文章编号:1005-0523(2023)05-0049-10

基于改进K-prototypes与GBDT的城市干道车辆出行群体辨识模型

梁 灯[1,3],蔡晓禹[2,3],彭 博[2,3],邢茹茹[1,3]

(1. 重庆交通大学交通运输学院,重庆 400074;2. 重庆交通大学智慧城市学院,重庆 400074;
3. 重庆交通大学山地城市交通系统与安全重庆市重点实验室,重庆 400074)

摘要:为了掌握城市干道交通运行规律,向交通管理部门制定相关交通需求管理政策提供理论依据,提出了一种基于组合模型的城市干道车辆出行群体辨识模型。基于青岛市胶州湾隧道过车数据,从出行强度、出行时间与出行习惯3个维度构建了出行特征指标体系以全面刻画车辆个体的出行行为。基于相关性分析剔除了冗余指标以避免对辨识研究的影响。针对混合属性出行特征指标数据,使用改进K-prototypes算法以有效地实现车辆出行群体划分,将其与GBDT算法相结合,建立了一种基于改进K-prototypes与GBDT的辨识模型,随机选取10 000个样本开展辨识研究。结果表明:研究道路存在5类车辆出行群体:高频通勤群体、低频通勤群体、营运群体、频次稳定群体与普通群体,对于这5类车辆出行群体,平均识别准确率为97.75%,最高识别准确率可达99.47%。

关键词:城市道路交通;群体辨识;出行特征;改进K-prototypes & GBDT

中图分类号:U491.4 **文献标志码**:A

本文引用格式:梁灯,蔡晓禹,彭博,等. 基于改进K-prototypes与GBDT的城市干道车辆出行群体辨识模型[J]. 华东交通大学学报,2023,40(5):49-58.

DOI:10.16749/j.cnki.jecjtu.20230508.017

Vehicle Travl Group Identification Model of Urban Arterial Road Based on Improved K-prototypes and GBDT

Liang Deng[1,2], Cai Xiaoyu[1,3], Peng Bo[1,3], Xing Ruru[1,2]

(1. College of Traffic and Transportation, Chongqing Jiaotong University, Chongqing 400074, China;
2. College of Smart City, Chongqing Jiaotong University, Chongqing 400074, China; 3. Chongqing Key Laboratory of Traffic System&Safety in Mountainous Cities, Chongqing Jiaotong University, Chongqing 400074, China)

Abstract: In order to identify the traffic operation law of urban arterial road and support basis for traffic management departments to formulate relevant traffic demand management policies, a vehicle travel group identification model of urban arterial road based on combined model was proposed. In this study, a travel characteristic indicator system was constructed from dimensions of travel intensity, travel time, travel habits for comprehensively describing the travel behavior based on the traffic bayonet data of Qingdao Jiaozhou Bay Tunnel. The redundant indicator was eliminated based on the correlation analysis to avoid the impact on identification research. For the mixed attribute travel characteristic indicator data, the improved K-prototypes algorithm was used to effectively classify the vehicle travel groups, and combined with GBDT, the identification model based on improved

收稿日期:2023-02-15
基金项目:重庆市技术创新与应用发展专项重点项目(CSTB2022TIAD-KPX0104)

K-prototypes and GBDT was established. By randomly selecting 10 000 samples to conduct identification research, the result shows that there are 5 vehicle travel groups for the road in this research, including high-frequency commuter groups, low-frequency commuter groups, operating groups, frequency stable groups, and ordinary groups. For the 5 vehicle travel groups, the average identification accuracy rate exceeds 97.75%, and the highest identification accuracy rate can reach 99.47%.

Key words: urban road traffic; group identification; travel characteristic; improved K-prototypes & GBDT

Citation format: LIANG D, CAI X Y, PENG B, et al. Vehicle travl group identification model of urban arterial road based on improved K-prototypes and GBDT[J]. Journal of East China Jiaotong University, 2023, 40(5): 49-58.

城市干道作为城市交通系统的重要组成部分，往往承担着很大比重的交通出行，交通拥堵也较多集中发生于城市干道，严重影响着居民的正常生活。对城市干道的车辆出行群体进行分类研究有助于掌握其交通运行规律，可为交通管理部门制定相关交通需求管理政策提供理论依据[1]，是保障城市交通系统正常运行的重要举措。

出行群体分类是将出行者划分为具有相似出行规律的各类群体，分类结果能够揭示出行者之间的共性与异质性。依据分类的方法，可分为基于监督学习的分类与基于聚类的分类。

基于监督学习的分类依赖于有标签的数据标定分类器的参数，在出行者类别难以人工标注的情况下，此类研究的开展多辅以额外的出行调查，如梁泉等[2]基于北京市连续1个月公交刷卡数据提取特征指标，结合RP调查结果构建了面向公交通勤乘客识别的BP神经元网络模型。崔洪军等[3]基于调查数据标定朴素贝叶斯分类器参数，进而对智能刷卡数据中缺少的出行目的属性加以补充。

不同的是，基于聚类的分类直接从数据构建出行特征指标，采用聚类算法实现出行者的自动划分，如Mohamed等[4]基于法国雷恩市的乘客刷卡数据，将乘客每周每天每小时对应的平均出行次数聚合成向量，使用K-means算法将乘客分为16类。刘凯[5]根据AFC刷卡数据的特点，基于DBSCAN算法提取乘客规律性特征，以两步聚类法将乘客分为3类。程小云等[6]基于AFC刷卡数据，针对工作日有出行的一卡通用户提取出行天数、出行集中度等特征，以GMM算法将乘客分为5类。

