基于 GBDT 的铁路事故类型预测及成因分析

钟敏慧 [1, 2] 张婉露 [1, 2] 李有儒 [1, 2] 朱振峰 [1, 2] 赵 耀 [1, 2]

摘 要 运用数据挖掘技术进行铁路事故类型预测及成因分析, 对于建立铁路事故预警机制具有重要意义. 为此, 本文提出一种基于梯度提升决策树 (Grandient boosting decision tree, GBDT) 的铁路事故类型预测及成因分析算法. 针对铁路事故记录数据缺失的问题, 提出一种基于属性分布概率的补全算法, 最大程度保持原有数据分布, 从而降低数据缺失对事故类型预测造成的影响. 针对铁路事故记录数据类别失衡的问题, 提出一种集成的 GBDT 模型, 完成对事故类型的鲁棒性预测. 在此基础上, 根据 GBDT 预测模型中特征重要度排序, 实现事故成因分析. 通过在开放数据库上进行实验, 验证了本文模型的有效性.

关键词 事故类型预测, 缺失补全, GBDT, 集成学习, 成因分析

引用格式 钟敏慧, 张婉露, 李有儒, 朱振峰, 赵耀. 基于 GBDT 的铁路事故类型预测及成因分析. 自动化学报, 2022, **48**(2): 470−478

DOI 10.16383/j.aas.c190630

GBDT Based Railway Accident Type Prediction and Cause Analysis

ZHONG Min-Hui[1, 2] ZHANG Wan-Lu[1, 2] LI You-Ru[1, 2] ZHU Zhen-Feng[1, 2] ZHAO Yao[1, 2]

Abstract The application of data mining technology in railway accident type prediction and cause analysis is of great significance to establish railway accident early warning mechanism. This paper proposes a gradient boosting decision tree (GBDT) based algorithm for railway accident type prediction and cause analysis. In order to solve the problem of data missing in railway accident record dataset, we propose a new data complement algorithm based on the attribute distribution probability, which can keep the distribution of original data as much as possible, thus reducing the impact of data missing on predicting railway accident type. To reduce the impact of unbalanced categories of data in railway accident dataset, an ensemble GBDT model is proposed to predict the types of accidents effectively and robustly. On these bases, according to the importance of features in GBDT prediction model, we complete the cause analysis of railway accidents. Experimental results on an open database show that our proposed method can predict the types and causes of railway accidents effectively.

Key words Prediction of railway accident type, missing data completion, GBDT, ensemble learning, cause analysis

Citation Zhong Min-Hui, Zhang Wan-Lu, Li You-Ru, Zhu Zhen-Feng, Zhao Yao. GBDT based railway accident type prediction and cause analysis. *Acta Automatica Sinica*, 2022, **48**(2): 470−478

近年来, 我国铁路事业高速发展, 在推动国民经济发展中发挥着至关重要的作用. 与此同时, 铁路安全问题也愈发受到重视. 在大数据时代, 如何利用铁路事故历史记录数据发掘有用信息, 建立事故预警机制, 对于推动铁路行业信息化, 提高运输效率, 防范安全隐患具有重要意义. 铁路事故类型预测和事故致因分析是建立事故预警机制的两个基础环节. 铁路事故预测利用历史事故记录估计和判断未来某种情况下是否会发生事故. 铁路事故成因分析通过分析事故发生时的客观环境与人为因素, 寻找造成事故的最可能原因, 从而采取针对性的预警防护手段. 因此, 利用铁路事故历史记录, 采用数据挖掘技术发掘其中有用信息, 进行铁路事故类型预测与成因分析具有重大现实意义.

铁路事故类型预测的本质是一个多分类问题. 常用的多分类模型有逻辑回归 (Logistic regression, LR)[1]、支持向量机 (Support vector machine, SVM)[2] 和决策树 (Decision tree, DT)[3] 等. 文献 [4] 利用决策树算法进行煤与瓦斯的突出预测. 然而, 这类分类器主要适用于简单、平衡的数据训练, 对

收稿日期 2019-09-11 录用日期 2020-01-17
Manuscript received September 11, 2019; accepted January 17, 2020
科技创新 2030-"新一代人工智能"重大项目 (2018AAA0102101), 中央高校基本科研业务费 (2018JBZ001), 国家自然科学基金 (61976018, 61532005) 资助
Supported by Science and Technology Innovation 2030 Major Program: New Generation Artificial Intelligence (2018AAA0102101), the Fundamental Research Funds for the Central Universities (2018JBZ001), National Natural Science Foundation of China (61976018, 61532005)
本文责任编委 王立威
Recommended by Associate Editor WANG Li-Wei
1. 北京交通大学信息科学研究所 北京 100044 2. 北京市现代信息科学与网络技术重点实验室 北京 100044
1. Institute of Information Science, Beijing Jiaotong University, Beijing 100044 2. Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044

于铁路事故记录这种复杂、类别失衡的高维数据, 训练较为困难, 且预测结果不够理想. 集成学习能够将多个模型集成以获取更好的预测结果, 对于不平衡数据的分类问题具有更好的有效性. 常用集成学习模型主要包括随机森林 (Random forest, RF)[5] 和梯度提升决策树 (Gradient boosting decision tree, GBDT)[6−7]. RF 基于 Bagging 思想[8], 并行集成基学习器, 模型简单, 计算开销小; 而 GBDT 则是基于 Gradient boosting 思想[6, 9], 对基学习器进行串行集成, 对数据拟合能力很强. 文献 [10−13] 分别使用以上模型进行预测.

铁路事故成因分析是对事故类型预测的反演. 常用的事故成因分析方法有复杂网络方法、灰色理论等. 文献 [14] 结合灰色综合关联度和信息熵, 利用熵分析事件不确定性的原理, 针对事故相关属性的重要度进行分析. 文献 [15] 运用多维关联规则提取技术找出事故成因关联规则. 上述事故成因分析方法对于值类别数较多的特征, 运算较复杂.

