

基于路网相似性的路段行程时间估计*

郑业晴¹, 朱欣焰^{1,2}, 张发明^{1†}, 芮维^{1,2}, 张东娟¹, 曾聪³

(1. 武汉大学 测绘遥感信息工程国家重点实验室, 武汉 430079; 2. 地球空间信息技术协同创新中心, 武汉 430079; 3. 香港理工大学 土地测量与地理资讯学系, 香港)

摘要: 虽然浮动车 GPS 数据量很大,但是在某些时段仍然有一些路段缺少实时浮动车数据,从而难以对行程时间进行估计。针对实时浮动车实时数据在估计路段行程时间时存在数据缺失的问题,提出了利用路网间属性和空间结构的相似性,从浮动车历史大数据中提取相似路段之间的时空关联特征,以目标路段与相似路段间的时空关联关系为输入,目标路段行程时间为输出,利用构建的三层神经网络模型进行数据缺失路段的行程时间估计。实验结果表明,路段行程时间估计值的平均绝对百分比误差可达到 30%,与 Naïve model (NM) 相比具有较好的估计精度,验证了从路网相似角度解决实时数据缺失的可行性。

关键词: 路网相似性; 大数据; 数据缺失; 时空关联; 人工神经网络

中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2018)06-1681-05

doi:10.3969/j.issn.1001-3695.2018.06.018

Link travel time estimation based on road similarity

Zheng Yeqing¹, Zhu Xinyan^{1,2}, Zhang Faming^{1†}, Guo Wei^{1,2}, Zhang Dongjuan¹, Zeng Cong³

(1. State Key Laboratory of Information Engineering in Surveying, Mapping & Remote Sensing, Wuhan University, Wuhan 430079, China; 2. Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China; 3. Dept. of Land Surveying & Geo-Informatics, Hong Kong Polytechnic University, Hong Kong, China)

Abstract: Despite the large amount of floating car data, there are still some links lack of real-time data during some certain period of time. Therefore, it is difficult to estimate the travel time. Considering the problem of sparse data when using floating car data estimate the travel time, this paper put forward a kind of inferred method based on big data of floating car. It designed a three-layer artificial neural network model, whose input information and output information were the spatio-temporal correlation and the target link travel time respectively, and obtained traffic spatio-temporal correlation relationship using historical big data of floating car and then inferred the travel time of target link. The experiment results indicate that the mean absolute percentage error of travel time can achieve to 30 percent. Compared with Naïve model, the accuracy of ANN model performs better, and it is practical to solve the problem of data missing from the aspect of road network similarity.

Key words: similarity of road network; big data; data sparsity; spatiotemporal association; artificial neural network (ANN)

0 引言

行程时间预测作为智能运输系统众多方面的一个关键环节,已成为智能运输系统重要理论研究的一部分^[1]。浮动车 GPS 接收器作为交通状态传感器,采集的信息包括实时速度、时间戳、经纬度坐标以及方位角等,一定程度反映了城市交通的运行状态,对城市路网中路段实时或准实时行程时间的估计发挥着重要作用^[2]。虽然浮动车 GPS 轨迹数据量大,但仍存在浮动车 GPS 采集信息的低频性以及行驶区域的局限性,在某些时段,某些路段存在数据稀疏甚至数据缺失的问题^[3]。因此,如何利用浮动车大数据推断路段行程时间是一个亟待解决的问题。

目前,已有许多基于模型的方法利用浮动车数据来估计路段行程时间,大致可分为卡尔曼滤波法、统计模型法、交通模拟等。Jula 等人^[4]提出了在随机动态网络中估计路段行程时间

的数学模型,并用卡尔曼滤波的方法估计未来时刻的路段行程时间;Jenelius 等人^[5]提出了一个统计模型用于城市路网行程时间的估计。然而,在数据缺失的情况下,上述方法已不能对路段的行程时间进行有效估计。半监督学习算法或转移学习算法^[4]、矩阵分解算法和协同过滤^[6]、基于相似性的聚类算法,分别在解决空气质量分析、城市油耗估计等存在的数据缺失性问题已有相关研究,但在交通领域,尤其是解决浮动车数据缺失情况鲜有研究。其中,最简单方便的就是前一时刻速度推估法,虽然简单易行,但是精度不高;多元线性回归拟合推估法^[7]精度相对较高,但此方法对数据的完整性有较高要求,在浮动车数据缺失的情况下,对模型的选择存在一定影响;还有研究^[8]利用基于朴素贝叶斯的估计模型,对稀疏路段的行程时间进行估计,算法复杂度较低,运算效率较高。

