基于非参数密度估计的不确定轨迹预测方法

程 媛 [1, 2] 迟荣华 [3] 黄少滨 [3] 吕天阳 [3, 4]

摘 要 随着大量移动设备的出现, 准确和高效的轨迹预测有助于提高面向位置的应用和服务的质量和水平. 针对现有方法对轨迹不确定性缺乏有效建模的问题, 提出了基于非参数密度估计的不确定轨迹终点预测方法. 在轨迹建模及模型训练阶段, 利用非参数估计对起点与终点相同的轨迹构建基于密度分布的不确定轨迹模型; 在轨迹预测阶段, 将待预测轨迹视为轨迹数据流, 并通过 KS (Kolmogorov-Smirnov) 检验方法与具有相同起点的不确定轨迹模型进行匹配, 其中匹配程度最高的不确定轨迹即为预测轨迹. 通过真实轨迹数据集上的实验表明, 与现有各类主要轨迹预测方法相比, 本方法在不同条件下的预测效率与准确性都有较明显优势.

关键词 轨迹预测, 不确定性, 非参数密度估计, KS 检验

引用格式 程媛, 迟荣华, 黄少滨, 吕天阳. 基于非参数密度估计的不确定轨迹预测方法. 自动化学报, 2019, **45**(4): 787−798

DOI 10.16383/j.aas.2018.c170419

Uncertain Trajectory Prediction Method Using Non-parametric Density Estimation

CHENG Yuan[1, 2] CHI Rong-Hua[3] HUANG Shao-Bin[3] LV Tian-Yang[3, 4]

Abstract With the popularization of a large number of mobile devices, the accurate and efficient trajectory prediction could help to improve the service quality of location-oriented applications. To solve the problem of less effectiveness existing in modeling for uncertain trajectories, we propose a method for predicting the destination of uncertain trajectories using the non-parametric density estimation method. In the modeling stage, the uncertain trajectory model between the same origin and destination is constructed with the method of non-parametric estimation to represent the density distribution feature. In the trajectory prediction stage, the trajectory to be predicted is regarded as a data stream. And it is matched with the uncertain trajectory having the same origin through the KS (Kolmogorov-Smirnov) hypothesis testing. Then the optimal matching uncertain trajectory is the prediction result and its destination is the predictive destination. The Experiments on real trajectory datasets indicate that the proposed method has obvious advantages in prediction efficiency and accuracy under different conditions, as compared to the existing trajectory prediction methods.

Key words Trajectory prediction, uncertainty, non-parametric density estimation, KS hypothesis testing

Citation Cheng Yuan, Chi Rong-Hua, Huang Shao-Bin, Lv Tian-Yang. Uncertain trajectory prediction method using non-parametric density estimation. *Acta Automatica Sinica*, 2019, **45**(4): 787−798

当前, 类似手机等智能终端的广泛普及, 使人类传感信息的获取变得更加容易, 而面向这种传感信息的应用和服务则随之快速增长且类型多种多样. 其中面向位置的服务在基于终端传感信息的应用中最为热门, 这类服务可实时获取用户的坐标和区域等位置信息. 所获取的位置信息的历史数据可用于对用户的移动轨迹进行预测, 而较准确的预测结果能够为用户提供所需的信息和帮助.

较为常见的轨迹预测研究是轨迹序列预测. Song 等[1] 在预测人类移动轨迹的研究工作中, 通过个体的轨迹信息熵的测量, 量化了人类移动行为的一般性规律, 并得出人的移动行为是可预测的. 其他移动轨迹预测方法还包括轨迹频繁模式挖掘[2−4]、基于复杂信息的行为模式挖掘[5−6] 和混合方法[7].

收稿日期 2017-07-28 录用日期 2018-01-08
Manuscript received July 28, 2017; accepted January 8, 2018
国家自然科学基金 (91546110), 黑龙江省自然科学基金 (F2017015), 黑龙江省普通高等学校青年创新人才培养计划 (UNPYSCT-2017079) 资助
Supported by National Natural Science Foundation of China (91546110), Natural Science Foundation of Heilongjiang Province (F2017015), and Training Program for Young Innovators in Heilongjiang General Institutes of Higher Education (UNPYSCT-2017079)
本文责任编委 曾志刚
Recommended by Associate Editor ZENG Zhi-Gang
1. 哈尔滨理工大学计算机科学与技术学院 哈尔滨 150080 2. 哈尔滨理工大学计算机科学与技术学院博士后流动站 哈尔滨 150080 3. 哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001 4. 审计署计算机技术中心国家仿真实验室 北京 100071
1. College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080 2. Postdoctoral Research Station, College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080 3. College of Computer Science and Technology, Harbin Engineering University, Harbin 150001 4. National Audit Simulation Laboratory of IT Center, National Audit Office, Beijing 100071

由于人的移动行为极其复杂, 可能会受到信号采集、客观环境、主观意图等多种不确定因素的影响, 使得对轨迹序列难以进行较准确的预测. 而移动行为复杂多变, 但一般情况下, 人的行为目的性较强, 移动轨迹终点较为明确, 因此移动轨迹预测的另一类研究问题就是移动轨迹的终点预测.

轨迹终点的预测方法一般是将待预测轨迹的数据流与历史轨迹数据进行匹配, 选择并筛选出最为相近的轨迹, 该轨迹的终点即为待预测轨迹的终点, 匹配方法也多基于常见的分类预测模型. De Brébisson 等[8] 以轨迹的起始坐标和背景等信息作为特征向量训练多层感知的神经网络, 通过匹配训练数据集中的轨迹实现对轨迹终点的预测. 对于训练过的相似轨迹, 模型可以较好地预测; 但对于新区域的轨迹, 模型需要新的训练数据集以及新的相关信息; 另外, 网络模型本身难以解释, 使其难以通过调整和控制网络模型适应新的数据, 限制了模型的适用性. Krumm 等[9] 和 Ziebart 等[10] 利用除轨迹以外的信息进行贝叶斯信念网的训练来预测轨迹的终点. Patterson 等[11] 也基于贝叶斯模型预测轨迹终点, 但考虑了如行走、乘车之间状态改变等形式的历史轨迹的运动模式信息. Monreale 等[12] 利用 T-pattern 决策树对移动轨迹建模, 通过匹配决策树中的路径预测轨迹终点.

随着轨迹预测研究的深入, 轨迹中的不确定性问题也逐渐被发现是限制轨迹预测准确性的重要因素, 现有研究大多从轨迹中位置间转移概率的角度来考虑不确定因素. Ashbrook 等[13] 在 GPS 数据上, 结合位置间的转移概率训练 Markov 模型, 并预测轨迹的终点. Gambs 等[14] 利用 Markov Chains 对轨迹位置序列中的兴趣点建模, 通过计算兴趣点间的转移概率预测轨迹的终点. 现有研究中利用的 Markov 和贝叶斯模型均可以描述位置之间的转移概率, 但其模型属于离散类型, 对于历史信息中尚未出现过的轨迹预测结果准确性有限. 因此, 乔少杰等[15] 以及 Besse 等[16] 利用高斯密度分布对轨迹建模, 并利用高斯回归过程对轨迹进行预测. 这种基于连续密度的模型不仅可以以概率的形式对轨迹的不确定建模, 还能对轨迹的终点给出连续形式的解, 相较于离散模型具有更高的预测准确性.

现有研究虽然采用了不同的预测方法, 但几乎均需要根据待检测的移动轨迹与历史轨迹数据的匹配程度进行轨迹终点的判断. 然而在真实的移动场景中, 人的移动行为具有较强的不确定性, 一是由于信息采集设备的能效问题可能导致录入的信息存在不同程度的损失或干扰, 使获取的信息与真实信息存在偏差; 再者由于人与社会的复杂性, 人的移动具有较强的不确定性, 使得在以同一目标为终点进行移动时可能出现任意不同的移动轨迹, 例如为了躲避交通的拥堵可能选择路程相对较远的路线, 或为了兴趣或特殊目的而选择比较随意的移动路线等. 可见, 真实的移动场景中的轨迹充满了各种各样的不确定性, 一个用户当前的移动轨迹可能与任意一条历史轨迹均不相同, 但实际上又可能与某条历史移动轨迹具有相同的起始点和终点. 针对这样的体现用户不确定移动行为的轨迹, 现有方法难以对终点进行较准确的预测. 对于移动轨迹的研究面临的问题来说, 一方面是无法准确地确定移动对象的移动轨迹细节 (例如移动对象下一个准确的移动位置), 另一方面是难以通过单一或简单的假设理论模型进行确定的建模, 这种复杂性与不确定性的存在, 使得理想条件下对移动轨迹的预测难以获得较准确的结果. 面对这种问题, 本文将结合不确定性的研究思路根据用户当前的移动行为其轨迹终点进行预测.

