文章编号:1009-6744(2020)05-0029-07 中图分类号:U491 文献标志码:A
DOI:10.16097/j.cnki.1009-6744.2020.05.005

基于出行特征的用地类型推断方法研究

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: 提出一种基于卷积神经网络推测城市交通小区内用地特征的算法,同时对交通小区内多种用地类型进行预测.选用公共交通出行数据集和网约车出行数据集,融合多种出行方式的出行特征对交通小区内用地特征刻画.提取交通小区内发生强度,吸引强度和产吸差强度3个指标作为模型输入,训练得到基于区域内出行特征双通道的卷积神经网络模型,采用网格寻优方法确定最优网络结构.选取北京市六环内交通小区作为研究对象,结果表明,本文算法能够同时推断交通小区内居住、工作和休闲用地特征,并获得各用地类型在小区内占比分布.

关键词: 城市交通;用地类型推测;卷积神经网络;交通小区;出行特征

Inferring Land Use Characteristics Using Travel Patterns

ZHANG Zheng[1], CHEN Yan-yan[1], LIANG Tian-wen[2]
(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 proposes a land use inferring method based on the convolutional neural network (CNN), which can infer multiple lane use types at the traffic analysis zones (TAZs) simultaneously. The study combines public transport mobility dataset and online car-hailing mobility dataset for inferring land use type. Generation intensity, attraction intensity, and difference between generation and attraction intensity are extracted from the travel dataset, which are then used to train the CNN. The optimal network structure is determined by grid search. The TAZs within the 6th Ring Road of Beijing are taken as examples for the analysis. The results indicate that the proposed method is able to estimate the proportion distribution of several land use types at the same time within the TAZs, such as resident, workplace and leisure land uses.

Keywords: urban traffic; inferring land use type; convolutional neural network (CNN); traffic analysis zone (TAZ); travel patterns

0 引 言

城市用地类型决定交通出行形态,影响人们的生活方式,并随着城市发展变化而变化[1].城市用地规划图不仅难以获取城市交通小区内各用地类型占比,也难以支撑实时交通需求分析.公交IC卡、网约车等时空出行数据,使根据出行特征采用机器学习方法研究城市区域用地成为可能.根据时空数据集的不同,城市用地特征分布推测方法可分为有监督和无监督两种方法.无监督方法利用无标签数据,采用数据挖掘方法得到区域出行特征,赋予该区域用地特征含义.LIU Y.[2]根据区域内巡游车供需差不同水平反映不同用地关系,构建上下车平衡向量判断城市内的用地类型.PAN G.[3]采用DBSCAN(Density-Based Spatial Clustering of Applications with Noise)聚类方法对巡游车轨迹信息进行聚类,根据聚类的区域出行

收稿日期:2020-05-10 修回日期:2020-06-25 录用日期:2020-06-28
基金项目:国家重点研发计划/National Key Research and Development Program of China(2017YFC0803903).
作者简介:张政(1992-),男,山东人,博士生. *通信作者:cdyan@bjut.edu.cn

特征赋予不同区域居住、办公等土地利用特征.有监督方法主要根据标签数据利用机器学习方法对均分的网格化城市区域用地类型推测.TOOLE J.[4]根据手机话单数据统计得到城市网格内人群特征,利用随机森林算法,推测城市网格化空间内的土地利用特征.ZHAN X.[5]利用社交网络签到数据比较有监督和无监督两种方法,基于有标签数据的随机森林方法对用地类型的推测能够取得较好效果.人工神经网络(Artificial Neural Network, ANN)因处理高维复杂数据方面的优势,逐渐被用来推测城市用地特征. ZHAO J.[6]利用ANN方法对公共自行车站点周边各种用地特征分布进行预测.人工神经网络方法对用地类型预测精度有较大提升,卷积神经网络(Convolutional Neural Network, CNN)是神经网络的扩展,处理交通时空数据时,能够捕获空间依赖性对推测结果的影响.MA X.[7]将路网速度看作图像,构建CNN网络提取路网速度时空特征,对路网交通速度进行预测.城市相邻区域内土地利用特征具有相似性,现有机器学习方法较难考虑相邻区域用地类型特征对结果的影响,缺乏对交通小区内多种混合用地类型分布同时推测.

