A, SINHA S. State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models[J]. Urban Water Journal, 2012, 9(2): 85. DOI: 10.1080/1573062X.2011.644566

[2] DAWOOD T, ELWAKIL E, NOVOA H M, et al. Water pipe failure prediction and risk models: state-of-the-art review[J]. Canadian Journal of Civil Engineering, 2020, 47(10): 1117. DOI: 10.1139/cjce-2019-0481

[3] KLEINER Y, RAJANI B. Comprehensive review of structural deterioration of water mains: statistical models[J]. Urban Water, 2001, 3(3): 131. DOI: 10.1016/S1462-0758(01)00033-4

[4] WILSON D, FILION Y, MOORE I. State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains[J]. Urban Water Journal, 2017, 14(2): 173. DOI: 10.1080/1573062X.2015.1080848

[5] NISHIYAMA M, FILION Y. Forecasting breaks in cast iron water mains in the city of Kingston with an artificial neural network model[J]. Canadian Journal of Civil Engineering, 2014, 41(10): 918. DOI: 10.1139/cjce-2014-0114

[6] SCHEIDEGGER A, LEITAO J P, SCHOLTEN L. Statistical failure models for water distribution pipes: a review from a unified

perspective[J]. Water Research, 2015, 83: 237. DOI: 10.1016/j.watres.2015.06.027

[7]TAIWO R, SEGHIER M, ZAYED T. Toward sustainable water infrastructure: the state-of-the-art for modeling the failure probability of water pipes[J]. Water Resources Research, 2023, 59(4): 12. DOI: 10.1029/2022WR033256

[8]赵洪宾，陈兵，伍悦滨. 给水管网漏失预测模型的研究[J]. 给水排水，2001，27(10)：94
ZHAO Hongbin, CHEN Bing, WU Yuebin. Study on leakage model of water distribution network[J]. Water & Wastewater Engineering, 2001, 27(10): 94. DOI: 10.13789/j.cnki.wwe1964.2001.10.033

[9]AL-ALI A M, LAURENT J, DULOT J P. Developing deterioration prediction model for the potable water pipes renewal plan-case of Jubail Industrial City, KSA[J]. Desalination and Water Treatment, 2020, 176: 324. DOI: 10.5004/dwt.2020.25539

[10]ABDELMAGEED S, TARIQ S, BOADU V, et al. Criteria-based critical review of artificial intelligence applications in water-leak management[J]. Environmental Reviews, 2022, 30(2): 280. DOI: 10.1139/er-2021-0046

[11]FAN Xudong, WANG Xiaowei, ZHANG Xijin, et al. Machine learning based water pipe failure prediction: the effects of engineering, geology, climate and socio-economic factors[J]. Reliability Engineering & System Safety, 2022, 219: 108185. DOI: 10.1016/j.ress.2021.108185

[12]CHEN T Y, VLADEANU G, YAZDEKHASTI S, et al. Performance evaluation of pipe break machine learning models using datasets from multiple utilities[J]. Journal of Infrastructure Systems, 2022, 28(2): 5022002. DOI: 10.1061/(ASCE)IS.1943-555X.0000683

[13]侯本伟，肖恒圣，吴珊. 考虑天气因素的给水管道漏损预测模型[J]. 哈尔滨工业大学学报，2022，54(2)：8
HOU Benwei, XIAO Hengsheng, WU Shan. Failure prediction model of water distribution pipelines considering weather factors[J]. Journal of Harbin Institute of Technology, 2022, 54(2): 8. DOI: 10.11918/202012047

[14]LATIFI M, ZALI R B, JAVADI A A, et al. Efficacy of tree-based models for pipe failure prediction and condition assessment: a comprehensive review[J]. Journal of Water Resources Planning and Management, 2024, 150(7): 33. DOI: 10.1061/JWRMD5.WRENG-6334

[15]ZALI R B, LATIFI M, JAVADI A A, et al. Semisupervised clustering approach for pipe failure prediction with imbalanced data set[J]. Journal of Water Resources Planning and Management, 2024, 150(2): 86. DOI: 10.1061/JWRMD5.WRENG-6263

[16]李航. 机器学习[M]. 北京：清华大学出版社，2022
LI Hang. Machine learning[M]. Beijing: Tsinghua University Press, 2022

[17]阎立华，马婷婷，田昊平. BP 神经网络对管网安全使用期的预测[J]. 沈阳建筑大学学报(自然科学版)，2008，24(5)：845
YAN Lihua, MA Tingting, TIAN Haoping. Research on prediction of safe using dates of water supplying pipes based on BP neural network[J]. Journal of Shenyang Jianzhu University (Natural Science), 2008, 24(5): 845

[18]DONG Xuefan, LIAN Ying, LIU Yijun. Small and multi peak nonlinear time series forecasting using a hybrid back propagation neural network[J]. Information Sciences, 2018, 424: 39. DOI: 10.1016/j.ins.2017.09.067

[19]MALDONADO S, PÉREZ J, WEBER R, et al. Feature selection for support vector machines via mixed integer linear programming[J]. Information Sciences, 2014, 279: 163. DOI: 10.1016/j.ins.2014.03.110

[20]BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5. DOI: 10.1023/A:1010933404324

[21]BRENTNALL A R, CUZICK J. Use of the concordance index for predictors of censored survival data[J]. Statistical Methods in Medical Research, 2018, 27(8): 2359. DOI: 10.1177/0962280216680245

（编辑　刘　彤）