

基于模糊理论的不确定轨迹模式挖掘

李帆, 夏士雄, 张磊

(中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116)

摘要: 轨迹模式挖掘是轨迹数据知识发现的一个重要课题. 由于定位设备的局限性, 采集到的轨迹数据具有不确定性. 着眼于不确定轨迹数据模式挖掘, 通过引入模糊集方法, 提出不确定轨迹数据模式挖掘方法. 首先用均匀网格划分对象的运动平面, 基于对象实际位置的概率分布计算轨迹点相对于每个网格的隶属度, 通过不确定轨迹兴趣区域发现算法, 得到所有的不确定兴趣区域, 然后在此基础上进行轨迹模式挖掘. 实验展示了所提出的不确定轨迹数据模式挖掘方法进行模式挖掘的效果.

关键词: 不确定轨迹数据; 模糊集; 网格; 不确定轨迹模式

中图分类号: TP312

文献标识码: A

文章编号: 1000-7180(2011)08-0070-04

Uncertain Trajectory Pattern Mining Based on Fuzzy Theory

LI Fan, XIA Shi-xiong, ZHANG Lei

(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: Trajectory pattern mining is an important topic of knowledge discovery. Due to the limitations of location device, the trajectory data are not certain. The paper focuses on mining uncertain trajectory patterns. With the introduction of fuzzy set, a novel method is proposed to deal with uncertain trajectory data pattern mining. Firstly, the work space is divided by a grid whose cells are size uniform. The degree of membership of uncertain points relative to cells is calculated based on the probability distribution of moving object's real position. Then obtain the regions through the uncertain regions of interest extracting method. Uncertain trajectory pattern mining is done based on regions of interest. Finally, the experiments show the effects of uncertain trajectory pattern mining using fuzzy theory.

Key words: uncertain trajectory data; fuzzy set; grid; uncertain trajectory pattern

1 引言

得益于具有 GPS 功能的设备的广泛使用, 人们能很容易收集到大量记录了移动对象地理位置的数据, 通过分析这些数据, 可得到移动对象的经常性运动轨迹, 表现为轨迹模式^[1], 利用这些模式可以进行下一步的研究或在某些商业领域里的应用.

轨迹聚类是用于发现轨迹模式的方法之一, 韩家炜提出“分割—组合”^[2]的方法进行轨迹聚类. 文献^[1, 3]提出通过把轨迹数据转化为兴趣区域的集

合, 利用时间序列模式挖掘方法进行轨迹模式挖掘. 文献^[4]针对不确定轨迹数据, 提出“轨迹组”的概念表示轨迹模式. 文献^[5]中提到对于不确定轨迹数据, 每个点应该对应一个 ϵ -邻域. 模糊理论是处理不确定的重要方法, 文献^[6]在处理不确定轨迹数据的时候, 首先把运动平面用网格划分, 引入直觉模糊集理论, 用隶属度、犹豫度来表达不确定点与网格单元格的关系.

局限于移动设备本身的特性和无线数据传输质量, 采集到的移动对象位置信息具有不确定性, 因此

收稿日期: 2011-05-13; **修回日期:** 2011-06-27

基金项目: 国家自然科学基金项目(50674086); 高等学校博士学科点专项科研基金项目(20060290508); 江苏省博士后基金(0802023C)

需要能解决位置不确定的移动对象的轨迹模式挖掘方法. 本文引入模糊集方法, 提出不确定轨迹数据模式挖掘方法.

2 轨迹模式挖掘

通常, 轨迹数据以四元组形式存储, (O_{obj}, t, x, y) , 其中 O_{obj} 表示移动对象, t 为采集时间, (x, y) 是时刻 t 对象 O_{obj} 的坐标. 相同 O_{obj} 的轨迹数据通过插值法进行轨迹重建, 可得到时间段 $[t_i, t_j]$ 内的运动轨迹.

轨迹是采集到的移动对象位置数据按照时间顺序通过轨迹重建得到的序列 $\langle T, A \rangle$, 其中 $T = \{(x_0, y_0), (x_1, y_1), \dots, (x_k, y_k)\}$ 是位置信息, $A = \{t_0, t_1, \dots, t_k\}$, 是与位置信息对应的数据采集时间.