本文采用常规公交和轨道交通组成的公交出行数据集和网约车出行数据集,构建基于CNN的深度学习模型,对城市交通小区内居住、工作和休闲用地类型占比进行推断.按照某一时间间隔将公交出行数据集和网约车出行数据集聚集到交通小区,选取交通小区产生强度、吸引强度和产吸强度差组成时空特征矩阵,并用兴趣点数据对小区内用地类型进行标注,将时空特征矩阵和标注数据作为CNN模型输入;确定CNN结构后,利用构建模型对北京市六环以内交通小区内各用地类型进行推断.

1 数据集介绍及处理

1.1 出行特征提取

交通小区单位时间内乘客产生量和吸引量能够反映该小区内用地特征[3],利用公交IC卡数据和网约车订单数据,提取交通小区层面的出行特征,用于CNN模型输入.公交IC卡数据包括地面公交和轨道交通用户刷卡数据,记录出行者匿名ID,出行起终点经纬度,线路及站点编号和时间戳.将个体在公交系统中完整出行链路信息融合,得到个体公交出行链[8],但公交站点在城市区域内位置固定,不能反映所有交通小区出行特征.网约车上下车位置不受限,故利用网约车订单数据为辅助数据作为模型输入.网约车订单数据记录出行者匿名ID,起终点经纬度和时间戳等.将公交出行链和网约车订单起终点坐标数据与城市交通小区进行空间匹配,得到表1所示样例数据集.

表 1 出行特征样例数据

Table 1 Sample of bus-chain and car-hailing dataset

用户ID	出发时间	到达时间	出发小区编号	到达小区编号	数据来源
Traveler1	2017-07-16 08:20:00	2017-07-16 08:59:00	61303	21005	公交出行链
Traveler2	2017-07-18 17:36:00	2017-07-18 18:20:00	51820	62503	网约车订单

按照1 h间隔,将出行OD数据集聚,选取交通小区层面需求发生强度,吸引强度,以及发生和吸引强度差作为CNN模型输入特征.

特征计算方法如下:对于任一小区 z_i , $i=\{1,2,\cdots,m\}$,在第 j 个时间间隔内的交通吸引量为 $z_{i,j}^{\mathrm{a}}$,交通发生量为 $z_{i,j}^{\mathrm{d}}$,则交通小区吸引强度,产生强度,发生和吸引强度差为

$$A_{i,j}^{T}=z_{i,j}^{\mathrm{a}}\Big/s_i \tag{1}$$

$$D_{i,j}^{T}=z_{i,j}^{\mathrm{d}}\Big/s_i \tag{2}$$

$$V_{i,j}^{T}=\left(z_{i,j}^{\mathrm{a}}-z_{i,j}^{\mathrm{d}}\right)\Big/s_i \tag{3}$$

式中:m 为研究范围内交通小区数量;s_i 为交通小区 z_i 的面积;T 为研究时段 n 天内不同日期编号,$T=\{1,2,\cdots,n\}$;$A_{i,j}^{T}$, $D_{i,j}^{T}$ 分别为在 T 日交通小区 i 在第 j 个时间间隔内的吸引强度、发生强度;$V_{i,j}^{T}$ 为在 T 日交通小区 i 在第 j 个时间间隔内发生强度和吸引强度之差.

公交出行数据集和网约出行数据集出行总量不同,分别按照式(1)~式(3)计算得到各自出行特征

时空矩阵，并对同一时刻不同小区间的强度数据按照标准差归一化处理.然后按照对应小区和时间段相加，得到输入时空特征矩阵.以小区吸引强度时空矩阵为例，计算输入时空特征矩阵，过程如图1所示.