针对数据缺失问题,本文提出了基于路网相似性特征神经网络模型,从历史交通大数据中提取目标路段和相似路段间

收稿日期: 2017-01-19; **修回日期:** 2017-04-13 **基金项目:** 测绘遥感信息工程国家重点实验室重点开放基金资助项目; 国家科技支撑资助项目(2012BAH35B03); 测绘遥感信息工程国家重点实验室专项科研经费资助项目

作者简介: 郑业晴(1993-),女,硕士研究生,主要研究方向为 WebGIS、交通 GIS; 朱欣焰(1963-),男,教授,博士,主要研究方向为空间信息服务、空间数据库等; 张发明(1989-),男(通信作者),博士研究生,主要研究方向为交通 GIS、时空大数据分析与应用(zhang_fa_ming@163.com); 芮维(1981-),男,副教授,主要研究方向为三维建模; 张东娟(1995-),女,硕士研究生,主要研究方向为 WebGIS; 曾聪(1993-),男,硕士研究生,主要研究方向为三维建模。

的时空关联关系,并用该模型进行路段行程时间的推断。对每一个目标路段,提取目标路段与相似路段的属性、空间结构特征并计算它们之间的相似度,提取相似路段的交通特性作为模型输入,目标路段行程时间作为输出,用构建的三层神经网络模型对提取的特征进行训练,最后用训练的神经网络模型对路段行程时间进行推断。实验结果表明,路网中两路段的属性、空间相似度越高,路段速度期望之间的相关性越强。因此,用相似路段来推断数据缺失路段的行程时间所得的结果与真实值更为接近。在现有数据质量有限的情况下,提出的神经网络模型与 NM(Naïve model)相比,能较好地利用目标路段与相似路段的交通关联关系推断路段行程时间,从而解决由于浮动车实时 GPS 数据缺失而不能有效估计路段行程时间的问题。

1 相似性特征提取

从数学角度来讲,相似性是指两个图形之间形状的完全相似。本文所讲的相似性是指局部路网间的差异程度,包括路网属性特征和路网空间结构之间的相似性。路网相似性特征衡量^[9,10]的判断需要通过两个局部路网之间的属性信息的相似和空间结构信息的相似进行判断。

1.1 属性相似性特征

对于属性相似性,从以下六个方面衡量路网在属性信息间的相似性。

a)路段的长度(length)。路段长度用符号 len 表示,单位为 m。

b)路段端点间欧氏距离(Euclid)。路段两端点间的欧氏距离用符号 euc 表示,单位为 m。

c)路段弯曲度(curve)。将弯曲度定义为路段长度 len 与路段端点间欧氏距离 euc 之间的比值,如式(1)所示,该比值越大,则该路段越弯曲;反之,该路段越平直,路段弯曲度用符号 cur 表示。

$$\text{cur} = \text{len}/\text{euc} \quad (1)$$

d)路段的方向(direction)。路段方向定义为路段的行驶方向,该道路为单向通车或者双向通车,路段方向用符号 dir 表示。若路段为单向通车,则数值化为 1,若为双向通车,则数值化为 2。

e)路段的车道数(lanes)。路段的车道数量用符号 lan 表示。

f)道路等级(rank)。道路为主干道、次干道或者支路,用符号 ran 表示。





1.2 结构相似性特征

结构相似性定义为两个对象在空间结构上存在的相似程度。对于路网空间结构相似性,本文从以下几个方面衡量它们的相似性。

1)路段入度(in degree)

入度指路网中与路段起点直接相连的路段个数,相对于交通流向而言,若路段入度为 0,则路网中路段的起点不与任何边相连。路段度的类型如表 1 所示。

表 1 路段的度类型

路段类型	路段的度	路段类型	路段的度
	0		2
	1		3

2)路段出度(degree)

