计算机科学
COMPUTER SCIENCE

基于隐马尔可夫模型的铁路出行团体关系预测研究

王欣，向明月，李思颖，赵若成

引用本文

王欣，向明月，李思颖，赵若成. 基于隐马尔可夫模型的铁路出行团体关系预测研究[J]. 计算机科学, 2022, 49(6A): 247-255.

WANG Xin, XIANG Ming-yue, LI Si-ying, ZHAO Ruo-cheng. Relation Prediction for Railway Travelling Group Based on Hidden Markov Model[J]. Computer Science, 2022, 49(6A): 247-255.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210500001

基于隐马尔可夫模型的铁路出行团体关系预测研究

王 欣[1] 向明月[2] 李思颖[2] 赵若成[3]
1 西南石油大学计算机科学学院 成都 610500
2 西南交通大学经济管理学院 成都 610031
3 伦敦大学伯贝克学院商业经济和信息学院 伦敦 WC1E 7HX
(xinwang@swpu.edu.cn)

摘 要 近年来,随着铁路交通网络和高铁技术的不断发展,铁路出行的快捷性和舒适性得到了大幅度提高,铁路出行被更多人选择,团队出行也变得更加普遍。旅客的出行行为通常会受同行旅客的影响,不同的出行团体有不同的出行偏好,如家庭团体出行时会考虑团体中的老人和小孩,更在意舒适度;年轻人组成的团体出行时会着重考虑体验感和新鲜感。因此,出行团体类型是研究该团体出行偏好的基础。基于此,文中提出了一种利用客票数据对铁路出行团体同行关系进行预测的方法。首先,基于铁路客票数据特点,提出了铁路出行团体同行次数的量化方法;然后,对隐马尔可夫模型在客票数据分析中的适用性进行了剖析,对基于隐马尔可夫模型的铁路出行团体关系预测问题进行了形式化定义。基于真实铁路购票数据,对构建的出行团体关系模型的预测准确性以及预测结果的一致性进行了验证,实验结果显示构建的模型的预测准确率高达96.38%,对于同一出行团体在不同时刻的预测结果的一致性达95%,由此认为所提方法能够高效且准确地预测铁路出行团体中的同行关系。

关键词:同行关系预测;铁路出行团体;隐马尔可夫模型
中图法分类号 F532.8

Relation Prediction for Railway Travelling Group Based on Hidden Markov Model

WANG Xin[1],XIANG Ming-yue[2],LI Si-ying[2] and ZHAO Ruo-cheng[3]
1 Southwest Petroleum University,School of Computer Science,Chengdu 610500,China
2 Southwest Jiaotong University,School of Economics and Management,Chengdu 610031,China
3 Birkbeck,University of London,School of Business,Economics and Informatics,London WC1E 7HX,UK

Abstract In recent years,with the continuous development of transportation network as well as technology in high-speed railway,the speed and comfort of railway travel have been greatly improved,more and more people choose to travel by railway. As a result,co-travel behaviors have become even common in rail trips. The travel behavior of passengers can be influenced by their peers,and different travel groups will present different travel preferences. For example,for a travelling group with family members,the elderly and children will be taken good care of,hence group members are more inclined to pursue comfort during the trip. When a few young people who are mutual friends form a travelling group,they care more about the sense of experience and freshness. Therefore,predicting the type of a travel group will be beneficial for learning travel preference of this group,e.g.,not only help transportation,tourism and other related industries to define their products and services that travel groups interest in,but also provide support for market decision-making in the railway transportation industry. Based onthis,this paper proposes a methodology for analyzing railway passengers' travelling behavioral using ticket booking data. Firstly,based on ticket booking data,it proposes the quantitative method of co-travel times of a travelling group. Secondly,it formalizes the prediction problem by incorporating Hidden Markov Model. Lastly,the accuracy and consistency of the model are verified with real-life data and experiment results show that the accuracy of our model can even reach 96.38%,in the meanwhile,the consistency is as high as 95%. Thus,we conclude that the proposed method can effectively and accurately predict the relationship of railway travel groups.

Keywords Co-travel relation prediction,Railway co-travel group,HMM

1 引言

交通运输业是经济建设的重要纽带。近年来随着高速铁路运输网络的不断扩展、配套设施的不断完善、出行速度的不断提升,铁路已成为人们出行的重要交通工具之一。为了提升服务质量,提高管理效率,铁路信息化系统得到了全面使用,并积累了海量的购票数据。这些数据蕴含着丰富的商业价值,如何挖掘出这些价值成为了当前众多学者关注的焦点。

目前对旅客出行行为的研究主要针对单一旅客。然而,随着人们与同事、朋友或家人共同出行行为的常态化,传统的面向单一个体的研究已难以适用。其原因在于:同行乘客对个体的出行行为有着重要影响,共同出行时团体中乘客的人数、乘客间的关系会对出行的目的、时间和交通工具的选择产生重要影响。因此,分析出行成员间的社会关系是研究铁路

通信作者:李思颖(1049810415@qq.com)

团体出行行为的重要内容之一。考虑到出行团体成员间的关系具有时序变化的特点，将隐马尔可夫模型应用于带有时序特征的出行团体序列中，有望实现团体成员间关系的预测，具有重要的理论与应用价值。

本文第2节介绍了研究现状；第3节介绍了出行团体预测模型的前期准备工作；第4节详细介绍了铁路出行团体关系预测模型的定义以及构建方式；第5节利用真实铁路客票数据进行出行团体同行关系预测；最后总结全文并展望未来。

2 相关研究

2.1 旅客出行行为研究

随着铁路运输的迅猛发展，选择铁路出行的乘客人数迅速增加，进而积累了大量的出行数据。如何从原始出行数据中抽取高级语义，进而更好地理解旅客的出行行为引起了人们的高度关注。针对此问题，研究人员提出了多种分析方法。

2.1.1 个体出行行为研究

Li等基于出租车出行目的地的周边环境特征对乘客的下车点进行了语义描述，构造出了带有时序特征的语义框架，并基于该框架对乘客的出行目的进行推断[1]。Chen等提出了一种名为Trip2Vec的模型，该模型提取了人类活动的背景，然后采用K-均值来聚合出行群体并解释出行目的[2]。Wang等提出了多源数据分析模型，用于计算出租车出行的特征参数，从而理解不同目的下的出行行为[3]。Deng等基于GPS出行数据、地理信息系统数据和出行者的社会人口统计特征，构建多个决策树对旅客的出行目的进行预测[4]。Zack等从带有地点标记的社交网络数据中提取出行时间、地点和相关人口统计学信息，利用这些信息来模拟不同生活场景中人们的出行目的[5]。针对智能卡出行数据，Bao等首先利用K-均值方法将自行车共享站点周围的POIs聚合为5种类型，然后构造隐含狄利克雷分布发现隐藏的自行车共享旅行模式和旅行目的[6]。为了预测当前和下一次旅行目的地，Cui等利用神经网络进行特征选择，然后基于贝叶斯网络对出行目的进行建模与分析[7]。