此外, 现有铁路事故记录数据存在严重的数据缺失问题, 在进行铁路事故类型预测和归因前, 首先需要对数据进行补全. 选择合适的补全方法对于提升预测结果的准确性有很大影响. 目前, 常用的补全方法主要包括均值填补法、最近距离填补法、回归填补法等[16−17]. 然而, 前两种方法在某种程度上会影响样本状态分布, 导致预测结果的偏差; 回归填补法仅适用于连续特征, 对于离散特征并不适用.

针对上述问题, 本文提出了一种基于 GBDT 的铁路事故类型预测及成因分析算法. 首先, 针对铁路事故数据缺失问题, 提出了一种基于属性分布概率的补全算法, 该算法最大程度地保持了原有的数据结构, 从而降低数据缺失对于类型预测造成的影响. 其次, 提出了一种基于 Bagging 的集成 GBDT 模型, 针对类别失衡的铁路事故历史记录数据能够进行高效训练, 得到准确的事故类型预测结果. 同时, 结合统计学习理论, 根据 GBDT 预测模型中的特征重要度排序, 实现事故致因分析. 算法整体框架如图 1 所示. 通过在公开的铁路事故数据库上进行实验, 验证了本文所提算法的有效性.

1　铁路事故缺失数据补全算法

在本节中, 我们主要介绍本文所提出的基于属性分布概率的缺失数据补全算法. 其中, 第 1.1 节给出本文所用符号的说明. 第 1.2 节对算法进行具体描述.

1.1　符号说明

为便于后文阐述, 首先对本文所用的一些符号

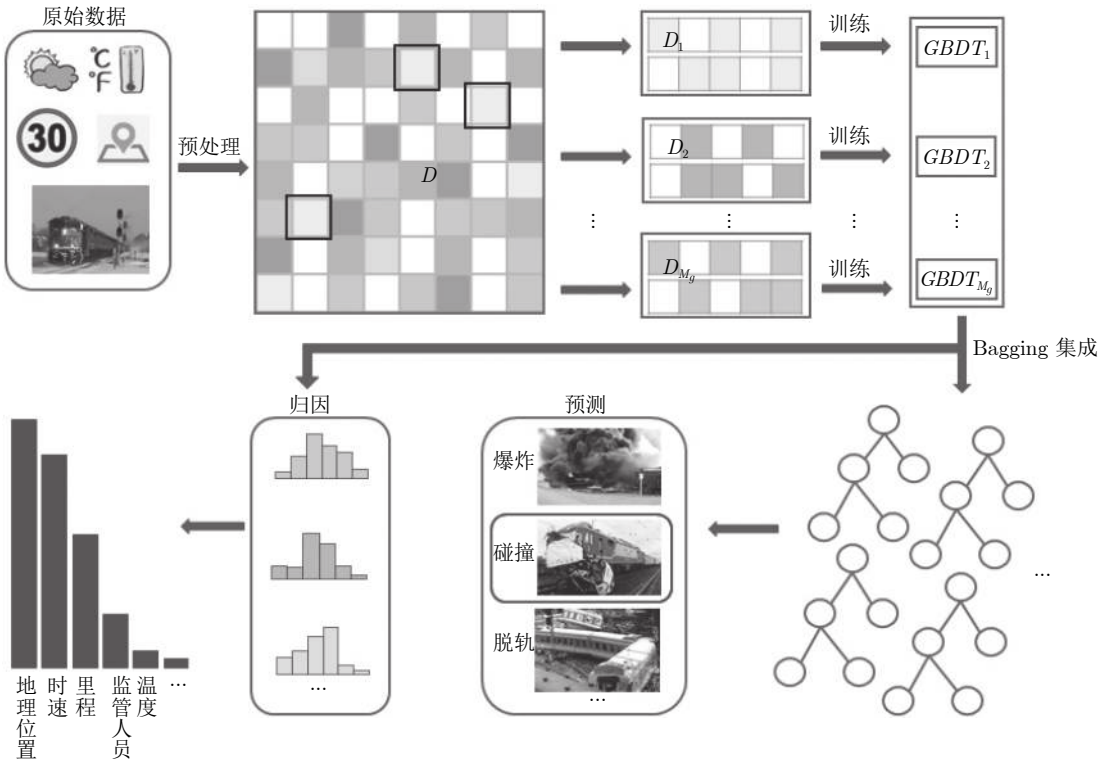

图 1　基于 GBDT 的铁路事故类型预测及成因分析框架

Fig. 1　The framework of GBDT-based railroad accident type prediction and cause analysis

进行说明. 令 $D \in \mathbf{R}^{N\times(p+1)}$ 表示记录条数为 N 的铁路设备事故数据集, 其中每条记录可表示为 $\boldsymbol{d} = [\boldsymbol{X}_i, y_i],\ 0 \le i \le N$.令 $X=[\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_N]^{\mathrm{T}} \in \mathbf{R}^{N\times p}$ 表示 N 条记录的 p 维特征空间, 其中 $\boldsymbol{X}_i = [x_i^1, x_i^2, \cdots, x_i^p] \in \mathbf{R}^{1\times p}$ 表示每一条记录的 p 维特征向量. $\boldsymbol{Y} = [y_1, y_2, \cdots, y_N]^{\mathrm{T}} \in \mathbf{R}^{N\times 1}$ 表示 N 条事故记录的类型向量, 其中, $y_i \in \{1, \cdots, C\}$, C 为事故类型总数量. 令 x^j 表示第 j 个特征, $1 \le j \le p$, 使用 a_j 表示 x^j 的取值. 若 x^j 是离散的类别型属性, 则类别 $a_j \in \{1, \cdots, k\}$, 其中 k 为 x^j 可取类别值的数量.

