文章编号:1000-5862(2021)02-0145-08

侧撞事故伤亡决定特征提取和影响因素分析

张志坚,江育斌,严利鑫
(华东交通大学交通运输与物流学院,江西 南昌 330013)

摘要:在交通事故中侧撞事故严重性最强,分析侧撞事故伤亡特征并得出决定性影响因素有助于降低事故风险. 该文先使用随机森林特征选择算法对交通事故的影响因素进行降维;然后,通过随机森林、神经网络、SVM 模型 3 种分类模型来验证降维效果,得出事故影响因素的重要程度;最后,构建了 3 种 Logit 模型来探究事故伤亡程度与影响因素之间的关系,得出各因素对事故严重程度的具体影响. 研究结果表明:当提取安全气囊、车辆类型、年龄、防护措施等 9 种重要程度较大的影响因素作为特征子集时,模型的准确率和有效率达到较高水平. 随机参数 Logit 模型分析的结果表明:在 T 字路口、驾驶员为男性等情况下事故率较低;车辆体型越大事故伤亡率越低.

关键词:交通事故;侧面碰撞;特征选择;随机森林模型;随机参数 Logit 模型
中图分类号:TP 311;U 491.31 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2021.02.06

0 引言

根据世界卫生组织的统计数据,全世界每年约有 135 万人死于交通事故,其中侧面碰撞事故和追尾事故约为所有事故类型的 1/2,而该占比在某些国家和地区中达 2/3. 据美国交通部报告称,2017 年在美国发生的严重伤亡事故中,侧撞事故约占死亡总数的 28%. 2014 年在中国发生的所有撞车事故中,车辆间的侧撞事故约占 41%,在各种车辆碰撞类型中排名第 1. 由此可见人员伤亡程度与事故碰撞类型存在较强的关联性,侧面车辆碰撞在交通事故中占比较大,且造成的人员伤亡程度更为严重. 由于侧面碰撞事故具有更高的危险性,进行侧面碰撞事故的相关研究很有现实意义.

事故的伤亡程度受到驾驶员、车辆、环境等多方面因素的影响,不同学者针对不同方面做了许多研究. Yuan Quan 等[1] 分析了人、车、路、环境变量对碰撞严重程度的影响,发现车辆类型、白天、车速等对事故伤亡率影响显著. A. K. Celik 等[2] 分析了恶劣天气、车辆、驾驶员年龄等因素对侧面碰撞事故严重程度的影响. 沈小燕等[3] 发现碰撞类型、路段类型、时间、道路线形等因素对道路运输事故的影响显著. M. Bareiss 等[4] 发现乘员的年龄、性别和 BMI 是侧面事故风险的重要影响因子. M. E. Kelley 等[5] 分析了乘员年龄、性别、身高、安全带的使用以及车辆类型等因素对侧撞事故的影响.

针对侧面碰撞事故,学术界逐渐有了较多研究,现有的研究主要选取了部分影响因素来分析特定问题,或综合考虑各方面特征进行整体分析,然而,交通环境的组成因素非常复杂,存在较多冗余特征. 因此考虑冗余特征的影响,通过特征选择方法进行降维来分析事故严重性的决定特征,可能会对相关结论做出一些补充. 由于特征选择算法只能对事故影响因素进行重要程度的排序,并提取出对事故影响较大的因素,但是这无法验证特征提取的有效性,因此使用分类模型进行验证也成为必要步骤.

随着机器学习算法和数据挖掘技术在学术界中的广泛应用,一些机器学习算法被广泛用来分析交通事故的主要影响特征,许多学者通过随机森林等特征选择算法对事故数据进行降维并分析、预测交通事故. Zhai Ben 等[6] 先应用随机森林算法来识别和排序最重要的变量,再基于重要变量来建立雾天车辆碰撞风险预测模型;朱林艳[7] 构建了随机森林模型来对特征变量重要性进行排序,并预测了行人受伤严重程度;黄兆国等[8] 针对车速、车距等雨天

收稿日期:2020-10-29
基金项目:国家自然科学基金(51805169)资助项目.
作者简介:张志坚(1978—),男,江西丰城人,副教授,博士,主要从事交通运输、运营与供应链管理研究. E-mail:zzjjxs@126.com

事故风险特征使用随机森林算法进行判别;王少华等[9]应用随机森林算法识别了车座损伤、机动车类型、车把旋转等是判定交通行为方式的关键特征变量;高珍等[10]利用随机森林模型预测了交通事故持续时间;K. Hamad 等[11]使用随机森林模型和人工神经网络模型进行了交通流的预测;Yu Bo 等[12]使用随机森林模型建立了应用于超速预测的模型,得出了交通事故影响因素的重要程度. 随机森林算法具有泛化性强、计算效率高、参数调整方便等优点,很适合于交通事故的特征提取和预测研究.

在得出事故的伤亡决策特征后,分析碰撞损伤严重程度与决策特征之间的关系就成为下一步的必要工作. 在分析决策变量与影响因素的交互作用方面,由于可以准确量化参数关系,所以 Logit 回归得到了广泛应用. 如 G. F. Ulfarsson 等[13]利用多项 Logit 模型分析了男性、女性驾驶员对事故严重程度影响的差异性;马壮林等[14]采用有序 Logit 分析隧道事故,发现事故的时间、超速、天气等事故属性对隧道事故严重程度有显著影响.

