软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(5):1048−1063 [doi: 10.13328/j.cnki.jos.004796]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种基于高斯混合模型的轨迹预测算法*

乔少杰[1], 金 琨[1], 韩 楠[2], 唐常杰[3], 格桑多吉[4], Louis Alberto GUTIERREZ[5]

[1](西南交通大学 信息科学与技术学院,四川 成都 610031)

[2](西南交通大学 生命科学与工程学院,四川 成都 610031)

[3](四川大学 计算机学院,四川 成都 610065)

[4](西藏大学 藏文信息技术研究中心,西藏 拉萨 850000)

[5](Department of Computer Science, Rensselaer Polytechnic Institute, New York, USA)

通讯作者: 韩楠, E-mail: hannan@swjtu.edu.cn

摘 要: 在智能交通控制系统、军事数字化战场、辅助驾驶系统中,实时、精确、可靠的移动对象不确定性轨迹预测具有极高的应用价值.智能轨迹预测不仅可以提供精准的基于位置的服务,而且可以提前监测和预判交通状况,进而推荐最佳路线,已经成为移动对象数据库研究的热点,亟需设计准确而高效的位置预测方法.针对现有方法的不足,提出了基于高斯混合模型的轨迹预测方法 GMTP,主要步骤包括:(1) 针对复杂运动模式利用高斯混合模型建模;(2) 利用高斯混合模型计算不同运动模式的概率分布,进而将轨迹数据划分为不同分量;(3) 利用高斯过程回归预测移动对象最可能的运动轨迹.GMTP 是高斯非线性概率统计模型,其优势在于:计算结果不仅是位置预测值,更是关于移动对象未来所有可能运动轨迹的概率分布,可以利用概率统计分布特性获得某种运动模式(如匀加速运动)下的位置预测.大量真实轨迹数据集上的实验结果表明:与相同参数设置下的高斯回归预测和卡尔曼滤波预测法相比,GMTP 的预测准确性平均提高了 22.2%和 23.8%,预测时间平均缩减了 92.7%和 95.9%.

关键词: 移动对象数据库;轨迹预测;高斯混合模型;运动模式

中图法分类号: TP18

中文引用格式: 乔少杰,金琨,韩楠,唐常杰,格桑多吉,Gutierrez LA.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. http://www.jos.org.cn/1000-9825/4796.htm

英文引用格式: Qiao SJ, Jin K, Han N, Tang CJ, Gesangduoji, Gutierrez LA. Trajectory prediction algorithm based on Gaussian mixture model. Ruan Jian Xue Bao/Journal of Software, 2015,26(5):1048−1063 (in Chinese). http://www.jos.org.cn/1000-9825/4796.htm

Trajectory Prediction Algorithm Based on Gaussian Mixture Model

QIAO Shao-Jie[1], JIN Kun[1], HAN Nan[2], TANG Chang-Jie[3], Gesangduoji[4], Louis Alberto GUTIERREZ[5]

[1](School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

[2](School of Life Science and Engineering, Southwest Jiaotong University, Chengdu 610031, China)

[3](College of Computer Science, Sichuan University, Chengdu 610065, China)

[4](Tibetan Information Technology Research Center, Tibet University, Lasa 850000, China)

[5](Department of Computer Science, Rensselaer Polytechnic Institute, New York, USA)

* 基金项目: 国家自然科学基金(61100045, 61165013); 教育部高等学校博士学科点专项科研基金(20110184120008); 教育部人文社会科学研究青年基金(14YJCZH046); 中央高校基本科研业务费专项资金(2682013BR023); 科学计算与智能信息处理广西高校重点实验室开放课题(GXSCIIP201407)

收稿时间: 2014-07-10; 修改时间: 2014-09-29; 定稿时间: 2014-11-25; jos 在线出版时间: 2015-02-02

CNKI 网络优先出版: 2015-02-02 15:24, http://www.cnki.net/kcms/detail/11.2560.TP.20150202.1524.005.html

Abstract: For intelligent transportation systems, digital military battlefield and driver assistance systems, it is of great practical value to predict the trajectories of moving objects with uncertainty in a real-time, accurate and reliable fashion. Intelligent trajectory prediction can not only provide accurate location-based services, but also monitor and estimate traffic to suggest the best path, and as such becomes an active research direction. Aiming to overcome the drawbacks of the existing methods, a new trajectory prediction model based on Gaussian mixture models called GMTP is proposed. The new model contains the following essential phases: (1) modeling the complex motion patterns based on Gaussian mixture models, (2) calculating the probability distribution of different types of motion patterns by using Gaussian mixture model in order to partition trajectory data into distinct components, and (3) inferring the most possible trajectories of moving objects via Gaussian process regression. The GMTP algorithm is naturally a Gaussian nonlinear statistical probability model and the advantage of the proposed model is that the result is not only a predicted value, but also a whole distribution beyond the future trajectories, therefore making it possible to infer the location in regard to some motion patterns, e.g., uniformly accelerated motion, by using statistical probability distribution. Extensive experiments are conducted on real trajectory data sets and the results show that the prediction accuracy of the GMTP algorithm is improved by 22.2% and 23.8%, and the runtime can be reduced by 92.7% and 95.9% on average, respectively, when compared to the Gaussian process regression model and Kalman filter prediction algorithm with similar parameter setting.

Key words: moving objects database; trajectory prediction; Gaussian mixture model; motion pattern

随着无线移动通信设备和无线车载传感器的发展,获取移动对象时空位置的手段更加多样化.在诸多重要应用领域中,如智能交通系统(ITS)、智能导航、数字化战场、物流配送、移动电子商务等,用户都需要及时查询和分析各种移动对象的轨迹位置信息,移动对象轨迹位置预测已经逐渐成为移动对象数据管理[1]中极为重要的研究方向.当前,数字化城市的公共交通车辆、出租车以及其他一些车辆都配备 GPS 及车载导航设备,可以将不同时刻采集到的车辆位置信息连接起来,构成完整的轨迹时间序列,进而挖掘其动态运动模式.对移动对象的轨迹信息挖掘的目的在于:在短时间内,可以提醒处于拥堵十字路口的驾驶员车辆前行的安全性;在长时间内,预测可能会发生交通堵塞的区域,及时做出调度,引导交通并提醒驾驶员及时做出路线调整.在一般情况下,移动对象总是周期性地向中央服务器发送其位置信息,然而在两次定位信息传递时间内,移动对象的具体位置信息和运动轨迹是无从得知的.此外,移动对象运动环境的多样化也使问题更加复杂.如何对移动对象位置信息准确预测,成为亟需解决的难点.当前已有一些研究成果,如移动对象的聚类、异常检测以及定位和运动趋势的预测.其中,对移动对象位置预测技术不断地完善,但由于理论和技术的不成熟,大多数模型不能很好地适应不断发展的移动计算技术的需要.诸如单一线性回归预测、神经网络预测、Markov 模型位置预测[2,3]等传统方法,因其理论的局限性,未能应用于对实时性、抗干扰性要求较高的移动通信环境中.

1 相关工作

移动对象按照空间分布特性被分为 3 类:(1) 运动道路不受限制,如飞机、轮船、带有传感器的鸽子等;(2) 运动轨道受限制,如公路行驶的车辆;(3) 运动道路轨迹分散,如移动用户.针对移动对象的位置预测主要分为过去轨迹的重现、当前和未来轨迹的预测.本文主要研究对象是运动道路轨迹分散的移动对象,同时解决该类移动对象的未来轨迹预测问题.当前,轨迹预测方法主要集中于轨迹频繁模式挖掘,通过挖掘频繁模式找出典型运动模式[4].代表性工作由 Morzy 等人[5]提出,提出一种结合前缀树 PrefixSpan 和频繁模式挖掘 FP-tree 算法挖掘移动对象动态运动规则,但是构建前缀树和FP-tree的时间代价较高.Jeung 等人[6]提出了一种综合考虑移动对象运动模式和运动函数的混合预测方法,然而提供的查询预测仅支持短期的查询结果.大多数轨迹预测方法基于轨迹的地理特性,而 Ying 等人[7]结合个体运动轨迹的语义特征和空间位置预测移动对象下一位置信息.这一方法的不足在于计算每条候选路径的 Semantic Score 时,代价较高.Zheng 等人[8]考虑个体的旅行经历和兴趣爱好,提出了一种 Hypertext Induced Topic Search 模型推测其感兴趣的运动路线,但这一方法主要应用于向用户推荐可能感兴趣的地点,不能给出一整条完整的运动曲线.本文的研究动机源于 Song 等人[9]在 Science 上发表的介绍预测人类移动性的工作,通过测量个体轨迹的信息熵,定量地给出了人类动态运动行为,具有 93%的可预测性.围绕这一工作,近期,Pan 等人[10]提出了基于多变元正态分布的最佳线性预测器,这一方法的不足在于预测会产生

延迟,不能应用于交通流的实时监控.Zhou 等人[11]提出了一种称为“semi-lazy”的预测方法,利用动态选取的参考轨迹构建预测模型,优点是可以基于少量的参考轨迹构建精准的模型.近期,针对受限路网中轨迹预测精度不高的问题,Qiao 等人[3]提出了一种基于隐马尔卡夫模型的轨迹预测算法 HMTP,从海量轨迹数据中提取隐状态和观察状态,根据不同类型的轨迹,自适应地预测最佳轨迹.与本文的工作不同点在于:本文提出了一种通用的轨迹预测算法,不仅仅适用于受限路网中移动对象的轨迹预测问题.