轨迹模式^[1]是轨迹的集合, 这些轨迹在相似的行程时间内覆盖了相似的地点序列. 轨迹模式为二元组 $\langle S, A \rangle$, $S = \{(x_0, y_0), (x_1, y_1), \dots, (x_k, y_k)\}$, $(x_i, y_i) \in R^2$, $0 \leq i \leq k$ 是位置序列, $A = \{a_1, a_2, \dots, a_k\}$, $a_j \in R$, $1 \leq j \leq k$ 是 S 中各点间的过渡时间. 轨迹模式可表示为 $(S, A) = (x_0, y_0) \xrightarrow{a_1} (x_1, y_1) \xrightarrow{a_2} \dots \xrightarrow{a_k} (x_k, y_k)$.

轨迹模式可看作对象频繁到达的空间区域的顺序序列, 因此轨迹模式可以表示为区域序列模式 $R_0 \xrightarrow{a_1} R_1 \xrightarrow{a_2} \dots \xrightarrow{a_k} R_k$, $(x_i, y_i) \in R_j$ 是 R_j 内的点.

3 不确定轨迹数据模式挖掘

由于移动设备本身的局限性和数据传输质量的影响, 在 t 时刻对移动对象位置的测量值与对象的真实位置有偏差, 假设移动对象的真实位置在以采样位置为均值的邻域内服从概率分布.

模糊理论用于描述对象个体与集合的关系. 数据的不确定性影响对象与集合的所属关系, 模糊集使用隶属度表示对象与集合的所属程度. 文中结合模糊集理论和移动对象位置的概率分布处理轨迹数据的不确定性.

3.1 不确定轨迹

采集的轨迹数据代表移动对象的期望位置, 通常情况下, 移动对象的真实位置处在期望位置为中心的某个范围之内, 可以用式(1)表示, 见图1(a).

$$F(p) = \begin{cases} (x_i, y_i), (x_i, y_i) \in \epsilon\text{-buffer} \wedge p \in \epsilon\text{-buffer} \\ 0, \text{other} \end{cases} \quad (1)$$

定义1 (不确定轨迹) 指由不确定点组成的序

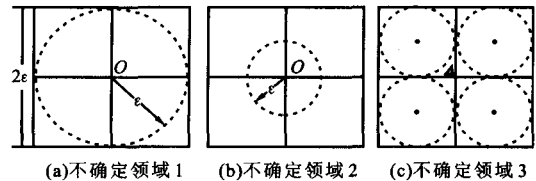


图1 移动对象的真实位置

列 $\langle UT, A \rangle$, 其中 $UT = \{r_0, r_1, \dots, r_k\}$ 是不确定点, 不确定点可表示为 ϵ -邻域, 即 $r_i = \{(x_{ij}, y_{ij}) \mid (x_{ij}, y_{ij}) \in \epsilon\text{-邻域}\}$, $0 \leq i \leq k, j \in \mathbb{Z}^+, A = \{t_1, t_2, \dots, t_k\}$ 表示采样时刻.

采用 ϵ -邻域表示不确定对象的位置, ϵ -邻域的大小是与移动对象的运动速度、测量设备的精度相关的量. 假设移动对象的真实位置在 ϵ -邻域内以测量值为均值服从高斯分布. 据“3 σ 法则”, 正态变量的值落在 $(\mu - 3\sigma, \mu + 3\sigma)$ 内几乎是肯定的, 基于不确定邻域 ϵ 大小的定义, 取 $\epsilon = 3\sigma$, 如图2所示.

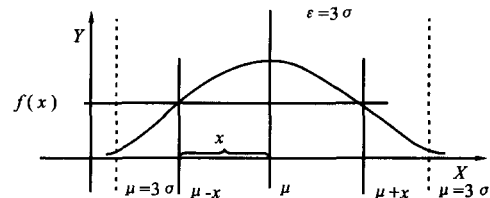


图2 不确定对象位置的高斯分布

本文基于模糊集理论, 提出不确定轨迹数据模式挖掘方法. 首先使用网格划分运动平面, 单元格的大小是基于不确定点邻域大小来确定的, 为了确保不确定点能完全落在某一个单元格内, 假定每个单元格为边长 $d = 2\epsilon$ 的正方形, 这样不确定点最多能覆盖四个单元格, 如图1(a)、(b)所示. 如果不确定点 P 的 ϵ -邻域部分或全部落在单元格 Cell 中, 那么称不确定点 P 属于单元格 Cell, 记做 $p \subseteq \text{Cell}$.

每个不确定点与相邻单元格的隶属关系通过隶属度确定. 设不确定点期望位置与其覆盖单元格中心位置距离为 $x \geq 0$, 那么有式(2):

$$\mu_p(x) = \begin{cases} 1 - \int_{\mu-x}^{\mu+x} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), & 0 \leq x \leq \epsilon \\ 0, & x > \epsilon \end{cases} \quad (2)$$

如果点在 Δ 区域, 如图1(c), 该区域中的点与任何单元格中心的距离都大于 ϵ , 称这样的点为奇异点, 需要根据移动对象的实际情况作不同处理.

