基于梯度提升决策树级联分类方法的城市轨道交通列车突发事件延误时间预测*

欧冬秀[1,2**]　张馨尹[1]　赵　源[2,3]　张　雷[4]　高博文[1]　吴宇森[1]

（1. 同济大学交通运输工程学院，201804，上海；2. 上海市轨道交通结构耐久与系统安全重点实验室，201804，上海；3. 上海轨道交通运营管理中心，200070，上海；4. 上海自主智能无人系统科学中心，201210，上海//第一作者，教授）

摘　要　为了精确预测城市轨道交通设备故障等突发事件致使的列车延误时间，提升应急处置效率和乘客引导服务水平，对地铁突发事件互联网发布数据和现场事故数据进行了关联融合，对面向不平衡的地铁事故数据随机欠采样，提出了一种基于 GBDT（梯度提升决策树）的级联分类预测方法，对地铁突发事件的延误时间进行预测。结果表明，GBDT 级联分类方法在延误时间容许偏差为 0～5 min 时的预测延误时间准确率，比现场发布的预测延误时间准确率高 20%～25%，比 GBDT 多分类预测方法准确率高 5%。

关键词　城市轨道交通；列车；突发事件；延误时间预测；级联分类方法；梯度提升决策树

中图分类号　U231.92

DOI：10.16037/j.1007－869x.2022.10.013

Urban Rail Transit Train Accident Delay Time Prediction Based on GBDT Cascade Classification Method

OU Dongxiu，ZHANG Xinyin，ZHAO Yuan，ZHANG Lei，GAO Bowen，WU Yusen

Abstract　To accurately predict train the delay time caused by accidents such as urban rail transit equipment failure，and improve the emergency response disposal efficiency and the passenger guidance service level，the association fusion of internet data and field data of metro accidents is carried out. According to the unbalanced metro accident random undersampling data，a cascade classification prediction method based on GBDT（gradient boosting decision tree）is proposed to predict the delay time of metro accidents. The results indicate that when the delay time allowable deviation of GBDT cascade classification method is 0～5 min，the predicted delay time accuracy of the method is 20%～25% higher than that released on site，and 5% higher than that of GBDT multi-classification prediction method.

Key words　urban rail transit；train；emergency accident；delay time prediction；cascade classification method；GBDT（gradient boosting decision tree）

First-author's address　School of Transportation Engineering，Tongji University，201804，Shanghai，China

城市居民日常出行对城市轨道交通的依赖毋庸置疑，而突发事件时有发生，经常致使列车延误。据统计，某城市在 2017 年至 2019 年地铁全线突发事故导致列车延误 5 min 以上高达 360 次，某条单线延误达 67 次，其中最长延误时间达 275 min，严重影响公众出行体验。为了减缓延误影响，地铁运营管理部门在信息平台实时播报地铁各条线路的运营情况和突发事件信息[1]，但现阶段播报的预报延误时间与实际延误时间相比具有较大偏差。如某日某条线路预报延误时间为 10 min 以上，而实际延误时间长达 128 min。精准的列车延误时间估计不仅能为乘客提供直观可信的地铁实时信息便于其调整出行路线，而且还能为运营管理部门调整运维方案、部署清客和救援工作提供基础数据支撑。因此，突发事故下地铁列车延误时间的预估研究对于提升地铁信息化服务水平具有重要意义。

一方面，学者们正研究运用数据驱动的人工智能方式进行故障诊断，实现智能运维和状态维修，从而降低故障发生率[2]；另一方面，学者们也在积极探索故障发生后降低列车延误影响的方法和技术。列车延误与设备维护、人员操作、外部环境、发生时段、客观综合等因素有关[3]。文献[4]从单因素、时空维度等方面研究了事故类型、运营时间、区

* 国家重点研发计划项目（2018YFB1201403）

** 通信作者

段等事件特征之间的关联性,但未对事件特征与延误时间的关联性做深入分析。文献[5]结合灰色模型和马尔科夫模型预测了列车晚点时间。文献[6-8]基于晚点列车运行数据进行了聚类分析,运用随机森林模型、支持向量机预测各类晚点列车的晚点时间。文献[9]构建了航班延误特征,基于 GBDT(梯度提升决策树)对航班延误进行分类预测。大量研究表明,分类预测方法能够对列车晚点时间进行可靠预测。

