**中文引用格式**:薛乐,于露,金龙哲,等. 基于 RF-Apriori 算法考虑填补缺失值的高速公路事故致因分析[J]. 中国安全科学学报,2025,35(4):211-218.

**英文引用格式**:XUE Le,YU Lu,JIN Longzhe,et al. Causal analysis of highway accidents considering filling in missing values based on RF-Apriori algorithm[J]. China Safety Science Journal,2025,35(4):211-218.

# 基于 RF-Apriori 算法考虑填补缺失值的高速公路事故致因分析*

薛 乐[1], 于 露**[1]讲师, 金龙哲[2]教授, 李 博[1]教授, 沈文进[1]

(1 大连交通大学 交通工程学院,辽宁 大连 116028;
2 北京科技大学 大安全科学研究院,北京 100083)

**中图分类号**:X928.02 **文献标志码**:A **DOI**: 10.16265/j.cnki.issn1003-3033.2025.04.0774

**资助项目**:辽宁省教育厅基本科研项目(LJKQZ20222462)。

**【摘 要】** 为改善高速公路交通安全状况,以法国2018—2022年的26 320条高速公路交通事故数据作为研究对象,选择3种具有代表性的算法填补数据中的缺失值,包括随机森林(RF)算法、期望最大化(EM)算法以及K最近邻(KNN)算法。并基于填补前后变量方差的变化比较不同填补算法对数据稳定性的影响,并运用Apriori关联规则算法对完成填补的事故数据进行不同严重程度等级的高速公路事故致因分析。结果表明:经缺失值填补后,RF算法稳定性更优,相较于原始数据训练的模型准确率提高5.66%,召回率提高9.22%,$F_1$分数提高9.91%。客车更易引发财产损失事故的发生;摩托车在限速较低的路段易引发受伤事故,在限速较高的路段易引发死亡事故,安全设备的使用情况对事故严重程度等级有较大关系。

**【关键词】** 随机森林(RF); Apriori算法; 缺失值; 高速公路; 事故致因; 数据填补; 关联规则

# Causal analysis of highway accidents considering filling in missing values based on RF-Apriori algorithm

**XUE Le[1], YU Lu[1], JIN Longzhe[2], LI Bo[1], SHEN Wenjin[1]**

(1 School of Transportation Engineering, Dalian Jiaotong University, Dalian Liaoning 116028, China;
2 Research Institute of Macro-Safety Science, University of Science and Technology Beijing, Beijing 100083, China)

**Abstract**: In order to improve the safety condition of highways, 26 320 highway traffic accident records in France from 2018 to 2022 were selected as the research object. Three representative algorithms were selected to impute missing values in the data, including the RF algorithm, the expectation-maximization (EM) algorithm, and the K-nearest neighbors (KNN) algorithm. The impact of different imputation algorithms on data stability was compared based on the changes in variable variance before and after imputation. The Apriori association rule algorithm was then applied to analyze the causes of highway accidents with different severity levels using the completed dataset. The results indicate that after missing

* 文章编号:1003-3033(2025)04-0211-08; 收稿日期:2024-11-14; 修稿日期:2025-01-08
** 通信作者:于 露(1990—),女,黑龙江大庆人,博士,讲师,主要从事交通安全系统工程等方面的研究。E-mail:yulubaobeihao@163.com。

value imputation, the RF algorithm demonstrates superior stability. Compared to the model trained on the original data, the accuracy is improved by 5.66%, the recall rate is increased by 9.22%, and the F1 score is enhanced by 9.91%. It is found that passenger vehicles are more likely to cause property damage accidents; motorcycles are prone to cause injury accidents on roads with lower speed limits and fatal accidents on roads with higher speed limits. The use of safety equipment is significantly related to the severity level of accidents.

**Keywords**: random forest(RF); Apriori algorithm; missing value; highway; accident cause; data filling; association rules

## 0 引 言

截止 2022 年底,全国(不含港澳台)高速公路运行里程达到 17.73 万 km,比上一年增长 0.82 万 km[1]。因高速公路具有行车速度快、通行能力大等优点,成为人们中长途出行的首要选择。然而,高速公路在给人们带来高效出行服务的同时,也潜在许多不安全因素。统计资料显示,我国近年来道路交通事故发生数、死亡人数呈现出上升的趋势[2]。高速公路较普通等级公路而言,一旦发生事故,造成的伤亡与财产损失也十分巨大[3]。学者们通过挖掘事故历史数据,运用传统事故致因理论或机器学习等方法研究事故的致因机制。但在数据采集、运输、存储过程中,往往会因为人员、技术、管理等问题造成数据的缺失,继而对后续的建模分析产生影响,导致模型准确性下降、误判风险上升。

