基于聚类的营运客车驾驶行为风险评估模型

李守义[1] 任方涛[1]
LI Shouyi REN Fangtao

摘 要 为了能够对营运客车进行驾驶行为风险控制，开展了城市道路营运客车驾驶行为风险评估方法的研究。首先，构建以连续急加速、连续急减速、急转弯、急刹车、单次急加速、单次急减速为特征的驾驶行为评价指标体系。然后，提出了使用熵权法计算风险发生频率与使用层次分析法计算事故损失程度的综合权重体系。最后，利用车载 GPS 模块，获取了 35 名驾驶员的营运客车驾驶数据，利用系统聚类算法实现危险驾驶行为等级划分。研究结果表明，加入连续急加速、连续急减速两个特征，系统聚类减少了主观因素的干扰，聚类效果更符合城市路况的风险特征。

关键词 系统聚类；营运客车；驾驶行为；风险评估；熵权法；层次分析法

doi：10.3969/j.issn.1672-9528.2024.08.012

0 引言

研究表明，驾驶员是绝大多数交通事故的主要因素[1]。交通事故主要因素为驾驶员的危险驾驶行为[2]，因此对于驾驶行为风险评估的研究具有重要的意义。

驾驶行为的数据通常来源于两个部分：驾驶模拟数据和自然驾驶数据。2000 年以前，主要采用半物理仿真驾驶模拟器构建交通场景，局限性在于环境压力感知失真、极端场景下驾驶数据难以获取、有效样本数量偏小[3]。随着互联网技术逐渐成熟，场景真实，采集手段多样化。多种传感器和移动智能终端支持实时采集技术，如采用车载传感器（陀螺仪或者加速器）提取纵向横向速度、加速度数据成本低，但是采集数据类型有限[4]。无人机提取基于视频的车辆轨迹数据[5]，数据采集全面，但延时性长、成本高。相比以上采集手段，基于卫星定位的驾驶数据采集技术具有成本低、系统性、实时性、客观的特点。梁秉毅等人[6]基于 GPS、控制器局域网总线（controller area network，CAN）开发公交车驾驶人驾驶风险动态评估系统。张志鸿[7]提出设置单次急加速事件的最短持续时间和连续急加速事件的合并时间阈值的设想。Chen 等人[8]提出基于车辆类型的差异，设计小轿车和大货车的“四急行为”差异性阈值的观点。Xian 等人[9]研究了路型和“四急”行为的相关性，并指出急减速与城市快速路路段交通事故最显著相关，当车辆加速度小于等于 -4 m/s^2 时碰撞概率剧增。危险驾驶行为建模基于固定参数统计模型，如 Logistic 回归、泊松分布等[10]。但是事故数据通常具有过度离散、零频次过高、时空关联、多层结构、异质性等特点，传统固定参数模型很难捕捉多种数据特征。机器学习算法已广泛应用在危险驾驶行为的风险评估中[11]，模型的适应性目前仍存在很大争议。高维特征向量的复杂度、样本量不足、局部最优解、可解释性不足是机器学习算法面临的共同问题。

针对上述问题，本文在自然驾驶数据采集的基础上，构建典型场景下营运客车危险驾驶行为风险评估模型。首先构建城市路况下营运客车危险驾驶行为评价指标体系，然后依据指标体系设计基于系统聚类的营运客车风险等级划分方法。

1 数据采集及预处理

1.1 数据采集及预处理

数据采集自车载 GPS 定位系统，采集某公司 35 辆宇通客车司机 60 天的驾驶数据，共 26 万余条数据，每个数据都包括车辆位置、车辆速度、车辆信号等参数。利用 Quantum GIS 软件，提取 GPS 数据中经纬度数据，投影至高德地图，确定轨迹覆盖武汉市区。

