DOI:10.3969/j.issn.1672-0032.2023.02.004

基于数据挖掘的道路交通事故成因分析

李虹燕,朱龙波,任宪通,许文雯
山东交通学院交通与物流工程学院,山东 济南 250357

摘要:为减少道路交通事故的发生,基于某市2020年的交通事故数据分析道路交通事故规律及成因,以照明条件、能见度和天气等9个因素为自变量,以无伤害、轻伤、重伤及死亡等4种交通事故严重程度为因变量,采用多元Logistic回归模型和有序多分类Logistic回归模型,分析影响交通事故严重程度的重要因素。对某市2021年第1季度的交通事故数据进行验证,结果表明:多元Logistic回归模型和有序多分类Logistic回归模型对交通事故严重程度的正确预测率分别为75.1%、75.0%。基于数据挖掘的道路交通事故成因分析可为交通管理部门治理交通环境、降低交通事故提供依据。

关键词:数据挖掘;事故成因;事故严重程度;多元Logistic回归;有序多分类Logistic回归
中图分类号:U491.31;TP311.13 **文献标志码**:A **文章编号**:1672-0032(2023)02-0020-08
引用格式:李虹燕,朱龙波,任宪通,等.基于数据挖掘的道路交通事故成因分析[J].山东交通学院学报,2023,31(2):20-27.
LI Hongyan,ZHU Longbo,REN Xiantong, et al. Analysis of road traffic accidents based on data mining [J]. Journal of Shandong Jiaotong University,2023,31(2):20-27.

0 引言

近年来,我国汽车保有量屡创新高,因交通事故造成的生命财产损失也在逐年增大。分析交通事故的规律及成因对保障出行安全、减少财产损失至关重要。当前,国内外学者多采用数据挖掘技术研究这一课题。Pande等[1]采用数据分析软件SAS中的卡方检验法分析美国佛罗里达州的交通事故数据,基于5种变量构建分析模型,发现照明不足情况下易发生严重交通事故。Siordia等[2]构建驾驶风险分类系统,模拟卡车在城市、山区、城际等环境的行驶情况,采用分类与回归树法、神经网络法和支持向量机等方法定量分析车辆行驶数据,采用专家评估法分类分析交通事故的影响因素。Geurts等[3]采用数据挖掘方法划定“黑色”区域,发现在划定区域内交通事故集中发生在路口左转、与行人碰撞、车辆失控和多雨天气等情况。王云等[4]采用改进的Apriori算法挖掘交通事故数据,寻找可能造成交通事故发生的因素与交通事故本身属性的联系。董立岩等[5]采用数据挖掘中的粗糙集手法设计2个决策表分析交通事故属性和影响因素,并将此算法应用到现实管理中。宗芳等[6]根据交通事故数据,采用有序多项选择模型计算天气、交通信号及车道等因素对受伤人数的贡献,结果表明道路在没有指示信息时受伤人数多于有信号指示信息时。

根据真实的交通事故数据寻找解决道路交通安全问题的途径,选择合适的交通事故数据挖掘方法,分析交通事故规律和导致事故发生的根本原因[7-10]。本文选取某市2020年的交通事故数据,采用多元Logistic回归分析与有序多分类Logistic回归分析方法挖掘道路交通事故的成因,分析交通事故数据中的规律及特征,研究不同影响因素对交通事故的影响程度,对交通系统建设提出针对性的改善建议,提升道路交通安全水平。

收稿日期:2022-05-18
第一作者简介:李虹燕(1981—),女,山东济宁人,工学硕士,主要研究方向为交通安全大数据分析,E-mail:e_lihy@126.com。

1　研究内容及方法

1.1　交通事故数据处理及编码

整理某市 2020 年交通事故案例数据，包括交通事故发生时间、天气、事故主要责任方、驾驶人驾龄等 103 项数据，以造成事故的主要责任方为筛选条件对交通事故数据进行初步筛选，最终选择 1910 起交通事故案例用于本研究。经过区分度（即项目效度，作为评价项目质量、筛选项目的主要依据）[11] 计算，从数据中选取特征区分度明显的照明条件、能见度、天气和车辆使用性质等 9 个因素为自变量，其中驾驶员年龄因素涉及数据面太广，自变量属性设为标度变量，其他自变量设为名义变量，自变量信息如表 1 所示。

