

基于网约车数据的城市区域出行时空 特征识别与预测研究

张政¹, 陈艳艳^{*1}, 梁天闻²

(1. 北京工业大学城市交通学院, 北京 100124; 2. 交通运输部公路科学研究院, 北京 100088)

摘要: 针对交通需求特征识别和需求预测问题, 构建改进的 LDA(Latent Dirichlet Allocation)城市区域内出行需求识别与预测组合模型, 快速识别城市区域内出行需求特征并对需求做出预测. 构建城市交通小区尺度内的空间和时间维度下的主要出行需求特征分布挖掘辨识方法, 以及数据集在不同时间尺度下时间维度出行特征构建及预测方法. 利用北京市三环内网约车出行订单数据, 验证模型的有效性和准确性. 结果表明, 模型能够对不同时间窗口下的区域出行需求特征进行辨识和预测, 取得较好的结果.

关键词: 城市交通; 交通需求识别和预测; LDA 模型; 网约车

Regional Travel Demand Mining and Forecasting Using Car-hailing Order Records

ZHANG Zheng¹, CHEN Yan-yan¹, LIANG Tian-wen²

(1. College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China;

2. Research Institute of Highway Ministry of Transport, Beijing 100088, China)

Abstract: This paper proposed a combined model that can fast mining traffic demand and prediction based on the Latent Dirichlet Allocation (LDA) analysis model. This combined analysis framework is able to deal with demand identification and prediction at the same time. The study first developed the traffic demand identification model at the traffic analysis zone (TAZ) scale for presenting the demand characteristics at both the spatial and temporal dimension. Then it proposed a prediction method under a multi-scale time window. The effectiveness and accuracy of the model was verified using car-hailing order data within Beijing's third-ring road. The results show that the model can identify and predict the regional travel demand under different time windows, and achieve good results.

Keywords: urban traffic; transportation demand identification and forecasting; Latent Dirichlet Allocation (LDA); car-hailing services

0 引言

网约车的逐步发展对城市居民的生活和出行方式产生较大影响. 城市不同区域网约车出行需求特征精准识别和预测, 对运营商优化车辆调度方案, 资源合理配置, 缓解高峰时段关键区域供需失衡, 减少环境污染等问题具有重要意义; 对交通规划者制定面向区域的需求管理政策, 开展可替

代的定制公共交通服务具有重要参考价值.

城市交通需求特征识别广泛采用聚类、矩阵分解等特征提取方法挖掘城市不同群体的出行特征规律. MA X. 等^[1]利用 DBSCAN(Density-Based Spatial Clustering of Applications with Noise)方法对北京市公交通勤者进行辨识; 曲昭伟等^[2]根据出租车 GPS 数据采用核密度方法对城市热点路段进

收稿日期: 2019-11-14

修回日期: 2020-02-17

录用日期: 2020-02-28

基金项目: 国家重点研发计划/National Key Research and Development Program of China(2016YFE0206800).

作者简介: 张政(1992-), 男, 山东人, 博士生.

*通信作者: cdyan@bjut.edu.cn

行辨识; CALABRESE F.^[3]采用矩阵分解方法对不同交通需求进行分类. 交通需求特征识别是预测的基础, 基于不同出行需求特征可分别构建预测模型. 交通预测方法主要包括参数方法和非参数方法, 参数方法如基于时间序列的统计模型自回归移动平均^[4]和泊松模型^[5]等, 而人工智能的发展使得机器学习方法逐渐成为解决短时交通需求预测的非参数方法, 如支持向量机^[6], 长短时记忆神经网络^[7]. JIANG S.^[8]提出基于最小二乘支持向量机的网约车需求预测方法. 谷远利^[9]提出基于深度学习网络的预测某区域内网约车短时供需差的框架, 该预测模型考虑时空因素和环境因素, 预测短时间内区域的出行需求. 郭宪^[10]提出基于多源数据的梯度决策回归树网约车需求预测方法.