一个用户的移动轨迹由其经过的若干个位置信息构成. 如果将一个用户的移动轨迹视为不确定的, 那么起始点和终点相同的若干移动轨迹则构成了表示两点间所有经过路线的不确定轨迹数据集, 即包含了两点间所有可能路线的样本集. 此时可以从该不确定样本集内的所有移动轨迹中获得用户在两点间的移动行为分布特征, 而这种行为的分布特征具体体现为轨迹中位置信息的分布特征. 如果认为移动轨迹中的位置点服从某种密度分布, 那么对于概率密度较高的位置, 分析当前用户的移动轨迹时会有较高的可能性选择该位置. 而非参数估计方法能够在不假设分布类型的前提下构建更符合数据实际分布特征的概率密度函数, 因此本文基于这种不确定性的研究思路, 提出一种基于非参数密度估计的面向不确定轨迹的终点预测方法. 首先基于非参数估计方法, 对已获取的两点间的所有移动轨迹数据构建符合其分布特征的概率密度函数; 然后在分析当前用户的移动轨迹时, 基于假设检验方法度量待检测轨迹与已有的不确定轨迹关于轨迹中位置点分布特征的相似性, 分析待检测轨迹未来可能经过的位置点以及可达的终点. 所提方法主要利用不确定性的分析方法, 结合所有历史移动轨迹信息的特点, 对移动轨迹进行建模, 使得模型能够基于概率分布函数体现历史数据中所有用户的移动行为特征, 最终达到能够对历史数据中尚未出现的轨迹也具有极好的识别能力, 即在更接近人类移动行为特征的条件下, 对其轨迹的未来移动位置和终点进行较准确的预测. 本文首先构建不确定移动轨迹模型; 进而提出不确定移动轨迹相似性度量方法; 然后据此对待检测的移动轨迹的终点进行预测; 接着通过实验验

证所提方法的有效性; 最后得出结论.

1　非参数密度估计

非参数密度估计是一种对先验知识要求最少, 完全依靠训练数据进行估计, 而且可用于任何形状密度函数的方法. 非参数密度估计方法能够在不假设数据分布形式的前提下获取符合分布特征的概率密度函数, 常用的非参数密度估计方法有直方图估计方法、核函数估计方法以及 K–近邻估计方法. 其中基于高斯核函数的非参数密度估计方法由于其较好的数学性质以及广泛的适用性, 在本文中予以采用. 基于如下核密度估计方法求得的概率密度函数能够在样本点与目标数据量足够大时, 收敛到任意一种密度函数[17].

$$G(x_j)=\sum_{i=1}^{N} q_i \mathrm{e}^{\frac{-\|\boldsymbol{x}_j-\boldsymbol{s}_i\|^2}{h^2}} \tag{1}$$

其中, q_i 为权重系数, h 为核函数的平滑参数 — 带宽 (bandwidth), $\{\boldsymbol{s}_i\}_{i=1,\cdots,N}$ 为服从某种未知分布的样本点.

假设数据 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 是 d 维向量, 并取自一个连续分布 $f(\boldsymbol{x})$, 在任意点 $\boldsymbol{x}$ 处的一种核密度估计定义为 $\hat{f}(\boldsymbol{x})=\frac{1}{nh^d}\sum_{l=1}^{n}K(\frac{\boldsymbol{x}-\boldsymbol{x}_l}{h^d})$, 其中 $f(\boldsymbol{x})$ 是一个 d 维随机变量的密度函数, $K(\cdot)$ 是定义在 d 维空间上的核函数, $K:\mathbf{R}^d\rightarrow\mathbf{R}$, 并满足条件 $K(\boldsymbol{x})\geq 0$, $\int K(\boldsymbol{x})\mathrm{d}u=1$. 其中, 带宽 h^d 对模型的光滑程度有较大的影响, 合适的取值可使模型具有更好的拟合效果. 在维度较高时, 带宽的取值需要参考每个维度上的取值分布, 可以通过核密度估计的边缘概率计算, 属于比较复杂的情况. 而在本文所提方法中假设表示轨迹点的经度和纬度的变量是相对独立的, 因此分别计算各个维度的 h 即可, 利用最小化均方误差方法, 当核函数为高斯核时, 可取最优带宽 $h_{\mathrm{opt}}=1.06\sigma n^{1/5}$.

2　不确定移动轨迹模型

理想的移动轨迹信息记录了移动对象的起点、终点、必要的路径信息以及较少的噪音或冗余, 这种移动模式简单且轨迹间的相似程度较高. 考虑到移动轨迹中的不确定性, 本文结合非参数密度估计方法对人的移动轨迹进行建模与预测. 将人的移动轨迹中的所有路线作为样本数据, 用于构建面向不确定性的移动轨迹模型, 其中的不确定性用能够体现轨迹样本分布的概率密度函数进行表示; 然后以该模型为基础, 结合个人当前的移动路线, 分析预测可能的移动方向与路径.

首先, 令数据集 D 表示已有的对象移动轨迹集合, 该集合中的每条数据记录了一个对象在一次移动行为中不同时间点采集的位置信息, 即每条记录由对象的位置信息序列构成. 那么一个移动对象的移动轨迹有序集可表示为 $D=\{T^1,T^2,\cdots,T^n\}$, 其中第 i 条移动轨迹序列 $T^i=\{\boldsymbol{u}_1^i,\boldsymbol{u}_2^i,\cdots,\boldsymbol{u}_d^i\}$, 表示为在 d 个时间点采集到的位置的有序集合, 且 u_j^i 为在第 j 个时间点采集的位置信息, 并可表示为一个三元组: $\boldsymbol{u}_j^i=(\boldsymbol{x}_j^i,\boldsymbol{y}_j^i,t_j^i)$, $1\leq j\leq d$, 其中 t_j^i 表示采样时间, $\boldsymbol{x}_j^i$, $\boldsymbol{y}_j^i$ 表示在 t_j^i 时间点采集的二维位置信息.

如果将轨迹数据集表示为 $D=\{(\boldsymbol{u}_1^1,\boldsymbol{u}_2^1,\cdots),(\boldsymbol{u}_1^2,\boldsymbol{u}_2^2,\cdots),\cdots,(\boldsymbol{u}_1^n,\boldsymbol{u}_2^n,\cdots)\}$, 可得采集到的轨迹位置集合 $U=\{\boldsymbol{u}_1^1,\boldsymbol{u}_2^1,\cdots,\boldsymbol{u}_1^2,\boldsymbol{u}_2^2,\cdots,\boldsymbol{u}_1^n,\boldsymbol{u}_2^n,\cdots\}$. 其中的位置采集信息可能包含实际坐标一致的位置, 也可能包含坐标相近的位置; 实际上在轨迹预测与规划中, 坐标相近的点可作为同一位置处理. 类似于公交站点的设置, 与某一站点相近的位置均属于该站点的可达范围, 若欲到达与某站点相近的目的地, 在该站点下车即可. 同样, 将集合 U 中坐标相近的采集位置信息作为同一地点处理, 得到整合后的轨迹点集 $V=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k\}$, 显然 $|V|\leq|U|$. 整合后的轨迹点集有助于重构历史轨迹路线, 以及分析同一出发地和同一目的地间的轨迹特征, 进而为后续轨迹预测奠定基础.

此时原始轨迹数据集 D 中的轨迹序列 T^i, $1\leq i\leq n$ 可表示为 $T^{i'}=\{\boldsymbol{v}_1^i,\boldsymbol{v}_2^i,\cdots,\boldsymbol{v}_{d'}^i\}$, 而且不同的轨迹序列间可能存在若干相同的轨迹点, 那么重构后的轨迹数据集可表示为 $D'=\{T^{1'},T^{2'},\cdots,T^{n'}\}$. 从 D' 中的轨迹序列即可提取若干具有相同起始点与终点的轨迹路线, 在这些路线中还极可能包含不同的轨迹点, 体现了不同用户在移动中的不同行为. 本文将这种具有相同起始点 $\boldsymbol{s}$ 终点 $\boldsymbol{e}$ 且经过不同轨迹点集的路线称为起始点 $\boldsymbol{s}$ 终点 $\boldsymbol{e}$ 的不确定轨迹, 具体描述如定义 1 所述.