但是在实际的交通环境中道路事故的引发会受到人、车、环境的综合影响,在每起事故中如性别、年龄等影响因素的作用效果可能会出现随机偏差. 学术界一般将这种随机偏差称为在建模过程中容易忽略的变量异质性. 随机参数 Logit 模型假设自变量的系数服从某种概率分布(如正态分布),并通过这种系数的随机性来描绘在不同风险情况下事故等级的差异性;近年来在交通安全领域中该模型逐渐得到了越来越广泛的应用. Liu Pengfei 等[15]使用随机参数 Logit 模型来探索在碰撞数据中未观察到的异质性;M. Rezapour等[16]使用随机参数 Logit 模型分析了车辆与路边交通障碍碰撞的事故特征;Hou Qinzhong 等[17]使用随机参数 Logit 模型解决了数据的异质性并提升了拟合优度,发现女驾驶员、黑暗无照明等因素对事故伤亡影响显著;陈昭明等[18]使用随机参数 Logit 模型分析了性别、能见度、车辆类型等因素对高速公路事故严重等级的影响. 在侧面碰撞事故方面,也有学者使用此模型进行研究,如 K. Haleem 等[19]使用随机参数 Logit 模型分析了侧面碰撞事故的变量特征,发现事故司机的年龄、交通量、车辆类型、撞击侧以及卡车的百分比等因素对侧面碰撞事故影响显著.

综上所述,提取出对事故严重程度有重要影响的决定特征,对于交通事故的分析和预测建模有重要意义. 本文首先使用随机森林等特征选择算法对影响事故的几类主要因素进行重要度排序;然后,使用机器学习分类算法对排序结果验证其有效性;最后,基于事故的主要影响因素使用 Logit 模型分析事故的伤亡程度与主要特征之间的关系,以期寻找事故的内部特征,为相关安全管理措施提供决策依据.

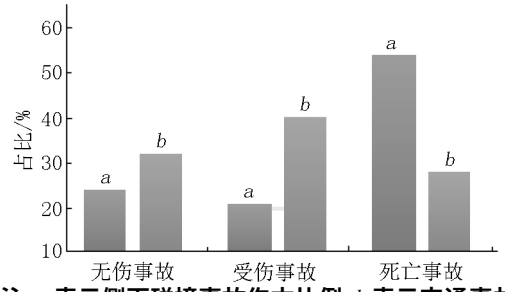
1 数据准备

基于美国死亡率分析报告系统(FARS)的交通事故数据库,提取了 2010—2017 年的事故数据,FARS 的交通事故数据库主要包括驾驶员信息、涉事车辆信息以及事故特征信息 3 个子数据库,通过提取和整合并借鉴已有研究成果,选择了反映事故驾驶员特征、车辆和事故特征等影响因素. 在原始数据中将事故严重程度分为死亡事故、重伤事故、轻伤事故及财产损失事故. 然而,由于事故严重程度是由相关部门根据现场情况记录的,重伤事故和轻伤事故的界限有事故个体差异和判断的主观性,所以可能并非完全合理. 因此,本文将 2 者合并为“受伤事故”,即将财产损失、受伤和死亡这 3 种事故严重程度作为建模因子,其他各变量的名称及具体含义如表 1 所示.

表 1 影响因素分类表

序号	变量名	符号	变量取值设定
1	严重程度	x_1	仅财产损失或无明显伤 = 1,轻伤或重伤 = 2,死亡 = 3
2	车辆位置	x_2	路中 = 1,路侧 = 2,路沿 = 3,路外 = 4
3	路口类型	x_3	非路口 = 0,十字路口 = 1,T 字路口 = 2
4	道路类型	x_4	市区 = 1,郊区 = 2
5	时间	x_5	0:01—6:00 = 1,6:01—12:00 = 2,12:01—18:00 = 3,18:01—24:00 = 4
6	车辆类型	x_6	轿车 = 1,SUV = 2,皮卡 = 3,卡车 = 4,其他 = 5
7	年龄	x_7	0 ~ 24 岁 = 1,25 ~ 45 岁 = 2,46 ~ 60 岁 = 3,60 岁以上 = 4
8	性别	x_8	男 = 1,女 = 2
9	防护措施	x_9	安全带 = 1,头盔 = 2
10	安全气囊	x_{10}	有 = 1,无 = 2
11	酒驾	x_{11}	有酒驾 = 1,无酒驾 = 2
12	药驾	x_{12}	有药驾 = 1,无药驾 = 2
13	光照条件	x_{13}	白天 = 1,夜晚有光照 = 2,夜晚无光照 = 3,黎明或黄昏 = 4
14	天气	x_{14}	晴 = 1,雨 = 2,阴 = 3,雪 = 4,大风 = 5,其他 = 6
15	工作日	x_{15}	工作日 = 1,休息日 = 2

该数据库将事故的碰撞类型分为 Angle、Front-to-Rear、Front-to-Front、Rear-to-ide、Rear-to-Rear、Sideswipe、Other 等 7 种，其中 Angle 包括左转、右转等各种车辆侧面受到碰撞的事故，将以这一类数据为基础进行研究. 通过提取到的侧面事故数据与所有碰撞类型的事故数据进行对比，从图 1 可以看出侧面碰撞事故的伤亡占比都较高.