表 1　自变量信息

自变量	自变量赋值	编码	代码	属性	自变量	自变量赋值	编码	代码	属性
照明条件	无照明	1	A_1	名义变量	路面情况	路面完好	1	E_1	名义变量
	有照明	2	A_2			凹凸	2	E_2	
能见度	≤50 m	1	B_1	名义变量		施工	3	E_3	
	>50~100 m	2	B_2			路障	4	E_4	
	>100~200 m	3	B_3			塌陷	5	E_5	
	>200 m	4	B_4			其他	6	E_6	
	其他	5	B_5		道路物理隔离	无隔离	1	F_1	名义变量
天气	晴	1	C_1	名义变量		中心隔离	2	F_2	
	雪	2	C_2			中心隔离加机动车非机动车（机非）隔离	3	F_3	
	雨	3	C_3			机非隔离	4	F_4	
	阴	4	C_4		驾驶人文化程度	初中毕业及以下	1	G_1	名义变量
	雾	5	C_5			高中毕业及以上	2	G_2	
	霾	6	C_6			其他	3	G_3	
	其他	7	C_7		驾驶人性别	男	1	H_1	名义变量
车辆使用性质	非营运	1	D_1	名义变量		女	2	H_2	
	营运	2	D_2		驾驶人年龄			I_1	标度变量

以无伤害 Z_1、轻伤 Z_2、重伤 Z_3 及死亡 Z_4 等 4 类交通事故严重程度为因变量，分析人、车、路、环境对交通安全的影响程度[12-13]。选取的 1910 起交通事故数据样本中无伤害、轻伤、重伤及死亡等 4 类交通事故分别为 1338、504、10、58 起，无伤害类交通事故的比例最高，其他依次为轻伤、死亡、重伤类交通事故。

1.2　Logistic 回归分析

采用多元 Logistic 回归和有序多分类 Logistic 回归分析交通事故严重程度的重要影响因素。

在多元 Logistic 回归分析中，参与回归分析的因变量为有序性分类变量，自变量为无序性分类变量，根据分析结果可得到因变量取特定值的概率与自变量的关系[14]。多元 Logistic 回归模型主要分析各类别目标变量与参照类别目标变量的对比情况，公式为：

$$\ln(P_j/P_J) = \beta_0 + \sum_{i=1}^{k}\beta_i x_i ,$$

式中：P_j 为第 j 类目标变量的概率，P_J 为第 $J(j \neq J)$ 类参照类别目标变量的概率，β_0 为样本观测值，β_i 为

模型的回归系数，x_i 为第 i 个自变量，k 为目标变量的类别数。

有序多分类 Logistic 回归模型多用来分析因变量存在的某种层次关系，所观测到的因变量数据变化趋势为：

$$y^* = \alpha + \sum_{t=1}^{k} \beta_t x_t + \varepsilon ,$$

式中：α 为常量，β_t 为模型的回归系数，x_t 为第 t 个自变量，ε 为误差项。

2 结果及分析

2.1 多元 Logistic 回归分析结果

采用软件 SPSS 设置变量类型，对交通事故数据进行单因素多元 Logistic 回归分析前需对自变量进行显著性检验，排除不显著的自变量，各自变量的显著性检验结果如表 2 所示。卡方 χ^2 越大，说明自变量与因变量的相关性越强。若显著性概率 $p<0.05$，说明变量显著相关；否则，则相关性不显著。由表 2 可知：天气和路面情况的 $p>0.05$，两者均属于不显著变量，其他自变量的 $p<0.05$，表明其他自变量均为显著变量。

表 2 各自变量的显著性检验结果

自变量	能见度	道路物理隔离	天气	性别	文化程度	路面情况	照明条件	年龄	车辆使用性质
χ^2	21.198	22.455	16.113	89.846	160.091	15.494	18.507	170.453	28.173
p	0.047	0.007	0.584	0	0	0.416	0	0	0

采用软件 SPSS 进行多元 Logistic 回归模型拟合度检验，结果如表 3 所示。由表 3 可知：多元 Logistic 回归模型的 $p<0.05$，说明多元 Logistic 回归模型对原始数据的拟合度较好，通过模型拟合度检验，可用于解释和分析影响交通事故严重程度的影响因素。