在数据填补方面,常用统计分析和机器学习填补,不同方法对不同数据的填补效果各有优异。DEB 等[4]提出一种利用数据内和数据间的相关性来填补缺失数据的数值或分类值方法;QI Hang 等[5]采用贝叶斯-高斯混合模型插补不同缺失率和缺失场景下的缺失速度数据;吴郁等[6]运用随机森林(Random Forest,RF)插补船舶碰撞事故数据的缺失值,并对比 Logistic 回归、Probit 回归、朴素贝叶斯方法,结果表明:RF 法误分率最低;陆嘉铭[7]、谢翘楚[8]、梅玉杰[9]等分别运用 RF 填补电力系统中出现的电力量测数据、电网数据的缺失数据,验证了 RF 对缺失数据填补的有效性。

在事故致因分析方面,严利鑫等[10]采用边际效用分析得出交通事故严重程度的显著影响因素;邱文利等[11]通过改进 Apriori 算法挖掘影响高速公路交通事故的关联规则,对比传统 Apriori 算法,关联规则准确性提升 81.3%,挖掘效率提升 86.5%;牛玥铧等[12]采用系统评价和 Meta 分析的优先报告项目标准从文献中分析人格特征与不安全驾驶行为之间的关系;王卓伦等[13]采用多期双重差分法和逐步法研究网约车对道路安全的影响,发现网约车会使城市道路交通事故的发生率提高 5.6%;于雷[14]建立主成分分析-灰色关联分析组合模型,分析简易程序交通事故和非简易程序交通事故致因,直接删除缺失数据,结果证明:该组合模型具有一定可靠性;孙维富[15]构建基于相对危险暴露量理论与 Logistic 回归分析的高速公路交通事故致因分析模型,删除缺失值、填补缺省值(均值、中位数、众数),研究表明:驾龄、是否超载等与高速公路交通事故的发生存在显著相关性。

综上,尽管学者已探索交通信息数据的填补方法,但尚未有专门针对高速公路交通事故信息数据的 RF 填补算法的研究和实际应用。在以往分析交通事故致因中,学者们往往采取简单的策略来处理缺失数据,如直接删除或使用缺省值进行填补,这种做法不仅限制事故致因分析模型的准确率,还对其精度产生严重的负面影响。

鉴于此,笔者利用 RF 算法补全高速公路交通事故数据的缺失值,并在此基础上,基于改进 Apriori 关联规则算法分析事故致因,以期改善高速公路交通安全状况提供理论支撑。

## 1 高速交通事故数据预处理

### 1.1 高速公路交通数据来源

考虑到我国高速公路交通事故数据获取困难的问题,选取法国道路交通事故数据集[16],从 2018—2022 年中的 276 187 条道路交通事故数据中提取出 26 320 条高速公路交通事故数据进行分析。剔除事故标签、经度、纬度等与事故致因模型相关性较小的变量后,得到 23 条事故自变量标签,涉及人员、车辆、道路、环境等信息,1 条事故因变量标签,按公安部 2017 年颁布的《道路交通事故处理程序规定》将其划分为财产损失事故、受伤事故和死亡事故 3 类,其中,财产损失事故 11 268 起,受伤事故 11 947 起,

死亡事故 3 105 起。

### 1.2　高速公路交通数据编码

高速公路交通事故数据变量维度较多且相对离散,不利于后续的建模分析。通过数据变换,将高速公路交通事故数据转换、合并成适合数据挖掘的描述形式[15]。具体变量定义及编码见表 1。