出度指路网中根据交通流向与路段终点相连的路段个数。若出度为 0,则路网中路段的终点不与任何边相连。

3)POI 强度(intensity, point of interest)

指单位面积上 POI 对交通的影响度。路段周边 POI 种类越多,不同类型 POI 密度越大,则 POI 强度越大,POI 对交通影响越大^[11],图 1 表示研究区域路网与 POI 之间的关联关系。该图说明 POI 的分布并不是均匀的,主要分布在主干道、次干道等,POI 对不同道路的影响程度不同。

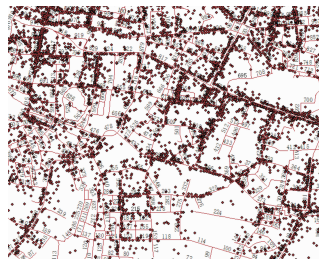


图 1 路网与 POI 分布之间的关系

假设 POI 类型可以分为 n 类,符号 p_i 、 c_{p_i} 、 w_{p_i} 和 s 分别表示 POI 的类别、某类 POI 的数量、某类 POI 对交通影响的权重以及缓冲区面积(km^2),则 POI 强度可根据式(2)计算。

$$\text{intensity} = \frac{\sum_{i=1}^n c_{p_i} \times w_{p_i}}{s} \quad i = 1, 2, \dots, n \quad (2)$$

根据 Esri 的 POI 分类标准^[12],结合 POI 对交通影响的特点,将现有 POI 类型分为 13 类,分别为餐饮、购物、住宿、出行、文体、娱乐、金融服务、教育、医疗、房产、旅游、企事业单位以及行政机构。由于不同时段、不同种类 POI 的活跃度不同,其对交通影响的权重不同,所以,不同种类 POI 对交通影响的权重是随时间变化的。每天不同时段不同类别 POI 的权重不同,工作日与非工作日的权重也不同,因此,定义随时段变化的 POI 权重为 1、2 两个等级,基本权重为 1,代表 POI 权重具有基本影响,2 代表相比之下 POI 具有较高权重,对交通具有较大影响,如式(3)所示。

$$w_{p_i} = \begin{cases} 1 & \text{基本权重} \\ 2 & \text{高峰时段权重} \end{cases} \quad (3)$$

在早高峰时段^[13](7:00~9:00),住宿、出行、教育、医疗、房产、企事业单位以及行政机构对交通影响较大,POI 具有较高权重;晚高峰时段(17:00~20:00),餐饮、购物、住宿、出行、教育、医疗房产、企事业单位、行政机构等对交通影响比较大,相应权重较高;在周末或者假期,餐饮、购物、出行、文体娱乐、生活服务、旅游等对交通具有更大的影响度。表 2 表示不同路段在不同时段的 POI 影响度。

表 2 不同时间段的 POI 影响度

link ID	早高峰影响度 (7:00~9:00)	晚高峰影响度 (17:00~20:00)	夜间影响度 (20:00~22:00)	其他时段 影响度
817	68	74	62	41
639	84	98	80	57
535	67	92	82	52

2 相似性度量标准

相似性度量(similarity),即计算个体间的相似程度,若相似性度量的值越大,说明个体间相似度越小,差异越大;反之,相似程度越大。

2.1 路段特征的向量表达

每一个路段具有属性特征和结构特征,只从语义上难以衡量两个路段之间存在的相似性。因此,为衡量两个路段之间的特征,首先需要对路段属性信息和空间信息分别数值化。同时,为便于衡量两个路段间的相似性,本文将路段的属性信息和空间结构信息分别表达为两个一维向量:属性向量(attribute vector)和空间结构向量(structural vector)。

根据以上定义,属性向量可表达为:属性向量=(路段的长度,路段端点间欧氏距离,路段弯曲度,路段的方向,路段的车道数,道路等级)。相应符号可表达如下:

$$\text{attVector} = (\text{len}, \text{euc}, \text{cur}, \text{dir}, \text{lan}, \text{ran})$$