3.2 不确定轨迹模式挖掘

不确定轨迹模式是有序的不确定点集合. 可对不确定兴趣区域进行序列模式挖掘得到轨迹模式.

提取兴趣区域需要计算每个单元格的轨迹密

度. 把每个单元格的隶属度和作为单元格的密度.

定义 2 (密度单元格) 给定密度门限值 δ , 如果网格单元格 Cell 包含的不确定点的隶属度和不小于门限值, 见式(3), 则称该单元格为密度单元格.

$$\text{density_Cell} = \{ \text{Cell} \mid \sum_{i=0}^n \mu_{pi} \geq \delta \wedge pi \subseteq \text{Cell} \} \quad (3)$$

定义 3 (不确定兴趣区域) 给定密度门限值 $\bar{\delta}$, 覆盖移动对象运动平面的 $n \times m$ 网格 G , 单元格的密度 ρ , 不确定兴趣区域定义为: (1) 相邻单元格的集合; (2) 单元格的平均密度不小于门限值 $\bar{\delta}$; (3) 各区域互不相交; (4) 区域是最大的, 即如果有不确定兴趣区域 R 存在, 则不存在 R' 使得 $R \subseteq R'$ 成立.

定义 4 (不确定轨迹模式) 是二元组 $\langle US, A \rangle$, $US = \{r_0, r_1, \dots, r_k\}$, $r_i = \{(x_{ij}, y_{ij}) \mid (x_{ij}, y_{ij}) \in \varepsilon\text{-buffer}\}$, $0 \leq i \leq k, j \in \mathbb{Z}^+$, 是二维平面上不确定点的序列, $A = \{a_1, a_2, \dots, a_k\}$, $a_j \in R, 1 \leq j \leq k$ 是时间间隔.

综合上述, 不确定轨迹模式挖掘首先需计算网格中每个单元格的密度; 然后通过对密度单元格进行扩展, 得到不确定兴趣区域; 在不确定兴趣区域基础上进行序列模式挖掘, 得到轨迹模式.

算法 1: 计算每个单元格的轨迹密度. 如果运动平面被划分为 $n \times m$ 单元格组成的网格, 共有 K 个轨迹点, 最坏情况下时间复杂度为 $O(\max\{4k, n * m\})$.

Algorithm1 computeDensity(UT, G_0, δ)

Input: 轨迹集合 UT , 网格 G , 单元格密度门限 δ

Output: 网格中每个单元格的密度.

foreach C_{ij} in G_0 do

$\rho_{c_{ij}} = 0$

Is Density $_{c_{ij}}$ = false

foreach UT_i in UT do

foreach r_i in UT_i do

get overlap cell set Cells of r_i

foreach cell in Cells do

compute the membership μ_{r_i} of r_i for cell 式(2)

$\rho_{\text{cell}} += \mu_{r_i}$

foreach C_{ij} in G_0 do

if $\rho_{c_{ij}} \geq \delta$ do Is Density $_{c_{ij}}$ = true 式(3)

return G

可以通过扩展密度网格得到兴趣区域^[1]. 扩展密度单元格, 直到平均密度小于门限值或者没有可以扩展的方向为止. 遍历完所有密度单元格后就得到兴趣区域的集合. 平均密度门限值可以是与轨迹

模式支持度正相关的量, 也可以单独指定.

4 试验及分析

实验展示了不确定轨迹数据模式挖掘效果. 数据来自于“GeoLife”项目收集的 GPS 数据, 完整数据包含 165 名用户两年内的户外位置记录. 随机抽取 36 名用户的记录作为实验数据, 实验环境为: T2250 @1.73 Hz CPU, 1 GB 内存, windows XP 系统.

由于兴趣区域的密度门限是与支持度正相关的量, 所以实验一比较不确定轨迹数据模式挖掘运行时间与支持度的关系. 假设时间容忍度 $\Delta\tau=1$, 网格边长 $d=0.5$, 不确定邻域 $\text{epsilon}=0.25$, 平均密度门限 $\delta=0.8$. 从图 3 上可以看到, 总的时间基本处于 32~33 s; 从总体来看, 随着支持度的增大, 运行时间呈减小的趋势; 支持度为 0.06 时运行时间出现一个震荡点. 支持度影响兴趣区域计算过程和模式挖掘过程, 随着支持度增加, 挖掘的兴趣区域数量变化, 从而导致挖掘出的轨迹模式变化, 所以时间会有波动, 但整体呈下降趋势.