综上, 交通需求特征识别和需求预测缺乏联系, 辨识方法和预测方法易受时空数据集颗粒度影响. 而基于特定需求特征的预测模型能够对结果进行较好的解析, 主题模型 LDA (Latent Dirichlet Allocation) 能够有效避免数据集稀疏和维数较高造成辨识结果差等问题, 且识别结果能够呈现出概率较高的出行特征. 本文以北京市三环路以内网约车出行订单需求和城市交通小区划分结果为基础, 构建基于主题模型 LDA 的城市出行需求辨识和预测组合分析框架; 分析城市不同区域内需求特征, 基于需求特征辨识结果构建多时间尺度下的组合预测模型; 对比常用预测方法, 验证模型的有效性和精度.

1 基于LDA的区域出行需求特征辨识与预测模型

1.1 问题描述

主题模型是以非监督学习的方式对数据集中隐含特征进行辨识的统计模型, 被广泛用于自然语言处理中的语义分析和文本挖掘. LDA 是一种常见的主题模型, 图1为标准 LDA 的概率图模型, 从概率分布角度表示时空数据集的生成过程. 数据集中每个元素生成过程为: 某时空数据集存在 D 个待挖掘隐含特征的研究对象, 描述研究对象 d ($d=1, \dots, D$) 共有 T 条数据记录, 其中某一条数据记录的数值记为 w , 该数值的主题编号为 z . 假设

所有研究对象共存在 K 种隐含出行特征, 每种特征可用主题 z 来概括, $z=1, \dots, K$. 那么该数据集的主题服从超参数 α 的狄利克雷 (Dirichlet) 分布. 研究对象 d 的主题服从参数为 θ_d 的多项 (multinomial) 分布, 第 t 个数值取值 w 服从参数为 ϕ_k 的 multinomial 分布. 当某一主题确定后, 该主题下数值服从超参数为 β 的 Dirichlet 分布.

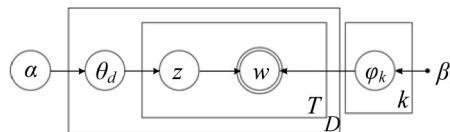


图1 LDA模型结构图

Fig. 1 Structure of LDA

利用网约车时空数据集挖掘得到城市不同区域内出行需求特征, 网约车时空数据集为海量用户的出行订单数据, 每条订单数据记录出行者的起点 O 、终点 D 、出发时间 S_1 、到达时间 S_2 等信息. 将订单数据集按照一定的时间颗粒度集聚到交通小区层面, 得到 $O \times D \times S$ 三维张量, 记为 χ , $\chi \in \mathbb{N}^{(u \times v \times n)}$; x_{ijt} 为在时间段 t 内, 从交通小区 i 到交通小区 j 的出行量. 此时, 交通小区为研究对象, 交通小区内某时段产生的到发客流为具体数值. 采用 LDA 模型发现交通小区层面隐含的经常性的出行特征, 分别得到 O, D, S 三个维度的主题分布 O', D', S' , 每个主题代表一种隐含的出行特征, 即各出行特征以概率分布形式呈现, 表示某种出行特征出现概率的大小. 符号含义如表1所示. 根据挖掘到的各维度隐含分布特征, 构建运算高效的出行需求预测模型.

表1 符号定义

Table 1 Symbol definition

符号	定义
u	出发交通小区数量
v	到达交通小区数量
n	时间窗口个数
O'	出发小区维度的主题分布
D'	到达小区维度的主题分布
S'	时间维度的主题分布

1.2 基于LDA的区域出行需求辨识方法

为发掘网约车时空数据集各维度隐含的主题分布特征, 构建图2所示改进 LDA 模型. 数据集 χ 中每个数值的生成过程如下.

