

中文引用格式:乔剑锋,王亚楠,吕淑然,等. 基于 K-means 和 LCA 的自动驾驶交通事故聚类分析[J]. 中国安全科学学报, 2025, 35(7):192-200.

英文引用格式:QIAO Jianfeng, WANG Yanan, LYU Shuran, et al. Cluster analysis of autonomous driving traffic accidents based on K-means and LCA[J]. China Safety Science Journal, 2025, 35(7):192-200.

基于 K-means 和 LCA 的自动驾驶交通事故聚类分析*

乔剑锋¹副教授,王亚楠¹,吕淑然¹教授,王汀¹副教授,夏学锋²

(1 首都经济贸易大学 管理工程学院,北京 100070;

2 中国石油集团测井有限公司 国际公司,北京 100101)

中图分类号:X951

文献标志码:A

DOI: 10.16265/j.cnki.issn1003-3033.2025.07.1025

【摘要】 为了深入挖掘自动驾驶汽车(AV)道路交通事故的内在规律,仅依靠单一事故描述因素的统计分析是不够的,还需要进一步挖掘由多个因素相互作用所体现的综合潜在类别。鉴于 AV 事故数据既包含结构化信息,又包含叙事文本的特点,在类型识别过程中创新性地提出将 K-means 聚类分析与潜在类别分析(LCA)相结合的方法,首先,使用 K-means 方法从叙事文本中提取关键信息;然后,将其作为 LCA 模型的输入,克服 LCA 仅能利用现有事故报告中的结构化信息这一局限性;最后,采用美国加利福尼亚州的 437 起 AV 交通事故验证组合方法的有效性。结果表明:AV 事故主要表现为 4 个综合类型;K-means 与 LCA 的组合方法能对含叙述文本的结构化信息实施高效的聚类分析。

【关键词】 K-means; 潜在类别分析(LCA); 自动驾驶; 聚类分析; 自动驾驶汽车(AV); 交通事故

Cluster analysis of autonomous driving traffic accidents based on K-means and LCA

QIAO Jianfeng¹, WANG Yanan¹, LYU Shuran¹, WANG Ting¹, XIA Xuefeng²

(1 School of Management Engineering, Capital University of Economics and Business, Beijing 100070, China; 2 International Branch, China Petroleum Corporation Logging Limited Company, Beijing 100101, China)

Abstract: To deeply explore the underlying patterns of road traffic accidents involving Autonomous Vehicles (AV), relying solely on the statistical analysis of individual accident description factors was insufficient. It was necessary to uncover further the comprehensive latent categories reflected by the interactions of multiple factors. Given that AV accident data contained structured information and narrative text, an innovative approach was proposed for type identification combining K-means clustering analysis and LCA. Specifically, the K-means method was used to extract key information from the narrative text, which was then fed into the LCA model to overcome the limitation of LCA being able to utilize only the structured information in existing accident reports. The effectiveness of this combined approach was verified using 437 AV traffic accidents in California, USA. The results show that AV accidents mainly manifest in four comprehensive types. The combined approach of K-means and LCA enables efficient

clustering analysis of structured information that includes narrative text.

Keywords: K-means; latent class analysis (LCA); autonomous driving; cluster analysis; autonomous vehicles (AV); traffic accidents

0 引言

自动驾驶汽车 (Autonomous Vehicle, AV) 为交通系统的高效性带来许多新机遇^[1]。然而,由于其社会关注度较高,一旦发生交通事故,易成为媒体和公众关注的焦点。如 2018 年美国亚利桑那州一辆 Uber AV 在公共道路上行驶致路人死亡,2021 年,一辆中国蔚来 AV 在高速公路上自动驾驶致车主死亡^[2]。这些交通事故引发公众对 AV 安全性的质疑,制约其商业化发展。通过聚类分析挖掘 AV 事故的隐藏模式,进而剖析其特征规律与致因机理,这对于有效遏制 AV 道路交通事故的发生具有重要意义。

由于国内厂商没有公开 AV 事故,相关研究工作开展较少,目前只有针对个别案例开展事故分析^[3]。国外主要针对美国加利福尼亚州交通部门 (California Department of Motor Vehicles, CADMV) 提供的 AV 事故进行研究。在 AV 事故类型分析方面,研究人员使用归纳分析法,针对碰撞前车辆运动状态重新定义了 15 种事故类型^[4]。基于“中国深度移动安全研究-交通事故”数据库,ZHOU Rui 等^[5]使用 K-means 将事故集聚类到 6 个碰撞情景中,发现十字路口是情景 1、2、3 的主要特征,而直道、三段式交叉口和匝道分别是情景 4、5、6 的主要特征。DAS 等^[6]使用潜在类别分析 (Latent Class Analysis, LCA) 揭示了 AV 事故的 6 个隐藏类型,如自动驾驶模式下无严重程度的碰撞、AV 停车状态下的两车碰撞等。基于韩国高速公路上的 AV 交通事故,CHAE 等^[7]采用隐含狄利克雷分布 (Latent Dirichlet Allocation, LDA) 将事故集分为 4 组,通过分析不同类型中的高频术语的分布来分析事故类型特征。ALAMBEIGI 等^[8]应用 LDA 确定了 AV 碰撞的 5 个主题,如 AV 停在十字路口的追尾碰撞、传统车辆 (Conventional Vehicle, CV) 从 AV 左侧超车时的侧刮碰撞等。