定义 1 (不确定轨迹). 移动对象的不确定轨迹 UT 由起始点 $\boldsymbol{s}$、终点 $\boldsymbol{e}$ 以及其间经过的轨迹点集 ns 和轨迹集 ts 构成, 表示为一个四元组: $UT=(\boldsymbol{s},\boldsymbol{e},ns,ts)$, 其中 $\boldsymbol{s},\boldsymbol{e}\in V$, $ns\subset V$, ts 是 D' 中所有经过 $\boldsymbol{s}$ 与 $\boldsymbol{e}$ 的轨迹的集合, 可表示为 $ts_{s,e}=T_{s,e}^{1'}\cup T_{s,e}^{2'}\cup\cdots\cup T_{s,e}^{n'}$.

一条不确定轨迹 UT 即是从移动轨迹数据集 D' 中提取的, 在起始点 $\boldsymbol{s}$ 和终点 $\boldsymbol{e}$ 之间的所有轨迹的集合; 并且起始点 $\boldsymbol{s}$、终点 $\boldsymbol{e}$ 以及轨迹中的轨迹点均属于轨迹点集 V. 即轨迹数据集 D' 中每条数据表示一条用户的移动轨迹序列 $T^{i'}$, 但在这些序列中包含的任意两点间都可能存在至少一条轨迹, 例如 $T_{i,j}^{1'}$ 表示在 $T^{1'}$ 中包含的 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 间的可达轨迹

路线. 因此若将 $\forall \boldsymbol{v}_i, \boldsymbol{v}_j \in V$ 作为不确定轨迹的起始点与终点, 并假设 D' 中每条轨迹序列中都包含这两个轨迹点间的路线, 可得 $\boldsymbol{v}_i$, $\boldsymbol{v}_j$ 间的不确定轨迹 $UT_{i,j} = \{(\boldsymbol{v}_i, \boldsymbol{v}_j, ts_{i,j}) | \boldsymbol{v}_i, \boldsymbol{v}_j \in V\}$, 其中 $ts_{i,j} = T_{i,j}^{1'} \cup T_{i,j}^{2'} \cup \cdots \cup T_{i,j}^{n'}$, 表示在历史数据中用户经过这两点时可能选择的所有轨迹路线.

通过对轨迹数据集 D' 的分析提取, 能够得到不确定轨迹数据集 $UTD = \{UT_{i,j}\}$, 其中每个元素表示一条不确定轨迹 $UT_{i,j} = \{(\boldsymbol{v}_i, \boldsymbol{v}_j, ts_{i,j}) | \boldsymbol{v}_i, \boldsymbol{v}_j \in V, i \neq j\}$. $ts_{i,j}$ 中并非仅有一种 $\boldsymbol{v}_i$,$\boldsymbol{v}_j$ 间的可达路线, 而是覆盖了采集信息中包含的所有经过两点间的用户路线; 因此能够体现不同用户在移动过程中的行为, 并为预测后续用户的移动轨迹提供较丰富的信息资源.

基于轨迹数据集虽然难于分析每个用户选择移动路线的目的, 但从 $ts_{i,j}$ 包含的移动轨迹中能够分析用户在两点间的移动行为分布特征. 如前所述由于多种原因, $ts_{i,j}$ 中的轨迹所经过的轨迹点是极其复杂和不确定的, 因此轨迹的密度分布一般不服从某种简单的假设分布, 例如正态分布、幂律分布等. 而非参数估计方法能够在不假设数据分布形式的前提下获取符合分布特征的概率密度函数, 本文即利用非参数估计中的核密度估计方法获取不确定轨迹中可体现用户行为分布特征的概率密度函数. 另外, 不同轨迹中的轨迹点及其数量均可能不同, 难以直接分析每条轨迹的分布密度; 而 ts 中的所有轨迹均由轨迹点组成, 因此分析两点间的用户移动行为特征时, 可将分析对象细化至轨迹点, 那么构建的概率密度函数实际描述的是在历史路径中轨迹点的分布情况.

定义 2 (密度式不确定轨迹). 定义 1 中的 $UT = (\boldsymbol{s}, \boldsymbol{e}, ns, ts)$ 表示起始点 $\boldsymbol{v}_s$ 和终点 $\boldsymbol{v}_e$ 间的不确定轨迹, 且 $\boldsymbol{v}_s, \boldsymbol{v}_e \in V$. 令 ts 中所有轨迹包含的轨迹点集合为 $ns_{s,e} = \{\boldsymbol{v}_1^{s,e}, \boldsymbol{v}_2^{s,e}, \cdots, \boldsymbol{v}_m^{s,e}\}$, $v_l^{s,e} \in V$, $1 \leq l \leq m$, 其中 m 为轨迹点的个数. 根据核密度估计方法可得 $ns_{s,e}$ 的概率密度函数为 $f^{s,e} = \frac{1}{mh^2} \sum_{l=1}^{m} K((\boldsymbol{x} - \boldsymbol{v}_l^{s,e})/h^2)$, 由于构建的是关于轨迹点的概率密度函数, 其中 $\boldsymbol{x}$ 为表示轨迹位置中经纬度的 2 维自变量, h^2 则表示 2 维的最优带宽. 那么起始点 $\boldsymbol{v}_s$ 和终点 $\boldsymbol{v}_e$ 间的密度式不确定轨迹为 $UTf = (\boldsymbol{s}, \boldsymbol{e}, f)$.

图 1 是以起始点 (39.138, 117.213) 和终点 (39.220, 117.161) 构成的轨迹及其密度分布. 该轨迹由多条经过起始点和终点的轨迹构成, 在该轨迹的前半部出现了不同的轨迹路径, 但后半程轨迹路径基本相同, 因此在密度分布中会出现位置点密度不均匀的情况.

可见为了提取移动轨迹数据集中利于预测分析用户移动轨迹行为的信息, 需要从中提取任意两点间的不确定轨迹并为其建模. 对于一条不确定轨迹而言, 其模型的构建不仅包括轨迹的起始点、终点和两点间用户选择的多种轨迹路线, 还包括这些路线中覆盖的轨迹点的分布特征, 即轨迹的不确定性体现在其中包含的多条可选轨迹路线, 而不确定性的描述则由能够体现轨迹点分布特征的概率密度函数来完成. 实际所构建的模型本身不再关注难于确定的无穷轨迹细节, 而是更多关注轨迹目标的终点以及可能出现的轨迹模式.

3 基于假设检验的不确定轨迹相似性度量

对不确定轨迹建模主要是试图根据当前用户的移动行为, 基于数据集中其他用户的历史移动行为特征, 判断其将来可能的移动路线与目的. 这个过程中的关键是不确定轨迹间的相似性, 即通过度量当前用户的移动轨迹与历史不确定轨迹间的相似性, 判断其未来可能的移动轨迹路线.

不确定轨迹的本质在于由作为样本轨迹的不同用户选择的多种轨迹路线所体现的不确定性, 而其不确定性可由轨迹点的概率密度函数描述, 表示移动对象在不同位置经过的可能性. 此外在轨迹预测的问题中, 预测方法的输入是动态采样的轨迹点, 是以流的形式输入的数据, 即轨迹数据流. 轨迹数据流是连续的且随时间不断动态增长的轨迹序列, 假设数据流 $T = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_t, \cdots\}$, $\boldsymbol{u}_t$ 是 t 时刻采集到的位置信息, $\boldsymbol{u}_{t+1}$ 是相对于 $\boldsymbol{u}_t$ 在 $t+1$ 时刻的新增的位置信息. 每采集到一个新的数据就会相应补充到原有序列的尾部, 并作为轨迹序列整体的一部分保存. 可见待预测的目标对象以数据流的形式表示, 并且无法一次性获取大量的数据样本以分析其分布特征. 所以在完成预测任务时难以直接比较两个概率密度函数间的相似性. 此时预测任务实为判断表示用户当前不完整移动行为的轨迹点的有限样本与已有不确定轨迹的密度分布间的匹配程度, 然后将匹配程度较高的不确定轨迹中的移动轨迹与目的地作为当前用户未来可能移动轨迹路线的预测结果. 能够实现这一目的的有效方法是假设检验方法.

