DOI:10.19850/j.cnki.2096-4706.2024.02.005

基于 K-means 的驾驶行为离散化特征聚类分析与研究

宋月亭，卢巍

（昆明文理学院 信息工程学院，云南 昆明 650221）

摘　要：为挖掘连续驾驶行为数据中潜在的特征关系，文章采用实际运输车辆连续的驾驶行为数据。首先通过相应的预处理和特征提取，获取对应车辆在相应时间段连续的驾驶行为数据；其次采用在离散标准数据集和连续且有噪声数据集中均有稳定表现的 K-means 聚类方法，对驾驶行为数据进行离散化聚类处理与分析；最后获得三类有代表性的驾驶行为：“平稳型驾驶”“冲动型驾驶”和“危险型驾驶”。此外，对驾驶行为中隐含的各类特征进行分析研究，为后续进一步根据驾驶行为数据进行数据挖掘之关联分析提供有力依据。

关键词：驾驶行为；聚类；离散化；特征分析

中图分类号：TP391　　**文献标识码：**A　　**文章编号：**2096-4706（2024）02-0017-04

Clustering Analysis and Research on Discretization Characteristics of Driving Behavior Based on K-means

SONG Yueting, LU Wei

(School of Information Engineering, The College of Arts and Sciences Kunming, Kunming 650221, China)

Abstract: To explore potential feature relationships in continuous driving behavior data, this paper collects continuous driving behavior data of actual transportation vehicles. Firstly, through corresponding preprocessing and feature extraction, obtain continuous driving behavior data of the corresponding vehicle during the corresponding time period; secondly, the K-means clustering method, which has stable performance in both discrete standard datasets and continuous noisy datasets, is used to make discretization clustering processing and analysis on driving behavior data; finally, three representative driving behaviors are obtained:“Steady Driving”“Impulsive Driving”“Dangerous Driving”. In addition, analyzing and studying the various hidden features in driving behavior provides a strong basis for further correlation analysis in data mining based on driving behavior data.

Keywords: driving behavior; clustering; discretization; characteristics analysis

0 引　言

随着经济与工业生产的蓬勃发展，以及人们日益增长的物质生活需求，我国汽车保有量、汽车驾驶员人数呈逐年稳步增长趋势。与此同时，驾驶量的增加带来了交通阻塞和环境污染，也使得近年来我国车祸数量明显增加 [1]。随着交通部门管控措施的加强以及国民素质的不断提升，交通事故发生率得到了一定程度的控制，但驾驶安全问题仍不容小觑。大部分交通事故的发生是由于驾驶员未能在驾驶过程中遵守交通规则，进行了如疲劳驾驶、超速驾驶等不良的驾驶行为操作 [2]。同时驾驶员的不良驾驶习惯，如急加速、急减速、急刹车、随意变道等，也为避让不及造成追尾事故埋下隐患。因此，根据实际驾驶中采集到的数据，分析驾驶员行车过程中隐含的各类驾驶行为特征以及其体现的相关关系，成为研究安全驾驶的重要内容。

收稿日期：2023-05-26

由于车载数据获取主要通过车辆传感器进行收集，因此在对获取到的车载数据进行分析时，除了需要对基础数据进行相应的处理和标准化外，还需要考虑到，此类数据是基于驾驶员行驶过程中每间隔一秒进行一次数据采集，致使大量数据形成连续的行车轨迹数据。因此，本文首先对大量的驾驶行为数据进行预处理和特征提取，进而针对连续的驾驶行为数据进行基于聚类分析的离散化处理，并通过聚类对不同驾驶行特征为进行安全性能分簇，最终形成三种驾驶行为特征。

1 驾驶行为数据特征处理

本文采用泰迪杯数据挖掘大赛中提供的部分运输车辆驾驶行为数据作为实验数据，选取了其中 40 辆车 2018 年 7 月 30 日至 2018 年 10 月 10 日的原始驾驶数据，数据主要涉及每辆间隔 1 秒时间下的转向角度、经纬度、转向灯情况、手刹脚刹情况、GPS 速度和里程等，如表 1 所示。