1.2 基于属性分布概率的补全算法

由于客观环境及人为原因等干扰因素, 导致铁路事故记录数据存在缺失, 对后续事故类型预测建模及成因分析有不利影响. 因此, 需对铁路事故数据进行缺失补全.

目前常用的补全方法包括均值补全、众数补全等. 然而, 由于铁路事故记录数据中的属性多为离散的类别型属性, 常规补全方法并不适用. 例如, 均值补全适用于连续的数值型属性; 众数补全适用于数据本身缺失较少, 其中需补全的属性的取值分布有明显偏好的情况, 对于取值分布较均衡的属性, 使用众数补全会改变原有属性取值的概率分布.

考虑到上述问题, 本文提出了一种基于属性分布概率的补全算法. 算法流程由算法 1 给出. 针对铁路事故记录数据中取值分布较均衡的离散、类别型属性 x^j , 计算现有数据下该属性所有取值 $a_j = n$ 出现的概率 P_j^n , 基于概率进行缺失值的填补, 从而在保持属性原有的分布的情况下, 完成对铁路事故数据的补全, 降低数据缺失对事故类型预测的影响.

P_j^n 计算公式如下:

$$P_j^n = \frac{A_j^n}{N_{ALL}} \tag{1}$$

表示当前 N_{ALL} 条事故记录下, 属性 x^j 取值为类别 n 的概率. A_j^n 表示属性 x^j 取值为类别 n 的个数.

算法 1. 基于属性分布概率的补全算法

输入. 待插补的特征 x^j 、取值 $a_j = 1, \cdots, k$ 的个数 $A_j^1, \cdots, A_j^k$ 、全部事故记录条数 N .

输出. 插补完成的特征 $\hat{x}^j$.

步骤 1. 计算事故记录中特征 x^j 存在的记录的条数 $N_{ALL} = \sum_{n=1}^{k} A_j^n$;

步骤 2. 计算事故记录中特征 x^j 空缺的记录条数 $N_{LACK} = N - N_{ALL}$;

步骤 3. for 特征 x^j 的所有取值 $(1, k)$

do

步骤 3.1. 计算特征 x^j 每一个取值出现的概率 $P_j^n = \frac{A_j^n}{N_{ALL}},\ n = (1, \cdots, k)$;

步骤 3.2. 计算每一个取值需要插补的次数 $S_j^n \leftarrow P_j^n \times N_{LACK}$;

步骤 4. for 特征 x^j 的所有取值 $(1, k)$

do

步骤 4.1. 将每一个要填补的取值按需要插补的次数扩展为集合 $T_j^n \leftarrow [a_j = n] * S_j^n,\ n = (1, \cdots, k)$, * 表示复制 S_j^n 次.

步骤 5. 将所有取值的集合合并为一个集合 $T_j = T_j^1 \cup T_j^2 \cup \cdots \cup T_j^k$;

步骤 6. for 每一个特征 x^j 的缺失位置 $(1, N_{LACK})$

do

步骤 6.1. 从 T_j 中随机无放回地取值填入空缺位置;

步骤 7. 输出插补完成的特征 $\hat{x}^j$.

2 铁路事故类型预测

铁路事故预测本质上是一个多分类问题. 由于铁路事故记录数据类别不均衡且属性多为离散值属性, GBDT 在处理这类数据时具有很好的有效性. 本章节详细介绍了基于改进 GBDT 的铁路事故类型预测模型. 其中, 第 2.1 节简要介绍了 GBDT 模型, 第 2.2 节对本文所提模型进行详细阐述.

2.1 GBDT 模型

GBDT 是基于 Boosting 算法[9] 的集成决策树模型. Boosting 算法依据上一次训练的残差生成基学习器. GBDT 在 Boosting 的基础上, 在残差减小的梯度方向上建立新的决策树[6–7]. GBDT 模型可表示为:

$$F_M(x) = \sum_{m=1}^{M} T(x; \Theta_m) \tag{2}$$

其中, $T(x; \Theta_m)$ 表示决策树, Θ_m 表示树的参数, M 为树的个数.

决策树 $T(x; \Theta_m)$ 的损失函数用 $L(\cdot)$ 表示, 在 GBDT 中, 损失函数为平方误差函数. 用 $T_{m-1}(x)$ 表示当前决策树, GBDT 通过最小化损失函数来确定下一棵决策树的参数 $\hat{\Theta}_m$.

$$\hat{\Theta}_m = \arg\min \sum_{i=1}^{N} L(y_i, T_{m-1}(x_i) + T(x_i; \Theta_m)) \tag{3}$$

2.2 基于 Bagging 的集成 GBDT 模型

由于铁路事故样本存在类别失衡的问题, 使用单一 GBDT 难以满足分类需求. 集成学习中的 Bagging 算法能够随机有放回地选择训练数据, 构

建基学习器, 然后将多个基学习器组合, 使用投票法或简单平均法计算分类结果[8, 18–19]. 文献 [20] 和 [21] 都是通过将多个分类器集成, 以获得更好的分类效果. 本文参考文献 [20], 提出一种基于 Bagging 的集成 GBDT 算法, 以 GBDT 作为基学习器, 利用 Bagging 算法将多个 GBDT 集成, 构造集成 GBDT 模型, 获得比单一 GBDT 优越的分类效果, 克服样本类别失衡对预测造成的影响, 实现铁路事故类型的精确预测.

算法流程如算法 2 所示. 对于输入的训练集 (X,y), 利用 Bootstrap 算法[8] 以采样率 α 随机采样 M_g 次, 得到 M_g 个训练子集, 从而构造 M_g 个 $GBDT_t$, $t=1,\cdots,M_g$; 对于每一个 $GBDT_t$ 的预测值 $\hat{y}_i$, 利用投票法, 选择 M_g 个 $GBDT_t$ 的预测结果中出现次数最多的预测值作为集成 GBDT 的最终预测结果 $\hat{y}_i$.