**表 1　变量定义及编码**

**Table 1　Variable definition and encoding**

| 变量类型 | 分类描述 | 编码 |
|---|---|---|
| 事故严重程度 | — | 0. 财产损失事故;1. 受伤事故;2. 死亡事故 |
| 事故原因 | — | 1. 超速;2. 疲劳驾驶;3. 违规变道;4. 制动不当;5. 违法停车;<br>6. 无证驾驶;7. 违规使用灯光;8. 其他影响安全行为;9. 其他操作不当 |
| 驾驶人特征 | 性别 | 1. 男;2. 女 |
| | 年龄/岁 | 1. (0,20];2. (20,30];3. (30,40];4. (40,50];5. (50,60];6. (60,100) |
| | 出行目的 | 1. 工作;2. 上学;3. 购物;4. 旅游;5. 其他 |
| | 安全设备使用情况 | 0. 未使用;1. 使用 |
| 车辆特征 | 车辆类型 | 1. 摩托车;2. 小型汽车;3. 客车;4. 货车;5. 其他 |
| | 撞击物 | 0. 撞击固定物;1. 撞击移动物 |
| | 撞击点 | 1. 正前方;2. 右前方;3. 左前方;4. 正后方;5. 右后方;<br>6. 左后方;7. 右侧;8. 左侧;9. 多方位 |
| | 发动机类型 | 1. 燃油;2. 混合动力;3. 电动;4. 氢气;5. 其他 |
| | 行驶意图 | 1. 不改变行车方向;2. 掉头、倒车、逆行;3. 2 车道之间;<br>4. 驶离高速;5. 变道;6. 紧急避让;7. 停车;8. 其他操作 |
| | 安全设备存在 | 0. 不存在;1. 存在 |
| 道路特征 | 车道总数 | 具体车道数 |
| | 道路坡度 | 1. 平坡;2. 缓坡;3. 陡坡;4. 陡峭坡 |
| | 道路线型 | 1. 直线;2. 左弯曲;3. 右弯曲;4. “S”型 |
| | 路表状况 | 1. 干燥;2. 潮湿;3. 积水;4. 积雪;9. 其他 |
| | 基础设施 | 0. 无;1. 隧道;2. 桥梁;3. 立交桥;4. 收费站;5. 其他 |
| | 最高限速/($km \cdot h^{-1}$) | ① 90;② 110;③ 130 |
| 环境特征 | 时间 | ① 0:00—6:00;② 6:00—12:00;③ 12:00—18:00;④ 18:00—24:00 |
| | 日期 | 1. 上旬;2. 中旬;3. 下旬 |
| | 季节 | 1. 春季;2. 夏季;3. 秋季;4. 冬季 |
| | 天气 | 1. 晴天;2. 小雨;3. 大雨;4. 雪;5. 雾;6. 强风;7. 烈日;8. 多云;9. 其他 |
| | 照明 | 1. 白天;2. 黄昏或黎明;3. 夜间无公共照明;4. 夜间有公共照明 |
| | 地点 | 1. 市区外;2. 市区内 |

## 2　缺失数据填补

### 2.1　数据缺失情况统计分析

统计分析数据缺失情况,缺失热力图如图 1 所示。从图 1 可以看出,最高限速、出行目的以及安全设备使用缺失值数量占比较大。事故因变量严重程度无缺失值存在,自变量中环境属性无缺失值存在,包含缺失值的自变量共有 16 个因素,具体变量缺失情况见表 2。

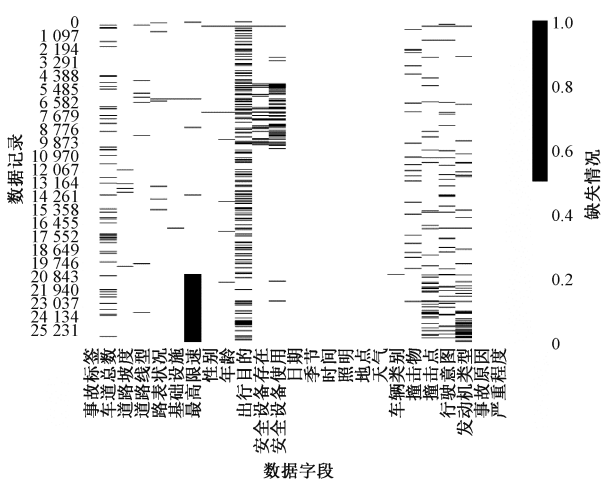


**图 1　缺失值热力图**

**Fig. 1　Heat map of missing values**

### 2.2　RF 算法

RF 算法基于决策树的集成算法,通过有放回地随机抽样形成多个训练样本,分别训练后得到相应数量的决策树组成 RF,由所有决策树投票或取均值

表 2 数据缺失情况统计

Table 2 Statistical of missing data

| 事故变量 | 缺失数量 | 缺失比率/% | 事故变量 | 缺失数量 | 缺失比率/% |
|---|---|---|---|---|---|
| 性别 | 294 | 1.1 | 行驶意图 | 2 654 | 10.0 |
| 年龄 | 738 | 2.8 | 发动机类型 | 2 128 | 8.0 |
| 出行目的 | 7 746 | 29.4 | 车道总数 | 1 568 | 5.9 |
| 安全设备存在 | 1 409 | 5.3 | 道路坡度 | 213 | 0.8 |
| 安全设备使用 | 3 400 | 12.9 | 道路线型 | 319 | 1.2 |
| 车辆类别 | 163 | 0.6 | 路表状况 | 548 | 2.0 |
| 撞击物 | 1 078 | 4.0 | 基础设施 | 643 | 2.4 |
| 撞击点 | 3 043 | 11.5 | 最高限速 | 6 218 | 23.6 |