受限于出行调查所耗费的大量人力、财力，基于监督学习的分类通常只进行是非通勤的判别[7-9]。相比之下，基于聚类的分类避免了出行调查，适合用于大样本数据集挖掘任务，因而成为目前大多数研究所采用的方法，国内外学者基于K-means[4,10,11]、K-means++[12-14]、DPC[15-16]、OPTICS[17]、LDA[18]、GMM[6,19]等算法开展出行群体分类研究，取得了丰富的研究成果。然而，目前的出行群体分类研究主要针对的是公交出行群体，对于车辆出行群体而言，其分类方法的选择亟需深入探讨。

分类方法的合适选择往往能有效地划分群体，而其实现的前提是构建出行特征指标体系以全面刻画出行者个体的出行行为。一般来说，应尽可能从多个维度构建指标，然而对于不同的数据而言，可获取的指标有所不同，需根据研究数据定制化选取指标。过车数据是某特定道路的车牌号识别数据，记录了车辆每次经过道路的车牌号、通行时间与方向、车辆类型等信息，具有准确性高、数据量大等优点，是研究车辆出行行为的良好数据源。

因此，本文以城市干道车辆出行群体为研究对象，基于过车数据构建多维出行特征指标体系以全面刻画车辆出行行为，考虑出行特征指标体系特点选取聚类算法，将聚类与监督学习相结合构建基于改进K-prototypes与GBDT的辨识模型，以期为交通管理部门制定相关交通需求政策提供理论依据。

1 车辆出行特征指标体系构建

1.1 数据来源

研究数据来源于青岛市胶州湾隧道2021年10月11日—31日(共计21 d)小型客车的过车数据。本文主要基于过车数据中车牌号、通行时间(具体到秒)与通行方向(2个方向)字段开展研究，数据格式如表1所示。对原始数据中车牌号缺失、车牌号

识别错误、车辆连续两次通行时间间隔过小等数据予以剔除，最终保留 160 万余条。

1.2　车辆出行特征指标体系构建

基于前期的统计分析发现，研究时间范围(21 d)内车辆以偶然出行为主，为确保所构建出行指标的有效性，本文只针对出行天数大于 3 的车辆构建指标。参考已有的研究并结合过车数据的特点，构建车辆出行特征指标体系，具体如表 2 所示。

表 1　过车数据主要字段

Tab.1　Primary field of the traffic bayonet data

Field name	Field comment
Vehicle plate	The identification of vehicle
Passage time	Time when the vehicle passes the detection section
Passage direction	Direction that the vehicle passes the detection section

部分出行特征指标的提取方法如下。

1) 首次/末次最频繁出行时段。首次/末次最频繁出行时段表征的是车辆多日出行活动中首次/末次出行最集中的时段；由于车辆的出行时间是连续值，需将连续的时间值划分为离散的时段。将 1 天 24 h 划分为 5 个时段：[00:00,06:30)，[06:30,10:00)，[10:00,16:30)，[16:30,19:30)，[19:30,24:00)，编号为 1~5。

2) 出行频次模式重复性。出行频次模式重复性表征车辆在多日出行活动中以相同出行频次进行的稳定程度，出行频次模式重复性定义为

$$H^{pc}(x)=-\sum_{x\in\Omega_x}p(x)\log_2 p(x) \tag{1}$$

式中：Ω_x 为出行频次模式 X 的取值空间；$p(x)$ 为 $X=x$ 的概率。需要说明的是，由于本文采用熵值作为出行频次模式重复性的度量，因此，当 H^{pc} 为 0 时，则代表车辆每天完全以相同频次出行；H^{pc} 大于 0 时，H^{pc} 越大则代表车辆重复性越低。

3) 出行时空模式重复率。出行时空模式重复率表征车辆以相同(时间-方向)模式出行的概率，以 30 min 为粒度将 24 h 划分为 48 个互不相交的时段，以(时间-方向)模式表示车辆的 1 次出行，则出行时空模式重复率 a 为