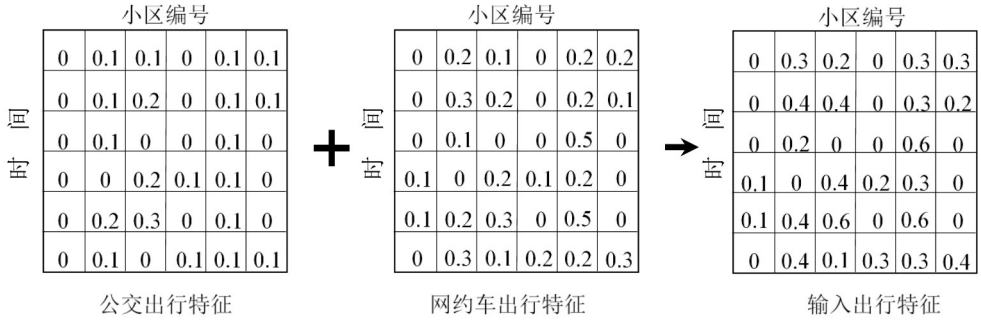


图1　输入出行特征计算过程(以小区吸引强度为例)

Fig. 1　Schematic of traffic analysis zone-based travel characteristic process (take $A_{i,j}^{T}$ as an example)

1.2 城市用地类型标注

交通小区内用地类型信息在缺乏用地规划图或获取各种用地类型占比困难时，可用POI数据根据TF-IDF(Term Frequency-Inverse Document Frequency)方法估计区域内用地类型[9].POI数据可通过百度地图API (Application Programming Interface)获得.获取得到POI数据包括居住、工作、休闲娱乐、公共服务、公交站点等18个类别，具体来说，每条POI数据包含POI所属类别，经纬度坐标和具体详细名称.相关研究表明，居住、工作和休闲娱乐用地与交通小区内居民出行特征有密切联系[4]，因此，本文提取居住(R)、工作(W)和休闲娱乐(L)这3种POI类别，利用TF-IDF方法对3种类别POI在各交通小区内的分布进行分析.TF-IDF计算过程为

$$\mathrm{tf}\left(p_c,z_i\right)=0.5+0.5\frac{f_{p_c,z_i}}{\max\left\{f_{p_{\mathrm{R}},z_i},f_{p_{\mathrm{W}},z_i},f_{p_{\mathrm{L}},z_i}\right\}} \tag{4}$$

$$\mathrm{idf}\left(p_c,Z\right)=\lg\frac{m}{\left|\left\{z_i\in\mathrm{Z}:p_c\in z_i\right\}\right|} \tag{5}$$

$$\mathrm{tf-idf}\left(p_c,z_i,Z\right)=\mathrm{tf}\left(p_c,z_i\right)\cdot\mathrm{idf}\left(p_c,Z\right) \tag{6}$$

式中：p_c为c类POI，$c=\{\mathrm{R},\mathrm{W},\mathrm{L}\}$；$f_{p_c,z_i}$为$p_c$在小区$z_i$中出现的频次；$Z$为所有交通小区$z_i$集合，有$Z=\{z_1,\cdots,z_i,\cdots,z_m\}$.

TF-IDF是衡量某词语在一篇文档中重要程度的方法，当某一词语在一篇文档中出现频率较高且在文档集合中出现次数较少时，该词语在此文档中较重要.近年来，随着交通大数据的涌现，该方法被广泛应用到交通小区功能属性特征分析等领域[10].TF-IDF计算包括TF和IDF两部分，TF表示某一词语在文档中出现的相对频次，IDF表示这篇文档中词语的分布在所有文集中的重要程度.将交通小区z_i类比为文档，研究范围内所有交通小区构成文集Z，交通小区内p_c点类比为不同类别单词.式(4)为TF计算方法，式(5)为衡量该p_c在所有小区中的重要程度，分母表示出现该p_c类别的交通小区的个数.故每个交通小区内用地类型特征可以通过计算该交通小区内各类POI的TF-IDF值获得.

2　基于卷积神经网络的区域用地类型推断

基于CNN深度学习模型，对区域内居住、工作和休闲3种用地类型同时进行推断，模型整体架构如图2所示.包括输入层、卷积层、池化层、平坦层、全连接层和输出层，各层设置具体如下.

(1) 输入层.

将处理得到的输入特征矩阵和基于POI的交通小区标注数据作为模型输入数据.

(2) 卷积层.