从数据来看，行进速度小于 30 km/h 的样本行驶时长占比为 90.61%，小于城市道路限速，故不分析超速行为。6:00—8:00、16:00—17:00、23:00—24:00 的样本数占总体比例为 30.9%、33.2%、8.6%，故重点分析这三个时间段。

原始数据存在数据缺失和不合理现象，需要对全部数据进行预处理。（1）删除无效数据，包括 ACC 状态为 0 的数据、与驾驶行为识别算法无关的数据（OBD 时间、海拔高度、OBD 速度、卫星数等）。（2）删除缺失严重数据、冗余数据，

1. 信阳学院 河南信阳 464000

[基金项目] 信阳学院校级项目“基于 DE-SVM 的驾驶分心行为识别算法研究”（2023-XJLYB-001）

包括采样时间间隔远大于车辆运行行为发生周期的数据、重复数据、经纬度偏移的飘点数据、车速超过 120 km/h 的数据。（3）填补缺失数据。采样时间间隔在 2 ～ 5 s 范围内的数据片段进行分段线性插值处理，计算对应加速度或角速度的插值。

1.2 驾驶行为特征选择

城市路况下连续急变速是营运客车的常见行为。营运客车由于自身惯性大，发动机动力性偏弱，司机不得已采取连续急减速、连续急加速的方式[12]。为了识别连续急变速对驾驶安全性的影响，将采集样本中的急加速、急减速特征分解成单次急加速、单次急减速、连续急加速、连续急减速。35 位司机的驾驶数据的急加速、急减速驾驶行为分布情况如图 1、图 2 所示。

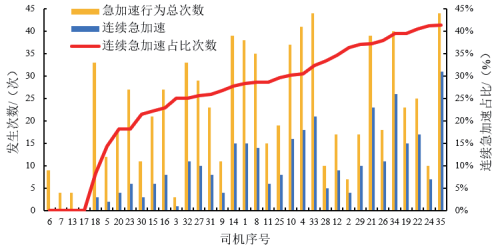


图 1 急加速样本分布情况

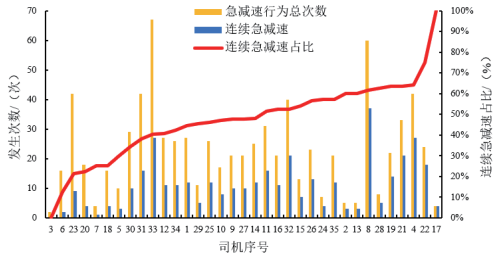


图 2 急减速样本分布情况

从图 1 和图 2 中可以看出，基于连续急变速行为在个体差异、数量分布、变速特性方面的显著差异，本文构建城市路况下营运客车危险驾驶行为的指标为：急刹车、急转弯、单次急加速、单次急减速、连续急加速、连续急减速。

驾驶数据的“加速度 - 速度”分布如表 1 所示。

表 1 “加速度 - 速度”分布

加速度区间	速度区间			
	$0 \le v \le 10$	$10 < v \le 20$	$20 < v \le 30$	$30 < v$
(4, +)	0.10%	0.02%	0.02%	0.01%
(3, 4]	0.10%	0.05%	0.02%	0.01%
(2.5, 3]	0.11%	0.05%	0.02%	0.01%
[-2.5, 0) (0. 2.5]	22.83%	42.82%	24.34%	9.16%
[-3, -2.5)	0.00%	0.01%	0.03%	0.05%
[-4, -3)	0%	0.01%	0.03%	0.06%
(-, -4)	0%	0.01%	0.02%	0.10%
速度分布	23.14%	42.98%	24.49%	9.39%

根据表 1 确定急变速行为的加速度阈值，v 为采集初始车速，V 为采集结束车速，a 为采集初始加速度，T 为采集持续时间，ω 为角速度，Δ 为角度变化。另外，急刹车、急转弯阈值界定参考文献 [13]，具体阈值设定见表 2。

