Journal of Computer Applications
计算机应用,2013,33(11):3119-3122,3133
ISSN 1001-9081
CODEN JYIIDU
2013-11-01
http://www.joca.cn

文章编号:1001-9081(2013)11-3119-04
doi:10.11772/j.issn.1001-9081.2013.11.3119

基于概率后缀树的移动对象轨迹预测

王　兴[1,2*],蒋新华[1,3],林　劼[2],熊金波[2]
(1.中南大学 信息科学与工程学院,长沙 410075;　2.福建师范大学 软件学院,福州 350108;
3.福建工程学院 下一代互联网应用技术研究中心,福州 350108)
(*通信作者电子邮箱 wangxing@csu.edu.cn)

摘　要:在移动对象轨迹预测中,针对低阶马尔可夫模型预测准确率不高、高阶模型状态空间膨胀的问题,提出一种基于概率后缀树(PST)的动态自适应变长马尔可夫模型预测方法。首先依时间先后将移动对象的轨迹路径序列化;然后根据移动对象的历史轨迹数据进行学习训练,计算序列上下文的概率特征,建立路径序列的概率后缀树模型,结合当前实际轨迹数据,动态自适应预测将来的位置信息。实验结果表明,该模型在二阶时取得最高的预测精度,随着阶数的增加,预测精度保持在82%左右,能取得较好的预测效果;同时空间复杂度呈指数级减少,大大节省了存储空间。该方法充分利用历史轨迹数据和当前轨迹信息预测未来轨迹,能够提供更加灵活、高效的基于位置服务。

关键词:变长马尔可夫模型;概率后缀树;历史轨迹;轨迹预测
中图分类号:TP311　　**文献标志码**:A

Prediction of moving object trajectory based on probabilistic suffix tree

WANG Xing[1,2*], JIANG Xinhua[1,3], LIN Jie[2], XIONG Jinbo[2]
(1. *School of Information Science and Engineering, Central South University, Changsha Hunan* 410075, *China*;
2. *Faculty of Software, Fujian Normal University, Fuzhou Fujian* 350108, *China*;
3. *Research Center for Next-Generation Internet Technology and Applications, Fujian University of Technology, Fuzhou Fujian* 350108, *China*)

Abstract: In the prediction of moving object trajectory, concerning the low accuracy rate of low order Markov model and the expansion of state space in high order model, a dynamic adaptive Probabilistic Suffix Tree (PST) prediction method based on variable length Markov model was proposed. Firstly, moving object's trajectory path was serialized according to the time; then the probability characteristic of sequence context was trained and calculated from the historical trajectory data of moving objects, the probabilistic suffix tree model based path sequence was constructed, combined with the actual trajectory data, thus the future trajectory information could be predicted dynamically and adaptively. The experimental results show that the highest prediction accuracy was obtained in second order model, with the order of the model increasing, the prediction accuracy was maintained at about 82% and better prediction results were achieved. In the meantime, space complexity was decreased exponentially and storage space was reduced greatly. The proposed method made full use of historical data and current trajectory information to predict the future trajectory, and provided a more flexible and efficient location-based services.

Key words: variable order Markov model; Probabilistic Suffix Tree (PST); history trajectory; trajectory prediction

0　引言

随着全球定位系统(Global Positioning System, GPS)定位技术和通信技术的发展,人们可以采集到更多的位置信息,基于位置的服务(Location Based Service,LBS)的应用越来越广泛,它包含两层含义:一是确定移动设备或用户所在的地理位置;二是提供与位置相关的各类信息服务,如导航服务、急救服务、信息查询服务等。此类服务可以为人们的生活和出行提供极大的便利,然而系统使用的上下文信息仅局限于用户的当前位置,缺乏智能性和灵活性。若能实时动态地预测用户将来的移动位置,则可以提供更加灵活的预报服务。如提供下一条路段拥堵的提醒服务、下一个服务区的停车场信息、下一个商场的商品广告信息、下一个景区的介绍以及下一个酒店的菜单信息等。