卷积核表示为w_j，j为卷积核编号，$j=\{1,2,\cdots,J\}$，共有J个卷积核.对于输入矩阵$\boldsymbol{x}$，卷积操作后得到的特征矩阵为

$$\boldsymbol{y}^{\mathrm{Conv}}=f\left(\boldsymbol{w}\otimes\boldsymbol{x}+\boldsymbol{b}\right) \tag{7}$$

式中：f为非线性激活函数；$\otimes$为矩阵点乘；$\boldsymbol{w}$为权重矩阵；$\boldsymbol{b}$为偏差量；$\boldsymbol{y}^{\mathrm{Conv}}$为卷积操作后的输出

矩阵.采用修正线性单元作为激活函数,卷积核尺寸大小会影响模型精度.

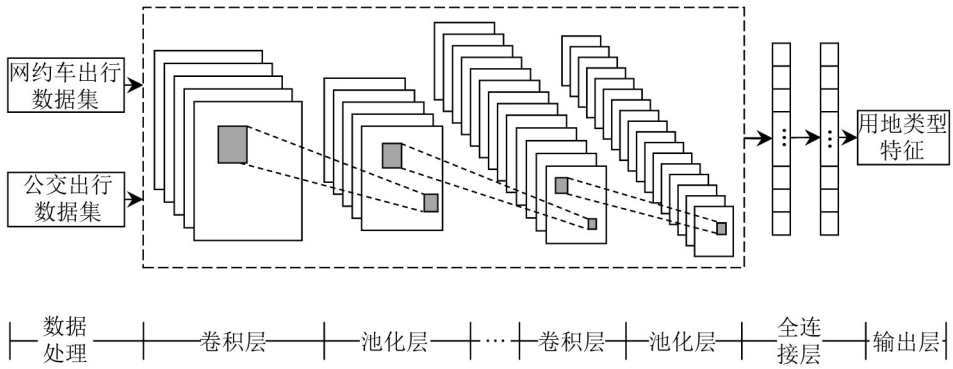


图2 用地类型推断模型架构

Fig. 2 CNN architecture for land use characteristics inferences

(3) 池化层.

对输入卷积层的结果进行池化,即提取卷积层中关键信息.池化层作用于卷积层感受区域,采用最大值方法对感受区域进行池化操作,即

$$\boldsymbol{y}^{\text{Pool}} = \max\left(\boldsymbol{x}_{(p,q)}\right) \tag{8}$$

式中:$\boldsymbol{x}_{(p,q)}$ 为作用于上层输入矩阵的感受区域;$\boldsymbol{y}^{\text{Pool}}$ 为池化操作后的输出矩阵.

(4) 全连接层.

经过卷积和池化操作,网格化数据被平坦化处理为一维向量数据.为防止模型过拟合,减少稀疏数据集对模型的影响,全连接层设置丢弃率以增强模型泛化能力,即

$$\hat{\boldsymbol{y}} = \boldsymbol{w} \cdot \hat{\boldsymbol{y}}^{\text{f}} + \boldsymbol{b} \tag{9}$$

式中:$\boldsymbol{w}$ 为权重矩阵;$\hat{\boldsymbol{y}}^{\text{f}}$ 为平坦化处理的数据矩阵;$\hat{\boldsymbol{y}}$ 为最终模型输出结果.

(5) 输出层.

输出交通小区内居住、工作和休闲用地类型分布.

为评价搭建CNN网络模型推断用地类型的精度,损失函数选用均方误差(Mean Square Error, MSE),即

$$L_{\text{MSE}} = \frac{1}{N}\sum_{k=1}^{N}\left(\hat{y}_k - y_k\right)^2 \tag{10}$$

式中:N 为样本总量;$\hat{y}_k, y_k$ 分别表示第 k 个样本的估计值和真实值.

CNN优化算法主要包括随机梯度下降(Stochastic Gradient Descent), RMSProp(Root Mean Square Prop)和Adam(Adaptive Moment Estimation),其中,Adam算法使用最广,故采用Adam优化算法对损失函数进行优化.

3 结果分析

选取2017年7月15~24日北京市六环以内的公交出行数据和网约车出行数据,基于出行特征对交通小区内用地类型推断.并获取同时期的POI数据集,提取居住、工作和休闲娱乐3种类别的POI数据计算小区内各用地类型占比.按照1 h间隔对出行数据进行聚合,得到公交出行数据集和网约车出行数据集,按式(1)~式(3)进行特征提取.将原始数据集按75%/25%划分为训练数据集和检测数据集,对于训练数据集,采用10重交叉验证的方法防止模型过拟合,增加模型泛化能力.