表 2 危险驾驶行为评价指标及阈值设定

指标	阈值
急刹车	$30\ km \cdot h^{-1} < v < 90\ km \cdot h^{-1}$、$a < -3\ m \cdot s^{-2}$、$T \le 5\ s$、$V = 0\ km \cdot h^{-1}$
急加速	$a \ge 3\ m \cdot s^{-2}$、$T \le 5\ s$
急减速	$a \le -3\ m \cdot s^{-2}$、$T \le 5\ s$；② $40\ km \cdot h^{-1} \le v$、$a \le -2.5\ m \cdot s^{-2}$、$1\ s \le T \le 4\ s$、$V \ne 0\ km \cdot h^{-1}$
急转弯	$0 < v < 30\ km \cdot h^{-1}$、$\omega \ge 1.57\ rad \cdot s^{-1}$；② $v \ge 30\ km \cdot h^{-1}$、$\Delta \ge 31°$ 、$5\ s \ge T \ge 3\ s$

2 驾驶行为特征权重设计

危险驾驶行为的风险来自两个方面：发生事故的可能性以及可能造成的严重程度[14]。因此，考虑不确定性和严重性程度设计权重。风险管理领域普遍采用经济损失衡量事故的严重度。文献 [15] 基于事故经济损失数据构建营运客车损失后果评价指数，文献 [9] 构建了城市路况下急加速、急减速、急转弯、急刹车特征变量的回归系数，二者都基于交通事故损失金额设计相对严重程度。本文结合熵权法和事故严重程度权重，进行特征权重的优化。

2.1 基于熵权法的权重设计

基于危险驾驶行为的发生频率采用熵权法设计权重，具体步骤如下。

（1）数据标准化处理。标准化矩阵 $\boldsymbol{Z}$ 中元素 z_{ij} 计算公式为：

$$z_{ij} = \frac{x_{ij} - Min\{x_{1j},...,x_{nj}\}}{Max\{x_{1j},...,x_{nj}\} - Min\{x_{1j},...,x_{nj}\}} \quad (1)$$

式中：x_{ij} 为第 i 个样本的第 j 个特征的值。

（2）将每个指标 j 下数值相同的样本定为一个微观状态，第 j 项指标下第 i 个微观状态所含样本数记为 n_{ij}。

（3）计算第 j 项指标下第 i 个微观状态所占比重 P_{ij}。

$$P_{ij} = \frac{n_{ij}}{n}(j = 1,2,...,m) \quad (2)$$

（4）计算每个指标的信息熵 e_j，归一化，得到每个指标的熵权 W_j。

$$e_j = -\frac{1}{\ln n_j}\sum_{i=1}^{n} P_{ij}\ln(P_{ij}) \quad (3)$$

$$W_j = \frac{1 - e_j}{\sum_{j=1}^{m}(1 - e_j)} \quad (4)$$

按上述熵权法公式分别计算以“急刹车、急转弯、单次急加速、单次急减速、连续急加速、连续急减速”和“急刹车、急转弯、急加速、急减速”两种特征下对应权重，其对比数据如表 3 所示。

表 3 熵权法权重对比

六评价指标	权重	四评价指标	权重
急转弯	0.211 331 65	急转弯	0.321 230 75
急刹车	0.359 027 07	急刹车	0.545 732 43
连续急减速	0.101 942 83	急减速	0.058 203 61
连续急加速	0.099 963 46		
单次急减速	0.110 903 76	急加速	0.074 833 21
单次急加速	0.116 831 23		

从表 3 中可以看出，四指标体系中急刹车权重为 0.546，而在六指标体系中急刹车权重为 0.359，四指标的权重分布中，单个指标权重过于集中，会导致出现危险驾驶行为的误判。综上所述，六指标体系下的权重分布较优。

2.2 基于损失强度的权重设计

文献 [15] 构建了营运客车的四指标体系行为严重度权重，文献 [9] 构建了城市路况下“四急”行为与风险的回归模型，回归系数反映特征的重要程度。参考上述观点，构建城市路况下营运客车急减速、急转弯、急加速、急刹车的严重度权重系数，见表 4。