移动对象的轨迹预测是近年来国内外学者研究的热点问题,具有较好的研究价值和广泛的应用前景。预测方法一般分为基于欧氏空间的轨迹预测和基于路网受限的轨迹预测。在实际中,大多数移动对象都是在受限路网中运动,不能在空间随意运动,所以基于欧氏空间的轨迹预测有一定的局限性。在受限路网中移动对象的轨迹预测研究中,有两类预测方法应用较广:一类是基于序列分析、模式的预测方法[1-5];另一类是基于马尔可夫模型的统计模型预测方法[6-10]。文献[6]使用混合马尔可夫模型进行移动路径预测;文献[7]使用马尔可夫链来进行轨迹的预测;文献[8]针对固定阶马尔可夫模型预测的局限,提出使用历史模式树进行自适应多阶马尔可夫模型位置预测,较定阶预测模型灵活,但其本质仍然是定阶的马尔可夫模型,在进行高阶模型预测时,存在空间复杂度较高的问题;文献[9]使用马尔可夫模型,根据车辆刚行驶过

收稿日期:2013-05-24;**修回日期**:2013-07-22。　**基金项目**:福建省重大专项(2011HZ0002-1);国家自然科学基金资助项目(61101139);福建省科技计划重点项目(2011H0002);福建省交通科技计划项目(201122)。

作者简介:王兴(1982-),男,湖北大冶人,讲师,博士研究生,主要研究方向:数据挖掘、智能交通;　蒋新华(1956-),男,湖南长沙人,教授,博士生导师,主要研究方向:无线宽带网络、智能交通;　林劼(1972-),男,福建三明人,副教授,博士,主要研究方向:数据挖掘、生物信息学;　熊金波(1981-),男,湖南益阳人,讲师,博士研究生,主要研究方向:网络安全、数据挖掘。

的路段预测其在短期内将到达的路段;文献[10]基于连续时间马尔可夫模型进行预测。在实际应用中,一阶模型仅仅依赖当前的信息,而忽略了过去的信息,预测精度不高;高阶模型随着阶数的增高,状态空间呈现指数级增长,空间复杂度高,同时,由于高阶模型覆盖率低,预测准确率反而会降低。

本文针对以上马尔可夫模型预测方法的不足,提出了一种基于概率后缀树(Probabilistic Suffix Tree,PST)的动态自适应变长马尔可夫模型预测方法,把移动对象的轨迹根据时间先后进行序列化,将序列分析方法和马尔可夫统计模型结合,根据移动对象的历史轨迹数据进行学习训练,计算上下文的概率特征,结合当前实际轨迹数据,动态自适应预测将来的位置信息。

1 相关定义和概念

定义 1 交通路网。交通路网由路段、交叉口及它们之间的拓扑结构组成,可以看成是一个有向图 $G=\langle V,R\rangle$,其中:V 表示交叉路口的集合,R 表示路段的集合。如图 1 所示。

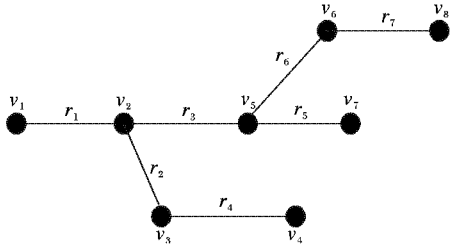

图 1 交通路网拓扑图

$R=\{r_1,r_2,r_3,\cdots,r_m\}$,其中 $r_i(1\leqslant i\leqslant m)$ 为路段。

定义 2 轨迹序列。移动对象在交通路网中按时间先后走过的路段序列,$T=\langle r_i,\cdots,r_j\rangle$,$r_k\in R$,轨迹序列中路段的个数为序列的长度,记为 $|T|$。

定义 3 后缀。给定轨迹序列 $T=\langle r_1,r_2,r_3,\cdots,r_n\rangle$,后缀序列 T_i 是指从 r_i 开始到轨迹序列末尾 r_n 的一段子序列,记为 $suffix(T_i)$,$1\leqslant i\leqslant n$。如 $suffix(T_1)=\langle r_1,\cdots,r_n\rangle$,$suffix(T_2)=\langle r_2,\cdots,r_n\rangle$,$suffix(T_3)=\langle r_3,\cdots,r_n\rangle$ 等均为 T 的后缀。

