·智能网联汽车场景感知与智能体验技术专题·

基于无监督机器学习的驾驶风格识别方法研究*

赵伯儒[1] 李想[1] 王琛越[1] 王鑫[2,3] 赵宗琴[2,3]

（1.重庆师范大学重庆国家应用数学中心，重庆市 401331；2. 重庆长安汽车股份有限公司，重庆市 400023；3. 智能汽车安全技术全国重点实验室，重庆市 401133）

【欢迎引用】赵伯儒，李想，王琛越，等．基于无监督机器学习的驾驶风格识别方法研究[J]. 汽车文摘，2025(8): 25-33.

【Cite this paper】ZHAO B R, LI X, WANG C Y, et al. Research on Driving Style Recognition Based on Unsupervised Machine Learning[J]. Automotive Digest (Chinese), 2025(8): 25-33.

【摘要】驾驶风格识别对于提升智能网联汽车个性化驾驶体验和优化能源利用具有重要意义。考虑到不同道路环境与不同驾驶风格之间的耦合关系，设计了一个级联传递框架充分利用采集的实车自然驾驶数据，将数据分割成不同物理意义的事件。以驾驶员ID作为伪标签，建立XGBoost模型从中学习驾驶风格的差异性，确定对于驾驶风格识别重要的特征及权重。基于混合专家系统的思想，采用WK-means算法对不同环境下的驾驶风格进行聚类，最终得到驾驶分数来衡量驾驶员的表现。对聚类后各类驾驶员驾驶数据的统计分析表明，该方法能有效实现不同驾驶风格驾驶员的聚类，为智能网联汽车技术的进一步发展奠定了基础。

关键词：驾驶风格；无监督聚类；特征选择；机器学习

中图分类号：U471.1　　**文献标志码：**A　　**DOI:** 10.19822/j.cnki.1671-6329.20240243

Research on Driving Style Recognition Based on Unsupervised Machine Learning

Zhao Boru[1], Li Xiang[1], Wang Chenyue[1], Wang Xin[2,3], Zhao Zongqin[2,3]

（1. National Center for Applied Mathematics in Chongqing, Chongqing Normal University, Chongqing 401331; 2. Chongqing Changan Automobile Co., Ltd., Chongqing 400023; 3. State Key Laboratory of Intelligent Vehicle Safety Technology, Chongqing 401133）

【Abstract】Driving style recognition plays a crucial role in enhancing personalized driving experiences and optimizing energy utilization in smart connected vehicles. Considering the relationship between different road environments and driving styles, a cascade delivery framework is designed to fully utilize real-world natural driving data and segment the data into events with distinct physical meanings. Using driver IDs as pseudo-labels, an XGBoost model learns differences in driving styles, identifying the key features and weights critical for recognition. Following the principles of a hybrid expert system, the WK-means algorithm clusters driving styles under varying conditions, ultimately generating driving scores to evaluate driver performance. The statistical analysis of the clustered data shows that this method effectively recognizes drivers with diverse driving styles, which lays the foundation for the further development of intelligent networked vehicle technology.

Key words: Driving style, Unsupervised cluster, Feature selection, Machine learning

0 引言

驾驶风格指驾驶员在驾驶车辆过程中表现出的习惯性方式，其通常伴随驾驶员的驾驶经验逐渐形成[1]。有效识别驾驶风格不仅可以提高驾驶员的操作体验，降低汽车燃油或电量消耗，而且对动力总成的优化控制具有重要作用。

在驾驶风格识别方面，国内外学者开展了大量研究。Xie等[2]采用随机森林结合特征工程的方法对驾驶行为进行了分类。Wang等[3]引入基于隐藏半马尔

*基金项目：智能汽车安全技术全国重点实验室开放基金课题（IVSTSKL-202302）；重庆市自然科学基金项目（CSTB2023NSCQ-LZX0160）。

可夫模型的贝叶斯非参数方法，从多维时间序列驾驶数据中提取原始驾驶模式，分析了驾驶员的行为和风格。Guo等[3]将驾驶风格与驾驶周期解耦，在对驾驶周期进行分类和识别的基础上，分析不同驾驶周期中油门踏板开度及其变化率，并建立模糊逻辑识别器来识别驾驶风格。Maria等[5]提出了一种基于动态聚类的驾驶风格识别和分析方法，其中聚类随着周围工况的变化而变化，以更好地识别驾驶风格。Moosavi等[6]采用卷积神经网络从轨迹中捕捉驾驶员行为的语义模式，使用循环神经网络发现模式之间的时间依赖性，从而识别驾驶风格。Choi等[7]将驾驶员风格识别问题转换为异常检测问题，找出训练后的模型结果在高回归误差和低回归误差的情况，并以此区分驾驶员的驾驶风格。Milardo等[8]提出了一个基于数据驱动的驾驶行为识别框架进行风格识别。Cai等[9]利用卷积神经网络和长短期记忆网络相结合的方法对驾驶风格进行分类。Xu等[10]采用离线和在线结合的方式识别驾驶风格，在离线部分使用主成分分析（Principal Component Analysis，PCA）和K-means算法学习典型的驾驶风格，在线部分使用局部最大似然技术进行在线驾驶风格识别。