2.1.2 团体出行行为研究

Lin等利用航空购票数据及航空公司的乘客信息系统建立乘客共乘网络图，并据此对乘客进行一系列特征提取，进而设计迭代分类算法，以实现对团体出行目的的预测[8]。Qian等通过客运票务数据对出行乘客进行团体划分，构建了带有时间维度的出行目的推断主题模型，用于对4种典型团体出行目的进行推断[9]。Zhou等观察到旅客在出行时，一条出行路线有多个目的地的情况越来越普遍，即出行决策会受到同行人的影响，于是构建了以家庭为单位的出行选择均衡模型，对早高峰出发时间进行分析[10]。而Jia等则通过早高峰出行模型的研究，探索拥挤收费对家庭出行行为的影响[11]。

2.2 旅客团体类型研究

为了使服务更具个性化，有必要对不同的乘客进行分类。通过观测发现，城市以外的出行一般以团体为主，因此推断出行团体关系类型具有重要意义。对此，Logesh等[12]收集用户在微博上展示的出行照片，挖掘其社会背景和人口统计学信息，对照片团体进行关系分类，并利用模型对用户的下一个出行地点进行预测。类似地，在文献[8]中，为了预测出行团体的出行目的，也采用了团体中个人和整体的人口统计学特征和出行特征对团体关系进行识别。Chen等[13]利用社交平台上照片数据所蕴含的背景语义，融合用户的个人属性以及所处团体的团体关系，提出了贝叶斯学习框架，并对用户进行个性化推荐。Elahe等[14]提出了一种考虑团体类型的群体模式发现方法，即挖掘不同团体类型的出行频繁模式，从而预测不同类型团体的出行地点。Wen等[15]基于民航数据构造了乘客之间的共乘网络，进一步利用分类算法对家庭关系和非家庭关系进行分类，最后采用加权社区检测算法实现家庭关系的发现。但目前对于团体关系的预测大多都是由标记数据预测的，因此有必要有效应用未标记数据进行群体关系预测。

2.3 基于隐马尔可夫模型的出行行为分析

在隐含关系预测问题上，隐马尔可夫模型有着优秀的表现。Erdem等[16]提出了混合隐马尔可夫模型，对获取到的低密度的GPS样本数据在道路网络上进行定位，并进行路径的重组，进而预测最有可能的出行路线。为了识别出社区中潜在的子群体及其演化规律，Ibrahima等[17]利用隐马尔可夫模型对子群体之间的演化关系进行了预测。Deng等[18]为了提高在异构蜂窝网络环境下热点地区切换的性能，利用隐马尔可夫模型对用户行为进行建模，进而对用户移动时间进行预测。Xiong等[19]认为在不同生命周期，人的行为、选择偏好都会产生改变，即出行模式是随着时间动态变化的，为了对离散的出行选择进行研究，构建了异质动态隐马尔可夫模型。

现有文献多为研究旅客出行行为，涉及共乘关系的研究是研究出行行为的中间环节，国内外学者较少单方面针对乘客共乘关系进行研究。但同行乘客之间的关系却是研究团体出行行为的基础之一，因此本选题专门针对乘客共乘关系进行研究。其次，目前有关团体出行的研究，大多是针对单个团体分别进行研究。本选题为了研究增加不同团体之间的关联性，同时对基于单个乘客的多个团体进行预测。

3 铁路出行团体关系预测模型构建准备

3.1 铁路共乘团体定义

在铁路购票订单中普遍存在一个票务信息包含多个出行人的情况。针对此，本文对共乘团体进行定义。

定义1 有相同订单号的乘客归属于同一个出行团体，该团体称为共乘团体。表1列出了3个共乘团体，其中G_i代表共乘团体i，每个共乘团体中乘客的购票订单号是一致的。

表1 共乘团体示例

Table 1 Example of co-travel group

共乘团体编号	订单号	乘客姓名
G_1	E000000001	李阳
	E000000001	张冰
	E000000001	王宇
G_2	E000000002	陈丽
	E000000002	张怡
	E000000002	杨新茹
G_3	E000000003	杜宏彩
	E000000003	杜宏霞
	E000000003	马学

3.2 量化共乘团体出行次数

由于本文以共乘团体为研究对象，因此需要对团体的共乘次数进行量化。本文提出以下共乘团体量化方法。

定义2 某团体中的乘客A和乘客B共同出行的次数记为N_{AB}。

定义3 乘客A和B同行次数N_{AB}除以乘客A的出行次数N_A可以刻画A乘客和同行乘客B同行的可能性大小，且

该可能性用$\frac{N_{AB}}{N_B}$来表示。

量化方法具体如下：共乘团体 G 中两个不同乘客可组成一个乘客对并对应一个同行次数。

Step1　团体 G 中乘客 i 与乘客 j 同行的概率如式(1)所示：

$$P(ij)=\frac{N_{ij}}{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}N_{ij}} \tag{1}$$

Step2　因此在有 n 个乘客的团体中，包含了$\frac{n(n-1)}{2}$个同行次数。由于同行次数反映了乘客与乘客之间同行的可能性，在对共乘团体进行量化时应该避免因某些乘客对同行次数过高导致团体出行量化结果虚高的情况，为此，本文利用团体中总出行次数的平均值$\overline{N_{ij}}$代替团体中乘客对的最高同行次数N_{ij}，即更新N_{ij}：

$$N_{ij}=\overline{N_{ij}}=\frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}N_{ij}}{\frac{1}{2}N(N-1)} \tag{2}$$

Step3　本文选用均值对团体出行次数进行量化，团体出行的量化如式(3)所示：

$$E(G)=\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}N_{ij}P(ij) \tag{3}$$

例 1　在一个共乘团体中有乘客 A，B 和 C，其中 A 与 B 的共同出行数为 7，A 与 C 的共同出行数为 3，B 与 C 的共同出行次数为 2，根据式(1)可以得到该共乘团体中两两乘客的同行概率：$P(AB)=0.583$，$P(BC)=0.167$，$P(AC)=0.2$。

从同行次数中可以看出，乘客 A 与乘客 B 有最大共乘次数，因此更新乘客 A 与乘客 B 的共乘次数为$\overline{N_{AB}}$：

$$\overline{N_{AB}}=\frac{7+3+2}{3}=4$$

该团体的共乘次数量化值 $E(g)$ 为：

$$\overline{N_{AB}}P(AB)+N_{BC}P(BC)+N_{AC}P(AC)=3.416$$

若不利用均值对该团体中的最高同行次数进行更新处理，团体出行次数$E^*(g)$的量化结果为：

$$N_{AB}P(AB)+N_{BC}P(BC)+N_{AC}P(AC)=5.02$$

而最高同行次数经过均值处理后，团体出行次数 $E(g)$ 的量化值为 3.416。可以看出，$E^*(g)$高于$E_{\overline{N_{AB}}}(g)$，避免了N_{AB}过高导致的团体出行量化次数的虚高。