算法 2. 基于 Bagging 的集成 GBDT 算法

输入. 训练集 (X,y)、基学习器 $GBDT$、迭代次数 M、Bootstrap 采样率 α.

输出. 集成 GBDT 预测值 $\hat{y}_i$.

步骤 1. for 每一轮迭代 $t\in(1,\cdots,M_g)$

do

步骤 1.1. 利用 Bootstrap 算法以采样率 α 随机采样, 得到训练子集 $(X,y)_t \leftarrow Bootstrap(\alpha,(X,y))$;

步骤 1.2. 利用 $(X,y)_t$ 训练 GBDT, 得到 $GBDT_t$;

步骤 2. 利用训练好的集成 GBDT 进行预测, 选择预测结果中出现最多次的预测值作为最终结果 $\hat{y}_i \leftarrow GBDT(X_i)=\text{argmax}_c\left[\sum_{t=1}^{M_g}\Pi(GBDT_t(X_i))\right]_{c\le C}$, 其中, $\Pi(c)$ 表示一个第 c 位为 1, 其余为 0 的 C 维向量.

3　铁路事故致因分析

铁路事故致因分析是铁路事故类型预测的反演, 通过对铁路事故发生时各种因素的分析, 能够推演事故发生的过程和解析事故因果关系, 以建立事故预警机制, 进行安全防范. 由于铁路事故记录数据特征维度较大, 传统致因分析方法[14–15] 并不适用. 在进行 GBDT 模型训练时, 可以输出特征重要度, 以分析哪些特征对预测结果存在关键影响. 因此, 本文结合统计分析的方法, 基于 GBDT 的特征重要度排序[6], 进行铁路事故致因分析.

对于某一特征 x^j 的全局重要度, 通过该特征在单棵决策树中重要度的平均值来衡量, 如式 (4) 所示.

$$\hat{J}_j=\frac{1}{M}\sum_{m=1}^{M}\hat{J}_j(T(x;\Theta_m)) \tag{4}$$

其中, $\hat{J}_j(T(x;\Theta_m))$ 表示特征 x^j 在单棵树上的重要度, 公式如下:

$$\hat{J}_j(T(x;\Theta_m))=\sum_{t=1}^{L-1}\hat{i}_t^2 1(v_t=x^j) \tag{5}$$

其中, L 表示树的叶子节点数量, $L-1$ 即为树的非叶子节点数量, v_t 表示与节点 t 相关联的特征, $\hat{i}_t^2$ 是节点 t 分裂之后的平方损失的减少值[6].

分析可得, 非叶子节点 t 在分裂时的 $\hat{i}_t^2$ 越大, 说明特征越重要. 根据重要度排序筛选出特征后, 按排序将特征分组累加代入预测模型重新训练, 以验证选择的可靠性.

4　实验结果及分析

4.1　实验数据设置及预处理

本文通过在美国联邦铁路管理局 (Federal Railroad Administration, FRA)[22] 公开的铁路设备事故数据上进行实验, 验证了本文所提算法的有效性.

实验数据采用 FRA 对外公布的 2016 年至 2018 年铁路设备事故数据. 数据集包含事故类型、事故发生具体时间、地点、日期、铁路编号等信息. 原始数据集统计信息见表 1, 共 5 434 条记录, 包含 144 个属性和 11 种事故类型. 11 种事故类型描述如表 2 所示, 其中, 类型 1 (Derailment) 记录数量最多, 类型 2 (Head on collision)、类型 6 (Broken train collision) 记录数量极少.

表 1　原始数据描述

Table 1　Description of original data

	Record	Accident type	Attribute
Number	5 434	11	144

表 2　事故类型描述

Table 2　Description of accident types

Type	Description
1	Derailment
2	Head on collision
3	Rearend collision
4	Side collision
5	Raking collision
6	Broken train collision
7	Hwy-rail crossing
8	RR grade crossing
9	Obstruction
10	Fire
11	Other impacts

原始数据存在严重属性缺失情况, 如表 3 所示. 本文首先通过多次数据清洗, 去除部分与实验结果无关性较强的属性, 最终保留 69 个属性. 这 69 个属性中, 共有 23 个属性存在缺失, 缺失属性均为类别型属性. 本文采用众数补全和第 1.2 节基于属性分布概率的补全算法两种方法进行数据补全. 统计每一个缺失属性取值的概率分布, 针对缺失属性类别分布较均衡的属性, 使用本文所提算法进行补全; 对于缺失值较少或类别分布有明显偏好的属性, 采用众数补全. 针对补全后的数据, 对类别型属性进行编码, 为后续模型训练做准备. 经过预处理后数据集的统计信息描述如表 4 所示.

表 3 数据集部分示例

Table 3 Examples of the dataset

Name	Description	Number	Type
RAILROAD	Railroad code	5 434	Object
CARS	Num. of cars carrying hazmat	5 434	Int64
TYPSPD	Train speed type	5 086	Object
TRNDIR	Train direction	5 161	Float64
TONS	Gross tonnage, excluding power units	5 434	Int64
TYPEQ	Type of consist	5 081	Object
EQATT	Equipment attended	5 074	Object
CDTRHR	Num. of hours conductors on duty	3 628	Int64
ENGHR	Num. of hours engineers on duty	4 201	Int64
TRKNAME	Track identification	5 434	Object

表 4 预处理后数据描述

Table 4 Description of preprocessed data

	Record	Accident type	Attribute dimension
Number	5 434	11	119

本文采用交叉验证的方式, 随机选择 80 % 作为训练集, 20 % 作为测试集.

4.2 度量标准

本文通过在美国联邦铁路管理局 (Federal Railroad Administration, FRA)[22] 公开的铁路设备事故数据上进行实验, 验证了本文所提算法的有效性. 本文采用均方误差函数 (Mean square error, MSE) 作为补全算法有效性的评价标准, 其定义如下:

$$MSE(x^j) = \frac{1}{N_{EMP}} \sum_{t=1}^{N_{EMP}} (a_{jt} - \hat{a}_{jt})^2 \tag{6}$$

其中, N_{EMP} 表示手动设置的空值的总数, a_{jt} 表示原始值, $\hat{a}_{jt}$ 表示插补后的值.