可以看出,探究 AV 事故中隐藏的事故类型和分析类型特征是当前研究的热点问题,研究人员已经使用 LCA、K-means、LDA 分析 AV 事故的综合类型。LDA 是自然语言处理经常使用的无监督聚类方法,当集合中的文档在内容上具有较高的相似性,

或者文档覆盖的主题范围重叠较多时,LDA 模型难以区分不同簇间的主题差异。常用的聚类方法包括层次聚类 (Hierarchical Clustering, HC) 和基于密度的带有噪声的应用空间聚类 (Density-Based Spatial Clustering of Applications with Noise, DBSCAN)。其中,HC 使用树状图来表示聚类结果,DBSCAN 能够识别出高密度的簇,但这 2 类方法的聚类目的和结果不是文中的考虑重点,而 K-means 算法原理和实现简单,目标是将数据集划分为 K 个群组,因此,文中选择 K-means 算法,探究 AV 叙事文本聚类之间的异质性,并利用类中心分析识别同一类型下 AV 事故的共同特征。传统的聚类分析方法如 K-means 和 HC,侧重于根据观测变量的相似性将个体分组,而 LCA 旨在通过统计关联分析,揭示结构化数据中不易观察到的群体结构,所以针对 AV 事故的结构化信息选择 LCA 模型。

总之,现有研究主要集中在单一事故类型的统计分析,或者使用单一聚类方法进行 AV 事故的综合类型研究。然而,每起 AV 事故既包含结构化信息,又包含文本信息。鉴于此,笔者拟提出 K-means 和 LCA 的组合聚类方法,首次从结构变量和非结构变量 2 个角度对 AV 事故进行综合类型分析。通过 K-means 聚类提取叙事文本中的关键信息,将其与结构变量组合成新的观测变量组,再利用 LCA 分析 AV 事故类型及其变量相关性,同时,以 K-means 结果验证和补充 LCA 结果,以期更充分地反映 AV 在新技术影响下的交通安全问题。

1 AV 事故数据和变量选择

AV 事故信息来源于 CADMV。自 2014 年 9 月起,如果制造商在加利福尼亚州的公共道路上进行 AV 测试,AV 制造商必须上报其交通事故。

1.1 数据收集和预处理

文中收集 2021 年 1 月至 2024 年 6 月期间的 437 起 AV 事故报告,并进行数据预处理,将原始事故信息导入电子表格,其中一列用于存储事故的描述信息,其余列则用于存储事故的结构化信息。

1.2 变量选择和统计分析

碰撞报告的一部分信息对分析事故致因来说可

以忽略,如制造商信息和事故发生时间、地点等。表1为12个类别变量所对应的不同枚举实例的事故统计分布情况。前5个类别变量描述AV的碰撞信息,类型变量6定义CV的行驶情况。类型变量7显示事故是否涉及弱势道路使用者(Vulnerable Road User, VRU),如行人、骑自行车者和使用滑板车的人。如果AV碰撞与VRU有关,则该项为“是”。类型变量8是人员受伤程度,包括所有道路交通参与者的受伤情况。类型变量9、10、11、12分别表示天气、照明、路面情况、道路情况。

表1 AV碰撞事故分布情况
Table 1 Distribution of AV crashes

类型变量	枚举实例	总数	占比/%
1 AV 驾驶状态	行驶	251	57.4
	停止	186	42.6
2 AV 驾驶模式	自动驾驶模式	252	57.7
	传统模式	185	42.3
3 AV 碰撞前的状态	停车	176	40.3
	直行	123	28.1
	转弯	48	11
	倒车	21	4.8
	减速/减速停车	39	8.9
	变道	8	1.8
	停车/泊车	12	2.7
	其他行驶状态	10	2.3
4 AV 碰撞类型	追尾	189	43.2
	侧刮	110	25.2
	侧撞	38	8.7
	正面碰撞	37	8.5
	撞物	37	8.5
	其他类型	26	5.9
	未损坏	29	6.6
5 AV 损坏程度	轻度	332	76
	中度	65	14.9
	重度	11	2.5
6 CV 碰撞前的状态	停车	13	3
	直行	188	43
	转弯	39	8.9
	倒车	27	6.2
	减速/减速停车	10	2.3
	变道	22	5
	停车/泊车	32	7.3
	超越其他车辆	17	3.9
	其他行驶状态	89	20.4
7 VRU	否	392	89.7
	是	45	10.3
8 人员受伤程度	未受伤	396	90.6
	受伤	41	9.4