空间结构向量可表达为:结构向量=(路段入度,路段出度,POI)。利用相应符号可表达如下:

$$\text{strVector} = (\text{degIn}, \text{degOut}, \text{POI})$$

2.2 相似度计算

计算个体间的相似程度有多种计算方法,一般有相似度度量方法和距离度量方法。余弦相似度是利用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小,与距离度量相比,余弦相似度更加注重两个向量在方向上的差异而非距离,其对绝对的数值不敏感。因此,本文中采用欧几里德距离相似度来衡量多维空间中两个向量的相似程度。

距离相似度:假设 X 和 Y 两个变量,它们都包含了 N 个维的特征,分别表达为向量形式,即 $X = (x_1, x_2, x_3, \dots, x_n)$, $Y = (y_1, y_2, y_3, \dots, y_n)$,则距离相似度^[14]的计算公式如下:

$$\text{sim}(X, Y) = \text{dis}(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (4)$$

分别根据相似度计算表达式计算属性相似度(attSim)和空间结构相似度(strSim),则相似度最终可表达为

$$\text{sim} = a \times \text{attSim} + (1 - a) \times \text{strSim} \quad (5)$$

为了突出属性或者空间结构方面的相似度,可通过调整 a 值来突出其重要性,本文中路网属性和空间结构具有同等重要性,因此 a 取值为 0.5。

3 神经网络模型

本研究选择神经网络模型作为训练模型。神经网络的类型很多,很多神经网络可以作为分析模型,本文选择目前应用最广泛的神经网络模型之一——BP 神经网络(back-propagation neural network)作为分析模型。神经网络模型结构如图 2 所示。

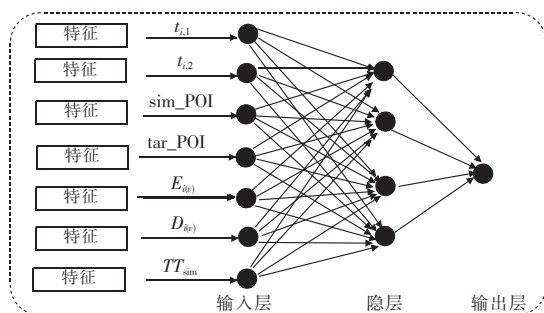


图2 BP神经网络模型结构

1) 输入层

$$X_i = (t_{i,1}, t_{i,2}, \text{sim_POI}_i, \text{tar_POI}_i, E_{i(v)}, D_{i(v)}, TT_{\text{sim}})$$

其中: x_i 代表输入层第 i 个向量; $t_{i,1}$ 表示第 i 个向量中车辆通

过相似路段和目标路段为一周中的哪一天; $t_{i,2}$ 表示第 i 个向量中车辆通过相似路段和目标路段为一天中的哪个时段; sim_POI_i 表示第 i 个向量中相似路段的 POI 影响度; tar_POI_i 表示第 i 个向量中目标路段的 POI 影响度; $E_{i(v)}$ 代表第 i 个向量中的速度期望; $D_{i(v)}$ 表示第 i 个向量中速度标准差; TT_{sim} 表示第 i 个向量中相似路段的行程时间。因此,神经网络输入包括一周中哪一天、一天中哪一时段、相似路段 POI 影响度、目标路段 POI 影响度、相似路段速度期望、相似路段速度方差、相似路段行程时间,输入层共有七个神经元。

2) 隐层

$$H(i) = \begin{bmatrix} h_1(i) \\ \vdots \\ h_n(i) \end{bmatrix} = \begin{bmatrix} \varphi(\sum_{j=1}^n \omega_{j,1} x_j(i) + b_1) \\ \vdots \\ \varphi(\sum_{j=1}^n \omega_{j,n} x_j(i) + b_n) \end{bmatrix} \quad (6)$$

其中: $h_n(i)$ 代表隐层第 n 个神经元的值; $\omega_{j,n}$ 代表输入层第 j 个神经元与隐层第 n 个神经元之间的链接权重; b_n 代表神经网络阈值; φ 代表一个转换函数,转换函数选择 S 形函数。

3) 输出层

$$y(i) = \varphi(\sum_{k=1}^n \omega_k h_k(i) + b) \quad (7)$$

其中: $y(i)$ 代表待估计时间段内目标路段的行程时间; ω_k 代表隐层第 k 个神经元与输出层神经元的连接权重; b 代表输出层的阈值; φ 代表转换函数。

