[18] 罗石贵,周伟.路段交通冲突技术研究[J].公路交通科技,2001,18(1):65-68.
[19] 高熙.机非混行车道最小宽度的研究探讨[J].科技信息,2013(16),387-388.
[20] 徐程.路段混合自行车交通运行特性与风险评估[D].长春:吉林大学,2016.
[21] Liu P, Marker S. Evaluation of contributory factors' effects on bicycle-car crash risk at signalized intersections [J]. Journal of Transportation Safety & Security, 2020, 12(1): 82-93.
[22] Kroll B. J, Ramey M. R. Effects of bike lanes on driver and bicyclist behavior [J]. Transportation engineering journal of the American Society of Civil Engineers. 1977, 103(2):243-256.

机动车交通事故预测模型评估与特征分析

——随机森林和 XGBoost 的比较

林 涛[1] 邵海鹏*[1,2] 方瑞韬[1]
(1.长安大学运输工程学院;2.生态安全屏障区交通网设施管控及循环修复技术交通运输行业重点实验室)

摘 要 交通安全问题一直被人们高度关注,交通安全机构依靠预测模型进行事故预测,但是现有模型的性能评估存在不足。本文基于随机森林和 XGBoost 算法分别进行了模型预测,通过八个预测性能指标同时进行模型的评估比较。此外,利用 SHAP 解释器对 XGBoost 模型进行解释。结果表明,准确度、敏感度、特异度、FPR 指标、精确度、NPV 指标、F1 分数和马修斯相关系数能够表现出模型在不同类别的预测性能,同时避免了模型过度拟合。随机森林模型和 XGBoost 模型在大部分测量指标的表现几乎都优于其他常用的数据挖掘模型。随机森林模型和 XGBoost 模型显著地提高了预测准确度,而没有提供额外的错误预测。与常用的机器学习算法相比,随机森林模型和 XGBoost 模型对交通事故不平衡数据的分析能力更佳。利用 SHAP 对模型进行特征依赖分析,为交通管理者提供合理的参考意见。

关键词 交通工程 交通安全 随机森林 XGBoost 预测准确度 SHAP 解释器

0 引言

交通事故是世界各国关注的一个交通安全问题。全球道路安全状况报告(Global Status Report on Road Safety)显示,随着大多数国家高速公路和机动车辆数量的迅速增加,世界上的事故总数大幅增加。美国国家公路交通安全管理局(NHTSA)的报告显示,2021 年上半年约有 2 万多人死于车祸,比 2020 年增加 18.4%。中国统计年鉴显示,2020 年我国道路交通事故万车死亡人数下降至 1.66 人,尽管近年万车死亡人数在逐年稳步下降,但是与发达国家相比依旧较高。跟据世界卫生组织报告,在世界范围内,每年有上百万人死于交通事故。如何提前预测交通事故,防止交通事故的发生以减少事故人员伤亡,一直是各国交通研究者所关注的问题。

机动车交通事故预测建模多年来一直是一个重要的安全研究课题。预测模型利用历史数据和数据统计,以此分析事故成因和预测事故发生的可能性,交通安全规划者依靠交通事故预测模型来分配未来几年的安全改进预算。因此,模型预测的准确性至关重要,尤其是在预测事故发生与事故严重程度方面。

非参数数据挖掘以及机器学习方法近年来广受欢迎,并以此建立预测模型并得到较好的预测结果。常见的方法包括:神经网络[1-2]、聚类分析[3-5]、分类和回归树分析[6-7]。大部分学者建议使用这种方法研究交通事故。但是,以往文献往往将重点注重于模型预测的准确率,较少研究者关注模型预测的综合指标,尤其是对于提高安全性至关重要的虚报率和严重程度的错误估计。因为

基金项目:国家重点研发计划(2019YFB1600300);中央高校基本科研业务费专项资金(300102219210,300102210201)。

大量的错误预测以及事故严重程度的高估或低估,不仅会浪费大量的资金和时间,还会使最需要重视最需要安全改进的地方发生误导。

随机森林(Random Forest,RF)方法是用于回归和分类树的集成方法。随机森林模型的一个已知优点是,它不仅可以限制过拟合,而且不会显著降低预测准确度。此外,随机森林是一个包含多棵决策树的分类模型,由于其结合了 Bootstrap 采样和随机属性选择的优点,在分类问题中取得了良好效果。随机森林不仅可以用于二分类,还可以用于多分类问题,其泛化能力好,一般不会出现过拟合,在有偏差的数据上,随机森林也能够获得较好的效果,随机森林算法具有更好的分类精度,且不会产生过拟合问题[8]。Dogru[9]利用随机森林进行交通事故进行分类预测,结果表明 RF 比其他监督机器学习(如 SVM、ANN 等)具有更佳的预测效果,模型评估指标为准确率、敏感度和特异度。Atumo[10]对密歇根州国际公路的交通事故热点识别和预测,同样验证了随机森林具有更好的预测性能,模型评估指标为敏感度和精确度和 F1 分数。Yassin[11]在随机森林的基础上,结合 K-means 聚类进行交通事故严重性预测,结果表明其模型预测准确度高达 99.86%,模型评估指标为准确度、敏感度、特异度、精确度和 F1 分数。

提升决策树是一种广泛而有效的集成学习方法,而 XgBoost 作为一种新型提升决策树算法,于 2016 年被提出,它能够自动利用 CPU 的多线程进行并行运算,可在缩短时间的同时提高精度。Meng[12]利用 XGBoost 方法,对多种数据源预测了事故的发生和持续时间,模型评估指标为准确率、敏感度和精确度。此外,Hamilton[13]和 Schlögl[14]表明,XGBoost 在预测事故可能性方面的表现优于其他几种机器学习技术,包括逻辑回归、贝叶斯正则化神经网络、SVM 和深度神经网络,Hamilton 的模型评估指标为准确率、灵敏度和特异度,Schlögl 的模型评估指标为准确率、灵敏度、特异度和 FPR 指标。Parsa[15]利用 XGBoost 对芝加哥高速公路交通事故进行事故预测,主要评估指标为准确率和灵敏度,结果表明 XGBoost 具有优异的预测能力,其预测准确率高达 99%。此外,XGBoost 还被用于预测交通事故的严重性,并取得了较好的效果,尤其是在使用空间数据时[16]。可以看出,随机森林模型和 XGBoost 在交通事故预测方面均能表现出良好的性能,但是以往研究对预测模型的全面评估还较少。