续表 1			
类型变量	枚举实例	总数	占比/%
9 天气	晴朗	364	83.3
	多云	45	10.3
	雨天	26	5.9
	雾天	2	0.5
10 照明	日光	293	67
	黎明	12	2.7
	路灯	127	29.1
	无路灯	5	1.1
11 路面情况	干燥	405	92.7
	潮湿	32	7.3
12 道路情况	坑洞	2	0.5
	障碍物	3	0.7
	狭窄	4	0.9
	其他道路情况	10	2.3
	无异常	418	95.7

注:加粗数字在文中被引用。

表1统计数据表明:① AV事故中人员受伤少并且车辆损坏程度低,90.6%的事故没有造成人员伤亡,82.6%的事故没有车辆损坏(6.6%)或损坏程度较轻(76%)。② 追尾碰撞是最常见的AV碰撞类型,占43.2%,其次是侧刮(25.2%)、侧撞(8.7%)和其他碰撞。相比文献[9],最新AV事故中追尾事故类型有所下降,反映了AV技术的进步。

2 LCA 和 K-means 模型

LCA可以有效解决结构化变量的聚类问题,通过分析显性数据之间的统计关系来识别总体样本的潜在类别。K-means可基于文本相似度,把具有相似类中心的文本聚在一起^[10]。

2.1 LCA

使用LCA模型需要经过3个步骤:模型构建、参数估计和模型拟合。该模型采用贝叶斯方法,检测显性数据中未观测到的潜在变量^[6]。假设C表示一个潜在的总体变量,LCA模型可定义为:

$$P_{O_l} = \sum_{k=1}^K P_{C_k} P_{O_l|C_k}$$

(1)

式中: $O_l(l=1,2,\cdots,L)$ 为第 l 个观测模式; P_{O_l} 为 O_l 模式在数据集中的概率; $C_k(k=1,2,\cdots,K)$ 为第 k 个潜在类别; P_{C_k} 为事故属于潜在类别 C_k 的先验概率。假设观测模式可以用 J 个独立的类型变量所表示(显变量),如表1的类型变量列,有 $J=12$,即每一起AV事故均是这12个类型变量的观察值。则有 $O_l = \{o_1, o_2, \cdots, o_j, \cdots, o_J\}$, o_j 是第 j 个变量的一

个枚举实例。不同的显变量有不同的枚举实例,如表 1 的枚举实例。假设第 j 个变量有 L_j 个枚举实例(或水平),枚举实例整形化后有 $o_j = 1, 2, \dots, L_j$ 。如,类型变量“AV 驾驶状态”有“行驶”和“停止”2 个枚举实例,模型后续处理中分别用“1”和“2”表示(枚举值)。观测模式的总数 $L = \prod_{j=1}^J L_j, P_{o_l|c_k}$ 表示联合条件概率:

$$P_{o_l|c_k} = \prod_{j=1}^J P(o_j | C_k)$$

(2)

式(1)中 P_{c_k} 和 $P_{o_l|c_k}$ 采用最大似然法估算。下式为观测模式 O_l 属于潜在类别 C_k 的概率(后验概率),表示随机抽取的一个样本(可以用 O_l 模式表示)属于某类型变量的概率。样本应被分配到概率最高的潜在类别中。

$$P_{c_k|O_l} = P_{c_k} P_{o_l|c_k} / P_{O_l}$$

(3)

LCA 模型拟合是通过对比聚类性能指标找到最合适的聚类数。性能指标量化计算的函数包括:赤池信息准则(Akaike Information Criterion, AIC)、贝叶斯信息准则(Bayesian Information Criterion, BIC)、赤池贝叶斯信息准则(Akaike's Bayesian Information Criterion, aBIC)和熵(Entropy)^[11]。AIC、BIC 和 aBIC 等信息准则的最小值和熵的最大值意味着最佳拟合,建议熵值大于 0.9^[12]。同时,LCA 模型的输入样本数应该大于 300,潜在类别应包含 5%以上的样本^[13]。

2.2 K-means

K-means 聚类目的是最小化组内距离^[14],所有类型的类内方差之和越小,表明聚类效果越好。

$$f = \min \sum_{k=1}^K v_k$$

(4)

式中: f 为目标函数; v_k 为第 k 个类的类内方差,如下式所示。

$$v_k = \sum_{x_n \in S_k} (x_n - \bar{c}_k)^2$$

(5)

式中: S_k 为第 k 个类型的样本集合; $\bar{c}_k$ 为第 k 个类

的类中心。K-means 算法包含 5 个步骤:①输入代表聚类总数的 K 值;②随机选择 K 个样本作为初始类中心;③通过计算欧氏距离,将每个数据样本分配到与其最接近的类中心;④计算每个聚类中所有样本的平均值并以此更新类中心;⑤重复步骤③和④,直到收敛。

初始类中心的选择对 K-means 的聚类结果有显著影响,为此,学者们提出 K-means++ 算法^[15],使初始类中心尽可能地分散,从而提高聚类的稳定性和准确性。聚类性能的衡量指标选用轮廓系数(Silhouette),取值范围在-1 和 1 之间,数值越大表示聚类效果越好。聚类试验采用开源 Python IDE,如自然语言工具包和 scikit-learn 等函数库。叙事文本采用词频-逆文本频率(Term Frequency Inverse Document Frequency, TFIDF)表示^[16]。