地铁事故数据具有低延误数量多、高延误数量少的特点,这种类别不平衡特性会影响机器学习算法的性能。对于类别不平衡数据集,文献[10]基于欠采样法提出基于自适应加权 Bagging-GBDT 分类算法,解决了数据集正负样本数目不均衡导致的分类算法预测准确率低的问题。目前,对地铁列车延误的预测方法较少地考虑了事故数据集的类别不平衡性,因此,基于不平衡事故数据对列车延误时间进行精细化预测的研究仍有待开展。

本文对地铁事故互联网发布数据和现场故障记录数据进行融合,并挖掘日期、时段、线路类别、致因等事故特征,及其对列车延误影响的关联关系。基于事故数据的不均衡特征,采用随机欠采样方法建立了基于 GBDT 的级联分类模型,并对突发事件引起的列车延误时间进行分级预测。

1 城市轨道交通列车运营延误影响及事故特征分析

为缓解突发事件影响,地铁运营管理部门在站内、互联网等多个平台发布事故信息,内容大致如下:“上海轨道交通 2 号线因信号设备故障,世纪大道站至南京东路站区间列车限速运行,预计晚点 10 min 以上,请乘客们及时调整出行……”实际运营中,地铁突发事件的播报延误时间远长于 10 min,且通常无法准时恢复运营。据上海轨道交通 2 号线 2017—2019 年的统计数据,地铁预报延误时间均为 20 min 以上;平均实际延误时间为 29 min,最大实际延误时间高达 275 min。由此可见,突发事件的预报延误时间和实际延误时间存在较大偏差,且两者偏差越大,对应急处置方案的制定和乘客引导的影响也越大。

如表 1 可见,轨道交通突发事件数据包含互联网发布数据和现场故障记录数据。互联网发布的非结构化文本信息含有丰富的信息:“2018-01-15T16:29:00,上海轨道交通 1 号线因信号设备故障,×站—×站区间列车限速运行,发车班次间隔延长,预计晚点 15 min 以上,请乘客们及时调整出行路径,以免耽误行程”“2018-01-15T16:49:00,1 号线信号设备故障已排除,全线运营正在逐步恢复中,给您出行带来不便,敬请谅解!”经文本抽取、挖掘等方法处理后获取如下特征元素:日期 d、时间 h、线路编号 l、预报延误时间 D_p、实际延误时间 D_r 等。现场故障记录数据包含如下特征元素:d、h、l、致因 c、延误时间 D_1、影响列车数 q 等。融合互联网数据与现场数据两个数据集,得到事故特征数据集:$\{d,h,l,c,D_p,D_r\}$。

表 1 上海轨道交通 1 号线事故特征元素取值示例

Tab. 1 Example of accident feature element values of Shanghai rail transit Line 1

特征元素	d	h	l	c	D_1/min	q
取值或说明	2018-01-15	16:10:00	1	通号	>15	4

2 GBDT 级联分类预测模型的建立

2.1 GBDT 级联分类预测模型

本文设计了一个面向不平衡数据的 GBDT 级联分类预测模型。模型构建过程中,组合多个学习器 $f(x)$ 构成层级分类器 $g(x)$,串联多个 $g(x)$ 的正例输出结果构成级联分类器 $G(x)$。

级联分类器是在每层分类器设置不同阈值划分样本并进行分类训练。若通过前一层分类器的测试样本满足下一层级阈值标准,则可进入下一层分类器测试,依次类推。

分类模型的输入为原始数据集 $s=\{x,y\}$,其中,x 为事故特征集,$x=\{d,h,l,c\}$,y 为 D_r 转换的二分类标签。分类模型的每层输出为预测延误时间 $\hat{y}$,z 为综合层级输出 $\hat{y}$ 得到的预测延误时间。

2.1.1 梯度划分

按照事件 D_r 划分“阶梯”级别,设置层级时间标准 $t_i, i\in[1,m]$,m 为层级数。判断输入的 D_r 与 t_i 的关系,将事故数据按层级时间标准进行划分。

