基于地铁售检票系统刷卡数据的乘客出行模式分析*

项 煜[1] 陈晓旭[2**] 杨 超[2,3] 段红勇[1]

(1. 河南省交通一卡通有限责任公司,450018,郑州;2. 同济大学道路与交通工程教育部重点实验室,201804,上海;
3. 同济大学城市交通研究院,200092,上海//第一作者,高级工程师)

摘 要 地铁自动售检票系统可以采集大量乘客刷卡数据,可提供更全面的地铁乘客时空信息。对乘客的出行模式分析有利于城市轨道交通运营企业预测地铁客流和制定运营策略。提出了分析地铁乘客出行模式的数据挖掘方法:对地铁刷卡数据进行预处理,根据其时空信息生成乘客出行链;分析反映乘客时空特性的聚类变量;利用 K-means 聚类算法对各聚类变量进行乘客聚类;分析潜在的乘客出行模式。以深圳地铁刷卡数据为例,对提出的地铁乘客出行模式分析方法进行了试验验证。

关键词 城市轨道交通;刷卡数据;乘客出行模式;K-means 聚类算法

中图分类号 F530.7

DOI:10.16037/j.1007-869x.2020.06.015

Analysis of Passengers Travel Patterns Based on Subway Automatic Fase Collection System Smart Card Data

XIANG Yu, CHEN Xiaoxu, YANG Chao, DUAN Hongyong

Abstract The subway Automatic Fare Collection systems collect tremendous amount of smart card data, which provides comprehensive spatial-temporal information about subway passengers. The analysis of passengers travel patterns benefits urban rail transit operation companies in predicting subway passenger flow and formulating operational strategies. A data-mining procedure to identify travel patterns of subway passengers was introduced: after pretreatment of smart card data, passengers travel chains were generated based on the spatial-temporal information of it; clustering variables that reflect spatial-temporal characteristics of passengers were analyzed; K-means clustering algorithm was adopted to cluster the passengers; the potential passengers travel patterns were then analyzed. Taking the Shenzhen subway smart card data as example, verification experiment was conducted on the proposed subway passengers traveling patterns analysis methodology.

Key words urban rail transit; smart card data; passenger travel patterns; K-means clustering algorithm

First-author's address Henan Transport Card Co., Ltd., 450018, Zhengzhou, China

0 引言

随着地铁 IC 卡在城市居民中的广泛使用,在地铁运营过程中其售检票系统产生了大量 IC 卡数据。通过 IC 卡数据可以研究乘客的出行模式。对乘客出行模式分析有利于城市轨道交通运营企业预测地铁客流和制定运营策略。

关于乘客出行模式的研究,文献[1]总结了关于刷卡数据的有关研究,得出的结论是,在个体层面上分析出行模式是一个非常具有挑战性的研究。文献[2-3]研究认为,大多利用刷卡数据进行乘客出行行为的研究都集中在对个体连续位置的预测上。文献[4-5]研究认为,使用智能刷卡数据分析地铁乘客的行为有助于降低运营成本和管理需求。文献[6]利用芝加哥交通管理局的智能刷卡数据分析乘客的公共交通使用情况和访问距离,分析结果可以为交通规划和市场研究提供有用的信息。文献[7]对加拿大运输局的刷卡数据进行了试验验证后,发现了公共交通用户的 4 种主要出行模式。文献[8]定义规律出行乘客为在典型的工作日出行两次或更多次的出行者,并发现公交卡种类可以影响出行模式。文献[9]以澳大利亚 BRT(快速公交)系统为例,将地理可视化方法应用于大型智能刷卡数据库,检查时空动态,并比较 BRT 乘客出行和其他公交出行的空间轨迹及其变化方式。文献[10]对新加坡交通系统进行研究,并提出利用多天智能刷卡数据分析乘客个体和集聚变异性的方法。文献

*河南省交通运输科技计划项目(2019G-2-2);中央高校基本科研业务费专项资金项目(22120180241);**通信作者

[11]使用地铁和公交刷卡数据来识别乘客的空间和时间出行模式的规律性。文献[12]建立了联合熵方法和修正的马尔可夫链模型,利用深圳地铁刷卡数据识别个体出行模式并预测乘客的出行规律。

地铁乘客的出行行为特征还需要进一步研究。本文提出了分析地铁乘客出行模式的数据挖掘方法。对地铁刷卡数据进行预处理,根据其时空信息生成乘客出行链,接着分析反映乘客时空特性的聚类变量,进而利用 K-means 聚类算法(以下简称"K-means")对上述聚类变量进行乘客聚类,来分析潜在的出行模式。本次研究基于深圳市 2014 年 5 月 12—16 日地铁刷卡数据进行研究,分析工作日乘客的出行模式。