国内学者对于驾驶风格的研究起步较晚，但也取得了一定成果。赵韩等[11]从车流密度与驾驶风格识别之间的耦合关系入手，提出一种考虑车流密度影响的驾驶风格多层次识别方法。金辉等[12]采用高斯混合模型（Gaussian Mixture Model，GMM）聚类算法对驾驶数据进行分析，建立基于Fisher判别的驾驶风格识别方法模型。董昊旻等[13]提出了基于半监督学习三协同训练算法对驾驶风格进行识别的方法。吕明等[14]通过自组织映射（Self-Organizing Map，SOM）神经网络分别对驾驶数据进行聚类分析，并建立了基于SOM神经网络的驾驶风格识别系统。宋函锟等[15]利用主成分分析降维算法及K-means聚类算法对行驶数据进行驾驶风格分类研究。黄江等[16]提出了一种基于多元特征参数与优化支持向量机相结合的驾驶员驾驶风格识别模型。柳祖鹏等[17]运用模糊数学的方法分析驾驶风格。梁科等[18]利用鲸鱼优化算法对驾驶数据进行特征选择，再利用基于长短期记忆网络（Long Short-Term Memory，LSTM）的自编码器获得用于谱嵌入的特征值和特征向量，并最终通过谱聚类对驾驶风格进行识别。秦大同等[19]提出了一种基于驾驶事件、谱聚类与随机森林相结合的算法识别驾驶风格。

然而，现有研究对驾驶风格的识别存在一定局限性。（1）用于驾驶风格的数据主要通过驾驶模拟器、测试车辆或智能手机等设备收集，存在难以模拟复杂的真实交通状况、测试成本高和数据收集困难的问题。（2）用于驾驶风格识别的特征大部分是手动选择特征或使用所有可用特征，该方法存在特征选择不够准确、聚类结果不够稳定等问题。（3）判断驾驶风格和行为通常具有较强的主观性，同一个驾驶员在不同环境、不同阶段的驾驶风格也可能有所不同。针对上述问题，本文对该领域的贡献主要是以下3个方面；（1）考虑到不同道路环境对不同类型驾驶风格的感知程度不同，设计了一个全新的级联传递框架，充分利用由重庆长安汽车股份有限公司（以下简称长安汽车）提供的实车自然驾驶数据研究不同道路环境自然驾驶条件下的驾驶风格。（2）在无关于驾驶风格真实标签的情况下，创建了一个有监督的机器学习模型，以驾驶员的ID作为伪标签，来学习驾驶风格的差异性，并通过对特征的重要性排序来选择用于驾驶风格聚类的特征。（3）考虑到驾驶环境与驾驶风格之间的强关联性，基于混合专家系统的思想，采用聚类方法为不同驾驶环境下的驾驶风格建立专家模型进行识别，并综合形成总体驾驶风格，输出驾驶分数来衡量驾驶员的表现。

1 数据采集

本文使用了一组由长安汽车收集的驾驶数据，以探索驾驶风格的无监督聚类。数据通过安装在车辆中的车载传感器和记录设备采集，包括车载单元、全球定位系统和车载惯性测量单元等。设备以10 Hz的频率收集数据，记录了多个驾驶相关变量，包括车速、加速度、加速踏板角度以及刹车踏板状态等。数据采集在不同城市的不同行驶环境中进行，涵盖了北京、重庆、西藏等28个省、自治区、直辖市的城市道路、高速公路和乡村道路等不同驾驶场景。为了确保数据的多样性和代表性，本文采用了968辆长安汽车S203车型长达10个月的实车自然驾驶数据，共有17个不同类型的驾驶参数，其中部分车辆驾驶参数如表1所示。

2 基于级联传递框架的事件划分

2.1 数据清洗

在数据实际采集过程中，可能存在环境变化、设备不稳定以及传输延迟的问题。为了便于后续驾驶

风格识别，需要先对数据进行预处理。在数据预处理阶段，首先进行了数据清洗，去除异常或无效的数据点。其次，使用插值方法对数据中存在的缺失值进行处理，填补了部分缺失的数据。

表1　部分车辆驾驶参数

采集信号中文名称	采集信号英文名称	信号说明
车速ESP*	ESP_VehicleSpeed	0~360 km/h
加速踏板	EMS_AccPedal	0°~90°
挡位	TCU_ActualGear	
横加速度	ESP_LatAccel	
纵加速度	ESP_LongAccel	

注：电子稳定程序（Electronic Stability Program，ESP）。

2.2　基于级联传递框架的事件划分

原始传感器数据具有高噪声特性，可以有效识别驾驶环境和风格所需的数据，但也无法避免大量无效信息，甚至部分信息可能干扰驾驶环境和风格识别。因此，有效利用实车自然驾驶数据，并从中提取出区分不同驾驶风格特征至关重要。

本文主要的研究目的是建立一个全新的级联传递框架以充分利用实车自然驾驶数据，将驾驶数据划分为不同的事件粒度大小，并使用无监督的机器学习方法对驾驶员驾驶风格进行聚类。考虑到驾驶风格和行为通常具有较强的主观性，同一驾驶员在不同环境、不同阶段的驾驶风格也可能存在差异。例如，在高速公路上谨慎的驾驶员可能在拥堵的环境中表现出激进的驾驶风格。因此，本文采用级联传递框架先对数据进行粗细粒度的事件划分，在此基础上进行聚类。具体可将驾驶事件划分以下4个类型：旅程（Journey，以下简称J事件）、行程（Drive，以下简称D事件）、加速（Accelerate，以下简称A事件）、刹车（Brake，以下简称B事件）。其中，J事件表示汽车行驶的整个旅程，一趟旅程事件可能包含多种驾驶场景，为粗粒度事件；D事件将J事件数据分为多段驾驶场景，即一段汽车速度从0开始，到速度为0结束的路程，为细粒度事件；A事件为汽车行驶路程中加速踏板被踩下（即其属性值不为0）的一段加速路程；B事件为路程中加速踏板未被踩下（即其属性值为0）的一段路程。基于级联传递框架的事件划分结构如图1所示。用于本文研究的J事件数量为15 854个、D事件数量为33 056个、A事件数量为515 638个以及B事件数量为490 047个。