$$a=\frac{n}{N} \tag{2}$$

式中：n 为重复出现模式的个数；N 为出行时空模式的总数。

为避免冗余特征指标的影响，需进行指标之间的相关性分析。对于混合属性出行特征指标体系，根据文献[20]的方法得到相关性结果如图 1 所示。

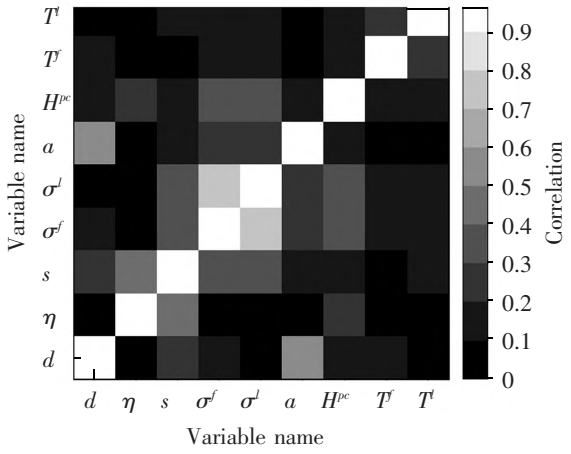


图 1　相关性分析结果

Fig.1　The result of the correlation analysis

表 2　车辆出行特征指标体系

Tab.2　Vehicle travel characteristic indicator system

Characteristic dimension	Characteristic indicator	Variable name	Value range	Type
Travel intensity	Travel days	d	[4,21]	Numerical
	Standard deviation of weekly travels	η	$[0,\eta_{max}]$	Numerical
	Average of daily travels	s	$[1,s_{max}]$	Numerical
Travel time	Standard deviation of first travel time	σ^f	$[0,\sigma^f_{max}]$	Numerical
	Standard deviation of last travel time	σ^l	$[0,\sigma^l_{max}]$	Numerical
Travel habits	Repetition rate of travel space-time pattern	a	[0,1]	Numerical
	Repeatability of travel frequency pattern	H^{pc}	$[0,H^{pc}_{max}]$	Numerical
	Most frequent first travel period	T^f	[1,2,3,4,5]	Classified
	Most frequent last travel period	T^l	[1,2,3,4,5]	Classified

由图1可知,“首次出行时间标准差”与“末次出行时间标准差”之间具有强相关性。本文将末次出行时间标准差剔除,最终保留出行天数、日均出行次数、周出行次数标准差、首次出行时间标准差、出行频次模式重复性、出行时空模式重复率与首/末次最频繁出行时段共8个指标开展辨识研究。

2 基于改进K-prototypes与GBDT的车辆出行群体辨识模型

2.1 改进的K-prototypes算法

K-prototypes是由Huang[21]提出的一种可有效解决混合型数据聚类问题的算法,本文构建的出行特征指标体系为数值和分类混合属性指标,采用该算法有较好的聚类效果。此外,针对原始算法相异度计算公式与初始聚类中心选取的不足加以改进。

给定样本量为n的数据集$D=[x_1,x_2,\cdots,x_n]$,样本i表示为$x_i=[x_i^1,\cdots,x_i^p,x_i^{p+1},\cdots,x_i^m]$,$x^1\sim x^p$为数值属性,$x^{p+1}\sim x^m$为分类属性。聚类过程中,类簇集合为$c=[c_1,c_2,\cdots,c_k]$,其中$k$为类簇个数,$k\geqslant 2$;聚类中心点集合表示为$z=[z_1,z_2,\cdots,z_k]$,其中$z_1$表示类簇$c_1$的中心点。

针对原始算法中数值属性部分仅使用简单欧氏距离求得样本之间的相异度,未考虑样本被分到类簇中各属性贡献大小的差异,本文使用标准差系数法以客观地确定各属性的权值。对于数值属性s,其标准差为

$$\sigma_s=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^s-\bar{x}^s)^2} \tag{3}$$

式中:x_i^s为样本i在属性s上的取值;$\bar{x}^s$为属性s的均值。则各属性的权重为:$w_s=\sigma_s/\sum\sigma_s$。

将数值属性部分的相异度定义为

$$d_r(x_i,x_j)=\left[\sum_{s=1}^{p}w_s(x_i^s-x_j^s)^2\right]^{1/2} \tag{4}$$

针对原始算法中分类属性部分仅使用属性出现频率最高的值作为聚类中心,而未考虑类簇中各属性取值的分布情况,根据参考文献[22]的思想将分类属性部分相异度定义为

$$d_c(x_i,x_l)=\sum_{s=p+1}^{m}\sum_{j}\frac{\delta(x_i^s,c_l^{js})}{|c_l|} \tag{5}$$

式中:$\delta(x_i^s,c_l^s)=\begin{cases}0, & x_i^s=c_l{}^{js}\\ 1, & x_i^s\neq c_l{}^{js}\end{cases}$,$x_i^s$为样本$i$在属性$s$上的取值,$c_l^{js}$为类簇$c_l$中样本$j$在属性$s$上的取值;$|c_l|$为类簇$c_l$中已有样本个数。

依据式(4)与式(5)将混合属性相异度定义为

$$d_g(x_i,z_l)=d_r(x_i,z_l)+\gamma d_c(x_i,c_l) \tag{6}$$

相异度计算公式改进后的K-prototypes算法的目标函数定义为

$$F(x,z)=\sum_{i=1}^{n}\sum_{l=1}^{k}u_{il}d_g(x_i,z_l) \tag{7}$$

式中:u_{il}为样本i对于类簇c_l的隶属度;u_{il}为0时,表示样本i样本未被划分到类簇c_l中;u_{il}为1时,表示样本i样本被划分到类簇c_l中。

针对原始算法初始聚类中心的选取采取随机方式导致聚类稳定性差,本文基于DPC算法生成的样本局部密度与距离的二维坐标决策图来选取初始聚类中心。DPC算法原理及具体步骤请参考文献[23-24]。综上得到改进K-prototypes算法的步骤如下

1)输入数据集D,计算样本之间距离以构建相异度矩阵,数值属性根据式(4)计算,分类属性采取汉明距离;

2)根据决策图选取初始聚类中心并输出;

3)根据式(6)计算样本点与各聚类中心的相异度,将各样本点划分到与其相异度最小中心点所对应类簇中;

4)基于类别划分后的数据,更新各类簇的中心。数值属性以该类簇中数值属性的平均值作为新的聚类中心;分类属性使用该类簇中出现频率最高的分类属性值作为新的聚类中心;