3.1 CNN网络结构的确定

采用网格搜索方法确定最优网络结构,即网络深度,卷积层卷积核大小,卷积核个数.模型精度评价指标选取MSE,MAPE和 R^2.网络结构中损失函数采用MSE,并采用Adam优化算法,设置优化算法学习率为0.001,全连接层丢弃率为0.3,卷积核移动步长为1步,激活函数为修正线性单元.得到不同网络结构模型精度如表2所示.由表2可得:随网络结构深度加深,对区域内用地类型推断结果精度越高;网络层数越深,产生过拟合的风险越大,如网络结构"卷积层(4)→卷积层(16)→卷积层(16)→卷积层(8)→全连接层"结构下产生了过拟合现象.同时,比较同一网络结构下卷积核尺寸大小发现,卷积核尺寸越小,模型推测效果越好,表明出行数据集在相邻短时间间隔内出行特征较为类似,时间间隔越长,出行特征波动越大,不利于

信息提取.

最终,用于交通小区用地类型特征推断的网络结构参数设置如下:卷积层(4)→卷积层(16)→卷积层(8)→全连接层,卷积核尺寸大小为2.为验证构建的卷积神经网络模型对用地类型推测效果,采用本文方法与机器学习模型分别对3种用地类型推测结果对比分析,结果如表3所示.由表3可得,Random Forest和SVR方法结果的 R^2 与CNN相近,但MSE和MAPE指标与CNN模型差别较大.

表2　不同网络结构下的精度分布

Table 2　Performance of different hidden layers and algorithms

网络结构	卷积核尺寸	MSE			MAPE/%			R^2		
		R	W	L	R	W	L	R	W	L
卷积层(4)→全连接层	2	0.38	0.29	0.39	12.63	12.73	13.06	0.43	0.38	0.33
	3	0.41	0.24	0.48	12.03	12.61	13.35	0.37	0.31	0.29
	4	0.42	0.44	0.39	12.25	12.39	13.19	0.32	0.36	0.34
	5	0.45	0.57	0.35	12.86	12.26	13.17	0.29	0.35	0.32
	6	0.37	0.11	0.38	12.48	12.04	13.67	0.25	0.32	0.26
卷积层(4)→卷积层(8)→全连接层	2	0.32	0.22	0.25	9.61	9.31	9.73	0.52	0.43	0.36
	3	0.36	0.26	0.34	9.05	9.24	9.79	0.48	0.4	0.32
	4	0.38	0.28	0.23	9.74	9.5	9.08	0.55	0.36	0.28
	5	0.34	0.25	0.25	9.54	9.69	9.28	0.52	0.32	0.22
	6	0.31	0.26	0.28	9.55	9.56	9.27	0.4	0.3	0.25
卷积层(4)→卷积层(16)→卷积层(8)→全连接层	2	0.106	0.127	0.131	6.83	7.04	8.25	0.78	0.71	0.66
	3	0.12	0.15	0.16	6.89	7.33	8.34	0.69	0.65	0.58
	4	0.15	0.136	0.18	7.07	7.21	8.09	0.77	0.59	0.55
	5	0.16	0.148	0.22	7.46	7.4	8.89	0.64	0.63	0.53
	6	0.15	0.16	0.23	7.94	7.47	8.85	0.65	0.52	0.49
卷积层(4)→卷积层(16)→卷积层(16)→卷积层(8)→全连接层	2	0.001	0.003	0.008	1.99	1.44	2.71	0.99	0.98	0.9
	3	0.001	0.002	0.01	2.32	1.75	5.2	0.93	0.89	0.87
	4	0.001	0.005	0.015	3.42	2.11	3.3	0.91	0.85	0.82
	5	0.002	0.003	0.012	1.77	1.71	2.03	0.9	0.88	0.79
	6	0.002	0.003	0.02	1.88	1.58	1.23	0.87	0.87	0.72