3 AV 事故聚类分析及安全提升建议

分别给出 K-means 聚类结果和基于 K-means 抽取关键特征的 LCA 聚类结果;通过讨论认识 K-means 和 LCA 组合方法的优势和不足;最后,基于聚类结果给出 AV 的通行效率和安全提升建议。

3.1 K-means 聚类分析

采用 K-means++ 算法,可以将 AV 事故的描述文本分割为 3 个聚类。当 $k=3$ 时 Silhouette 取最大值 0.015 1。表 2 为 3 个类中心的 TFIDF 值较大的前 15 个特征词。聚类 1 包含 203 个样本,主要指“多车道下 AV、CV 和 VRU 混行情况下的 AV 交通事故”。事发前 AV 主要驾驶意图是右转,道路交通的典型参与者有大卡车或自行车。聚类 2 包含 156 个样本,主要指“在交叉路口当 AV 停止时发生的追尾事故”。AV 在交叉路口遇到信号指示灯的情况下,AV 发生的追尾事故。聚类 3 包含 79 个样本,主要指“当 AV 或 CV 处于停车状态,而 AV 或 CV 倒车时发生的侧刮”。由此可知:主要事故类型是侧刮,事故发生场地是停车场。

表 2 K-means 的聚类中心

Table 2 Clustering centers for K-means

聚类 1(203 个样本)		聚类 2(156 个样本)		聚类 3(79 个样本)	
特征词	TFIDF	特征词	TFIDF	特征词	TFIDF
车道 (lane)	0.089 2	停止 (stop)	0.103 4	停车 (park)	0.194 9
右 (right)	0.064 6	后部 (rear)	0.099 4	侧面 (side)	0.110 0
转弯 (turn)	0.060 5	保险杠 (bumper)	0.092 6	手动 (manual)	0.073 1
大卡车 (truck)	0.055 5	灯 (light)	0.091 4	前部 (front)	0.065 9

续表 2

聚类 1(203 个样本)		聚类 2(156 个样本)		聚类 3(79 个样本)	
停止(stop)	0.051 0	交叉路口(intersection)	0.074 6	镜子(mirror)	0.063 4
前(front)	0.048 0	自动(autonomous)	0.067 2	倒车(reverse)	0.062 2
手动(manual)	0.047 0	在...之后(behind)	0.064 2	平行(parallel)	0.049 1
侧面(side)	0.044 1	红(red)	0.061 7	场地(lot)	0.047 5
后部(rear)	0.039 8	轻微(minor)	0.054 1	轻微(minor)	0.046 9
在...上(onto)	0.037 2	交通(traffic)	0.052 4	卷入(involve)	0.044 7
自动(autonomous)	0.037 0	靠近(approach)	0.043 5	后部(rear)	0.043 1
卷入(involve)	0.035 0	英里每小时(mph)	0.039 0	右(right)	0.042 9
街(boulevard)	0.033 0	转弯(turn)	0.038 1	自动(autonomous)	0.039 4
自行车(cyclist)	0.030 8	卷入(involve)	0.037 8	空间(space)	0.039 0
在...之后(behind)	0.030 5	在...之前(front)	0.036 1	马路牙子(curb)	0.037 9

注:聚类 1 表示第 1 个聚类,以此类推;加粗字体代表类型的主要特征,并在文中引用;表 3 同。

通过类中心分析,进一步抽取关键类型信息。关键类型信息的选择标准是特征词具有类型特征并且在大部分事故均有相关描述。例如:聚类 2 的特征词有“交叉路口”。通过对所有事故的分析,归纳出事故发生的典型地点包括交叉路口、正常道路、停车场、路边停车位、道路汇入汇出口等。聚类 1 的特征词有“车道”,车道指 AV 行驶在哪个车道。通过分析事故描述信息,发现 64.3% 的事故没有明确说明事故发生时 AV 所处的车道信息,所以忽略此类型信息。忽略的特征词还包括:①结构化信息中已有的信息,如聚类 1 的“停止”、“转弯”;②介词,如聚类 1 的“前”、“在...之上”;③普通常用且没有类型特征含义的词,如聚类 1 的“卷入”、聚类 2 的“交通”。综上,在 AV 叙事文本中仅抽取事故发生地点作为后续综合类型分析的新增输入变量。

3.2 K-means 和 LCA 组合方法的聚类分析

基于 3.1 节抽取的事故发生地点和原始选定的 12 个类型变量,采用 LCA 模型进行聚类分析。聚类数设定为 2-8,性能指标曲线如图 1 所示。聚类数分别为 2、4、7 时,BIC 值为最低(9 307)、aBIC 值为最低(8 892)、AIC 值为最低(8 689),即 3 个聚类结果的拟合效果均可以接受。但当聚类数为 7 时,个别潜在类别的样本数不足总样本的 5%,不足以拟合 LCA 模型;聚类数为 2 时,聚类 1 表示“AV 在行驶时发生的撞击事故”;聚类 2 表示“AV 停车时的追尾碰撞事故”。相比 2 个聚类,4 个聚类划分颗粒度更细,类型信息更具体,各观测变量的聚类响应占比见表 3。表中 4 个聚类各自包含的样本占比均高于 5%,样本可以拟合 LCA 模型,同时熵值为最大值 0.954(大于 0.9)。