引入级联传递框架可以更好地处理复杂的驾驶数据，通过逐层缩小数据的粒度，使每层级数据具有一定现实意义，从而确保驾驶风格特征能够在合适的层次上被提取。此外，该框架使聚类算法可以更灵活地适应和识别不同环境的驾驶风格，为后续在各层级事件中传递聚类结果奠定了基础。

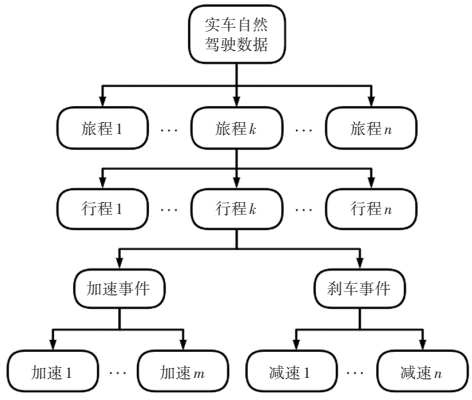


图1　基于级联传递框架的事件划分结构

3　重要特征的选择和排序

在研究驾驶风格过程中，特征工程在驾驶数据分析中具有重要作用。通过提取有效的特征，可以揭示不同驾驶风格之间的差异性，为进一步的驾驶风格识别奠定基础。

3.1　特征工程

3.1.1　基于统计函数的特征提取

基于统计函数的特征提取方法旨在从驾驶数据中获取有意义的信息，并通过计算各种统计度量描述不同驾驶行为的特征。统计度量主要包括均值、方差、最大值、最小值以及中位数，可以反映驾驶行为的集中趋势、变异程度以及行为的极端情况。部分基于统计函数的特征提取方法如表2所示。

表2　部分基于统计函数的特征提取方法

序号	特征名称	缩写	定义
1	平均值	mean	$Y_{mean}=\frac{1}{n}\Sigma_{i=1}^{n}x_i$
2	最大值	max	$Y_{\max}=\max(x_i)$
3	标准差	std	$Y_{std}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-Y_{mean})^2}$

注：i为数据元素的索引，$x=\{x_1,x_2,\cdots,x_n\}$表示一个数据样本，n为数据元素的总数。

3.1.2　基于信号处理的频域特征提取

可以将车辆采集的驾驶数据视为时间序列数据，采用信号处理的方法进行分析。基于信号处理的频域特征提取方法，结合傅里叶变换或其他相关的频域转换技术，将驾驶数据由时域转换为频域。上述特征主要包括峰度、偏度、小波能量、频谱熵以及平均频

率。通过这种转换，能够分析不同频率范围内驾驶行为的能量分布和频谱特性，获得关于驾驶行为更加细粒度的信息，揭示其在不同频率条件下的行为特征和动态模式，对研究驾驶风格的时变特性和频率相关行为具有重要意义。部分基于信号处理的特征提取方法如表3所示。

表3　部分基于信号处理的特征提取方法

序号	特征名	缩写	定义
1	峰度	kurt	$Y_{kurt}=\frac{\sum_{i=1}^{n}(x_i-Y_{mean})^4}{(n-1)Y_{std}^4}$
2	偏度	skew	$Y_{skew}=\frac{\sum_{i=1}^{n}(x_i-Y_{mean})^3}{(n-1)Y_{std}^3}$
3	频谱熵	spectral entropy	$Y_{spectral\_entropy}=-\sum_{k=1}^{m}P(k)\log_2 P(k)$

注：k为频率分量的索引，$P(k)$为第k个频率分量的概率，m为频率分量的总数。

3.1.3　基于驾驶数据流的时序特征提取

为了量化驾驶员的驾驶行为，本文参考了Mohammadnazar等[20]提出的时变驾驶波动概念，通过在驾驶员层面创建时间序列数据流来捕捉瞬时驾驶行为的变化，使用时变驾驶波动作为驾驶风格激进程度的替代度量。驾驶波动显示了驾驶数据偏离常态的程度，较高的驾驶波动表示驾驶员在横向和纵向上的速度、加速和减速波动均较大，进而表明驾驶员具有较高的不稳定性。时变驾驶波动的计算方法可表示为：

$$v_f=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i-\bar{r})^2} \tag{1}$$

$$r_i=\ln\left(\frac{x_i}{x_{i-1}}\right)\cdot 100 \tag{2}$$

式中：x_{i-1}为观测值x_i之前的观测值，r为x_{i-1}到x_i的增长率，$\bar{r}$为参数r的平均值。

3.2　重要特征选择

对于原始驾驶参数，本文采用上述3种特征提取方法，共生成了262个特征。目前大多数研究采用手动选择重要特征或直接使用全部特征进行聚类，这种方法存在特征选择不够准确、聚类结果不稳定，难以解释和理解等问题。此外，无监督机器学习算法缺乏通用的特征重要性计算方法。为了确定可以有效描述驾驶风格的特征选择方法，依靠机器学习建模和特征重要性对特征进行排序。由于在许多自然驾驶数据中无法获得驾驶风格的真实标签，本文针对驾驶数据的时间序列特性，将时间序列数据转换为驾驶事件切片，并将驾驶员ID作为伪标签用于模型训练，从而将无监督特征选择转换为监督特征选择的问题。最终，通过对特征进行重要性排序来选择用于驾驶风格聚类特征。该特征选择方法的优点在于，即使原始数据中不存在关于驾驶风格或行为的标签，驾驶员ID可以在一定程度上表征驾驶员的驾驶风格，因此将驾驶员ID作为伪标签训练模型，创建了一个有监督的机器学习任务，以选择关键特征，捕捉个体之间驾驶行为的显著差异。此外，本文还采用了一种与模型无关的重要性排序方法来计算特征的重要度。排列重要性表明，当目标特征被随机排列的值替代时，模型性能将下降，该方法被广泛用于解释机器学习模型。