2.1.2 层级分类器 $g(x)$

2.1.2.1 面向不平衡数据的随机欠采样

将事故数据按层级时间标准划分为负样本和正样本。事故数据表现出标签不平衡的问题。采用随机欠采样方法实现正、负样本平衡,具体方法为:对于每一层级 $i(i\in[1,m])$,对数据进行随机欠

采样,进而得到 k 个相互独立的正负样本平衡的数据集,每个数据集记为 $s_{i,j}$(j 为数据集编号),训练得到 k 个 GBDT 学习器 $f_{i,j}(x)$($i\in[1,m]$,$j\in[1,k]$),组合 k 个 GBDT 学习器的结果得到最终分类结果。

2.1.2.2　学习器 $f(x)$

采用 GBDT 作为学习器 $f(x)$,GBDT 是一种基于 CART(分类与回归决策树)的集成学习模型。该模型串行训练 1 组弱学习器(CART 决策树),将预测延误时间逐步拟合逼近真实值。对于二分类模型,对样本进行正、负分类,采用 sigmoid 函数计算得到类别[11]。

输入样本集为 $s_{i,j}=(x,y_i)$,$i\in[1,m]$,$j\in[1,k]$。其中,x 为输入特征,$x=\{d,h,l,c\}$;y_i 为对应样本 x 的实际延误标签。对第 i 层级含有 n 个样本的数据集训练学习器。GBDT 模型 $f(x)$ 的构建步骤如下:

步骤 1　初始化学习器 $f(x)$,并采用对数损失函数,调整决策树参数使得损失函数 $L(y_i,f(x))$ 达到最小。

$$\begin{cases}L(y_i,f(x))=-y_i\ln(1+e^{-\hat{y}_i})-(1-y_i)\cdot\ln(1+e^{\hat{y}_i})\\ f(x)=T(x;\theta_1)\end{cases}\tag{1}$$

式中:

$\hat{y}_i$——模型对样本 x 的预测延误标签;

θ_1——决策树参数。

步骤 2　利用损失函数的负梯度 r_i 拟合残差,调整决策树的参数目标使损失函数达到最小,并更新模型 $f(x)$。

$$\begin{cases}r_i=-\dfrac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\\ f(x)=T(x;\theta_1)+T(x;\theta_2)\end{cases}\tag{2}$$

式中:

θ_2——决策树参数。

步骤 3　重复步骤 2 完成 $L-1$ 次迭代,并通过 sigmoid 函数完成概率计算,实现类别判别。

$$f(x)=\sigma\left[\sum_{i=1}^{L}T(x;\theta_i)\right]\tag{3}$$

式中:

θ_i——决策树参数。

2.1.2.3　集成学习分类模型

以均等投票机制组合同类别且彼此之间无强关联的学习器,如图 1 所示。采用学习器进行二分类预测,则投票机制定义为:超过半数学习器及层级分类器输出正例,即判定延误时间大于该层级时间标准 t_i。则层级分类器 $g_i(x)$ 可表示为:

$$g_i(x)=\begin{cases}1, & \sum_{j=1}^{k}f_i(x)>\dfrac{k}{2},\ i\in[1,m]\\ 0, & \text{其他}\end{cases}\tag{4}$$

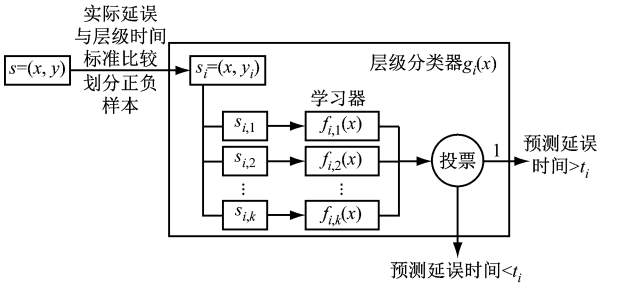


图 1　层级分类器 $g(x)$ 的结构

Fig. 1　Structure of hierarchical classifier $g(x)$

2.1.3　级联分类器 $G(x)$

级联通过正例输出串联所有层级分类器,以实现多分类的效果。GBDT 多分类即对含有多个分类标签的样本进行分类。相较于 GBDT 多分类,GBDT 级联分类预测模型在各层级分别进行类别数据的平衡处理,各层级之间彼此独立,可以同时训练。如图 2 所示,当预测延误时间大于该层级时间标准时,层级分类器预测输出正例,进入下一层级分类器进行预测;若层级分类器预测输出负例,则终止计算。通过正例串联各层级分类器预测结果,得到预测延误时间区间。