非参数数据挖掘及机器学习模型的可解释性较差,一般被认为是黑箱模型,无法了解样本特征值是如何影响最终的预测结果。近年来,一些研究已经开始利用 SHAP 解释器(Shapley Additive exPlanation)对模型进行进一步地分析。SHAP 最初是由 Shapley[17]提出,它基于博弈论,它提供了一个可以衡量模型中特征重要性的方法。2017 年,Lundberg 和 Lee[18]用 Python 开发了一个实用程序包,能够计算不同技术的 SHAP,包括 LightGBM、GBoost、CatBoost、XGBoost 模型。在交通安全方面,Mihaita[19]使用 SHAP 分析不同特征对事故持续时间的影响。Parsa[15]使用 SHAP 对基于 XGBoost 的交通事故检测模型特征进行分析和解释。

综上所述,本文研究的目的是评估随机森林模型和 XGBoost 模型等常用的数据挖掘方法在预测机动车交通事故中的应用,计算模型预测的准确度、敏感度、特异度、FPR 指标、精确度、NPV 指标、F1 分数和马修斯相关系数,并分析比较尤其是事故严重程度预测中容易出现虚报的指标,可为帮助相关部门进行安全优化改进从而减少不必要的资金支出。在本文中,采用了六种常用的数据挖掘方法,对决策树(DT)、随机森林(RF)、XGboost、支持向量机(SVM)、朴素贝叶斯(NB)和逻辑回归(Logistic)进行了比较,并评估各预测模型的全面性能。最后,利用 SHAP 解释器对 XGBoost 模型的预测结果进行进一步解释,在提高该模型的可解释性的同时,为交通安全管理者提供合理的意见。

1 数据

1.1 数据预处理

本研究中使用的数据为 2015—2020 年陕西省某市所发生的所有机动车交通伤亡事故。剔除含有缺失值的数据后,最后共保留 6977 条机动车事故数据作为研究对象,期间,有 4998 起伤人事故和 1979 起死亡事故。关于因变量的设置,借鉴国外对事故发生和严重程度的研究[20-21],在本文采用了二分类方法,将事故严重程度分为 2 类:伤人事故和死亡事故。

数据集分为两部分:随机选择 60% 作为训练数据集的观测值,其余 40% 的观测值作为测试数

据集。训练数据集中的样本数为4186,测试数据集中的样本数为2791。数据集中共有32个变量。参考以往文献[22]并综合考虑各方面影响,最终使用26个自变量来构建随机森林模型及XGboost模型,具体如表1所示。

表中仅有年龄、驾龄、事故发生时间、涉事机动车数量是连续变量,而其他的变量均为分类变量。本研究使用分类树建立随机森林模型和XGboost模型,并运用Python3.7进行建模和后续的数据分析。

自变量分类表 表1

变量名	分类	赋值	变量名	分类	赋值
事故时间	无	连续变量	天气	晴	1
事故形态	侧翻	1		阴	2
	成员跌落或抛出	2		雨	3
	刮撞行人	3		雾	4
	滚翻	4		雪	5
	碾压行人	5		其他	6
	碰撞后碾压行人	6	能见度	200m以上	1
	碰撞静止车辆	7		100~200m	2
	碰撞运动车辆	8		50~100m	3
	其他车辆间事故	9		50m以下	4
	其他车辆与行人事故	10	照明条件	白天	1
	其他事故	11		黎明、黄昏	2
碰撞程度	无	0		夜间有路灯	3
	其他	1		夜间无路灯	4
	同向剐蹭、护栏	2	肇事者性别	男	0
	对向剐蹭、护栏	3		女	1
	其他角度碰撞、侧面碰撞角度不确定	4	年龄	其他	连续变量
	侧面碰撞同向	5	户口性质	非农业	0
	侧面碰撞对向	6		农业	1
	追尾碰撞	7	交通方式	驾(驶)非机动车	1
	侧面碰撞直角	8		驾摩托车	2
	正面碰撞	9		驾驶其他机动车	3
道路类型	高速公路	1		驾驶汽车	4
	一级道路	2	驾龄	无	连续变量
	城市快速路	3	血液酒精含量	无	0
	二级道路	4		0~19	1
	一般城市道路	5		20~79	2
	三级道路	6		80以上	3
	四级道路	7	安全带、头盔使用情况	无	0
	其他	8		有	1
道路线型	平直	1	地形	平原	1
	一般坡	2		丘陵	2
	一般弯	3		山区	3
	一般弯坡	4	涉事机动车数量	无	连续变量
	陡坡	5	事故原因	其他违法	1
	连续陡坡	6		非违法过错	2
	急弯	7		非机动车违法	3
	一般弯陡坡	8		机动车违法	4
	一般坡急弯	9	事故道路横断面位置	其他	1
	急弯陡坡	10		人行道、人行横道	2

续上表

变量名	分　类	赋　值	变量名	分　类	赋　值
路面状况	路面完好	1	事故道路横断面位置	非机动车道	3
	凹凸	2		机非混合道	4
	路障	3		机动车道	5
	施工	4	道路物理隔离	无隔离	0
	塌陷	5		中心隔离与机非隔离	1
	其他	6		机非隔离	2
路面结构	沥青	1		中心隔离	3
	水泥	2	中央隔离措施	无	0
	沙石	3		波形护栏	1
	土路	4		柔性护栏	2
	其他	5		混凝土护栏	3
路表情况	干燥	1		金属护栏	4
	潮湿	2		绿化带	5
	泥泞	3		活动护栏	6
	油污	4		隔离墩(柱)	7
	积水	5	道路安全隐患督办等级	无	0
	漫水	6		县级	1
	冰雪	7		市级	2
	其他	8		省级	3
路侧防护设施	无	0	道路安全属性	正常路段	1
	波形护栏	1		已经治理但仍存在隐患道路	2
	柔性护栏	2		正在治理隐患路段	3
	混凝土护栏	3		已排查尚未治理隐患路段	4
	金属护栏	4		尚未排查隐患路段	5
	绿化带、行道树	5	交通控制方式	无控制	0
	活动护栏	6		标志标线	1
	护栏墩(柱)	7		信号灯、民警指挥	2
	其他	8		其他	3

