2020 年 第 20 卷 第 19 期
2020, 20(19): 07877-04
科 学 技 术 与 工 程
Science Technology and Engineering
ISSN 1671—1815
CN 11—4688/T

引用格式:崔洪军, 张晓阳, 朱敏清. 基于隐马尔可夫模型的公交乘客出行链识别方法[J]. 科学技术与工程, 2020, 20(19): 7877-7880
Cui Hongjun, Zhang Xiaoyang, Zhu Minqing. Recognition method of public passenger trip chain based on hidden markov model[J]. Science Technology and Engineering, 2020, 20(19): 7877-7880

交通运输

基于隐马尔可夫模型的公交乘客出行链识别方法

崔洪军[1], 张晓阳[1], 朱敏清[2]
(1. 河北工业大学土木与交通学院,天津 300401; 2. 河北工业大学建筑与艺术学院,天津 300401)

摘 要 以公交 IC 卡数据为基础,结合公交站点周边土地类型及乘客时空转移序列构建隐马尔可夫模型对乘客的出行目的进行识别,继而实现完整乘客公交出行链的提取。采用石家庄某一工作日的公交 IC 卡数据来实现模型,并与同日期的公交乘客出行调查数据对比,结果表明模型与实际调查结果较为吻合,工作日公交乘客出行链以通勤类出行链为主,总占比达 81.87%,通勤活动具有很强的时效性。本文模型为公交 IC 卡数据的深度研究提供了一定的基础。

关键词 交通运输系统;IC 卡;隐马尔可夫模型;出行链;出行特征
中图法分类号 U121; **文献标志码** A

Recognition Method of Public Passenger Trip Chain Based on Hidden Markov Model

CUI Hong-jun[1], ZHANG Xiao-yang[1], ZHU Min-qing[2]
(1. School of Civil and Transportation Engineering, Hebei University of Technology, Tianjin 300401, China;
2. School of Architecture & Art Design, Hebei University of Technology, Tianjin 300401, China)

[**Abstract**] Based on the bus IC card data, a Hidden Markov model was constructed to identify the traveling purpose of the passenger. By combining the land types around the bus station and the passenger space-time transfer sequence, the model extracted a complete bus traveling chain for passengers. The model was implemented by using Shijiazhuang's bus IC card data for a working day and compared it with the bus passenger travel survey data of the same date. The results show that the model is consistent with the actual survey results. The bus passenger traveling chain is dominated by commuting travel chains, with a total proportion of 81.87%, and commuting activities have a strong timeliness. The proposed model provides a certain basis for in-depth research on public transit IC card data.

[**Key words**] communications and transportation system; IC card; hidden Markov model; trip chains; travel characteristics

IC 卡的广泛应用在各大城市中产生了海量的乘客出行信息数据,基于 IC 卡数据样本量大、信息储存量高、更新速度快、较手工收集数据方法更为准确且成本低廉等特点,可用其对乘客的出行特征、交通运营情况等进行描述及刻画。通过对 IC 卡和 GPS 数据进行识别分析可获得更为完整的公交乘客出行信息,同时对于公共系统的规划发展有着重要的意义。

目前在中国运营的公交车辆存在着两种计费方式:一票制上车打卡方式(如天津公交)及分段计价上下车均打卡方式(如北京公交);而一票制的计费方式在中国被大多城市所采用。一票制计费缺少乘客的下车站点具体信息,因而无法直接利用所得数据推算乘客出行信息。如何高效、准确地补全乘客的下车站点信息成了研究的热点及难点。国外对基于 IC 卡数据的乘客出行研究相对较早:Zhao[1]基于出行链思想,结合自动收费系统(AFC)、自动定位系统(AVL),实现了公交-地铁、地铁-地铁两种出行方式的下车站点的推导。Alex[2]基于出行链模型对公交乘客下车站点进行了判断,并生成了种子矩阵以实现对不同规模 OD 矩阵的预测。Farzin[3]基于巴西圣保罗的公交乘客出行数据,以总数据量的 5% 为数据样本,推算了乘客的下客站点。Barry 等[4]提出经典出行链假说,并以此推断了纽约市公交乘客的下客站点。较之国外,中国对此方面的研究起步稍晚。胡郁葱等[5]通过数据挖掘技术