大多数轨迹预测问题针对的都是受限路网中移动对象的位置查询和预测[12].Tao 等人[13]提出一种基于未知模型的预测方法,弥补了线性预测的缺陷,将轨迹数据存储在 TPR*树中,使用递归函数进行预测.然而,这一方法忽略了主、客观不确定性因素的影响.Qiao 等人[14]充分考虑了影响移动对象位置变化的运动速度、方向和路况信息,设计了一种基于时间连续贝叶斯网络的连续轨迹序列预测算法.实验证明,这一方法在轨迹预测的准确性和时效性上均优于朴素预测算法.为了进一步提高预测结果的准确性和时效性,文献[12]提出了一种基于轨迹时间连续贝叶斯网络的轨迹预测算法 TPMO,考虑了移动速度和方向对移动对象动态运动行为的影响.与本文工作的不同点在于:本文所提出的预测算法可以解决轨迹离散状态预测准确性较差的问题,并且对于移动对象轨迹的预测,可以依据概率模型精确度量其预测误差.

从时空轨迹数据中挖掘运动模式的多样性,对移动对象位置预测也是至关重要的.轨迹运动模式建模需要对状态空间进行离散化分析,Hu 等人[15]将轨迹模型视为离散状态点之间转变的过渡,离散状态分析方法的不足在于:需要对大量时空数据进行离散化处理,并且需要分析离散数据点之间的关联.Feng 等人[16]基于隐马尔可夫模型重现移动对象轨迹序列,这一方法无需考虑对象缺失的观察状态信息,但是预测准确率不高.Gaffney 和 Smyth 提出对原始连续轨迹基于原型的聚类方法[17],通过增加高斯噪声,将具有相似轨迹部分的移动对象聚合在一起,使用最大似然原理实现无监督学习.这是一种比较新颖和实用的轨迹数据分析方法,在本文的后续工作中,借鉴其利用 EM 算法来确定轨迹聚类的簇个数,并考虑了聚类簇中轨迹点离散空间和时间偏移.轨迹位置预测中,稳定线性回归如卡尔曼滤波对短时间内(1 步或 2 步)的预测有比较稳定准确的判断[18],而对于长时间(未来 5 步以上)的预测,仅当处理的轨迹数据是无噪声点的情况才比较有效.高斯过程回归方法对于处理具有噪声点的数据有着比较好的预测效果,是一种基于非参数的概率性方法,而且所使用的训练数据集规模比较小[19].Qiao 等人[20]对轨迹数据构建双层索引,利用 RoI(region-of-interest)检查算法对轨迹进行划分,构建频繁轨迹模式树预测移动对象最可能运动模式.针对上述研究的不足,本文提出了基于高斯混合模型(Gaussian mixture model,简称 GMM)的轨迹建模方法,利用概率模型刻画运动轨迹.其优势在于:可以摆脱轨迹离散状态分析方法的弊端,并且对于移动对象轨迹的预测,可以依据概率模型精确度量其预测误差.

2 基本概念及模型框架

本节首先给出几个主要概念,然后介绍本文所提出模型的框架及工作原理.

已知移动对象数据库 D,其中存储大量运动对象在不同时间采样点的位置信息,位置点在时间上的有序集合称为轨迹,用 $D=\{Trj_1,Trj_2,...,Trj_n\}$表示,轨迹的数量定义为$|D|$.本文对轨迹作如下定义:

定义 1(轨迹序列). 移动对象原始轨迹序列 $Trj=\{s_1,s_2,...,s_d\}$表示有序的离散轨迹点$\{s_i=(x_i,y_i,t_i)|1\leqslant i\leqslant d\}$所构成的序列,其中,$t_i$表示时间戳,$i\in[1,d)$,$t_i<t_{i+1}$,$(x_i,y_i)$表示移动对象的 2 维空间坐标.

定义 2(轨迹矢量集). 对欧式空间 2 维平面 X 轴和 Y 轴方向进行建模,利用两个方向上的轨迹矢量表示轨迹数据:

$$D=\{Trj_1,Trj_2,...,Trj_n\}=\{(s_x^1,s_y^1),(s_x^2,s_y^2),...,(s_x^n,s_y^n)\}=\{(s_x^1,s_x^2,...,s_x^n)^T,(s_y^1,s_y^2,...,s_y^n)^T\}=\{X,Y\},$$

其中,$s_x^i=(x_{i1},x_{i2},...,x_{id})^T$表示第 i 条轨迹在 X 方向上的投影矢量集,$s_y^i=(y_{i1},y_{i2},...,y_{id})^T$表示第 i 条轨迹在 Y 方向上的投影矢量集,(s_x^i,s_y^i)称为轨迹 Trj_i 的矢量集,$\{X,Y\}$称为轨迹数据集 D 上的轨迹矢量集.

定义 3(高斯过程回归,Gaussian processes regression). 对于输入训练数据集 $D=\{(x_i,y_i)\}_{i=1}^N=(X,y)$,其中,$x_i\in R^d$ 为 d 维输入矢量,$X=[x_1,x_2,\ldots,x_n]$为 $d\times n$ 维输入矩阵,$y_i\in R$ 为相应的输出标量.已知输入数据集合 X,可构成一个随机变量集合$\{f(x_1),f(x_2),\ldots,f(x_N)\}$,且具有联合高斯分布.该高斯过程全部统计数字特征可由均值函数 $m(x)$

和协方差函数 $k(x,x')$确定,即

$$f(x)\sim GP(m(x),k(x,x')) \tag{1}$$

其中,$m(x)=E[f(x)]$,$k(x,x')=E[(f(x)-m(x))(f(x')-m(x'))]$.本文采用的核函数为标准指数协方差函数,即

$$Cov(f(x),f(x'))=k(x,x')=\theta_0^2\exp\left(-\frac{(x-x')^2}{2\theta_1^2}\right)+\sigma^2\delta_{ij} \tag{2}$$

其中,θ_0 为指数权重,θ_1 表示长度规模,δ_{ij} 是狄拉克函数.当 $i=j$ 时,函数 $\delta_{ij}=1$,否则为 0.

定义 4(轨迹高斯混合模型). 轨迹高斯混合模型中,每个状态用一个高斯过程函数表示,即,数据是由多个高斯过程模型线性组合产生的,用公式(3)、公式(4)来表示轨迹点在二维坐标系下的 GMM 模型.假设数据点(x,y)是多维运动矢量,其随机生成来自 M 个相互独立的高斯过程线性组合的总体 $G=\{GP_1,GP_2,...,GP_M\}$,且每个高斯过程对应的权重分为为$\omega_1,\omega_2,...,\omega_M$,混合而成的总体分布为 G,于是,x 的概率密度函数为

$$p(x\mid\lambda)=\sum_{j=1}^{M}\omega_j GP(x\mid\mu_j,\Sigma_j) \tag{3}$$

$$p(y\mid\lambda)=\sum_{j=1}^{M}\omega_j GP(y\mid\mu_i,\Sigma_i) \tag{4}$$

其中,$GP(\cdot)$表示高斯过程的概率密度函数,M 表示高斯过程数量,ω_j 为第 j 个高斯过程的权重,且 $\sum_{j=1}^{M}\omega_j=1$. μ表示此密度函数中心点,Σ表示协方差矩阵.参数用集合表示为$\lambda=\{\omega_i,\mu_i,\Sigma_i\}$,$i=1,\ldots,M$.

定义 5(预测误差). 轨迹预测时,将测试轨迹数据集输入预测模型得到预测输出轨迹,其中,输入测试轨迹为部分轨迹点的集合,预测输出点由图 1 中虚线表示.

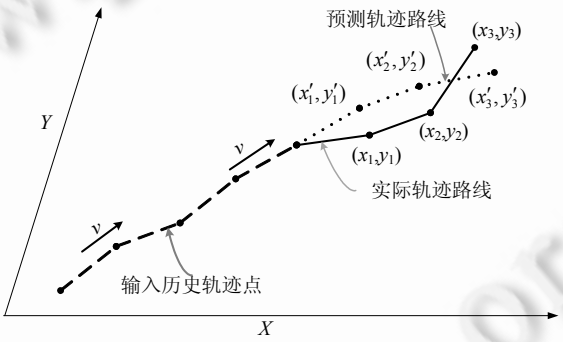


Fig.1 An example of predicted trajectories

图 1 预测轨迹示例

采用公式(5)所示的均方根误差 RMSE 来计算预测轨迹点与实际轨迹点的几何空间误差:

$$RMSE=\frac{\sum_{i=1}^{k}\sqrt{(x_i'-x_i)^2+(y_i'-y_i)^2}}{k} \tag{5}$$

其中,(x_i,y_i)表示真实位置,(x_i',y_i') 表示预测位置,k 为预测轨迹点的数量.