表 3 多元 Logistic 回归模型拟合度检验结果

项目	−2 倍对数似然	χ_2	自由度	p
仅截距	2 534.170			
最终	2 211.016	323.154	24	0

注：仅截距是指多元 Logistic 回归模型自变量值为 0 时因变量的值。最终是指导入数据后的拟合结果。

计算得到多元 Logistic 回归模型的考克斯−斯奈尔伪 R^2（可决系数）、内戈尔科伪 R^2、麦克法登伪 R^2 分别为 0.220、0.286、0.169，结果均偏低，说明原始变量变异的解释程度一般，拟合程度中等。

以死亡类交通事故严重程度为对照组，分析无伤害、轻伤及重伤等 3 类交通事故严重程度的相对发生概率。因照明条件、能见度、道路物理隔离、车辆使用性质等自变量均为有序性分类变量，需对变量的各类别进行哑变量处理，将分类变量转化为数值变量，按照特征类别进行编码。即设各类别中最后 1 个类别的回归系数 $\beta=0$，其他类别的 β 都以最后 1 个类别为参照，同时将优势比 O_R（$O_R=e^{\beta}$）设为效应指标，分析自变量的各类别对因变量的影响程度，O_R 越大，对应的自变量越重要[15]。当 $\beta>0$ 时，$O_R>1$，说明自变量为危险因素；当 $\beta<0$ 时，$O_R<1$，说明自变量为保护性因素；当 $\beta=0$，即 $O_R=1$，说明该自变量与因变量无关。自变量的不同取值对交通事故严重程度的影响检验如表 4~6 所示。

表 4 无伤害交通事故严重程度的影响检验

数值变量	β	p	O_R	数值变量	β	p	O_R
[能见度等于 1]	−2.674	0.990	0.069	[文化程度等于 2]	3.218	0	24.984
[性别等于 1]	0.633	0.184	1.883	[能见度等于 2]	−14.666	0	4.27×10^{-7}
[车辆使用性质等于 1]	−0.861	0.122	0.423	[能见度等于 3]	−14.475	0	5.17×10^{-7}
[照明条件等于 1]	−0.797	0.020	0.450	[能见度等于 4]	−14.408	0	5.53×10^{-7}
[文化程度等于 1]	2.750	0	15.649	年龄	−0.052	0	0.95

表 5　轻伤交通事故严重程度的影响检验

数值变量	β	p	O_R	数值变量	β	p	O_R
年龄	0.002	0.882	1.002	[照明条件等于 1]	-0.410	0.254	0.661
[性别等于 1]	-0.770	0.109	0.465	[车辆使用性质等于 1]	0	0.997	0.998
[文化程度等于 1]	3.270	0	26.315	[能见度等于 1]	-3.980	0.985	0.019
[文化程度等于 2]	2.903	0	18.234				

表 6　重伤交通事故严重程度的影响检验

数值变量	β	p	O_R	数值变量	β	p	O_R
年龄	-0.025	0.365	0.976	[照明条件等于 1]	1.480	0.057	4.391
[道路物理隔离等于 1]	0.453	1.000	1.572	[车辆使用性质等于 1]	11.768	0.977	2.907
[道路物理隔离等于 2]	0.580	0.999	1.786	[能见度等于 1]	-2.352	0.999	0.095
[道路物理隔离等于 3]	1.711	0.998	5.534	[能见度等于 2]	-2.375	0.999	0.093
[性别等于 1]	-0.614	0.514	0.541	[能见度等于 3]	-1.331	1.000	0.264
[文化程度等于 1]	2.303	0.005	10.008	[能见度等于 4]	-1.031	1.000	0.356
[文化程度等于 2]	2.076	0.132	7.970				

$p<0.05$ 说明对应自变量的取值具有统计意义,对因变量不同分类水平有显著影响。由表 4 可知:文化程度、年龄、照明条件、车辆使用性质对交通事故严重程度有显著影响,在轻伤及重伤交通事故中文化程度低的驾驶人占总人数的比例较大;驾驶人年龄越大,事故严重程度越低;有照明条件下交通事故严重程度较低;车辆为营运用途时交通事故严重程度较低。