1 地铁刷卡数据预处理

本次研究使用的数据来自深圳地铁,包括地铁 IC 卡数据(2014 年 5 月 12—16 日)。深圳地铁线网由 5 条线路和 118 个站点组成,截至 2014 年底,运营里程为 178 km,年度客流量为 8.6 亿人次。地铁 IC 卡数据格式如表 1 所示。

表 1 深圳地铁 IC 卡数据格式

列名	数据含义	数据示例
Card Id	IC 卡的用户 ID	280045973
Timestamp	用户刷卡日期和时间	20140131221857
IN_OUT	乘客进出站标识	IN
Line	地铁线路名称	5
Station	地铁站名称	深圳北站
Fee	结算费用	15
Equipment	AFC(自动售检票)设备号	268004129

首先对地铁数据中的异常数据进行处理,异常数据包括完全相同的记录、刷卡时间字段有缺失的记录。清理前数据的共有 22 427 293 条记录,有 133 369 条异常数据在数据处理过程中被删除。

数据清理完成后,进行乘客出行链生成。乘客出行链定义为出行者每天进行的一系列出行,这是很有效的分析乘客出行模式的方法[14]。在这项研究中,一周五天工作日的数据用于构建乘客的出行链。在乘客出行链生成的过程中,可能由于缺少记录,无法构建一些出行。例如,有些乘客只在入口处有记录或者只在出口处有记录,这样的出行被舍弃。乘客出行链样例如表 2 所示。

表 2 深圳地铁乘客的出行链样例

Card ID	Timestamp	IN_OUT	Station
321922453	20140512083145	IN	益田站
321922453	20140512084350	OUT	福田站
321922453	20140512175515	IN	福田站
321922453	20140512180809	OUT	益田站

2 乘客出行模式分析方法

本研究的重点是挖掘乘客出行规律和出行模式。2 种主要挖掘方法为:通过对乘客出行行为的分析,定义了若干聚类变量来反映时空变异性和活动模式;利用 K-means 基于上述聚类变量对乘客进行聚类,以识别潜在的乘客出行模式。

2.1 聚类变量

为了使用聚类方法进行乘客出行模式的识别,需要构造能够反映乘客出行特征的聚类变量。智能刷卡数据中含有乘客的出行时空数据,可以提取一些特定变量来表征乘客的出行行为,以此构建的聚类变量应该能够实现区别乘客出行模式。本研究选取的聚类变量如下:

1)不同起点/终点占比。是指在一段时间内,乘客出行的起点或终点的差异程度,可以反映乘客出行的空间分布特征。在一周内的工作日中,如果乘客每天第一次出行的起点都相同,或者最后一次出行的终点都相同,那么该乘客被推断为通勤者。该变量是空间出行差异性的指标,可用于推断乘客出行的规律性。例如,乘客一周内出行了 5 天,不同起点占比含义为:占比 0,表示每天第一次出行的起点都相同,意味着该乘客的出行具有很强的空间规律性;占比 1/5,表示其中一天第一次出行起点和其它天不同;占比 2/5,表示其中两天第一次出行起点和其它天不同;占比 1,表示所有天第一次出行的起点都不相同,意味着该乘客的出行非常不规律。通过同样的方法,可以计算不同终点占比。不同起点/终点占比越小,乘客出行空间规律性越强。

2)行程时间。出行行程时间会影响乘客的出行方式选择。居住地和工作地距离较近的乘客偏向于乘坐地铁,而另外的乘客倾向于其它出行方式。关于行程时间的变量,本次研究选择乘客出行的最大行程时间、平均行程时间、最小行程时间作为聚类变量。

3）出行频率。乘客出行频率与乘客的出行模式密切相关，可以反映出行的规律性。针对出行频率，选择乘客日均出行次数和每周出行天数作为聚类变量。

4）出行开始/结束时间。乘客出行的开始时间和结束时间可以反映乘客的出行时间特性。本次研究选取乘客当天第一次出行的开始时间和当天最后一次出行的结束时间作为反映时间分布特性的聚类变量。