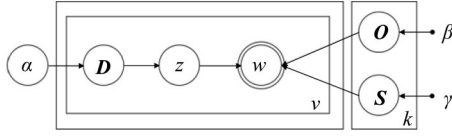


图2 改进LDA概率图模型

Fig. 2 Structure of modified LDA model

(1) 对任一主题 $z(z=1, \dots, K)$ 有:

(a) 对任一出发主题分布 $O'_z, O'_z \sim \text{Dirichlet}(\beta)$;

(b) 对任一时间主题分布 $S'_z, S'_z \sim \text{Dirichlet}(\gamma)$.

(2) 对任一到达交通小区 $j(j=1, \dots, v)$ 有:

(a) $D'_j \sim \text{Dirichlet}(\alpha)$;

(b) 对于任一到达小区 j , 共有 N_j 条出行记录, 出行记录 $i(i=1, \dots, N_j)$ 生成过程为

① 该记录主题编号 z_{ij} 服从 $z_{ij} \sim \text{multinomial}(D'_j)$;

② 该记录来自出发小区 $o_{ij} \sim \text{multinomial}(O'_{z_{ij}})$;

③ 该记录的时间戳 $t_{ij} \sim \text{multinomial}(S'_{z_{ij}})$.

图2中, α, β, γ 为 Dirichlet 超参数, 采用吉布斯采样方法对改进 LDA 模型进行参数估计^[11]. 吉布斯采样方法包括初始化和更新迭代两部分, 超参数需在初始化过程中按照经验设定某初始值, 更新迭代过程的全概率公式为

$$p(z_{ijt} = k | O', D', S', \alpha, \beta, \gamma) \propto \frac{O'_{ik} + \alpha}{\sum_k O'_{ik} + \alpha k} \times \frac{D'_{kj} + \beta}{\sum_j D'_{kj} + \beta v} \cdot \frac{S'_{kt} + \gamma}{\sum_t S'_{kt} + \gamma n} \quad (1)$$

式中: z_{ijt} 为在 t 时间下由小区 i 出发, 到达小区 j 的主题编号; O'_{ik}, D'_{kj} 和 S'_{kt} 分别表示第 i 出发小区, 第 j 到达小区和第 t 时间段属于主题 k 的个数. 全概率公式为每次更新采样过程中生成某个主题的概率. 主题个数 K 取值利用混合度指标^[11]来衡量, 当混合度较小时, 模型效果较优.

采样经过一定次数后, 采样过程趋于稳定, 主题概率分布不再变化, 此时各维度的主题概率分布为

$$O'_{ik} \propto \frac{O'_{ik} + \alpha}{\sum_k O'_{ik} + \alpha k} \quad (2)$$

$$D'_{kj} \propto \frac{D'_{kj} + \beta}{\sum_j D'_{kj} + \beta v} \quad (3)$$

$$S'_{kt} \propto \frac{S'_{kt} + \gamma}{\sum_t S'_{kt} + \gamma n} \quad (4)$$

如果时空数据集时间尺度不确定, 需要得到不同时间尺度下各维度隐含的出行特征, 直接做法是分别按照对应的时间尺度重复上述参数推断过程, 计算成本较高. 吉布斯采样过程对所有尺度分配某一主题的概率是等价的, 故构建基于某最小时间尺度下的主题分布为基准, 构建任意时间尺度下的时间维度主题分布. 记某时间窗口大小为 l_h , 则任意时间窗口表示为

$$l_h = l_0 \cdot L^{(h)} \quad (5)$$

式中: l_0 为基准时间窗口大小; $L^{(h)}$ 为时间窗口增长因子, h 为时间窗口大小编号, $h=0, 1, 2, \dots, H$, 通常情况下 $L=2$ ^[12]. 因为对主题分布矩阵中每个元素分配主题的概率是相同的, 即每个时间颗粒度下估计潜在主题的过程相同, 故对于时间颗粒度为 l_h 的时间主题矩阵分布有

$$S_{kt}^{(h)} \propto \sum_{m=1}^{S^{(0)}} S_{k,t-l_h+m}^{(0)} \quad (6)$$

式中: $S_{kt}^{(h)}$ 为在时间窗口为 h 下的时间主题概率分布, $m=l_h/l_0$. 根据最小时间窗口矩阵分析得到主题矩阵 $S^{(0)}$, 可以得到不同时间窗口下时间维度的主题特征分布矩阵, 基于此可呈现在多时间窗口下的主题概率的分布特征情况.