假设检验是用来判断样本与样本、样本与总体的差异是由抽样误差引起还是本质差别造成的统计推断方法. 其基本原理是先对总体的特征做出某种假设, 然后通过抽样研究的统计推理, 对此假设应该被拒绝还是接受做出推断. 常用的假设检验方法有 u- 检验法、t 检验法、$\aleph^2$ 检验法 (卡方检验)、F- 检验法以及秩和检验等. 这些假设检验方法需要对数据分布进行某种假设, 而不确定轨迹的密度分布一

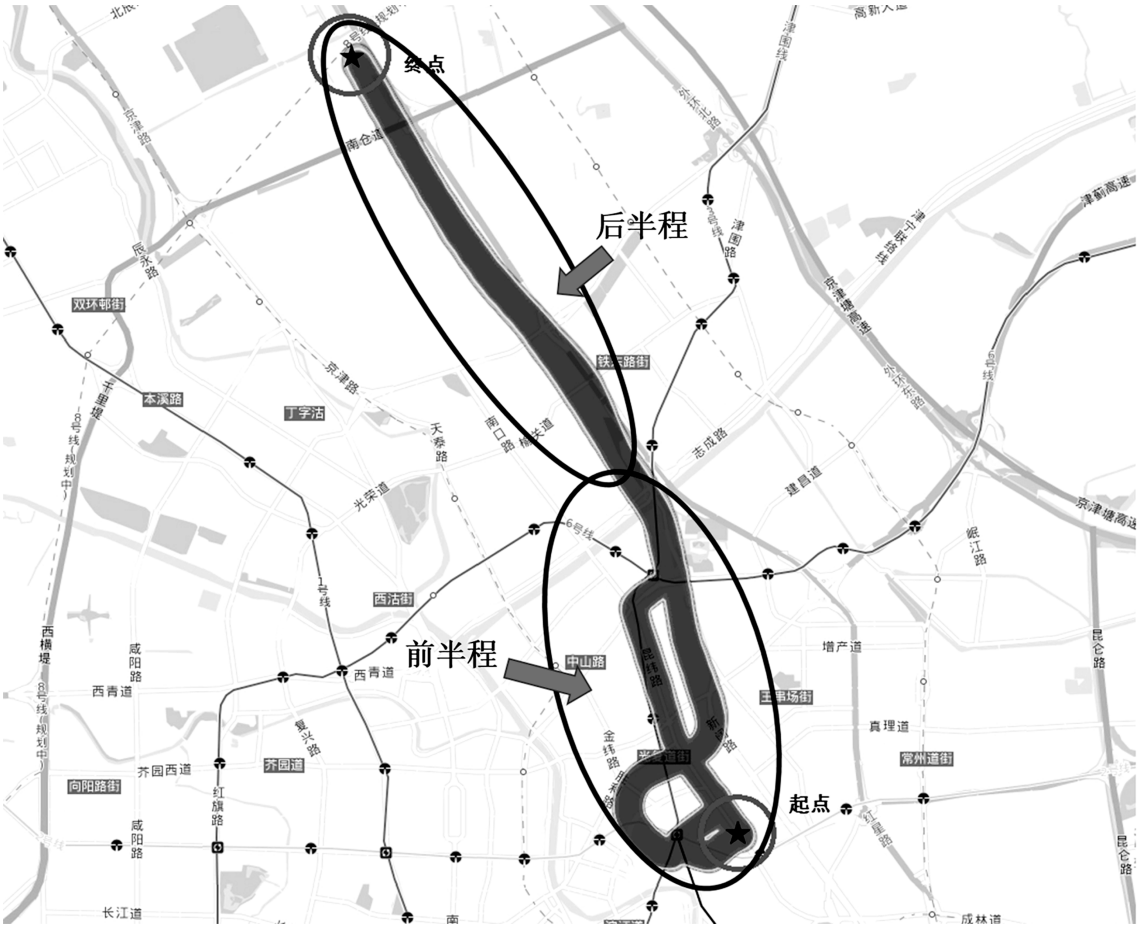


图 1　不确定轨迹示意图

Fig. 1　Uncertain trajectory

是其分布情况预先未知, 二是其一般并不符合理想密度分布 (例如正态分布、指数分布等), 因此一般的假设检验方法并不足以分析本文面对的问题. 非参数估计方法能够在不对数据分布进行预先假设的基础上, 根据数据自身特点分析其分布特征, 有助于解决本文面对不确定轨迹的预测问题. 而 KS (Kolmogorov-Smirnov) 检验作为一种非参数的假设检验方法, 不仅具有很好的稳健性, 而且分析过程中不依赖均值的位置, 对尺度化不敏感, 即使样本不服从正态分布, 也依然具有良好的敏感性, 有助于完成表示当前用户行为的有限轨迹点样本与表示已有轨迹不确定性的密度分布间的匹配问题. 因此本文将基于 KS 检验进行不确定轨迹间的相似性度量.

KS 检验基于累积频数分布, 检验数据分布是否与某种理论分布相似, 或比较两个数据分布是否有显著性差异. KS 检验的核心依然是检验两个样本的密度分布是否一致, 但无需对数据的分布类型进行预先假设, 而且随着用户移动行为的持续进行, 获取的表示移动轨迹的数据逐步积累, 使得 KS 检验这样基于累积频数分布的方法能够在此过程中逐步获得接近数据真实分布的结论.

由于本文分析对象为二维的轨迹点, 因此根据二维正态分布的假设检验方法, 基于 KS 检验分析二维随机变量的分布情况. 通过轨迹点不同维度的密度分布以及累积分布的差异, 能够体现由轨迹点描述的轨迹间的差异. 那么基于 KS 检验方法判断轨迹关于轨迹点密度分布的一致性, 就能够说明轨迹间的相似性. KS 检验方法的应用需要考虑的重要问题是变量是否独立, 位置信息中的分量间是否独立决定了方法的合理性和复杂程度. 对于位置分量的独立性问题, 难以通过理论证明独立性是轨迹位置信息中的固有性质, 那么就需要通过实验验证变量独立性的存在, 因此本文通过卡方检验对样本进行随机抽取检验, 检验结果说明位置样本中表示经纬度的分量间相互独立. 另外, 在准确性损失较小的前提下, 假设位置信息分量间相互独立可大幅度降低模型和计算的复杂性, 在实现较高预测准确度的前提下进行快速轨迹预测. 出于上述考虑, 本文假

设位置信息中的分量间是独立的, 对于二维随机变量的 KS 检验, 只需要分别对各分量的分布进行检验即可, 因此本文只给出一维随机变量的 KS 检验, 然后再推广到二维随机变量的情况.

根据 KS 检验的定义, 假设随机变量 $\boldsymbol{x}$ 取自于样本, 目标密度函数为 f, 则有假设 H_0 : 总体 X 服从分布 f, 检验统计量为

$$Z = \sqrt{n}\max(|f_n(\boldsymbol{x}_{i-1}) - f(\boldsymbol{x}_i)|, |f_n(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)|)$$

若 H_0 为真, 则 Z 依分布收敛于 Kolmogonov 分布, 即 $Z \xrightarrow{d} Ko = \sup|B(f(x))|$, 可得

$$P(Ko \leq \boldsymbol{x}) = 1 - 2\sum_{i=1}^{+\infty}(-1)^{i-1}\mathrm{e}^{-2i^2\boldsymbol{x}^2}$$

对于轨迹对象, 假设轨迹数据流 $T = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, 其中 $\boldsymbol{x}_1$ 是轨迹的起点, $\boldsymbol{x}_n$ 是移动对象在时刻 n 的位置, 另有某密度式不确定轨迹 $UTf = (\boldsymbol{s}, \boldsymbol{e}, f)$. 若 $\boldsymbol{x}_1 = \boldsymbol{s}$, 根据 KS 检验判断 T 是否可能来自于不确定轨迹 UTf 的密度分布, 即判断 T 来自的总体是否服从概率密度函数为 f 的分布, 具体步骤如下:

1) 检验的假设定义如下: H_0 来自的总体服从概率密度函数为 f 的分布, 说明当前用户轨迹可能来自于数据集中的不确定轨迹 UTf 的密度分布; H_1 来自的总体不服从概率密度函数为 f 的密度分布.

2) 构造 f 的累积概率密度分布 F: $F(\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}} f(t)\mathrm{d}t$; 根据核密度函数公式, $F(\boldsymbol{x})$ 可表示为

$$F(\boldsymbol{x}) = \int_{-\infty}^{x} \frac{1}{mh^2}\sum_{l=1}^{m} K\left(\frac{t - v_l^{s,e}}{h^2}\right)\mathrm{d}t$$

其中, $K(\cdot)$ 表示核函数.

3) 计算累计分布 F'.

4) 计算统计量

$$Z = \sqrt{n}\max(|F'(\boldsymbol{x}_{i-1}) - F(\boldsymbol{x}_i)|, |F'(\boldsymbol{x}_i) - F(\boldsymbol{x}_i)|)$$

5) 根据 KS 检验临界值表[18], 若 $Z > Z(n, \alpha)$, 则拒绝 H_0 假设, 反之则接受.

若 H_0 假设成立, 说明 T 可能来自于具有相同起点的不确定轨迹 UTf 的密度分布, 预测 T 的轨迹终点时, 可参考 UTf 轨迹的终点.