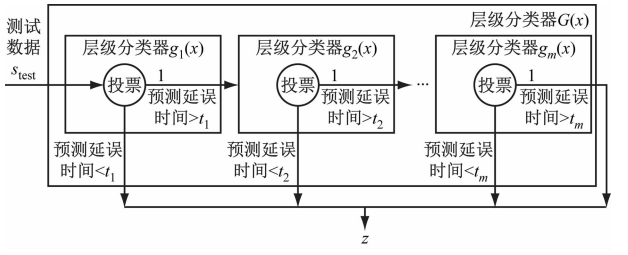


图 2　级联分类器 $G(x)$ 的结构

Fig. 2　Structure of cascade classifier $G(x)$

2.2　模型训练与评测

2.2.1　模型训练

确定合理的训练集和测试集样本量,在训练集中通过有标签的样本来寻找 1 组使得损失函数取值最小的模型参数。模型参数主要包括 GBDT 框架参数 φ(宏观参数,包括基学习器的个数和权重缩减系数等)和 CART 决策树参数 θ(微观参数,包括决策树的深度、节点数及使用特征数量等参数)。采

用网格搜索寻优方法对参数空间进行求解，并以对数似然损失函数作为评判标准，求得最佳模型参数。

2.2.2 预测结果评测

乘客对延误时间具有一定容许度。定义 n_t 为预测延误时间 z_t 与 D_r 之差同容许偏差 ξ 的大小关系。z_t 与 D_r 之差在 ξ 内为预测准确。则准确率 R_{acc} 的计算公式为：

$$\begin{cases} n_t = \begin{cases} 1, & D_{r,t} - z_t \leqslant \xi \\ 0, & D_{r,t} - z_t > \xi \end{cases} \\ R_{acc} = \left(\sum_{t=1}^{N} n_t \right) / N \end{cases} \tag{5}$$

式中：

t——样本编号，$t \in [1, N]$，N 为样本总数；

$D_{r,t}$——第 t 个样本的实际延误时间。

3 实例验证

3.1 数据分析与处理

经数据清洗与融合，获得某城市 2017 年 1 月 1 日至 2019 年 12 月 31 日地铁事故互联网数据与现场数据 265 条。为有效利用事故特征，本文将事故特征（事故日期、事故时段、事故线路、事故致因）进行细致划分：事故日期划分为工作日故障和非工作日故障；事故时间划分为高峰期故障和非高峰期故障；事故致因主要划分为车辆故障、通号故障、供电故障和客观故障（包含运营组织、安全管理）。

事故线路采用 *K*-means 算法进行聚类。将事故线路分为事故高发线路、事故中发线路、事故低发线路等 3 类，见图 3。

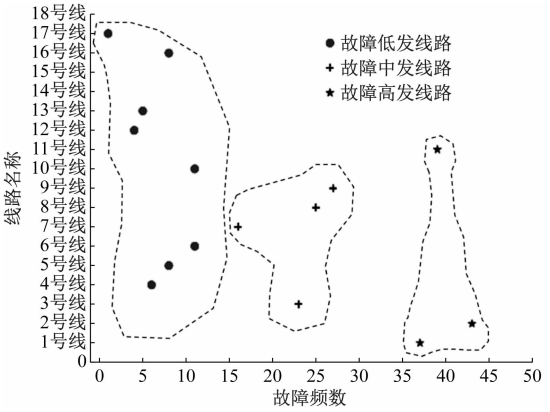
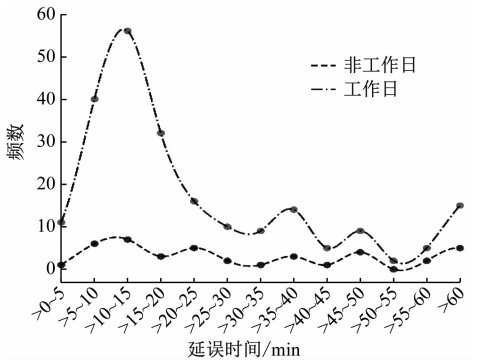