得到最终结果[17]。RF 基本流程如图 2 所示。

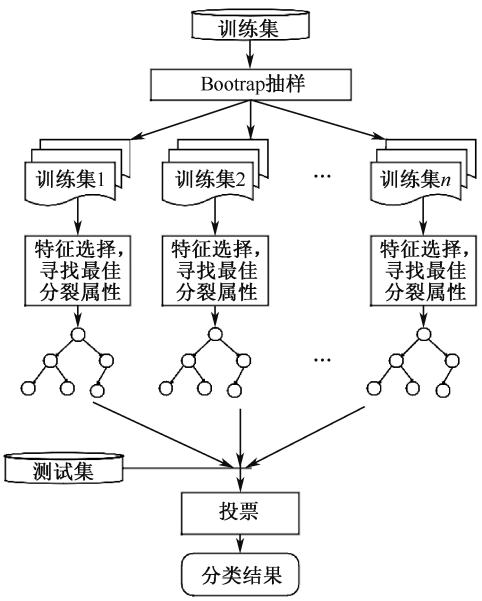


图 2 RF 基本流程

Fig. 2 Basic flowchart of RF

## 2.3 算法对比

现有数据填补方法分为统计学方法和机器学习方法[18]。为更加全面地验证 RF 算法的数据填补效果，选取统计学填补方法中的期望最大化(Expectation Maximization，EM)填补算法、机器学习填补方法中的 K 最近邻(K-Nearest Neighbor，KNN)填补算法作为对比。选取自变量方差作为对比指标。

使用同样的运行环境，16 个含有缺失值的自变量在经过不同算法填补后的方差见表 3。在缺失值的填补过程中，填补的值往往是根据数据中已有的信息计算得到。若填补后的数据方差变小，则说明填补后的数据较原始数据更加稳定，变异性减小[19]。由表 3 可知：通过不同算法填补缺失值后，有 15 个自变量经由 RF 算法填补后的方差最小，从统计学意义上证明 RF 算法填补数据更加稳定。

表 3 自变量方差对比

Table 3 Comparison of variances of independent variables

| 事故自变量 | 原始数据 | RF | EM | KNN |
|---|---|---|---|---|
| 车道总数 | 3.840 | 3.620 | 3.792 | 3.723 |
| 道路坡度 | 0.320 | 0.318 | 0.318 | 0.318 |
| 道路线型 | 0.405 | 0.401 | 0.402 | 0.402 |
| 路标状况 | 0.402 | 0.394 | 0.397 | 0.396 |
| 基础设施 | 1.096 | 1.071 | 1.077 | 1.077 |
| 最高限速 | 0.767 | 0.628 | 0.768 | 0.594 |
| 性别 | 0.188 | 0.185 | 0.186 | 0.186 |
| 年龄 | 1.785 | 1.736 | 1.792 | 1.753 |
| 出行目的 | 2.734 | 1.955 | 2.772 | 2.284 |
| 安全设备存在 | 0.012 | 0.011 | 0.013 | 0.011 |
| 安全设备使用 | 0.242 | 0.216 | 0.234 | 0.230 |
| 车辆类别 | 0.223 | 0.222 | 0.222 | 0.223 |
| 撞击物 | 0.152 | 0.147 | 0.153 | 0.149 |
| 撞击点 | 5.441 | 4.844 | 5.096 | 5.010 |
| 行驶意图 | 3.900 | 3.522 | 3.665 | 3.627 |
| 发动机类型 | 0.452 | 0.420 | 0.438 | 0.447 |

## 2.4 模型结果对比

仅分析填补前后的数据方差变化判断填补效果不够全面，具有一定的局限性。为进一步分析模型填补效果，构建极端梯度提升(Extreme Gradient Boosting，XGboost)事故严重程度预测模型，选取准确率、召回率、$F_1$ 分数作为模型评价指标，并运用不同填补算法对比分析填补的数据。

基于 Python3.11 构建 XGboost 模型，采用贝叶斯寻优算法确定模型最佳参数，以相同的运行环境和模型参数设置，使用测试集样本检验模型的泛化能力，不同算法填补后模型混淆矩阵如图 3 所示。对比不同算法填补数据在 XGboost 模型中的评价指标，见表 4。

由图 3 和表 4 可知：利用统计学方法和机器学习方法填补缺失值后，模型指标均得到一定改善。其中，经 RF 算法填补后的数据训练效果最优。这表明：对于当前的数据集而言，RF 算法在数据填补方面展现出最佳的效果，显著提升模型的预测性能。

表 4　XGboost 模型评价指标

Table 4　XGboost model evaluation indicators

| 数据填补方法 | 评价指标 | | | | | |
|---|---|---|---|---|---|---|
| | 准确率 | 较原始数据性能提升/% | 召回率 | 较原始数据性能提升/% | $F_1$ 分数 | 较原始数据性能提升/% |
| 原始数据 | 0.646 4 | — | 0.552 0 | — | 0.563 2 | — |
| RF 算法 | 0.683 0 | 5.66 | 0.602 9 | 9.22 | 0.619 0 | 9.91 |
| KNN 算法 | 0.656 9 | 1.62 | 0.554 5 | 0.45 | 0.566 4 | 0.57 |
| EM 算法 | 0.654 3 | 1.22 | 0.553 8 | 0.33 | 0.565 5 | 0.41 |