2.2　有序多分类 Logistic 回归分析结果

采用有序多分类 Logistic 回归分析, 对有序多分类 Logistic 回归模型进行平行线检验,若 $p>0.05$,说明平行线假设成立;否则,平行线假设不成立,检验结果如表 7 所示。对有序多分类 Logistic 回归模型进行拟合度检验,结果如表 8 所示。

表 7　有序多分类 Logistic 回归模型的平行线检验

项目	-2 倍对数似然	X_2	自由度	p
原假设	2 211.016			
常规	2 184.626	26.390	48	0.995

表 8　有序多分类 Logistic 回归模型拟合信息

项目	-2 倍对数似然	X_2	自由度	p
仅截距	2 534.170			
最终	2 211.016	323.154	24	0

由表 7 可知:检验得到 $p=0.995>0.05$,说明自变量在各回归模型的效应相同,符合有序多分类 Logistic 模型的比例优势假设条件,可采用此有序多分类 Logistic 回归模型。由表 8 可知:最终模型拟合得到 $p<0.05$,说明有序多分类 Logistic 回归模型有统计意义,通过拟合度检验。

计算得到有序多分类 Logistic 回归模型的考克斯-斯奈尔伪 R^2、内戈尔科伪 R^2、麦克法登伪 R^2 分别为 0.156、0.202、0.115,结果均偏小,说明原始变量变异的解释程度一般,有序多分类 Logistic 回归模型的拟合程度中等。

估算各数值变量的参数,结果如表 9 所示。其中,标准误差是因变量各实际值与估计值间的平均差异程度,表明估计值对各实际值代表性的强弱;越小,回归方程的代表性越强,用回归方程估计或预测的结果越准确。自由度是以样本的统计量估计总体参数时,样本中能自由取值的变量个数。

表9 各自变量的参数估算

数值变量	估算	标准误差	瓦尔德χ^2	自由度	p	95%置信区间	
						下限	上限
年龄	0.050	0.004	136.274	1	0	0.041	0.058
[道路物理隔离等于1]	0.381	0.270	1.995	1	0.158	−0.148	0.909
[道路物理隔离等于2]	0.702	0.284	6.112	1	0.013	0.145	1.258
[道路物理隔离等于3]	0.660	0.303	4.748	1	0.029	0.066	1.253
[文化程度等于1]	−0.454	0.149	9.225	1	0.002	−0.747	−0.161
[文化程度等于2]	−1.274	0.219	33.86	1	0	−1.704	−0.845
[照明条件等于1]	0.562	0.153	13.546	1	0	0.263	0.862
[照明条件等于2]	0						
[性别等于1]男性	−1.197	0.140	−0.922	1	0	0.229	0.398
[性别等于2]女性	0						
[车辆使用性质等于1]	0.919	0.23	15.993	1	0	0.469	1.370

由表9可知:各数值变量不同分类水平在模型中的回归系数对因变量不同分类水平有显著性影响,且具有统计意义;年龄、道路物理隔离、文化程度、照明条件、性别及车辆使用性质对交通事故严重程度有显著性影响。

有序多分类Logistic回归分析同样需对变量的各类别进行哑变量处理,将O_R作为效应指标分析自变量的各类别对因变量的影响程度[16-17]。各自变量的O_R如表10所示。

表10 各自变量的优势比

数值变量	β	标准误差	假设检验			O_R	O_R的95%瓦尔德置信区间	
			瓦尔德χ^2	自由度	p		下限	上限
[道路物理隔离等于1]	0.381	0.269 5	1.995	1	0.158	1.463	0.863	2.482
[道路物理隔离等于2]	0.702	0.283 8	6.112	1	0.013	2.017	1.156	3.518
[道路物理隔离等于3]	0.660	0.302 8	4.748	1	0.029	1.934	1.069	3.502
[性别等于1]	−1.197	0.140 4	72.638	1	0	0.302	0.229	0.398
[文化程度等于1]	−0.454	0.149 4	9.225	1	0.002	0.635	0.474	0.851
[文化程度等于2]	−1.274	0.219 0	33.860	1	0	0.280	0.182	0.429
[文化程度等于3]	0[a]					1.000		
[照明条件等于1]	0.562	0.152 8	13.546	1	0	1.755	1.301	2.368
[照明条件等于2]	0[a]					1.000		
[车辆使用性质等于1]	0.919	0.229 9	15.993	1	0	2.508	1.598	3.935
[车辆使用性质等于2]	0[a]					1.000		
年龄	0.050	0.004 3	136.274	1	0	1.051	1.042	1.060