表3　不同模型推测结果比较

Table 3　Comparison analysis on land use inference using different models

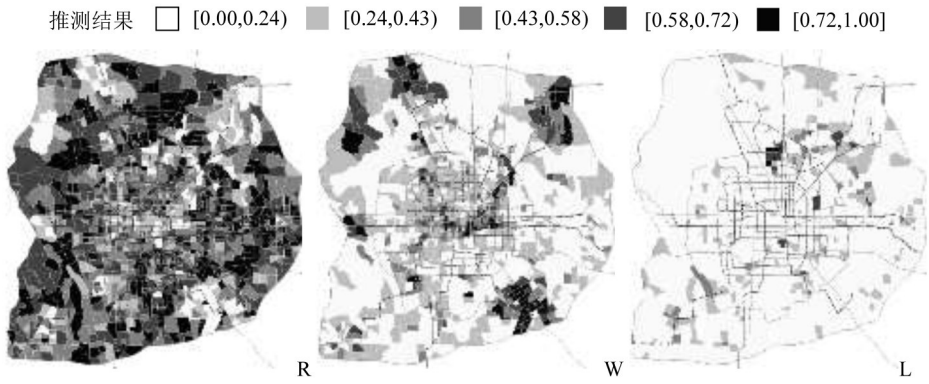
模　型	MSE			MAPE/%			R^2		
	R	W	L	R	W	L	R	W	L
CNN	0.106	0.127	0.131	6.83	7.04	8.25	0.78	0.71	0.66
SVR	0.36	0.47	0.58	12.77	15.71	16.03	0.77	0.68	0.67
Random Forest	0.32	0.42	0.60	10.88	12.58	13.23	0.79	0.70	0.62

3.2　用地类型特征辨识结果分析

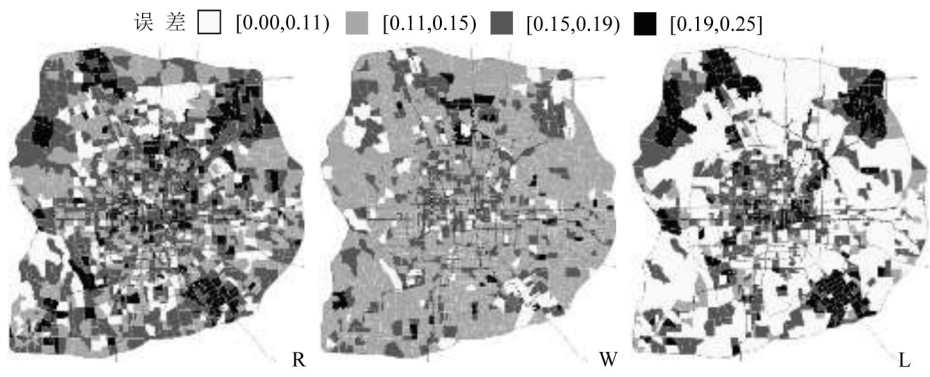
CNN网络模型对区域用地类型推测结果空间分布如图3所示.图3(a)为模型推测的交通小区内3种用地类型占比分布,图3(b)为推测误差.由图3和表3可得,构建CNN网络模型对区域内用地类型推测可以取得较好结果.可利用公共交通系统出行数据集,网约出行数据集和所构建的网络模型对城市其他区域内用地类型进行推测.

六环内大部分交通小区内主要用地成分为居住用地类型,样本多反映了居住用地类型出行特征,故对居住用地类型推测结果比其他两种用地类型准确;以工作用地类型为主导的交通小区呈现集中分布趋势,主要集中分布在上地,亦庄,望京,CBD;休闲用地类型为主的交通小区主要为城市公园,样本数量较少,误差较其他两种用地类型大.对比用地类型占比分布和与其对应的误差分布可得,当交通小区内某种用地类型组分占比超过60%,推测结果误差较小,即如果区域内存在某

种用地类型特征明显高于其他两种，则其误差较小；相反，如果区域内各用地类型占比没有明显差别时，往往导致利用CNN模型对区域内用地类型的推测结果较差；当用地组分占比小于30%时，误差较大.