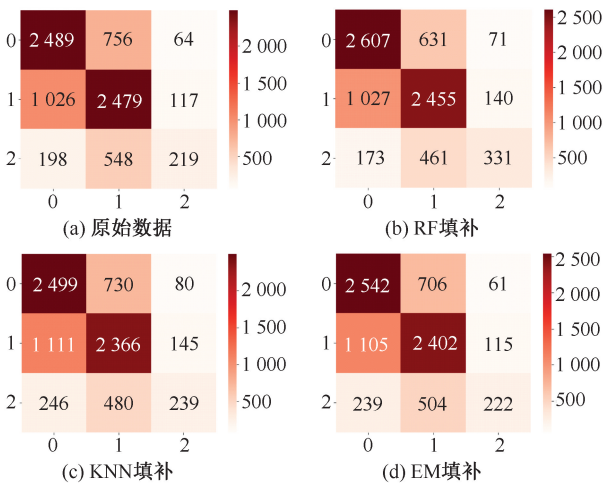


图 3　不同算法填补后模型混淆矩阵

Fig. 3　Matrix diagram of model confusion filled in by different algorithms

## 3　Apriori 算法事故致因分析

在完成数据填补的基础上,进一步建立事故致因挖掘模型,以事故严重程度为因变量,其余除事故标签以外的变量作为自变量进行建模分析。

### 3.1　Apriori 关联规则挖掘算法

传统的统计分析难以发现事故内部变量之间的潜在特征[20],因此,引入 Apriori 关联规则算法,通过深入挖掘内部各要素,发现内部各要素之间的潜在特征。Apriori 关联规则算法是最为经典的挖掘大型数据库中关联规则的算法,旨在从大量数据中寻找事物之间的隐含关系。通过迭代扫描出数据库中所有满足最小支持度的频繁项集,再利用生成的频繁项集构造满足最小置信度和提升度的关联规则[21]。事故影响因素间的关联关系由支持度、置信度和提升度 3 个指标体现[22]。

1) 支持度是指几个相关联的事故数据特征在数据集中同时出现的次数占总数据集的比重。

2) 置信度是指一个事故特征出现后,另一个事故特征也出现的概率,也就是事故特征的条件概率。

3) 提升度是指关联规则 $X \Rightarrow Y$ 的置信度与 $Y$ 的支持度之比。

### 3.2　关联规则结果分析

最小支持度阈值和最小置信度阈值对关联规则输出结果影响极大,若阈值过小,会导致大量低频偶然数据包含其中,且加长运行时间;若阈值过大,则会导致强关联规则数量过少,无法进行有效分析[23]。为对比缺失值填补对关联规则挖掘结果的影响,分别使用剔除缺失值的数据集和经过 RF 算法填补缺失值的数据集,初步将最小支持度设置为 0.01、最小置信度设置为 0.6,生成关联规则支持度置信度散点图,如图 4 所示,颜色的深浅代表提升度的高低。

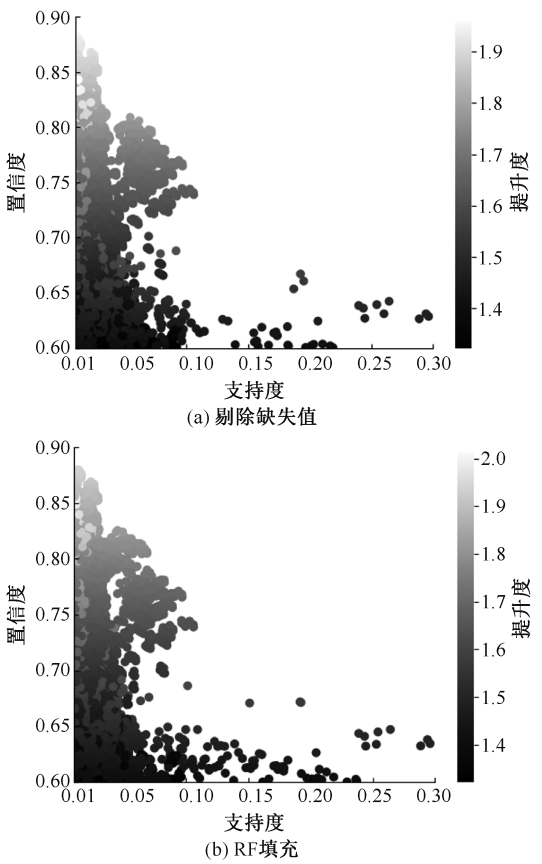


图 4　支持度置信度散点对比

Fig. 4　Scatter comparison chart of support and confidence

从图 4 可以看出,在最小支持度和最小置信度阈值相同的情况下,通过 RF 算法进行缺失值填补的数据提升度高于剔除缺失值的数据,说明经 RF 算法填补的数据在挖掘结果上效果更优。