收稿日期: 2020-01-16; **修订日期**: 2020-02-13
基金项目: 国家自然科学基金(51678212)
第一作者:崔洪军(1974—),男,汉族,河北定州人,博士,教授。研究方向:交通工程。E-mail:cuihjl1974@163.com。

获取公交 OD 矩阵。陈峥嵘[6]利用智能公交数据处理方法对公交客流 OD 进行研究。胡继华等[7]结合出行链模型对公交乘客的下车站点进行概率推算。吴祥国[8]提出可根据乘客多日出行链进行下车站点判断,但并未提出相应的算法,只能通过人工识别。

由以上研究成果可知,国外研究人员由于研究数据较为丰富、全面,故利用出行链方法得到的推算结果其准确性较高;中国因行政区划、数据兼容性等原因,可用于研究的数据通常不够完整。因此,完全搬用国外对于乘客完整出行信息的获取方法不切合实际。本研究将根据公交乘客 IC 卡信息对乘客出行目的进行识别,进而提取公交乘客的完整出行链。

1 乘客出行链提取

提取出行链,需对乘客的出行目的进行完善。翁剑成等[9]通过整合多源数据,提出了出行链提取的四阶段法,朱亚迪等[10]基于概率图模型对乘客的出行链进行了提取,并分析了乘客的出行特征。本研究基于广义出行链思想构建连续隐马尔可夫模型(CHMM),进而对乘客的出行链进行提取并分析其出行目的。

1.1 模型构建

隐马尔可夫模型由马尔可夫模型及高斯混合分布共同构成,用五个元素来描述,其中包括隐藏状态集合 0、可观测状态集合、初始状态概率矩阵 $\boldsymbol{\pi}$、隐含状态转移概率矩阵 $\boldsymbol{A}$、观测状态转移概率矩阵 $\boldsymbol{B}$。其转化机理如图 1 所示。

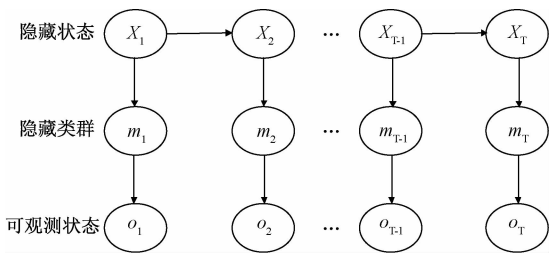


图 1 连续隐马尔可夫链转化机理

Fig. 1 The mechanics of determining states and observations behind a CHMM

选定的出行活动序列在其初始状态时的概率集如式(1)所示:

$$\boldsymbol{\pi} = \{\pi_i\} = \{P(x_i = i)\}, \quad i = 1,2,\cdots,N \tag{1}$$

式(1)中:在初始状态下,x_i 为出行链活动序列的初始状态变量,为出行活动次数,i 表示总数 N 中的第 i 个出行活动;π_i 为第 i 个出行活动在初始状态下的概率,$\boldsymbol{\pi}$ 是初始概率的一个向量;两个连续马尔可夫过程间的转移概率矩阵如式(2)所示:

$$\boldsymbol{A} = \{a_{ij}\} = \{P(x_t = j \mid x_{t-1} = i)\}, \quad i,j = 1,2,\cdots,N \tag{2}$$

式(2)中:x_t 为出行链活动序列中的第 t 个状态,a_{ij} 为在第 $t-1$ 个状态下活动 i 变为第 t 个状态下活动 j 的转移概率。在出行链中,出行目的 m_t 为隐藏状态,从隐藏状态变为可观察状态的输出概率为 g_{xik}。因可观察变量 o_t 为连续变量,将其分成 K 类可观察状态。在此,假设状态 k 对应的可观察参数服从均值、方差分别为 μ_k、σ_k 的高斯分布,其分布矩阵见式(3):

$$\boldsymbol{G} = \{g_{ik}\} = \{P(m_t = k \mid x_t = i)\}, \quad i = 1,2,\cdots,N,\ k = 1,2,\cdots,K \tag{3}$$