本文所提出模型的系统框架如图 2 所示,其工作原理分为 3 个步骤:

(1) 利用 ETL 技术将历史轨迹预处理后转化为轨迹矢量存储于数据库中;

(2) 对不同运动模式轨迹数据进行 GMM 聚类分析,利用最大似然估计 EM 算法求得聚类模型参数,使其基于历史数据模型概率达到最大化,获得 M 个聚簇;

(3) 利用最小二乘法和高斯混合回归模型训练得到预测模型 GMTP,根据输入的新轨迹数据预测移动对象未来最可能的运动轨迹.

本文第 3 节形式化给出移动对象简单和复杂轨迹运动概率模型.第 4 节给出高斯混合模型回归预测理论基础.第 5 节详细阐述 GMM 轨迹预测模型训练和学习原理.第 6 节介绍基于 GMM 的轨迹预测算法.第 7 节通过

对比实验检验所提算法的性能优势.第 8 节总结全文并对未来工作进行展望.

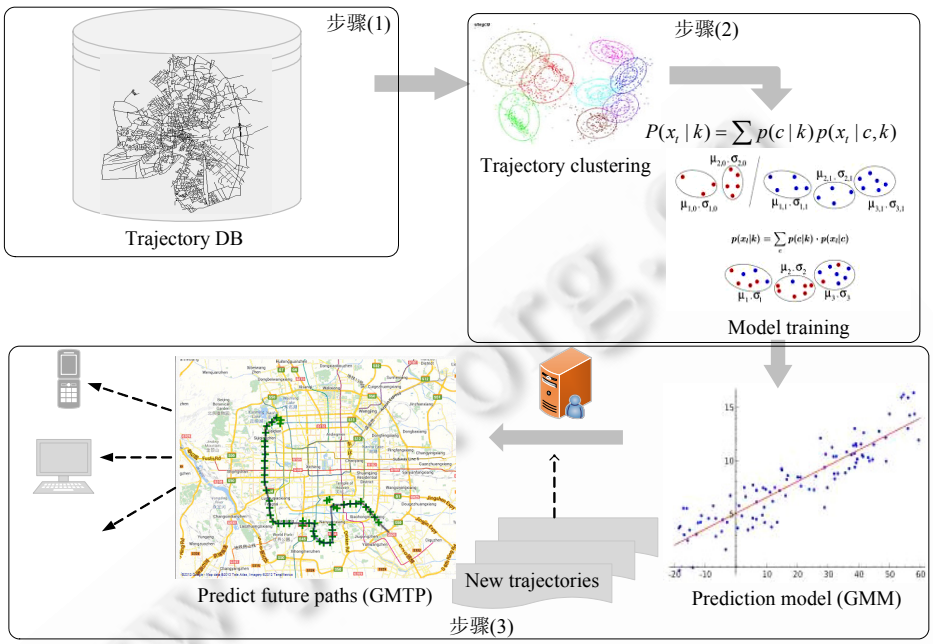


Fig.2 Trajectory prediction schema for moving objects based on GMM

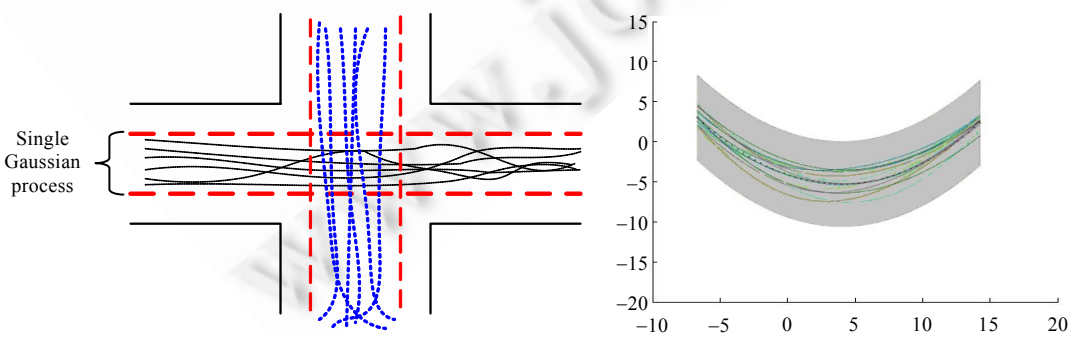
图 2 基于 GMM 的移动对象轨迹预测框架

3 运动模式概率模型

本文将单一运动模式利用高斯过程 GP 表示,而复杂场景中多种运动模式利用高斯混合模型 GMM 建模.

3.1 简单轨迹运动模式

在简单轨迹运动模式场景中,许多轨迹具有相同运动模式,可以用一个高斯过程 GP 表示.可以认为移动物体在 x 和 y 方向上变化是相互独立的,一条轨迹需要两个高斯过程(x 和 y 方向)来表示.图 3(a)中点连接的虚线表示利用高斯过程均值表示的典型运动模式,宽虚线内部分为用协方差刻画的不确定性运动波动范围.如图 3(b)所示,不同种类实线代表不同轨迹,这些轨迹运动模式比较相近,可由单一高斯过程来表达.



(a) 单一轨迹运动模式 (b) 单一运动模式 GP 刻画

Fig.3 Pattern modeling for single motion patterns based on GP

图 3 单一运动模式 GP 建模

本文将定义 1 中的一条轨迹划分为两个方向上的 d 维矢量 $X=(x_1,x_2,...,x_d)^T$ 和 $Y=(y_1,y_2,...,y_d)^T$,d 表示轨迹观测点个数.对于 N 条具有相同运动模式的轨迹,利用高斯过程建模,典型运动模式的轨迹概率模型为

$$P(X)=\prod_{n=1}^{N}GP(x_n\mid\mu_x,\Sigma_x) \tag{6}$$

$$P(Y)=\prod_{n=1}^{N}GP(y_n\mid\mu_y,\Sigma_y) \tag{7}$$

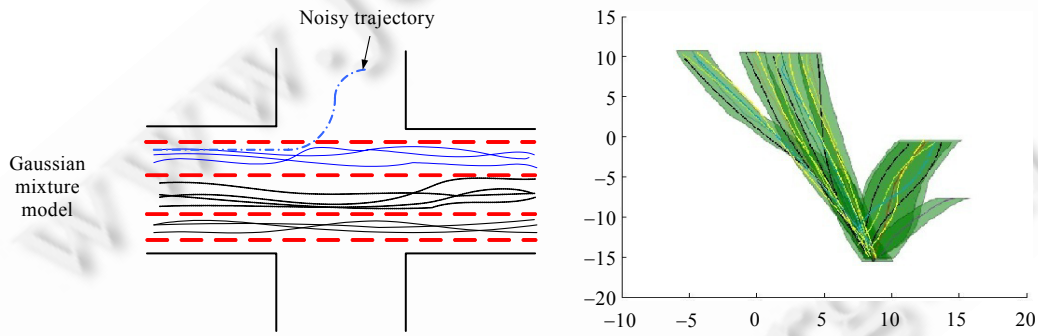
$D=\{X,Y\}$,表示训练轨迹集.$GP(x_i|\mu_x,\Sigma_x)$表示 X 方向轨迹特征矢量 x_i 符合高斯过程的轨迹模式概率函数;Y 方向上的高斯概率函数类似,定义如下:

$$GP(x_i\mid\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^d\mid\Sigma\mid}}\exp\left\{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right\} \tag{8}$$

其中,(x_i,y_i)为轨迹矢量集合;μ,Σ表示轨迹高斯过程的典型运动模式在 X 方向上的均值和协方差矩阵,其中,均值 $\mu=[E(x_1),E(x_2),...,E(x_d)]^T$,协方差矩阵 $\Sigma=(C_{ij})_{dxd},C_{ij}=Cov(x_i,x_j)$.

3.2 复杂轨迹运动模式

在包含较为复杂的运动模式场景下,典型运动模式不止一个,如图 4(a)所示,很难用一个高斯过程来刻画,两条宽虚线之间部分代表一个高斯过程.图 4(b)中,不同粗细的线条表示不同类型轨迹,需要利用多个高斯过程,即,高斯混合模型表示.注意,图 4(a)中有一条指向上方的点划线代表一条异常轨迹.



(a) 复杂轨迹运动模式 (b) 复杂轨迹运动模式 GMM 刻画

Fig.4 Pattern modeling for complex motion patterns based on GMM

图 4 复杂运动模式 GMM 建模

在具有多种轨迹模式的场景中,一条轨迹可能隶属于多个轨迹模式,准确地刻画一条轨迹需要采用混合模型.例如,对于轨迹 $Trj_n=(x_n,y_n)$,X 和 Y 方向上矢量在所有 M 个轨迹模式中出现的总概率是由不同运动模式的高斯概率混合而成:

$$p(x_n\mid\lambda)=\sum_{i=1}^{M}\omega_i GP(x_n\mid\mu_{x,i},\Sigma_{x,i}) \tag{9}$$

$$p(y_n\mid\lambda)=\sum_{i=1}^{M}\omega_i GP(y_n\mid\mu_{y,i},\Sigma_{y,i}) \tag{10}$$

其中,$GP(X_n|\mu_{x,i},\Sigma_{x,i})$表示轨迹矢量 X_n 相对第 i 个运动模式状态的高斯概率函数;M 表示混合轨迹模式的状态数量;ω_i 表示第 i 个轨迹模式的权重且 $\sum_{i=1}^{M}\omega_i=1$;$\mu_{x,i},\Sigma_{x,i}$ 和 $\mu_{y,i},\Sigma_{y,i}$ 分别表示第 i 个轨迹模式状态在 X 和 Y 方向上的均值和协方差,参数 $\lambda=\{\omega_i,\mu_i,\Sigma_i\}$,$i=1,\ldots,M$.