(a) 各用地类型推测结果分布



(b) 各用地类型推测误差分布

图3 推测及误差分析结果

Fig. 3 Inferring result and error analysis in study area

4 结 论

本文构建基于CNN深度学习模型对交通小区内几种用地类型同时推断的方法.基于用地类型推测结果，可用于交通规划用地类型合理性评估，研究交通需求特征与用地类型的相关关系等，对于面向区域需求管理的交通政策制定具有重要意义.主要结论如下：选取交通小区发生强度，吸引强度和发生吸引强度差作为用于推断用地类型的特征；将公交和网约车出行数据集融合引入CNN深度学习模型，采用网格寻优方法确定网络结构，可对交通小区内居住，工作和休闲用地类型同时进行推测；深度学习对用地类型推测效果优于机器学习算法.实例分析结果表明：交通小区内用地类型分布会影响推测结果，各种用地类型分布失衡将损失占比较小用地类型的推测结果；同时，各种用地类型平衡分布情况推测结果误差较大.后续研究将讨论不同深度学习模型对推测结果的影响，应考虑不同用地类型标签分布对推测结果的影响程度.

参考文献：

[1] VAN ACKER V, WITLOX F. Commuting trips within tours: how is commuting related to land use?[J]. Transportation, 2010, 38(3): 465-486.

[2] LIU Y, WANG F, XIAO Y, et al. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai[J]. Landscape and Urban Planning, 2012, 106(1): 73-87.

[3] PAN G, QI G, WU Z, et al. Land-use classification using taxi GPS traces[J]. Ieee T Intell Transp, 2013, 14 (1): 113-123.

[4] TOOLE J L, ULM M, GONZáLEZ M C, et al. Inferring land use from mobile phone activity[C]// Proceedings of the ACM SIGKDD International Workshop on Urban Computing, Beijing: Association for Computing Machinery, 2012: 1-8.

[5] ZHAN X, UKKUSURI S V, ZHU F. Inferring urban land use using large-scale social media check-in data[J]. Networks and Spatial Economics, 2014, 14: 647-667.

[6] ZHAO J, FAN W, ZHAI X. Identification of land-use characteristics using bicycle sharing data: A deep learning approach[J]. Journal of Transport Geography, 2020, 82(102562).

[7] MA X, ZHUANG D, HE Z, et al. Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction[J]. Sensors, 2017, 17(4): 818-834.

[8] 赵晋. 基于精细化人群分类的公交路径选择模型研究 [D]. 北京: 北京工业大学, 2017. [ZAHO J. Research on public transport route selection model based on meticulous population classification[D]. Beijing: Beijing University of Technology, 2017.]

[9] SPÄRCK JONES K. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 2004, 60(5): 493-502.

[10] YUAN J, ZHENG Y, XIE X. Discovering regions of different functions in a city using human mobility and POIs[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing: Association for Computing Machinery, 2012: 186-194.

上接第 20 页

[7] 黄新婷. 产业结构变动对交通运输需求的影响预测及实证分析[D]. 西安:长安大学, 2016. [HUANG X T. Prediction and empirical analysis of the impact of industrial structure changes on transportation demand [D]. Xi'an: Chang'an University, 2016.]

[8] 国务院. 国务院关于中西部地区承接产业转移的指导意见[EB/OL]. (2010-09-06) [2020-09-18]. http://www.gov. cn/zhengce/content/2010-09/06/content_1536.htm. [State Council. Guiding opinions of the State Council on undertaking industrial transfer in central and western regions[EB/OL].(2010- 09- 06) [2020- 09- 18]. http://www.gov.cn/zhengce/content/2010-09/06/content.Htm]

[9] SUN J W. Changes in energy consumption and energy intensity: A complete decomposition model[J]. Energy Economics, 1998, 20(1): 85-100.

[10] 刘红光, 刘卫东, 刘志高. 区域间产业转移定量测度研究: 基于区域间投入产出表分析[J]. 中国工业经济, 2011(6): 79- 88. [LIU H G, LIU W D, LIU Z G. Quantitative measurement of interregional industrial transfer based on interregional input-output table[J]. China Industrial Economy, 2011(6): 79-88.]

[11] NORTON R D, REES J. The product cycle and the spatial decentralization of American manufacturing[J]. Regional Studies, 2007,13(2): 141-151.

[12] 胡秋阳. 投入产出分析: 理论、应用和操作[M]. 北京: 清华大学出版社, 2019. [HU Q Y. Input output analysis: Theory, application and operation[M]. Beijing: Tsinghua University Press, 2019.]