4　基于隐马尔可夫模型的铁路共乘团体关系预测

本文方法是基于隐马尔可夫模型实现的。下文从问题定义、预测方法实现两个方面分别展开介绍。

4.1　共乘团体关系预测问题定义

首先对本文方法所采用的隐马尔可夫模型进行简要介绍。

4.1.1　隐马尔可夫模型的回顾

隐马尔可夫模型(HMM)是一种研究一组观测值和生成这组观测值背后的一系列隐含事件之间因果关系的概率模型。HMM 包含以下组成部分：

(1)$S=s_1,s_2,\cdots,s_N$，包含 N 个隐含状态的集合。

(2)$A=a_{11},a_{ij},\cdots,a_{NN}$，表示状态转移矩阵，其中的每一个$a_{ij}$代表从$s_i$转移到$s_j$的概率，并且对任意的 i 有$\sum_{j=1}^{N}a_{ij}=1$。

(3)$O=o_1,o_2,\cdots,o_T$，包含 T 个可观测值的集合。

(4)$B=b_i(o_t)$，一系列可观测值出现的可能性，也就是发射概率，代表了可观测值o_t产生自隐含状态s_i的概率。

(5)$\pi=\pi_1,\pi_2,\cdots,\pi_N$，表示 N 个隐含状态的初始概率分布。其中，π_i代表马尔可夫链从隐含状态s_i开始的概率。如果π_i为 0，则表示隐含状态不可能出现在一条马尔可夫链的最开始阶段。对任意的 i 仍然有$\sum_{i=1}^{N}\pi_i=1$。

具体来说，隐马尔可夫模型解决的问题可以分为 3 类：

(1)概率问题。给定 HMM 参数 $\lambda=(A,B,\pi)$和观测状态序列 O，得到该观测状态序列出现的概率 $P(O|\lambda)$。

(2)解码问题。给定一组可观测值序列 O 和 HMM 参数 $\lambda=(A,B,\pi)$。反推产生该序列的最有可能的隐含状态序列 Q。

(3)学习问题。给定一组可观测值序列 O，得到模型参数 $\lambda=(A,B,\pi)$。

4.1.2　模型主要问题介绍

本文中主要涉及到的是解码问题和学习问题。

(1)解码问题。解码问题的实现需要利用到维特比算法。维特比算法从解决问题的思路上看类似于选择最优路径，在使用时需要给定可观测值序列以及 HMM 参数 $\lambda=(A,B,\pi)$，得到最有可能的团体关系序列。

根据维特比算法结构图(见图 1)，现有一组观测值序列 $y=(y_1,y_2,y_3,\cdots,y_n)$，隐含状态集合 $x=(x_1,x_2,x_3)$，需要找到从初始观测值y_1经过y_2，y_3等中间观测值到达y_n的最大可能状态路径，也就是产生这一组观测值序列对应的隐含状态集合。y_i的隐含状态可能值为x_{i1}，x_{i2}，x_{i3}(x_{ij}表示第 i 个观测值由隐含状态 j 产生)。

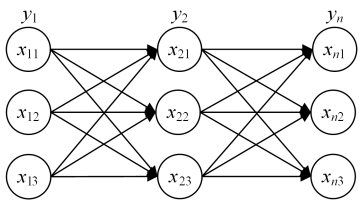


图 1　维特比算法图结构

Fig. 1　Graph structure of Viterbi algorithm

对于最初时刻的观测值y_1，计算其在 3 种不同隐含状态下的产生概率，如式(4)所示：

$$P(y_{1i})=P(y_j\text{由}x_i\text{产生})\times\max_{s=1,2,3}\{x_s|t=0\} \tag{4}$$

其中，i 的取值为 1，2，3，下同。

对于中间观测值y_j，计算其由 $j-1$ 时刻以及当前时刻的隐含状态发出的概率，如式(5)所示：

$$P(y_{ji})=P(y_j\text{由}x_i\text{产生})\times\max_{s=1,2,3}\{P(y_{j-1}\text{由}x_s\text{产生})\times P(x_s\rightarrow x_i)\} \tag{5}$$

(2)学习问题。学习问题的解决需要实现鲍姆-韦尔奇(Baum-Welch)算法。在使用时需要给定一组 HMM 参数 $\lambda=(A,B,\pi)$的初始解和可观测序列。

Baum-Welch 算法的目的为在有缺失值(隐含状态 I 未知)的情况下找到$\bar{\lambda}$(概率模型的参数)，使得一组观测数据 O 出现的概率最大。即：

$$\bar{\lambda}=arg\max_{\lambda}\sum_{d=1}^{D}\sum_{I}(\log P(O,I|\lambda))P(I|O,\lambda)$$

其中，D 为输入的可观测序列的数量。

4.1.3　问题定义

本文的主要目的是识别共乘团体中乘客间的关系类别，如识别一个共乘团体中的成员是家庭关系或者非家庭关系。

然而，直观的观察购票订单并不足以确定共乘团体成员间的关系。分析铁路购票订单可以发现，订单中有明显的时间信息和出行人信息。因此受隐马尔可夫模型的启发，我们可以以单个乘客为纽带，构建带有时序特征的共乘团体序列，对序列中的团体关系进行预测。

结合隐马尔可夫模型可以理解为，通过观测值推断出观测值背后的隐含事件。为此，本文对拟研究的问题做出如下形式化描述：

首先，以单一乘客为中心提取出行序列，即以乘客 P 为中心，对包含该乘客的 m 个共乘团体按照其出行时间的先后进行排序而形成的一条共乘团体序列：

$$Seq(p)=g^1,g^2,g^3,\cdots,g^m \tag{6}$$

其中，$g^i \neq g^j$，g^t 的出发时间早于 g^{t+1} 的出发时间（$t\in[1,m]$）。

针对序列中的每一个共乘团体 g^t，可以从其铁路购票订单中选取一组描述该共乘团体的属性 $F(g^t)$。定义 $L(g^t)$ 为 $F(g^t)$ 所属的类别，因此 $L(g^t)$ 可视为铁路出行场景下的可观测值，一条共乘团体序列即可转换为一条可观测值序列。

通过构造以单个乘客为中心的铁路隐马尔可夫模型，拟研究的问题（简称为 CIP）可描述为：识别出可观测值序列对应的团体关系类别序列。

由分析可知，CIP 的解决对应着隐马尔可夫模型中的解码问题，即给定一条可观测值序列以及 HMM 参数 $\lambda=(A,B,\pi)$，得到最有可能的团体关系序列。

例 3 给出了解决 CIP 后的结果。

例 3　以乘客陈一为中心提取出包含该乘客的所有共乘团体，如表 2 所列，可以看到该乘客一共有 3 次出行记录。对以乘客陈一为中心提取出的所有共乘团体按照其出发时间进行排序，得到陈一的共乘团体序列。然后利用隐马尔可夫模型解决 CIP 问题，得到该出行序列中每一个共乘团体最有可能关系的类别，如图 2 所示。

表 2　共乘团体（以陈一为中心）

Table 2　Co-travel group (take Chen Yi as center)

车次	出发日期	出发地	目的地	订单号	车厢	座位号	乘客身份证号	乘客姓名
D527	2016/06/01	成都	重庆	E00000011	05	07A	51192319740826 ****	陈一
						07B	51192319780207 ****	杨琳
						07C	51192320020618 ****	陈好
D633	2016/07/03	广州	西安	E00000026	02	26A	51192319740826 ****	陈一
						26B	51170119760507 ****	李想
						26C	51172319730609 ****	赵天
D923	2016/10/03	成都	上海	E00000057	03	16A	51192319740826 ****	陈一
						16B	51302719721213 ****	张虹
						16C	51302719700209 ****	陈新
						16D	51172319730609 ****	赵天