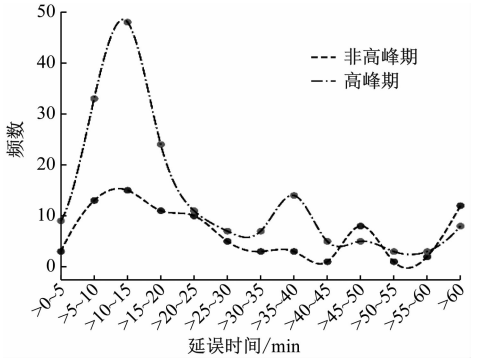
图 3 地铁事故线路聚类图

Fig. 3 Diagram of metro accident line cluster

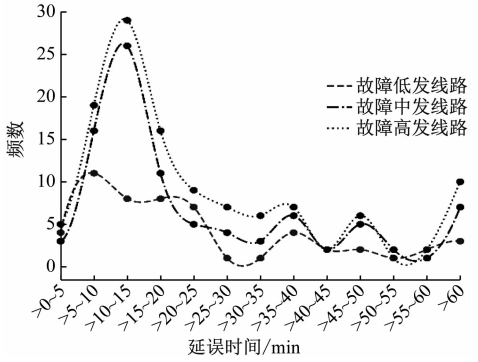
如图 4 所示，通过分析各事故特征下不同延误时间的事故频数，得到延误时间与事故特征之间的



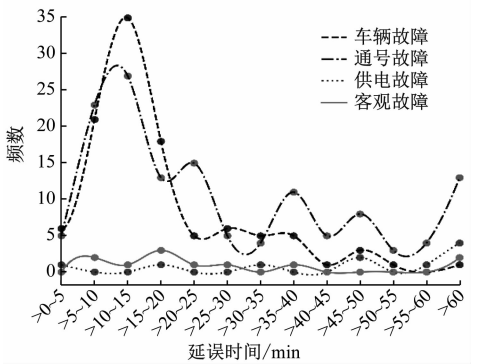
a）工作日与非工作日的事故频数



b）高峰期与非高峰期的事故频数



c）低发、中发、高发的事故频数



d）各类事故致因下的事故频数

图 4 地铁不同事故特征下的事故频数-延误时间关系

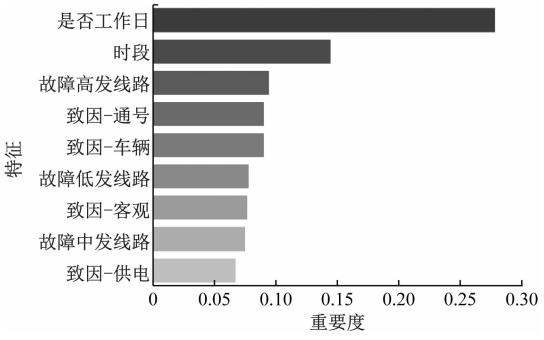
Fig. 4 Relationship of accident frequency-delay time under different metro accident characteristics

相关性:取 5 min 作为延误时间粒度,工作日事故延误时间多为>5 ~20 min,非工作日事故延误时间多为>5 ~15 min;高峰期事故延误时间多为 >5 ~20 min,非高峰期事故延误时间多为 >5 ~15 min;故障高发线路事故延误时间多为 >5 ~15 min,故障中发线路事故延误时间多为 >5 ~20 min,故障低发线路事故延误时间多为 >5 ~10 min;车辆故障延误时间多为 >5 ~20 min,通号故障延误时间多为 >5 ~15 min,供电故障和客观故障延误时间多为 >5 ~20 min。