1.1 马尔可夫模型

假设 S 是一个由有限个状态序列组成的集合。令 $S=\{X_n,n=1,2,3\}$,则有:

$$P(X_{n+1}=j\mid X_n=i_n,X_{n-1}=i_{n-1},\cdots,X_1=i_1)=P(X_{n+1}=j\mid X_n=i_n,X_{n-1}=i_{n-1},\cdots,X_{n-L+1}=i_{n-L+1}) \quad (1)$$

则该马尔可夫模型称为 L 阶马尔可夫模型。当前状态序列的概率由过去的 L 个已知状态序列的概率决定。当 $L=1$ 时为标准马尔可夫模型。本文中,L 是动态自适应变化来决定的,所以也称之为变长马尔可夫模型。

1.2 概率后缀树

概率后缀树(PST)最早由 Ron 等[11]提出,采用后缀树作为存储结构,树中的边用序列集中的一个符号标记,每条边都是不同的符号。概率后缀树中的节点存放一个符号串,根节点为 Root,叶子节点存放所有的后缀序列串,中间节点中的符号串可由本节点遍历到根节点所经过的边的符号标记生成符号串。同时,每个节点都对应一个概率向量,依次存放该节点符号序列的下一个符号出现的条件概率,通过统计下一个符号出现的相对频率计算得到。对于一个给定的序列 accactact,对应的深度 $L=3$ 的概率后缀树如图 2 所示。

其中第一层节点 c 中的概率向量为[1/4,1/4,1/2],分别表示在符号串 c 出现时,下一个可能出现的符号为 a,c,t 的条件概率 $p(a|c)=1/4$,$p(c|c)=1/4$,$p(t|c)=1/2$。其中 $p(t|c)=ct$ 出现的次数/c^* 出现的次数,观察给定序列,ct 出现的次数为 2,c^* 表示以 c 开头的长度为 2 的序列个数,有 cc,ca,ct,ct 为 4,因此,$p(t|c)=2/4=1/2$。

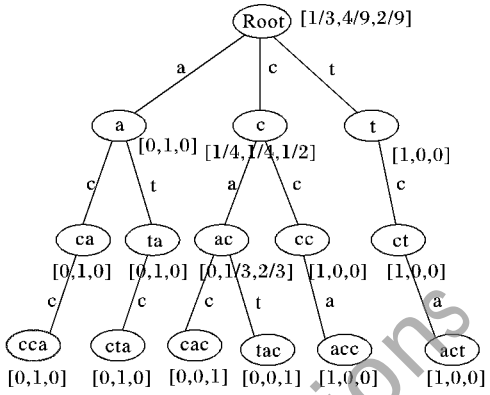

图 2 深度 $L=3$ 的概率后缀树

PST 模型的实质是一种变长马尔可夫模型。同传统的马尔可夫模型相比,概率后缀树模型能够更有效地利用和处理高阶序列信息,但随着训练数据规模的增大,状态树空间仍会急剧膨胀,复杂度为 $O(n^3)$。针对此问题,学者们提出了相应的 PST 模型的改进算法[12-15],可以在 $O(n)$ 线性复杂度开销内完成树的构造。本文中概率后缀树的构造方法以文献[14]为基础,构造树的时间复杂度和空间复杂度为 $O(n)$。

2 基于 PST 模型的轨迹预测

2.1 轨迹数据的抽象化处理

为了对交通路网中的移动对象进行有效预测,必须对移动对象的历史轨迹数据进行建模和分析,由于移动对象在路网中按一定的时间间隔发送 GPS 位置信息,采集到的 GPS 数据含有经纬度信息,把 GPS 中 WGS-84 空间坐标转换成平面坐标,投影到地图网格中,对路网进行空间网格划分后,便可根据空间网格与道路经纬度信息,建立道路与浮动车轨迹点的对应关系,实现浮动车数据与具体道路的投影关联。如图 3 所示,即是根据时间先后,把 GPS 数据点在 GoogleEarth 中可视化形成的一条轨迹数据图。按图 1 路网的拓扑结构,结合定义 1,用 A、B、C 和 D 表示道路路网中的顶点,则可把轨迹数据分割成路段序列来表示,即可得到 AB、BC、CD 路段序列。