3.2.1　去除特征的多重共线性

特征的多重共线性描述了2个或多个特征之间的高度相关性或强线性依赖。在存在多重共线性的情况下，特征之间的高度相关性将影响特征的重要性排序结果，使得模型难以判定特征对目标变量的贡献。虽然多重共线性的特征可能仍提供部分信息，但不足以对数据分析和建模的有效性产生显著影响。因此，在进行特征选择和排序之前，需要先处理特征的多重共线性问题，以确保选择的特征具有代表性和独立性，从而提高模型的预测能力和稳定性。本文通过对特征的Spearman相关系数矩阵进行分层聚类来识别共线特征。通过相关性分析，可将高度相关的特征聚类成组。为了解决特征间的多重共线性问题，从每组中选择一个代表性特征，并构建为新的特征子集。

考虑到粗粒度的J事件所提取的特征较难有效表征驾驶员的驾驶风格，本文对细粒度的D事件、A事件和B事件分别采用分层聚类方法来识别共线特征组。在每个组中，保留了最常用且能够充分表征驾驶风格的特征，而删除了其余特征。图2展示了D事件特征层次聚类的部分结果，显示出D事件中速度时间序列的标准差、方差、引擎速度的平均值、中位数和频谱熵高度相关。由于速度标准差可以衡量驾驶行为中速度的变化程度，且在文献中被广泛使用，因此予以保留，而其他4个特征则属于研究领域的新特征。在本文的研究中，共删除了240个特征，最终得到一个包含22个无多重共线性特征的新特征集合。

3.2.2　特征重要性排序

本文将新的特征集合作为输入，通过XGBoost分类器训练驾驶员识别模型。每个驾驶员的驾驶数据被分成不同粗细粒度的事件，使用分层抽样的方法将数据分为训练集和测试集。模型的预测标签为驾驶

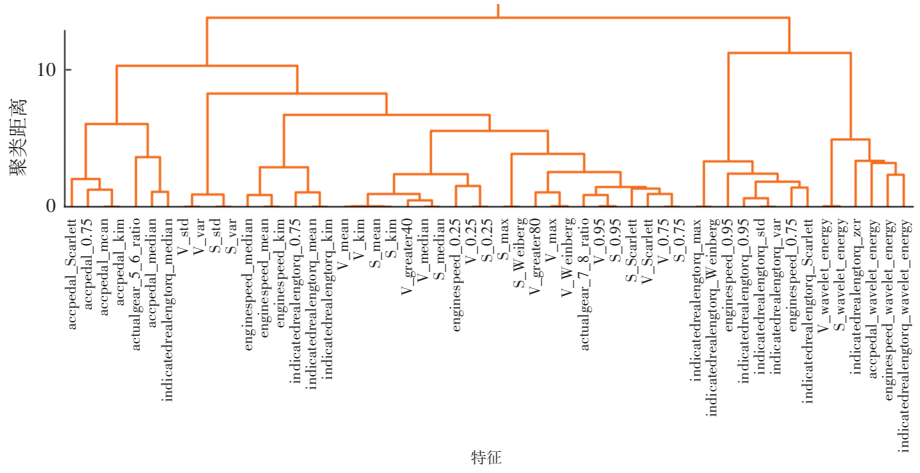


图2　D事件特征层次聚类的部分结果

员的ID，通过Python库ELI5计算每个特性的重要性排序。XGBoost分类器的训练过程重复了5次，使用特征的平均重要权重对特征进行排序，其中权重最高的特征被视为关键特征。该特征选择方法的优势为即使原始数据未包含关于驾驶风格或技能水平的标签，只要包含驾驶员的身份信息，就能够通过监督机器学习任务选择捕捉个体之间驾驶行为显著差异的特征。本文分别对D事件、A事件和B事件进行了上述的特征重要性排序操作，为后续驾驶风格聚类提供重要特征的权重。表4为D事件的特征重要性排序及权重。

表4　D事件的特征重要性排序及权重

序号	特征名称	缩写	权重	序号	特征名称	缩写	权重
1	J事件环境类别	env_class_j	0.092	12	速度波峰数	V_peak	0.040
2	横向加速度标准差	lataccel_std	0.065	13	加速度偏度	acc_skew	0.040
3	油门开度标准差	accpedal_std	0.065	14	单位时间挡位转换次数	unit_change	0.038
4	加速度第四分之一位数	acc_0.25	0.061	15	速度平均标准误差	V_sem	0.038
5	D事件环境类别	env_class_d	0.056	16	挡位转换次数	actualgear_shift	0.038
6	油门开度中位数	accpedal_median	0.051	17	速度频谱熵	V_spectral_entropy	0.037
7	横向加速度平均值	lataccel_mean	0.050	18	事件级驾驶时变波动	speed_volatility_event_mean	0.034
8	速度标准差	V_std	0.050	19	B事件速度标准差平均值	B_mean_speed_std	0.033
9	速度平均值	V_mean	0.048	20	横向加速度偏度	lataccel_skew	0.029
10	纵向加速度中位数	longaccel_median	0.046	21	加速度平均值	acc_mean	0.025
11	横向加速度小波能量	lataccel_wavelet_energy	0.041	22	速度平均频率	V_mean_frequency	0.023

4　驾驶风格聚类

4.1　K-means

K-means是一种无监督的聚类算法，其目标是将数据样本划分为k个聚类，并通过迭代优化的方法找到每个聚类的质心，使得每个样本都能聚集在最近的质心周围。K-means方法通过最小化以下定义的误差函数来得到最终输出的聚类结果，误差函数可表示为：