注:0[a]为无效值。

由表10可知:1)在性别对交通事故严重程度的影响中,男性驾驶人导致交通事故的严重程度是女性驾驶人的0.302倍,表明女性驾驶人更易造成严重交通事故;2)在照明条件对交通事故严重程度的影响中,无照明条件下造成交通事故的严重程度是有照明的1.755倍,表明无照明条件下更易造成严重交通

事故;3)在文化程度对交通事故严重程度的影响中,初中毕业及以下的驾驶人造成交通事故的严重程度是高中毕业及以上驾驶人的 2.268 倍,表明文化程度较低的群体更易造成严重交通事故;4)在道路隔离对交通事故严重程度的影响中,造成严重交通事故的概率从大到小依次为中心隔离、中心隔离加机非隔离、无隔离、机非隔离,说明机非隔离能有效遏制严重交通事故的发生;5)在车辆使用性质对交通事故严重程度的影响中,非营运车辆造成的交通事故严重程度是营运车辆的 2.508 倍,表明非营运车辆更容易造成严重交通事故。

结合多元 Logistic 回归模型和有序多分类 Logistic 回归模型的分析结果可知:自变量为女性驾驶人、无照明、文化程度较低、无道路隔离及非营运车辆时,发生严重交通事故的概率更高。

3　交通事故严重程度预测模型

3.1　多元 Logistic 回归模型的拟合方程

基于多元 Logistic 回归分析,得到无伤害交通事故的拟合方程为:

$$\ln Z_1 = 30.379 - 0.797A_1 + 2.750G_1 + \cdots - 14.408B_4 - 0.052I \ ,$$

轻伤交通事故的拟合方程为:

$$\ln Z_2 = 27.954 - 0.414A_1 + 3.273G_1 + \cdots - 15.752B_4 - 0.002I \ 。$$

交通事故严重程度分别为无伤害和轻伤时,多元 Logistic 回归模型的回归系数如表 11 所示。

表 11　无伤害和轻伤时多元 Logistic 回归模型的回归系数

交通事故数值变量	多元 Logistic 回归模型的回归系数		交通事故数值变量	多元 Logistic 回归模型的回归系数	
	无伤害	轻伤		无伤害	轻伤
[能见度等于 5]	0[b]	0[b]	[文化程度等于 2]	3.218	2.903
[能见度等于 4]	-14.408	-15.725	[文化程度等于 1]	2.750	3.270
[能见度等于 3]	-14.475	-15.547	[性别等于 2]	0[b]	0[b]
[能见度等于 2]	-14.666	-15.882	[性别等于 1]	0.633	-0.766
[能见度等于 1]	-2.674	-3.982	[道路物理隔离等于 4]	0[b]	0[b]
[车辆使用性质等于 2]	0[b]	0[b]	[道路物理隔离等于 3]	-11.188	-10.546
[车辆使用性质等于 1]	-0.861	-0.002	[道路物理隔离等于 2]	-12.222	-11.772
[照明条件等于 2]	0[b]	0[b]	[道路物理隔离等于 1]	-11.843	-11.634
[照明条件等于 1]	-0.797	-0.414	年龄	-0.052	0.002
[文化程度等于 3]	0[b]	0[b]	截距	30.379	27.954

注:0[b] 为无效数据。

以某市 2021 年第 1 季度的交通事故数据为例进行实例验证,结果如表 12 所示。由表 12 可知:多元 Logistic 回归模型对交通事故的严重程度为无伤害和轻伤的拟合方程结果和实际测算结果基本吻合。