表 4 损失强度及权重换算

指标	风险指 [16]	回归系数	损失强度指数	换算权重
急刹车	5	−1.957	7	0.333 333 33
急转弯	5	−1.816	8	0.380 952 38
急加速	2	−1.880	2	0.095 238 09
急减速	5	−2.707	4	0.190 476 19

本文基于六指标体系，还需要确定连续急加速和连续急减速两个特征的权重。连续急加速和连续急减速两个特征权重大小，采用层次分析法，以驾驶数据中连续急加速和连续急减速行为发生次数为依据，建立对应权重关系。

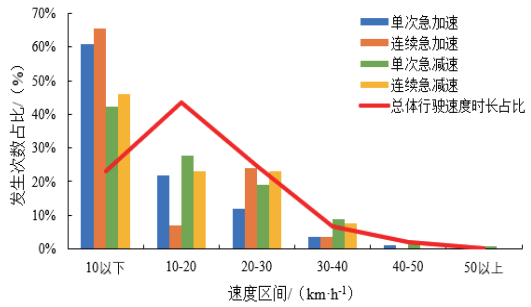


图 3 急变速行为速度区间分布

从图 3 可知，单次急加速行为集中在速度 0 ～ 20 km/h 区间，连续急加速行为集中在速度 0 ～ 30 km/h 区间。单次急变速与连续急变速的速度区间存在显著差异。从表 5 可知，连续急变速行为时长分布大部分位于 2 ～ 3 s 时间段内。综上所述，单次急变速和连续急变速在速度、时长两个因素存在差异。

表 5 连续急变速行为时长对比

持续时长		1 s	2 s	3 s	4 s	5 s
连续急加速	次数	15	189	72	25	5
	占比	4.90%	61.76%	23.53%	8.17%	1.63%
连续急减速	次数	13	235	88	13	0
	占比	3.72%	67.34%	25.21%	3.72%	0.00%

以此构建判断矩阵，用于统计驾驶数据中单次急变速和连续急变速次数。总共有四个评价因素单次急加速、连续急加速单次急减速、连续急减速，共有五位评委打分，采用 1 ～ 4 分标度法，设计的判断矩阵如表 6 所示。

表 6 急变速行为判断矩阵

	单次急加速	连续急加速	单次急减速	连续急减速
单次急加速	1	1/2	1/3	1/4
连续急加速	2	1	1/2	1/3
单次急减速	3	2	1	1/2
连续急减速	4	3	2	1

将上述判断矩阵，以及驾驶数据导入 SPSS，计算得到四个因素的权重值 q_j 见表 7，判断矩阵的最大特征值为 4.031，平均随机一致性指标（RI）值为 0.89，一致性比例（CR）值为 0.011 6 < 0.1，因此满足一致性检验，计算权重具有一致性。

表 7 基于层次分析法的权重对比

六评价指标	权重	四评价指标	权重
急转弯（θ_1）	0.333 333 333	急转弯（θ_1）	0.333 333 333
急刹车（θ_2）	0.380 952 381	急刹车（θ_2）	0.380 952 381
连续急减速（θ_3）	0.027 265 766	急减速（θ_3）	0.095 238 095
连续急加速（θ_4）	0.045 737 277		
单次急减速（θ_5）	0.079 190 495	急加速（θ_4）	0.195 238 095
单次急加速（θ_6）	0.133 506 462		