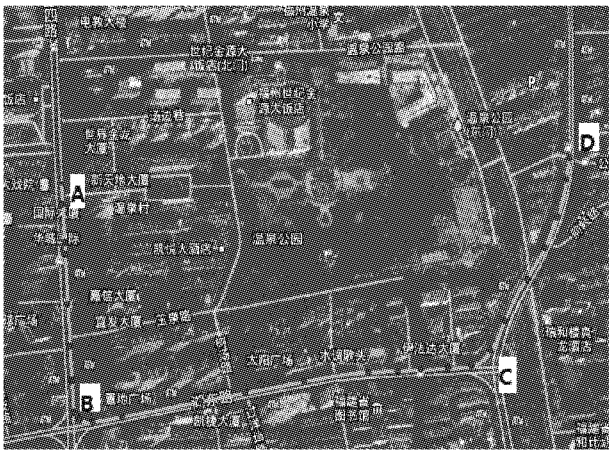

图 3 GPS 数据点轨迹

2.2 轨迹预测的思想和原理

将坐标位置信息转化为路段序列之后,序列中的路段序列是一个有限的状态序列,把轨迹中的每一个路段当成一种

状态,根据历史轨迹序列,训练路段序列之间的上下文信息进行建模,由于马尔可夫模型的特性,其下一个状态的出现的概率仅与当前的若干个状态有关,本文采用马尔可夫模型进行建模,统计路段序列在历史数据中出现的次数来计算相对频率,得出概率转移矩阵。概率的计算公式如下:

$$P(r_{k+1} \mid r_1, r_2, \cdots, r_k) = \begin{cases} \dfrac{TFr_{k+1}}{N}, & k = 0 \\ \dfrac{TFr_{1,r_2,\cdots,r_{k+1}}}{TFr_{1,r_2,\cdots,r_k}}, & k > 0 \end{cases} \tag{2}$$

其中:k 表示轨迹序列中连续出现的 k 个路段;TF 用来计算路段出现的频率;TFr_n 表示路段 r_n 在历史轨迹数据中出现的次数;N 表示所有的路段 r_i 出现的次数总和;$TFr_{1,r_2,\cdots,r_k}$ 表示路段序列$\langle r_1, r_2, \cdots, r_k\rangle$ 在历史轨迹数据中出现的次数;$TFr_{1,r_2,\cdots,r_{k+1}}$ 表示路段序列$\langle r_1, r_2, \cdots, r_k, r_{k+1}\rangle$ 在历史轨迹数据中出现的次数。

预测时,针对马尔可夫模型的低阶模型预测精度低的缺陷,本文引入多阶模型,采用最近的 k 个历史状态进行预测。

给定当前 k 个路段的上下文序列$\langle r_1, r_2, \cdots, r_k\rangle$,则可以使用转移概率矩阵来预测下一个路段 r_{k+1}。取其最大条件概率的下一个路段作为预测的结果,即:

$$r_p = \arg\max_{r_{k+1}} \{P(r_{k+1} \mid r_1, r_2, \cdots, r_k)\} \tag{3}$$

其中:$\langle r_1, r_2, \cdots, r_k\rangle$ 为当前轨迹序列,r_{k+1} 为下一个可能的路段,则 r_p 为预测结果。

根据上面的描述,根据给定的 k 个最近的状态(即移动对象经过的 k 个最近的路段序列),即可预测第 $k+1$ 个状态的值(即下一个路段 r_{k+1} 的结果),k 值也即是马尔可夫模型的阶数。

若给定的轨迹序列$\langle r_1, r_2, \cdots, r_k\rangle$在历史轨迹数据中没有匹配时,则取历史轨迹序列中,能够支持的最大阶数 $t(t < k)$ 进行预测。计算公式如下:

$$P(r_{k+1} \mid r_1, r_2, \cdots, r_k) = P(r_{k+1} \mid r_{k-t+1}, \cdots, r_k) = \begin{cases} \dfrac{TFr_{k+1}}{N}, & t = 0 \\ \dfrac{TFr_{k-t+1,\cdots,r_{k+1}}}{TFr_{k-t+1,\cdots,r_k}}, & t > 0 \end{cases} \tag{4}$$

2.3　PST 模型的建立和预测

基于移动对象的路段序列轨迹数据,采用概率后缀树模型进行建模,主要分为轨迹序列的训练和预测两部分,主要的算法描述与分析如下。

2.3.1　轨迹序列的训练

采用如下的数据结构进行节点索引:

```
typedef struct pst_struct
{
  char * str;                                  //节点字符串
  int * Pro                                    //概率向量
  struct pst_struct ** son;                    //指向儿子节点指针
} pst_node;
typedef pst_node * pst_type;
```