于是,对于轨迹训练集 $D=\{X,Y\}$,整个轨迹训练集高斯混合模型似然函数为

$$P(X\mid\lambda)=\prod_{n=1}^{N}p(x_n\mid\lambda) \tag{11}$$

$$P(Y\mid\lambda)=\prod_{n=1}^{N}p(y_n\mid\lambda) \tag{12}$$

模型中如何计算复杂运动模式中的参数 λ 是一个关键步骤,本文通过使轨迹训练模型(公式(11)、公式(12))

概率达到最大,从已知的训练轨迹矢量集 $D=\{X,Y\}$中学习训练出最佳参数λ,即,利用概率最大的方式选出构成运动模型的概率密度最大的那一组参数,为轨迹回归分析预测时所用.λ的求取方法将在第 5 节详细介绍.

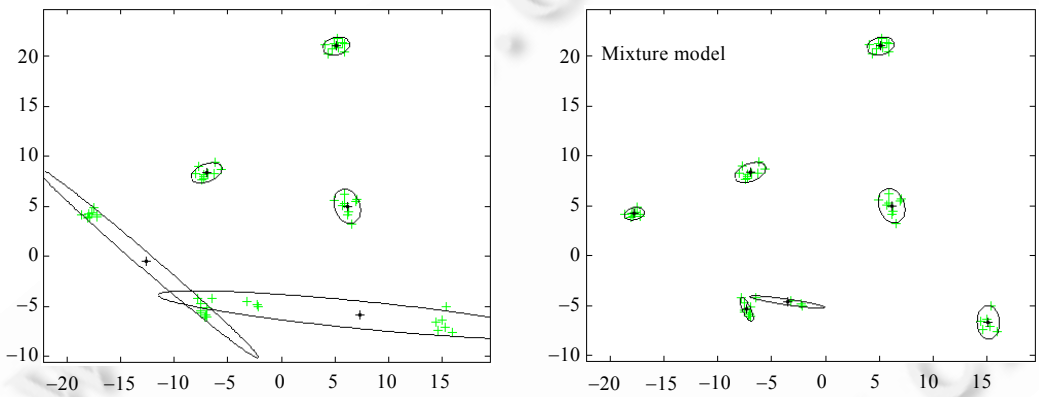
4 高斯混合模型回归预测理论

模型预测原理为:根据轨迹数据推出 GMM 的概率分布;应用高斯混合概率模型聚类获得 M 个 Component (对应了 M 个 cluster);最后,应用回归模型进行预测.聚类过程中数据点的生成需要满足以下条件:

(1) 个数据点都是在所有类别区域中随机生成的.

(2) 每个数据点属于类别 i 的概率 w_i 满足 $\sum_{i=1}^{M} w_i = 1$,M 为聚簇的个数,w_i 为每个类的先验概率(权重).模型中,M 值的设定不是人工设定的经验值,而是利用 k-means 算法初始化,通过训练集数据进行模型学习确定(如图 5 所示),因此,聚类的结果将更加客观地反映数据的本质特性.

(3) 如果一个随机生成数据点 x 可能属于类 I,那么,类 I 关于该观测数据点 x 的概率密度函数为 $f(x|\theta_i)$.



(a) k-means 对数据初始化聚类 (b) 利用 GMM 重新聚类结果

Fig.5 An example of trajectory clustering via GMM

图 5 基于 GMM 的轨迹聚类结果示例

通过图 5(b)可以看出:与图 5(a)相比,GMM 模型对轨迹点聚类的经过更加精细,可以将重叠区域的轨迹点进一步区分,划分到正确的簇中,保证下一步轨迹预测的精度.此外,轨迹聚类的另一个作用是去除噪声数据,作用是提高算法的准确性,而且可以提高时间性能.对历史轨迹数据点聚类分析(即,模型训练学习过程,参见第 5 节),然后利用高斯混合回归模型(Gaussian mixture regression,简称 GMR)进行预测分析,具体步骤如下:

假设训练数据集为 $D_{train}=(x,y)$,其中,输入数据为 x,输出数据为 y.测试数据集 $D_{test}=(x^*,y^*)$,输入测试数据为 x^*,预测输出 y^*,那么$[y,y^*]^T$的联合概率密度函数遵从如下的 GMM 模型:

$$p_{yy^*}(y,y^*)=\sum_{i=1}^{M}\omega_i GP(y,y^*\mid \mu_i,\Sigma_i) \tag{13}$$

其中,$\mu_i=[\mu_{iy},\mu_{iy*}]^T$, $\mu_i=[\mu_{iy},\mu_{iy^*}]^T$, $\Sigma_i=\begin{bmatrix}\Sigma_{iy} & \Sigma_{iyy^*}\\ \Sigma_{iy^*y} & \Sigma_{iy^*}\end{bmatrix}$, 并且满足 $\sum_{i=1}^{M}\omega_i=1$. 联合概率密度函数表示为

$$p_{yy^*}(y,y^*)=\sum_{i=1}^{M}\omega_i GP(y^*\mid y,f_i\hat{}(y),\sigma_i^2) \tag{14}$$

其中, $f_i\hat{}(y)=E[y^*\mid y]=\mu_{iy^*}+\Sigma_{iy^*y}\Sigma_{iyy}^{-1}(y-\mu_{iy}),\sigma_i^2=Var[y^*\mid y]=\Sigma_{iy^*}-\Sigma_{iy^*y}\Sigma_{iyy}^{-1}\Sigma_{iyy^*}$.

边缘密度 p_y 和条件密度 $p_{y^*|y}$ 可分别利用公式(13)和公式(14)求得,y 的边缘密度函数为

$$p_y(y)=\int_{p_{yy^*}}(y,y^*)\mathrm{d}y=\sum_{i=1}^{M}\omega_i GP(y,\mu_{iy},\Sigma_{iy}) \tag{15}$$

条件密度函数为

$$p_{y^*|y}(y^*\mid y)=\sum_{i=1}^{M}\Phi_i(y)GP(y,f_i\hat{}(y),\sigma_i^2) \tag{16}$$

其中,混合权重计算公式如下:

$$\Phi_i(y)=\frac{\omega_i GP(y,\mu_{iy},\Sigma_{iy})}{\sum_{i=1}^{M}\omega_i GP(y,\mu_{iy},\Sigma_{iy})} \tag{17}$$

因此,y^*关于 y 的回归函数,即,y^*的预测值为

$$\overline{y^*}=f\hat{}(y)=E[y^*\mid x,y,x^*]=\sum_{i=1}^{M}\Phi_i(y)f_i\hat{}(y) \tag{18}$$

对应的方差为

$$v(y)=Var[y^*\mid y]=\sum_{i=1}^{M}\Phi_i(y)[f_i\hat{}(y)^2+\sigma_i^2]-[\sum_{i=1}^{M}\Phi_i(y)f_i\hat{}(y)]^2 \tag{19}$$

公式(18)称为轨迹高斯混合回归模型,其基本思想是:首先,利用公式(16)对轨迹数据利用概率密度函数建模,通过 GMM 对训练轨迹数据进行聚类分析;然后,利用 EM 算法估计相应参数,依据符合正态分布数据的条件分布得到 M 个高斯分量的回归函数;最后,利用公式(18)将回归函数加权混合完成轨迹回归预测.

5 模型训练学习原理

使用 GMM 对复杂运动模式建模就是要准确估计模型的参数λ,GMM 参数估计最常用的一种方式是最大似然估计(expectation-maximization,简称 EM).EM 算法在迭代中改善模型参数估计,在每次迭代中不断地增加模型估计参数λ与观测训练轨迹 X 方向矢量 x_i(由若干轨迹点构成)的匹配概率.这里讨论 X 方向,Y 方向情况类似.即,每次迭代使公式(11)达到 $P(X|\lambda^{k+1})>P(X|\lambda^k)$,其中,$k$ 表示迭代的次数.通过不断地学习,获得最佳匹配训练轨迹矢量集 $X=\{x_1,x_2,...,x_n\}$.迭代训练的目的即为找到一组λ,使得 $P(X|\lambda)$最大,如公式(20)所示:

$$\hat{\lambda}=\arg\max_{\lambda}P(X\mid\lambda) \tag{20}$$

求解使 $P(X|\lambda)$达到最大的参数λ值的过程如下:

为了便于求解,通常利用 $\log P(X|\lambda)$代替 $P(X|\lambda)$求取最大值.文中 GMM 模型可以解释为对具有不同典型运动模式的轨迹建模,单个高斯过程 $GP(\cdot)$分量表示某一种简单运动模式出现的概率.