1.2　不平衡数据处理

某一类数据稀少的数据集被定义为不平衡数据[23]。交通事故数据就是典型的不平衡数据,因为在交通事故中无人伤亡的数量是远高于伤亡事故的,而伤人事故又高于死亡事故,在其他事故领域中,也表现出同样的现象。当使用传统方式预测时,往往偏向对有充足数据量的类别进行预测。这会导致对数据量较大的类别表现出良好的预测性能,但对数据量稀少的类别预测能力相对较差,甚者无法预测罕见事件,这可以在以前的研究[24-26]中得到类似的结果。处理不平衡数据有几种常用的技术,分别是数据加权、过采样和欠采样。以往文献中,过采样容易导致模型过拟合,为了解决过拟合问题,文献常使用 SMOTE 法[27],即利用少数类的每个数据生成新的数据集,使少数类数据量与多数类数量持平,但该方法适合连续变量的数据集,与 SMOTE 相近的还有 ADASYN 法。而欠采样技术往往会大致大量数据的丢失,从而导致模型的精度大幅下降。

指定先验概率通常可为模型中的不平衡数据获得更好的预测结果,一般来说,先验概率由每个类别的频率来表示[28]。当增加不平衡数据的先验概率时,也增加后验概率,从而改变不平衡数据的分类边界,使得更多的样本被分类到该类别中。

考虑到数据集中的本身特点,因此,本文的解决办法是对不平衡数据增加权重,在模型训练时调整损失函数的惩罚系数,使模型能平等对待多数的伤人事故与少数的死亡事故这两个不平衡类别。

2 模型建立与评价

2.1 随机森林

随机森林模型是一种用于分类和回归的集成方法。由随机森林模型做出的决策是基于由许多决策树做出的决策集合。在本研究中,决策树拆分的算法选择 ID3 信息熵算法。ID3 信息熵算法测量信息增益,以确定是否应该选择属性变量作为拆分器以及节点是否应该进一步拆分。假设一个变量 S 有 c 个不同的值,S 的熵 $E(S)$ 的计算公式见式(1):

$$E(S) = \sum_{i=1}^{C} -p_i \log_2 p_i \tag{1}$$

式中:p_i——取某个确定值的概率。

如果变量在某个属性上被划分为子集 $s_1, s_2, \cdots, S_c$,期望熵(EH)可以衡量变量 S 分裂后计算结果数学期望的不确定性,计算公式见式(2):

$$\mathrm{EH} = \sum_{i=1}^{C} \frac{a_i}{a} \times (-p_i \log_2 p_i) \tag{2}$$

式中:a_i——每个子集 $s_1, s_2, \cdots, s_c$ 中的样本总数;

a——父节点 S 中的样本总数。

信息增益 IG(S,A)是将属性 A 分割成 S 个集合之前到之后熵差的度量[29]。换句话说,在属性 A 上拆分集合 S 后,S 中的不确定性减少的值。信息增益 IG(S,A)按等式(3)计算:

$$\mathrm{IG}(S,A) = E(S) - \mathrm{EH} \tag{3}$$

信息增益为零的节点被认为是不需要进一步分裂的终端节点。决策树总是对训练数据进行完美分类,随着决策树不断分裂数据,树变得越来越大,对训练数据集的准确率会有所提高,但其应用于测试数据集时,所做出的预测结果不理想。决策树倾向于过度拟合训练数据,这在应用完整数据集时可能会产生较差的结果。RF 的最大优势之一是它可以在不显著降低预测准确度的情况下限制过度拟合。

RF 模型主要基于两种方法构建:袋装法和随机子空间。袋装法是使用 Bootstrapping 随机重采样得到训练样本,然后对每个 Bootstrapping 样本进行学习的过程。Bootstrapping 是一种带替换的统计随机重采样方法,用于处理不平衡数据。随机子空间随机选择属性变量来分割 RF 模型中每个决策树的节点。而不被 Bootstrapping 采用的样本被称为袋外样本,常被用于模型选择和评估。具有较低袋外 OOB 误差率的 RF 模型是最优的结果,因为它可以提供较高的预测准确度。通常,包含更多树的 RF 模型具有较低的袋外误差率,也不会导致过度拟合,但在某些情况下,过多的决策树会存在缺点[30]。

在本研究中,使用袋装法的 RF 模型作为预测工具。此外,还进行了完整的预测准确度分析,以更好地理解所提出的 RF 方法的预测性能。对于模型训练,选取 60% 数据进行训练,40% 进行模型测试,本文使用了 Scikit-learn 库中的随机森林包。通过五折交叉验证选择最优超参数值:树的个数 n estimators:200;最大深度 max depth:9;使用袋装法 bootstrap:True;分裂算法 criterion:entropy 信息熵 ID3 算法;不平衡数据加权 class weight:balanced;叶子节点最少样例数 min samples leaf:2;分裂内部节点所需最少样例数 min samples split:2。

2.2 XGBoost

XGBoost 是一种高效的梯度提升决策树。决策树的结构类似于具有根节点(最顶端的节点)、内部节点和叶节点(末端节点)的树,决策树算法常使用简单的规则,从根节点开始分支,经过内部节点,最后在叶子中结束。相比之下,梯度增强决策树是一种集成学习技术,它使用一系列决策树,其中每个决策树从先前的树中学习,并影响下一棵树来改进模型并构建强大的学习者。详细可参考陈天奇和 Guestrin 的相关研究[31]。

给定一个具有 n 个样本的数据集,存在独立变量 x_i,并且这些变量中的每一个都具有 m 个特征,因此 $x_i \in R^m$。对于每一个变量,都有相应的因变量 y_i,$y_i \in R$。树集合模型使用自变量和 k 个函数预测因变量 $\overset{\$}{y}_i$:

$$\overset{\$}{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{4}$$

式中:f_k——一个具有叶分值的独立的树结构;

F——树的空间。

目标是最小化公式(5):

$$L(\phi) = \sum_i l(\overset{u}{y}_i, y_i) + \sum_k \Omega(f_k) \tag{5}$$

式中:l——损失函数;

Ω——惩罚模型复杂性的系数,并且:

$$\Omega(f) = \Upsilon T + \frac{1}{2}\lambda \parallel \omega_i \parallel^2 \tag{6}$$

式中：T——叶节点个数；

ω_i——第 i 个叶节点的得分。

通过解方程式(4)～式(6)，ω_j^* 的最佳值和相应值为：

$$\omega_j^* = -\frac{\sum_{i\in Ij} \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1})}{\sum_{i\in Ij} \partial^2_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) + \lambda} \tag{7}$$

$$L_{(q)}^{\sim t} = -\frac{1}{2}\sum_{j=1}^{T} \frac{(\sum_{i\in Ij} \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}))^2}{\sum_{i\in Ij} \partial^2_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) + \lambda} + \lambda T \tag{8}$$

因为在实践中，很难为所有可能的树结构计算该值，所以使用了以下公式：

$$L_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in IL} \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}))^2}{\sum_{i\in IL} \partial^2_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) + \lambda} + \frac{(\sum_{i\in IR} \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}))^2}{\sum_{i\in IR} \partial^2_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) + \lambda} - \frac{(\sum_{i\in I} \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}))^2}{\sum_{i\in I} \partial^2_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) + \lambda}\right] - \gamma \tag{9}$$