4 轨迹预测算法

现有的轨迹预测研究中的一般预测目标是根据轨迹数据集中的历史轨迹信息, 预测移动对象起始点至终点的详细确切的轨迹, 试图为用户构建其未来可能经过的完全轨迹模式. 但在实际情况中, 影响人的移动行为的因素较多, 导致移动轨迹极其复杂而且容易出现与历史轨迹不同情况的移动模式. 因此本文主要对用户移动轨迹的最终目的地进行预测, 而非对轨迹中的每一个轨迹点进行精准预测.

轨迹终点的预测需要根据当前用户已经形成的移动轨迹数据流, 在不确定轨迹集中检索所有具有相同起始点且密度分布特征相同的不确定轨迹. 首先根据当前已形成的部分轨迹数据流 $T = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, 确定轨迹的起点 $\boldsymbol{s} = \boldsymbol{x}_1$; 在不确定轨迹数据集 UD 中检索所有满足起点为 $\boldsymbol{s}$ 的条件的 k 个不确定轨迹 $UTf_i = (\boldsymbol{s}, e_i, f_i)$, $1 \leq i \leq k$; 然后利用 KS 检验方法对轨迹数据流 T 与 f_i, $1 \leq i \leq k$ 依次进行假设检验, 检验结果为真的密度分布函数所对应的不确定轨迹可作为分析当前用户移动轨迹的依据, 即将不确定轨迹的终点作为待分析移动轨迹的终点的可能预测结果.

在实际的分析过程中, 用户移动行为中每一次被采集到的轨迹点都作为输入轨迹数据流的新增数据, 由于利用了 KS 检验基于累积分布的特点, 随着移动轨迹中采集到的轨迹点的增长以及输入轨迹数据流样本数量的增多, 分析轨迹分布的 KS 检验精度能够不断提高, 得到的结果不断贴近数据实际分布, 使得预测结果的准确性不断提高. 另外如果检测结果为没有在轨迹数据集中找到与以 $\boldsymbol{s}$ 为起始点的轨迹数据流 $T = \{\boldsymbol{s}, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t\}$ 相匹配的不确定轨迹, 可以考虑是否在数据集中存在不确定轨迹与当前数据流中的部分轨迹相匹配, 即将轨迹数据流进行截取得到 $\{\boldsymbol{x}_2, \cdots, \boldsymbol{x}_t\}$, 并作为输入轨迹数据流再次进行预测, 直到发现匹配结果或无历史数据可匹配为止. 该预测算法的具体步骤下:

算法 1. UDTM(D, T^*)

输入. 历史移动轨迹数据集 $D = \{T^1, T^2, \cdots, T^n\}$, T^*: 移动轨迹.

输出. 匹配的不确定轨迹集 $\{UDf^*\}$.

步骤 1. 令 U 为 D 中轨迹内包含的所有位置信息的集合.

步骤 2. 构建不确定轨迹集 UD.

1) $V = \mathrm{Kmeans}(U, k)$ //利用 K-Means 聚类将数据集中相似的位置归为一类, 形成约简后的轨迹点集.

2) **for** each $T_u \in D$ **do**

$T^u \xrightarrow{V} T^v$ //利用约简后的轨迹点对原有轨迹进行.

end

3) **for** each $\boldsymbol{v}_i$, $\boldsymbol{v}_j$ in V **do**

$w_{i,j} = T_1^v \cup T_2^v \cup \cdots$ //从约简后的轨迹集中计算任意两点间的可达轨迹.

end

4) $UD = \{(\boldsymbol{v}_i, \boldsymbol{v}_j, w_{i,j}) | \boldsymbol{v}_i, \boldsymbol{v}_j \in V\}$　//构建不确定轨迹集 UD, 包含任意两点间的不确定轨迹.

5) **for** each $w_{i,j} \in UD$ **do**

$f_{i,j} = f^{s,e}(w_{i,j})$　//根据定义 2 对每个不确定轨迹计算表示其密度分布的概率密度函数.

end

6) $UDf = \{(\boldsymbol{v}_i, \boldsymbol{v}_j, f_{i,j}) | i, j \in k\}$

//构建密度式不确定轨迹集, 包含任意两点间的密度式不确定轨迹.

步骤 3. 移动轨迹终点预测.

1) **while** $(UDf^* \neq \emptyset || T^* == \emptyset)$

for each $UT \in UDf$ where $\boldsymbol{s} = \boldsymbol{x}_1$

if $K_{S(T^*,UT)} == 1$ //利用 KS 检验判断 T^* 来自与其具有相同起点的不确定轨迹的密度分布.

then

$UDf^* = UDf^* + UT$

end if

$T^* = T^* - T^*(1)$ //去除第一个元素后的集合重新作为待检验对象.

end for

end while

2) Return UDf^*.

算法首先构建不确定轨迹集, 该过程中先基于聚类方法对原始轨迹的位置点进行约简, 得到约简后的轨迹点集, 时间复杂度为 $\mathrm{O}(Nkt)$, 其中 N 为数据集 D 的大小, 且 $k, t \ll N$; 而后构建不确定轨迹集, 该步骤需要遍历轨迹数据集, 时间复杂度为 $\mathrm{O}(nk^2)$, 其中 n 为轨迹个数, k 为聚簇个数; 再进一步计算表示每条不确定轨迹中轨迹点分布特征的概率密度函数, 从而得到密度式不确定轨迹集, 所需时间复杂度为 $\mathrm{O}(nk^2)$. 可见构建不确定轨迹集的时间复杂度基本为线性时间. 在轨迹预测阶段, 主要进行当前移动轨迹与数据集中所有具有相同起始点的密度式不确定轨迹关于分布是否一致的检验; 已知轨迹点总数为 N, 不确定轨迹总数为 k^2, 得不确定轨迹平均长度为 N/k^2, 则核密度估计函数的计算时间复杂度为 $\mathrm{O}(N/k^2)$, 若基于 KS 检验得到的具有相同起点的不确定轨迹个数为 m 以及待预测轨迹的长度为 t, 则预测算法的时间复杂度为 $\mathrm{O}(tm(N/k^2))$.

图 2 是轨迹预测示意图. 下面通过图 2 解释基于所提预测算法对当前用户的轨迹终点进行预测的过程.

假设轨迹起点为 $A(\boldsymbol{x}_0, \boldsymbol{y}_0)$, 而且已经获得轨迹数据集中的所有不确定轨迹及其概率密度函数. 若不确定轨迹集中包含不确定轨迹 (A, B, f_{AB}), (A, C, f_{AC}) 和 (A, D, f_{AD}), 首先可得由 A 可能到达的终点包括轨迹点 $\{B, C, D\}$; 随着用户移动行为的持续进行, 采集到移动轨迹数据流 $\{(\boldsymbol{x}_0, \boldsymbol{y}_0), (\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2)\}$, 根据当前的数据流, 经 KS 检验得到当前移动轨迹与不确定轨迹 (A, B, f_{AB}) 的密度分布并不一致的结论, 此时的候选终点集缩减为 $\{C, D\}$; 在采集到更丰富的轨迹数据集 $\{(\boldsymbol{x}_0, \boldsymbol{y}_0), (\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), (\boldsymbol{x}_3, \boldsymbol{y}_3), (\boldsymbol{x}_4, \boldsymbol{y}_4)\}$ 后, 经 KS 检验可得轨迹 (A, C, f_{AC}) 的密度分布与当前轨迹并不一致. 最终可判断 A 的可能终点为 D.

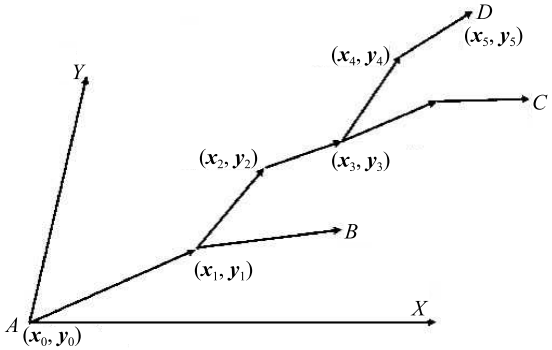


图 2　轨迹预测示意图

Fig. 2　Trajectory prediction

5　算法实施与性能分析

本文将从算法实施过程以及与几种轨迹预测方法间的结果对比两个角度, 验证所提算法的有效性与实用性.

5.1　算法实施

所提方法是一种轨迹终点预测方法, 通过构建模型、训练模型, 再经轨迹识别算法, 实现对轨迹终点的预测. 其中有两个关键问题需要解决: 一是并非任意两点间的轨迹均适合轨迹预测, 因此如何提取合理的轨迹, 使得轨迹能够真正描述一个移动行为, 作为轨迹模式识别的基础; 二是如何确定待预测轨迹数据流的长度, 如上所述随着待预测轨迹采集样本的增加可以提高轨迹识别的准确度, 但同时也会增加预测复杂性和预测周期, 所以确定合理的待预测轨迹数据流长度可以使预测方法更加高效.