提升度的高低反映关联规则中前项与后项的相关性,提升度>1 且越高表明正相关性越高,提升度<1 且越低表明负相关性越高,提升度=1 表明没有相关性。

结合图 4b,将最小置信度修改为 0.75,可筛选出提升度>1.7 的关联规则,高提升度关联规则更能表现出事故属性与严重程度之间的强关联性。

为更加清晰直观地体现出导致不同严重程度的影响因素,分别从财产损失事故、受伤事故、死亡事故 3 个严重程度属性方面进行挖掘。

1) 财产损失事故关联规则挖掘。筛选后关联规则后项为财产损失事故的规则有 845 条,财产损失事故关联规则挖掘结果见表 5。由表 5 可知:客车在撞击移动物体时,使用安全设备或在直线道路上行驶时,发生财产损失事故的概率将扩大 1.94 倍;撞击点为车辆后方时,易造成财产损失事故。可能由于客车在高速公路上被要求的限速值较低,加之安全设备的使用对其严重程度有一定的降低作用。

**表 5　财产损失事故关联规则挖掘结果(部分)**

**Table 5　Mining results of association rules for property loss accidents (partial)**

| 前项 | 后项 | 支持度 | 置信度 | 提升度 |
|---|---|---|---|---|
| {直线道路,撞击移动物,客车} | {严重程度-财产损失事故} | 0.02 | 0.83 | 1.94 |
| {安全设备使用,撞击移动物,客车} | {严重程度-财产损失事故} | 0.02 | 0.83 | 1.94 |
| {小型汽车,男性,撞击车辆右后方} | {严重程度-财产损失事故} | 0.02 | 0.81 | 1.88 |
| {白天,男性,撞击车辆右后方} | {严重程度-财产损失事故} | 0.01 | 0.79 | 1.85 |
| {使用安全设备,平坦道路,客车} | {严重程度-财产损失事故} | 0.01 | 0.78 | 1.83 |
| {工作,撞击车辆左后方,男性} | {严重程度-财产损失事故} | 0.01 | 0.78 | 1.81 |

2) 受伤事故关联规则挖掘。筛选后关联规则后项为受伤事故的规则有 1 955 条,受伤事故关联规则挖掘结果见表 6。从表 6 可以看出,在高速公路行驶时,摩托车更易引发受伤事故。其中,驾驶摩托车在限速 90 km/h 的 2 车道之间行驶时,其发生受伤事故的概率将扩大 1.93 倍。

**表 6　受伤事故关联规则挖掘结果(部分)**

**Table 6　Mining results of association rules for injury accidents (partial)**

| 前项 | 后项 | 支持度 | 置信度 | 提升度 |
|---|---|---|---|---|
| {摩托车,限速 90 km/h,2 车道之间行驶} | {严重程度-受伤事故} | 0.02 | 0.87 | 1.93 |
| {使用安全设备,摩托车,2 车道之间行驶} | {严重程度-受伤事故} | 0.03 | 0.86 | 1.90 |
| {工作,摩托车,路表潮湿} | {严重程度-受伤事故} | 0.01 | 0.86 | 1.89 |
| {白天,摩托车,路表潮湿} | {严重程度-受伤事故} | 0.01 | 0.84 | 1.85 |
| {摩托车,隧道} | {严重程度-受伤事故} | 0.01 | 0.82 | 1.80 |
| {摩托车,限速 90 km/h,30~40 岁} | {严重程度-受伤事故} | 0.01 | 0.82 | 1.80 |

3) 死亡事故关联规则挖掘。由于严重程度为死亡事故的数据样本占总样本的比例较少,按所设置的支持度和置信度未能挖掘出后项为死亡事故的关联规则。经过多次试验后,指定最小支持度为 0.02,最小置信度为 0.1,得到 1 770 条关联规则。死亡事故关联规则挖掘结果见表 7。从表 7 可以看出,在限速 130 km/h 的高速公路行驶时,撞击移动物体更易造成死亡事故的发生。其中,在限速 130 km/h 的道路上行驶时,未使用安全设备的情况撞击移动物体,其发生死亡事故的概率将扩大 2.26 倍。