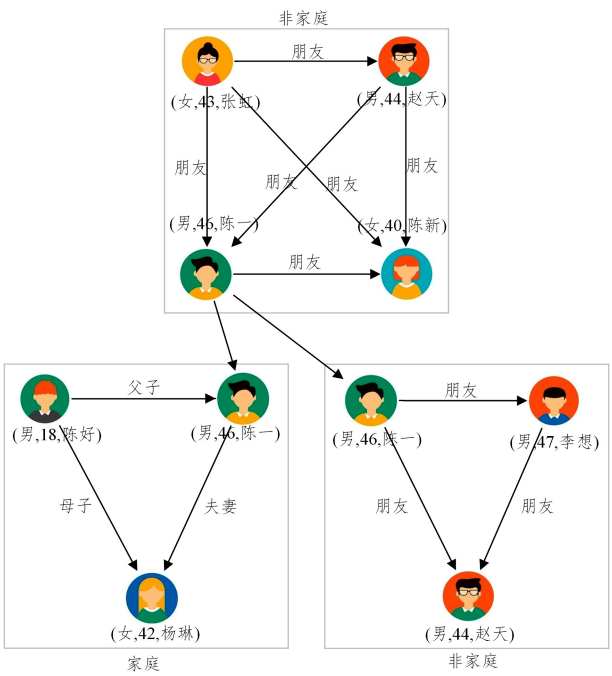


图 2　共乘团体关系预测结果

Fig. 2　Predicting results of co-travel group relations

图 2 中，每个虚线框代表一个共乘团体，框外的标注为该团体的关系类别；框内的节点代表共乘团体中包含的乘客，每个乘客都带有姓名、年龄和性别等属性。直线相连的乘客之间存在共同出行的关系，线上的标注为其关系类别；而双向箭头连接着同一个乘客，该乘客为出行序列的中心。

4.2　HMM 参数学习

4.2.1　参数定义

若要解决 CIP 问题，首先需要解决隐马尔可夫模型参数获取问题，进而需要解决的是隐马尔可夫模型中的学习问题，即鲍母韦尔奇算法。因此本文根据铁路购票订单中包含的信息的特点以及隐马尔可夫模型的需要对铁路场景下的隐马尔可夫模型中的参数做出如下定义。

（1）隐含状态集合 S：隐含状态不能直接观察到，是观测值的发出者。基于本文的目的，设定共乘团体关系包含家庭关系和非家庭关系两种类别，于是隐含状态集合可以表示为 $S=\{G_y,G_n\}$，其中 G_y 和 G_n 分别代表家庭关系和非家庭关系。

（2）观测值状态集合 O：在铁路购票订单中能直接观察到的有：购票信息、个人信息以及团体信息。从以上信息中提取出一组特定属性 $F(g_i)=(f_{gs},f_{gd},f_{ad},WhetherFN_g)$ 作为购票订单的量化表示，其中：

1）f_{gs} 表示共乘团体中包含的乘客人数。若人数为两人，$f_{gs}=2$ 人，否则 $f_{gs}=2$ 人以上。

2）f_{gd} 表示共乘团体中包含的乘客性别。如果该团体包含两种性别，$f_{gd}=1$，否则 $f_{gd}=0$。

3）f_{ad} 表示共乘团体中乘客的最大年龄和最小年龄之间的差。如果该差值不大于 20，$f_{ad}=$没有年龄差，否则 $f_{ad}=$有年龄差。

4）$WhetherFN_g$ 表示共乘团体中乘客是否存在最大姓氏。如果有大于等于半数的顾客拥有一样的姓氏，则 $WhetherFN_g=$有同姓氏，否则 $WhetherFN_g=$无同姓氏。

为了便于表示，本文用类别标记L_i代表第 i 种属性值组合，所有可能出现的属性值集合形成了可观测值状态集合$O=\{L_1,L_2,L_3,\cdots,L_n\}$。

(3)状态转移矩阵 $\boldsymbol{A}$：矩阵 $\boldsymbol{A}$ 中的a_{ij}代表了从s_i转移到s_j的概率，其计算式如式(7)所示：

$$a_{ij}=\frac{P(S_T=s_i,S_{T+1}=s_j)}{P(S_T=s_i)}=\frac{\gamma(s_i,s_j)}{\gamma(s_i)} \tag{7}$$

其中，$S_T(S_{T+1})$代表在 $T(T+1)$时刻的隐含状态为$s_i(s_j)$，$\gamma(s_i,s_j)$代表从s_i转移到s_j的数量，$\gamma(s_i)$代表从隐含状态s_i转移到任意隐含状态的数量。

(4)观测值发射矩阵 $\boldsymbol{B}$：每一类观测值都是由某一种隐含状态s_i发出的，每一类隐含状态产生不同类型观测值有不同的可能性。矩阵 $\boldsymbol{B}$ 中的元素b_{jk}代表了隐含状态s_j产生观测值o_k的概率。

b_{jk}的计算式如式(8)所示：

$$b_{jk}=\frac{P(O_t=L_t|S_i=s_i)}{P(S_i=s_i)}=\frac{\tau(L_t|s_j)}{\tau(s_j)} \tag{8}$$

其中，$O_t=L_t$代表第 t 时刻观测值为L_t类别的属性值组合，$\tau(L_t|s_j)$代表隐含状态s_j产生观测值为L_t类别的属性值组合的数量，$\tau(s_j)$代表了隐含状态s_j产生的观测值总量。

(5)状态初始概率 π。某个状态的初始概率指该状态在 $T=0$ 时刻出现的概率，可以等价表示为 $P(H_1=s)$，$\forall s\in S$，其中H_1为隐含状态序列中的第一个元素。π_{s_i}为隐含状态s_i的样本数量与样本总量的比值，如式(9)所示：

$$\pi_{s_i}=\frac{|s_i|}{|D|},s_i\in S \tag{9}$$

其中，$|s_i|$为隐含状态s_i出现的次数，$|D|$为所有隐含状态的出现次数。

4.2.2　观测值序列定义

算法学习出的模型参数是使得输入数据有最大概率发生的一组参数。为了比较不同输入数据对参数学习效率的影响，本文根据预测序列定义了两种输入序列 Seq。

(1)以单个乘客为中心，提取出包含该乘客的所有共乘团体，形成共乘团体序列，其中共乘团体序列中包含的团体按照各自出行时间的先后顺序进行排序，即：

$$Seq(p)=g^1,g^2,g^3,\cdots,g^m \tag{10}$$

其中，$g^i\neq g^j(1\leqslant i\leqslant j\leqslant m)$。

(2)随机选取起始共乘团体，后续共乘团体与其前一个共乘团体包含至少一个相同乘客，即：

$$lineSeq=g^1,g^2,g^3,\cdots,g^m \tag{11}$$

其中，g^i与$g^{i+1}(1\leqslant i\leqslant i+1\leqslant m)$包含至少一个相同乘客，且$g^i\neq g^j(1\leqslant i\leqslant j\leqslant m)$。