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu(C_i)) \tag{3}$$

式中：E表示所有样本到其对应聚类中心的总距离，C_i表示第i个聚类，$\mu(C_i)$为第i个聚类的质心，$d(x,\mu(C_i))$为数据样本x与$\mu(C_i)$之间的距离。

K-means算法中的常见的距离度量方式主要包括欧几里得距离、曼哈顿距离、闵可夫斯基距离和切比雪夫距离。本文使用欧几里得距离作为数据样本与聚类质心之间的距离度量方式，其公式可表示为：

$$d(x,\mu) = \sqrt{\sum_{k=1}^{n} (x_k - \mu_k)^2} \tag{4}$$

式中：$\mu = \{\mu_1, \mu_2, \cdots, \mu_n\}$为一个聚类的质心，$d(x,\mu)$为$x$到$\mu$的距离。

4.2 WK-means

WK-means(Weighted K-means)是一种基于加权的K-means聚类算法,用于对数据进行聚类分析。与传统的K-means算法相比,WK-means在计算样本间距离时考虑了样本的权重信息,从而在聚类过程中能够更加准确地反映样本的相似性和距离。

在传统的K-means算法中,每个样本都被视为具有相等的权重,而WK-means则引入了样本权重的概念。样本权重可以根据数据的特点或先验知识进行分配,以突出某些样本的重要性或减弱某些样本的影响。从而WK-means可以根据不同样本的权重信息更好地调整聚类中心,得到更精确的聚类结果。

WK-means与传统的K-means算法的区别在于样本间距离的计算方式。在传统的K-means算法中,使用欧氏距离或其他距离度量方式衡量样本之间的相似性。而在WK-means中,采用样本间的加权距离计算样本的相似性。加权距离可以通过对距离度量进行加权求和的方式求得,其中样本权重决定了不同样本在距离计算中的贡献度。本文基于加权的欧几里得距离计算WK-means的距离,其计算过程可表达为:

$$d(x,y)=\sqrt{\sum_{i=1}^{n} w_i(x_i-y_i)^2} \tag{5}$$

式中:$d(x,y)$表示样本x和样本y之间的加权欧氏距离,w_i表示样本的权重,x_i和y_i分别表示样本x和样本y在第i个特征上的取值。

4.3 聚类数量的选择

从驾驶风格分类的角度看,每个集群可以代表一种驾驶风格。聚类数量的确定通常依赖研究人员基于经验的主观判断,或通过聚类质量度量进行客观选择。本文结合了上述2种方法,采用主观判断方法将数据中的驾驶环境划分为3个聚类。然后,针对驾驶风格分类任务,采用聚类质量度量的方法确定聚类数量,该划分方法更符合以往对驾驶环境和驾驶风格分类的研究。

本文根据轮廓系数(Silhouette Coefficient Score)和卡林斯基指数(Calinski-Harabasz Index)进行聚类数量的选择。其中,轮廓系数是一种用于度量聚类结果紧密度和分离度的指标。对于每个数据点,轮廓系数计算了该点与同簇内其他点的平均距离以及其与最近邻簇内所有点的平均距离。该指标的取值范围为[-1,1]。取值越接近1,表明该样本聚类越紧密且与其他簇分离度较高;取值越接近-1表示样本聚类效果差;取值为0表示样本在2个簇的边界区域。数据点x_i的轮廓系数S计算公式可表示为:

$$S_{x_i}=\frac{b_{x_i}-a_{x_i}}{\max\left(a_{x_i},b_{x_i}\right)} \tag{6}$$

式中:b_{x_i}为样本x_i到最近其他簇的所有样本的平均距离,a_{x_i}为样本x_i到同一簇内其他样本的平均距离。

卡林斯基指数基于簇内紧密度和簇间分离度对评估聚类质量进行评判。该指数通过比较簇间协方差和簇内协方差来度量聚类的紧密性和分离度,以此判断聚类效果。卡林斯基指数的值越大,表示簇内方差大、簇间方差小,聚类效果越好。卡林斯基指数C的计算公式可表示为:

$$C=\frac{\frac{Tr(B)}{k-1}}{\frac{Tr(W)}{N-k}} \tag{7}$$

式中:B表示簇间协方差矩阵,W表示簇内协方差矩阵,Tr表示矩阵的迹,k表示簇的个数,N表示总样本数。

5 试验结果

本试验旨在验证所提方法在不同环境下对驾驶风格识别的有效性。首先,对上文所述不同粒度大小的事件数据进行环境聚类,分析每个聚类簇的特征,以确定各簇对应的具体环境。其次,计算每种环境下数据的轮廓系数和方差比率准则作为评价标准,通过指标的结果综合确定最佳驾驶风格聚类类别数和最优聚类算法。然后,在各环境下分别应用算法进行风格聚类,并对每个聚类簇的特征进行详细分析,以刻画该聚类簇所代表的驾驶风格类型。最后根据不同环境下的风格聚类簇,计算每位驾驶员的驾驶得分,以综合量化其驾驶表现。

首先通过K-means算法对不同粒度大小的事件进行环境聚类,每个事件被分配一个数字(1、2或3)以表示驾驶事件所在的环境。粗粒度事件的聚类结果会传递给细粒度事件,每个聚类所代表的驾驶环境是基于数值判断的。图3展示了D事件环境聚类箱体图,其中环境聚类3的速度均值中位数最高,其次是环境聚类2,而环境聚类1的速度均值中位数最低。当将驾驶环境分成3类时,在速度均值维度上出现了显著差异,这进一步证实了聚类算法可以有效地将原始数据按照驾驶环境特征分成高速、中速和慢速3类场景。