4 实验验证

4.1 实验数据

本文选择的研究区域路网如图 3 所示,该研究区域内的路网主要由城市主干道、次干道和支路构成。

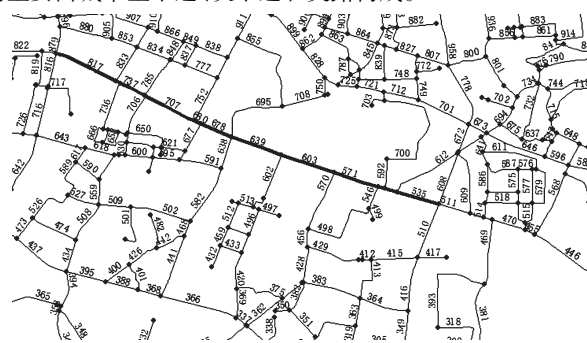


图3 珞瑜路中的部分路段

本实验选择珞瑜路中的部分路段 817, 743, 737, 707, 680, 678, 639, 571, 547, 545, 535 作为实验路段;利用 2014 年 3 月 1 日到 7 月 30 日历史浮动车数据作为实验数据,包括车辆进入路段时刻、车辆 ID、经度、纬度、车辆行驶速度、车辆行驶方向以及车辆载客状态,进行详细分析计算,验证以上所提出模型的有效性。

4.2 相似路段特征提取

选择研究区域内的两两路段作为研究目标,首先将每条路段的属性特征和空间结构特征进行归一化处理,计算两两路段之间的距离相似度。从前面讨论可知,POI 对交通产生的影响随时间变化,因此路段相似度也是随时间变化的,分别为早高峰时段相似度、晚高峰时段相似度、夜间时段相似度以及其他

时段相似度。

计算实验区域内路段间不同方向、不同时段的两两相似度,表 3 表示路段 817 和 639 在不同方向不同时间段的相似度。从表中可知,在四个不同时段中,相同方向两路段的相似度值在 0.15 左右,相反方向两路段的相似度值在 0.5 左右,因此,路段在不同时段的相似度相近,可用同一值替代。

表 3 路段 817 和 639 在不同方向不同时间段的相似度

时间段	相同方向相似度	相反方向相似度
早高峰	0.151 199	0.495 058
晚高峰	0.178 311	0.504 001
夜间	0.153 545	0.495 780
其他时段	0.137 273	0.490 984

本实验利用 2014 年 7 月 21 日到 7 月 27 日一周的历史数据计算两路段速度期望之间的 Pearson 系数,并进行相关性检验。

表 4 表示路段 817 与 639 之间不同方向交通流的相关性。从表中可得到,有 95% 的可能性表明路段 817 与 639 同一方向的速度期望之间具有一定的正相关性,相反方向不相关,这与实际交通情况相符。图 4 表示路段 817 与 639 在 2014 年 5 月 3 日 0 时到 20 时期间的平均速度变化折线图,可见两条折线的变化趋势大致相同,同样反映了路段 817 与 639 在同一方向上的速度期望之间具有一定的正相关关系。

表 4 路段间相关系数

目标路段	相似路段	目标路段方向	相似路段方向	相关系数 r	显著性水平 p
817	639	+1	+1	0.422 68	$*1.786\ 3e-05$
		+1	-1	-0.041 27	0.689 74
		-1	+1	-0.054 19	0.599 96
		-1	-1	0.267 73	$*0.008\ 361\ 1$

* 表示在 0.05 置信水平上显著相关(双侧)

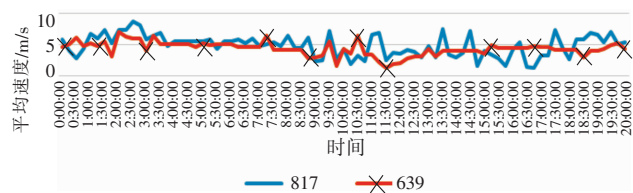


图 4 路段 817 与 639 之间速度期望的相关关系

表 5 表示路段 817 与 639 不同方向之间的相似度与 Pearson 相关系数之间的关联关系。通过分析区域路网内两两路段之间不同方向的路段相似度与相关性之间的关系可以发现,总体上距离相似度值越小,相关系数越大,即路网中的两路段属性、空间相似度越高,路段速度期望之间的相关性越强。因此,用相似路段来推断数据缺失路段的行程时间所得的结果与真实值越接近。