式中：$I = I_L \cup I_R$。作为决策树算法，XGBoost 不受多重共线性的影响。因此，即使两个变量在一个系统中捕捉到相同的现象，两个变量都可以保留下来。

对于模型训练，选取 60% 数据进行训练，40% 进行模型测试，本文使用了 XGBoost 包的分类器和 Scikit-learn 库。五折交叉验证后选择的最优超参数值为：迭代次数 n estimators：200；最大深度 max depth：9；随机抽取样本比例 subsample：0.9；学习率 eta：0.1；不平衡数据加权 scale pos weight：6；子采样参数 colsample bytree 和 colsample bylevel 分别取 0.8 和 0.7；节点分裂最小损失函数下降值 gamma：1.0；alpha 和 lambda 分别是 L1 和 L2 正则化项的权重，均取 1.0。

2.3　模型评估

分类准确度是评估模型的最常使用的标准。混淆矩阵(Confusion Matrix)可以用作二元分类问题，混淆矩阵有四个元素：真正类(True Positives，TP)、真负类(True Negatives，TN)、假正类(False Positives，FP)和假负类(False Negatives，FN)，这四个元素表示观察到的结果和预测的结果是否一致，具体如表 2 所示。TP 和 TN 分别为事故是否发生的正确预测；FP 和 FN 检验结果分别为事故是否发生的错误预测。

混淆矩阵　　表 2

混淆矩阵		实际情况	
		Negative(0)	Positive (1)
预测情况	Negative (0)	TN	FN
	Positive (1)	FP	TP

一般来说，预测准确度按公式(10)计算，公式(10)是文献[30]中最常见的指示预测模型验证的公式，其中 TP、TN、FN、FP 代表上述四种类别中每个类别的预测结果个数。

$$\text{Accurancy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{10}$$

分类准确度方程提供了总体预测准确度测量。然而，准确度并不衡量每个目标类别(发生事故与未发生事故)的预测准确性。由于数据是不平衡数据，大多数观测值属于未发生事故的类别，整体预测准确度主要由模型未发生事故类别的性能来表示。模型在预测发生事故类别的准确度可能被模型在预测总体准确度方面的优异结果所掩盖。因此，引入了敏感度、特异度、精确度、NPV 和 F1 分数，以分析模型对不平衡数据更全面的表现。

敏感度定义为式(11)[30]，它通过描述模型在实际结果为正时模型预测也为正的效果。换句话说，它代表了模型预测正类(TP)与实际总正类(TP 和 FN)的百分比。

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

可以看出，敏感度忽略了假正类(FP)的描述。一个模型可以通过牺牲大量的假正类(FP)预测来增加它的敏感度。如等式(12)中所定义的，特异度(Specificity)评估当实际结果为负时，模型预测也为负类的结果是否良好。假正类的比率 FPR

(False Positive Rate)被称为反向真负类比率[30],如等式(13)中所示。等式(13)总结了当实际结果为负时,预测结果为正的频率。敏感度、特异度和 FPR 都侧重于检测观测指标的覆盖范围,而不是着眼于模型预测性能。对于不平衡的事故数据,占优势的数据将不产生虚报率。因此,即使特异度和 FPR 考虑假警报,它们也不能单独作为性能评估的指标。

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{12}$$

$$\text{FPR} = 1 - \text{Specificity} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{13}$$

为了评估预测模型预测技能,引入了由等式(14)定义的精确度[30],也被称为 PPV(Positive Predictive Value)指标。精确度精确描述了一个模型在预测正类方面的准确程度。它评估 TP 预测相对于模型预测总正类(TP 和 FP)预测的百分比。该指标是对模型预测正类的性能评价。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

在样本中存在不平衡数据的情况下,检查精确度和特异度非常重要。大多数交通事故数据是不平衡的数据,其中有许多事件不发生(类别 0)的情况,而事件发生(类别 1)的情况很少,在本次研究中,类别 0 对应的是伤人事故,而类别 1 对应的是死亡事故,伤亡事故本身就是不平衡数据。如前所述,大量的不发生样本表明类别 0 的模型预测率高,导致大量的 TN,因此特异度对 FN 不敏感,通常产生较高的值。精确度通过对少数类的正确预测来揭示一个模型的真正事件预测能力。

式(15)中描述的 NPV(Negative Predictive Value)评估模型在预测事故为多数量类别 0 的真实表现[32]。它评估 TN 预测相对于模型预测总负类(TN 和 FN)预测的百分比。精确度和 NPV 是两个可以用来评估模型的真实预测性能的度量。因此,敏感度、特异度和 FPR 侧重于检测模型准确预测的覆盖范围,而精确度和 NPV 侧重于评估模型的实际预测性能。

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{15}$$

同时总结精确度和敏感度的综合分数也很重要。由等式(16)定义的 F1 分数计算精确度和敏感度的谐波平均值。它总结了模型的综合准确度,1 表示完美的精确度和敏感度,0 表示最坏的情况。此外,研究指出 F1 分数对数据不平衡也很敏感[33]。

$$\begin{aligned}\text{F1} &= \left(\frac{\text{sesitivity}^{-1} + \text{precison}^{-1}}{2}\right)^{-1} \\ &= 2\,\frac{\text{sesitivity} \times \text{precison}}{\text{sesitivity} + \text{precison}}\end{aligned} \tag{16}$$

在本研究中,马修斯相关系数(Matthews Correlation Coefficient,MCC)被选择为二分类的预测性能的另一个综合衡量指标。MCC 最早由 BW Matthews 引入,由于它对不平衡数据的检测结果更可靠、更客观,同时在一定程度上弥补 F1 分数的不足[34-35],因此在生物医学中被大量使用。MCC 可以看作是二元变量 Pearson 相关性的离散化[34]。MCC 由等式(17)定义。它是一个介于 -1 和 1 之间的值,其中 1 表示完美预测,-1 表示预测和观察完全不一致。

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP}+\text{FP}) \cdot (\text{TP}+\text{FN}) \cdot (\text{TN}+\text{FP}) \cdot (\text{TN}+\text{FN})}} \tag{17}$$

本研究提出了一个完整的预测能力评估,以避免过度乐观的验证和使用上述所有测量的误导性错觉,然而以往文献在敏感度、特异度、精确度、NPV 和 F1 分数很少被同时应用评估,马修斯相关系数在交通事故预测领域的应用也极少。最后,引用以往文献中常用的 ROC 曲线与 AUC 指标进行补充对比。