对于铁路事故类型预测模型, 本文采用准确率 Accuracy、查准率 Precision、查全率 Recall 和 F1-score 作为评价指标.

分类准确率计算公式为:

$$Accuracy = \sum_{i \in C} \frac{N_i}{N} \left(\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \right) \tag{7}$$

查准率计算公式为:

$$Precision = \sum_{i \in C} \frac{N_i}{N} \left(\frac{TP_i}{TP_i + FP_i} \right) \tag{8}$$

查全率计算公式为:

$$Recall = \sum_{i \in C} \frac{N_i}{N} \left(\frac{TP_i}{TP_i + FN_i} \right) \tag{9}$$

F1-score 计算公式为:

$$F1\_score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

其中, C 表示所有事故类型的总数; N_i 表示事故类型为 i 的样本个数, N 表示样本总个数; TP_i 表示被正确预测为第 i 类的个数; TN_i 表示被正确预测不为第 i 类的个数; FP_i 表示被错误预测为第 i 类的个数; FN_i 表示被错误预测为不为第 i 类的个数.

4.3 补全算法对比实验结果及分析

为验证基于属性分布概率的补全算法的有效性, 本文将所提算法与插值法 (Interpolation completer)、众数补全 (Mode completer) 两种补全方法进行比较. 基于属性分布概率的补全算法最大程度地保持了原始数据的分布结构. 以特征 TRNDIR 为例. 特征 TRNDIR 有 4 种取值, $a_j \in \{1, 2, 3, 4\}$ 表示火车运行的四个方向. 表 5 展示了使用三种方法进行补全后与补全之前 4 种取值的概率分布. 从表 5 可以看出, 使用插值法与众数补全法补全后, 造成该特征某一取值过多, 破坏了原本的数据分布, 而本文所提算法完全不改变原有的概率分布, 从而减少了由于数据缺失对铁路事故类型预测带来的影响.

为进一步定量分析基于属性分布概率的补全算法的有效性, 本实验以均方误差函数 (MSE) 作为评价标准, 对 3 种补全方法进行对比. 以 TRNDIR、ENGHR、CDTRHR (特征描述见表 3) 三个特征为例, 随机从数据集中选择 100 条记录, 设置以上三个特征的值为空, 用三种补全算法依次进行补全, 记录每一种补全算法 MSE. 共进行 10 次实验, 取 10 次 MSE 之和的平均值进行对比, 实验结果如图 2 所示. 由图 2 可得, 基于属性分布概率的补全算法 MSE 明显低于其他两种方法, 表明本文所提算法具有很好的有效性.

表 5　三种方法补全前后特征 *TRNDIR* 取值分布
Table 5　Distribution of the attribute *TRNDIR* values before and after three completion methods

Algorithm	$a_j=1$	$a_j=2$	$a_j=3$	$a_j=4$
Before completion	0.22	0.20	0.31	0.27
Interpolation	0.21	0.19	0.30	0.30
Mode	0.21	0.19	0.34	0.26
Our algorithm	0.22	0.20	0.31	0.27

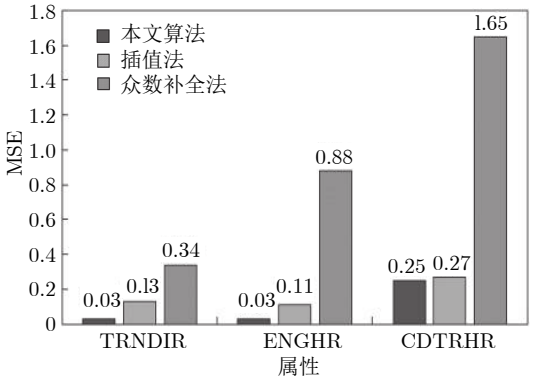

图 2　三种补全方法结果对比
Fig. 2　Comparison of three methods results

4.4　模型集成实验结果及分析

为验证构造基于 Bagging 的集成 GBDT 模型时不同因素的影响, 本节对不同参数下集成 GBDT 的效果进行了对比, 以确定进行铁路事故类型预测任务时的最佳参数设置.

基于 Bagging 的集成 GBDT 模型需要调优的参数可分为两类, 包括 Bagging 框架参数和 GBDT 参数, GBDT 参数又包括 Boosting 框架参数和决策树参数. 其中, Bagging 框架参数包括最大迭代次数, 即集成的 GBDT 数量, 以及最大采样率; Boosting 框架参数包括最大迭代次数, 即子树的最大数量, 以及学习步长等; 决策树的参数主要包括树的深度.

首先, 利用网格搜索法对单一 GBDT 的参数进行调优. 经过调优后, GBDT 的迭代次数为 100, 学习步长为 0.2, 决策树最大深度为 6. 此时 GBDT 在测试集上的预测准确率为 84.1 %.

得到最优 GBDT 参数组合后, 对集成 GBDT 的 Bagging 框架参数进行调优, 考虑运行效率和分类性能, 以选择合适的 GBDT 数量及采样率. 在实验中, 首先确定 GBDT 个数, 分别用 5、10、15、20、25、30 个 GBDT 进行集成, 此时最大采样率设置为 0.9, 以在测试集上预测结果的准确率和模型训练时间作为评价标准, 结果如图 3 所示. 数量为 1 时表示不进行集成, 仅用单一 GBDT 进行预测. 可以看出, 当 GBDT 个数增加时, 模型预测准确率呈上升趋势, 表明使用 Bagging 进行集成的方法确实有效. 当 GBDT 个数为 15 和 25 时, 模型预测准确率最高, 达到 85 % ~ 85.2 %, 比单一 GBDT 预测准确率高出约 1 个百分比, 但使用 15 个 GBDT 训练的时间是使用 25 个 GBDT 训练时间的 1/2. 综合考虑分类效果和性能, 最终预测模型使用 15 个 GBDT 进行集成. 为进一步确定采样率, 分别将采样率设置为 0.6、0.7、0.8、0.9、1.0 进行实验, GBDT 的数量设置为 15, 以预测结果的准确率作为评价标准. 最终结果如表 6 所示. 可以看出, 当采样率为 0.9 时, 模型预测准确率最高.