对于活动序列中出行目的为 i 的第 t 个出行,其观察参数为 o_t 的概率可由式(4)表示:

$$P_i(o_t) = \sum_{k=1}^{K} g_{ik} f(o_t \mid m_{ik}, s_{ik}) \tag{4}$$

式(4)中:$\mu_{ik} = \mu_k$,$\sigma_{ik} = \sigma_k$,$i = 1,2,\cdots,N$,$k = 1,2,\cdots,K$,由以上向量集合组成的可观测状态参数集合由式(5)表示:

$$\begin{aligned}\hat{L}(\lambda) &= P(o_1,o_2,\cdots,o_t \mid \lambda) = \\ &P(o_1,o_2,\cdots,o_t \mid [\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{G},\{\mu_k\},\{\sigma_k\}]) = \\ &\prod_{l=1}^{M}\Big\{\sum_{x_1,x_2,\cdots,x_{T^l}}\prod_{t=1}^{T^l}\Big[\sum_{k=1}^{K} g_{x_t k} f(o_t \mid \mu_k,\sigma_k)\Big]\times \\ &\Big(\prod_{t=1}^{T^l} a_{x_{t-1}x_t}\Big)\Big\}\end{aligned} \tag{5}$$

式(5)中:$\hat{L}$ 为该模型的最大似然函数;M 为出行链的条数;$a_{x_{t-1}x_t}$ 为状态间的转移概率;T^l 为第 l 个出行活动链中活动的个数,将向量集 $\boldsymbol{\lambda} = [\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{G},\{\mu_k\},\{\sigma_k\}]$ 进行优化,取得最大值。

1.2 模型求解

采用前向后向算法可对可观测状态对应的出行链概率进行计算,在得到概率后,可采用 Baum-Welch 算法对隐马尔可夫模型进行参数估计,继而通过 Viterbi 算法根据参数估计结果完成出行链提取。具体推导过程参见参考文献[11]。

2 数据处理

2.1 数据结构

研究数据来源为石家庄市公交 IC 卡及公交 GPS 数据,其中 IC 卡基础数据为数据卡 ID 编号、刷卡时间、车辆号、车辆线路号等,GPS 数据包括车辆终端号(车辆号)、GPS 系统时间、经纬度数据、GPS 接收时间等;同时 IC 卡基础数据中的车辆号对应 GPS 数据中的车辆终端号。只选取研究所需数据,

数据结构如表1所示。

表1　研究数据描述

Table 1　Data description for the study

字段名称	数据类型	字段描述
CARD_NO	字符	用户卡号
RIDING_TIME	字符	交易时间
LINE_NO	字符	线路号
BUS_NO	字符	车辆号
UPDOWN	字符	车辆运行方向(上下行)
PACK_DATETIME	字符	进出站时间(由进出站类型判断)
BUSSTOP_NO	字符	站点编号
BUSSTOP_NAME	字符	站点名称
BUSSTOP_LNG	数字	站点经度
BUSSTOP_LAT	数字	站点纬度

2.2　基于时间数据匹配的上车站点识别

对于上车站点的识别,中国学者有较多的研究[12-13]。利用基于时间数据匹配的识别方法:将IC卡刷卡数据中的交易时间与GPS系统中的进出站时间相比对,若打卡交易时间在车辆进出站的时间范围内,则认为该时刻车辆所处站点为乘客的上车站点。即对于任意刷卡记录,其交易时间 t_{ri} 与车辆的进出站时间 t_{aj}、t_{dj} 满足

$$t_{aj} \leqslant t_{ri} \leqslant t_{dj} \tag{6}$$

则在此进出站时间时车辆所对应的站点为乘客的上车站点。

3　模型验证与分析

3.1　模型验证

采用2018年某一周工作日石家庄智能公交卡数据对模型进行实现,并在同一时期同区域公交站点附近进行了居民出行调查,调查所得数据用以与模型识别结果进行对比分析。

按照前述模型,根据乘车时间,结合乘车站点周边土地利用情况进行算法优化,对可观测状态进行聚类,聚类结果如表2所示。可观测状态的六个状态类分别对应五个不同的活动类,分别为通勤(W)、外出办公(B)、购物(S)、回家(H)、其他(O),对应关系见表3。可以看出以通勤和回家为目的的出行活动具有很强的时效性,而其他类型的活动出行时间分布则较为广泛。