表 1 部分车辆原始行驶数据

Vehicleplatenumber	Location_time	Direction_angle	lng	Right_turn_signals	Foot_break	…	Gps_speed
AA00001	2018/8/7 10:24:36	110	115.849 5	0	0	…	15
AA00001	2018/8/7 10:24:37	108	115.849 6	0	0	…	17
AA00001	2018/8/7 10:24:38	116	115.850 1	0	0	…	33
AA00001	2018/8/7 10:24:39	120	115.851 2	0	0	…	34
AA00001	2018/8/7 10:24:40	120	115.851 3	0	0	…	33
…	…	…	…	…	…	…	…

首先，由于可能存在传感器定位偏移、数据精度缺失、数据传输故障等问题，需要将原始数据进行预处理。针对数据传输故障等导致的缺失数据，结合前后时间点数据进行均值填充；针对数据传输故障等情况，对重复时间或重复设备号数据进行删除；同时对明显的速度异常、转向角度异常、里程异常等情况进行分析修正，对于较短时间片段内的异常，利用前后时间点下的均值进行替换，对于超过一定时间段（5 s 以上）的异常值，进行删除；同时，对手刹和脚刹状态进行数据区分，以便区分该刹车行为致使车辆处于停止状态还是减速行驶状态。

然后，由于仅含有车辆对应时刻行驶状态，因此需要对上述异常数据进行分析，获取车辆在两次停止状态间，相应时间片段内的急加速、急减速、刹车、疲劳驾驶、平均车速、最高车速等驾驶行为特征数据。查阅相关资料，并参照行业经验，设定对应时间片段内，当 $a > 3\ \text{m/s}^2$，且其时间域在 $0 < t < 3$ s 时为急加速；当 $a \leqslant -3\ \text{m/s}^2$，且其时间域在 $0 < t < 3$ s 时为急减速；当 $a < -4\ \text{m/s}^2$，且使得 3 s 后 $v < 0.5$ m/s 时为急刹车；当连续驾驶时间 $T_{\text{work}} > 4$ h 且休息时间 $T_{\text{rest}} < 20$ min，或一天累计驾驶时间 $T > 8$ h 时为疲劳驾驶 [3]。

最终获得具备相应驾驶行为特征的连续驾驶行为片段，如表 2 所示。

表 2 部分行驶片段特征数据

时间片 /s	平均速度 /（km/h）	最高速度 /（km/h）	急加速 / 次	急减速 / 次	急刹车 / 次	疲劳驾驶 / 次
3 634	63.78	78.3	2	1	0	0
731	45.41	66.0	4	0	1	0
14 832	78.12	90.5	13	7	4	1
7 309	70.67	87.4	6	3	2	0
4 795	59.36	86.0	0	0	0	0
…	…	…	…	…	…	…

2 K-means 聚类算法及其离散化检验

常用的数据离散化方法有等宽法、等频法。等宽法通过划分相同宽度的区间对连续数据进行划分，简单直观，但对数据分布要求较高，各类别下数目容易不均。相较于等宽法，等频法避免了类分布不均匀的问题，但同时也有可能将两个非常接近的数值划分到不同的区间，以满足等频对每个区间数据个数的要求。相较于上述两种方法，在对连续数据进行离散化转换时，可以通过聚类算法将连续变量进行聚类划分处理，根据聚类结果将某一类连续属性值表述为其潜在的某种特征类型。连续数据的离散化过程主要包括确定离散区间准则和将数据属性按照一定规则划分 [4]。由于 K-means 聚类算法基于数据间距进行分析，综合考虑了各连续数据点的邻近性，因此 K-means 聚类算法在解决连续数据离散化中有较为不错的表现 [5]。简单易懂、时间复杂度低的 K-means 算法为数据离散化提供了极高的计算效率。该算法具体步骤如下：