PST 的构造过程主要分为两个步骤:

1)初始化 PST 只包含一个根节点,根节点的概率向量值为每个符号元素在所有符号中出现的相对频率。将所有相对频率超过阈值 *pro_min* 的符号作为候选节点。

2)循环迭代扩充每个候选节点,直到树的深度为 H:

a)计算每个候选节点的后续符号概率向量;

b)假如当前候选节点的符号串为 s,若存在一个符号 r 属于 R,且 rs 的相对频率超过 *pro_min*,则将标记为 rs 的节点作为该节点的候选子节点。

算法 1　概率后缀树的建立 Train_PST。

```
输入:路段集 R = {r1,r2,r3,…,rm},历史轨迹序列训练集 T = {T1,T2,…,Tn},树的深度 H,最小频度 pro_min。
输出:概率后缀树 PST。
Algorithm Train_PST(R,T,H)
BEGIN
  Q←empty, current_depth←1,size←element numbers of R
                                        //Q 为队列,初始化为空
  For(ri in R)
    Pro[ri]←Compute Frequency of ri according to Scanning T
  PST←root                                        //创建根节点
  If (pro[ri] >pro_min)
    Add ri to Queue
  While(Q not empty && current_depth < H))
    BEGIN
      s = delete_queue(Q);        //得到当前候选节点符号串 s
      Pro[rs]←Compute All Frequency of rs according to
        Scanning T
      If (Pro[rs] >pro_min)
      Add rs to Queue
      PST←rs                                      //创建候选子节点
    END
return PST;
```

数据集的训练　T 中有 n 条历史轨迹序列数据,每个序列不论长短各占一行,存放在相应的训练文件中。

提取训练集 T 中的路段,形成路段集 R,根据 R 集中的路段元素建立概率后缀树,先建立根节点,然后依次逐层扩展建立候选子节点,构造每个节点时,需依据训练集中的序列信息统计节点的下一个路段元素出现的相对频率,从而计算出各节点的概率向量,同时,索引各节点的指针,形成层次关系。当树的当前深度达到最大的深度时,训练结束。理论上,训练时,树的深度 H 可以等于序列的最大长度,但从实际考虑,在预测时是根据序列中最近的 k 个路段数据来进行的($k \leqslant L$),所以实际建立的树深度为 $L(L \leqslant H)$。一是为了减少计算和存储开销;二是太过久远的历史数据对未来预测影响较小。另外,设置相应的阈值进行剪枝,如最小频率 *pro_min* 阈值。当概率向量中的值小于该阈值时,相应的下一个元素出现的概率很低,可以从树中剪掉,在建树的过程中节省了开销。

2.3.2　轨迹序列的预测

概率后缀树中存储了相应的轨迹序列上下文的统计概率信息。给定一个预测序列,从根节点出发,沿着序列倒序的方向访问 PST 中的节点,匹配到相应的节点,取最大的概率对应的路段,即为预测的下一个路段。

算法 2　轨迹序列的下一路段预测 Predict_PST。

```
输入:概率后缀树 PST, 给定的预测序列 T = {r1,r2,r3,…,rn}。
输出:下一路段。
Algorithm Predict_PST(PST,T)
BEGIN
  For(each suffix of T)
    C = Get longest suffix of T reverse matched PST;
  Search_tree(C);              //根据后缀上下文序列 Context,
                    //从根节点到叶子节点的顺序倒序搜索
  r = arg max(p(rn+1 | C));
  return r;
END
```

预测的过程如下:对于给定的序列 $T = \{r_1, r_2, r_3, \cdots,$

$r_n\}$,T 的后缀为 $suffix(T)$，按 $suffix(T_1)$,$suffix(T_2)$,…,$suffix(T_n)$ 的顺序去倒序从根节点开始搜索概率后缀树，找到匹配的节点，该过程返回一个最长后缀序列倒序匹配成功的节点。取得该节点的概率向量中最大概率值对应的路段，即为预测结果。例如，给定预测序列 $T=\{r_1,r_2,r_3,r_4,r_5\}$，$T$的后缀依次为 $suffix(T_1)=\{r_1,r_2,r_3,r_4,r_5\}$，$suffix(T_2)=\{r_2,r_3,r_4,r_5\}$，$suffix(T_3)=\{r_3,r_4,r_5\}$，$suffix(T_4)=\{r_4,r_5\}$，$suffix(T_5)=\{r_5\}$。搜索 PST 树，按后缀序列的倒序序列进行，即$\langle r_5,r_4,r_3,r_2,r_1\rangle$，首先从 $suffix(T_1)$ 开始：