2.4 模型解释

本研究使用 Lundberg 和 Lee[18] 提出的 SHAP 程序包对 XGBoost 模型进行解释。SHAP 基于博弈论和局部解释来估计每个特征的贡献。假设一个 XGBoost 模型,其中一个组 N(包含 n 个特征)用于预测输出,在 SHAP 中,各特征的贡献 Φ_i 是基于其边际贡献进行分配的[17],详细可参考 Lundberg 和 Lee 的相关研究[18]。Shapely 值通过以下公式(18)确定:

$$\Phi_i = \sum_{S \subseteq N \setminus i} \frac{|S|!(n - |S| - 1)!}{n!} \left[v(S \cup \{i\}) - v(S)\right] \tag{18}$$

式中:n——特征的总数;

S——任何特征 N 的子集;

$v(S)$——S 的贡献值。

二元特征 g 的线性函数根据以下可加性特征属性方法定义[31]:

$$g(z') = \Phi_0 + \sum_{i=1}^{M} \Phi_i z_i' \tag{19}$$

式中，$z' \in \{0,1\}^M$ 观察到特征时取值为1，否则为0，M 为输入特征的个数。

3　结果与分析

3.1　结果对比

RF模型和XGBoost模型预测测试数据集的混淆矩阵如表3所示。表4显示了六个模型的各个指标比较结果。很明显，这两个模型的预测准确度均超过了85%，达到较好的预测分类结果，二者的特异度明显高于预测的敏感度，虽然模型能够表现出良好的预测准确度和特异度，但不意味着它们具有很强的预测能力。这种现象的根本原因是机动车交通事故数据是不平衡数据集。因此需要使用其他指标进行综合对比。

RF模型和XGBoost模型的混淆矩阵　表3

RF模型	实际伤人事故(0)	实际死亡事故(1)
预测伤人事故(0)	2172(TN)	225(FN)
预测死亡事故(1)	146(FP)	248(TP)
XGBoost模型	实际伤人事故(0)	实际死亡事故(1)
预测伤人事故(0)	2208(TN)	168(FN)
预测死亡事故(1)	165(FP)	250(TP)

表4显示，RF模型和XGBoost模型在大部分测量指标的表现优于其他数据挖掘模型。朴素贝叶斯模型虽然NPV指标是所有模型中最高的，然而，该模型显著牺牲了模型的精确度，仅为0.3717，比其他模型低了至少20%，也就是说，该模型假正类的数据很高，意味着模型会高估机动车事故严重程度，误导事故结果。决策树模型的精确度是所有模型中最高的，但该模型显著牺牲了模型的敏感度和NPV指标，即该模型假负类的数据很高，意味着模型会低估机动车事故严重程度，将严重误导事故结果。

RF模型在特异度和精确度方面分别超过XGBoost模型约0.65%与2.70%。但是，XGBoost模型在准确度、敏感度、NPV指标这三个方面分别超过RF模型约1.36%、7.38%、2.32%。也就是说，XGBoost模型在防止假负类(FN)错误方面具有更好的性能；RF模型在防止假正类(FP)错误方面具有更好的性能，二者在数据预测能力上各有优势。RF模型和XGBoost模型在准确度、敏感度、特异度和FPR均优于其他四个模型，在精确度和NPV指标上也具有很好的表现。

六个数据挖掘模型的指标对比　表4

项目	Logistic	Bayes	DT	SVM	RF	XGBoost
Accuracy	0.7621	0.8799	0.7420	0.7832	0.8671	0.8807
Sensitivity	0.3377	0.5816	0.3041	0.3825	0.5243	0.5981
Specificity	0.9238	0.9079	0.9280	0.9136	0.9370	0.9305
FPR	0.0762	0.0921	0.0720	0.0864	0.0630	0.0695
Precision	0.6280	0.3717	0.6421	0.5901	0.6294	0.6024
NPV	0.7854	0.9586	0.7584	0.8198	0.9061	0.9293
F1	0.4392	0.4535	0.4127	0.4641	0.5721	0.6002
MCC	0.3288	0.4021	0.3049	0.3483	0.4971	0.5313

为了确定模型对不平衡类别数据的真实预测能力，应同时考虑评估F1得分，因为F1分数对数据不平衡不仅敏感，而且同时考查了精确度和敏感度。与RF模型相比，XGBoost模型具有更高的F1分数，提高了约2.81%，这表明XGBoost模型在综合预测能力方面具有更好的性能。此外，另一个综合衡量指标马修斯相关系数也应该得到重视。XGBoost模型的MCC最高，最接近与1，说明RF模型和XGBoost模型在不平衡数据中具有了良好的预测能力。在其他模型中，Logistic模型、决策树和支持向量机的MCC相近，说明它们对不平衡数据的预测能力相近；而朴素贝叶斯模型仅次于RF模型和XGBoost模型，说明该模型在不平衡数据中具有不错的预测能力。

综合来看，朴素贝叶斯的综合表现仅次于RF模型和XGBoost模型。Logistic模型具有最高的特异度和较好的精确度；而支持向量机SVM在特异度方面也较高，在F1评分和MCC中与Logistic模型

相近,这表明 Logistic 模型与支持向量机 SVM 在不平衡数据的总体预测能力是相似。值得注意的是,支持向量机 SVM 在 F1 分数中高于朴素贝叶斯模型,但是在 MCC 中却得到了相反的结果,这是因为在 F1 分数中没有反应真负类(TN),从而导致支持向量机 SVM 的 F1 分数产生了虚高的结果。

总之,敏感度、特异度、FPR 指标、精确度、NPV 指标和 F1 分数能够表现不同类别的预测性能。然而,他们倾向于高估不平衡数据集的预测能力。精确度能够表明模型对少数类别的真实预测能力,但在事故分析以及以往的文献中应用较少,同时,NPV 指标和 F1 分数同样很少被同时应用评估。此外,准确度确实受到了敏感度、特异度、FPR、精确度和 NPV 指标的影响,但它容易高估不平衡数据的预测能力。F1 评分可以作为谐波平衡指标,它对数据不平衡很敏感,但是也会产生具有误导性的虚高结果,为此,本研究引入考察了混淆矩阵所有元素综合性更强的马修斯相关系数指标,它在一定程度上弥补 F1 分数的不足。综上所述,XGBoost 模型在不平衡数据的模型预测性能中表现稍微好于 RF 模型。