表 5 路段 817 与 639 相似度与 Pearson 相关系数关联关系

目标路段	相似路段	目标路段方向	相似路段方向	距离相似度	相关系数
817	639	+1	+1	0.155 082	0.422 68
		+1	-1	0.496 456	-0.041 267
		-1	+1	0.496 456	-0.054 2
		-1	-1	0.155 082	0.267 73

综合路段之间的相似性以及相关性,将路段数据按距离相似度值由小到大排序,当路段间相似度的阈值为 0.27 时,有 20% 的路段存在较强相似性并且相关,这符合二八定律的原理^[15],即小部分路段具有高度相似性以及相关性。因此,当两

路段之间的距离相似度小于该阈值时,可认为两路段间具有一定的相似性。

根据前面的描述,选择两相似路段之间的特征值(时段 1($t_{i,1}$),时段 2($t_{i,2}$),目标路段 POI 影响度(POI),相似路段 POI 影响度,相似路段的速度期望(E),相似路段速度方差(D),相似路段的行程时间(TT_{sim}))作为模型输入特征,目标路段行程时间(TT_{tar})作为模型输出特征,提取相似路段 817 与 639 的输入输出特征值,特征值如表 6 所示。

表 6 路段 817 与 639 的输入与输出特征值

t_1	t_2	tar_POI	sim_POI	TT_{sim}	E	D	TT_{tar}
7	1	256.859 4	443.035 6	100.997 516	4.71	0	103.938 415
6	93	256.859 4	443.035 6	117.167 069	4.06	0.89	75.980 721
3	36	426.010 8	652.894 6	200.716 582	2.37	0	143.992 732
2	93	256.859 4	443.035 6	105.010 662	4.53	0.04	79.995 962

4.3 模型训练与验证

在利用神经网络模型进行路段行程时间估计之前,神经网络模型的训练是必不可少的过程。将获得的整个数据集分成两部分,训练神经网络数据和评估数据,其中训练神经网络数据分成训练数据集、验证数据集和测试数据集,分别用于神经网络训练、验证和测试。训练数据集用于神经网络的训练,验证数据集用于防止神经网络出现过拟合现象,测试数据集则用于神经网络性能的测试,评估数据则用于验证神经网络模型的性能。

为了评估神经网络模型的性能,引进平均绝对百分比误差(mean absolute percentage error, MAPE)和均方根误差(root mean square error, RMSE)作为评估的性能指标,如式(8)(9)所示。

$$MAPE = 100\% \times \frac{1}{n} \sum_{i=1}^n \frac{|t_{fc,i} - t_{true,i}|}{t_{true,i}} \quad (8)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (t_{fc,i} - t_{true,i})^2} \quad (9)$$

其中: $t_{fc,i}$ 表示某时刻路段行程时间的预测值; $t_{true,i}$ 表示某时刻路段行程时间的实际值。

本实验的数据集分割情况如表 7 所示,神经网络的训练函数选择 Levenberg-Marquardt 算法^[16],学习速率设置为 0.01,以保持神经网络的稳定。表 7 的结果表明在不同数据量情况下,通过 ANN 模型预测的路段行程时间平均绝对百分比误差约为 30%,并且数据量越大,平均绝对百分比误差和均方根误差越小,预测精度越高。

表 7 神经网络模型训练和评估数据集

数据总量	ANN 训练			ANN 评估		
	训练数据 (百分比)	验证数据 (百分比)	测试数据 (百分比)	评估数据 (百分比)	MAPE	RMSE
5 284	2 644(50%)	528(10%)	528(10%)	53(1%)	0.318 5	58.839 8
				264(5%)	0.309 3	57.244 6
				528(10%)	0.302 9	54.682 1