例 4：依旧以表 2 为例，进行整个模型的实现过程展示。表 3 为以陈一为中心的共乘团体序列。为了对每个共乘团体进行描述得到模型的可观测值，需要提取相关属性，因此本例拟定对从每个共乘团体中提取出的属性组合进行量化表示，如式(12)所示：

$$\begin{aligned}G_i&=attribution_i\\&=[size_{G_i},agediff_{G_i},gendernum_{G_i},familyname_{G_i}]\end{aligned} \tag{12}$$

其中，$size_{G_i}$描述共乘团体 i 中包含的乘客数量，$agediff_{G_i}$表示共乘团体 i 中最大年龄与最小年龄之间的差值，$gendernum_{G_i}$为共乘团体中存在的乘客性别，$familyname_{G_i}$表示同姓氏的人数在该共乘团体中人数中的占比(当同姓氏的人数占团体人数的一半及以时其值为 1，其余值为 0)。

表 3　序列共乘团体

Table 3　Co-travel sequence

订单号	姓名	性别	团体编号
E00000011	陈一	男	G_1
	杨琳	女	
	陈好	女	
E00000026	陈一	男	G_2
	李想	女	
	赵天	男	
E00000057	陈一	男	G_3
	张虹	女	
	陈新	男	
	赵天	男	

为了减少冗余，本文对年龄的差值进行了分段处理，如表 4 所列，定义年龄差每增加 20 岁隔代数加一。

表 4　年龄差划分

Table 4　Age difference

年龄差值	阶段定义
[0,20]	无隔代
[20,80]	有隔代

假设模型 $\lambda=(S,O,A,B,\pi)$通过学习算法获取到如下参数值。

隐含的团体关系状态有两种：家庭和非家庭(分别用 F 与 NF 表示，以下同)，隐含状态集合表示为：

$S=\{F,NF\}$

结合表 2 对表 3 中的每个共乘团体进行设定的属性值提取后，设置了 3 种类别的观测值，表 5 列出了每一类观测值及其对应的具体的属性值组合。因此观测值集合可以表示为：

$O=(attribution_1,attribution_2,attribution_3)$

表 5　观测值集合展示

Table 5　Observation set display

观测值类别	观测值
attribution1	(3，男女，有隔代，1)
attribution2	(3，男，无隔代，0)
attribution3	(4，男女，无隔代，1)

初始状态概率分布矩阵：

$$\pi=\begin{pmatrix}F=0.3\\NF=0.7\end{pmatrix}$$

隐含转态转换概率矩阵：

$$A=\begin{pmatrix}a_{F\text{-}F}=0.7 & a_{F\text{-}NF}=0.3\\a_{NF\text{-}F}=0.4 & a_{NF\text{-}NF}=0.6\end{pmatrix}$$

观测值发射矩阵(公式中 $attribution$ 简写为 $attr$)：

$$B=\begin{pmatrix}F_{attr_1}=0.5 & F_{attr_2}=0.2 & F_{attr_3}=0.3\\NF_{attr_1}=0.2 & NF_{attr_2}=0.6 & NF_{attr_3}=0.2\end{pmatrix}$$

对表 3 中的 3 个共乘团体按照设定的属性值进行提取后，以陈 * 为中心乘客对包含其的团体按照团体出行时间排序后得到如图 3 所示的铁路数据中隐马尔可夫模型示例，可观测值序列为：

$O_{set}=(attribution_1, attribution_2, attribution_2)$

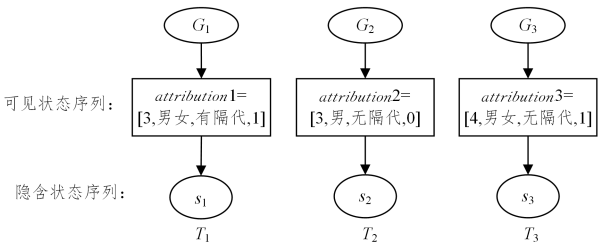


图 3 铁路数据中隐马尔可夫模型示例

Fig. 3 Example of Hidden Markov Model in Railway Data

基于以上给出的模型参数,利用维特比算法解决 CIP,过程如下:

Step1 G_1 的属性值类别为观测值集合中的 $attribution_1$,根据家庭关系和非家庭关系产生这种属性值的概率对 G_1 的团体关系概率进行计算:

$$P(G_1\rightarrow F)=P(attribution_1\mid F)\times P(\pi_F)=0.15$$

$$P(G_1\rightarrow NF)=P(attribution\,1\mid NF)\times P(\pi_{NF})=0.14$$

由结果可知,G_1 为家庭关系的概率大于其为非家庭关系的概率,因此 G_1 更有可能为家庭关系,G_1 的局部最优路径为 F。

Step2 G_2 的属性值类别为观测值集合中的 $attribution_2$,同理对 G_2 的两种团体关系概率进行计算:

$$P(G_2\rightarrow F)=P(attribution_2\mid F)\times\max\{P(G_1\rightarrow F)\times a_{F\text{-}F}, P(G_1\rightarrow NF)\times a_{NF\text{-}F}\}$$
$$=0.0525$$

$$P(G_2\rightarrow NF)=P(attribution_2\mid NF)\times\max\{P(G_1\rightarrow F)\times a_{F\text{-}NF}, P(G_1\rightarrow NF)\times a_{NF\text{-}NF}\}$$
$$=0.0168$$

可以看出,G_2 为家庭概率的可能性更高,到此时 G_2 更有可能为家庭关系,G_2 的局部最优路径为:$F\rightarrow F$。

Step3 G_3 的属性值类别为观测值集合中的 $attribution_3$,同理对两种团体关系概率进行计算:

$$P(G_3\rightarrow F)=P(attribution_3\mid F)\times\max\{P(G_2\rightarrow F)\times a_{F\text{-}F}, P(G_2\rightarrow NF)\times a_{NF\text{-}F}\}$$
$$=0.00735$$

$$P(G_3\rightarrow F)=P(attribution_3\mid NF)\times\max\{P(G_2\rightarrow F)\times a_{F\text{-}NF}, P(G_2\rightarrow NF)\times a_{NF\text{-}NF}\}$$
$$=0.00945$$

G_3 为非家庭概率的可能性更高,因此 G_3 更有可能为非家庭关系。

图 4 给出了 3 个共乘团体回溯的路径,从 G_3 的两种关系中选择可能性更高的一种,即非家庭关系开始回溯寻找最大可能的隐含状态路径。G_3 的非家庭关系计算时选择的是 G_2 的家庭关系,于是 G_2 图的隐含状态为家庭关系,同理找到 G_1 的隐含状态液为家庭关系,于是可观测序列中 3 个共乘团体的团体关系序列最大可能为:家庭-家庭-非家庭。

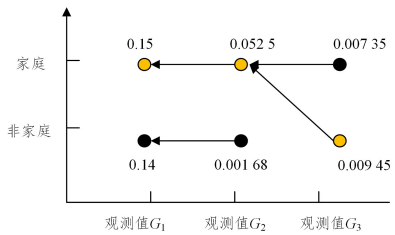


图 4 维特比算法最优路径选择

Fig. 4 Viterbi algorithm optimal path selection

5 算法实验

本节对提出的铁路共乘团体关系预测方法进行了实验验证。首先对真实数据中的共乘团体的团体出行次数进行量化,然后将本文提出的基于隐马尔可夫模型的关系预测模型运用到实际的数据集中,并对其准确性和一致性进行验证。