5)重复步骤3)和5),直到目标函数值收敛或者达到预设的迭代次数为止,输出聚类结果。

为验证本文改进K-prototypes算法的有效性与可行性,使用UCI数据库的真实的混合属性数据集Statlog Heart(SH)、Credit Approval(CA)与Australia Credit Approval(ACA)进行验证,选取K-prototypes(KP)与Fuzzy K-prototypes(FKP)算法进行比较。验证数据集的描述如表3所示。

为了评估算法聚类效果,采用正确率(AC)与类精度(PE)作为评价指标,其定义如下

$$AC=\frac{1}{t}\sum_{i=1}^{k}t_i \tag{8}$$

表 3　验证数据集描述

Tab.3　Description of validation data sets

Data set	Number of samples	Number of numerical attributes	Number of classified attributes	Number of clusters
SH	270	6	7	2
CA	690	6	9	2
ACA	690	6	8	2

$$PE = \frac{1}{k}\sum_{i=1}^{k}\frac{t_i}{ct_i} \tag{9}$$

式中：t 为数据集的样本个数；k 为类簇数；t_i 为第 i 个类簇中被正确划分的样本数；ct_i 为聚类结果中第 i 个类簇的样本数。AC 与 PE 的值越大代表聚类效果越好。

由于量纲的不同，各属性的取值差异性较大，为了增加聚类准确性与减少计算复杂性，聚类之前，本文对验证数据集的数值属性部分数据均采取 min-max 归一化处理，即将取值控制在[0,1]，公式如下

$$x^*=(x-x_{\min})/(x_{\max}-x_{\min}) \tag{10}$$

式中：x^* 为归一化后的取值；$x_{\min}$，$x_{\max}$ 为数据集中对应数值属性下的最小值与最大值。

对于随机选取初始聚类中心的 KP 与 FKP 算法，在各数据集上给出聚类个数后均重复实验 50 次取均值作为最终结果。在各数据集上将所有算法共有的参数 γ 设置为对应数据集中分类属性个数与数值属性个数的比值[25]，此外，FKP 算法中的模糊因子均设置为 2，本文算法中 DPC 算法中近邻占比 P_d 均设置为 1.5%，即将对应数据集的距离 δ 按升序排列后位于第 1.5%的值作为截断距离 d_c 取值。

在各数据集上的实验结果比较如表 4 所示。从表 4 可以看出，本文改进后的 K-prototypes 的聚类效果要明显优于 KP 与 FKP 算法，由此验证了本文改进 K-prototypes 算法的有效性与可行性。

表 4　验证数据集上的实验结果比较

Tab.4　Comparison of experimental results in validation data sets

Data set	Evaluating indicator	KP algorithm	FKP algorithm	Algorithm in this study
SH	*AC*	0.762	0.765	0.848
	PE	0.774	0.773	0.848
CA	*AC*	0.722	0.729	0.790
	PE	0.723	0.736	0.804
ACA	*AC*	0.778	0.775	0.854
	PE	0.767	0.787	0.857

2.2　GBDT 算法

GBDT 全称梯度提升决策树，是统计学习性能最好的方法之一[26]，在解决各种领域如城市交通、电力、医学等领域的分类问题和回归问题上均表现出优异的性能。GBDT 的核心思想是利用损失函数的负梯度在当前模型的值作为算法中的残差近似值，通过不断拟合残差从而使残差不断减少。在分类问题中，GBDT 采用交叉熵损失函数或者对数似然损失函数，GBDT 算法原理请参考文献[26]。

2.3　车辆出行群体辨识模型

车辆出行群体辨识包括 3 个部分：出行特征指标数据集构建、基于改进 K-prototypes 的车辆出行群体划分与基于 GBDT 的车辆出行群体识别。车辆出行群体辨识流程如图 2 所示。