4.4 模型对比

为了验证模型的有效性,本文与 NM^[17]进行对比,NM 可以用式(10)表示。

$$\hat{Z}_t = Z_{(t-1)} + \varepsilon_t \quad (10)$$

其中: \hat{Z}_t 表示 t 时刻的预测值; $Z_{(t-1)}$ 表示 $t-1$ 时刻的历史值; ε_t 表示均值为零的残差。

将 ANN 训练模型的性能指标与 NM 进行对比,用三组不

同数据分别预测缺失时段的行程时间,并计算两性能指标。表8表示在不同数据量的情况下,两模型的MAPE值与RMSE值的对比结果。表中结果显示,当选取10%的数据量作为预测数据时,ANN模型的MAPE误差约为0.3,而NM明显高于ANN模型,约为0.4。同样,ANN模型的RMSE误差约为54.7, NM为75.7,精度大幅低于ANN模型。

表8 ANN模型与NM的行程时间评估性能指标

数据量(百分比)	53(1%)	264(5%)	528(10%)
MAPE	ANN模型	0.318 5	0.309 3
	NM	0.721 8	0.429 6
RMSE	ANN模型	58.839 8	54.682 1
	NM	108.313 0	75.744 9

在图5中用直方图更加直观地反映了两模型在评估精度上的差异。由图可见,随着数据量的增大,ANN模型和NM的误差值都有降低趋势,而且ANN模型的两项误差值均低于NM,具有更好的预测精度。由于现有数据质量问题,目前只能达到现有精度,但结果显示,ANN模型的平均绝对百分比误差和均方根误差仍优于NM。

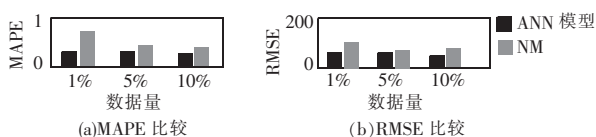


图5 ANN模型与NM的行程时间评估性能指标直方图

5 结束语

浮动车的运行状态在一定程度上可以反映道路交通拥堵状况,因此,可以利用浮动车的交通信息对路段行程时间进行推断。以往的研究大多基于数据充足的条件下推断行程时间,但是在实时数据缺失的情况下,已有方法很难对数据缺失路段的行程时间进行有效估计,并且多数估计方法通过仿真数据实现,应用价值不高。本研究首次提出利用属性信息、空间信息衡量路段间的相似性,并利用历史交通大数据提取相似路段间的时空关联关系,通过设计的三层神经网络模型对路段行程时间进行推断的方法。实验结果表明,在数据缺失情况下,用该模型对目标路段行程时间进行推断,能达到较好的实验效果,验证了本文方法的有效性。本文现有的模型方法仍受限于交通设施、车道占用等特殊因素的影响,因此后续研究需要综合考虑这些因素,以提高路段行程时间推断的准确度。同时,由于现有数据质量问题,目前只能达到现有精度。随着数据质量的改善,相信未来的研究能达到更高的估计精度,并实现对区域性路网行程时间的估计,拓宽应用范围。

参考文献:

- [1] 杨兆升. 关于智能运输系统的关键理论——综合路段行程时间预测的研究[J]. 交通运输工程学报, 2001, 1(1): 65-67, 89.
- [2] 张发明, 朱欣焰, 芮维, 等. 利用浮动车大数据进行稀疏路段行程时间推断[J]. 武汉大学学报: 信息科学版, 2017, 42(1): 56-62.
- [3] Zheng Yu, Liu Furui, Hsieh H P, et al. U-Air: when urban air quality inference meets big data [C]//Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2013: 1436-1444.
- [4] Julia H, Dessouky M, Ioannou P A. Real-time estimation of travel times along the arcs and arrival times at the nodes of dynamic stochastic networks [J]. IEEE Trans on Intelligent Transportation Systems, 2008, 9(1): 97-110.
- [5] Jenelius E, Koutsopoulos H N. Travel time estimation for urban road networks using low frequency probe vehicle data [J]. Transportation Research, Part B: Methodological, 2013, 53(4): 64-81.
- [6] Zhang Fuzheng, Wilkie D, Zheng Yu, et al. Sensing the pulse of urban refueling behavior [C]//Proc of ACM International Joint Conference on Pervasive and Ubiquitous Computing. New York: ACM Press, 2013: 13-22.
- [7] 刘春, 黄美娟, 杨超. 浮动车数据缺失道路的速度推估模型与实现 [J]. 同济大学学报: 自然科学版, 2010, 38(8): 1255-1260.
- [8] 王晓蒙, 彭玲, 池天河. 基于稀疏浮动车数据的城市路网交通流速度估计[J]. 测绘学报, 2016, 45(7): 866-873.
- [9] 袁晶. 大规模轨迹数据的检索、挖掘和应用[D]. 合肥: 中国科学技术大学, 2012.
- [10] 袁冠. 移动对象轨迹数据挖掘方法研究[D]. 徐州: 中国矿业大学, 2012.
- [11] 塔娜, 柴彦成, 关美宝. 建成环境对北京市郊区居民工作日汽车出行的影响[J]. 地理学报, 2015, 70(10): 1675-1685.
- [12] 张玲. POI的分类标准研究[J]. 测绘通报, 2012(10): 82-84.
- [13] Zhang Faming, Zhu Xinyan, Guo Wei, et al. Analyzing urban human mobility patterns through a thematic model at a finer scale [J]. ISPRS International Journal of Geo-Information, 2016, 5(6): D10: 10.3390/ijgi5060078.
- [14] Deza M M, Deza E. Encyclopedia of distances [M]//Encyclopedia of Distances. Berlin: Springer, 2009: 1-583.
- [15] Kiremire A R. The application of the Pareto principle in software engineering [EB/OL]. (2011-10-19). http://www2.latech.edu/~box/ase/papers2011/Ankunda_termpaper.PDF.
- [16] Zhang Faming, Zhu Xinyan, Hu Tao, et al. Urban link travel time prediction based on a gradient boosting method considering spatiotemporal correlations [J]. International Journal of Geo-Information, 2016, 5(11): DOI: 10.3390/ijgi5110201.
- [17] Cheng Tao, Wang Jiaqi, Haworth J, et al. A dynamic spatial weight matrix and localized space-time autoregressive integrated moving average for network modeling [J]. Geographical Analysis, 2014, 46(1): 75-97.

(上接第1671页)

- [5] 周水庚, 周傲英, 曹晶, 等. 一种基于密度的快速聚类算法[J]. 计算机研究与发展, 2000, 37(11): 1287-1292.
- [6] 王桂芝, 王广亮. 改进的快速 DBSCAN 算法[J]. 计算机应用, 2009, 29(9): 2505-2508.
- [7] Zhao Yanchang, Cao Jie, Zhang Chengqi, et al. Enhancing grid-density based clustering for high dimensional data [J]. Journal of Systems and Software, 2011, 84(9): 1524-1539.
- [8] Karypis G, Han E H, Kumar V. Chameleon: hierarchical clustering using dynamic modeling [J]. Computer, 1999, 32(8): 68-75.
- [9] 黄红伟, 黄天民. 基于网格相对密度差的扩展聚类算法[J]. 计算机应用研究, 2014, 31(6): 1702-1705.

- [10] 刘淑芬, 孟冬雪, 王晓燕. 基于网格单元的 DBSCAN 算法[J]. 吉林大学学报: 工学版, 2014, 44(4): 1135-1139.
- [11] 冯玲, 刘克剑, 唐福喜, 等. 一种基于网格查询的改进 DBSCAN 算法[J]. 西华大学学报: 自然科学版, 2016, 35(5): 25-29.
- [12] 张枫, 邱保志. 基于网格的高效 DBSCAN 算法[J]. 计算机工程与应用, 2007, 43(17): 167-169, 184.
- [13] Piegl L A, Tiller W. Algorithm for finding all k, nearest neighbors [J]. Computer-Aided Design, 2002, 34(2): 167-172.
- [14] <http://cs.joensuu.fi/sipu/datasets/> [EB/OL].
- [15] <http://www.pudn.com/downloads219/sourcecode/math/detail1030717.html> [EB/OL].
- [16] Khalid S, Razzaq S. TOBAE: a density-based agglomerative clustering algorithm [J]. Journal of Classification, 2015, 32(2): 241-267.