1.3 多时间尺度下的区域出行需求预测方法

若通过主题分析得到时间维度的主题概率分布是明显的时间序列, 利用自回归(AR)方法进行预测. 对于某一时间窗口下, 利用前 ω 时段下的记录完成 t 时刻的预测, 即

$$S_{kt} = \lambda_1 S_{kt-1} + \dots + \lambda_\omega S_{kt-\omega} + \varepsilon_t \quad (7)$$

式中: λ_n 为回归系数, $n=1, \dots, \omega$; ε_t 为误差项.

多个层面的时间窗口, 基于最小时间尺度下分布, 实现不同时间尺度下时间维度主题的估计. 则有

$$S_{kt}^{(h)} = \sum_{h=0}^H \sum_{n=1}^{\omega} \lambda_{nk}^{(h)} S_{k,t-n}^{(h)} + \varepsilon_t \quad (8)$$

式中: $\lambda_{nk}^{(h)}$ 为第 h 个时间窗口下, 第 n 个回归系数; $S_{k,t-n}^{(h)}$ 为第 h 个时间窗口下, $t-n$ 时段的主题分布. 完成时间维度主题矩阵趋势变化的预测后, 还需要推算区域的交通出行需求.

$$\hat{x}_{ijt} = a \bar{x}_i \sum_{k=1}^K O'_{ik} \cdot D'_{kj} \cdot S_{kt}^{(h)} \quad (9)$$

式中: a 为采用数据集的时间长度; $\bar{x}_i$ 是平均每个时间窗口下从 i 小区出发的平均出行需求; $\hat{x}_{ijt}$ 表示区域出行需求。

式(9)是利用出发主题分布矩阵, 到达主题分布矩阵和时间主题分布矩阵, 基于对时间主题矩阵的预测结果, 预测未来区域出行需求。

2 案例分析

选用北京市三环路以内2017年7月17~30日滴滴出行网约车订单数据验证模型的有效性。网约车时空订单数据集每条数据记录每一订单编号, 上车地点经纬度, 下车地点经纬度, 上车时间, 下车时间等信息。为研究不同时间尺度下区域出行特征分布及预测结果, 选取15, 30, 60 min这3个时间窗口, 然后根据订单经纬度坐标集聚到三环内交通小区层面, 得到 χ_1 , χ_2 和 χ_3 这3个数据集。

2.1 区域出行特征识别结果分析

对数据集 χ_1 按照图2改进LDA模型框架进行

主题分析, 当主题数 $K=3$ 时, 改进LDA模型取得较优效果。图3(a)为空间层面(出发小区, 到达小区)主题分布结果, 3种主题分别为低, 中, 高不同出行强度在空间上的分布。为验证模型结果的合理性, 从数据周期内随机选取1 d数据, 进行统计分析, 结果如图3(b)所示。模型结果与统计结果整体趋势相差不大, A, B, C, D, E这5处地点, 两种方法结果相异或相似, 利用POI(Point of interest)数据计算这5处地点的土地利用性质^[13]。A主要为高等教育用地, 此类用地约占该交通小区的67.8%; B和E主要为交通枢纽用地, 此类用地约占交通小区面积的86.8%; C主要为公园绿地用地类型, 约占该交通小区的76.1%。这4处地点出行主要是随机出行, 故某日出行数据统计结果与数据集总体模型结果存在一定差别。两者结果相似的D类交通小区, 主要为商业写字楼, 此类交通小区产生的出行较为规律, 故两种方法结果差别较小。