5.1 实验数据简介

本文的实验数据来自于 XX 铁路局 2015 年 1 月到 2016 年 11 月的铁路乘客购票订单,一共包含了 164 944 条购票信息,涵盖了上万名铁路出行乘客。其中有 42 237 个订单带有团体关系的标记(即有家庭关系和非家庭关系的关系属性)。订单数据中,包含订单号(sequence_no)、车次(train_code)、发车时间(train_date)、始发地(from_station)、目的地(to_station)、车厢号(coach_no)、座位号(seat_no)、乘车人身份证号(id_no)、乘车人姓名(name)、乘车人出生日期(birth)及乘车人性别(sex)共 11 个字段,如表 6 所列。

表 6 原购票数据格式

Table 6 Format of original purchasing data

车次	出发日期	出发地	目的地	车厢	座位号	订单号	身份证号	姓名	出生日期	性别
D0101	2015-03-13	岳池	成都	04	011A	EA00000001	51292319470207 ****	李阳	19900719	女
D0101	2015-03-13	岳池	成都	04	011C	EA00000001	51292319680714 ****	张冰	19550409	女
D0101	2015-03-13	岳池	成都	04	011D	EA00000001	51162119890815 ****	王宇	19860707	女
D2208	2015-04-07	成都东	重庆北	15	017C	EA00000007	51622119860707 ****	刘明娜	19860707	女
D2208	2015-04-07	成都东	重庆北	15	017D	EA00000007	51092119800620 ****	刘明慧	19800620	女
D0505	2015-04-12	重庆	成都	07	0028	EA00000003	23262319600523 ****	杜宏彩	19860707	男
D0505	2015-04-12	重庆	成都	07	0029	EA00000003	23260319800303 ****	杜宏霞	19800620	女
D0505	2015-04-12	重庆	成都	07	2230	EA00000003	23262319580211 ****	马学	19591224	女

本文以表 1 中提取出的共乘团体 G_1,G_3 为例,根据表 3 这两个共乘团体的购票订单以及实验过程所需属性对数据集进行处理,得到实验所需数据。如图 5 所示,共乘团体中的每一条购票订单被划分为 3 个部分:乘客信息、行程信息以及订单信息。乘客信息包含乘客姓名、身份证号码、年龄及性别等个人属性;行程信息包含出发时间、出发地以及目的地等出行属性;订单信息包含购票订单号、车次、车厢号及座位号等票据信息。为保护隐私,本文对乘车人身份证号进行了 md5 加密,且本文例子中出现的所有旅客的个人信息均为虚构。

包含 m 个共乘团体的数据集可以表示为:

$$D=(g_1, g_2, \cdots, g_m) \tag{13}$$

为了对本文方法进行验证,一共选取 4 2237 个团体进行

训练和测试。以单一乘客为中心，提取出所有包含该乘客的不重复团体，并依据其出行时间的先后进行排序，得到 67 308 条出行序列，去掉重复的出行序列（因为每个订单包含多个乘客，在数据集中这些乘客可能再次同行，即被包含在同样的订单中）后仍然有 5510 条出行序列，相关数据统计结果如表 7 所列。

乘客信息

姓名	身份证号	性别	出生日期	共乘团体
李洋	512923196702O7****	女	19470207	
张冰	512923196807l4****	女	19680714	G_1
王宇	511621198908l5****	女	19890815	
杜宏彩	232623196005l3****	女	19600523	
杜宏霞	232603198003O3****	女	19800303	G_3
马学	232623196802l1****	女	19580211	

团体信息

出发时间	出发地	目的地	共乘团体	团体出行次数
2015-3-13	岳池	成都	G_1	E_{G1}
2015-4-12	重庆	成都	G_3	E_{G3}

订单信息

订单编号	车次	出行人数	车厢	共乘团体
E00000001	D0101	3	07	G_1
E00000003	D0505	3	04	G_3

图 5　购票信息拆分示例

Fig. 5　Example of splitting purchase data

表 7　数据统计

Table 7　Statistics of data

数据类型	数量
共乘团体	42 237
以单个乘客为中心的出行序列	67 308
去重后的出行序列	5510

5.2　实验数据拟合

根据团体量化公式，首先对带有关系属性标记的 42 237 个共乘团体进行出行次数的量化，发现共乘团体出行次数分布在 1～24 次之间。量化后的次数分布如表 8 所列，出行次数在一次和两次之间的共乘团体的数量为 24 062，超过了总团体数的 50%，而 3～24 次的总和约小于总共乘团体的 50%。

表 8　团体出行次数统计

Table 8　Statistics of group co-travel times

出行次数区间	[1,2]	[3,24]
包含团体数量	24062	18175

对两个出行次数区间中的共乘团体分别按照其量化后的出行次数从高到低排序。

具有稳定关系的一群人出现在同一个团体中的概率高于相互之间不熟悉的一群人，如具有亲戚关系的一群人一起出现的次数高于突然被安排出差的几个人一起出现的概率。于是本文规定出行次数在[1,2]之间的团体为低量化订单，出行次数在[3,24]之间的团体为高量化订单。

需要从带有关系属性标记的共乘团体中选取训练样本。为了增强样本特征，即家庭与非家庭团体的特征更加明显，本文对高量化订单进行排序，从排好序的高量化订单序列中选择具有家庭关系的出行团体，共提取出 3617 个订单作为家庭关系样本；对低量化订单进行排序，从排好序的低量化订单序列中选择具有非家庭关系的出行团体，共提取出 17 558 个订单。由于与家庭关系训练样本数值相差过大，为了保持样本的平衡性，仅选取前 20%的订单，即 3512 个订单作为非家庭关系样本。

对于这 7129 个团体，利用上文定义的属性值集合 $F(g_i)=(f_{gs},f_{gd},f_{ad},WhetherFN_g)$ 进行表示，共得到 16 类观测值（因为订单编号为团体标识，所以进行团体出行次数量化时去除了订单编号这一属性），同上文对年龄阶段的定义一样，认为乘客之间的年龄差超过 20 岁则乘客之间存在隔代，否则认为是同代属性值类别，如表 9 所列。

表 9　属性值类别

Table 9　Category of property value

类别	属性值集合	类别	属性值集合
L_1	(2 人,1,没有年龄差,无同姓氏)	L_9	(2 人以上,1,有年龄差,无同姓氏)
L_2	(2 人,1,没有年龄差,有同姓氏)	L_{10}	(2 人以上,1,有年龄差,有同姓氏)
L_3	(2 人,1,有年龄差,无同姓氏)	L_{11}	(2 人以上,0,没有年龄差,有同姓氏)
L_4	(2 人,1,有年龄差,有同姓氏)	L_{12}	(2 人以上,0,有年龄差,无同姓氏)
L_5	(2 人,0,没有年龄差,有同姓氏)	L_{13}	(2 人以上,0,有年龄差,有同姓氏)
L_6	(2 人,0,有年龄差,无同姓氏)	L_{14}	(2 人以上,1,没有年龄差,无同姓氏)
L_7	(2 人,0,有年龄差,有同姓氏)	L_{15}	(2 人以上,0,没有年龄差,无同姓氏)
L_8	(2 人以上,1,没有年龄差,有同姓氏)	L_{16}	(2 人,0,没有年龄差,无同姓氏)