表 12　多元 Logistic 回归模型实例验证

事故严重程度	事故量	拟合方程结果	实际测算结果
无伤害	922	6.827	6.675
轻伤	792	6.981	6.675

3.2　有序多分类 Logistic 回归模型的拟合方程

基于有序多分类 Logistic 回归模型,得到无伤害交通事故的拟合方程为:

$$\ln Z_1 = 1.187 + 0.562A_1 - 0.984E_1 + \cdots - 0.124B_4 + 0.050I \ ,$$

轻伤交通事故的拟合方程为:

$$\ln Z_2 = 3.936 + 0.562A_1 - 0.984E_1 + \cdots - 0.124B_4 + 0.050I \ ,$$

重伤交通事故的拟合方程为：

$$\ln Z_3 = 4.107 + 0.562A_1 - 0.984E_1 + \cdots - 0.124B_4 + 0.050I \text{ 。}$$

有序多分类 Logistic 回归模型的回归系数如表 13 所示。

表 13 有序多分类 Logistic 回归模型的回归系数

年龄	［车辆使用性质等于 1］	［照明条件等于 1］	［路面状况等于 5］	［路面状况等于 3］	［路面状况等于 1］	［文化程度等于 2］	［性别等于 1］	［天气等于 6］	［天气等于 5］
0.050	0.919	0.562	3.180	−0.532	−0.984	−1.274	−1.197	−0.107	0.258
［天气等于 4］	［天气等于 3］	［天气等于 2］	［天气等于 1］	［道路物理隔离等于 3］	［道路物理隔离等于 1］	［能见度等于 4］	［能见度等于 2］	［伤害程度等于 3］	［伤害程度等于 1］
−0.715	−0.961	0.157	−0.743	0.660	0.381	0.124	−0.004	4.107	1.187

同样以某市 2021 年第 1 季度的交通事故数据为例进行实例验证，结果如表 14 所示。

表 14 有序多分类 Logistic 回归模型实例验证

事故严重程度	事故数	拟合方程结果	实际测算结果
无伤害	922	6.827	6.675
轻伤	792	9.540	6.675
重伤	49	9.711	3.892

由表 14 可知：有序多分类 Logistic 回归模型对无伤害和轻伤交通事故的预测结果和实际测算结果基本吻合；重伤交通事故的拟合方程结果与实际测算结果差异较大，可考虑扩大数据样本容量后修正模型。

经计算，多元 Logistic 回归模型和有序多分类 Logistic 回归模型对交通事故严重程度的正确预测率分别为 75.1%、75.0%，说明 2 种模型对无伤害和轻伤事故的预测较为准确。

4 结束语

采用多元 Logistic 回归模型和有序多分类 Logistic 回归模型识别影响交通事故严重程度的多个自变量，通过回归模型中的回归系数和优势比，得到不同自变量对交通事故严重程度的贡献。对多类别目标变量构建交通事故严重程度预测模型，揭示因变量与自变量间的内在联系。

通过定量分析交通事故数据可科学预测交通事故发生概率，为交通管理部门治理交通设施与道路环境、提高驾驶人安全意识提供依据。本文选取的自变量不够全面，Logistic 模型可能存在偏差，且选取的车辆因素较少，仅分析车辆的使用性质无法得到更多车辆因素对交通事故严重程度的影响。未来研究可通过剔除交通事故数据中噪音变量等干扰因素，寻找影响交通事故严重程度的重要变量，构建更为快速且准确的预测模型，研究各因素间交互作用的影响。

参考文献：

[1] PANDE A, ABDEL-ATY M. Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool[J]. Safety Science, 2009,47(1):145-154.

[2] SIORDIA O S, DIEGO I M D, CONDE C,et al. Driving risk classification based on experts evaluation[C]//Proceedings of the 2010 IEEE Intelligent Vehicles Symposium. La Jolla, CA, USA:IEEE,2010:1098-1103.

[3] GEURTS K, THOMAS I, WETS G. Understanding spatial concentrations of road accidents using frequent item sets[J]. Accident Analysis & Prevention, 2005,37(4):787-799.

[4] 王云，苏勇. 关联规则挖掘在道路交通事故分析中的应用[J]. 科学技术与工程，2008(7):1824-1827.
WANG Yun, SU Yong. Application of association rules in the analysis of traffic accident[J]. Science Technology and Engineering,2008(7):1824-1827.

[5] 董立岩，刘光远，苑森淼，等. 数据挖掘技术在交通事故分析中的应用[J]. 吉林大学学报(理学版)，2006,44(6):951-955.