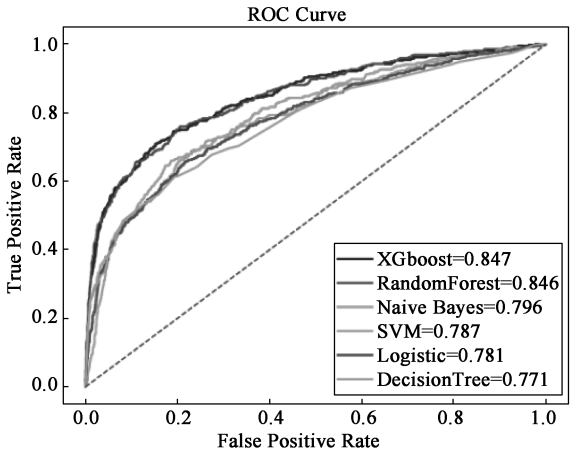

图 1 ROC 曲线与 AUC 指标

六种模型的 ROC 曲线如图 2 所示。图 3 中可以看出,决策树、朴素贝叶斯、Logistic 回归、支持向量机、RF 模型与 XGBoost 模型的 AUC 指标均保持在0.75~0.85 之间。其中,RF 模型和 XGBoost 模型的 AUC 值均接近0.85,分别为0.846 和0.847,说明这两种模型均能达到较好的预测效果,且两者效果极其相近;XGBoost 模型比 RF 模型的 AUC 值高0.001,说明 XGBoost 模型的表现稍微优于 RF 模型。

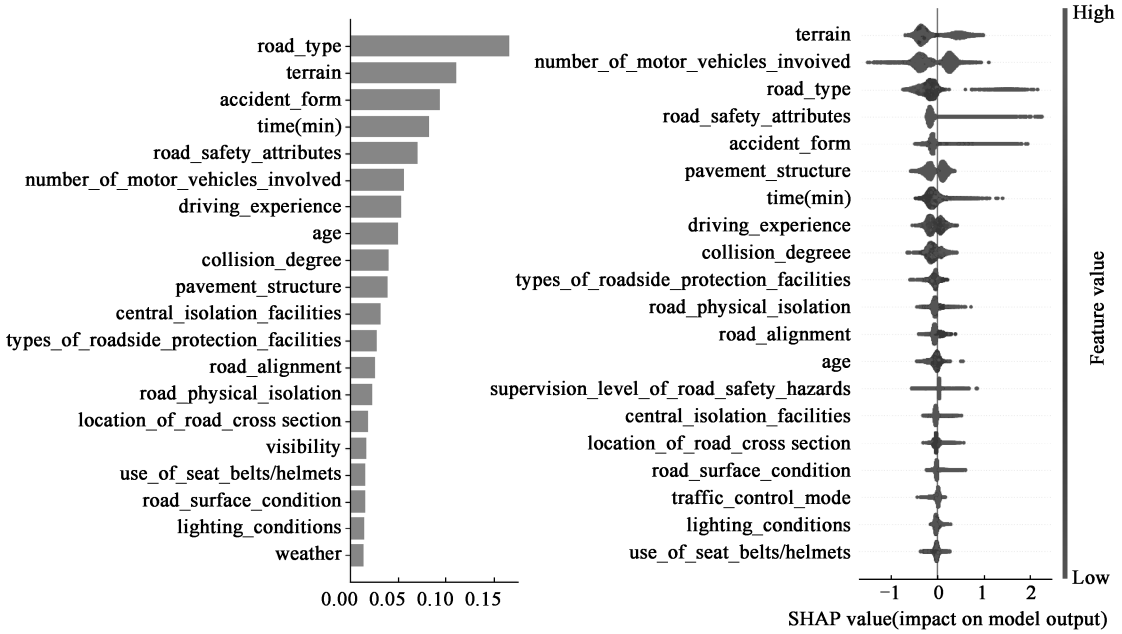

图 2 RF 和 XGBoost 模型的变量重要度排序

将变量重要性分别按照模型平均精度下降(Mean Decrease Accuracy)作排序,然后再提取导致严重的死亡事故影响因素做排序。图 3 左边为 RF 模型中各变量重要度的排序,图 3 右边为 XGBoost 模型利用 SHAP 解释器输出的各变量重要度的排序。结果表明 RF 与 XGBoost 的变量重要性排序结果类似,可以看出两种模型的结果相近,道路类型、道路所处地形、事故时间、涉事机动车数量和道路安全属性是最重要的影响变量,其次,年龄、驾龄、碰撞程度、能见度和佩戴头盔及系

安全带对事故严重程度的影响较大。

3.2　特征依赖分析

在图3a)和图3b)中,在x轴绘制数据特征的值(主要选取了连续变量的特征),y轴为该特征的SHAP值,由于篇幅有限,本研究选取具有代表性的两个特征图进行分析。

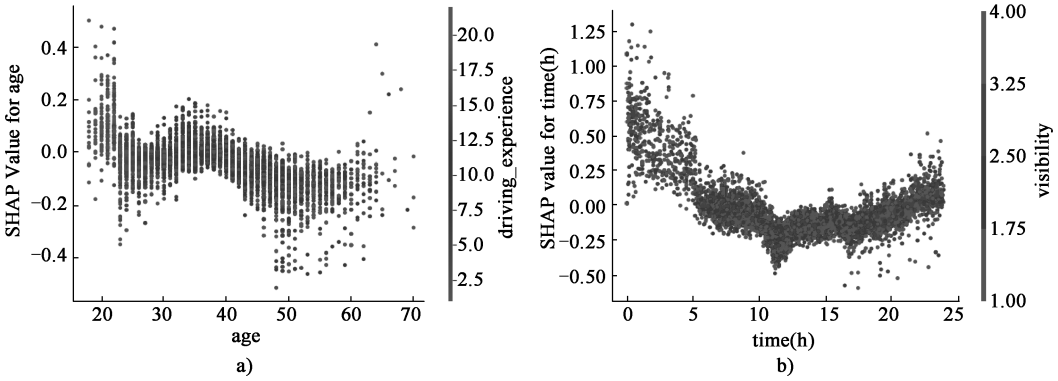

图3　SHAP依赖分析

图3a)中显示了的驾驶人年龄与驾龄对机动车交通事故严重程度的影响。可以看出,驾驶人年龄在35岁以前,驾龄大部分集中在5年以内,年龄超过40岁以后,驾龄普遍较长。有趣的是,年龄在18~23周岁之间,SHAP值普遍较高(大于0),过了23周岁以后,出现了断崖式的下降(小于0),这意味着年轻且无驾驶经验的驾驶员易导致交通事故,这与以往的研究结果相似[36]。然而,年龄在35周岁附近时,SHAP值有所上升,这可能与驾驶人的社会压力有关。年龄在40周岁以上的驾驶人群体的SHAP值普遍低于0,在此区间的驾驶人驾龄普遍在10年以上,说明随着年龄的增加,具有丰富驾驶经验的驾驶人发生交通事故的概率较低,但也存在极少数高龄驾驶人发生严重交通事故的现象,这同样与以往的研究结果相似[37]。