对各环境聚类的数据进一步分析,根据上文中得到的特征重要性排序结果,本文选取了横向加速度标准差(lataccel_std)、时变驾驶波动(speed_volatility_event_mean)、速度均值标准误差(V_sem)、加速度标准

差(acc_std)和横向加速度平均值(lataccel_mean)5个特征衡量各环境中的驾驶员风格,衡量结果如表5所示。高速场景各项特征指标均低于低速和中速场景,说明高速场景中,驾驶员的激进行为和驾驶波动性低于其他场景。例如,高速、中速和低速场景的特征speed_volatility_event_mean均值分别为1.525、4.148和8.887,其余特征也表现出类似的趋势。由于低速和中速环境的行驶场景多为城市、乡村等,所处的环境和道路条件变化较大,如交通信号灯以及与其他道路使用者(行人、自行车和摩托车)有更复杂的交互,因此,低速和中速环境的驾驶时变波动值会高于高速环境。同样,低速环境相较于中速环境拥有更高的驾驶时变波动值,这可能是由于低速环境的行驶场景更加拥堵且复杂多变。该结果表明,按照不同驾驶环境对驾驶风格进行识别的方法具有可行性。

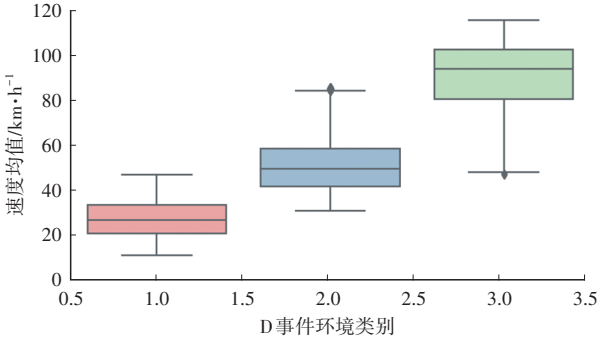


图3　D事件环境聚类箱体图

表5　对每个环境中驾驶员风格的衡量结果

特征	环境聚类1				环境聚类2				环境聚类3			
	最小值	最大值	平均值	标准差	最小值	最大值	平均值	标准差	最小值	最大值	平均值	标准差
lataccel_std/m·s^{-2}	0.122	0.985	0.434	0.193	0.198	1.009	0.476	0.191	0.161	1.176	0.418	0.234
speed_volatility_event_mean/m·s^{-1}	4.318	16.256	8.887	2.735	1.481	7.504	4.148	1.454	0.356	5.637	1.525	1.194
V_sem/m·s^{-1}	0.356	1.984	0.936	0.381	0.312	1.269	0.694	0.235	0.123	0.527	0.296	0.105
acc_std/m·s^{-1}	0.288	0.852	0.551	0.132	0.279	0.718	0.468	0.106	0.087	0.785	0.349	0.159
lataccel_mean/m·s^{-1}	0.007	0.231	0.085	0.056	0.004	0.168	0.052	0.039	0.001	0.197	0.0458	0.044

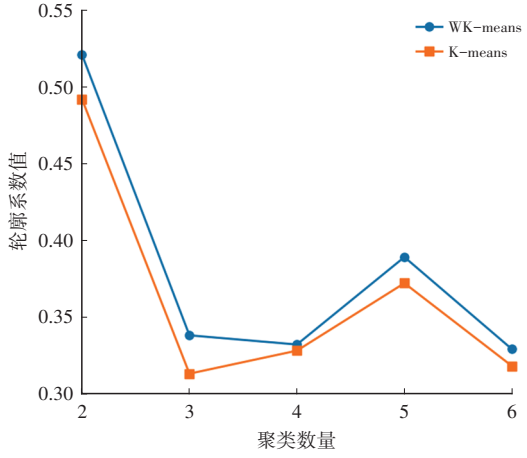
虽然实车自然驾驶数据中包含大量字段,但并非所有字段均为识别驾驶风格的必要特征。因此,有必要选择一部分有价值的属性作为训练特征。一方面,特征选择能够增强模型的泛化能力,降低模型过拟合的风险;另一方面,减少不必要的字段数量可以实现降维,提高训练效率。本文将上文挑选出的22个重要特征及权重作为驾驶风格训练的输入。

数据集中的每个元组代表一个驾驶事件,驾驶事件被分类到距离最近的聚类中。同时,数据经过标准化处理,以确保特征具有可比性。从驾驶风格分类的角度来看,每个聚类代表一种驾驶风格。图4展示了WK-means和K-means方法不同簇数的轮廓系数图和方差比率准则图。与K-means方法相比,WK-means方法在轮廓系数和方差比率准则上的表现更优,表明该方法提供了更高质量的聚类。尽管在聚类类别数为2时,轮廓系数和方差比率准则值最高,但由于该类别数不符合实际需求。因此,本文最终选择以5作为聚类类别数,并采用WK-means方法作为专家模型对不同环境下的驾驶风格进行聚类分析。