实验数据选用了轨迹研究领域普遍使用的公开数据集, 由微软亚洲研究院采集的 GPS 轨迹数据 Geolife[19]. 该数据由 Geolife 记载了 5 年内 182 名用户的 GPS 坐标信息, 包括经纬度、海拔和时间点. 该数据集包含 17 621 个轨迹, 24 874 410 条位置信息, 总距离 1 292 951 千米, 总持续时间 50 176 小时. 这些轨迹由不同的 GPS 记录器记录.

为了提取合理的轨迹数据, 需要合理地确定轨迹的起始点与终点, 起始点与终点之间的连续位置点则是一条合理的轨迹. Geolife 轨迹数据是有明确的开始记录时间和结束记录时间的, 所以可以根据此信息确定轨迹. 另外, 在数据集中任意一条轨迹内还可能包含多条子轨迹, 这些子轨迹虽然不属于原始数据中实际意义上的完整轨迹, 却极可能成为分析其他用户移动行为的依据, 因此有必要提取这些子轨迹的信息. 例如对于用户中途的休息、吃饭等行为, 可能将原始轨迹拆分成几个部分. 因此为了获取更丰富的轨迹数据, 首先对移动轨迹中的停留时间的频次进行统计, 得到如图 3 所示的全部数据中所有用户轨迹的停留时间频次分布图.

图 3 中数据显示停留时间与其对应的频数在双对数坐标系中趋于线性, 属于典型的幂律分布; 在该数据集中一般的统计轨迹周期为 1 s 或 5 s, 超过 5 s 的可视为移动的停顿, 而当停留时间大于 10 s 时, 停留的频数明显下降. 因此本文将移动行为中停留时间超过 10 s 的停留点视为驻点, 若驻点属于某轨迹内的轨迹点, 则将其用于获取原有轨迹的子轨迹; 即根据驻点来定义轨迹预测中不确定轨迹的起点或终点; 从轨迹数据集中获取经过任意两个驻点的轨迹集合, 计算这些轨迹内包含的轨迹点的概率密度函数, 从而分析包含的轨迹点的分布特征, 并构建以选取驻点为起始点和终点的不确定轨迹模型.

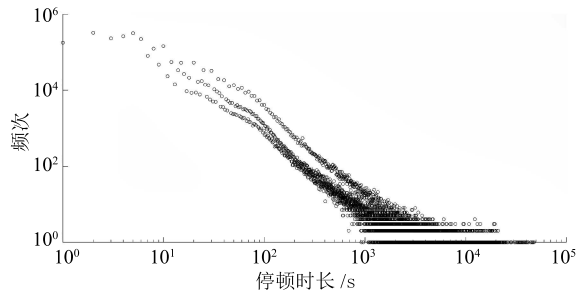


图 3 停留时间的频次分布图

Fig. 3 Frequency distribution of marking time

另一个关键是如何选取合理的轨迹数据流长度, 以提高预测的准确度和效率. 在理想情况下 (不确定轨迹接近典型理论分布), 根据预测算法的思想, 设原假设为待预测轨迹数据流来自于某不确定轨迹的分布. 若经过 KS 检验, 原假设被拒绝, 可排除该不确定轨迹的候选. 对于 KS 检验而言, 当样本数量较多时, 能够得到较为准确的检验结果, 并易于区分密度分布相近的不确定轨迹, 却同时影响预测的效率. 但当样本数量较少时, 假设检验的灵敏度则较差. 因此待预测对象的轨迹数据流的规模直接影响了预测结果. 所以本文尝试分析不确定轨迹密度分布在不同的显著水平下, 可供拒绝原假设的最大样本数量比. 显著性水平选取 0.01 ∼ 0.9, 在不同的显著性水平下, 原假设被拒绝的平均轨迹数据流长度统计如图 4 所示. 可见在显著性水平为 0.01、样本占总体平均 4 % 时即可拒绝不符合的不确定轨迹, 所需最大的数据流长度为 7 %; 而当显著性水平为 0.05 时, 最大所需长度为 4 %, 因此为了提高算法的准确性, 本文选择显著性水平为 0.05 以及样本规模为 7 % 的参数作为轨迹识别条件.

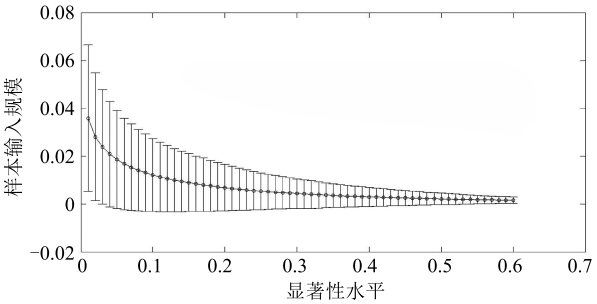


图 4 不同显著性水平下的有效样本规模

Fig. 4 Significance level of different data scale

然而在实际的轨迹数据分析中, 经常可以发现轨迹一般情况下并不接近任何典型理论分布. 如图 5 所示, 图中的不确定轨迹 u-t (Uncertain trajectory) 以及待预测子轨迹 s-t (Sub trajectory) 的经度和纬度的累积密度并不接近任何典型密度分布, 其最大累积误差也大于 KS 检验的指标值. 在这种情况下, KS 检验方法常常会较早的拒绝假设, 使得任何轨迹都无法获得匹配结果, 因此本文根据 KS 检验的方法, 以最大累积误差为指标值, 选择待预测轨迹与不确定轨迹进行匹配, 选取指标值最小的为最佳匹配对象, 进而预测轨迹终点.

根据上述分析, 本文将轨迹预测分为两种情况: 一是在 KS 检验过程中接受原假设的结果 (例如密度分布接近典型理论分布的情况) 进行直接预测; 二是在 KS 检验失效时, 将最小累积密度误差作为指标值, 选取误差最小的作为最佳的匹配对象进行间接预测. 通过两种方法的结合, 预测方法能够针对不同情况进行轨迹终点预测. 图 6 描述了 Geolife 轨迹数据集上预测的准确度分析结果, 可见在样本规模 40 % 时, 预测准确度即可达到 70 %, 并且随着待预测轨迹数据流规模的不断增加, 所提方法的预测准确度不断提高. 而且预测结果的方差较小, 说明所提预测方法不仅具有较高的预测精度, 而且具有较好的稳定性.