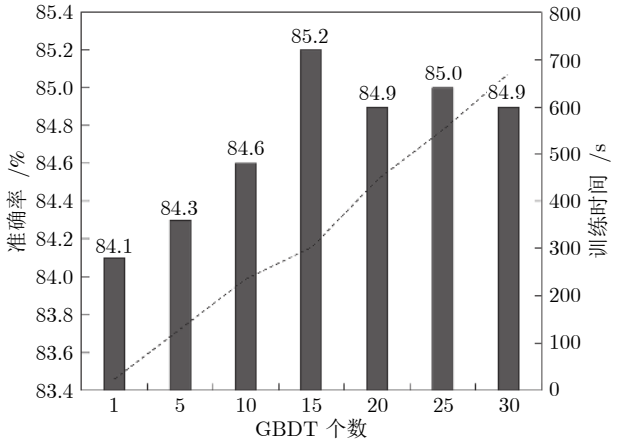

图 3　不同 GBDT 集成个数下分类准确率
Fig. 3　Accuracy of classifiers with different number of GBDT

表 6　不同采样率下集成 GBDT 分类准确率
Table 6　Accuracy of classifiers with different sampling rates

α	0.6	0.7	0.8	0.9	1.0
Accuracy	0.841	0.846	0.845	0.852	0.848

4.5　模型选择实验结果及分析

为进一步验证基于 Bagging 的集成 GBDT 模型的有效性, 本实验中使用相同的训练集, 分别对 DT[3]、RF[5]、ET[8]、GBDT[6–7] 和集成 GBDT (Ensemble GBDT) 进行训练, 在相同测试集上进行测试, 对比预测结果. 共进行 10 次实验, 取 10 次结果的平均值作为最终结果. 其中, 根据第 4.4 节实验, 集成 GBDT 的参数设置为: 集成的 GBDT 个数为 15, 采样率为 0.9, 每个 GBDT 的迭代次数为 100, 学习步长为 0.2, 决策树最大深度为 6.

分类结果用混淆矩阵表示, 如图 4 所示. 分类

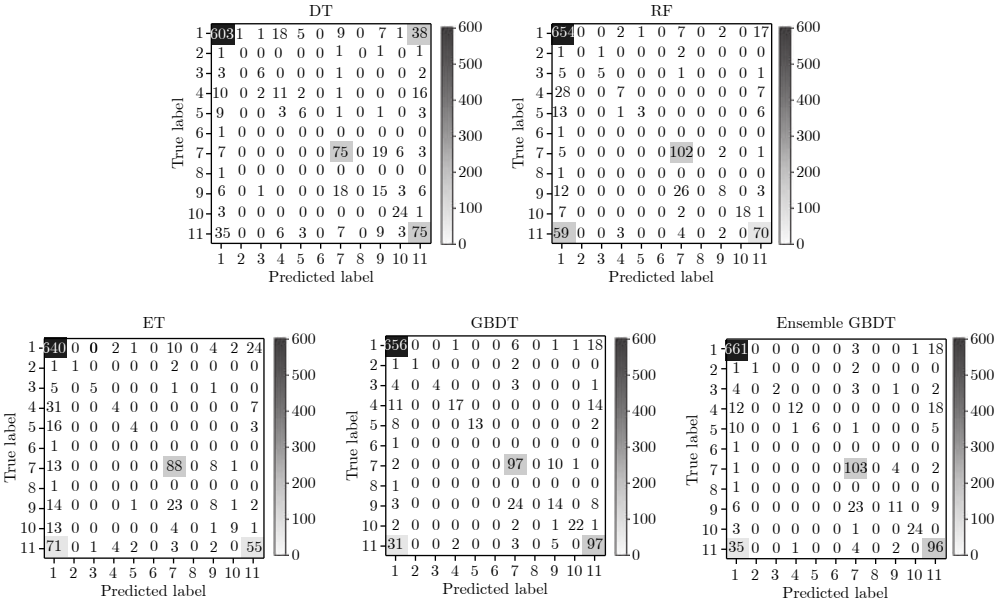

图 4 混淆矩阵

Fig. 4 Confusion matrix

效果如表 7 所示. 从表 7 可以看出, 单一 GBDT 的分类 F1-score 较其他 3 种分类器高出 10 % ~ 14 %; 进行集成后查全率和召回率比单一 GBDT 提高了约 1 %, 效果最佳.

表 7 各分类器性能对比

Table 7 Performance comparison of classifiers

Classifier	Accuracy	Precision	Recall	F1
DT	0.728	0.73	0.73	0.73
RF	0.773	0.74	0.77	0.75
ET	0.734	0.70	0.73	0.71
GBDT	0.841	0.84	0.84	0.84
Ensemble GBDT	0.852	0.85	0.85	0.85

因集成 GBDT 在每个 GBDT 训练时随机选择训练样本, 降低了样本类别失衡造成的影响, 且将 Bagging 与 Boosting 结合的方式充分考虑了模型过拟合问题, 提高了模型的泛化能力, 从而提高了分类的准确率, 故而效果最优.