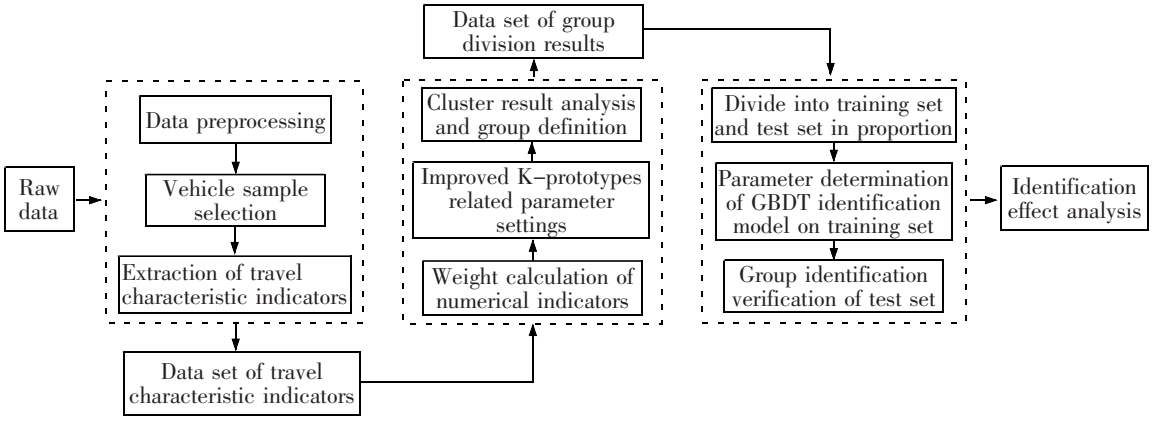


图 2　车辆出行群体辨识流程图

Fig.2　Structure of vehicle travel group identification method

3 实验分析

3.1 数据准备

经过统计分析，研究数据中出行天数大于3的车辆共73 124辆，占总体的19.6%，但该部分车辆出行量占总体的58.7%，针对该部分车辆进行分类可有效研究掌握道路出行规律。此外，由于需要构建相异度矩阵以选取初始聚类中心，样本量大会导致计算时间过长，本文随机选取10 000个车辆样本开展辨识研究。对于出行特征指标数据中的数值型部分，采取Min-max归一化处理，处理方式见式(10)。

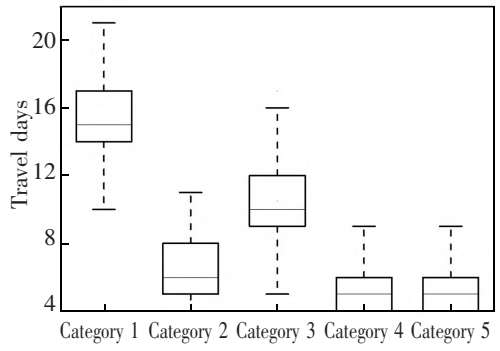
3.2 车辆出行群体划分

基于归一化后的特征指标数据集，使用标准差系数法求得各数值指标的权重。K-prototypes算法中参数设置为0.333，迭代次数设置为100次。DPC算法中近邻占比P_d设置为1%。得到初始聚类中心5个，而最终聚类结果如表5所示。绘制不同群体的出行特征指标的分布情况，具体如图3所示。

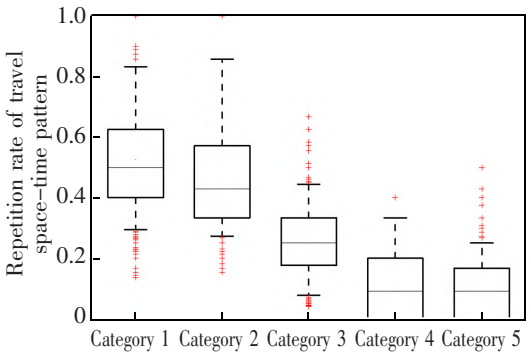
表5 聚类结果

Tab.5 The result of clustering

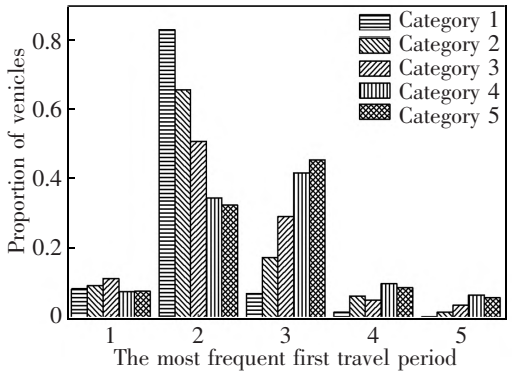
Variable name of the indicator	Cluster centers				
	1	2	3	4	5
d	15.627	6.539	10.883	5.010	5.322
η	1.502	1.901	2.278	1.254	1.608
s	1.849	1.728	1.713	1.486	1.617
σ^f	2.334	2.082	4.522	3.303	3.846
a	0.526	0.482	0.252	0.099	0.092
H^{pc}	0.498	0.275	1.033	0	0.929
T^f	2	2	2	3	3
T^l	4	4	4	3	3



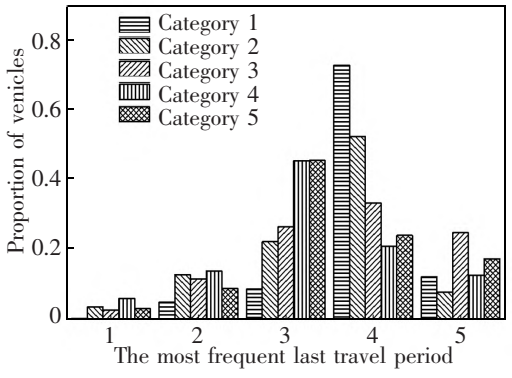
(a) Distribution of travel days



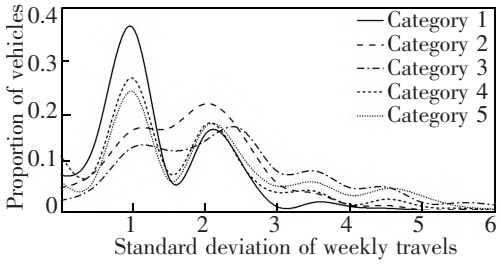
(b) Distribution of repetition rate of travel space-time pattern



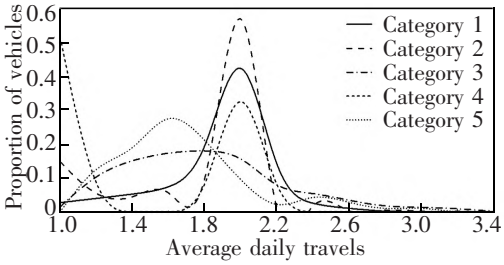
(c) Distribution of the most frequent first travel period