以团体中单个乘客为中心，在这 7129 个共乘团体中提取出包含该乘客的所有共乘团体，并按照其各自的出发时间对提取出的团体进行排序，剔除掉仅包含一个共乘团体的序列，最后得到 930 条以单个乘客为中心按照团体出行时间排序的出行序列。本文选取 500 条出行序列作为训练样本数据，430 条作为测试集数据。根据上文定义的初始解中每个参数的计算公式，得到以下几种参数：

(1)两种团体关系的初始概率 π。在 500 条训练样本中共包含了 1420 个共乘团体，其中具有家庭关系的共乘团体有1211 个，非家庭关系的共乘团体为 209 个，分别计算其在总团体数中所占的比例，得到 π_{G_y} 和 π_{G_n}，结果如表 10 所列。

表 10　初始概率

Table 10　Initial probability

π_{G_y}	π_{G_n}
0.8529	0.1471

(2)团体关系转化概率 A。在 500 条序列中，统计序列中从家庭类别G_y（非家庭类别G_n）分别转移至非家庭类别G_y和家庭类别G_n的总量，计算其占G_y（G_n）总转移量的比例，结果如表 11 所列。

表 11　转移概率

Table 11　Probability transition

	G_y	G_n
G_y	0.99720	0.00280
G_n	0.03279	0.96703

(3)两种团体关系产生属性值类别L_i的概率$\boldsymbol{B}$。经统计后发现,在样本数据中一共出现了16种类型的属性值组合,于是观测值集合为$O=\{L_1,L_2,L_3,\cdots,L_{16}\}$。为了对这16种观测值在不同关系下出现的概率$P(L_i|G_{s_i})$进行计算,首先分别统计这16种观测值在1211个家庭关系团体与209个非家庭关系团体中出现的次数,然后分别计算其在两种关系中所占的比例,于是得到观测值发射矩阵$\boldsymbol{B}$,结果如表12所列。

表12 观测值发射概率

Table 12 Emission probability of observation value

L_i	$P(L_i\|G_y)$	L_i	$P(L_i\|G_n)$
L_1	0.2632	L_1	0
L_2	0.0016	L_2	0
L_3	0.0111	L_3	0
L_4	0.0037	L_4	0
L_5	0.0043	L_5	0
L_6	0.0233	L_6	0
L_7	0.0032	L_7	0
L_8	0.1485	L_8	0.0797
L_9	0.2456	L_9	0.0664
L_{10}	0.2574	L_{10}	0.0019
L_{11}	0.0074	L_{11}	0.0493
L_{12}	0.0111	L_{12}	0.0474
L_{13}	0.0196	L_{13}	0.0190
L_{14}	0	L_{14}	0.4478
L_{15}	0	L_{15}	0.2846
L_{16}	0	L_{16}	0.0039

5.3 模型预测准确度实验

本文基于上述求得的初始解,利用鲍母-韦尔奇算法对模型参数进行学习,利用学习得到的参数,用维特比算法进行预测。对于某一条出行序列中所包含的共乘团体,如果维特比算法的预测结果与订单所标注的团体关系一样,则认为预测正确,并标记为TR;否则这个预测结果被认为错误,且标记为FR。

鲍母-韦尔奇算法除设置初始解外,还需要预先输入一组观测序列Obs。于是本文基于$Seq(p)$和$lineSeq$两类序列类型,分别构造序列Obs_1和Obs_2,并分别对这两种观测序列进行学习,得到模型参数。随后利用维特比算法进行准确率评估。

5.3.1 基于Obs_1的准确度

Obs_1为上文所定义的第一种观测序列,即以某一个乘客为中心提取出包含该乘客的所有不重复共乘团体,并对这些共乘团体按照其出行时间进行排序而得到的出行序列。

为了验证模型的有效性以及出行序列训练集大小对模型的影响,本文设置了3种Obs_1:1)$Obs_{[1,100]}$;2)$Obs_{[1,200]}$;3)$Obs_{[1,300]}$。下标$[i,j]$中的i代表该序列为Obs_1类型的观测序列,j代表出行序列训练集中包含出行序列的数量。

本文设置的3种Obs_1的出行序列训练集分别包含了100,200和300条出行序列,从训练集中选取对应数量的出行序列作为其训练数据,与初始解一起代入鲍母-韦尔奇算法对模型的参数进行学习,经过学习得到参数集分别为:1)$Param_{[1,100]}$;2)$Param_{[1,200]}$;3)$Param_{[1,300]}$。每种Obs_1会选取3组其对应大小的出行序列训练集进行参数学习,因此每一个参数集中包含了3组参数(A,B,π)。

在准确率预测阶段,本文在测试集中随机选取3组大小为200条的出行序列测试集,简化表示为TG_1,TG_2,TG_3。在维特比算法中代入学习得到的参数对3组出行序列测试集中包含的共乘团体的关系进行预测,每一组出行序列测试集在不同的参数上表现出的平均准确率如表13所列。

表13 基于Obs_1的预测准确率

Table 13 Prediction accuracy based on Obs_1

(单位:%)

Parameters	TG_1	TG_2	TG_3
$Param_{[1,100]}$	73.98	74.63	73.54
$Param_{[1,200]}$	77.54	77.36	78.37
$Param_{[1,300]}$	83.79	84.58	87.29

可以看出:1)$Param_{[1,300]}$预测TG_3的平均准确率达到了87.29%;2)当训练集扩大后,预测平均准确率也在增加,这一点与推论相符。

5.3.2 基于Obs_2的准确度

Obs_2为上文所定义的第二种观测序列类型,同Obs_1,本文设置了3种Obs_2:1)$Obs_{[2,100]}$;2)$Obs_{[2,200]}$;3)$Obs_{[2,300]}$。下标$[i,j]$中的i代表该序列为Obs_2类型的观测序列,j代表出行序列训练集中包含出行序列的数量。

经过学习得到参数集分别为:1)$Param_{[2,100]}$;2)$Param_{[2,200]}$;3)$Param_{[2,300]}$。每种Obs_2会选取3组其对应大小的出行序列训练集进行参数学习,因此每一个参数集中包含了3组参数(A,B,π)。

在准确率测试阶段,同样本文在测试集中随机选取3组大小为200条的出行序列测试集,简化表示为TG_1,TG_2,TG_3。在维特比算法中代入学习得到的参数对3组出行序列测试集中包含的共乘团体的关系进行预测,每一组出行序列测试集在不同的参数上表现出的平均准确率如表14所列。

表14 基于Obs_2的预测准确率

Table 14 Prediction accuracy based on Obs_2

(单位:%)

Parameters	TG_1	TG_2	TG_3
$Param_{[2,100]}$	84.13	84.42	85.62
$Param_{[2,200]}$	86.62	87.42	88.13
$Param_{[2,300]}$	95.63	95.42	96.38