将模型识别结果对乘客出行目的进行统计,并与实际调查结果进行对比,对比情况见表4。

根据对比结果,模型对以通勤和回家为目的的活动识别结果与调查结果吻合度较高,误差率分别为12%、6%,而对其他类型的活动识别结果则偏差略大,这也表明在石家庄公交车出行是居民较为主要的通勤方式,而在进行其他活动时人们往往更倾向于选择其他更为便捷交通方式,而这可能也是导致模型对其他类型出行活动识别结果不够准确的因素之一。

表2　可观测状态分类结果

Table 2　Classification results of observable states

	平均上车时间(标准差)	站点类型					
		典型居住区	典型办公区	居住主导区	办公主导区	交通枢纽区	其他
状态1	7:18(0:48)	0.681	0.003	0.192	0.042	0.081	0.001
状态2	9:27(1:23)	0.156	0.471	0.131	0.094	0.105	0.043
状态3	12:06(0:43)	0.001	0.699	0.016	0.201	0.065	0.018
状态4	13:43(0:39)	0.577	0.129	0.144	0.082	0.059	0.009
状态5	17:52(1:17)	0.087	0.681	0.026	0.183	0.021	0.002
状态6	19:41(1:34)	0.135	0.624	0.043	0.187	0.004	0.007

表3　可观测状态与出行目的对应结果

Table 3　Correspondence between observable status and travel purpose

状态	通勤	外出办公	购物	回家	其他
状态1	0.702	0.062	0.297	0	0.228
状态2	0.287	0.595	0.621	0	0.163
状态3	0	0.151	0.001	0.385	0.187
状态4	0	0.192	0.016	0	0.241
状态5	0.009	0	0.033	0.422	0.129
状态6	0.002	0	0.032	0.193	0.052

表4　模型识别结果与调查结果对比

Table 4　Comparison of model recognition results with survey results

出行目的	客流量占比	
	模型识别结果	调查结果
通勤	0.439	0.391
外出办公	0.116	0.082
购物	0.023	0.049
回家	0.387	0.411
其他	0.035	0.067

3.2　结果分析

根据模型识别结果,各活动之间的转移概率如表5所示。活动间的转移概率可在一定程度上反映居民出行链中活动与活动间的相互关系。可以看出,乘客通勤之后最可能的行程是回家;外出办公之后行程为回家的居多,返回工作地的次之;购物之后回家的可能性最高,而在基于家的活动之后通勤出行概率最大;在进行其他类型的活动之后往往会继续进行该类活动。各类型出行链占比如表6所示。由表6可知,通勤活动是公交出行链中最主要的部分。

表 5 活动间转移概率

Table 5 Transition probability of activities

	通勤	外出办公	购物	回家	其他
通勤	0	0.225	0.089	0.681	0.005
外出办公	0.311	0.167	0.003	0.496	0.023
购物	0	0	0	0.922	0.078
回家	0.813	0.006	0.015	0	0.166
其他	0	0	0.006	0.243	0.751

表 6 各类出行链占比

Table 6 Proportion of various travel chains

活动类型	出行链结构	比例/%
通勤类	H-W-H	54.72
	H-W-B-H	17.37
	H-W-B-W-H	9.78
购物类	H-S-H	8.30
	H-S-O-H	6.38
其他类	H-O-H	3.05
	H-O-O-H	0.40

4 结论

利用公交 IC 卡数据,基于乘客乘车时间及乘车站点周边土地类型,构造隐马尔可夫模型对其出行目的进行识别,进而提取公交车乘客出行链,从出行链角度对乘客出行特征进行研究。得出如下结论。

(1)利用乘客公交 IC 卡数据,结合站点周边土地类型构建隐马尔可夫模型,对乘客出行目的进行识别,进而提取完整出行链以研究出行特征。模型识别结果与实际调查结果相比吻合度较好。

(2)在工作日,公交车乘客以通勤为目的的出行最多,其他类型的活动选择公交以外的出行方式的可能性更大。

(3)公交车乘客的通勤出行具有很强的时效性,其他类型的活动出行时间分布则无固定时间。

参 考 文 献

1 Zhao J H. The planning and analysis implications of automated data collection systems: rail transit OD matrix inference and path choice modeling examples[D]. Cambridge: Massachusetts Institute of Technology, 2004.