$P(X|\lambda)$是参数λ的非线性函数,直接求其最大值非常困难,可以将公式(11)转化为下式进行间接求取:

$$J(\lambda,\lambda')=\sum_{i=1}^{M}p(x,i\mid\lambda)\log p(x,i\mid\lambda') \tag{21}$$

其中,$\lambda'=\{\omega_i',\mu_i',\Sigma_i'\}$ 为模型的另一组参数,i=1,2,...,M 为混合分量序号,$p(x,i|\lambda)$表示在参数λ条件下,一条轨迹在 X 方向上的特征矢量 x 属于第 i 种运动模式的概率密度.于是,对公式(21)进行运算,得到:

$$J(\lambda,\lambda')-J(\lambda,\lambda)=\sum_{i=1}^{M}p(x,i\mid\lambda)\{\log p(x,i\mid\lambda')-\log p(x,i\mid\lambda)\}=\sum_{i=1}^{M}p(x,i\mid\lambda)\log\frac{p(x,i\mid\lambda')}{p(x,i\mid\lambda)} \tag{22}$$

对于函数 $f(x)=\log x$,具有性质:该函数在点$(x,f(x))|_{x=1}$ 处切线方程为 $\varphi(x)=x-1$.于是,$f(x)\leqslant\varphi(x)$,且当 x=1 时等号成立.对于公式(22),有:

$$J(\lambda,\lambda')-J(\lambda,\lambda)\leqslant\sum_{i=1}^{M}p(x,i\mid\lambda)\left[\frac{p(x,i\mid\lambda')}{p(x,i\mid\lambda)}-1\right]=\sum_{i=1}^{M}\{p(x,i\mid\lambda')-p(x,i\mid\lambda)\} \tag{23}$$

记为

$$J(\lambda,\lambda')-J(\lambda,\lambda)\leqslant p(x,\lambda')-p(x,\lambda) \tag{24}$$

当且仅当$\lambda=\lambda'$时等号成立.显然,只要 $p(x,\lambda')<p(x,\lambda)$,就有 $J(\lambda,\lambda')<J(\lambda,\lambda)$.由于 $J(\lambda,\lambda')$与 $p(x,\lambda')$具有相同的单调性,因此,使 $p(x,i|\lambda')$递增的过程就是使 $J(\lambda,\lambda')$递增,可以对 $p(x,\lambda)$取关于λ的微分,如下:

$$\nabla_\lambda p(x,\lambda)=\nabla_\lambda\sum_{i=1}^{M}p(x,i\mid\lambda)=\sum_{i=1}^{M}(\nabla_\lambda p(x,i\mid\lambda))=\sum_{i=1}^{M}p(x,i\mid\lambda)(\nabla_\lambda\log p(x,\lambda)) \tag{25}$$

结合公式(21)和公式(25),有下列等式成立:

$$\nabla_{\lambda} p(x,\lambda)=\nabla_{\lambda} J(\lambda,\lambda')|_{\lambda'=\lambda} \tag{26}$$

可见,当$\lambda'=\lambda$时,$J(\lambda,\lambda')$极值与 $p(x,\lambda)$极值处于相同点λ.可见,$J(\lambda,\lambda')$与 $p(x,\lambda)$不仅单调性一致,而且二者极值点也相同,那么可以通过迭代收敛使 $J(\lambda,\lambda')$趋于最大,得到新的模型参数λ'.

该迭代算法重估推导如下:将公式(11)代入公式(21),可得:

$$J(\lambda,\lambda')=\sum_{n=1}^{N}\sum_{i=1}^{M}\Phi_n(i)\log\omega_i' GP'(x_n\mid\mu_{x,i}',\Sigma_{x,i}') \tag{27}$$

其中,$\Phi_n(i)=p(x_n,i|\lambda)=p(x_n|\lambda)p(i|x_n,\lambda)$.

下面需要进一步计算轨迹矢量属于运动模式 i 的概率,求出公式(27)中参数 $\lambda'=\{\omega_i',\mu_{x,i}',\Sigma_{x,i}'\}$ 的偏导数为 0 时的值.这个步骤称为 *E*-step(求期望)和 *M*-step(求极值),具体方法如下.

(1) *E*-step:轨迹 $Trj_n=(x_n,y_n)$在 X 方向上轨迹矢量 x_n 属于轨迹模式状态 i 的概率为

$$p(i\mid x_n,\lambda)=\frac{\omega_i p(x_n\mid i,\lambda)}{p(x_n\mid\lambda)}=\frac{\omega_i GP(x_n\mid\mu_{x,i},\Sigma_{x,i})}{\sum_{k=1}^{M}\omega_k GP(x_n\mid\mu_{x,k},\Sigma_{x,k})} \tag{28}$$

(2) *M*-step:利用最大期望法求取 GMM 参数的迭代估计公式:

$$\omega_i'=\frac{1}{T}\sum_{n=1}^{N}p(i\mid x_n,\lambda),\mu_i'=\frac{\sum_{n=1}^{N}p(i\mid x_n,\lambda)x_n}{\sum_{n=1}^{N}p(i\mid x_n,\lambda)},\Sigma_i'=\frac{\sum_{n=1}^{N}p(i\mid x_n,\lambda)x_n^2}{\sum_{n=1}^{N}p(i\mid x_n,\lambda)}-\mu_i'^2 \tag{29}$$

其中,*E*-step 是假设知道各个高斯模型的参数.然后估计每个高斯模型的权值.*M*-step 是基于估计的权值,回过头再去确定高斯模型的参数.重复上述两个步骤,直到波动很小,近似达到极值.

6 轨迹预测模型

6.1 工作原理

在第 3 节中,根据轨迹模式状态单一和多样性的特征建立了两种情况下不同的模型.基于此,GMTP 算法利用最小二乘法的线性回归与高斯混合回归模型 GMR 相结合,针对不同的运动模式分别进行轨迹预测.

在处理时空轨迹数据时,预测未来位置点采用经典的线性回归方法是比较有效的.通过已知的历史轨迹来预测第 1 步,表示为 $\overline{S_{d+1}}=f(S_d,S_{d-1},\ldots,S_{d-k+1})$;然后,利用历史轨迹预测进行第 2 步预测.以此类推,第 n 步($d+n$ 位置)预测表示为 $\overline{S_{d+n}}=f(S_{d+n-1},S_{d+n-2},\ldots,S_{d-k+n})$,$f$ 表示 GMR 回归预测方程(见公式(18)).$\{S_1,S_2,\ldots,S_d\}$为已知有序的历史轨迹离散点,$S_i=(x_i,y_i,t_i)$,k 为回归预测时输入的历史轨迹长度.可见,预测下一时刻轨迹位置时,需要将上一步预测值作为新一步预测的历史轨迹点,采用迭代的方式不断地预测未来的位置点.

在实际预测中,先对历史轨迹建模和聚类,然后进行回归预测.在 X 和 Y 方向对移动对象分别进行预测,单独建模得到各自轨迹预测的回归函数 f 为 $\overline{x_{d+1}}=f_x(x_d,x_{d-1},\ldots,x_{d-k+1})$, $\overline{y_{d+1}}=f_y(y_d,y_{d-1},\ldots,y_{d-k+1})$.

将训练数据集 $D=\{(x_i,y_i)\}_{i=1}^{N}=(X,Y)$ 分为 $D_x=\{(x_{i-1},\Delta x_i)\}_{i=2}^{N}$ 和 $D_y=\{(y_{i-1},\Delta y_i)\}_{i=2}^{N}$ 两个方向分别处理.如,在已知轨迹位置信息$\{x_1,x_2,\ldots,x_d\}$预测下一个位置点 x_{d+1},即,$\overline{x_{d+1}}=f_x(x)+\varepsilon_{x,d}$,将其转为由$\Delta x=(\Delta x_2,\Delta x_3,\ldots,\Delta x_d)$预测求取下一个位置增量$\Delta x_{d+1}$,其中,$\Delta x_d=x_d-x_{d-1}$,$\varepsilon$为轨迹噪声,$\varepsilon\sim N(0,\sigma^2)$,求取公式如下:

$$\overline{\Delta x_{d+1}}=f_x(\Delta x)+\varepsilon_{x,d}.$$

利用已知的观测值 $x=[x_1,x_2,\ldots,x_{d-1}]$,$\Delta x=[\Delta x_2,\Delta x_3,\ldots,\Delta x_d]$,预测 $\overline{\Delta x_{d+1}}$ 值,进而得到 $d+1$ 位置点在 X 方向的坐标 $x_{d+1}=x_d+\overline{\Delta x_{d+1}}$.这一过程的关键在于如何得到增量回归估计函数 $f_x(\Delta x)$.在简单运动模式和复杂运动模式情况下,采用 GMR 预测回归方程(见公式(18)),得到如下回归方程:

$$\overline{\Delta x_{d+1}}=f_x(\Delta x)=\sum_{i=1}^{M}\Phi_i(x)f_i^{\wedge}(\Delta x) \tag{30}$$