4.6 特征选择实验结果及分析

本文根据特征重要度进行特征选择, 将选择特征按重要度排名分组累加, 代入模型重新训练, 以进行事故致因分析. 为验证特征选择的有效性及可靠性, 在本节实验中进行了不同特征组合对事故类型预测结果的对比. 将单一 GBDT 训练中, 特征重要性大于 0.001 的特征筛选出来用于集成 GBDT 模型训练, 训练后的模型在测试集上分类准确率提高了 1.6～1.8 个百分点, 表明基于 GBDT 的特征选择具有一定的可靠性. 为进一步分析所选特征的正确性, 将特征按重要度降序排列, 以十个为一组依次累加代入集成 GBDT 模型训练, 结果如图 5 所示. 实验结果表明, 随着特征数量的增多, 分类准确率呈现上升趋势且逐步逼近于使用全部特征训练所得准确率, 说明所选特征符合重要度排序. 当特征数量大于 30 时, 分类准确率趋于平稳, 表明之后增加的特征对预测结果几乎没有影响, 进一步验证了特征符合重要度排序.

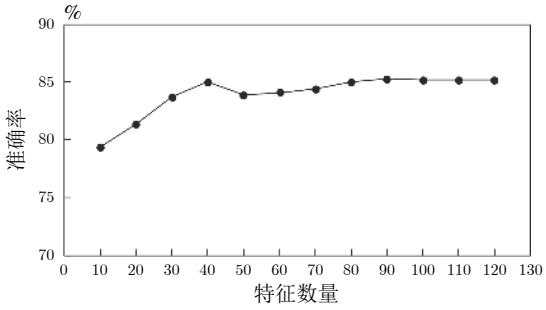

图 5 不同特征数量下预测结果

Fig. 5 Prediction results of classifier with different features

为进行铁路事故成因分析, 本文选择排名前 15 的特征进行总结, 如表 8 所示. 将这 15 个特征按共性可划分为地理位置 (Location)、速度 (Speed)、里程 (Mileage)、天气 (Weather)、载货 (Freight)、监管人员因素 (Manager) 六大类. 以脱轨 (Derail-

ment) 和碰撞 (Collision) 两类事故为例, 综合进行事故成因分析, 结果如图 6 所示. 由图 6 可知, 铁路事故发生与地理位置、列车行驶速度等有重要联系, 温度和人员因素与事故发生也有一定联系. 结果符合常规事故成因, 具有可靠性.

表 8　重要度排名前 15 的特征

Table 8　Features of top 15 in importance

No.	Name	Description
1	Latitude	Latitude in decimal degrees
2	Longitude	Longitude in decimal degrees
3	CNTYCD	FIPS county code
4	HIGHSPD	Maximum speed
5	TRKNAME	Track identification
6	RRCAR1	Car initials (fist involved)
7	TEMP	Temperature in degrees fahrenheit
8	MILEPOST	Milepost
9	STATION	Nearest city and town
10	TRNSPD	Speed of train in miles per hour
11	RRCAR2	Car initials (causing)
12	SUBDIV	Railroad subdivision
13	ENGHR	Num. of hours engineers on duty
14	CDTRHR	Num. of hours conductors on duty
15	TONS	Gross tonnage

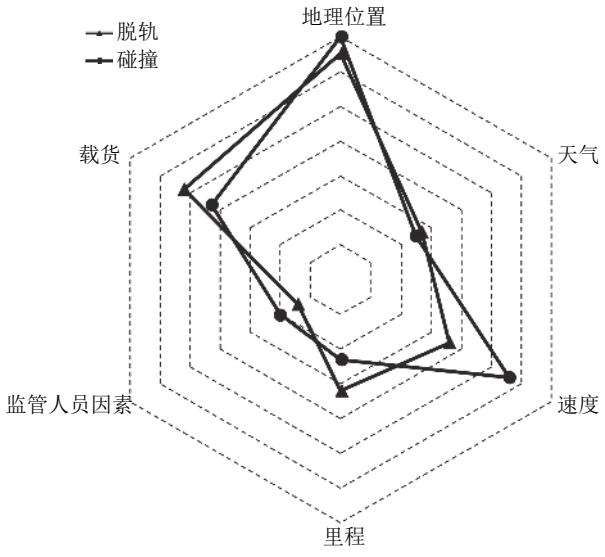

图 6　两类事故致因中不同因素的比例

Fig. 6　Proportion of different factors in causes of two types of railroad accidents

5　结论

本文提出了一种基于 GBDT 的铁路事故类型预测和成因分析模型. 针对铁路事故记录数据缺失的问题, 提出基于属性分布概率的补全算法, 以保持原有数据结构, 减少数据缺失对预测结果的影响. 由于铁路事故数据存在类型失衡等问题, 对预测结果也存在很大影响. 为此, 本文基于 Bagging 对 GBDT 进行集成, 提高了单一 GBDT 的预测精度. 同时, 结合统计分析的方法, 根据特征重要度进行特征选择, 进而对特征进行分析和总结, 推测铁路事故成因, 减少了人力的投入. 实验证明, 本文方法具有很好的可靠性和有效性.

References

1　Ming L. Data mining: concepts, models, methods, and algorithms. *IIE Transaction*, 2004, **36**(5): 495−496

2　Feng Shi-Yong. *Regression Analysis Method*. Beijing: Science Press, 1974
(冯士雍. 回归分析方法. 北京: 科学出版社, 1974)

3　Rutkowski L, Jaworski M, Pietruczuk L, Duda P. Decision trees for mining data streams based on the gaussian approximation. *IEEE Transactions on Knowledge and Data Engineering*, 2013, **26**(1): 108−119

4　Li Ding-Qi, Cheng Yuan-Ping, Wang Hai-Feng, Wang Liang, Zhou Hong-Xing, Sun Jian-Hua. Coal and gas outburst prediction based on improved decision tree ID3 algorithm. *Journal of China Coal Society*, 2011, **36**(4): 619−622
(李定启, 程远平, 王海峰, 王亮, 周红星, 孙建华. 基于决策树 ID3 改进算法的煤与瓦斯突出预测. 煤炭学报, 2011, **36**(4): 619−622)