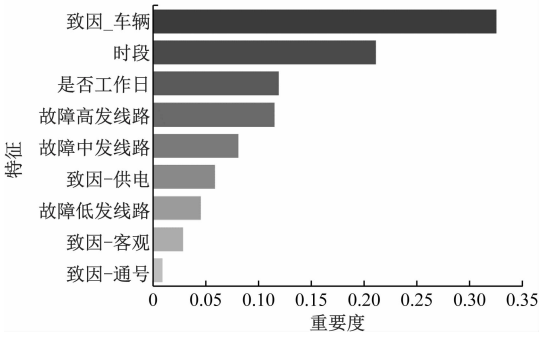
3.2　结果分析与模型评价

为平衡模型的复杂程度和有效性,本文选取层级数量 m 为 3,层级基分类器个数 k 为 3,地铁运营部门播报延误时间分别为 10 min、15 min、20 min 及以上(延误时间为 5 min 以内未公示),故设置每个层级时间标准分别为 $t_1 = 10$ min、$t_2 = 15$ min、$t_a = 20$ min。将事故日期、事故时段、事故线路和事故致因等作为自变量,将实际延误时间作为模型的因变量,即根据实际延误时间是否大于 t_1、t_2、t_3,对其进行二分类转换为 y_1、y_2、y_3。将数据集按 8∶2 划分为训练集和测试集,对模型进行训练和测试。

模型训练完成后,得到层级分类器的特征重要度,如图 5 所示。由图 5 可见,对事故延误是否大于 10 min 的预测与事故日期和事故时段有较大关联,对事故延误是否大于 15 min 的预测与事故致因和事故时段的关联程度较高。



a) 事故延误时间为 10 min



b) 事故延误时间为 15 min

图 5　不同事故延误时间下层级分类器的特征重要度

Fig. 5　Feature importance of hierarchical classifier under different accident delay times

为确定模型性能,将本文提出的 GBDT 级联分类方法预测延误时间与互联网预报延误时间、GBDT 多分类方法预测延误时间的准确率进行对比。GBDT 多分类预测方法是将 CART 决策树作为弱分类器,采取一对多策略,对每个类别训练一定数量分类器,从而进行多分类预测。

乘客对延误时间预测的容许偏差 $\xi = \{0,5,10,15\}$。对比不同 ξ 时互联网预报、GBDT 多分类方法、GBDT 级联分类方法下延误时间的准确率,如图 6 所示。由图 6 可见,延误时间在 0 ~5 min 容许偏差范围内,GBDT 级联分类方法预测延误时间的准确度较互联网预报高 20% ~25%,较 GBDT 多分类方法的准确度高 5%。延误时间在 10 min 和 15 min 等较大容许偏差范围内,GBDT 级联分类方法预测延误时间的准确率达 95%,且较互联网预报准确率高 5% ~20%,较 GBDT 多分类方法准确度高 5% ~10%。但对于乘客在城市轨道交通实际运营中较高的服务品质需求,若延误时间存在较大偏差,则很难被乘客接受。GBDT 级联分类模型进行了梯度划分,并分层级对不平衡数据进行了随机欠采样,保证了数据类别的平衡性,有效改善了不平衡数据在分类预测问题中准确率低的问题。因此,相比 GBDT 多分类方法,GBDT 级联分类方法预测延误时间的准确率得以提升。

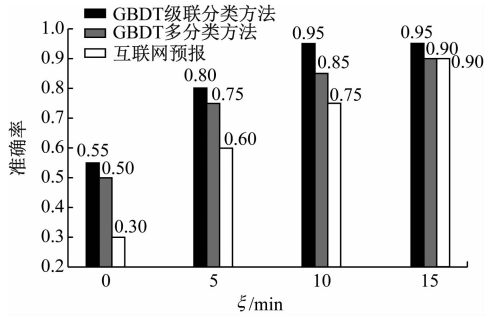


图 6　不同容许偏差下延误时间预测准确率对比

Fig. 6　Comparison of accurate delay time predictions under different allowable deviations

4　结语

本文关联融合了地铁事故的互联网数据和现场数据,并对数据特征进行了分析。基于事故数据

聚类得出高发、中发、低发事故线路类型,以及事故延误时间与事故日期、事故致因和事故时段的关联程度较高。

本文所提出的 GBDT 级联分类模型通过梯度划分层级结合分层随机欠采样保证了事故数据类别的平衡性,改善了数据不平衡造成的分类预测不准确问题,并通过梯度级联层级分类器精细化地预测了突发事件下的轨道交通延误时间。该方法所预测的延误时间在 0 ~5 min 容许偏差范围内比互联网预报的准确率提升了 20% ~25%,比 GBDT 多分类预测方法的准确率提升了 5%,由此可见延误时间预测准确率得到了显著提升。