其中, $f_i^{\wedge}(\Delta x)=K_i(x^*,x)(K_i(x,x)+\sigma^2 I)^{-1}\Delta x$, $\Phi_i(x)=\dfrac{\omega_i GP(\Delta x;0,K_i(x,x))}{\sum_{i=1}^{M}\omega_i GP(\Delta x;0,K_i(x,x))}$. 从轨迹 $Trj=\{s_1,s_2,...,s_n\}$中提取部分历史轨迹$\{s_1,s_2,...,s_d\}$回归求得未来 d+1 时刻位置的 X 和 Y 方向上的预测增量 $\overline{\Delta x_{d+1}}$,$\overline{\Delta y_{d+1}}$, 即可得到该位置点的预测值:

$$\overline{s_{d+1}}=(\overline{x_{d+1}},\overline{y_{d+1}},t_{d+1})=((x_d+\overline{\Delta x_{d+1}}),(y_d+\overline{\Delta y_{d+1}}),t_{d+1}).$$

此外,在已知历史轨迹 d+1 时刻位置的真实值 $s_{d+1}=(x_{d+1},y_{d+1},t_{d+1})$的基础上,可以定量求得 d+1 时刻位置的预测误差Δs_{d+1} 为

$$\Delta S_{d+1}=\sqrt{(\overline{x_{d+1}}-x_{d+1})^2+(\overline{y_{d+1}}-y_{d+1})^2}=\sqrt{(x_d+\overline{\Delta x_{d+1}}-x_{d+1})^2+(y_d+\overline{\Delta y_{d+1}}-y_{d+1})^2} \tag{31}$$

6.2 算法实现

本文提出的基于高斯混合模型的轨迹预测算法如下:

算法 1. 基于高斯混合模型的轨迹预测算法——GMTP.

输入:训练轨迹数据集 $D_{train}=\{T_1,T_2,...,T_n\}$;

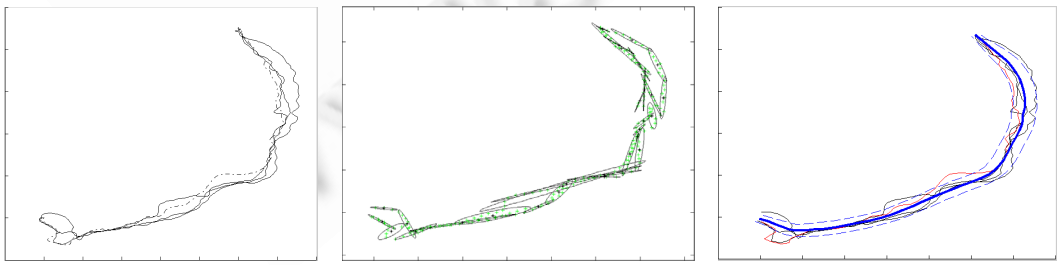
测试轨迹数据集 $D_{test}=\{T_1^*,T_2^*,...,T_m^*\}$.

输出:轨迹预测误差均值 RMSE.

1. $T^*=\{s_1,s_2,...,s_n\}$; //已知轨迹序列
2. $model\_Inital=Kmeans(D_{train})$; //$k$-means 算法初始化模型
3. $m=Train(D_{train},model\_inital)$; //迭代训练生成新模型,参数优化后为$\lambda=\{\omega_i,\mu_i,\Sigma_i\}$
4. $k=n-d$; //求取预测步数,d 观测轨迹点数量
5. for i=1 to k
6. $p=Predict(m)$; //预测连续 k 步未来轨迹点
7. $e[i]=CalRMSE(p,p_r)$; //计算每步的预测误差,p_r 表示真实轨迹点
8. end
9. $RMSE=\left(\sum_{i=1}^{k}e[i]\right)\Big/k$; //求取误差均值

基于 GMM 的轨迹预测模型的优势在于,模型参数可以根据不同历史数据自适应训练获取.其主要问题在于:模型本身是非参数性(无需人为设置参数)求解,导致其计算量随着数据增加变大.在模型训练中,参数都是通过最优化边缘似然获取的,每一次梯度计算都要对协方差矩阵进行求逆运算,时间复杂度为 $O(n^3\times m)$,n 表示轨迹数量,m 表示梯度计算的次数.虽然模型训练代价高,但是预测的时间复杂度为 $O(n^2)$,n 表示预测轨迹的数量.

图 6 给出针对具有复杂运动模式的轨迹进行预测的结果示例:首先,将真实轨迹数据(如图 6(a)所示,其中,实线表示训练轨迹,虚线表示测试轨迹)进行 GMM 建模和训练,得到相应的聚类结果(如图 6(b)所示);然后,利用 GMTP 预测轨迹未来位置点(如图 6(c)所示),得到粗实线的连续轨迹.



(a) 真实轨迹 (b) GMM 轨迹聚类结果 (c) GMTP 预测结果

Fig.6 An example of trajectory prediction results based on GMM

图 6 基于 GMM 的轨迹预测结果示例

7 实验与性能分析

为了检验本文所提出的轨迹预测方法的性能,设计实现了基于 GMM 的轨迹预测算法 GMTP、基于高斯回归的预测算法 GPR(主要用于预测具有单一运动模式的轨迹)、基于 Kalman filter(卡尔曼滤波)的预测算法 Kalman.其中,Kalman filter[18]是一种应用普遍的回归预测方法.本节比较上述 3 种算法的准确性和时间性能.实验运行在主频为 2.66GHz 的 AMD Athlon 5000+CPU 上,内存为 4GHz,操作系统为 Windows 7,实验平台为 MATLAB R2012a.实验数据来源于 MIT 停车场行驶车辆数据[21],收集了 40 453 条真实轨迹数据.导致噪声数据产生的主要原因是数据采集和数据记录错误.噪声数据的处理主要针对训练数据集,本文主要采用简单地图坐标匹配算法,将噪声点垂直于坐标轴投影到主干轨迹上.实验中所用参数及设置见表 1.

Table 1 Parameter setting

表 1 参数设置

参数	值
训练轨迹数量	39 453
测试轨迹数量	100, 200, 300, …, 900, 1000
噪声数据比例	0.05, 0.1, 0.15, …, 0.35, 0.4
历史轨迹输入长度(轨迹点数)	10, 20, 30, 40, 50
预测轨迹长度(轨迹点数)	50, 40, 30, 20, 10
GMM 模型 GP 混合分量个数	5, 10, 15

7.1 预测准确性和时间比较

本节对 3 种预测算法在不同的测试集 100~1000 条测试轨迹下进行预测分析,应用定义 5 的预测误差计算方法,实验结果取每种测试集下所有轨迹预测误差(RMSE)的平均值,评价预测的准确性.如图 7 所示:与另外两种算法相比,随着测试轨迹数量的增加,GMTP 预测误差最小,一直保持在 40m 以下,相对较长的真实轨迹,预测精度较高且比较稳定.分析实验结果可知:GPR 预测的误差比 GMTP 平均高出 10m 左右,但比 Kalman filter 预测更为精确.原因在于:GPR 对于处理较为简单运动模式轨迹预测的精度较好,而实验中采用的数据集中轨迹模式较为复杂且种类繁多,很难用单一高斯过程描述,而基于 GMM 的模型通用性较好.基于 Kalman filter 的轨迹预测仅仅对轨迹进行线性回归预测,没有对不同的轨迹模式进行聚类分析,因此预测误差最大.与 GPR 和 Kalman filter 算法相比,GMTP 预测准确率平均提高了 22.2%和 23.8%.

为了进一步验证本文方法的性能,观察算法的预测时间代价.图 8 中,GMTP 预测时间非常小,与 GPR 和 Kalman filter 算法相比,平均缩减了 92.7%和 95.9%.因为 Kalman filter 对每条轨迹的下一个位置预测都要将前一个点的位置信息代入回归分析,当预测的轨迹数量增加时,预测时间会呈线性增长,如图 8 所示.而 GMTP 预测时可以同时对具有统一模型参数的轨迹进行预测,利用高斯混合模型刻画的轨迹仅需一次性预测即可,因此当预测轨迹数量增加时,只要没有增加较多的轨迹运动模式,预测时间就不会产生较大的增加和波动.当训练集轨迹数量达到一定规模时,轨迹运动模式较为丰富,可以包含大部分运动模式,利用 GMM 训练得到的 GMTP 预测模型已经能够处理各种复杂运动模式的轨迹预测问题.由于处在受限的道路交通环境下,测试集中所增加轨迹的运动模式与之前的模式很多都是一致的.仅仅是测试集轨迹数量增加而无变化的运动模式增加时,训练得到模型中 GP 混合分量个数等都不会有太多变化,因此,GMTP 预测误差波动较小.

为了进一步验证本文所提算法的性能优势,与已有的性能较好的轨迹预测算法进行比较,包括文献[3]提出的预测算法 TPMO、文献[12]提出的预测算法 HMTP.上述两项工作采用命中率作为预测有效性的评价指标,而本文采用均方根误差 RMSE 评价轨迹预测的准确性,无法在一个衡量尺度上对比,因此,本文比较 3 种算法的预测时间性能.采用 MIT 停车场数据,预测轨迹数量在 100~1000 范围内变化,实验结果如图 9 所示.