注：a 表示侧面碰撞事故伤亡比例，b 表示交通事故总体伤亡比例.

图 1　不同类型事故的伤亡占比

2　模型选择

特征选择方法广泛应用于数据处理和建模领域中，可以去除冗余特征，得到最优特征子集. 特征选择方法使得数据挖掘算法效果提升，并具有较高泛化能力. 特征选择方法主要有过滤(Filter)法和包裹(Wrapper)法 2 种，过滤法独立于后续的机器学习算法，它先对数据集进行特征选择，再训练学习器，但不能保证选出最优化特征子集. 而包裹法直接把最终将要使用的学习器的性能作为特征子集的评价依据，可以选择最有利的特征子集，效果一般较好，但时间复杂度相对较高，不适用于高维数据. 由于在本研究中特征维度较小，包裹法可以获得较好的训练效果，故本文采用的特征选择方法为包裹法.

2.1　随机森林算法

随机森林(Random Forest，RF)算法是利用多棵决策树分别训练数据集并集成预测的一种包装型特征选择算法，是 Bagging 算法之一. 该算法采用的是 Bootstrap 随机有放回抽样，从事故数据集中随机抽取数据来构造多棵决策树，再根据决策树的结构进行投票，从而得出最终预测结果. 交通事故受驾驶员、车辆、环境等多方面因素的影响，数据变量的构成非常复杂，噪声也较多. 而随机森林算法具有特征选择和分类预测的功能，且具有泛化性良好、对噪声不敏感的特性，能在分析时得出变量的重要度评分(Variable Importance Measure，VIM)[20]. 随机森林算法根据变量的重要度评分来对特征进行评估，评分来源于计算每个特征对随机森林里每棵决策树做的贡献值后取平均值，对特征之间的贡献值进行比较、排序. 特征对每棵树的贡献通常可以用基尼指数(Gini index)或者袋外数据(Out of Bag，OOB)错误率作为评价指标来衡量[20].

基尼指数表示在样本集合中某个随机选中的样本被分错的概率. Gini 指数(数值型常数)越小表示选中的样本被分错的概率越小. 将特征重要度评分用 V 来表示，Gini 指数用 G_I 来表示. 设有 M 个特征 $X_1, X_2, \cdots, X_m$，计算出每个特征 X_j 的 Gini 指数评分 V_j^{Gini}，即第 j 个特征在 RF 所有决策树中节点分裂不纯度的平均改变量，Gini 指数的计算公式为

$$G_{I_m} = \sum_{k=1}^{|K|} \sum_{k' \neq k} k p_{mk} p_{mk'} = 1 - \sum_{k=1}^{|K|} p_{mk}^2,$$

其中 K 表示类别总数，p_{mk} 表示在节点 m 中类别 k 所占的比例，即任意从节点 m 中随机抽取 2 个样本，其类别标记不一致的概率.

特征 X_j 在节点 m 处的重要性用节点 m 分枝前后的 Gini 指数变化量表示，其计算公式为

$$V_{jm}^{\text{Gini}} = G_{I_m} - G_{I_l} - G_{I_r},$$

其中 G_{I_l} 和 G_{I_r} 分别表示分枝后 2 个新节点的 Gini 指数.

最后，把所有求得的重要性评分做归一化处理：

$$V_j^{\text{Gini}^*} = V_j^{\text{Gini}} / \sum_{i=1}^{c} V_i^{\text{Gini}},$$

其中 $\sum_{i=1}^{c} V_i^{\text{Gini}}$ 是所有特征的增益之和，V_j^{Gini} 是特征 X_j 的基尼指数.

2.2　Boruta 算法

Boruta 算法是基于随机森林算法构建的包装型分类器，使用平均下降精度来衡量变量的重要性. Boruta 算法考虑了在随机森林算法中决策树的平均准确度损失的波动，通过引入阴影特征向样本集合中添加随机性并从随机样本集中分析结果，可以减少由随机波动和相关性所产生的误差. 算法的具体构建流程如下：

(i) 创建阴影特征，首先复制原始特征集，对每个特征值随机混合，构建出随机组合的阴影特征，拼接到真实特征中，构成新的特征集；

(ii) 在新的特征集合上运行随机森林分类器，基于平均下降精度来分别计算原始特征和阴影特征的重要度评分，在 Boruta 算法中的特征重要度评分用 Z 分数来表示；

(iii) 对比原始特征的 Z 分数是否高于其阴影特征的最大 Z 分数,Z 分数低于其阴影特征最大 Z 分数的特征被认为是不重要的,将其从特征集中删除;

(iv) 遍历所有特征,筛选出可用于建模的最优特征子集.

在 Boruta 算法中的重要度评分是基于 RF 模型的袋外误差定义的,袋外误差的计算公式为

$$M_{OOB} = (y_i - \hat{y}_{iOOB})^2/N,$$

其中 M_{OOB} 为随机森林的袋外误差,y_i 是样本值(模型数据的输入项),$\hat{y}_{iOOB}$ 是样本 y_i 的袋外样本的预测值.