可以看出:1)$Param_{[2,300]}$预测TG_3的平均准确率达到了96.38%;2)当训练集扩大后,预测平均准确率也在增加,这一点与推论是相符的。

通过比较两种结构的观测序列在模型中的性能发现,Obs_2优于Obs_1,这是由于在组成结构上Obs_2是以多个中心乘客发散形成的出行序列,而Obs_1是以单个中心乘客延伸出的出行序列,Obs_2会长于Obs_1,这一点也从侧面证明了训练集越大,模型的性能就越好。

5.4 模型预测一致性实验

一个共乘团体包含至少两个乘客,而维特比算法中的观测序列提取是以乘客为中心的,因此该团体可能出现在多个出行序列中,对团体关系预测的一致性进行判定是很有必要的。本文定义的一致性判定规则为同一共乘团体在不同出行序列中利用维特比算法预测出的团体关系是否一致,若一致则称为同关系,否则为异关系。

基于此,本文随机选取了1100个不重复共乘团体,这些共乘团体都出现至少两次,对包含这些团体的出行序列进行团体关系预测,判断结果如表15所列,1044个团体在不同的序列中表现为同一种团体关系,属于同关系数据集(简化为SRS),只有56个共乘团体在预测时出现了不同的团体关系,

属于异关系数据集（简化为 DRS）。

表 15　团体关系预测一致性

Table 15　Prediction consistency of group relationship

（单位：%）

	SRS	*DRS*
团体比例	95	5
团体数量	1 044	56

结束语　目前对旅客出行的研究大部分着眼于对旅客出行方式、旅客出行目的以及旅客出行时间等方面，针对旅客同行关系方面的研究还很少，特别是对铁路共乘团体出行关系的预测更是欠缺，而明确团体出行关系是大部分旅客出行研究的基础。针对以上情况，本文确定了以预测铁路共乘团体出行关系为核心。综合对相关文献的整理与研究，本文提出了基于铁路的隐马尔可夫模型，以单个用户为中心提取出带有时序的出行序列，对序列中的共乘团体进行出行关系预测。结合铁路出行场景的特点：1）对隐马尔可夫模型中的各项参数以及计算公式进行了重定义；2）提出共乘团体出行次数的量化标准，对数据进行拟合；3）利用人口统计学相关信息对铁路购票订单中的共乘团体进行统一描述，从中提取本文实验所需的数据。最终将其应用于真实铁路数据中，进行共乘团体出行关系的预测。两类实验结果证明，本文方法能够高效且准确地完成共乘团体出行关系的预测。

本文未来的研究主要分为两方面开展：一方面完善多种关系的判定方法，旅客之间的同行关系是多种多样的，需要丰富共乘团体之间的社会关系；另一方面丰富旅客间同行关系的表达，不仅考虑静态特征（同行频次等），还要延长研究时间跨度，考虑动态特征（如出行间隔时间、大部分出行所选时期、团体出行的距离、是否往返以及是否为节假日出行等）来研究出行特征。

参 考 文 献

[1] LI A, AXHAUSEN K W. Trip purpose imputation for taxi data [C]// 8th Swiss Transport Research Conference. 2018.

[2] CHEN C, LIAO C, XIE X, et al. Trip2Vec: a deep embedding approach for clustering and profiling taxi trip purposes[J]. Personal and Ubiquitous Computing, 2019, 23(1): 53-66.

[3] YANG S, WENG J, CHEN Z, et al. Taxi travel purpose estimationand characteristic analysis based on multi-sourcedata and semantic reasoning a case study of beijing[C]// Web Information SystemsEngineering WISE 2013 Workshops. 2013: 474-492.

[4] DENG Z W, JI M H. Deriving rules for trip purpose identification from GPS travel survey data and land use data: A machine learning approach [C] // Seventh International Conference on Traffic and Transportation Studies. 2018: 768-777.

[5] ZHU Z, ULF B, GERHARD T. Inferring travel purpose from crowd-augmented human mobility data[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'17). 2014: 44-49.

[6] BAO J, XU C, LIU P, et al. Exploring Bikesharing Travel Patterns and Trip Purposes Using Smart Card Data and Online Point of Interests[J]. Networks and Spatial Economics, 2017, 17(4): 1231-1253.

[7] CUI Y. Forecasting current and next trip purpose with social media data and Google Places[J]. Transportation Research Part C: Emerging Technologies, 2018, 97: 159-174.

[8] LIN Y, WAN H, JIANG R, et al. Inferring the Travel Purposes of Passenger Groups for Better Understanding of Passengers [J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(1): 235-243.

[9] QIAN J P, SHAO C F, LI J. Trip Purpose Inference of Group Passengers Based on Ticket Sales Data[J]. Journal of Transportation Systems Engineering and Information Technology, 2020, 20(6): 99-105.

[10] ZHOU C X, XIAO L L. The analysis of travel behavior during morning rush hour considering household travels[J]. Systems Engineering-Theory & Practice, 2020, 40(12): 3220-3229.

[11] JIA Z, WANG D Z, CAI X. Traffic managements for household travels in congested morning commute[J]. Transport Research Part E, 2016, 91: 173-189.

[12] SUBRAMANIYASWAMY V, VIJAYAKUMAR V, LOGESH R, et al. Intelligent Travel Recommendation System by Mining Attributes from Community Contributed Photos[J]. Procedia Computer Science, 2015, 50: 447-455.

[13] CHEN Y Y, CHENG A J, HSU W. Travel Recommendation by Mining People Attributes and Travel Group Types From Community-Contributed Photos[J]. IEEE Transactions on Multimedia, 2013, 15(6): 1283-1295.

[14] NASERIAN E, WANG X, DAHAL K, et al. Personalized location prediction for group travellers from spatial-temporal trajectories[J]. Future Generation Computer Systems, 2018, 83: 278-292.

[15] WAN Y H, WAN Z W, LIN Y F, et al. Discovering family groups in passengers social networks[J]. Journal of Computer Science and Technology, 2015, 30: 1141-1153.

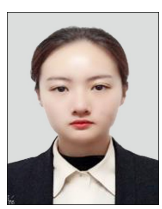
[16] OZDEMIR E, TOPCU A E, OZDEMIR M K. A hybrid HMM model for travel path inference with sparse GPS samples[J]. Transportation, 2018, 45(1): 233-246.

[17] GUEYE I, NDONG J, SARR I. An Accurate Probabilistic Model for Community Evolution Analysis in Social Network[C]// The 11th International Conference on Signal-Image Technology & Internet-Based Systems(SITIS). IEEE, 2015: 343-349.

[18] TU S. HMM-based User Behavior Prediction Method in Heterogeneous Cellular Networks [J]. International Journal of Performability Engineering, 2018: 14(9): 2163.

[19] XIONG C. The analysis of dynamic travel mode choice: a heterogeneous hidden Markov approach [J]. Transportation, 2015, 42(6): 98-106.

WANG Xin, born in 1981, Ph.D, professor, Ph. D supervisor, is a member of ACM, IEEE, CCF and CAAI. His main research interests in clude knowledge discovery in database, artificial in telligence, machine learning and data mining.



LI Si-ying, born in 1996, postgraduate. Her main research interest includes data mining.