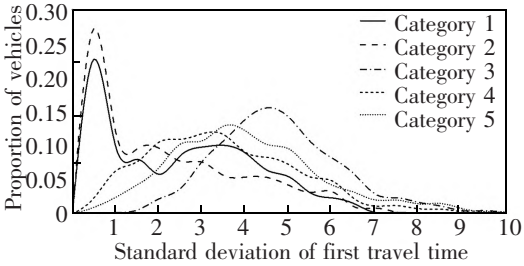
(d) Distribution of the most frequent last travel period



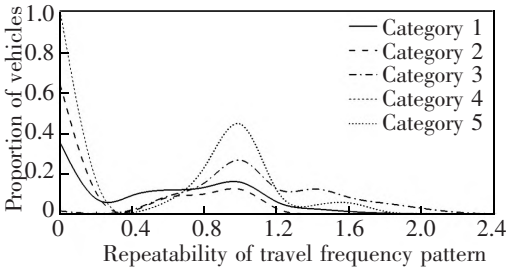
(e) Distribution of the standard deviation of weekly travels



(f) Distribution of the average daily travels



(g) Distribution of standard deviation of the first travel time



(h) Distribution of repeatability of the travel frequency pattern

图 3　不同群体出行特征指标分布

Fig.3　The distribution of travel characteristic indicator for different groups

根据聚类结果与不同群体的出行特征指标分布情况定义群体。

1) 类别 1 与类别 2：类别 1 与类别 2 出行特征基本相似，日均出行次数集中于 2 次、首次出行时间稳定、出行时空模式重复率高、出行频次模式重复性强、首/末次最频繁出行时段分别集中于 2 和 4，与通勤群体的特征相符合。与类别 1 相比，类别 2 出行天数少，周出行次数波动性较大。综上，将类别 1 定义为高频通勤群体，类别 2 定义为低频通勤群体。

2) 类别 3：出行天数较多、日均出行次数分布散乱且其日均出行次数大于 2 的车辆数占比大于其他类别、周出行次数与首次出行时间波动性大、出行时空模式重复率低、出行频次模式重复性弱、首/末次最频繁出行时段分布较为散乱，与道路上出行的出租车、网约车的特征相符合。综上，结合实际将类别 3 定义为营运群体。

3) 类别 4：类别 4 出行天数少、日均出行次数小、周出行次数的标准差较小、首次出行时间标准差较大、出行时空模式重复率基本为 0、首次最频繁出行时段集中在 2 与 3、末次最频繁出行时段集中在 3。但其出行频次模式重复性为 0，说明此类车辆在其多日出行活动中每天以相同频次出行。综上，将类别 4 定义为频次稳定群体。

4) 类别 5：类别 5 在出行天数、出行时空模式重复率、首次最频繁出行时段、末次最频繁出行时段上、周出行次数标准差分布与类别 4 相似，但在日均出行次数、首次出行时间标准差、出行频次模式重复性的分布与类别 3 相似，说明此类群体并无明显的规律。综上，将类别 5 定义为普通类。

3.3　车辆出行群体识别

基于 3.2 节聚类后获取群体划分结果数据集，根据 2.2 节所提算法开展群体识别研究。

由于 GBDT 算法无法直接处理分类型指标，建立识别模型前需要对首/末次最频繁出行时段指标进行 One-hot 编码处理，以 0 和 1 来解释属性。首次最频繁出行时段编码处理示例如表 6 所示。

本文采用 Python 机器学习库的中 Sklearn 模块进行车辆出行群体识别建模，由于本文的车辆出行群体识别是多分类问题，因此，将损失函数设置为交叉熵损失函数，其他部分参数[27]如表 7 所示。

将数据集按照 8∶2 的比例划分为训练集与测试集。为了避免过拟合与欠拟合问题的出现，在训练集上，以识别准确率为评价指标，针对 Learning_rate

表6 首次最频繁出行时段编码

Tab.6 Coding of the most frequent firsts travel period

Vehicle umber	T^f	One-hot coding				
		1	2	3	4	5
1	3	0	0	1	0	0
190	3	0	0	1	0	0
1 156	2	0	1	0	0	0

表7 GBDT部分参数

Tab.7 Partial parameters of GBDT

Parameter	Comment
Learning_rate	For the same fitting effect, smaller value of learning_rate means that more basic models need to be iterated
N_estimators	Taking too small value is easy to cause under-fitting. Taking too large value is easy to cause over-fitting
Subsamples	The sampling proportion used to build the basic model, generally set to [0.5,0.8]
Max_depth	Maximum depth of each basic model, which depends on the complexity of data

与N_estimators进行调参。其中,Max_depth设置为5,Subsamples设置为0.8。调参步骤如下。

1) 将Learning_rate的初始值设置为0.3,采用5折交叉验证法对N_estimators寻优。

2) 降低参数Learning_rate的值,并按比例增加参数N_estimators,采用5折交叉验证法寻找使得识别准确率最高的参数组合。重复此步骤,得到不同Learning_rate下的最优N_estimators值,如表8所示。

表8 不同Learning_rate下的最优组合

Tab.8 Optimal combination under different Learning_rate

Learning_rate	N_estimators	Identification accuracy rate
0.3	100	0.974 6
0.15	200	0.974 3
0.1	300	0.974 7
0.05	600	0.974 8
0.01	3 000	0.976 1