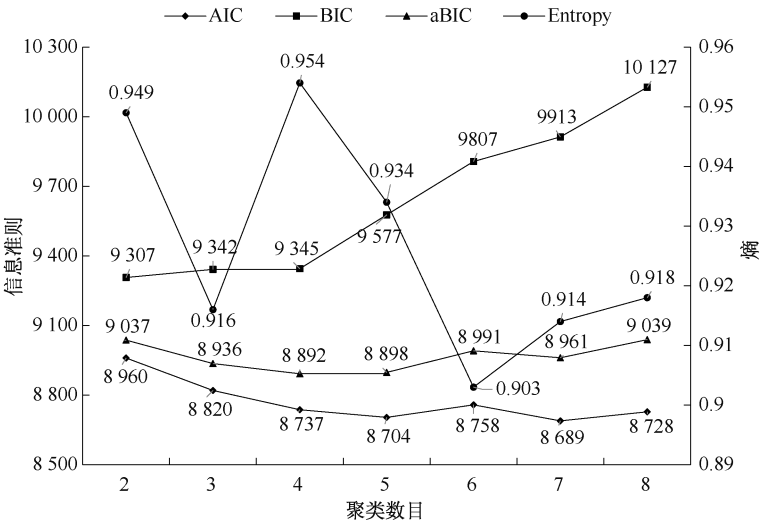


图 1 不同聚类数目的性能指标值

Fig. 1 Performance index values for different numbers of clusters

聚类1包含170例事故(占38.9%),主要指“AV停止时的追尾事故”。事故发生前100%的AV处于停止状态,59.7%的事故是追尾事故,59.1%的CV在直行,52.5%的事故发生在交叉路口。

聚类2包含158例事故(36.2%),主要指“AV行驶时发生的追尾或侧撞事故”。典型场景是在汇入汇出路口(17.1%),AV减速(20.5%)、转弯(21.7%)或变道(4.4%),CV直行(52.9%)或转弯(16%)时,发生的侧撞事故(13.3%)。事故发生地点也包括交叉路口(51.4%)或直行道路(28.1%),AV也可能在直行(43.2%)。

聚类3包含85例事故(占19.4%),主要指“AV行驶时发生的侧刮和撞物事故”。典型场景包括2个:①在停车场(14.1%)或路边停车位(17.4%),AV倒车(22.4%)CV停车(11%)时发生的侧刮事故(30.9%);②在直行道路(45.2%)AV直行(48%)发生的撞物事故(27.7%)。

聚类4包含24例事故(占5.5%),主要指“AV在复杂路况以及复杂汇车行为时发生的碰撞事故”。AV主要处于自动驾驶模式(70.6%),事故容易发生在交叉路口(53.6%)或汇入/汇出路口(12.3%)。典型场景包括2个:AV直行(50.2%)CV变道(12.4%)或转弯(16.6%)时发生的正面碰撞(16.6%)和AV直行发生的撞物事故(20.8%)。与聚类3的撞物场景不同,聚类4的事故路面潮湿(100%)或在路灯下行驶(62.5%)。

3.3 基于 K-means 和 LCA 组合方法的讨论

为充分认识 K-means 和 LCA 的组合方法,从5个方面进行讨论:

1) K-means 聚类结果是 LCA 聚类结果的验证和补充。AV 事故的叙事文本是结构化信息的一种重复表达或理清关键变量关系的一种方式,所以 K-means 聚类结果是 LCA 聚类结果的一种有效验证手段。如 K-means 发现的聚类2“在交叉路口当 AV 停止时发生的追尾事故”,与 LCA 聚类1“AV 停止时的追尾事故”的类型描述一致;K-means 发现的聚类3“当 AV 或 CV 处于停车状态,而 AV 或 CV 倒车时发生的侧刮”,与 LCA 聚类3的典型场景一致(在停车场或路边停车位,AV 倒车 CV 停止时发生的侧刮事故)。AV 叙事文本也可以理解为是对结构信息的有益补充,所以 K-means 聚类结果也是 LCA 聚类结果的补充。如 K-means 发现的聚类1“多车道下 AV、CV 和 VRU 混行情况下的 AV 交通事故”是 LCA 聚类结果的补充,因为 LCA 聚类中没有添加