输入：K 个聚类簇数目；D：包含有 n 个对象的数据集。

输出：K 个簇的集合及类编号。

1）随机选取 D 中 K 个样本作为初始聚类中心。

2）计算每个样本与初始聚类中心间的距离并根据距离分配相应的聚类簇。

3）移动聚类中心，选定新的聚类中心为聚类簇重心。

4）重复步骤 2），直至目标函数最优 E 值最小且不再发生变化，则算法结束。

其中，样本间的距离计算采用欧氏距离，其计算公式如下：

$$\text{Dist}(X_i, X_j) = \sqrt{(X_i - X_j)^T (X_i - X_j)} \tag{1}$$

聚类中心计算公式如下：

$$c_k = \frac{1}{n_k} \sum_{x_i \in c_k} x_i \tag{2}$$

目标函数最优 E 值计算公式如下：

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_j} \text{Dist}(X_j, c_i) \tag{3}$$

为检验该算法针对不同数据聚类性能，本文利用数据集对 K-means 算法进行分析与评估，选取人工合成连续化数据集 R15[6] 和标准数据集 Seeds，其中 R15 数据集包含 2 个维度、15 种类别的 600 份样本，并含有桥接噪声或随机噪声；Seeds 数据集包含 7 个

维度，3 种类别的 210 份样本。评价指标本文选取准确度 ACC（Clustering Accuracy），具体为：

$$\mathrm{ACC}=\frac{1}{n}\sum_{i=1}^{n}y_i \tag{4}$$

其中，y_i 为第 i 簇中聚类正确的数据点个数，准确度取值范围在 0 到 1 之间，值越大表示聚类结果越准确 [7]。

由于 K-means 聚类算法每次随着聚类中心选取不同，对结果会产生一定波动 [8-10]。为避免实验结果的偶然性，提高实验准确性，本文选择在每个数据集上运行 20 次，取每次聚类结果评价指标 ACC 的平均值作为最终结果。实验结果如表 3 所示。

表 3　K-means 算法在不同数据集下的准确度

评价指标	R15 数据集	Seeds 数据集
ACC	0.817 5	0.879 8

根据上述实验结果可以看出，K-means 算法对数据标准且离散化的数据集以及包含噪声数据且连续化的数据集均能有效果不错的、稳定的准确率。说明该方法针对驾驶行为这类连续化且存在噪声的海量数据，进行离散化处理并根据聚类效果分析不同类别下的隐性特征是有效可行的。

3　驾驶行为离散化特征聚类分析

上述实验已经检验了 K-means 聚类算法对连续化数据的稳定聚类性能，因此，将处理后的驾驶行为特征数据带入进行操作及分析。首先结合轮廓系数确定该驾驶行为数据集的最佳聚类簇数。轮廓系数是评价聚类效果好坏的一种简单评价方式，假设有一点 i，记 i 向量到其所属的簇中的其他所有点的平均值为 $a_{i,\ i}$ 向量到某一不包含该点的簇中的所有点的平均距离的最小值为 b_i，可将 i 向量轮廓系数表示为：

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{5}$$

轮廓系数取值范围在 -1 到 1 之间，值越趋近于 1 表示内聚度和分离度都越好。将所有点的轮廓系数求平均值，就是该数据集聚类结果的总体轮廓系数。

将不同聚类簇数在驾驶行为特征数据集上进行聚类，根据其轮廓系数结果可以看出，当聚类簇数为 3 时，轮廓系数最大，如图 1 所示，表明针对该驾驶行为数据集最佳聚类簇数为 3。

因此将该驾驶行为数据集聚类簇数设定为 3，采用平均速度、最高速度、急加速、急减速、急刹车、疲劳驾驶作为特征项，采用 K-means 算法进行离散化聚类分析。为避免聚类中心点的选取对实验结果带来的偶然性，提高实验准确性，本文在该驾驶行为数据集中运行 20 次，将每次运行结果的聚类中心数据取平均值，作为最终的聚类中心结果。离散化聚类处理效果如图 2 所示，为呈现聚类效果，仅利用平均速度和最大速度进行绘图，但在聚类划分过程中选取的特征数据为前文处理后获得的 6 个特征项。