3.4 群体识别结果分析

基于3.3节得所有Learning_rate下的最优组合(0.01,3 000)在训练集上建立GBDT识别模型,在测试集上得识别结果如图4所示。

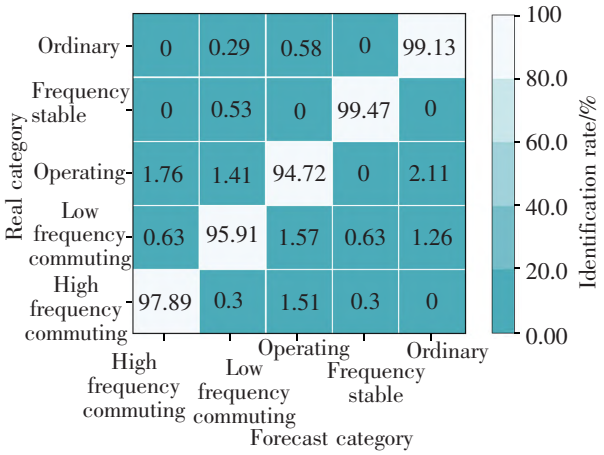


图4 群体识别混淆矩阵图

Fig.4 Lithology confusion matrix of group identification

由图4可知,5类车辆出行群体识别准确率分别为97.89%,95.91%,94.72%,99.47%与99.13%。平均识别准确率约为97.42%,最高可达99.47%。

4 结论

1) 针对城市干道车辆,在考虑过车数据特点的基础上,从出行强度、出行时间与出行习惯3个维度构建了全面刻画车辆个体出行行为的出行特征指标体系,并基于相关性分析剔除了冗余的指标,保障了所构建指标体系的合理性。

2) 针对城市干道车辆出行群体的含有数值型与分类型指标的出行特征指标数据,使用相异度计算公式改进与基于密度峰值聚类算法的二维(局部密度-距离)决策图,选取初始聚类中心的改进K-prototypes算法划分得到了5类典型群体:高频通勤群体、低频通勤群体、营运群体、频次稳定群体与普通群体。基于群体划分结果数据集,在训练集上确定了GBDT群体识别模型的关键参数,测试集上进行了群体识别验证,对于这5类群体,平均识别准确率为97.42%,最高可到99.47%,识别效果良好。

3) 采用改进K-prototypes与GBDT的组合模型可有效辨识城市干道车辆出行群体,有助于掌握城市干道的交通运行规律,可为交通管理部门制定相关出行需求管理政策提供理论依据。

参考文献：

[1] 黄正国. 基于车牌识别数据的车辆出行特征研究[D]. 成都：西南交通大学，2019.
HUANG Z G. Research on vehicle travel feature based on license plate recognition data[D]. Southwest Jiaotong University，2019.

[2] 梁泉，翁剑成，林鹏飞，等. 基于个体出行图谱的公共交通通勤行为辨别方法研究[J]. 交通运输系统工程与信息，2018，18(2)：100-107.
LIANG Q，WENG J C，LIN P F，et al. Public transport commuter identification based on individual travel graph[J]. Journal of Transportation Systems Engineering and Information Technology，2018，18(2)：100-107.

[3] 崔洪军，赵锐，朱敏清，等. 基于朴素贝叶斯分类器的乘客出行属性分析[J]. 科学技术与工程，2020，20(11)：4572-4576.
CUI H J，ZHAO R，ZHU M Q，et al. Travel attributes analysis of passengers based on nave bayes classifier[J]. Science Technology and Engineering，2020，20(11)：4572-4576.

[4] MOHAMED E，ETIENNE C，JOHANNA B，et al. Understanding passenger patterns in public transit through smart card and socioeconomic data：A case study in rennes，france [C]//New York：The International Workshop on Urban Computing，2014.

[5] 刘凯. 地铁乘客出行规律分析及目的地预测方法研[D]. 北京：北京交通大学，2019.
LIU K. Analysis of metro passenger travel law and study of destination prediction method[D]. Beijing Jiaotong University，2019.

[6] 程小云，张学宇，薛顺然，等. 基于多维属性的轨道交通出行行为分类方法[J]. 交通运输工程与信息学报，2020，18(4)：166-174.
CHEN X Y，ZHANG X Y，XUE S R，et al. Method of analyzing rail transit travel behavior based on multidimensional attributes[J]. Journal of Transportation Engineering and Information，2020，18(4)：166-174.

[7] 翁小雄，吕攀龙. 基于 GBDT 算法的地铁 IC 卡通勤人群识别[J]. 重庆交通大学学报(自然科学版)，2019，38(5)：8-12.
WENG X X，LYU P L. Commuter crowd identification based on GBDT algorithm [J]. Journal of Chongqing Jiaotong University(Natural Science)，2019，38(5)：8-12.

[8] TAKAHIKO KUSAKABE，YASUO ASAKURA. Behavioral data mining of transit smart card data：A data fusion ap proach[J]. Transportation Research Part C，2014，46：179-191.

[9] 孙世超，杨东援. 基于朴素贝叶斯分类器的公交通勤人群辨识方法[J]. 交通运输系统工程与信息，2015，15(6)：46-53.
SUN S C，YANG D Y. Identification of transit commuters based on naive bayesian classifier[J]. Journal of Transportation Systems Engineering and Information Technology，2015，15(6)：46-53.