Z 分数(变量重要度)用于评估每个影响因素的重要性,Z 分数的计算方法为

$$Z_{\text{score}} = \overline{M_{OOB}}/S_{OOB},$$

其中 Z_{score} 为 Z 分数(它表示算法得出的重要度评分,该值越大重要度越高),$\overline{M_{OOB}}$ 是袋外误差的平均值(它表示分类误差),S_{OOB} 为袋外误差的标准差.

筛选结果以阴影特征重要性的 Z 分数最大值为筛选指标,若特征变量的值大于 Z 分数,则该特征被认为是重要的;否则,该变量被认为是不重要的.

3 事故决策属性的选取和验证

为验证 Boruta 算法和 RF 算法的训练效果,加入了递归特征消除(recursive feature elimination,RFE)筛选算法作为对照组,RFE 是经典向后选择算法的一种,在特征选择方法中运用广泛. 本文采用在 R 软件中的 Boruta 和 randomForest 软件包分别构建 Botura 算法和 RF 算法提取事故的属性重要程度. 这 2 种算法都基于随机森林思想,首先需要设定每个节点处样本预测器的个数 m_{try} 的初始值(通常为变量个数的均方根),以基尼不纯度或袋外数据误差最小为依据确定最佳的决策树数量 n_{tree}. 从图 2 可以看出,当 $n_{tree} > 300$ 时,模型的分类准确度趋于稳定,在 $n_{tree} = 400$ 之后波动较小,且错误率水平达到最低,所以为了在保证效能的情况下减少决策树的数量并减少运行时间,本文取 $n_{tree} = 400$. 这 3 种属性选择算法对影响因素重要程度的排序如图 3 所示.

特征选择算法只能对事故影响因子进行重要排序,无法对最优的提取结果进行确定. 而加入分级模型则能够有效解决这一问题. 通过选择不同数量属性计算其分级准确率,最后选择最优属性组合以期得到最优识别结果. 本文采用 RF、SVM、ANN 这 3 种常用的分类器模型对在不同属性选择方法下的识别效果进行评价. 图 4 ~ 图 6 为不同分类器在采用不同特征选择算法时的分类结果.

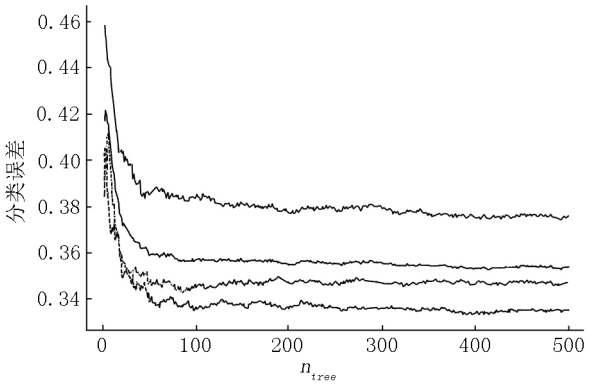
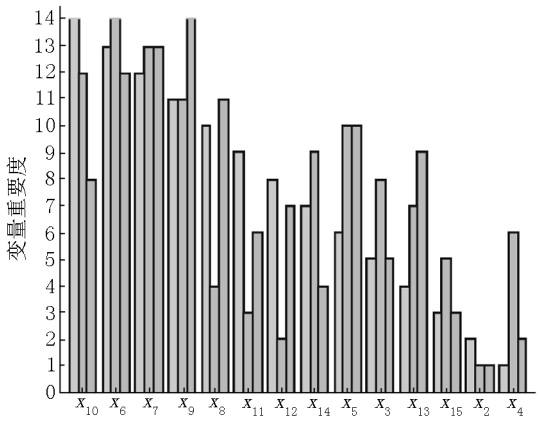


图 2 分类误差与决策树数量的关系



注:在每个变量的柱状图中左侧为 Boruta 算法,中间为 RF 算法,右侧为 RFE 算法.

图 3 变量重要度对比

由图 4 可知,在使用 Boruta 算法时 RF 的分类准确率较高,当选择 7 个特征时 RF 分类算法的准确率最高,且从选择第 10 个特征起,分类准确率下降幅度增大. 由图 5 可知,3 种特征选择算法对 SVM 的优化效果无明显差距,当选择 10 个特征时均能使 SVM 分类准确率在 60% 以上,再增加特征对分类准确率提升较小. 由图 6 可知,当采用 Boruta 算法时,选择 9 个特征能够使 ANN 的分类准确率最高,此时 ANN 的分类准确率也高于其他分类算法. 从图7 可知,使用 Boruta 算法得到特征排序的最优分类结果略优于 RF 特征选择算法和 RFE 特征选择算法.

当使用 Boruta 算法选择 9 个特征时,ANN 分类算法的准确率在几种算法中达到最高,继续增加特征数量对 ANN 和 SVM 分类算法的准确率影响不大,反而会降低 RF 分类算法的准确率. 所以最终选择安全气囊、车辆类型、年龄、防护措施、性别、酒驾、药驾、天气、时间等 9 个特征作为决策特征属性.