1）若没有成功匹配，则去掉一个最久远的元素即 r_1，一般情况下，时间最久的历史信息对未来的预测影响较小，按序列 $suffix(T_2)$ 倒序进行再次搜索匹配，即为$\langle r_5,r_4,r_3,r_2\rangle$，依次类推，直到匹配成功一个最长的后缀序列为止。

2）成功匹配后，定位到节点，在对应的节点处，从概率向量中找到一个最大概率值，所对应的路段即为预测的下一个路段。

3 实验

3.1 实验数据准备

为了验证算法的预测性能，进行模拟实验，实验硬件平台为 Intel Core i3-2310M CPU 2.1 GHz，内存为 2 GB；软件平台为 Windows XP，Visual C++.NET 2010，C++ 语言。实验数据来源于某省某市区 2013-01-03 至 2013-02-03 这一个月的浮动车数据，浮动车数据按时间先后存放在轨迹文件中，根据地图匹配算法，把 GPS 数据点呈现在地图上，假设每条轨迹数据不形成重复的回路和环路，将轨迹数据进行预处理，截取一段作为轨迹数据，取 10 000 条轨迹数据。其中随机抽取 80% 的轨迹数据进行训练，每条轨迹序列长度取值范围为 [30,100]，剩下 20% 的轨迹数据进行预测，预测的轨迹序列长度取值范围为[1,10]。实验中主要参数如表 1 所示。

表 1 实验中主要参数

参数	值	含义
m	50 ~ 300	路段集 R 中路段的个数
n	10000	训练集 S 中轨迹序列的条数
$\mid T\mid$	30 ~ 100	训练集 S 每条轨迹序列的长度
H	1 ~ 10	PST 的深度
L	1 ~ 10	给定预测路段序列的长度
pro_min	0.001	相对频率阈值

3.2 实验结果及分析

实现了 3 种模型，文献[8]中的历史模式树模型，马尔可夫模型和 PST 模型，实验随机进行 3 次，实验结果的预测平均值如图 4 所示。

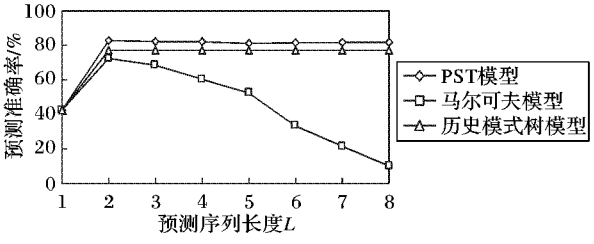

图 4 3 种模型 L 取 1 ~ 8 的平均预测结果

从实验结果可以看出，PST 模型在长度序列为 2 时，到达最好的预测效果，这与传统马尔可夫模型的结论是一致的。当给定的序列长度大于 2 时，PST 模型的预测效果保持相对不变，波动不大，能够自适应序列的长度进行有效预测，预测效果良好，而传统马尔可夫模型在阶数越大时，预测精度逐步降低。另外，本文的预测结果趋势与文献[8]提出的自适应多阶马尔可夫模型一致，平均预测精度稍有提高。原因在于本文中的模型使用概率阈值 pro_min 剔除掉了覆盖率较低的轨迹序列；而文献[8]中采用历史模式树记录所有的历史轨迹序列，存在一些覆盖率较低的轨迹序列，即在历史训练数据中出现次数较少的序列。