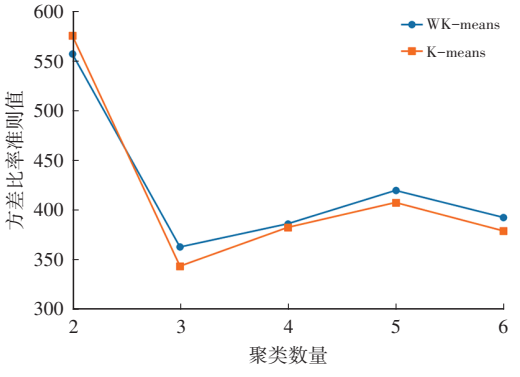
在对数据集执行WK-means聚类时,每个事件被分配一个类别编号(例如1、2、3、4或5)来表示驾驶事件所在聚类。本文从22个聚类特征中选取了5个特征作为衡量标准,并通过计算每个聚类的特征平均值确定所代表的驾驶风格。表6展示了高速、中速和低速环境中驾驶风格分类特征的平均值。该表显示,不同环境下的驾驶风格聚类结果具有较为明显的差异,通过WK-means聚类算法能够将不同环境下的数据分为5个类别,每个类别代表一种驾驶风格。例如,在中速场景中,风格4和风格5的speed_volatility_event_mean特征均值分别为5.804和5.81,对应的lataccel_std特征均值分别为1.072和0.614。由此可见,时变驾驶波动值较大的风格类别,其横向加速度标准差值也较大,表现出更为激进的驾驶风格。类似的趋势也出现在其他特征中,如acc_std和V_mean等。上述结果进一步验证了WK-means算法可以有效聚类不同环境下的驾驶风格。

为了进一步量化驾驶风格,分别针对激进驾驶、冒险驾驶、平稳驾驶、温和驾驶和保守驾驶分配1、2、3、4、5的值。因此,驾驶风格被视为一个离散变量,取值范围为1~5。考虑到驾驶员在不同环境的驾驶风格存在差异,其驾驶风格评分随之变化。为了实现对驾驶表现进行统一量化评估,本文定义了“驾驶分数”指标。该指标是指每种环境类别值的平均值。具体而言,在不同环境类别中,根据驾驶员ID对驾驶事件进行汇总,为每个驾驶员计算3个驾驶得分,范围为1~5,以反映驾驶员在不同环境下的表现。与离散的

驾驶风格不同，驾驶分数是一个连续变量，可以取1~5之间的任何值，这种连续性使得对驾驶员的表现评估更加精细化和灵活。



(a)不同簇数下的轮廓系数



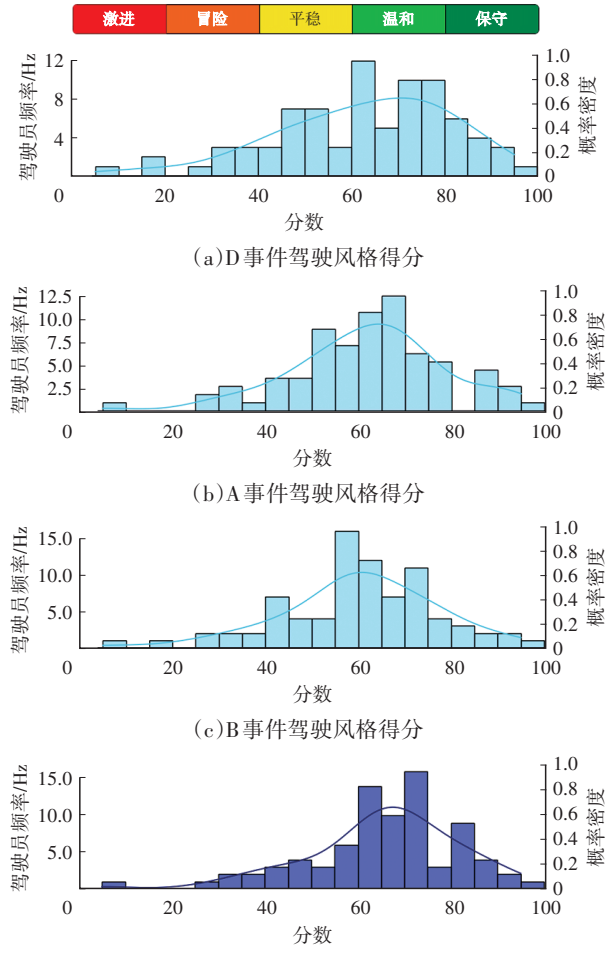
(b)方差比率准则

图4　聚类方法比较

表6　各环境中驾驶风格分类特征的平均值

环境	聚类	lataccel_std /m·s^{-2}	speed_volatility_event_mean /m·s^{-1}	V_sem /m·s^{-1}	acc_std /m·s^{-2}	lataccel_mean /m·s^{-2}
高速环境	风格1	0.429	10.048	1.392	0.709	0.07
	风格2	0.322	9.324	0.576	0.402	0.087
	风格3	0.64	9.457	0.829	0.577	0.125
	风格4	0.32	6.455	1.096	0.518	0.063
	风格5	0.958	10.294	0.708	0.633	0.094
中速环境	风格1	0.515	5.182	1.38	0.682	0.062
	风格2	0.374	3.697	0.653	0.415	0.046
	风格3	0.404	2.967	0.788	0.429	0.042
	风格4	1.072	5.804	0.458	0.594	0.062
	风格5	0.614	5.81	0.618	0.533	0.071
低速环境	风格1	0.306	0.581	0.153	0.165	0.096
	风格2	0.469	0.355	0.121	0.089	0.117
	风格3	0.487	2.606	0.305	0.419	0.026
	风格4	1.462	5.83	0.334	0.782	0.055
	风格5	0.354	1.199	0.34	0.352	0.043

如图5所示，不同粒度事件下驾驶分数的直方图和总体驾驶分数直方图表明，表现良好和优秀的驾驶员数量高于表现较差和极差的驾驶员数量。总体驾驶分数直方图的分布形态进一步佐证了这一趋势，这与当前有关激进驾驶和保守驾驶员比例的认知和研究相一致。