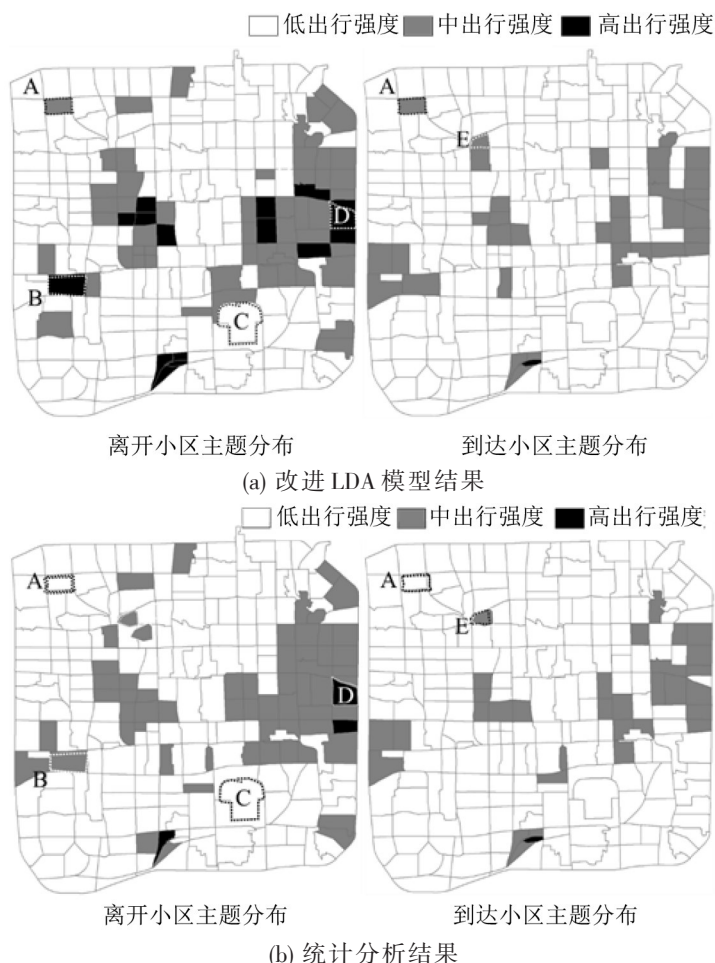


图3 出发小区和到达小区主题分布

Fig. 3 Topics distribution of origin and destination traffic analysis zones.