在图3b)中,本文选择事故时间(单位:h)作为特征来确定能见度从高(200m以上取值为1)到低(50m以下取值为4)的影响。可以看出,大部分能见度较高的时间集中在早晨8点至下午6点之间,意味着在早上至下午的时候,一般能见度也较高,是比较符合当地的天气气候变化的。在这个时间段中,SHAP值均低于0,表明良好的能见度能显著降低机动车交通事故的严重程度。在18点至24点时,SHAP值有明显的提升趋势,超过21点后,SHAP值高于0,而且能见度显著降低,点的密度大幅增加,这说明晚上的机动车交通事故的严重程度较高。在凌晨0点至5点之间,虽然点的密度明显比其他时间段稀疏,但是SHAP值显著高于其他时间段,意味着此时间段的机动车交通事故发生次数虽然较少,但是事故一旦发生,事故死亡概率会很高。

结合事故影响因素排序图与特征依赖分析,对于交通部门来说,首先需要对各类型的驾驶员群体开展系统的安全教育培训,尤其针对年龄在20周岁左右的年轻驾驶员群体,规范全体驾驶员的安全意识,同时加强道路设施的完善。对于行政执法部门来说,需要加大执法力度,合理规范设置道路路段行驶速度,以减少因超速行驶或速度差过大导致的道路交通事故发生。对于广大机动车驾驶员来说,自觉并严格遵守交通规则,加强自身交通安全意识,并尽可能减少夜间不必要的机动车驾驶出行。

4　结语

本文提出了通过八个预测性能指标进行数据挖掘模型的评估,利用2015—2020年陕西省某市的机动车交通安全事故数据,建立了六种数据挖掘模型,以此得到混淆矩阵并计算出各预测性能指标。分析结果表明:

(1)准确度、敏感度、特异度、FPR指标、精确度和NPV指标能够显示出模型在不同类别的预测性能。RF模型和XGBoost模型在大部分测量指标的表现优于其他数据挖掘模型。随机森林模型和XGBoost模型显著地提高了预测准确度,而没有提供额外的错误的正向和反向预测。

(2)综合以上六个指标、F1评分、MCC指标和AUC指标,发现XGBoost模型在机动车交通事故不平衡数据的严重程度预测性能中表现稍微比

RF 模型更好。

(3)针对事故伤亡程度的重要度排序和特征依赖分析,提出相关建议以缓解机动车交通事故的严重程度。

(4)传统机器学习模型的可解释性差,本研究利用 SHAP 解释器,使 XGBoost 模型的结果是可解释的。

(5)研究也存在一些不足,选择的变量过多,模型准确度相对不高,需要进一步精简和优化模型;此外,卷积神经网络与前馈神经网络方法可作为新的交通事故预测模型研究方向。

本研究有助于相关部门更有效地改进分配安全预算,降低安全改进成本,并避免不必要的资金支出。

参考文献

[1] Chiou Y C. An artificial neural network-based expert system for the appraisal of two-car crash accidents[J]. Accident Analysis & Prevention, 2006,38(4):777-785.

[2] Zeng Q, Huang H L. A Stable and Optimized Neural Network Model for Crash Injury Severity Prediction [J]. Accident Analysis and Prevention 2014,73:35-81.

[3] Luca M D, Mauro R, Russo F, et al. Before-after Freeway Accident Analysis Using Clustering Algorithms[J]. Social and Behavioral Science 2011,20:723-731.

[4] Luca M D, Mauro R, Russo F, et al. Road Safety Management Using Bayesian and Clustering Analysis [J]. Social and Behavioral Science 2012,54:1260-1269.

[5] Ona J D, Lopezz Z, Mujalli r, et al. Analysis of Traffic Accident on Rural Highway Using Latent Class Clustering and Bayesian Networks [J]. Accident Analysis and Prevention 2013, 51: 1-10.

[6] Magazzu D, Comelli M, Marinoni A. Are Car Drivers Holding a Motorcycle License Less Responsible for Motorcycle-car Crash Occurrence? A Non-parametric Approach[J]. Accident Analysis and Prevention 2006, 38 (3):65-70.

[7] Kashani A T, Rabieyan R, Besharati M M. A Data Mining Approach to Investigate the Factors Influencing the Crash Severity of Motorcycle Pillion Passengers [J]. Journal of Safety Research 2014,51:93-101.

[8] Jahangiri A. Investigating Violation Behavior at Intersections using Intelligent Transportation Systems: A Feasibility Analysis on Vehicle/Bicycle-to-Infrastructure Communications as a Potential Countermeasure[M]. 2015,170.

[9] Dogru N, Subasi A. Traffic accident detection using random forest classifier[C]//2018 15th learning and technology conference (L&T). IEEE,2018:40-45.

[10] Atumo E A, Fang T, JIANG X. Spatial statistics and random forest approaches for traffic crash hot spot identification and prediction [J]. International journal of injury control and safety promotion,2021:1-10.

[11] Yassin SS. Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach [J]. SN Applied Sciences,2020,2(9):1-13.

[12] Meng H L, Wang X H. Expressway crash prediction based on traffic big data. 2018 Int. Conf. Signal Process. Mach. Learn. 1-6.

[13] Mousa S R, Bakhit P R, ISHAK S. An extreme gradient boosting method for identifying the factors contributing to crash/near-crash events: a naturalistic driving study[J]. Canadian Journal of Civil Engineering,2019,46.

[14] Schlögl M, Stütz R, Laaha G, et al. A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset[J]. Accident Analysis & Prevention, 2019:134-149.

[15] Parsa A B, Movahedi A, Taghipour H, et al. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis [J]. Accident Analysis & Prevention,2020,136:105405.

[16] Mokoatle M, Marivate V, Esiefarienrhe M. Predicting Road Traffic Accident Severity using Decision Trees and Time-Series Calendar

Heatmaps[C]. 2019 IEEE Conference on Sustainable Utilization and Development in Engineering and Technologies (CSUDET). IEEE,2019:11-17.

[17] Shapley L S. 17. A value for n-person games [M]. Princeton University Press,2016.