采用 GBDT 级联分类方法预测延误时间不仅能为乘客提供更为可信的地铁实时信息,还能为地铁运营管理部门调整运维方案、部署清客和救援等工作提供基础数据支撑。后续可进一步引入成熟的实时数据处理软件,实现地铁线路延误时间的在线预测。

参考文献

[1] 宋雨洁. 地铁发布晚点信息存在的问题和优化建议[J]. 都市快轨交通,2020(3):157.
SONG Yujie. Release of information metro delays: problems and suggestions for optimization[J]. Urban Rapid Rail Transit, 2020(3):157.

[2] OU D, XUE R, CUI K. A data-driven fault diagnosis method for railway turnouts[J]. Transportation Research Record Journal of the Transportation Research Board, 2019, 2673(4):448.

[3] 解熙,蒲琪. 城市轨道交通列车延误统计指标及评价指标体系研究[J]. 城市轨道交通研究, 2018(4):75.
XIE Xi, PU Qi. Research on statistics index and evaluation index system of urban rail transit delays[J]. Urban Mass Transit, 2018(4):75.

[4] 钱蕾,周玮腾,韩宝明. 城市轨道交通运营突发事件数据可视化分析[J]. 铁道科学与工程学报,2020(4):1025.
QIAN Lei, ZHOU Weiteng, HAN Baoming. Visual exploration of emergency operation events in urban rail transit[J]. Journal of Railway Science and Engineering, 2020(4):1025.

[5] 孙延浩,张琦,任禹谋,等. 基于改进灰色马尔科夫模型的列车晚点预测[J]. 计算机仿真,2020(6):117.
SUN Yanhao, ZHANG Qi, REN Yumou, et al. Prediction of train delay time under speed limit condition based on improved Grey-Markov Model[J]. Computer Simulation, 2020(6):117.

[6] 胡瑞,徐传玲,冯永泰,等. 广深高速铁路列车分类型晚点预测[J]. 中国安全科学学报, 2019(增刊2):181.
HU Rui, XU Chuanling, FENG Yongtai, et al. Prediction of different types of train delay of Guangzhou-Shenzhen high-speed railway[J]. China Safety Science Journal, 2019(S2):181.

[7] 周晓昭,张琦,许伟. 不同限速下基于随机森林的列车区间运行时分预测研究[J]. 铁道运输与经济,2018(2):18.
ZHOU Xiaozhao, ZHANG Qi, XU Wei. A study on the Random Forest-based predictor of trains' running time in different sections[J]. Railway Transport and Economy, 2018(2):18.

[8] 张朴,孟令云,李宝旭. 基于支持向量机的高速铁路列车晚点演化预测[J]. 电气技术, 2019(增刊1):1.
ZHANG Pu, MENG Lingyun, LI Baoxu. Prediction of high-speed railway train delay evolution based on machine learning[J]. Electrical Engineering, 2019(S1):1.

[9] 刘金元,丁勇,李涛. 基于梯度提升决策树的航班延误分类预测[J]. 数学的实践与认识,2018(4):1.
LIU Jinyuan, DING Yong, LI Tao. Classification of flight delay based-on GBDT[J]. Journal of Mathematics in Practice and Theory, 2018(4):1.

[10] 李新鹏,高欣,何杨,等. 不平衡数据集下基于自适应加权 Bagging-GBDT 算法的磁盘故障预测模型[J]. 微电子学与计算机,2020(3):14.
LI Xinpeng, GAO Xin, HE Yang, et al. Prediction model of disk failure based on adaptive weighted bagging-GBDT algorithm under imbalanced dataset[J]. Microelectronics & Computer, 2020(3):14.

[11] 李昕迪, 陈万忠. 基于 FSWT 和 GBDT 的癫痫脑电信号分类研究[J]. 吉林大学学报(信息科学版),2019(2):186.
LI Xindi, CHEN Wanzhong. Classification of epileptic EEG signals based on Frequency Slice Wavelet Transform and Gradient Boosting Decision Tree[J]. Journal of Jilin University (Information Science Edition), 2019(2):186.

(收稿日期:2021 - 07 - 08)