图4为 χ_1 在时间维度主题分析结果,根据式

(6)得到 χ_2 和 χ_3 的主题分布,如图5和图6所示。

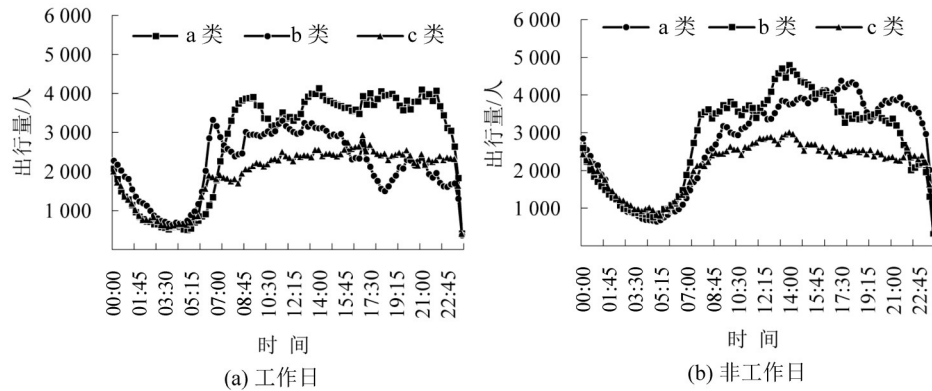


图4 时间窗口为15 min的时间维度主题分布

Fig. 4 Topic distribution on temporal dimension under 15 min

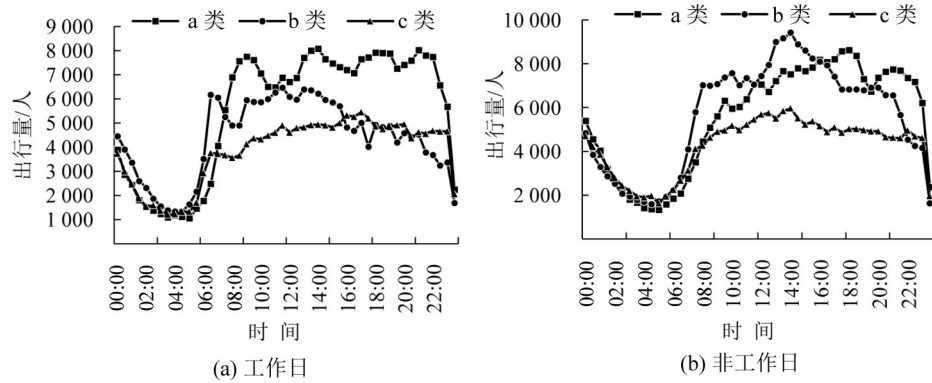


图5 时间窗口为30 min的时间维度主题分布

Fig. 5 Topic distribution on temporal dimension under 30 min

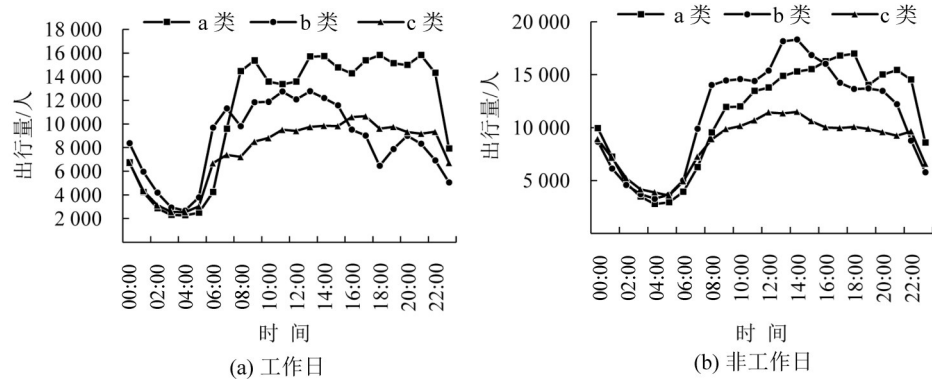


图6 时间窗口为60 min的时间维度主题分布

Fig. 6 Topic distribution on temporal dimension under 60 min

由图4~图6可看出,工作日与非工作日出行强度变化趋势不同.随着时间窗口变大,时间主题分布逐渐趋于平缓,各类出行强度大小关系逐渐变得明显.a类出行特征在工作日内始终能够维持较高的出行强度,在非工作日内存在明显的单峰出行强度特征;b类在工作日出行强度整体低于a类,在非工作日b类同样存在单峰,但与a类单峰出现时段不一致;c类表示较低的出行强度分布,

在1 d内变化较为平稳,且工作日和非工作日出行强度特征差别不大.

2.2 区域出行预测结果分析

为评价改进LDA构建的预测模型效果,选取均方根误差(Root Mean Squared Error, RSME)和平均绝对百分比误差(Mean Absolute Percentage Error, MAPE)作为衡量误差的指标,分别评价在不同时间尺度下模型的预测效果,如表2所示.

表 2 预测精度评价

Table 2 Forecasting accuracy evaluation

模 型	15 min		30 min		60 min	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
AR	37.22	0.385	35.61	0.355	32.87	0.323
LDA-AR	36.20	0.262	32.87	0.258	30.32	0.250

由表2可知,时间尺度的大小会影响预测精度,时间窗口越大,需求越平滑,预测精度越高.改进的LDA-AR预测模型的精度比传统的AR模型高.