马尔可夫模型的空间复杂度随着阶数的增加呈指数趋势增长，设路段序列集 R 有 m 个元素，训练序列集 S 中训练序列的总长度为 N，预测序列长度为 L，则马尔可夫模型的空间复杂度为 $S1=O(m^L)$。而 PST 概率后缀树的空间复杂度只跟训练序列的总长度有关，空间复杂度为 $S2=O(N)$。在本文实验中，根据表 1 的数据，假定：$m=100$，$\mid T\mid=100$，$N=n*\mid T\mid=10000\times100$，则根据计算，当 $L=3$ 时，PST需要的状态空间和马尔可夫模型相同。令 $K=S1/S2$，随着预测序列长度 L 的增加，马尔可夫模型空间急剧膨胀。如图 5 所示。

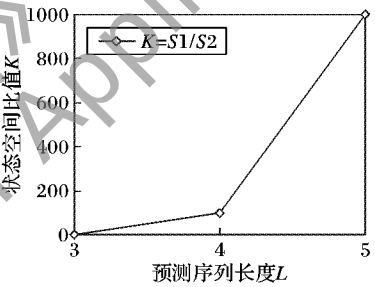

图 5 两种模型所需状态空间的比值

另外，pro_min 的设置对预测精度有一定的影响。阈值过大，虽可以减少训练数据的时间和空间开销，但会导致预测精度明显降低；阈值过小，开销加大，且不一定能保证预测精度的有效提高。因此，适当的阈值，可以减少训练数据的时间和空间开销，同时能保证一定的预测精度。

4 结语

本文针对定长马尔可夫模型预测方法的不足，提出了一种基于概率后缀树的动态自适应变长马尔可夫模型预测方法，给出了模型的算法和分析过程，详细阐明了轨迹训练和预测步骤和方法，采集交通路网中的浮动车轨迹数据进行实验。实验结果表明，该方法能取得较好的预测效果，有效解决传统马尔可夫高阶模型状态空间膨胀问题。该方法有效利用历史轨迹数据进行学习，结合当前的轨迹进行下一步轨迹预测，为轨迹预测提供了一种新的思路，能更好地提供基于位置的服务。下一步将进一步改进提高模型的性能，从两个方面进行：一是对算法的数据结构进行改进，提高算法的效率，达到更少的开销；二是从数据建模的角度，引入其他的度量尺度参数，如速度，行驶方向等，提高预测精度。

参考文献：

[1] ASAHARA A, MARUYAMA K, SATO A, *et al*. Pedestrian-movement prediction based on mixed Markov-chain model [C]// Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM Press, 2011: 25 - 33.

(下转第 3133 页)

算法对于较大规模的网络进行社区划分时稳定性有待提高,这也是下一步要做的工作。

4　结语

针对目前大部分基于智能进化算法的社区发现方法存在的问题,本文提出了一种基于免疫遗传的复杂网络社区划分算法,利用改进的字符编码方式结合种群初始化方法和遗传算子在不需要任何先验信息的情况下自动确定网络社区数并获得最优社区划分方案;在种群初始化和交叉算子中采用启发式方法使算法搜索的解空间更靠近问题的最优解空间,提高了算法的收敛速度,增强了算法的性能;基于免疫原理,在依据适应度选择的机制上,增加基于浓度的调节因子,保持了群体多样性,有效抑制了种群的退化现象。以上 4 个方面的工作使 CDIGA 能够更有效地发现复杂网络的社区结构。通过对 CDIGA 仿真实验的结果分析,表明该算法可自动获得最优社区数和社区划分方案,并且有效提高了社区划分的质量。本文进一步要做的工作一是通过对更多真实网络上的运行结果进行参数分析,更加合理地设定算法中采用的各参数,从而提高算法的稳定性;二是增强算法局部搜索能力,进一步提高算法的精度。

参考文献:

[1] 罗锦坤,元昌安,杨文,等.基于基因表达式编程算法的复杂网络社区结构划分[J].计算机应用,2012,32(2):317 - 321.

[2] KERNIGHAM B W, LIN S. An efficient heuristic procedure for partitioning graphs [J]. Bell System Technical Journal, 1970, 49(2): 292 - 307.

[3] POTHEN A, SIMON H, LIU P, *et al.* Partitioning sparse matrices with eigenvectors of graphs [J]. SIAM Journal on Matrix Analysis and Applications, 1990, 11(3): 430 - 452.

[4] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821 - 7826.

[5] NEWMAN M E J. Fast algorithm for detecting community structure in networks [J]. Physical Review E, 2004, 69(6): 066133.