2.3 优化后的权重设计

本文结合熵权法和层次分析法，对特征权重进行优化设计。为了避免熵权法中只考虑到发生次数频率为权重计算依据，同时层次分析法有一定的主观性，专家意见难以统一的缺点，结合两种方法的优缺点，对权重综合考量，设计新的权重公式见式（5），所求权重见表 8。

$$Q_j = \frac{W_j \times q_j}{\sum_{j=1}^{m} W_j \times q_j} \tag{5}$$

表 8 AEW-AHP 综合权重对比

六评价指标	AEW-AHP 权重	四评价指标	AEW-AHP 权重
急转弯（θ_1）	0.294 808 481	急转弯（θ_1）	0.319 850 151
急刹车（θ_2）	0.572 393 339	急刹车（θ_2）	0.621 013 667
连续急减速（θ_3）	0.011 632 447	急减速（θ_3）	0.016 558 131
连续急加速（θ_4）	0.019 134 110		
单次急减速（θ_5）	0.036 754 965	急加速（θ_4）	0.016 558 131
单次急加速（θ_6）	0.065 276 659		

3 驾驶行为危险识别

3.1 系统聚类

本文采用系统聚类（hierarchical cluster method）对驾驶员样本数据集进行聚类分析。系统聚类是一种基于无监督学习的机器学习方法，更符合危险驾驶行为的随机性特点，计算公式为：

$$d_{ij}=\sqrt{\sum_{i=1}^{m}(x_{ik}-x_{jk})^2}\,(k=1,2,...,n) \tag{6}$$

式中：d_{ij}（欧几里德距离）表示 G_p 类中的第 i 类样本与 G_q 中第 j 类样本之间的相似度，x_{ik}、x_{jk} 分别为样本 i 和样本 j 的第 k 特征值。

3.2 基于系统聚类的聚类结果分析

利用优化的六指标体系，以及利用 AEW-AHP 综合权重对 35 名驾驶员驾驶数据进行系统聚类，根据不同聚类数取值的轮廓系数，确定最佳聚类数值，轮廓系数图如图 4 所示。使用 Python 进行聚类结果生成，聚类树状图如图 5 所示。