2.2　聚类方法

聚类是把相似的对象通过静态分类的方法分成不同的组别或者更多的子集，让在同一个子集中的成员对象都有相似的一些属性。聚类方法包括多种算法，通常分为监督聚类和无监督聚类。监督聚类方法要求输入数据包含每个组成员的已知信息，而无监督聚类方法旨在对没有标签的数据对象进行分类。根据本研究中的数据形式，数据缺乏已知的标签，应采用无监督的聚类方法来挖掘地铁乘客的出行模式。无监督聚类方法可以基于输入数据的相似性找到出行行为聚类。无监督聚类算法主要分为分层算法和分区算法。部分无监督聚类算法优缺点[15]分析见表 3。

表 3　部分无监督聚类算法的优缺点分析

无监督聚类算法	代表性算法	优点	缺点
分层算法	BIRCH；CURE；Chameleon	容易确定聚类数；无输入参数	时间复杂度高；不适合大规模数据处理；对异常值敏感
划分算法	K-means；CLARANS	可以高效处理大规模数据；运算速度快	需要判断聚类数；初始值的设定对结果有影响

2.3　K-means 聚类算法

如表 3 所示，分层算法稳健性很低，并且对异常值的灵敏度很高。由于将个体分配给群集不是迭代的，分层算法无法调整潜在的错误分类。相比之下，划分算法通过优化局部或全局定义的目标函数来生成观察组，因此在涉及大规模数据集的研究中是被优选采用的。

K-means 作为一种计算有效的方法，适合于该研究。地铁刷卡数据规模较大，并且从数据中构建的聚类变量都是数值型。因此，本次研究选择 K-means 对上述聚类变量进行聚类，从而识别乘客出行模式。然而，K-means 有 2 个缺点：对初始种子的依赖性和难以选择聚类的数量。对于第一个缺点，可以通过重复迭代来调整它以找到最佳结果；对于第二个缺点，可以应用聚类评估的 Silhouette Coefficient（轮廓系数，以下简称为“SC”）来找到最佳簇数。

3　试验验证分析

3.1　聚类变量的分布

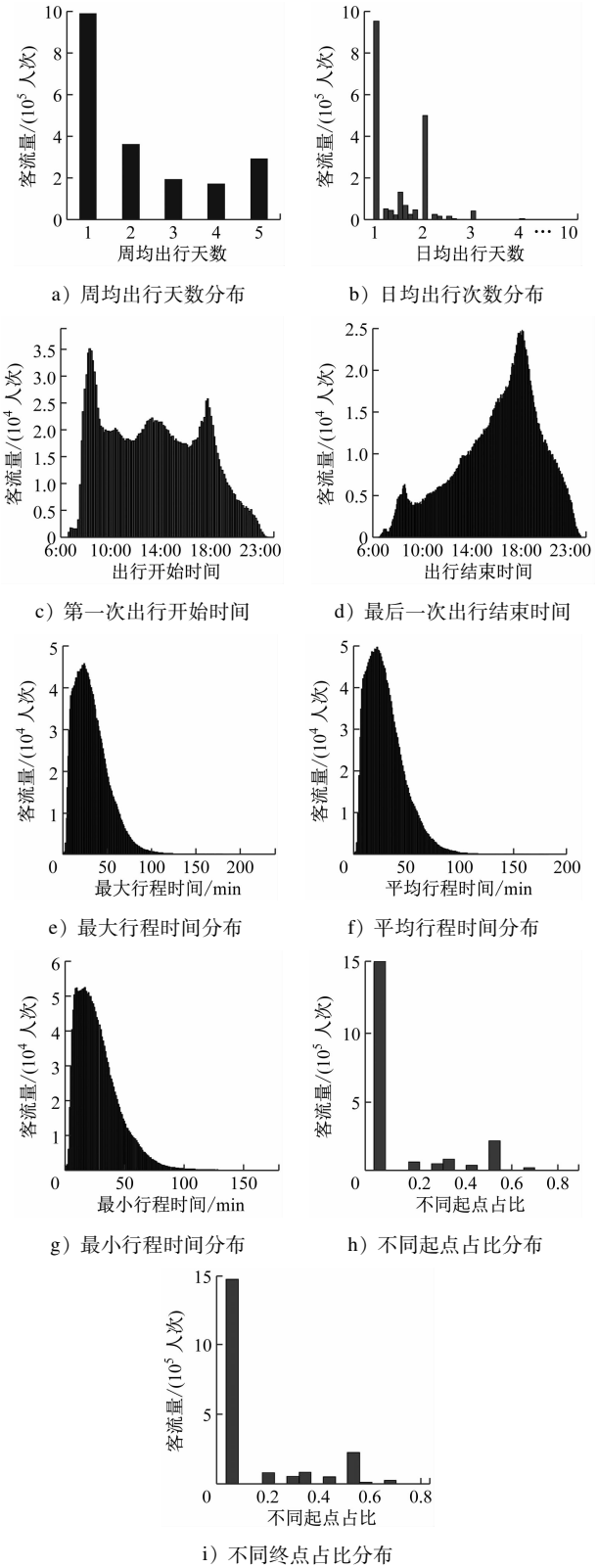
图 1 为一周工作日中所有聚类变量的分布情况。由图 1a）可见，在所有地铁乘客中，在一周内仅只有一天出行的乘客数量是最高的。值得注意的是，一周工作日中有 5 天出行的乘客数量相当可观，这意味着有较多的上班通勤乘客。由图 1 b）可见，很多乘客日均出行 1 次，而日均出行 2 次的乘客也是很多的。日均出行 2 次的乘客可能去上学或者工作，因为学生和工作者大多数是日均 2 次出行，并且大多出行都是从家出发到达目的地，最后从目的地返回到家。由图 1 c）可见，7：00—9：00 存在一个明显的早高峰，13：30—14：30 存在一个午高峰，18：00—19：00 存在一个晚高峰。由图 1 d）是可见，很多乘客在 16：00—20：00 之间完成了他们的最后一次出行。由图 1 e）—图 1 g）可见，大多数乘客在 60 min 内完成出行。由图 1 h）—图 1 i）可见，在不同起点占比和不同终点占比中，占比为 0 的占据高比例。