**表 7　死亡事故关联规则挖掘结果(部分)**

**Table 7　Mining results of association rules for fatal accidents (partial)**

| 前项 | 后项 | 支持度 | 置信度 | 提升度 |
|---|---|---|---|---|
| {撞击移动物,限速 130 km/h,未使用安全设备} | {严重程度-死亡事故} | 0.02 | 0.43 | 3.64 |
| {撞击移动物,限速 130 km/h,路表干燥} | {严重程度-死亡事故} | 0.02 | 0.42 | 3.57 |
| {撞击移动物,白天,直线道路} | {严重程度-死亡事故} | 0.02 | 0.32 | 2.67 |
| {撞击移动物,道路平坦,未使用安全设备} | {严重程度-死亡事故} | 0.04 | 0.27 | 2.26 |
| {男性,未使用安全设备,车道总数 4} | {严重程度-死亡事故} | 0.03 | 0.25 | 2.10 |
| {限速 130 km/h,路表平坦,车道总数 4} | {严重程度-死亡事故} | 0.03 | 0.24 | 2.04 |

综上所述,为降低高速公路事故的发生率以及严重程度等级,提出以下建议:①针对客车在撞击移动物体时,易引发财产损失事故,应提醒客车司机谨防疲劳驾驶,防止事故的发生;②针对摩托车在限速 90 km/h 的 2 车道之间行驶时,其发生受伤事故的提升度最高问题,应加强对摩托车在高速公路行驶的监管,对于摩托车可以行驶的路段,设置摩托车专用车道,禁止摩托车长时间占用其他车道行驶;控制摩托车限速值,防止摩托车超速行驶;加强安全设备使用情况的监管和处罚力度。

## 4　结　论

1) 采用 RF 算法、统计学中 EM 算法以及机器学习中的 KNN 算法填补事故数据中的缺失值,对比填补前后的变量方差变化差异,在 16 个含有缺失值的自变量中,有 15 个自变量经 RF 算法填补后的结果最优。

2) 采用 XGboost 模型进行严重程度预测,对比不同算法填补数据的准确率、召回率、$F_1$ 分数可知:RF 填补训练效果相较于其他算法填补训练效果更优,对比原始数据训练模型,其准确率提高 5.66%,召回率提高 9.22%,$F_1$ 分数提高 9.91%。

3) 客车易引发财产损失事故,原因可能由于客车在高速公路行驶时,容易疲劳驾驶,导致事故的发生;摩托车在限速较低的路段易引发受伤事故,在限速较高的路段易引发死亡事故;安全设备的使用情况对事故严重程度等级有较大关系。

## 参 考 文 献


[1] 国家统计局. 中国统计年鉴[M]. 北京:中国统计出版社,2023:515.

[2] 吕能超,王玉刚,周颖,等. 道路交通安全分析与评价方法综述[J]. 中国公路学报,2023,36(4):183-201.
LYU Nengchao,WANG Yugang,ZHOU Ying, et al. Review of road traffic safety analysis and evaluation methods[J]. China Journal of Highway and Transport, 2023,36(4):183-201.

[3] 杨洋,袁振洲,陈治,等. 基于改进系统工程决策理论的高速公路交通安全评价研究[J]. 北京交通大学学报,2022,46(3):34-48.
YANG Yang,YUAN Zhenzhou,CHEN Zhi, et al. Study on freeway traffic safety evaluation based on improved system engineering decision theory[J]. Journal of Beijing Jiaotong University,2022,46(3):34-48.

[4] DEB R,LIEW W A. Missing value imputation for the analysis of incomplete traffic accident data[J]. Information Sciences, 2016,339:274-289.

[5] QI Hang,ZHAO Xiaohua,YAO Ying,et al. BGCP-based traffic data imputation and accident detection applications for the national trunk highway[J]. Accident Analysis and Prevention,2023,186:DOI:10.1016/J.AAP.2023.107051.

[6] 吴郁,张金奋,范存龙,等. 基于随机森林的船舶碰撞事故缺失数据插补[J]. 武汉理工大学学报:交通科学与工程版,2019,43(6):1 120-1 124.
WU Yu,ZHANG Jinfen,FAN Cunlong,et al. Imputation of missing values for ship collision accident data based on random forest[J]. Journal of Wuhan University of Technology:Transportation Science &,Engineering,2019,43(6):1 120-1 124.

[7] 陆嘉铭,奚增辉,瞿海妮,等. 电力量测数据缺失补齐方法研究与实践[J]. 电力大数据,2023,26(7): 40-49.
LU Jiaming, XI Zenghui, QU Haini,et al. Research and practice on power measurement data missing value Imputation methods[J]. Power Systems and Big Data,2023,26(7): 40-49.

[8] 谢翘楚,姚毅. 电网历史数据缺失及补录研究[J]. 四川理工学院学报:自然科学版,2017,30(2):21-25.
XIE Qiaochu,YAO Yi. Research on the data missing and data completion of power grid[J]. Journal of Sichuan University of Science & Engineering :Natural Science Edition,2017,30(2):21-25.