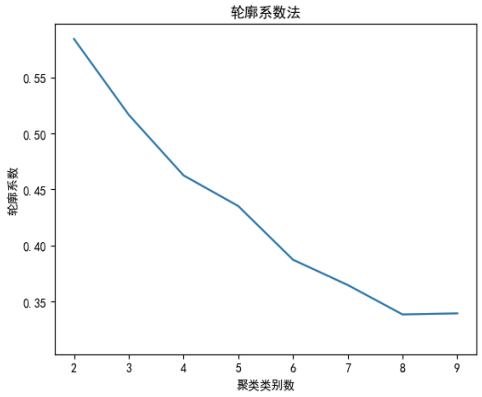


图 4 聚类的轮廓系数

由图 4 可知，将整体驾驶员样本分为两类效果最好，但在实际分类中，聚类数过少并不利于对真实数据的分类分析，因此确定最佳聚类数值为 3。

由图 5 可知，驾驶员聚类结果被分为 3 类时，其占比分别为 5.71%、11.43%、82.86%。聚类结果驾驶员驾驶行为指标均值如图 6 所示。

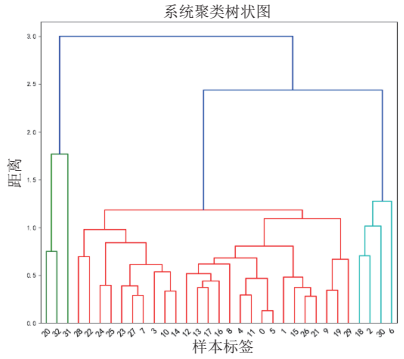


图 5 系统聚类树状图

根据图 6 中各类指标参数平均值结果，发现第一类样本的急转弯指标参数平均值是 3 类中最高的，说明此类驾驶员在转弯时，未采取减速措施直接进行转弯行为，或习惯急打方向盘以进行掉头和转弯操作，而急转弯行为极易引起车辆侧翻和侧滑事故，故将第 1 类分为高风险型。第二类样本的急刹车、单次急减速、连续急减速指标参数平均值都是 3 类中最高的，说明此类驾驶员的驾驶习惯为频繁采取激烈减速行为，城市道路交通参与者复杂，急刹车行为极易造成车辆制动失控，从而引发连环事故，而连续急减速行为易误导后车混淆行车意图，从而引发追尾事故，故将第 2 类分为中风险型。相对而言，第 3 类样本的六类指标参数平均值都是 3 类中最低，说明此类驾驶员为确保行车安全，驾驶习惯相对保守，故将第 3 类分为普通型。普通型的危险驾驶行为发生频率较低，说明普通型驾驶员在驾驶过程中对安全车速范围控制准确，保持速度稳定的能力较强，而中风险型驾驶员对车速控制能力最差。不同风险类型驾驶员的驾驶行为习惯特性分类结果也验证了本文提出的六指标体系对于提高驾驶员驾驶风险划分准确度是有效的。

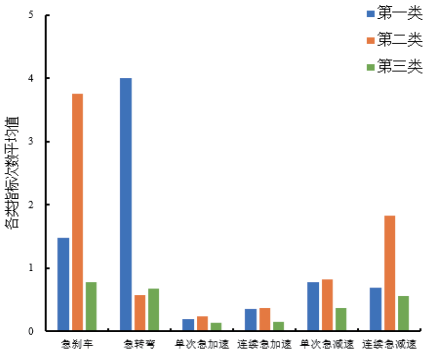


图 6 各类指标参数平均值

4 聚类结果分析对比

4.1 不同指标体系下聚类结果分析

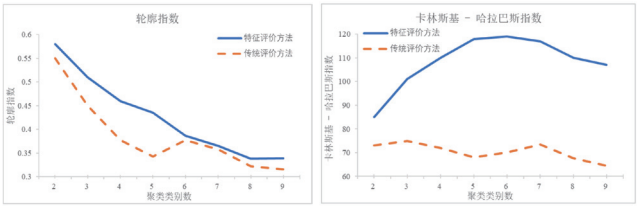
为进一步分析六指标体系的聚类结果，将聚类结果与四指标体系的聚类结果进行对比。两种聚类方法除了是否考虑细分急变速行为影响计算输入参数外，样本数据、特征指标与聚类方法均保持一致，对比结果如表 9 所示。

表 9 两种指标的驾驶员聚类结果对比

指标体系	六指标体系			四指标体系		
驾驶员类型	高风险	中风险	普通	高风险	中风险	普通
人数占比 /%	5.71	11.43	82.86	5.71	28.57	65.71
样本编号	22、34	3、7、20、32	0、1、2、4、5、6、8、9、10、11、12、13、14、15、16、17、18、19、20、21、23、24、25、26、27、28、29、30、31、33	22、34	0、3、7、9、13、17、20、31、32、33	1、2、4、5、6、8、10、11、12、14、15、16、18、19、21、23、24、25、26、27、28、29、30

由表 9 可知，有 29 名（82.86%）的驾驶员被两种方法归为同一类，而 6 名（17.14%）驾驶员被归为不同类型。是否细分急变速行为会显著影响聚类结果，将急加速和急减速行为细分为单次急加速、连续急加速、单次急减速和连续急减速行为后，在改进的六指标体系中将 6 名被归为中风险的驾驶员归类为普通型。这表明细分急变速行为会对驾驶员聚类结果产生显著影响，在样本数据集内部，样本特征具有更好的可比性。

系统聚类是一种无监督聚类，难以直接验证改进指标体系的准确性，故选取卡林斯基 - 哈拉巴斯指数（CHI）和轮廓系数进一步评价两种指标体系的聚类效果。CHI 定义为组间离散度与组内离散度的比值，CHI 越大，表明聚类效果越好。聚类效果比较如图 7 所示。