AV/CV 行驶的车道情况,所以在 LCA 聚类结果中没有此类信息。

2) K-means 和 LCA 组合方法弥补了 LCA 模型的缺陷。使用原始12个类型变量,采用 LCA 也可以给出4个聚类结果。其结构变量的聚类响应占比与基于13个类型变量的聚类结果大体一致。但采用组合聚类分析方法,可以添加事故发生地点这一重要输入变量,弥补 LCA 模型不足。基于13个类型变量的聚类结果包含较全的关键信息,使聚类结果信息更丰富。例如聚类1的事故发生在交叉路口,聚类2的事故发生在汇入汇出路口、交叉路口或直行道路,聚类3的事故发生在停车场或路边停车位,聚类4的事故发生在交叉路口和汇入汇出路口。

3) 基础聚类方法建议选择 LCA 模型。LCA 优点是选取的结构变量均是 AV 事故的关键信息,聚类结果易于解释,因为每一个枚举实例的含义具有明确的所指对象。缺点是使用模型前必须确定类型变量,并指定类型变量的枚举实例范围。如在试验前没有考虑某一重要类型变量,LCA 的分类结果是不能反映该类型变量对类型的影响。而 K-means 不会遗漏关键信息,通过 TFIDF 进行文本向量化表示后,重要的类型特征信息会充分表现出来。在高维空间中,由于基于欧氏距离的文本相似度值趋于一致,K-means 聚类算法的性能会显著下降。通过分析 K-means 的类中心可以获得类型特征,但特征信息所指对象并不明确。例如“倒车”不能说明是 AV 倒车还是 CV 倒车。所以组合方法不以 K-means 聚类为基础,而通过 K-means 抽取关键信息。

4) K-means 和 LCA 都假设变量相互独立,缺少变量之间相互影响的分析。为弥补上述缺陷,下面基于 LCA 的聚类结果,主观分析类型变量之间的相互影响关系。表3中,聚类1表明 AV 停止,影响后车 CV 的直行,从而导致追尾事故;聚类2表明 AV 行驶缓慢,影响后车 CV 的直行或转弯,从而导致侧撞类等事故;聚类3表明 AV 倒车导致的侧刮类事故,以及在自动驾驶模式下 AV 直行或转弯时发生的撞物类事故;聚类4表明路面潮湿或灯光不足的环境下容易导致 AV 事故。

5) LCA 模型面临样本数少、观测类型多的挑战。首先,LCA 需要足够的样本量来确保每个潜在类别都有足够的观测数据。NYLUND-GIBSON 等^[17]建议理想情况下应该有300~1000个案例,文中符合总体样本量的需要。如果潜在类别样本量少,导致无法充分捕捉其特征,可采用的规避策略是

表 3 结构变量的聚类响应占比

Table 3 Proportion of clustered responses for structural variables

%

类型变量	实例	聚类 1	聚类 2	聚类 3	聚类 4
1 AV 驾驶状态	行驶	0	95.4	99	66.8
	停止	100	4.6	1	33.2
2 AV 驾驶模式	自动驾驶模式	66.6	58.7	34.2	70.6
	传统模式	33.4	41.3	65.8	29.4
3 AV 碰撞前的状态	停车	90.3	8.6	0	37.3
	直行	1	43.2	48	50.2
	转弯	1.6	21.7	10.6	8.3
	倒车	1.2	0	22.4	0
	减速/减速停车	0	20.5	6.6	4.2
	变道	0.6	4.4	0	0
	停车/泊车	5.3	0	3.5	0
	其他行驶状态	0	1.6	8.8	0
4 AV 碰撞类型	追尾	59.7	44.6	11.7	29.4
	侧刮	22.7	25.4	30.9	20.8
	侧撞	8.2	13.3	0	12.5
	正面碰撞	7	10.2	5.8	16.6
	撞物	0	5.4	27.7	20.8
	其他类型	2.3	1.1	23.9	0
5 AV 损坏程度	未损坏	5.9	5.2	11.5	4
	轻度	76.9	75.3	76.7	71
	中度	14.8	17.2	9	20.8
	重度	2.3	2.3	2.9	4.2
6 CV 碰撞前的状态	停车	0	2.3	11	0
	直行	59.1	52.9	0	16.8
	转弯	5.7	16	0	16.6
	倒车	13	0	2.4	12.3
	减速/减速停车	4.7	1.3	0	0
	变道	2.9	8.9	0	12.4
	停车/泊车	1.2	0	34.1	4.4
	超越其他车辆	6.5	2.5	2.4	0
	其他行驶状态	7	16.2	50.1	37.4
7 VRU	否	89.7	88	90	100
	是	10.3	12	10	0
8 人员受伤程度	未受伤	88.2	88.9	96.1	100
	受伤	11.8	11.1	3.9	0
9 天气	晴朗	91.3	90.2	78	0
	多云	6.8	8.6	22	5.2
	雨天	1.3	0.6	0	94.8
	雾天	0.6	0.6	0	0
10 照明	日光	69.5	68.1	68.6	37.5
	黎明	2.9	1.3	5.8	0
	路灯	27.5	27.5	25.6	62.5
	无路灯	0	3.2	0	0
11 路面情况	干燥	97	100	96.5	0
	潮湿	3	0	3.5	100