[10] CHEN H，YANG C，XU X. Clustering vehicle temporal and spatial travel behavior using license plate recognition data [J]. Journal of Advanced Transportation，2017(7)：1-14.

[11] YANG C，YAN F F，UKKUSURI S V. Unraveling traveler mobility patterns and predicting user behavior in the Shenzhen metro system[J]. Transport metrica A：Transport Science，2018，14(7)：576-597.

[12] 陈君，田朝军，赵清梅，等. 基于时空行为规律挖掘的公交乘客分类方法[J]. 交通运输工程学报，2021，21(5)：274-285.
CHEN J，TIAN C J，ZHAO Q M，et al. Bus passenger classification method based on spatial and temporal behavior regularity mining[J]. Journal of Traffic and Transportation Engineering，2021，21(5)：274-285.

[13] MA X，WU Y J，WANG Y，et al. Mining smart card data for transit riders travel patterns[J]. Transportation Research Part C：Emerging Technologies，2013，36：1-12.

[14] 马新露，雷小诗，马筱栎，等. 基于高速公路收费数据的车辆分类研究——以重庆市为例[J]. 交通运输研究，2021，7(1)：73-80.
MA X L，LEI X S，MA X L，et al. Vehicle classification based on expressway toll data：A case study of Chongqing [J]. Transport Research，2021，7(1)：73-80.

[15] 梁野，吕卫锋，杜博文. 基于峰值密度聚类的公交出行目的分类模型[J]. 哈尔滨工程大学学报，2018，39(3)：541-546.
LIANG Y，LYU W F，DU B W. Classification model of public transport trip purpose based on density peak clustering[J]. Journal of Harbin Engineering University，2018，39(3)：541-546.

[16] 杜蕊. 轨迹数据驱动的城市干线影响区车辆出行特征辨识及建模[D]. 重庆：重庆交通大学，2020.
DU R. Identification and modeling of vehicle travel characteristics in affected areas of urban arterial road driven by trajectory data[D]. Chongqing：Chongqing Jiaotong University，2020.

[17] VENUGOPAL S，DIVYA D. Transit passenger segmentation based on the travel patterns mined from smart card data using Optics algorithm[J]. International Journal of Advanced Information Science and Technology，2016，5(5)：

49–56.

[18] 王长硕,蒲英霞. 基于Labeled–LDA模型的居民群体分类与出行特征分析[J]. 计算机应用与软件,2022,39(11):17–24.
WANG C S,PU Y X. Analysis of classification and activity characteristics of urban residents based on Labeled–LDA model[J]. Computer Applications and Software,2022,39(11):17–24.

[19] JI Y J,CAO Y,LIU Y,et al. Research on classification and influencing factors of metro commuting patterns by combining smart card data and household travel survey data[J]. Iet Intelligent Transport Systems,2019,13(10):1525–1532.

[20] 赵超. 混合属性聚类算法的研究及应用[D]. 秦皇岛:燕山大学,2017.
ZHAO C. Research on clustering algorithm for mixed attributes and application[D]. Qinghuangdao:Yanshan University,2017.

[21] HUANG Z. Extensions to the K–means algorithm for clustering large data sets with categorical values[J]. Data Mining and Knowledge Discovery,1998,2(3):283–304.

[22] 石鸿雁, 徐明明. 基于平均差异度的改进K–prototypes聚类算法[J]. 沈阳工业大学学报,2019,41(5):555–559.
SHI H Y,XU M M. Improved K–prototypes clustering algorithm based on average difference degree[J]. Journal of Shenyang University of Technology,2019,41(5):555–559.

[23] RODRIGUEZ A,LAIO A. Clustering by fast search and find of density peaks[J]. Science,2014,344(6191):1492–1496.

[24] 王森,邢帅杰,刘琛. 密度峰值聚类算法研究综述[J]. 华东交通大学学报,2023,40(1):106–116.
WANG S,XING S J,LIU C. Survey of density peak clustering algorithm[J]. Journal of East China Jiaotong University,2023,40(1):106–116.

[25] 欧阳浩,戴喜生,王智文,等. 基于信息熵的粗糙K–prototypes聚类算法[J]. 计算机工程与设计,2015,36(5):1239–1243.
OU Y H,DAI X S,WANG Z W,et al. Rough K–prototypes clustering algorithm based on information entropy[J]. Computer Engineering and Design,2015,36(5):1239–1243.

[26] 李航. 统计学习方法[M]. 北京:清华大学出版社,2012.
LI H. Statistical learning methods[M]. Beijing:Tsinghua University Press,2012.

[27] 战友,邓强胜,罗志伟,等. 基于GBDT的沥青路面抗滑性能感知模型研究[J]. 土木工程学报,2023,56(2):121–232.
ZHAN Y,DENG Q S,LUO Z W,et al. Research on GBDT–based asphalt pavement skid resistance perception model[J]. China Civil Engineering Journal,2023,56(2):121–232.

第一作者:梁灯(1996—),男,硕士研究生,研究方向为交通信息工程及控制。E–mail:3192513626@qq.com。

通信作者:蔡晓禹(1979—),男,教授,博士生导师,研究方向为深度学习交通视频检测与状态识别。E–mail:caixiaoyu@cqjtu.edu.cn。

(责任编辑:吴海燕)