（a）轮廓系数图　　（b）卡林斯基 - 哈拉巴斯指数图

图 7 聚类效果比较

从图 7 可知，改进的六指标体系的聚类效果评价指标 CHI 值都优于传统四指标体系的聚类结果，基于六指标体系的驾驶员聚类结果在类内聚合和类间分离方面都有更好的表现。

5 总结

针对现有路况不细分及风险强度因素缺乏等问题，构建城市路况下营运车辆的危险驾驶行为识别指标体系；基于熵权法和层次分析法，设计频率和强度两个维度，构建城市路况下营运客车个体危险驾驶行为评估方法，使用系统聚类法对群体样本聚类，较好地解决危险驾驶行为风险等级划分标准的问题。聚类结果证明：相较于四指标体系，基于频率和强度特征的六评价指标体系，其分类特征与实际风险特征吻合，在实现风险精细化管理过程中具有较强的实际应用价值。局限性是样本数量偏少、采用主观意见定义风险的损失强度特征，均会影响结果的准确性；扩大样本规模，优化算法中风险强度特征描述的量化指标，是下一步的研究方向。

参考文献：

[1] 郑恒杰，熊昕，张上．基于车联网数据挖掘的驾驶员行为分析 [J]. 信息通信 ,2019(8):52-55.

[2] 韩天园，田顺，吕凯光，等．基于文本挖掘的重特大交通事故成因网络分析 [J]. 中国安全科学学报 ,2021,31(9):150-156.

[3]MESKEN J, LAJUNEN T, SUMMALA H. Interpersonal violations, speeding violations and their relation to accident involvement in Finland[J].Ergonomics,2002,45(7):469-483.

[4] 薛清文，蒋恩明，陆键．基于轨迹数据的危险驾驶行为识别方法 [J]. 中国公路学报 ,2020,33(6):84-94.

[5] 赵炜华，边浩毅，王丽华，等．大客车驾驶人高速公路行车安全意识评价 [J]. 公路与汽运 ,2013(6):72-78.

[6] 梁秉毅，朱伟，张秀芸，等．基于大数据的公交驾驶员驾驶风险评估研究 [J]. 工业控制计算机 ,2018,31(9):123-125.

[7] 张志鸿．基于 OBD 数据分析的驾驶行为研究 [D]. 西安：长安大学 ,2017.

[8]CHEN S, XUE Q, ZHAO X, et al. Risky driving behavior recognition based on vehicle trajectory[J].International journal of environmental research and public health,2021,18:17-31.

[9]XIAN H, HOU Y, WANG Y, et al.Influence of risky driving behavior and road section type on urban expressway driving safety[J].Sustainability,2022,15(1):15010398.

[10] 覃文文，李欢，李武，等．货车驾驶人驾驶行为与行车安全研究进展 [J]. 交通运输系统工程与信息 ,2022,22(5):55-74.

[11] 孙川，吴超仲，褚端峰，等．基于车联网数据挖掘的营运车辆驾驶速度行为聚类研究 [J]. 交通运输系统工程与信息 ,2015,15(6):82-87.

[12] 黄海南，沈正航，徐锦强．基于无人机视频采集的车辆驾驶行为分析实验设计 [J]. 实验技术与管理 ,2023,40(10):85-90.

[13] 王雪松，徐晓妍．基于自然驾驶数据的危险事件识别方法 [J]. 同济大学学报 (自然科学版),2020,48(1):51-59.

[14] 柳鹏飞，陆见光，徐磊，等．公路货运危险驾驶行为智能预测技术研究 [J]. 汽车技术 ,2024(3):56-62.

[15] 胡立伟．基于模糊小波神经网络的营运客车运行风险评估模型研究 [J]. 安全与环境学报 ,2020,3(20):862-871.

【作者简介】

李守义（1993—），男，河南信阳人，硕士，助教，研究方向：机器学习、机器视觉。

任方涛（1992—），男，河南信阳人，硕士，助教，研究方向：水质检测技术、电磁波探测技术。

（收稿日期：2024-05-27）