续表 3

类型变量	实例	聚类 1	聚类 2	聚类 3	聚类 4
12 道路情况	坑洞	0	0	2. 4	0
	障碍物	0	0	1. 2	8. 3
	狭窄	0. 6	0. 6	2. 3	0
	其他道路情况	1. 2	0	9. 4	0
	无异常	98. 2	99. 4	84. 7	91. 7
13 发生地点	十字路口/交叉路口	52. 5	51. 4	23. 3	53. 6
	直行道路	21. 8	28. 1	45. 2	25. 9
	停车场	3. 5	1. 3	14. 1	8. 2
	汇入汇出路口	6. 4	17. 1	0	12. 3
	路边停车位	15. 8	2. 1	17. 4	0

使潜在类别样本数占比高于 5%。不同于自动分类(有监督学习),样本少会导致含样本少的类型训练不充分从而导致分类精度低;文中求解无监督聚类问题,样本集具有一定的统计意义即可使用。其次,观测类型变量多会增加模型训练复杂度。为此,文中删除一些观测变量,如厂商名,为了更好地分析 AV 事故的共性特征。试验结果表明:现有 13 个类型变量的聚类结果较为稳定。由于贝叶斯估计利用先验信息来引导参数估计,特别是在样本量较小的情况下,也可以提供稳定、可靠的参数估计。最后,观测变量多使综合类型的命名难以覆盖全部类型变量所表达的信息。文本使用事故占比较大或相比其他类型采用较突出的类型枚举实例对综合聚类进行命名,如聚类 1 是“AV 停止时的追尾事故”,其中 AV 停止的事故数占比 100%,追尾事故类型突出占比 59.7%。

3.4 AV 的通行效率和安全提升建议

结合 K-means 和 LCA 的聚类响应结果,针对 AV 的通行效率和安全提升方面提出 4 个改进建议:

- 1) AV 需要预测周边所有其他道路交通使用者的驾驶行为,通过综合评判,调整其自身行驶策略。聚类 1 表明:AV 缺少预测后车驾驶行为,由于提前制动或制动太快导致后车追尾。
- 2) AV 需要快速给出路径规划并快速通行,平衡礼让行为和行驶效率,避免不必要的减速。聚类 2 表明:AV 行驶速度低导致其侧刮、追尾类等事故。
- 3) AV 需要准确地识别停车位和周围车辆,并提升自动泊车技术。聚类 3 表明:AV 倒车或泊车

时容易与周边车辆发生侧刮事故。

4) AV 传感器技术需进一步提升。聚类 3 表明:AV 直行时对狭窄路面和障碍物的监测不足,尤其是转弯情况下,需要提前监测转弯方向的道路和环境,以免撞物事故的发生;同时聚类 4 表明:AV 在灯光较暗情况下对障碍物的监测性能有所下降^[18],同时对特殊环境(如湿滑路段)的监测感知不足。

4 结 论

1) 在聚类分析中,K-means 和 LCA 的组合方法弥补了 LCA 模型仅能使用结构变量的缺陷。采用 K-means 对叙事文本进行聚类分组,并通过类中心特征分析识别出新的类型变量,有助于丰富 LCA 模型的信息输入,使综合类型包含 AV 事故的关键信息。

2) 聚类分析揭示了 AV 道路交通事故的类型及其成因规律。在交叉路口,AV 因制动较快而易被紧随其后的 CV 追尾;在汇入汇出路口、交叉路口或直行道路上,AV 行驶速度较慢,而 CV 驾驶员操之过急,导致追尾或侧撞事故;在停车场或路边停车位,AV 在倒车或直行时因环境监测和路径规划技术不成熟,易发生侧刮或撞物事故;在复杂路况及复杂汇车行为场景下,AV 的障碍物检测和规避技术不足,从而引发碰撞事故。

3) K-means 和 LCA 组合方法的局限是假设事故描述因素独立,后续可以深入研究因素的相互影响关系^[19]以及因素的逻辑时序关系等^[20]。另外,观测类型变量多会增加模型参数估计的复杂度,影响模型聚类结果的可靠性,并增加综合类型的命名难度,后续可以结合降维技术继续分析聚类结果。

参 考 文 献

[1] 赵晓华, 陈浩林, 李振龙, 等. 不同情景下自动驾驶接管行为的影响特征[J]. 中国公路学报, 2022, 35(9): 195-214.