实验结果表明:GMTP 略优于基于隐马尔可夫模型的预测算法 HMTP,明显优于基于时间连续贝叶斯网络的预测模型 TPMO;GMTP 的平均预测时间近似为 HMTP 的 1/10,是 TPMO 的 1/2.原因在于:GMTP 是一种高斯非线性概率统计模型,结合最小二乘法与高斯混合回归算法进行预测,模型训练时间代价很低;HMTP 算法训练

模型过程中需要利用 HMM 模型提取隐状态和观察状态,代价相对较高;而 TPMO 算法训练过程中不但需要利用热点区域挖掘算法对轨迹进行聚类,并且需要构建轨迹时间连续贝叶斯网络,极其耗时.

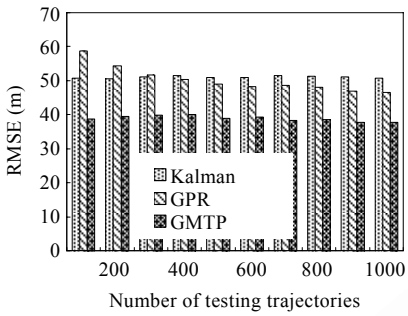


Fig.7 Prediction error comparison

图 7 预测误差比较

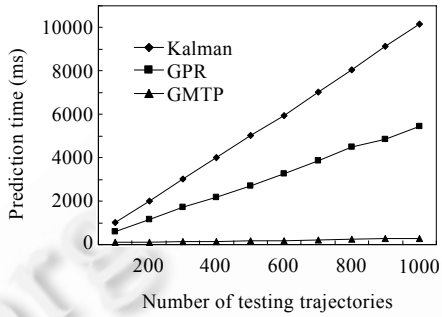


Fig.8 Prediction time comparison

图 8 预测时间比较

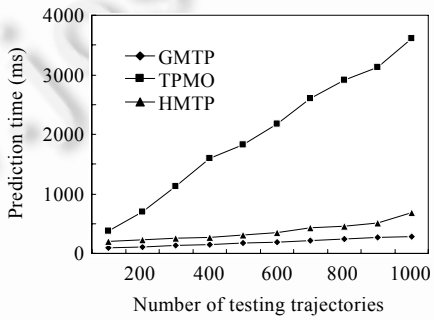


Fig.9 Prediction time comparison between typical trajectory prediction algorithms

图 9 通用轨迹预测算法时间性能比较

7.2 预测误差分析

GMTP 高斯混合分量个数初始化采用 *k*-means 方法,目标是选取最合适的混合分量.值得注意的是,如果选择的混合分量过少,则算法不能充分地对数据建模,进而对测试轨迹进行有效的预测判断;如果选择太多高斯混合分量,则会增加轨迹预测时间代价.图 10 为确定了训练轨迹数量,在不同的高斯混合分量个数下,比较不同轨迹输入长度(即,已知轨迹点个数)对预测精度的影响.图 11 为确定了混合分量个数,当训练轨迹数量不断增加时,观察不同轨迹输入长度下误差的变化.注意,大规模训练轨迹的预测误差结果类似,不再赘述.

由图 10、图 11 可以看出:

(1) 在不同高斯过程分量个数和训练轨迹数据集下,观测的历史轨迹输入点个数越多,算法预测误差就越低.由于预测模型有了更多的历史轨迹数据点信息,包含更多的运动模式,轨迹聚类更加精确,预测误差降低.

(2) 图 10 中,当 GP 混合分量个数从 5 增加到 15 时,其预测误差变化不大.针对本文采用的 GPS 轨迹数据集,实验得出:当 GP 混合分量达到 5 时,算法的预测误差收敛.值得注意的是,不同轨迹数据集其运动模式个数不同,因此需要通过实验探寻最佳的混合分量个数.

(3) 图 11 中,当训练轨迹数量增加时,预测误差波动较小.原因参见第 7.1 节图 7 的实验结论,进一步证明了 GMTP 算法的稳定性.

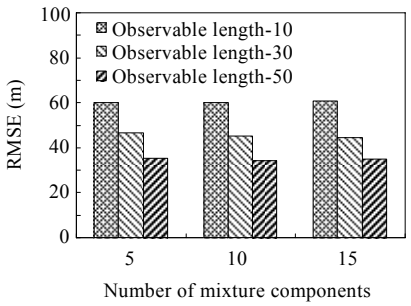


Fig.10 Prediction error comparison with distinct mixture components

图 10 不同混合分量下预测误差的比较

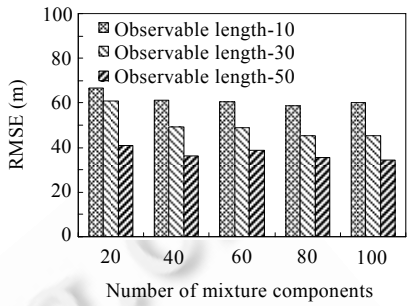
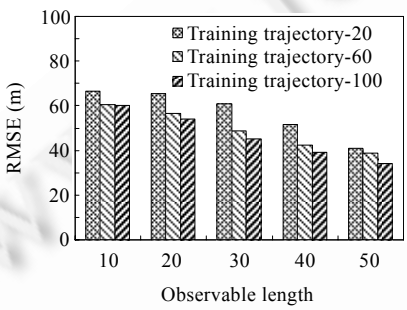


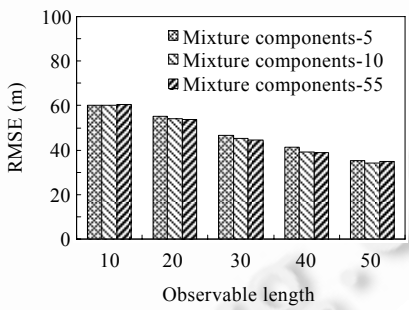
Fig.11 Prediction error comparison under different training datasets

图 11 不同训练集下预测误差的比较

当有足够历史轨迹点信息时,预测误差会相应降低,因此,历史轨迹输入长度变化对预测误差具有很大影响.图 12(a)为确定了 GP 混合分量个数,观察不同输入历史轨迹长度下,训练轨迹数量变化对预测精度的影响.图 12(b)是训练轨迹集确定,分析不同的历史轨迹输入长度下,GP 混合分量个数变化对预测误差的影响.



(a) GP 混合分量个数=5



(b) 训练轨迹数量=100

Fig.12 Prediction error under different lengths of input historical trajectories

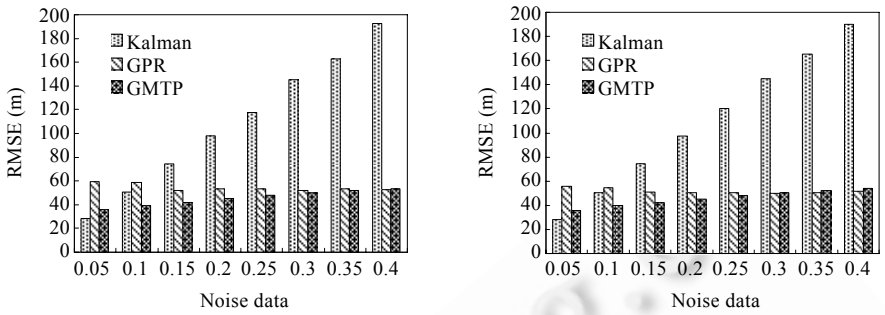
图 12 不同输入历史轨迹长度下的预测误差

由图 12 可以看出:预测模型的输入历史轨迹长度越大,其预测误差就越低.如图 12(a)所示，对于不同训练轨迹集,当输入历史轨迹长度增加时,误差不断降低,与客观事实相符.由图 12(b)可以看出:当混合分量个数从 5 增加到 15 时,相同历史轨迹输入长度下预测误差变化不大,与前述实验结论一致.

7.3 算法抗干扰性分析

大多数 GPS 轨迹数据中均存在一些噪声数据,对预测精度产生很大的影响.本节将验证 3 种预测算法对噪声数据的抗干扰性.分别选取训练轨迹数量为 100 和 200 条进行实验,结果如图 13 所示,其中,横轴表示噪声点所占比例.

实验得出:卡尔曼滤波算法对噪声数据的变化比较敏感,随着噪声增大,预测误差不断增大,近似成线性.而对于 GMTP 和 GPR 算法,噪声数据的增加对预测误差并无太大的影响.其原因在于:Kalman filter 是自回归滤波器,它利用卡尔曼滤波器从一系列有噪声的观察数据中估计出被观察过程的内部状态,因此噪声数据变化对误差产生很大的影响.而对于 GMTP 和 GPR 模型,预测之前的轨迹聚类模型已经将噪声轨迹点进行有效的归类,进行预测时算法能够很好地区分出噪声轨迹,从而不会影响其他轨迹数据的预测.