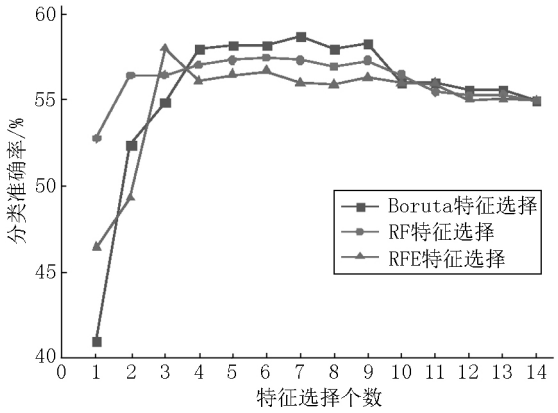


图 4　RF 分类结果

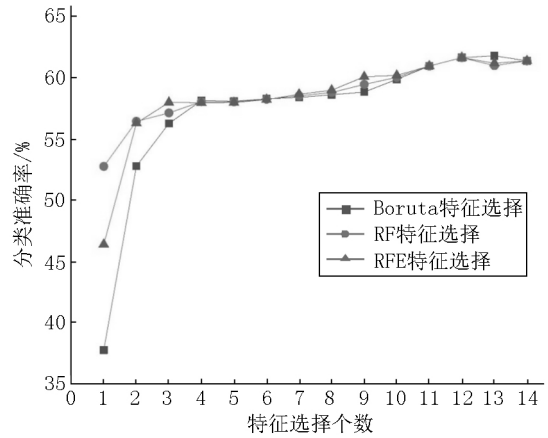


图 5　SVM 分类结果

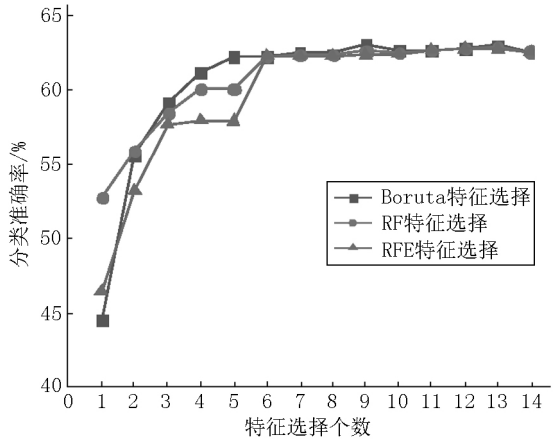


图 6 ANN 分类结果

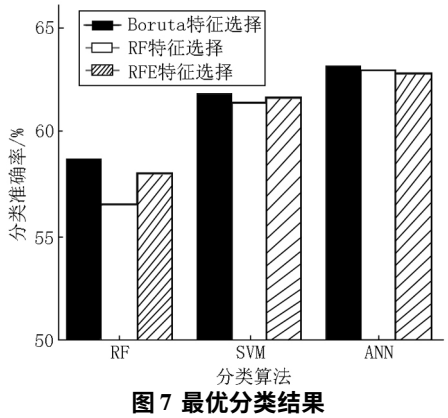


图 7 最优分类结果

4　事故特征分析

4.1　随机参数 Logit 模型

学者们运用了 Logistic 回归模型、多项 Logit 模型、异方差 Logit 模型、随机参数 Logit 模型等来分析事故严重等级. 不同的 Logit 模型存在不同的侧重点. 其中 Logistic 回归主要用于二分类情况的建模; 多项式 Logit 模型主要用于分析决策变量有 3 种因素的建模; 异方差 Logit 模型用变量来描述在指标中存在的方差，以分析在建模过程中存在的异方差性; 随机参数 Logit 模型用服从随机分布(如正态分布) 的系数来刻画自变量的随机性，并通过这种随机系数来描绘不同风险情况下事故等级的差异性.

由于本文的建模目标决策变量是多分类的，同时影响因素情况较为复杂可能会存在异质性，因此，针对上述 Logit 模型的各自特性，使用随机参数 Logit 模型来对事故严重等级和事故特征之间的关系进行敏感性分析.

本文研究对象是事故严重程度，即事故分别是财产损失事故、受伤事故和死亡事故的概率. 模型的效用函数(即决定事故严重程度的函数) 可表示为

$$S_{in}=\boldsymbol{\beta}_i X_{in}+\varepsilon_{in},$$

其中 S_{in} 为事故 n 对严重程度 i 的效用函数，n 为事故编号，X_{in} 为事故严重程度的影响因素集合，$\boldsymbol{\beta}_i$ 为事故严重程度影响因素的参数向量，ε_{in} 为误差项.

在多项 Logit 模型中设定各个参数向量 $\boldsymbol{\beta}_{in}$ 都为固定参数，它表示在模型中每个数据案例记录的事故影响因素对于决策变量的作用效果都保持一致，没有考虑到不同事故对象在同一事故因素情况中的随机性影响. 但是在实际的交通环境中道路交通事故的引发会受到人、车、环境的综合影响，在每起事故中各个影响因素的作用效果可能会出现随机偏差，学术界一般将这种随机偏差称为在建模过程中容易忽略的变量异质性. 如在分析驾驶员性别因素的影响时，不同驾驶员的驾龄、生理和心理条件、驾驶水平、应急反应能力等都存在不同，因此同种性别的驾驶员在驾驶过程中由于受到其他相关因素的综合影响而引发不同严重程度的事故，即性别因素对于事故严重等级的影响可能存在差异，这种差异就是变量的异质性.