2 Alex C. Bus passenger origin-destination matrix estimation using automated data collection systems[D]. Cambridge: Massachusetts Institute of Technology, 2006.

3 Farzin J M. Constructing an automated bus origin-destination matrix using farecard and global positioning system data in Sao Paulo, Brazil [J]. Transportation Research Record: Journal of the Transportation Research Board, 2008(2072):30-37.

4 Barry J, Freimer R, Slavin H. Use of entry-only automatic fare collection data to estimate linked transit trips in New York City[J]. Transportation Research Record Journal of the Transportation Research Board, 2009(2112):53-61.

5 胡郁葱, 梁杰荣, 梁枫明. 基于 IC 卡数据挖掘获取公交 OD 矩阵的方法[J]. 交通信息与安全, 2012, 30(4): 66-70.
Hu Yucong, Liang Jierong, Liang Fengming. A way to get bus regional OD matrix based on mining IC card information[J]. Journal of Transport Information and Safety, 2012, 30(4): 66-70

6 陈峥嵘. 智能公共交通系统数据分析方法与应用研究[D]. 南京: 东南大学, 2012.
Chen Zhengrong. Research on data analysis method and application of intelligent public transport system[D]. Nanjing: Southeast University, 2012.

7 胡继华, 邓 俊, 黄 泽. 结合出行链的公交 IC 卡乘客下车站点判断概率模型[J]. 交通运输系统工程与信息, 2014, 14(2): 62-67.
Hu Jihua, Deng Jun, Huang Ze. Trip-chain based probability model for identifying aligting stations of smart card passengers[J]. Journal of Transportation Systems Engineering and Information Technology, 2014, 14(2): 62-67.

8 吴祥国. 基于公交 IC 卡和 GPS 数据的居民公交出行 OD 矩阵推导与应用[D]. 济南: 山东大学,2011.
Wu Xiangguo. Derivation and application of OD matrix of resident bus trip based on bus IC card and GPS data[D]. Jinan: Shandong University, 2011.

9 翁剑成, 王 昌, 王月玥, 等. 基于个体出行数据的公共交通出行链提取方法[J]. 交通运输系统工程与信息, 2017, 17(3): 67-73.
Weng Jiancheng, Wang Chang, Wang Yueyue, et al. Extraction method of public transit trip chains based on the individual riders' data[J]. Journal of Transportation Systems Engineering and Information Technology, 2017, 17(3): 67-73.

10 朱亚迪, 陈 峰, 王子甲, 等. 基于概率图模型的乘客出行链提取方法[J]. 吉林大学学报(工学版), 2019, 49(1): 60-65.
Zhu Yadi, Chen Feng, Wang Zijia, et al. Extraction method of passenger travel chain based on probability map model[J]. Journal of Jilin University (Engineering and Technology Edition), 2019, 49(1): 60-65.

11 Han G, Sohn K. Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model [J]. Transportation Research Part B Methodological, 2016, 83: 121-135.

12 杨万波, 王 昊, 叶晓飞, 等. 基于 GPS 和 IC 卡数据的公交出行 OD 推算方法[J]. 重庆交通大学学报(自然科学版), 2015, 34(3): 117-121.
Yang Wanbo, Wang Hao, Ye Xiaofei, et al. OD matrix inference for urban public transportation trip based on GPS and IC card data [J]. Journal of Chongqing Jiaotong University (Natural Sicence), 2015, 34(3): 117-121.

13 李海波, 陈学武, 陈峥嵘. 基于公交 IC 卡和 AVL 数据的客流 OD 推导方法[J]. 交通信息与安全, 2015, 33(6): 33-39, 95.
Li Haibo, Chen Xuewu, Chen Zhengrong. A method for estimating origin-destination matrix of public transit based on smart card and AVL data[J]. Journal of Transport Information and Safety, 2015, 33(6): 33-39, 95.