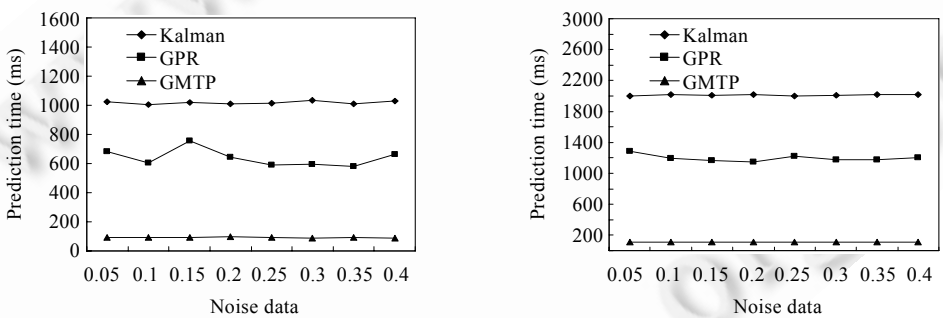


(a) 训练集轨迹数量=100 (b) 训练集轨迹数量=200

Fig.13 Prediction error comparison under different portion of noise data

图 13 不同比例噪声数据下预测误差比较

为了进一步验证算法的时间性能,对 3 种预测模型在不同噪声数据下的时间代价进行比较.由图 14 可以看出,卡尔曼滤波算法的时间代价比较高.原因在于:卡尔曼滤波模型假设 k 时刻的真实状态是从 k–1 时刻的状态演化而来,所以每一步都要依据上一步来分析确定,其时间效率比较低.而 GMTP 和 GPR 算法在预测之前已经将轨迹频繁模式挖掘出来,用均值点构成的轨迹代表大量具有相同运动模式轨迹的预测路线,因此时间效率较高.此外,相比 GPR 算法,GMTP 预测的时间代价更低,进一步证明了算法的时效性.



(a) 训练集轨迹数量=100 (b) 训练集轨迹数量=200

Fig.14 Prediction time comparison under different portion of noise data

图 14 不同比例噪声数据下预测时间的比较

8 结论与展望

移动对象不确定性轨迹预测是一个崭新和充满挑战性的研究课题,针对当前轨迹预测精度不高、局限于道路网络、预测实时性不好的不足,本文提出了一种基于高斯混合模型的轨迹预测方法.该算法可以利用高斯混合模型对移动对象复杂运动模式建模,统计不同运动模式的概率分布,进而将轨迹划分为不同高斯过程分量,实现准确和高效的位置预测.这一方法的优点在于:算法运行过程中无需设置大量的参数,可以通过数据自身利用概率统计分布特性获得各种运动模式下的位置信息.

未来的研究工作包括:

(1) 与国内交通部门合作,将本文所提模型应用于卡口摄像头采集到的交通数据,辅助跟踪和定位机动车(主要针对未装备 GPS 定位系统的交通工具);

(2) 充分考虑客观因素对移动对象位置预测的影响,如红绿灯、交通堵塞,提高预测算法对环境因素的自适应性.

References:

[1] Meng XF, Ding ZM. Mobile Data Management: Concepts and Techniques. Beijing: Tsinghua University Press, 2009. 185−200 (in Chinese).

[2] Asahara A, Sato A, Maruyama K, Seto K. Pedestrian-Movement prediction based on mixed Markov-chain model. In: Proc. of the 19th ACM SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. New York: ACM Press, 2011. 25−33. [doi: 10.1145/2093973.2093979]

[3] Qiao SJ, Shen DY, Wang XT, Han N, Zhu W. A self-adaptive parameter selection trajectory prediction approach via hidden Markov models. IEEE Trans. on Intelligent Transportation Systems, 2015,16(1):284−296. [doi: 10.1109/TITS.2014.2331758]

[4] Mamoulis N, Cao HP, Kollios G, Hadjieleftheriou M, Tao YF, Cheung DW. Mining, indexing, and querying historical spatiotemporal data. In: Proc. of the 2004 ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2004. 236−245. [doi: 10.1145/1014052.1014080]

[5] Morzy M. Mining frequent trajectories of moving objects for location prediction. In: Proc. of the 5th Int'l Conf. on Machine Learning and Data Mining in Pattern Recognition. LNCS 4571, Heidelberg: Springer-Verlag, 2007. 667−680. [doi: 10.1007/978-3-540-73499-4_50]

[6] Jeung H, Liu Q, Shen HT, Zhou XF. A hybrid prediction model for moving objects. In: Proc. of the 24th Int'l Conf. on Data Engineering. Washington: IEEE Computer Society, 2008. 70−79. [doi: 10.1109/ICDE.2008.4497415]

[7] Ying JC, Lee WC, Weng TC, Tseng S. Semantic trajectory mining for location prediction. In: Proc. of the 19th ACM SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. New York: ACM Press, 2011. 34−43. [doi: 10.1145/2093973.2093980]

[8] Zheng Y, Zhang LZ, Xie X, Ma WY. Mining interesting locations and travel sequences from GPS trajectories. In: Proc. of the 18th Int'l Conf. on World Wide Web. New York: ACM Press, 2009. 791−800. [doi: 10.1145/1526709.1526816]

[9] Song CM, Qu ZH, Blumm N, Barabsi AL. Limits of predictability in human mobility. Science, 2010,327(5968):1018−1021. [doi: 10.1126/science.1177170]

[10] Pan TL, Sumalee A, Zhong RX, Indra-Payoong N. Short-Term traffic state prediction based on temporal-spatial correlation. IEEE Trans. on Intelligent Transportation Systems, 2013,14(3):1242−1254. [doi: 10.1109/TITS.2013.2258916]

[11] Zhou JB, Tung KH, Wu W, Ng WS. A "semi-lazy" approach to probabilistic path prediction in dynamic environments. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2013. 748−756. [doi: 10.1145/2487575.2487609]

[12] Qiao SJ, Peng J, Li TR, Zhu Y, Liu LX. Uncertain trajectory prediction of moving objects based on CTBN. Journal of University of Electronic Science and Technology of China, 2012,41(5):759−763 (in Chinese with English abstract).

[13] Tao YF, Faloutsos C, Papadias D, Liu B. Prediction and indexing of moving objects with unknown motion patterns. In: Proc. of the 2004 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2004. 611−622. [doi: 10.1145/1007568.1007637]

[14] Qiao SJ, Tang CJ, Jin HD, Long T, Dai SC, Ku YC, Chau M. PutMode: Prediction of uncertain trajectories in moving objects databases. Applied Intelligence, 2010,33(3):370−386. [doi: 10.1007/s10489-009-0173-z]

[15] Hu WM, Xiao XJ, Fu ZY, Xie D, Tan TN, Maybank S. A system for learning statistical motion patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006,28(9):1450−1464. [doi: 10.1109/TPAMI.2006.176]

[16] Feng T, Guo YF, Huang KZ, Ji J. Behavior trajectory restoration algorithm based on hidden Markov models. Computer Engineering, 2012,38(12):1−5 (in Chinese with English abstract).

[17] Gaffney SJ, Robertson AW, Smyth P, Camargo SJ, Ghil M. Probabilistic clustering of extratropical cyclones using regression mixture models. Climate Dynamics, 2007,4(29):423−440. [doi: 10.1007/s00382-007-0235-z]

[18] Deng HB, Zhang L, Wu Y, Zhou J, Liu F. Research on track estimation based on Kalman filtering algorithm. Transducer and Microsystem Technologies, 2012,31(5):4−7 (in Chinese with English abstract).

[19] Sung HG. Gaussian mixture regression and classification [Ph.D. Thesis]. Houston: Rice University, 2004.

[20] Qiao SJ, Han N, Zhu W, Gutierrez LA. TraPlan: An effective three-in-one trajectory prediction model in transportation networks. IEEE Trans. on Intelligent Transportation Systems, 2014. [doi: 10.1109/TITS.2014.2353302]

[21] http://www.ee.cuhk.edu.hk/~xgwang/MITtrajsingle.html

附中文参考文献:

[1] 孟小峰,丁治明.移动数据管理:概念与技术.北京:清华大学出版社,2009.185−200.

[12] 乔少杰,彭京,李天瑞,朱焱,刘良旭.基于 CTBN 的移动对象不确定轨迹预测算法.电子科技大学学报(自然科学版),2012,41(5):759−763.

[16] 冯涛,郭云飞,黄开枝,吉江.基于隐马尔可夫模型的行为轨迹还原算法.计算机工程,2012,38(18):1−5.

[18] 邓胡滨,张磊,吴颖,周洁,刘枫.基于卡尔曼滤波算法的轨迹估计研究.传感器与微系统,2012,31(5):4−7.

乔少杰(1981−),男,山东招远人,博士,副教授,CCF 高级会员,主要研究领域为移动对象数据库,轨迹数据挖掘.

唐常杰(1946−),男,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,数据挖掘.

金琨(1990−),男,硕士生,主要研究领域为移动对象数据库,轨迹预测.

格桑多吉(1972−),男,副教授,主要研究领域为数据挖掘.

韩楠(1984−),女,博士,工程师,主要研究领域为移动对象数据库,生物信息学.

Louis Alberto GUTIERREZ(1980−),男,博士,Researcher,主要研究领域为数据挖掘.