DONG Liyan, LIU Guangyuan, YUAN Senmiao, et al. Application of data mining to trafficaccidents analysi[J]. Journal of Jilin University (Science Edition), 2006,44(6):951-955.

[6] 宗芳,许洪国,张慧永. 基于 Ordered Probit 模型的交通事故受伤人数预测[J]. 华南理工大学学报(自然科学版),2012,40(7):41-45.
ZONG Fang, XU Hongguo, ZHANG Huiyong. Forecast of injury number due to traffic accident based on Ordered Probit Model [J]. Journal of South China University of Technology (Natural Science Edition),2012,40(7):41-45.

[7] 毛应萍,于丰泉,孙烨垚,等. 道路交通事故数据挖掘分析技术及应用研究[J]. 交通与运输,2020,33(增刊 2):106-111.
MAO Yingping, YU Fengquan, SUN Yeyao, et al. Road traffic accident data mining and applicantion[J]. Traffic & Transportation, 2020,33(Suppl. 2):106-111.

[8] 许洪国. 道路交通事故分析与处理[M]. 2 版. 北京:人民交通出版社,2004.

[9] 裴玉龙. 道路交通安全[M]. 北京:人民交通出版社,2007.

[10] 刘运通. 道路交通安全指南[M]. 北京:人民交通出版社,2004.

[11] European Environment Agency. Traffic accident fatalities[R]. Luxembourg:Commission of the European Communities,2001:3-8.

[12] 马壮林,邵春福,李霞. 基于 Logistic 模型的公路隧道交通事故严重程度的影响因素[J]. 吉林大学学报(工学版),2010,40(2):423-426.
MA Zhuanglin, SHAO Chunfu, LI Xia. Analysis of factors affecting accident severity in highway tunnels based on logistic mode[J]. Journal of Jilin University (Engineering and Technology Edition),2010,40(2):423-426.

[13] 沈斐敏,张荣贵. 道路交通事故预测与预防[M]. 北京:人民交通出版社,2007.

[14] 李世民,孙明玲,关宏志. 基于累积 Logistic 模型的交通事故严重程度预测模型[J]. 交通标准化,2009(2/3):168-171.
LI Shimin, SUN Mingling, GUAN Hongzhi. Prediction model cumulative logistic for severity of road traffic accident[J]. Transportation Standardization, 2009(2/3):168-171.

[15] 王义婷,贺玉龙,孙小端. 高原双车道公路事故形态影响因素分析[J]. 黑龙江交通科技,2017,40(11):8-10.

[16] 刘海珠. 道路交通事故严重程度影响因素分析及预测模型建立[D]. 长春:吉林大学,2014.
LIU Haizhu. The analysis of influencing factors of crash severity and the establishment of prediction model[D]. Changchun:Jilin University,2014.

[17] HONG L, SUDHEER C, JOHN R, et al. Mining and analysis of traffic safety and roadway condition data[C]. Washington:46th Annual Transportation Research Forum,2005.

Analysis of road traffic accidents based on data mining

LI Hongyan, ZHU Longbo, REN Xiantong, XU Wenwen

School of Transportation and Logistics Engineering, Shandong Jiaotong University, Jinan 250357, *China*

Abstract: In order to reduce the possibilities of traffic accidents on road, the regularities and causes of road traffic accidents are analyzed according to the traffic accident data of a city in 2020. Therefore, the multiple logistic regression model and the ordered multiple logistic regression model are used to analyze the important factors affecting the severity of traffic accidents. The 9 factors such as lighting conditions, visibility, weather and so on as independent variables while the 4 traffic accident severity degrees such as no injury, light injury, serious injury and death as dependent variables are introduced into the two models. Based on the two models, the traffic accident data of a city in the first quarter of 2021 are testified, and the results show that the correct prediction rates of traffic accident severity by the multiple logistic regression model and the ordered multiple logistic regression model are 75.1% and 75.0% respectively. This cause analysis of road traffic accidents based on data mining could provide grounds for traffic control authorities to improve traffic environment and to reduce traffic accidents in the future.

Keywords: data mining; accident causes; accident severity prediction; multiple logistic regression; ordered multiple logistic regression

(责任编辑:郭守真)