3 结 论

本文将LDA模型引入到识别交通时空数据集隐含出行特征问题中来,有别于LDA得到单一维度主题分布,增加多维度信息作用于数据集观测值,可得到到达小区,出发小区和时间维度下的主题分布特征;同时基于最小时间尺度下的主题分布构建任意时间尺度下的需求分布预测模型,提高计算效率;并选取北京市三环路以内网约车时空数据集,验证模型的有效性和精度.结果表明,构建的基于最小时间窗口下的多尺度分布特征识别和预测方法能够识别不同时间颗粒度下的区域需求特征,实现区域间的需求变化趋势预测.本文研究对城市交通规划,公共交通组织运营等方面具有重要意义.

城市交通需求辨识与组合预测模型只对总体需求进行了预测,拟在后续研究中考考虑不同主题下预测结果的对比分析.

参考文献:

- [1] MA X, LIU C, WEN H, et al. Understanding commuting patterns using transit smart card data[J]. Journal of Transport Geography, 2017(58): 135-145.
- [2] 曲昭伟,王鑫,宋现敏,等.基于出租车GPS大数据的城市热点出行路段识别方法[J].交通运输系统工程与信息, 2019, 19(2): 238-246. [QU Z W, WANG X, SONG X M, et al. Urban hotspot travel section identification method based on taxi GPS large data[J]. Journal of Transportation Systems Engineering and Information Technology, 2019, 19(2): 238-246.]
- [3] CALABRESE F, READES J, RATTI C. Eigenplaces: Segmenting space through digital signatures[J]. Ieee Pervas Comput, 2010, 9(1): 78-84.
- [4] JAMIL M S, AKBAR S. Taxi passenger hotspot prediction using automatic ARIMA model[C]. 2017 3rd International Conference on Science in Information Technology. Bandung, Indonesia, IEEE, 2017.
- [5] MOREIRA-MATIAS L, GAMA J, FERREIRA M, et al. Predicting taxi-passenger demand using streaming data [J]. Ieee T Intell Transp, 2013, 14(3): 1393-1402.
- [6] LI Y, LU J, ZHANG L, et al. Taxi booking mobile app order demand prediction based on short-term traffic forecasting[J]. Transportation Research Record, 2017, 2634(1): 57-68.
- [7] MA X, TAO Z, WANG Y, et al. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data[J]. Transportation Research Part C: Emerging Technologies, 2015(54): 187-197.
- [8] JIANG S, CHEN W, LI Z, et al. Short-term demand prediction method for online car-hailing services based on a least squares support vector machine[J]. IEEE Access, 2019, DOI: 10.1109/ACCESS.2019.2891825.
- [9] 谷远利,李萌,芮小平,等.基于深度学习的网约车供需缺口短时预测研究[J].交通运输系统工程与信息, 2019, 19(2): 223-230. [GU Y L, LI M, RUI X P, et al. Short-term forecasting of supply-demand gap under online car-hailing services based on deep learning[J]. Journal of Transportation Systems Engineering and Information Technology, 2019, 19(2): 223-230.]
- [10] 郭宪,沈吟东.基于梯度提升回归树的网约车需求预测 [C]. 2018世界交通运输大会,北京,2018. [GUO X, SHEN Y D. Prediction for e-hail taxi demands based on gradient boosting regression trees[C]. 2018 World Transport Convention, Beijing, China, 2018.]
- [11] HEINRICH G. Parameter Estimation for Text Analysis [R]. Darmstadt, Germany, 2005.
- [12] Matsubara Y, Sakurai Y, Faloutsos C, et al. Fast mining and forecasting of complex time-stamped events[C]. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 2012.
- [13] YUAN N J, ZHENG Y, XIE X, et al. Discovering urban functional zones using latent activity trajectories[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(3): 712-725.