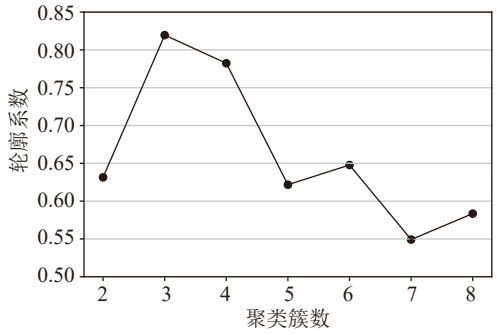

图 1　驾驶行为数据中不同聚类簇数的轮廓系数

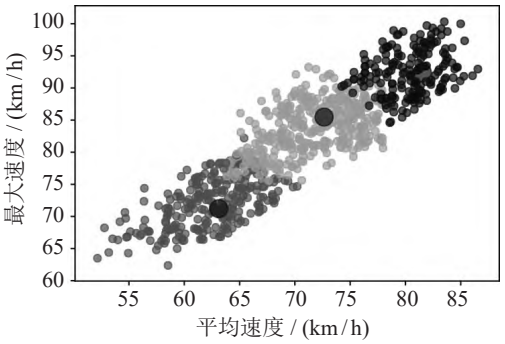

图 2　驾驶行为数据离散化聚类处理效果图

因此，基于 K-means 算法对驾驶行为进行特征离散化聚类，形成 3 类安全性能驾驶行为分簇。各类数据簇的聚类中心点结果如表 4 所示。根据其数据特征，本文将其概括为平稳型驾驶、冲动型驾驶和危险型驾驶。

表 4　驾驶行为特征聚类中心数据

类型	平均速度 /（km/h）	最高速度 /（km/h）	急加速 / 次	急减速 / 次	急刹车 / 次	疲劳驾驶 / 次
平稳型驾驶	63.12	71.22	2.4	2.3	1.8	0
冲动型驾驶	72.64	85.47	8.6	7.9	5.7	0.3
危险型驾驶	81.53	92.38	13.4	11.5	9.2	1.4

根据表中聚类中心点特征数据来看，第一类整体车速都相对缓和平稳，没有过多的急加速、急减速和急刹车，同时不存在疲劳驾驶这项危险驾驶行为，可将此类驾驶行为归为“平稳型驾驶”；第二类车速相对第一类有一定的提高，相对还算稳定，和前一类别相比，在急加速、急减速和急刹车方面次数有明显增多，说明此类驾驶行为在行车过程中，驾驶人员经常习惯性猛踩刹车急停或猛踩油门加速，尽管最高速度在正常限速范围下，此类行为在某些情况下有可能酿

成交通事故，同时在该类别中，有少量驾驶人员存在一定的疲劳驾驶情况，可将此类驾驶行为归为“冲动型驾驶”；第三类整体车速过快，平均车速较高，同时急加速、急刹车等情况最多，在反复的猛踩油门加速过程中，使得中心点最高车速较为接近路段限速范围。同时在该类别中，普遍存在疲劳驾驶的情况且次数较多，另外根据具体数据样本，还监测到有数次超速情况发生。此类疲劳驾驶、超速行驶等危险驾驶行为是导致交通事故发生的重要因素，由此可将此类驾驶行为归为“危险型驾驶”。

4 结 论

本文针对驾驶行为的连续化原始数据，通过相应的数据预处理对驾驶行为特征进行提取，结合 K-means 聚类方法，对驾驶行为数据集进行离散化聚类处理与分析，最终在该数据集上获得三类代表性驾驶行为归类。根据每个类别中心点特征和该类别下数据样本特征，对其相应的驾驶行为特征数据进行分析，挖掘驾驶行为下的隐性特征，分别将其三个类别归为“平稳型驾驶”“冲动型驾驶”和“危险型驾驶”。根据驾驶数据可以看出，“平稳型驾驶”体现了大部分防御性驾驶人员的驾驶行为习惯，整体驾驶较为缓和平稳；“冲动型驾驶”体现了当下很多情绪急躁驾驶人员的行车习惯，尽管车速不算太快，但是习惯性猛踩油门或猛踩刹车，容易造成追尾，存在一定的安全隐患；“危险型驾驶”体现了一些存在“路怒症”或是对自己极度自信的驾驶人员的行车习惯，由于存在疲劳驾驶和超速行驶等危险行为，极易发生交通事故。因此，可以看出本文通过基于 K-means 的驾驶行为离散化特征分析，根据实际行车数据有效地进行了不同代表性驾驶行为的划分，为后续进一步根据驾驶行为数据进行关联分析数据挖掘提供了依据。

参考文献：

[1] 王万丰 . 我国道路交通安全事故统计分析 [J]. 中国安全生产，2020，15（3）：52-53.