- ZHAO Xiaohua, CHEN Haolin, LI Zhenlong, et al. Influence characteristics of automated driving takeover behavior in different scenarios[J]. China Journal of Highway and Transport, 2022, 35(9):195-214.
- [2] 郭延永, 刘佩, 袁泉, 等. 网联自动驾驶车辆道路交通安全研究综述[J]. 交通运输工程学报, 2023, 23(5): 19-38.
- GUO Yanyong, LIU Pei, YUAN Quan, et al. Review on research of road traffic safety of connected and automated vehicles[J]. Journal of Traffic and Transportation Engineering, 2023, 23(5):19-38.
- [3] 陈吉清, 翁楚滨, 兰凤崇. 智能车辆换道潜在冲突分析与风险量化方法[J]. 汽车工程, 2021, 43(11): 1 565-1 576, 1 586.
- CHEN Jiqing, WENG Chubin, LAN Fengchong. Potential conflict analysis and risk quantification method of intelligent vehicle lane change[J]. Automotive Engineering, 2021, 43(11):1 565-1 576, 1 586.
- [4] LIU Qian, WANG Xuesong, WU Xiangbin, et al. Crash comparison of autonomous and conventional vehicles using pre-crash scenario typology[J]. Accident Analysis and Prevention, 2021, 159;DOI:10.1016/j.aap.2021.106281.
- [5] ZHOU Rui, HUANG Helai, LEE J, et al. Identifying typical pre-crash scenarios based on in-depth crash data with deep embedded clustering for autonomous vehicle safety testing[J]. Accident Analysis and Prevention, 2023, 191;DOI: 10.1016/j.aap.2023.107218.
- [6] DAS S, DUTTA A, TSAPAKIS I. Automated vehicle collisions in California: applying Bayesian latent class model[J]. IATSS Research, 2020, 44(4):300-308.
- [7] CHAE Q, KIM J, JANG J, et al. Development of risk-situation scenario for autonomous vehicles on expressway using topic modeling[J]. Journal of Advanced Transportation, 2022, 2022:1-19.
- [8] ALAMBEIGI H, MCDONALD A D, TANKASALA S R. Crash themes in automated vehicles: a topic modeling analysis of the California department of motor vehicles automated vehicle crash database[J]. arXiv: Applications, 2020;DOI: 10.48550/arXiv.2001.11087.
- [9] GOODALL N J. Comparison of automated vehicle struck-from-behind crash rates with national rates using naturalistic data[J]. Accident Analysis and Prevention, 2023, 154;DOI:10.31224/osf.io/vbrx5.
- [10] 杨文臣, 周燕宁, 田毕江, 等. 基于聚类分析和 SVM 的二级公路交通事故严重度预测[J]. 中国安全科学学报, 2022, 32(5):163-169.
- YANG Wenchen, ZHOU Yanning, TIAN Bijiang, et al. Traffic accident severity prediction for secondary highways based on cluster analysis and SVM model[J]. China Safety Science Journal, 2022, 32(5):163-169.
- [11] QIAO Jianfeng, LI Yu. Resource leveling using normalized entropy and relative entropy[J]. Automation in Construction, 2018, 87:263-272.
- [12] LIN Zijing, FAN Wei. Exploring bicyclist injury severity in bicycle-vehicle crashes using latent class clustering analysis and partial proportional odds models[J]. Journal of Safety Research, 2020, 76:101-117.
- [13] WELLER B E, BOWEN N K, FAUBERT S J. Latent class analysis: a guide to best practice[J]. Journal of Black Psychology, 2020, 46(4):287-311.
- [14] 杨慧敏, 石琴, 陈一锴, 等. 基于混合聚类的农村公路单车事故影响因素分析[J]. 中国安全科学学报, 2020, 30(8):129-136.
- YANG Huimin, SHI Qin, CHEN Yikai, et al. Influencing factors for single vehicle accidents on rural highways based on hybrid clustering approach[J]. China Safety Science Journal, 2020, 30(8):129-136.
- [15] ARTHUR D, VASSIIVITSKII S. K-means++: the advantages of careful seeding[C]. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, 2007: 1 027-1 035.
- [16] QIAO Jianfeng, WANG Changfeng, GUAN Shuang, et al. Construction-accident narrative classification using shallow and deep learning[J]. Journal of Construction Engineering and Management, 2022, 148(9);DOI:10.1061/(asce)co.1943-7862.0002354.
- [17] NYLUND-GIBSON K, CHOI A Y. Ten frequently asked questions about latent class analysis[J]. Translational Issues in Psychological Science, 2018, 4(4):440-461.
- [18] 罗崎瑞, 张道文, 周华, 等. 面向智能汽车预期功能安全的驾驶场景评价[J]. 中国安全科学学报, 2022, 32(8): 140-145.
- LUO Qirui, ZHANG Daowen, ZHOU Hua, et al. Evaluation on driving scenarios for safety of intended functionality of intelligent vehicles[J]. China Safety Science Journal, 2022, 32(8):140-145.
- [19] YE Wanbao, WANG Chuanlin, CHEN Fuxiang, et al. Approaching autonomous driving with cautious optimism: analysis of road traffic injuries involving autonomous vehicles based on field test data[J]. Injury Prevention, 2020, 27(1): 42-47.
- [20] SONG Yu, CHITTURI M, NOYCE D. Impact of event encoding and dissimilarity measures on traffic crash characterization based on sequence of events[J]. Accident Analysis and Prevention, 2023, 185;DOI:10.1016/j.aap.2023.107016.



作者简介: 乔剑锋 (1977—),男,内蒙古呼和浩特人,博士,副教授,主要从事安全数据文本挖掘以及安全风险预警和评价方面的研究。E-mail: qiaojianfeng@cueb.edu.cn。