[6] 何东晓,周栩,王佐,等.复杂网络社区挖掘—基于聚类融合的遗传算法[J].自动化学报,2010,36(8):1160 - 1170.

[7] 朱大勇,侯晓荣,张新丽.遗传聚类的社团发现[J].智能系统学报,2009,4(1):81 - 84.

[8] SHANG R H, BAI J, JIAO L C, *et al.* Community detection based on modularity and an improved genetic algorithm [J]. Physica A: Statistical Mechanics and its Applications, 2013, 392(5): 1215 - 1231.

[9] SHI C, YAN Z Y, WANG Y, *et al.* A genetic algorithm for detecting communities in large-scale complex networks [J]. Advances in Complex Systems, 2010, 13(1): 3 - 17.

[10] LIU X, LI D Y, WANG S L, *et al.* Effective algorithm for detecting community structure in complex networks based on GA and clustering [C]// Proceedings of the 7th International Conference on Computational Science. Berlin: Springer, 2007: 657 - 664.

[11] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks [J]. Physical Review E, 2004, 69(22): 1 - 15.

[12] AGUSTIN-BLAS L E, SALCEDO-SANZ S, JIMENEZ-FERNANDEZ S, *et al.* A new grouping genetic algorithm for clustering problems [J]. Expert Systems with Applications, 2012, 39(10): 9695 - 9703.

[13] 周世兵,徐振源,唐旭清.新的 K-均值算法最佳聚类数确定方法[J].计算机工程与应用,2010,46(16):27 - 31.

[14] TASGIN M, HERDAGDELEN A, BINGOL H. Community detection in complex networks using genetic algorithms [EB/OL]. [2013-04-16]. http://arxiv.org/abs/0711.0491.

[15] 郭世泽,陆哲明.复杂网络基础理论[M].北京:科学出版社,2012:270 - 271.

(上接第 3122 页)

[2] GIDÓFALVI G, BORGELT C, KAUL M, *et al.* Frequent route based continuous moving object location and density prediction on road networks [C]// Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM Press, 2011: 381 - 384.

[3] YING J J C, LEE W C, WENG T C, *et al.* Semantic trajectory mining for location prediction [C]// Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM Press, 2011: 34 - 43.

[4] ABRAHAM S, SOJAN LAL P. Spatio-temporal similarity of network-constrained moving object trajectories using sequence alignment of travel locations [J]. Transportation Research Part C: Emerging Technologies, 2012, 23: 109 - 123.

[5] 张伟,柳先辉,丁毅,等.基于支持向量回归的多时间序列自回归方法[J].计算机应用,2012,32(9):2508 - 2511.

[6] 余雪岗,刘衍珩,魏达,等.用于移动路径预测的混合 Markov 模型[J].通信学报,2006,27(12):61 - 69.

[7] 彭曲,丁治明,郭黎敏.基于马尔可夫链的轨迹预测[J].计算机科学,2010,37(8):189 - 193

[8] 吕明琪,陈岭,陈根才.基于自适应多阶 Markov 模型的位置预测[J].计算机研究与发展,2010,47(10):1764 - 1770.

[9] KRUMM J. A Markov model for driver turn prediction [EB/OL]. [2013-04-22]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.153.2524&rep=rep1&type=pdf.

[10] GIDÓFALVI G, DONG F. When and where next: individual mobility prediction [C]// Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems. New York: ACM Press, 2012: 57 - 64.

[11] RON D, SINGER Y, TISHBY N. The power of amnesia: Learning probabilistic automata with variable memory length [J]. Machine Learning, 1996, 25(2/3): 117 - 149.

[12] BEJERANO G, YONA G. Variations on probabilistic suffix trees: statistical modeling and prediction of protein families [J]. Bioinformatics, 2001, 17(1): 23 - 43.

[13] LEONARDI F G. A generalization of the PST algorithm: modeling the sparse nature of protein sequences [J]. Bioinformatics, 2006, 22(11): 1302 - 1307.

[14] LIN J, JIANG Y, ADJEROH D. The virtual suffix tree [J]. International Journal of Foundations of Computer Science, 2009, 20(6): 1109 - 1133.

[15] LIN J, ADJEROH D, JIANG B H. Probabilistic suffix array: efficient modeling and prediction of protein families [J]. Bioinformatics, 2012, 28(10): 1314 - 1323.