[2] XING Y，LYU C，WANG H J，et al. Driver Activity Recognition for Intelligent Vehicles: A Deep Learning Approach [J].IEEE Transactions on Vehicular Technology，2019，68（6）：5379-5390.

[3] 廖纪勇 . 基于聚类和关联规则的驾驶行为分析与研究 [D]. 昆明：昆明理工大学，2021.

[4] 孙吉贵，刘杰，赵连宇 . 聚类算法研究 [J]. 软件学报，2008，19（1）：48-61.

[5] 张良均，杨坦，肖刚，等 .MATLAB 数据分析与挖掘实战 [M]. 北京：机械工业出版社，2015.

[6] 于彦伟，贾召飞，曹磊，等 . 面向位置大数据的快速密度聚类算法 [J]. 软件学报，2018，29（8）：2470-2484.

[7] NIE F P，WANG C L，LI X L. K-Multiple-Means: A Multiple-Means Clustering Method with Specified K Clusters [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.Anchorage：ACM Press，2019：959-967.

[8] 江文奇，黄容，牟华伟，等 . 面向大规模数据精简的聚类中心点优化和 FCM 算法设计 [J]. 数学的实践与认识，2021，51（17）：144-151.

[9] 罗兴隆，贺兴时，杨新社 . 二分 k-means 锚点提取的快速谱聚类 [J]. 计算机工程与应用 .2023，59（16）：74-81.

[10] 姜子超 . 基于秃鹰搜索算法优化 K-Means 的动态特征子集聚类研究 [D]. 哈尔滨：东北林业大学，2022.

作者简介：宋月亭（1995. 10—），女，汉族，山东济宁人，助教，硕士，主要研究方向：人工智能与数据挖掘。

（上接 16 页）

[2] 宋永生，黄蓉美，王军 . 基于 Python 的数据分析与可视化平台研究 [J]. 现代信息科技，2019，3（21）：7-9.

[3] 钟机灵 . 基于 Python 网络爬虫技术的数据采集系统研究 [J]. 信息通信，2020（4）：96-98.

[4] 任妮，吴琼，栗荟荃 . 数据可视化技术的分析与研究 [J]. 电子技术与软件工程，2022（16）：180-183.

[5] 谢美英 . 基于 Anaconda 的婴儿用品数据爬取及可视化分析 [J]. 现代信息科技，2021，5（14）：90-93.

[6] 冯洪熙，王林，魏嘉银，等 . 基于回归分析的网络招聘信息爬取及可视化 [J]. 现代信息科技，2021，5（10）：1-5.

[7] 刘宇韬，施莉，刘诗含 . 基于 TF-IDF 与 Word2vec 的用户评论分析研究 [J]. 成都航空职业技术学院学报，2022，38（4）：89-92.

[8] 钟晓旭 . 基于 Web 招聘信息的文本挖掘系统研究 [D]. 合肥：合肥工业大学 .

[9] 殷漫漫 . 基于电商化妆品评论主题的挖掘研究——以京东平台化妆品为例 [J]. 营销界，2022（21）：161-163.

[10] 冯晓磊 . 基于 Python 的拉勾网网络爬虫设计与实现 [J]. 现代信息科技，2023，7（6）：85-87+91.

[11] 陈佳楠 . 招聘网站中数据分析类岗位的现状及其影响因素 [D]. 桂林：广西师范大学，2020.

[12] 刘畅 . 基于 Web 文本挖掘的数据分析岗位需求研究 [J]. 中国管理信息化，2018，21（10）：76-79.

[13] 涂晓彬 . 基于大数据技术的网络招聘岗位需求分析方案 [J]. 信息技术与信息化，2022（12）：31-34.

作者简介：王姣姣（1994—），女，汉族，河南洛阳人，助教，硕士，研究方向：大数据技术、计算机应用；姚华平（1976—），女，汉族，河南洛阳人，讲师，硕士，研究方向：软件工程、计算机应用。