5.2 算法性能分析

为了说明本文所提方法的有效性, 本文在 Geo-

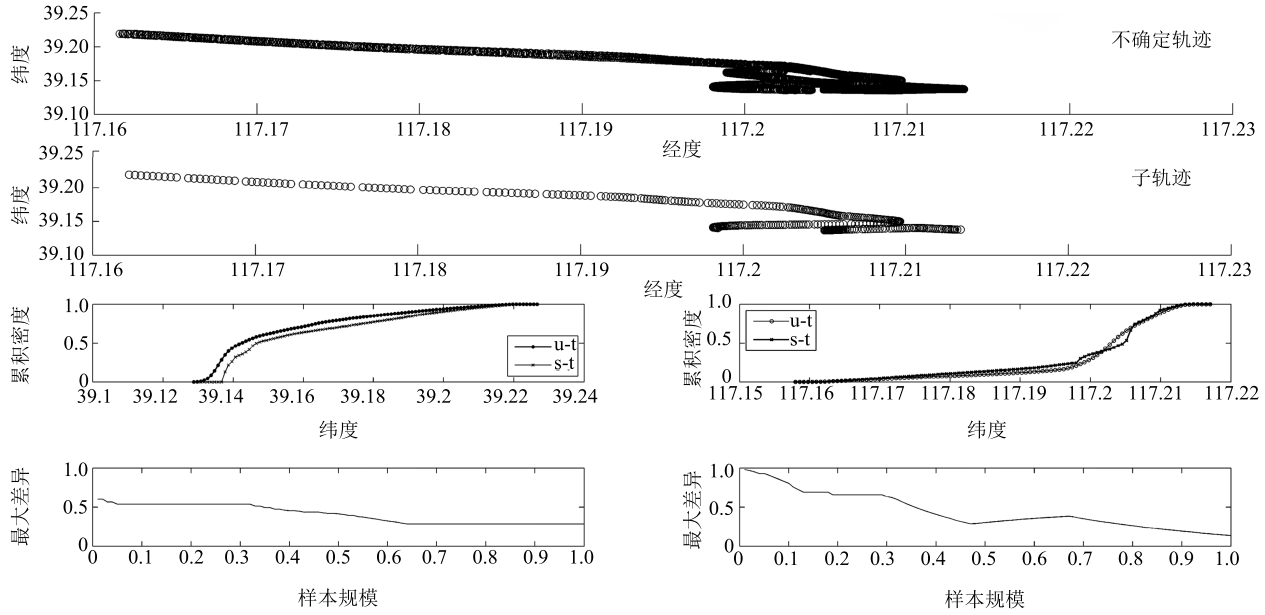


图 5　不确定轨迹预测的累计密度及其误差变化

Fig. 5　Accumulation density and error of uncertain trajectory prediction

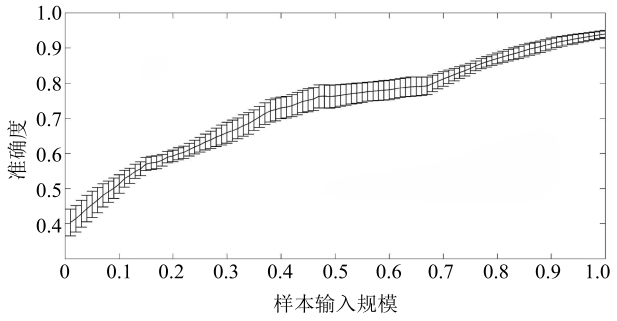


图 6　预测算法准确度分析

Fig. 6　Prediction method accuracy

life 和 T-Drive 数据集[19] 上, 与马尔科夫方法 (Markov-based methods, MBM)[14]、贝叶斯网络方法 (Bayesian network-based methods, BNM)[20]、回归方法 (Regression-based methods, RBM)[21] 和神经网络方法 (Neural network-based methods, NNM)[21] 从预测准确性与方法运行效率两方面进行对比. 为了保证对比结果的公平性, 对比算法的参数选择参照对应文献中效果最优的参数. 基于 Markov 的方法根据文献 [14] 中算法的执行分析采用 3 阶 Markov 模型. 基于回归的方法根据文献 [2] 采用基于高斯核的回归模型. 基于神经网络的方法采用文献 [21] 中基于多层感知机的网络模型 MLP, 网络模型中的隐层数量为 5, 训练模型时的学习率、冲量因子以及学习次数分别为 0.3、0.2 和 500 次.

为了验证算法在不同数据条件下的准确性, 在 Geolife 和 T-Drive 数据集上进行算法性能的对比分析. 验证采用十折交叉验证法, 将数据按 70 : 30 的比例划分, 其中 70 % 作为训练数据集, 30 % 作为测试数据集, 并以准确度为指标评价预测的准确性. 但这 4 种预测方法一般都是短期预测, 而本文方法是直接预测轨迹终点的长期预测; 为了能够统一尺度, 对于测试数据, 分别获取待预测轨迹的 30 %、60 % 和 90 % 作为输入数据, 进而对于 3 种不同长度的待预测轨迹数据, 对比各算法的预测准确性.

表 1 和表 2 分别记录了两个实验数据集上进行交叉验证的几种算法的预测结果准确性, 进行十折交叉验证的实验次数为 120 次, 由于篇幅限制仅列出部分结果. 从表中数据可以看出, 所提算法在测试数据集上均能获得较高的准确性, 而且随着待预测轨迹数据的丰富, 即待预测轨迹的 60 % 和 90 % 作为输入数据时, 所得准确性相对更高, 说明随待预测轨迹数据规模的增加, UDTM 的预测准确性也能逐渐提高.

为了更客观地比较几种算法的预测效果, 将表 1 和表 2 中的数据进行统计分析, 获取每种情况下所得准确性的均值与方差, 对比结果如图 7 和图 8 所示. 在预测过程中, 输入轨迹数量为 30 % 时, 对比算法的准确性只有 50 % 左右, 而 UDTM 可以达到 70 % 左右, 当输入轨迹数量为 60 % 时, UDTM 的预测准确度则可达到 0.8 ∼ 0.9, 并且 UDTM 预测准确性的方差为 ±0.0173, 说明 UDTM 方法在十折交叉验证中具有较好的稳定性. 相比其他算法, UDTM 算法在 T-Drive 和 Geolife 数据集中都具有较好的准确度, 并且在 T-Drive 数据集上的准确度更

表 1 Geolife 数据集上各算法的预测准确性对比
Table 1 Prediction accuracy comparison of several methods on Geolife

样本规模	MBM	BNM	RBM	NNM	UDTM
	0.496	0.51	0.434	0.552	0.671
	0.515	0.511	0.491	0.549	0.663
	0.506	0.489	0.495	0.548	0.653
30 %	···	···	···	···	···
	0.508	0.524	0.467	0.561	0.660
	0.523	0.498	0.426	0.548	0.673
	0.502	0.495	0.464	0.547	0.674
	0.674	0.630	0.618	0.694	0.860
	0.652	0.634	0.633	0.716	0.827
	0.654	0.660	0.665	0.749	0.855
60 %	···	···	···	···	···
	0.696	0.643	0.585	0.729	0.847
	0.687	0.632	0.627	0.717	0.861
	0.650	0.654	0.644	0.732	0.861
	0.793	0.749	0.761	0.861	0.916
	0.807	0.745	0.729	0.861	0.897
	0.794	0.800	0.750	0.861	0.900
90 %	···	···	···	···	···
	0.799	0.784	0.771	0.863	0.890
	0.775	0.780	0.706	0.860	0.894
	0.802	0.767	0.771	0.839	0.900

表 2 T-Drive 数据集上各算法的预测准确性对比
Table 2 Prediction accuracy comparison of several methods on T-Drive

样本规模	MBM	BNM	RBM	NNM	UDTM
	0.519	0.495	0.49	0.593	0.705
	0.511	0.513	0.444	0.596	0.699
	0.477	0.512	0.482	0.579	0.708
30 %	···	···	···	···	···
	0.520	0.510	0.467	0.586	0.719
	0.535	0.506	0.488	0.601	0.721
	0.521	0.505	0.480	0.594	0.702
	0.691	0.659	0.620	0.780	0.898
	0.680	0.688	0.683	0.747	0.895
	0.673	0.675	0.634	0.767	0.889
60 %	···	···	···	···	···
	0.685	0.675	0.654	0.793	0.897
	0.675	0.681	0.660	0.772	0.879
	0.680	0.644	0.607	0.761	0.902
	0.841	0.798	0.751	0.915	0.969
	0.805	0.779	0.761	0.879	0.944
	0.857	0.808	0.777	0.910	0.948
90 %	···	···	···	···	···
	0.839	0.823	0.694	0.893	0.963
	0.790	0.797	0.721	0.901	0.961
	0.804	0.786	0.740	0.888	0.961

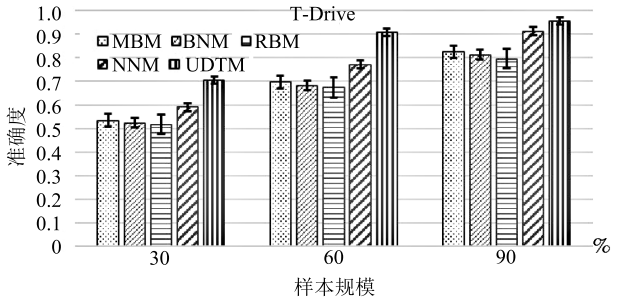


图 7 T-Drive 数据集的准确性验证
Fig. 7 Accuracy verification on T-Drive

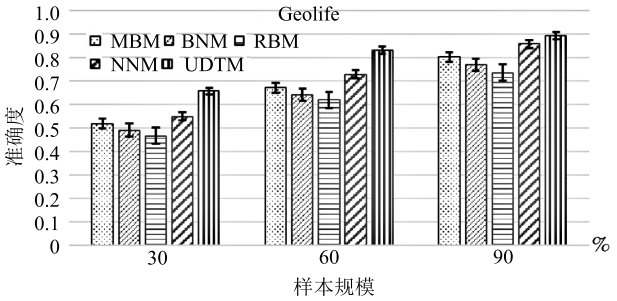


图 8 Geolife 数据集的准确性验证
Fig. 8 Accuracy verification on Geolife

高, 其原因是 Geolife 的样本都是个人的移动轨迹, 相比 T-Drive 中的汽车轨迹具有更强的随机性, 因此预测准确性略低.