为考虑变量的异质性，在建模时给变量参数向量 $\boldsymbol{\beta}_i$ 增加 1 个随机项 $\boldsymbol{\alpha}_i v_{in}$，以体现该因素对事故严重程度影响的随机变化性，即

$$\boldsymbol{\beta}_{in}=\boldsymbol{\beta}_i+\boldsymbol{\alpha}_i v_{in},$$

其中 $\boldsymbol{\beta}_{in}$ 为在事故 n 中影响因素 X_{in} 对事故严重程度 i 的参数向量，$\boldsymbol{\alpha}_i$ 为系数矩阵(它表示各随机参数间协方差及潜在相关性) ，v_{in} 为均值为 0、协方差矩阵

为单位阵的随机项(服从标准正态分布),它表示未观测到的数据异质性.

4.2 拟合效果检验

为检验随机参数 Logit 模型的拟合效果,本文分别建立了随机参数 Logit 模型、多项 Logit 模型和异方差 Logit 模型,并分别计算了 3 种模型的 AIC(赤池信息准则)和 McFadden R^2 值.

AIC 和 McFadden R^2 是衡量统计模型拟合优良性的常用标准,通常 AIC 数值越小表示模型的拟合优度越好. McFadden R^2 取值范围为 [0,1],McFadden R^2 越接近于 1 说明模型的分析结果越优.

从表 2 中 McFadden R^2、AIC 这 2 个拟合指标值可以看出,随机参数 Logit 模型的 AIC 最小且 McFadden R^2 值最接近 1,拟合效果更优,因此采取随机参数 Logit 模型来分析事故特征更为合理.

表 2 模型拟合效果对比

	McFadden R^2	AIC
随机参数 Logit 模型	0.277	27 346.88
异方差 Logit 模型	0.273	27 493.45
多项 Logit 模型	0.174	27 785.47

4.3 参数评估

表 3 列出了随机参数 Logit 模型的参数评估结果. 表 3 仅针对具有统计学意义的事故因素给出了估计系数 B、z 值和相对风险比 r(relative risk ratio,RRR)[21]. B 表示不同事故等级严重性相比的差异,z 值表示显著性,相对风险比 r 是由估计系数得出的表示不同影响因素对事故发生率的影响.

1) 路口类型. 由表 3 可知,相比其他类型的路面情况,T 型路口事故的受伤概率和死亡概率更低,分别是其他路口类型的 0.791 和 0.799. 潜在原因是:T 型路口的车辆路线相比十字路口等更为固定,驾驶员在到达路口时会更加谨慎;且在一般情况下,驾驶员在交叉口区域内的行驶速度较低,碰撞动能较小,也更不容易发生伤亡事故.

2) 时间. 12:01—18:00 和 18:01—24:00 的死亡事故发生率低于其他时段,且 18:01—24:00 的死亡事故发生率低于 12:01—18:00. 潜在原因可能是傍晚和夜间由于光照条件相比于白天更差,驾驶员的驾驶行为更加谨慎,车速也可能有所降低.

表 3 参数评估结果

变量	受伤事故			死亡事故		
	B	z 值	r	B	Z 值	r
T 字路口	−0.234 772***	−3.548 5	0.791	−0.224 834**	−2.864 8	0.799
雨天	0.441 553***	5.270 2	1.556	0.675 838***	6.851 5	1.966
雪天	0.582 727***	4.260 0	1.791	0.839 831***	5.162 6	2.316
多云	0.327 453***	5.383 0	1.387	0.370 708***	5.101 4	1.449
12:01—18:00				−0.422 526**	−3.108 6	0.655
18:01—24:00				−0.595 514***	−5.094 3	0.551
SUV	−0.298 310***	−4.747 1	0.742	−1.512 082***	−19.061 7	0.220
货车				−1.163 598***	−10.011 2	0.312
皮卡	−0.423 534***	−6.845 6	0.654	−1.707 856***	−22.048 3	0.181
卡车	−1.738 116***	−23.219 2	0.176	−4.109 794***	−28.877 1	0.016
两轮车	−1.208 322***	−6.376 9	0.299	−1.291 886***	−6.565 2	0.275
46~60 岁	0.175 383**	2.710 1	1.191	0.711 483***	8.962 3	2.036
60 岁以上	0.306 590***	3.930 9	1.359	2.033 314***	23.464 0	7.637
女性驾驶员	0.393 921***	7.759 1	1.483	0.445 295***	7.542 6	1.560
无安全带	1.152 694***	10.534 9	3.168	2.221 691***	19.469 0	9.226
无头盔	1.189 455***	8.052 9	3.284	2.220 583***	14.742 5	9.217
无安全气囊	−1.377 918***	−31.171 1	0.252	−1.259 856***	−23.080 3	0.284
酒驾				0.519 173***	4.300 5	1.680
药驾	0.314 893**	2.583 8	1.370	0.658 138***	4.987 2	1.931

注:**、***分别表示在 5% 和 1% 的显著水平上显著.