[18] Lundberg S M, Lee S I. A unified approach to interpreting model predictions [C]// Proceedings of the 31st international conference on neural information processing systems. 2017, 4768-4777.

[19] Mihaita A S, Liu Z, Cai C, et al. Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting [J]. arXiv preprint arXiv:1905.12254,2019.

[20] Kumar S. Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India [J]. European Transport Research Review, 2017,9(2):1-10.

[21] Ona J. Analysis of traffic accidents on rural highways using latent class clustering and Bayesian networks[J]. Accident Analysis and Prevention,2013,51(2):1-10.

[22] 李英帅,张旭,王卫杰,等.基于随机森林的电动自行车骑行者事故伤害程度影响因素分析[J].交通运输系统工程与信息,2021,21(01):196-200.

[23] Da Cieslak, Chawla N V. Learning Decision Trees for Unbalanced Data [C]. Machine Learning and Knowledge Discovery in Databases, European Conference, ECML/PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part I. Springer-Verlag,2008.

[24] Chang LY, Chen W C. Data Mining of Tree-based Models to Analyze Freeway Accident Frequency [J]. Journal of Safety Research, 2005,36(3):65-75.

[25] Pande A, Aty Ma, Das A. A Classification Tree Based Modeling Approach for Segment Related Crashes on Multilane Highways[J]. Journal of Safety Research,2010,41:391-398.

[26] Chang L Y, CHIEN J T. Analysis of Driver Injury Severity in Truck-involved Accidents Using a Non-parametric Classification Tree Model[Jl. Safety Science,2013,51:17-22.

[27] Han H, Wang W Y, Mao B H. Borderline-SMOTE:a new over-sampling method in imbalanced data sets learning [C]//International conference on intelligent computing. Springer, Berlin, Heidelberg,2005:878-887.

[28] Zhou X Y, Lu P, Zheng Z J, et al. Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree [J]. Reliability Engineering & System Safety, 2020, 200:106931.

[29] Prajwala T R. A Comparative Study on Decision Tree and Random Forest Using R Tool [J]. International Journal of Advanced Research in Computer and Communication Engineering January,2015.

[30] Rachman A, Chandima R M. Machine Learning Approach for Risk-Based Inspection Screening Assessment[J]. Reliability Engineering & System Safety,2019,185:518-532.

[31] Chen T Q, Guestrin C. Xgboost:a scalable tree boosting system [J]. International Journal of Intelligence Science,2016:785-794.

[32] Zhou X, Lu P, Zheng Z, et al. Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree[J]. Reliability Engineering & System Safety,2020,200:106931.

[33] Matthews B W, Matthews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme[J]. Biochimica et Biophysica Acta,1975,405(2):442-451.

[34] Boughorbel S, Jarray F, El-anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric [J]. PloS one, 2017,12(6):e0177678.

[35] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification

evaluation[J]. BMC genomics,2020,21(1):1-13.
[36] 贾云帆,张李斌,段亚妮,等. 驾驶人员路怒情绪与驾驶风格及攻击行为相关分析[J]. 中国公共卫生,2016,32(10):1373-1377.
[37] 胡江碧,曹新涛. 道路交通事故肇事驾驶员特征分析[J]. 中国公路学报,2009,22(06):106-110.

1980—2020 年中国交通事故空间分布特征研究

应江龙[1,2] 郭建群[1,2] 蒋仲廉[*1] 初 晓[1,2] 余诚强[1,2,3]
(1. 武汉理工大学国家水运安全工程技术研究中心;
2. 武汉理工大学交通与物流工程学院;3. 闽江学院海洋研究院)

摘 要 随着我国交通运输行业的蓬勃发展,交通安全问题得到了更多学者的关注。本文根据 EM-DAT(Emergency Events Database)数据库,以发生在 1980—2020 年间的道路、水路、航空、轨道交通运输重大级以上事故数据为基础,采用重心分析和标准差椭圆统计方法分析交通事故空间特征,研究结果表明:道路交通事故数量占总事故数比例约 49.7%,其人口损失空间分布由东南-西北向逐渐向西扩张,发展为东-西方向分布;水路交通事故人口损失由最初的东北-南分布逐渐逆时针旋转,逐步发展为南-北方向分布;航空交通人口损失保持东南-西北方向分布特征,但东西向分布范围逐渐收缩;轨道交通人口损失由东北-西南方向分布逐渐转变为东南-西北方向,且各方向上的分布范围都大幅收缩。本文研究结果可为交通事故防控及应急资源配置提供参考。

关键词 交通事故 空间分布特征 重心分析 标准差椭圆 EM-DAT 数据库

0 引言

交通运输是国民经济基础性、先导性、战略性产业和重要服务业,降低行业风险、避免发生重大事故一直是行业和学术界关注的重点问题之一。Li 等[1]针对城市内机动车碰撞事故提出了一种基于地理信息系统(Geographic Information System,GIS)的贝叶斯方法,并通过此方法对德克萨斯州哈里斯县五年内的汽车碰撞事故数据开展分析,识别了潜在高碰撞风险的路段。Acharya 等[2]研究了韩国近岸水域水上交通事故空间分布,通过 GIS 直观展示了安全事故的空间分布情况,确定了事故高发区和安全缺陷区。Huang 等[3]基于全球综合船舶信息系统(Global Integrated Shipping Information System,GISIS)中的海上伤亡和事故(Marine Casualties and Incidents,MCI)数据,分析了 2002—2011 年间的水路交通事故空间分布,利用 GIS 实现了结果可视化分析,研究结果表明:大约 51.1% 的水路交通事故发生在距离大陆 25mile 范围内,62.2% 的事故发生在距离大陆 50mile 范围内。Nezval 等[4]使用核密度估计(Kernel Density Estimation,KDE)方法分析了捷克境内火车与野生动物碰撞数据,最终识别出了 208 个碰撞热点,并通过风险参数对上述热点进行了排序,为铁路部门完善、提升安全防护措施奠定了基础。作为免费开放的灾害数据库,EM-DAT 数据库得到了国内外学者的广泛关注。Shen 等[5]利用 EM-DAT 数据库记录的技术灾难相关数据,建立了风险预测模型,开展了全球各国技术灾难人口损失、受伤害人数、经济损失等预测和验证,结果整体符合预期。目前,针对交通运输事故时空分布的相关研究,通常采用政府公报或者商业数据库中的数据,无法在保证基础数据的完整性和一致性,且获取难度较大;EM-DAT 数据库在数据标准化及

基金项目:国家自然科学基金项目(52071250, 51709220),中央高校基本科研业务费专项资金(2018IVB078, 2019III096CG)资助。
1mile = 1609.344m。