[9] 梅玉杰,李勇,周王峰,等. 基于机器学习的配电网异常缺失数据动态清洗方法[J]. 电力系统保护与控制,2023,51(7):158-169.
MEI Yujie,LI Yong,ZHOU Wangfeng,et al. Dynamic data cleaning method of abnormal and missing data in a distribution network based on machine learning[J]. Power System Protection and Control,2023,51(7):158-169.

[10] 严利鑫,胡鑫辉,刘清梅,等. 道路交通事故严重程度预测及致因分析[J]. 华东交通大学学报,2024,41(5):65-73.
YAN Lixin,HU Xinhui, LIU Qingmei, et al. Road traffic accident severity prediction and causation analysis[J]. Journal of East China Jiaotong University, 2024,41(5):65-73.

[11] 邱文利,杨海峰,张少波,等. 基于改进 Apriori 算法的高速公路交通事故关联分析[J]. 中外公路,2024,44(3):227-235.
QIU Wenli,YANG Haifeng,ZHANG Shaobo, et al. Correlation analysis of highway traffic accidents based on improved Apriori algorithm[J]. Journal of China & Foreign Highway,2024,44(3):227-235.

[12] 牛玥铧,聂百胜. 人格特征对驾驶人驾驶行为的影响研究[J]. 中国安全科学学报,2024,34(2):117-123.
NIU Yuehua,NIE Baisheng. Study on influence of personality traits on drivers´ driving behavior[J]. China Safety Science Journal,2024,34(2):117-123.

[13] 王卓伦,林岩. 中国网约车服务对道路交通事故的影响研究[J]. 中国安全科学学报,2024,34(2):209-216.
WANG Zhuolun,LIN Yan. Research on impact of ride-hailing services on road traffic crashes in China[J]. China Safety Science Journal,2024,34(2):209-216.

[14] 于雷. 基于数据挖掘的道路交通事故分析及预防对策研究[D]. 重庆:重庆交通大学,2022.
YU Lei. Analysis and preventive countermeasures of road traffic accidents based on data-driven[D]. Chongqing: Chongqing Jiaotong University,2022.

[15] 孙维富. 基于数据挖掘的高速公路交通事故分析及预防对策研究[D]. 长春:吉林大学,2018.
SUN Weifu. Study on the freeway traffic accident analysis and countermeasures based on data mining[D]. Changchun: Jilin University,2018.

[16] 法国政府数据开放平台. 法国政府数据开放平台[EB/OL]. [2024-10-14]. https://www.data.gouv.fr/fr/.

[17] BREIMAN L. Random forests[J]. Machine Learning,2001,45(1):5-32.

[18] JADHAV A,PRAMOD D,RAMANATHAN K. Comparison of performance of data imputation methods for numeric dataset[J]. Applied Artificial Intelligence,2019,33(10):913-933.

[19] 周备,张莹,张生瑞,等. 基于生成对抗网络的追尾事故数据填补方法研究[J]. 交通运输系统工程与信息,2024,24(1):132-137,198.
ZHOU Bei,ZHANG Ying,ZHANG Shengrui, et al. Research on rear-end crash data imputation methods based on generative adversarial networks[J]. Journal of Transportation Systems Engineering and Information Technology,2024,24(1):132-137,198.

[20] 牛毅,李振明,樊运晓. 基于数据挖掘的高速公路货车交通事故影响因素关联分析研究[J]. 安全与环境工程,2020,27(4):180-188.
NIU Yi, LI Zhenming, FAN Yunxiao. Correlation analysis of influencing factors of truck traffic accidents on expressways[J]. Safety and Environmental Engineering, 2020,27(4):180-188.

[21] FAYYAD U, PIATETSKY-SHAPIRO G, SMYTH P. From data mining to knowledge discovery in databases[J]. AI Magazine, 1996, 17(3): 37-54.

[22] 何丽瑶. 基于 Apriori 及 XGBoost 算法的道路交通事故分析研究[D]. 苏州:苏州大学,2020.
HE Liyao. Analysis and research of road traffic accidents base on Apriori and XGBoost algorithms[D]. Suzhou :Soochow University,2020.

[23] 黄锦,王梓豪,陈曾惠,等. 基于 Apriori 关联算法的城市综合体停车需求影响因素关联分析[J]. 福建交通科技,2024(3):81-86.
HUANG Jin,WANG Zihao,CHEN Zenghui, et al. Correlation analysis of factors affecting parking demand in urban complexes based on Apriori correlation algorithm[J]. Fujian Transportation Science & Technology,2024(3):81-86.

**作者简介：** 薛 乐 (1997—),男,山东济宁人,硕士研究生,主要研究方向为道路交通安全、事故致因分析等。E-mail:xuele1116@163.com。