3) 天气. 雨雪和多云天气都会增加事故发生概率,其中雨天的事故发生概率约是雪天的 1.151 倍(1.791/1.556). 这可能是因为雪天的能见度和路面状况都最为恶劣,驾驶员会更谨慎地驾驶,更不容易采取一些危险驾驶行为;而雨天由于路面湿滑,制动距离也更长,且雨天路面情况要优于雪天,雨天的出现频率往往也远多于雪天,不能引起驾驶员的足够重视,从而导致相比雪天事故发生率雨天反而更高.

4) 性别. 由表 3 可知,相比于男性驾驶员,女性驾驶员的受伤事故和死亡事故的发生率分别是男性

的 1.483 和 1.560 倍. 这反映了男性与女性驾驶员在驾驶风格、经验与风险感知方面的差异.

5) 年龄. 60 岁以上的驾驶员事故死亡率远高于其他年龄段的驾驶员(它是其他年龄段的 7 倍以上，$r = 7.637$). 这可能主要是因为随着年龄增大驾驶员的反应能力的减弱和身体素质的下降. 而 46 ~ 60 岁这一年龄段的驾驶员事故伤亡率相比于较年轻驾驶员，伤亡率已经有所上升，文献[22]研究表明驾驶员的驾驶能力会随年龄上升而减弱. 这说明 45 岁以上驾驶员的驾驶能力已经有所下降.

6) 酒驾和药驾. 酒驾和药驾的驾驶员死亡事故发生率分别是正常驾驶员的 1.680 和 1.931 倍，这说明酒驾和药驾的危害性，同时提醒驾驶员药驾的潜在危害会更大.

7) 车辆类型. 当事故车辆为卡车时，发生受伤事故和死亡事故的概率远低于事故车辆为其他类型车辆的情况($r = 0.176$，$r = 0.116$)；潜在原因是：大型车辆行驶速度普遍较低，加之大型车辆驾驶室普遍较高(在一定程度上可保护驾驶员)，从而降低了事故伤害程度. 皮卡和 SUV 的安全系数次于卡车但优于普通轿车和两轮车等；潜在原因是：SUV 和皮卡相比于轿车具有更大的重量和体型，防撞性能也比小轿车更好. 从表 3 的相对风险比可以看出，车辆类型对于事故严重程度的差异性很大，不同类型的车辆伤亡率差异较大，车辆体型越大事故伤亡率越低.

8) 防护措施. 安全带和头盔对事故严重程度的影响显著. 未戴头盔和未系安全带会显著增加事故的伤亡率(分别是普通驾驶员的 3.168 和 3.284 倍)，由此可见不使用防护措施会极大地增强事故风险. 在安全气囊弹出的侧面碰撞事故中，事故的伤亡率却远高于安全气囊未弹出的侧面碰撞事故；潜在原因是：安全气囊只有在受到较为猛烈的撞击时才会弹出，未弹出安全气囊的事故车辆撞击往往并不严重，因此伤亡率较低；而弹出安全气囊事故的碰撞动能往往较大，安全气囊对于侧面碰撞的防护效果又较差，导致此类事故的伤亡率较高.

5　结语

1) 采用 Boruta、RF、RFE 这 3 种特征选择算法按不同特征对伤亡程度的影响重要度进行了排序，通过采用 RF、SVM、ANN 这 3 种分类方法对事故伤亡程度的决策特征进行确定. 发现通过特征选择可以减少事故预测的变量维度，且能保证预测精度，可以显著降低建模难度. 降维结果显示：安全气囊、车辆类型、年龄、防护措施、性别、酒驾、药驾、天气、时间这 9 个特征对事故伤亡程度具有决定性影响.

2) T 型路口相比其他路段的事故发生率较低. 12:01—18:00 和 18:01—24:00 的死亡率低于其他时段. 雨雪和多云天气都会增加事故发生概率. 女性驾驶员的伤亡事故发生率都更高. 60 岁以上的驾驶员事故死亡率远高于其他年龄段的驾驶员，46 ~ 60 岁这一年龄段的驾驶员事故伤亡率略高于较年轻驾驶员. 卡车发生受伤事故和死亡事故的概率远低于事故车辆为其他类型车辆的情况，车辆体型越大事故伤亡率越低. 安全带和头盔对事故严重程度的影响显著，在侧面碰撞事故中，安全气囊弹出的事故伤亡率高于安全气囊未弹出的事故.

3) 研究成果可为相关部门制定事故预防措施和交通环境改善方案提供决策依据. 依据本文研究成果可知：在能见度下降的雨雪天气和路面湿滑时，加装可变标识对驾驶员做出实时提醒；加强女性和 45 岁以上驾驶员的驾驶能力和交通安全意识培训；增强侧面安全气囊的研发和推广力度以及车辆的侧面防护性能；考虑小型车更易受创，着重增强小型车辆的防护性能.

6 参考文献

[1] Yuan Quan，Xu Xuecai，Xu Mingchang，et al. The role of striking and struck vehicles in side crashes between vehicles：Bayesian bivariate probit analysis in China [J]. Accident Analysis and Prevention，2020，134：105324.

[2] Celik A K，Oktay E. A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars Provinces of Turkey [J]. Accident Analysis and Prevention，2014，72：66-77.