(a)D事件驾驶风格得分

(b)A事件驾驶风格得分

(c)B事件驾驶风格得分

(D)总体事件驾驶风格得分

图5　不同粒度事件下驾驶分数直方图和总体驾驶分数

6　结束语

本文旨在解决智能网联汽车技术领域中的驾驶风格识别问题，充分利用实车自然驾驶数据，深入研究了不同驾驶环境中自然驾驶条件下的驾驶风格。通过设计了一个级联传递框架，将数据分割成不同粒度大小的事件，为后续分析提供了坚实的基础。在特征提取方面，采用的方法主要包括基于统计函数、信号处理以及驾驶数据流的特征提取方式，获得用于驾驶环境和驾驶风格聚类的特征。进而建立了一个具有监督的机器学习模型，以驾驶员ID作为伪标签，从中学习驾驶风格的差异性，确定对于驾驶风格识别重要的特征及权重。基于混合专家系统的思想，采用

WK-means算法对不同环境下的驾驶风格进行聚类，最终得到驾驶分数来衡量驾驶员的表现。研究发现，不同环境条件下，激进驾驶、平稳驾驶以及保守驾驶的感知和阈值存在差异。例如，在高速环境中激进的驾驶风格可能在低速环境中被视为平稳驾驶。此外，不同粒度事件的驾驶分数直方图表明，驾驶员总体表现良好，激进风格的驾驶员数量比平稳和保守驾驶风格的驾驶员数量少，这与人们对于驾驶风格的认知一致。未来的研究可以丰富数据集，加入不同车型的数据，以增强模型的鲁棒性；可以考虑驾驶风格的时间评价，以评估驾驶风格随时间的变化；此外，还可以结合音、视频数据在驾驶风格的基础上进一步研究驾驶行为和驾驶意图，这对智能网联汽车技术发展具有积极意义。

参 考 文 献

[1] LYU N, WANG Y, WU C, et al. Using Naturalistic Driving Data to Identify Driving Style Based on Longitudinal Driving Operation Conditions[J]. Journal of Intelligent and Connected Vehicles, 2022, 5(1): 17-35.

[2] XIE J, ZHU M. Maneuver-Based Driving Behavior Classification Based on Random Forest[J]. IEEE Sensors Letters, 2019(10): 7002104.

[3] WANG W, XI J, ZHAO D. Driving Style Analysis Using Primitive Driving Patterns with Bayesian Nonparametric Approaches[J]. IEEE, 2018(10): 2986-2998.

[4] QIUYI GUO, ZHAO Z, SHEN P, et al. Adaptive Optimal Control Based on Driving Style Recognition for Plug-in Hybrid Electric Vehicle[J]. Energy, 2019, 186(12): 115824.

[5] A M V N D Z, A F M, B J S, et al. Dynamic Clustering Analysis for Driving Styles Identification[J]. Engineering Applications of Artificial Intelligence, 2021, 97(1): 104096.

[6] MOOSAVI S, MAHAJAN P D, PARTHASARATHY S, et al. Driving Style Representation in Convolutional Recurrent Neural Network Model of Driver Identification[J/OL]. (2021-02-11)[2025-07-14]. https://arxiv.org/abs/2102.05843.

[7] CHOI Y A, PARK K H, PARK E, et al. Unsupervised Driver Behavior Profiling Leveraging Recurrent Neural Networks[J/OL]. (2021-08-11)[2025-07-14]. https://arxiv.org/abs/2108.05079.

[8] MILARDO S, RATHORE P, SANTI P, et al. A Data-Driven Framework for Driving Style Classification[C]. International Conference on Advanced Data Mining and Applications. Springer, 2022.

[9] CAI Y, ZHAO R, WANG H, et al. CNN-LSTM Driving Style Classification Model Based on Driver Operation Time Series Data[J]. IEEE Access, 2023(11): 16203-16212.

[10] XU T, WU K, ZHU Y, et al. Driving Style Recognition at First Impression for Online Trajectory Prediction[J]. IFAC PapersOnLine, 2023, 56(2): 11287-11292.

[11] 赵韩, 刘浩, 邱明明, 等. 考虑车流密度影响的驾驶风格识别方法研究[J]. 汽车工程, 2020, 42(12): 1718-1727.

[12] 金辉, 吕明. 基于改进Fisher判别的起步工况驾驶风格研究[J]. 北京理工大学学报, 2020, 40(3): 262-266.

[13] 董昊旻, 张维轩, 王文彬, 等. 基于Tri-Training的驾驶风格分类算法[J]. 汽车技术, 2021(4): 6-11.

[14] 吕明, 张滢, 冯先泽. 基于SOM神经网络的多工况驾驶风格识别[J]. 汽车实用技术, 2021, 46(2): 108-112.

[15] 宋函锟. 基于NGSIM数据库的驾驶风格聚类研究[J]. 汽车实用技术, 2022, 47(24): 40-45.

[16] 黄江, 李雨涵, 吴盛斌, 等. 基于多元特征参数与改进SVM算法的驾驶风格识别研究[J]. 重庆理工大学学报(自然科学), 2022, 36(11): 8-19.

[17] 柳祖鹏, 罗陈怡, 严运兵. 考虑车辆跟车及换道交互参数的驾驶风格识别[J]. 武汉理工大学学报:交通科学与工程版, 2023, 47(2): 209-213.

[18] 梁科, 陈华晟, 潘明章, 等. 采用双向LSTM自编码器的驾驶风格谱聚类识别研究[J]. 重庆理工大学学报(自然科学), 2023, 37(10): 28-37.

[19] 秦大同, 陈沫机, 曹宇航, 等. 基于驾驶事件的驾驶风格分类与识别方法研究[J]. 中国机械工程, 2024, 35(9): 1534-1541.

[20] MOHAMMADNAZAR A, ARVIN R, KHATTAK A J. Classifying Travelers' Driving Style Using Basic Safety Messages Generated by Connected Vehicles: Application of Unsupervised Machine Learning[J]. Transportation Research Part C Emerging Technologies, 2021, 122: 102917.

（责任编辑　梵玲）