同时分析了这些聚类变量周末的分布情况。这些聚类变量值的范围有所改变，但其在周末分布的特征与工作日相类似。由此表明，上述聚类变量可以表征这些地铁乘客出行的时间和空间特性。

3.2　乘客出行模式

SC 通过结合内聚度和分离度两种因素，来对聚类效果好坏进行评价。SC 值越大，说明聚类效果越合理。不同聚类数量对应的 SC 的计算结果如图 2 所示。由图 2 可见，聚类数为 5 时的 SC 值最大，因此确定最佳聚类数量为 5。将工作日地铁乘客出行模式分为 5 种，每种出行模式的聚类变量的平均值如表 4 所示。由表 4 可见，出行模式 4 的乘客数占比最高，出行模式 2 的乘客数占比最少。

出行模式 1 的乘客为高频通勤用户，周均出行天数为 3.57 天，日均出行次数为 2.08 次，乘客当天



a）周均出行天数分布　b）日均出行次数分布

c）第一次出行开始时间　d）最后一次出行结束时间

e）最大行程时间分布　f）平均行程时间分布

g）最小行程时间分布　h）不同起点占比分布

i）不同终点占比分布

图 1　一周工作日中聚类变量的分布情况

第一次出行平均开始时刻是 9:25，当天最后一次出行结束时刻为 18:46。通勤用户的行程时间较长，不同起点占比和终点占比的比例为 0.08 和 0.13，

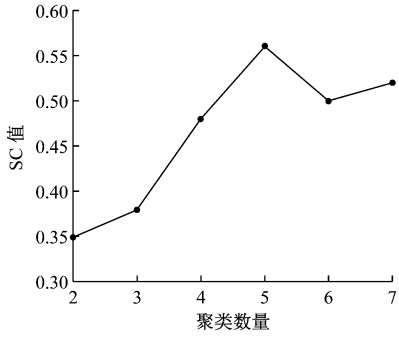


图 2　不同聚类数量下的 SC 值

说明乘客出行的空间位置具有很强的规律性。结合乘客出行时间特征，可以合理推断该类乘客为通勤出行。

出行模式 2 的而乘客为上午出行用户，周均出行天数为 1.99 天，日均出行次数为 1.27 次；大约在 8:20 开始第一次出行，10:30 完成当天最后一次出行；出行的不同起点占比和不同终点占比分别为 0.06 和 0.15。

出行模式 3 的乘客为中午出行用户，周均出行天数为 2.09 天，日均出行次数为 1.41 次；当天第一次出行平均开始时刻是 12:48，当天最后一次出行结束时刻为 14:38；不同起点占比和终点占比的比例为 0.23 和 0.25。总体而言，该类乘客的出行在时间和空间维度上具有很高的规律性。

出行模式 4 的乘客为下午出行用户，周均出行天数为 2.05 天，日均出行次数为 1.43 次；大约在 16:02 开始第一次出行，18:14 完成当天最后一次出行；出行的不同起点占比和不同终点占比分别为 0.19 和 0.17。