5　Breiman L. Random forest. *Machine Learning*, 2001, **45**(1): 5−32

6　Friedman J H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 2001, **29**(5): 1189−1232

7　Friedman J H. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 2002, **38**(4): 367−378

8　Zhou Zhi-Hua. *Machine Learning*. Beijing: Tsinghua University Press, 2016.
(周志华. 机器学习. 北京: 清华大学出版社, 2016.)

9　Schonlau M. Boosted regression (boosting): An introductory tutorial and a Stata plugin. *The Stata Journal*, 2005, **5**(3): 330−354

10　Weng Xiao-Xiong, Lv Pan-Long. Subway IC card commuter crowd identification based on GBDT algorithm. *Journal of Chongqing Jiaotong University (Natural Science)*, 2019, **38**(5): 8−12
(翁小雄, 吕攀龙. 基于 GBDT 算法的地铁 IC 卡通勤人群识别. 重庆交通大学学报 (自然科学版), 2019, **38**(5): 8−12)

11　Mursalin M, Zhang Y, Chen Y H, Chawla N V. Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. *Neurocomputing*, 2017, **241**: 204−214

12　Cheng J, Li G, Chen X H. Research on travel time prediction model of freeway based on gradient boosting decision tree. *IEEE Access*, 2018, **7**: 7466−7480

13　Ma X, Ding C, Luan S, Wang Y, Wang Y P. Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method. *IEEE Transactions on Intelligent Transportation Systems*, 2017, **18**(9): 2303−2310

14 Su H W, Zhang W J, Li Z H. Analysis and prediction of water traffic accidents in jingtang port based on improved GM(1, 1) model. In: Proceedings of the 37th Chinese Control Conference (CCC). New York, USA: IEEE, 2018.2212−2217

15 Das S, Sun X D. Investigating the pattern of traffic crashes under rainy weather by association rules in data mining. In: Proceedings of the 93rd Transportation Research Board (TRB) Annual Meeting. Washington D.C., USA: Nation Academy of Sciences, 2014

16 Jin Yong-Jin. *Statistical Processing of Missing Data*. Beijing: China Statistics Press, 2009.
(金勇进. 缺失数据的统计处理, 北京: 中国统计出版社, 2009.)

17 Jin Yong-Jin. Data loss and processing in survey (I) data missing and impact. *Journal of Applied Statistics and Management*, 2001, **20**(1): 59−62
(金勇进. 调查中的数据缺失及处理 (I)-缺失数据及其影响. 数理统计与管理, 2001, **20**(1): 59−62)

18 Collell G, Prelec D, Patil K R. A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing*, 2018, **275**: 330−340

19 Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, 2012, **42**(4): 463−484

20 Zhu Zhen-Feng, Tang Jing-Yuan, Chang Dong-Xia, Zhao Yao. GBDT based hierarchical model for commodity distribution prediction. *Journal of Beijing Jiaotong University*, 2018, **42**(2): 9−13, 45
(朱振峰, 汤静远, 常冬霞, 赵耀. 基于 GBDT 的商品分配层次化预测模型. 北京交通大学学报, 2018, **42**(2): 9−13, 45)

21 Yang Lian-Bao, Li Ping, Xue Rui, Ma Xiao-Ning, Wu Yan-Hua, Zou Dan. Intelligent classification of faults of railway signal equipment based on imbalancd text data mining. *Journal of the China Railway Society*, 2018, **40**(2): 59−66
(杨连报, 李平, 薛蕊, 马小宁, 吴艳华, 邹丹. 基于不平衡文本数据挖掘的铁路信号设备故障智能分类. 铁道学报, 2018, **40**(2): 59−66)

22 Federal Railroad Administration Office of Safety Analysis [Online], available: https://safetydata.fra.dot.gov/OfficeofSafety/Default.aspx, June 1, 2019

钟敏慧 北京交通大学信息科学研究所硕士研究生. 主要研究方向为计算机视觉, 机器学习.
E-mail: mhzhong@bjtu.edu.cn
(**ZHONG Min-Hui** Master student at the Institute of Information Science, Beijing Jiaotong University. Her research interest covers computer vision and machine learning.)

张婉露 北京交通大学信息科学研究所硕士研究生. 主要研究方向为计算机视觉, 深度学习.
E-mail: wlzhang@bjtu.edu.cn
(**ZHANG Wan-Lu** Master student at the Institute of Information Science, Beijing Jiaotong University. Her research interest covers computer vision and deep learning.)

李有儒 北京交通大学信息科学研究所硕士研究生. 主要研究方向为数据挖掘, 机器学习.
E-mail: liyouru@bjtu.edu.cn
(**LI You-Ru** Master student at the Institute of Information Science, Beijing Jiaotong University. His research interest covers data mining and machine learning.)

朱振峰 北京交通大学信息科学研究所教授. 2005 年获中国科学院自动化研究所模式识别国家重点实验室工学博士学位. 主要研究方向为图像视频分析与理解, 计算机视觉, 机器学习. 本文通信作者.
E-mail: zhfzhu@bjtu.edu.cn
(**ZHU Zhen-Feng** Professor at the Institute of Information Science, Beijing Jiaotong University. He received his Ph. D. degree in pattern recognition and intelligence system from the Institute of Automation, Chinese Academy of Sciences in 2005. His research interest covers image and video understanding, computer vision and machine learning. Corresponding author of this paper.)

赵 耀 北京交通大学信息科学研究所教授, 所长. 1996 年获北京交通大学工学博士学位. 主要研究方向为图像与视频编码, 数字水印与取证, 视频分析及理解, 人工智能.
E-mail: yzhao@bjtu.edu.cn
(**ZHAO Yao** Professor and director at the Institute of Information Science, Beijing Jiaotong University. He received his Ph. D. degree from Beijing Jiaotong University in 1996. His research interest covers image/video coding, digital watermarking and forensics, video analysis and understanding and artificial intelligence.)