时间复杂性是轨迹预测算法执行的关键指标, 好的预测算法应在具有较高预测准确度的同时兼具较快的响应速度. 根据 UDTM 方法的模型定义, UDTM 方法不是完全根据样本的个数, 而是根据样本的聚集情况估计核函数的数量和权重确定模型的核函数的个数, 由于模型的这种设计, 使得算法在执行时具有较高的效率. 图 9 描述了几种对比算法的预测时间, UDTM 虽然并未具有最快的预测时间, 如图 9 所示随着输入轨迹数据流的增多, UDTM 的时间也并未显著增长. 其中 RBM 与 NNM 由于需要训练模型优化参数, 使其在建模时时间耗费较多; 而 UDTM 只需要确定非参数模型中不同核的权重, 但为了保证预测精度, 需要首先对不确定轨迹进行

建模, 仍具有一定时间复杂度, 因此相对于 BNM 和 MBM 而言需要的预测时间较长. 另外在预测阶段, 由于 RBM 需要将轨迹中的位置点进行回归计算, 所以当轨迹数量增加时, 预测所需时间会呈指数型增长, 而 MBM、BNM、NNM 以及 UDTM 由于在预测阶段都是一次性计算, 且模型本身不会更改, 所以预测时间并不会随轨迹数量的增长而显著增加.

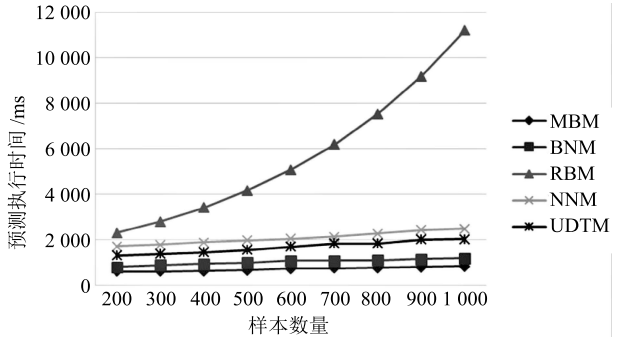


图 9　算法执行效率比较

Fig. 9　Algorithms execution efficiency comparison

虽然所提轨迹预测算法 UDTM 并非具有最快的预测效率, 但并未较其他算法相差过多, 并且也并不会随着输入数据的增多使得预测时间显著增长, 因此 UDTM 仍然具有较高的效率. 另外即使在用户轨迹数据流并不多的情况下, UDTM 相较于其他算法还能够较准确地预测待分析轨迹的终点, 即在准确性方面具有较强的优势.

6　结论

针对移动对象轨迹的随机性与不确定性问题, 本文提出了一种基于非参数统计方法的不确定轨迹预测算法. 首先利用核密度估计方法对轨迹数据集中的不确定轨迹进行建模, 通过概率密度函数表示其不确定性的分布特征, 基于非参估计的方法能够获得较客观的符合数据实际分布的不确定轨迹分布情况; 然后利用 KS 检验的方法分析输入轨迹与已知的不确定轨迹分布间的匹配关系, 从而根据相匹配的不确定轨迹预测输入轨迹的目标终点. 所提方法充分考虑轨迹中体现的移动用户行为的不确定性, 对历史轨迹中任意可能的位置进行计算, 因此在实际预测问题中, 能够获得较好的建模和预测能力, 在真实数据集上的实验也验证了所提方法的有效性和可靠性.

References

1 Song C M, Qu Z H, Blumm N, Barabási A L. Limits of predictability in human mobility. *Science*, 2010, **327**(5968): 1018−1021

2 Mamoulis N, Cao H P, Kollios G, Hadjieleftheriou M, Tao Y F, Cheung D W. Mining, indexing, and querying historical spatiotemporal data. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, USA: ACM, 2004. 236−245

3 Morzy M. Mining frequent trajectories of moving objects for location prediction. In: Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition. Leipzig, Germany: Springer, 2007. 667−680

4 Jeung H, Liu Q, Shen H T, Zhou X F. A hybrid prediction model for moving objects. In: Proceedings of the 24th International Conference on Data Engineering. Cancun, Mexico: IEEE, 2008. 70−79

5 Ying J J C, Lee W C, Weng T C, Tseng V S. Semantic trajectory mining for location prediction. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Chicago, USA: ACM, 2011. 34−43

6 Zheng Y, Zhang L Z, Xie X, Alma W Y. Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain: ACM, 2009. 791−800

7 Qiao S J, Shen D Y, Wang X T, Han N, Zhu W. A self-adaptive parameter selection trajectory prediction approach via hidden Markov models. *IEEE Transactions on Intelligent Transportation Systems*, 2015, **16**(1): 284−296

8 De Brébisson A, Simon É, Auvolat A, Vincent P, Bengio Y. Artificial neural networks applied to taxi destination prediction. In: Proceedings of the 2015 International Conference on ECML PKDD Discovery Challenge. Aachen, Germany: CEUR-WS.org, 2015. 40−51

9 Krumm J, Horvitz E. Predestination: inferring destinations from partial trajectories. In: Proceedings of the 8th International Conference on Ubiquitous Computing. Orange County, USA: Springer, 2006. 243−260

10 Ziebart B D, Maas A L, Dey A K, Bagnell J A. Navigate like a cabbie: probabilistic reasoning from observed context-aware behavior. In: Proceedings of the 10th International Conference on Ubiquitous computing. Seoul, Korea: ACM, 2008. 322−331

11 Patterson D J, Liao L, Fox D, Kautz H. Inferring high-level behavior from low-level sensors. In: Proceedings of the 5th International Conference on Ubiquitous Computing. Seattle, WA, USA: Springer, 2003. 73−89

12 Monreale A, Pinelli F, Trasarti R, Giannotti F. Wherenext: a location predictor on trajectory pattern mining. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France: ACM, 2009. 637−646

13 Ashbrook D, Starner T. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 2003, **7**(5): 275−286

14 Gambs S, Killijian M O, Del M N, Cortez P. Next place prediction using mobility markov chains. In: Proceedings of the 1st Workshop on Measurement, Privacy, and Mobility. Bern, Switzerland: ACM, 2012. Article No.3

15 Qiao Shao-Jie, Jin Kun, Han Nan, Tang Chang-Jie, Gesangduoji, Gutierrez L A. Trajectory prediction algorithm based on Gaussian mixture model. *Journal of Software*, 2015, **26**(5): 1048−1063
(乔少杰, 金琨, 韩楠, 唐常杰, 格桑多吉, Gutierrez L A. 一种基于高斯混合模型的轨迹预测算法. 软件学报, 2015, **26**(5): 1048−1063)

16 Besse P C, Guillouet B, Loubes J M, Royer F. Destination prediction by trajectory distribution based model. *IEEE Transactions on Intelligent Transportation Systems*, 2018, **19**(8): 2470−2481

17 Willard K E, Connelly D P. Nonparametric probability density estimation: improvements to the histogram for laboratory data. *Computers and Biomedical Research*, 1992, **25**(1): 17−28

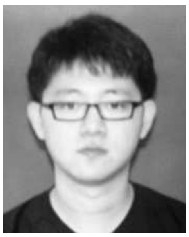
18 Wang Xing, Chu Ting-Jin. *Non-parametric Statistics* (2nd Edition). Beijing: Tsinghua press, 2014. 361
(王星, 褚挺进. 非参数统计 (第 2 版). 北京: 清华大学出版社, 2014. 361.)

19 Yuan J, Zheng Y, Xie X, Sun G Z. Driving with knowledge from the physical world. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2011. 316−324

20 Ferrer G, Sanfeliu A. Bayesian human motion intentionality prediction in urban environments. *Pattern Recognition Letters*, 2014, **44**: 134−140

21 Bui D T, Tuan T A, Klempe H, Pradhan B, Revhaug I. Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, 2016, **13**(2): 361−378

程 媛 哈尔滨理工大学计算机科学与技术学院讲师. 主要研究方向为数据挖掘, 不确定性研究.
E-mail: changuang7@sina.com
(**CHENG Yuan** Lecturer at the College of Computer Science and Technology, Harbin University of Science and Technology. Her research interest covers data mining and uncertainty research.)



迟荣华 哈尔滨工程大学计算机科学与技术学院博士研究生. 主要研究方向为机器学习, 不确定性研究. 本文通信作者.
E-mail: chironghua@126.com
(**CHI Rong-Hua** Ph. D. candidate at the College of Computer Science and Technology, Harbin Engineering University. His research interest covers machine learning and uncertainty research. Corresponding author of this paper.)

黄少滨 哈尔滨工程大学计算机科学与技术学院教授. 主要研究方向为分布式计算与仿真, 模型检测, 数据集成.
E-mail: huangshaobin@hrbeu.edu.cn
(**HUANG Shao-Bin** Professor at the College of Computer Science and Technology, Harbin Engineering University. His research interest covers distributed computing and simulation, model checking, and data integration.)

吕天阳 审计署计算机技术中心国家仿真实验室高级工程师. 主要研究方向为复杂网络, 计算机审计.
E-mail: raynor1979@163.com
(**LV Tian-Yang** Senior engineer at the National Audit Simulation Laboratory of IT Center, National Audit Office. His research interest covers complex network and computer-aided audit.)