出行模式 5 的乘客为低频夜间出行用户，周均出行天数为 1.6 天，日均出行次数为 1.26 次，因此这类乘客低频率乘坐地铁。该类乘客当天第一次出行的平均时刻是 19:30，结束一天的最后一次出行时刻为 20:47。该类乘客的不同起点占比和不同终点占比分别为 0.09 和 0.08，说明该类乘客的起点和终点分别是接近一致的。晚上出行用户的行程时间一般较短。

4　结语

对乘客出行模式进行研究有利于城市轨道交通运营企业预测地铁客流和制定运营策略。本文提出了分析地铁乘客出行模式的数据挖掘方法。首先对地铁刷卡数据进行预处理，根据其时空信息

表 4　一周中工作日各出行模式下聚类变量的平均值

出行模式	周均出行天数	日均出行次数	第一次出行开始时刻	最后一次出行结束时刻	最大行程时间/min	不同起点占比	不同终点占比	最小行程时间/min	平均行程时间/min	人数占比/%
1	3.57	2.08	09:25	18:46	50.91	0.08	0.13	40.07	45.38	18.25
2	1.99	1.27	08:20	10:30	31.99	0.06	0.15	28.51	30.29	16.15
3	2.09	1.41	12:48	14:38	34.01	0.23	0.25	29.33	31.67	20.34
4	2.05	1.43	16:02	18:14	36.34	0.19	0.17	31.28	33.78	25.63
5	1.60	1.26	19:30	20:47	31.76	0.09	0.08	29.03	30.39	19.63

生成乘客出行链,接着分析反映乘客时空特性的聚类变量,进而利用 K-means 对各聚类变量进行乘客聚类,然后分析潜在的出行模式。着重研究了乘客在工作日的出行模式,并分析了相应模式出行特征。研究结果显示,大多数地铁乘客的出行行为具有较高的规律性。本研究结果可以辅助城市轨道交通运营企业深入理解乘客出行行为,从而有针对性地制定差异化的服务措施。在下一步的研究中,要强化聚类变量的构建,因为其仍然是识别地铁乘客出行模式的关键,另外还要对聚类方法进行进一步的探索。

参考文献

[1] PELLETIER M P, TRÉPANIER M, MORENCY C. Smart card data use in public transit: A literature review[J]. Transportation Research Part C: Emerging Technologies, 2011, 19(4): 557.

[2] 陈君, 吕玉坤, 崔美莉. 基于出行模式的公交 IC 卡乘客下车站点判断方法[J]. 西安建筑科技大学学报(自然科学版), 2018, 50(1):23.

[3] 张鹏, 张国武. 城市轨道交通乘客下车时间特性分析与建模[J]. 城市轨道交通研究, 2011(11):80.

[4] MORENCY C, TRÉPANIER M, AGARD B. Analysing the variability of transit users behaviour with smart card data[C]// 2006 IEEE Intelligent Transportation Systems Conference. Toronto:IEEE, 2006: 44.

[5] 鲁放, 韩宝明, 蔡晓春. 城市轨道交通常乘客行为研究[J]. 城市轨道交通研究, 2012, 15(2):39.

[6] UTSUNOMIYA M, ATTANUCCI J, WILSON N. Potential uses of transit smart card registration and transaction data to improve transit planning[J]. Transportation Research Record, 2006(1): 118.

[7] AGARD B, MORENCY C, TRÉPANIER M. Mining public transport user behaviour from smart card data[J]. IFAC Proceedings Volumes, 2006, 39(3): 399.

[8] LEE S, HICKMAN M D. Travel pattern analysis using smart card data of regular users[C]//Proceedings of the 90th Annual Meeting of the Transportation Research Board. Washington DC: the Transportation Research Board ,2011:4.

[9] TAO S, CORCORAN J, MATEO-BABIANO I, et al. Exploring Bus Rapid Transit passenger travel behaviour using big data [J]. Applied geography, 2014, 53: 90.

[10] ZHONG C, MANLEY E, ARISONA S M, et al. Measuring variability of mobility patterns from multiday smart-card data [J]. Journal of Computational Science, 2015, 9: 125.

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

[13] 蔡于. 影响城市轨道交通运营的乘客行为分析[J]. 城市轨道交通研究, 2008(6):47.

[14] MCGUCKIN N, NAKAMOTO Y. Trips, chains, and tours: using an operational definition[C]//National Household Travel Survey: Understanding Our Nation's Travel (NHTS). Washington DC: NHTS, 2004.

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

(收稿日期:2019 - 09 - 12)