[3] 沈小燕，魏珊珊，冯煜清. 基于机器学习的危险货物道路运输事故影响因素分析 [J]. 交通信息与安全，2020，38(15)：113-119，128.

[4] Bareiss M，Gabler H C. Estimating near side crash injury risk in best performing passenger vehicles in the United States [J]. Accident Analysis and Prevention，2020，138：105434.

[5] Kelley M E，Talton J W，Weaver A A，et al. Associations between upper extremity injury patterns in side impact motor vehicle collisions with occupant and crash characteristics [J]. Accident Analysis and Prevention，2019，122：1-7.

[6] Zhai Ben，Lu Jiantao，Wang Yanli，et al. Real-time prediction of crash risk on freeways under fog conditions [J]. International Journal of Transportation Science and Technology，2020，9(4)：287-298.

[7] 朱林艳. 基于随机森林的行人交通事故研究 [D]. 武汉：华中科技大学，2018.

[8] 黄兆国,过秀成,贾亮. 基于随机森林的雨天车辆跟驰风险行为研究 [J]. 交通信息与安全,2020,38(1):27-34.
[9] 王少华,黄建玲,陈艳艳,等. 自行车驾驶员交通行为方式判定研究 [J]. 重庆交通大学学报:自然科学版,2020,39(6):19-24.
[10] 高珍,柯阿香,余荣杰,等. 基于随机生存森林的交通事件持续时间预测 [J]. 同济大学学报:自然科学版,2017,45(9):1304-1310.
[11] Hamad K,Al-Ruzouq R,Zeiada W,et al. Predicting incident duration using random forests [J]. Transportmetrica A:Transport Science,2020,16(3):1269-1293.
[12] Yu Bo,Chen Yuren,Bao Shan. Quantifying visual road environment to establish a speeding prediction model:an examination using naturalistic driving data [J]. Accident Analysis and Prevention,2019,129:289-298.
[13] Ulfarsson G F,Mannering F L. Differences in male and female injury severities in sport-utility vehicle,minivan,pickup and passenger car accidents [J]. Accident Analysis and Prevention,2004,36(2):135-147.
[14] 马壮林,张祎祎,杨杨,等. 公路隧道交通事故严重程度预测模型研究 [J]. 中国安全科学学报,2015,25(5):75-79.
[15] Liu Pengfei,Fan Wei. Exploring injury severity in head-on crashes using latent class clustering analysis and mixed logit model:a case study of North Carolina [J]. Accident Analysis and Prevention,2020,135:105388.
[16] Rezapour M,Wulff S S,Ksaibati K. Examination of the severity of two-lane highway traffic barrier crashes using the mixed logit model [J]. Journal of Safety Research,2019,70:223-232.
[17] Hou Qinzhong,Huo Xiaoyan,Leng Junqiang,et al. Examination of driver injury severity in freeway single-vehicle crashes using a mixed logit model with heterogeneity-in-means [J]. Physica A:Statistical Mechanics and Its Applications,2019,531:121760.
[18] 陈昭明,徐文远,曲悠扬,等. 基于混合 Logit 模型的高速公路交通事故严重程度分析 [J]. 交通信息与安全,2019,37(3):42-50.
[19] Haleem K,Gan A. Effect of driver's age and side of impact on crash severity along urban freeways:a mixed logit approach [J]. Journal of Safety Research,2013,46:67-76.
[20] 李光华,李俊清,张亮,等. 一种融合蚁群算法和随机森林的特征选择方法 [J]. 计算机科学,2019,46(11A):212-215.
[21] Vajari M A,Aghabayk K,Sadeghian M,et al. A multinomial logit model of motorcycle crash severity at Australian intersections [J]. Journal of Safety Research,2020,73:17-24.
[22] Osman M,Mishra S,Paleti R. Injury severity analysis of commercially-licensed drivers in single-vehicle crashes:accounting for unobserved heterogeneity and age group differences [J]. Accident Analysis and Prevention,2018,118:289-300.

The Decisive Feature Extraction and Main Influencing Factors Analysis of Side Impact Accidents

ZHANG Zhijian,JIANG Yubin,YAN Lixin
(School of Transportation and Logistics,East China Jiaotong University,Nanchang Jiangxi 330013,China)

Abstract: Side collisions are the most serious in traffic accidents. Analyzing the characteristics of injuries and deaths of side collisions and deriving decisive factors can help to reduce the risk of accidents. Firstly,the random forest algorithm is used to reduce the dimensionality of the influencing factors of traffic accidents in this paper. Secondly,the dimensionality reduction effect is verified through three models of classification such as random forest,neural network and SVM,and the importance of the accident influencing factors are obtained. Finally,three Logit models are constructed to explore the relationship between the accident severity and the influencing factors,and the specific impacts of each factor on the severity of the accident are obtained. The results show that the accuracy and efficiency of the model reach a relatively high level when the nine important influencing factors are selected such as airbag,vehicle type,age,and protective measures. The results of the random parameter Logit model show that under the circumstances of lighting at night,T-shaped intersections,the accident rate is lower. The larger the vehicle size is,the lower the accident casualty rate becomes.

Key words: traffic accident; side collision; feature selection; random forests model; random parameter logit model

(责任编辑:曾剑锋)