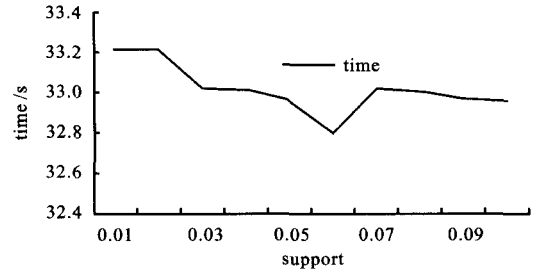


图 3 时间与支持度

实验二展示了不确定兴趣区域个数与设定的平均密度门限的关系. 假设支持度 $\text{support}=0.01$, 时间容忍度 $\Delta\tau=1$, 网格边长 $d=0.5$, 不确定邻域 $\text{epsilon}=0.25$. 从图 4 可以看到, 随着平均密度门限的增大, 能满足密度门限的网格区域集合减少, 兴趣区域数量减少.

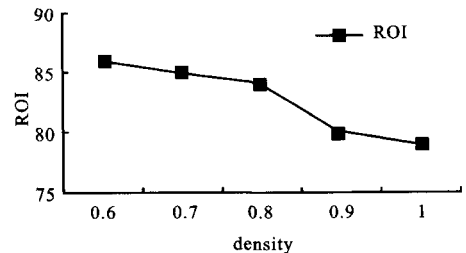


图 4 兴趣区域个数与密度门限

实验三展示了轨迹数量和兴趣区域个数的关系. 假设支持度 $\text{support}=0.01$, 时间容忍度 $\Delta\tau=1$,

网格边长 $d=0.5$, 不确定邻域 $\epsilon=0.25$, 平均密度门限 $\text{density}=0.8$. 随着轨迹数量增加, 兴趣区数量增加, 如图5所示.

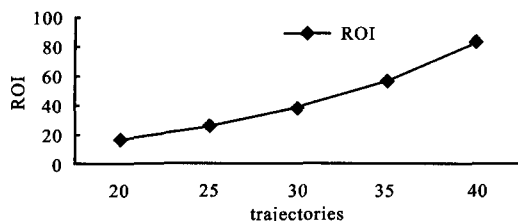


图5 兴趣区域个数与轨迹条数

5 结束语

本文针对轨迹数据的不确定性, 引入模糊集理论, 提出不确定轨迹数据模式挖掘方法. 方法使用网格划分对象运动平面; 针对每个不确定点用隶属度度量该点属于单元格的程度; 统计出每个单元格隶属度的和作为该单元格的密度; 扩展密度单元格得到兴趣区域; 在兴趣区域的基础上挖掘轨迹模式. 最后使用“GeoLife”项目收集的数据进行试验, 展示了本文提出的方法的效果.

参考文献:

[1] Fosca Giannotti, Mirco Nanni, Dino Pedreschi, et al.

Trajectory pattern mining [C]// Proc. KDD, 2007. USA; California, 2007.

[2] Lee Jae Gil, Han Jiawei, Whang Kyu Young. Trajectory clustering: a partition-and-group framework [C]// Proc. ACM SIGMOD Int. Conf. China; Beijing, 2007.

[3] Marcin Gorawski, Pawel Jureczek. Regions of interest in trajectory data warehouse [C]// ACIIDS 2010. Berlin; Springer Verlag, 2010.

[4] Jiong Yang, Meng Hu. TrajPattern: mining sequential patterns from imprecise trajectories of mobile objects [C]// EDBT 2006. Germany; Munich, 2006.

[5] Trajcevski G, Wolfson O, Hinrichs K, et al. Managing uncertainty in moving objects databases [J]. ACM TODS. 2004, 29(3): 463-507.

[6] Nikos Pelekis, Ioannis Kopanakis, Evangelos E Kotsifakos, et al. Clustering uncertain trajectories [M]. Berlin; Springer-Verlag, 2010.

作者简介:

李帆 男, (1986-), 硕士研究生. 研究方向为知识发现与数据挖掘.

夏士雄 男, (1961-), 博士, 教授, 博士生导师. 研究方向为智能信息处理.

张磊 男, (1977-), 博士, 副教授. 研究方向为数据挖掘.

(上接第69页)

统由2个压缩仿射变换 ω_1, ω_2 构成, 即原图像大小为11.6 kB, 但求解出分形压缩最优解 IFS 只需要存储 $2 \times 6 = 12$ 个实数, 占用48 B的存储空间, 压缩率约为4%.

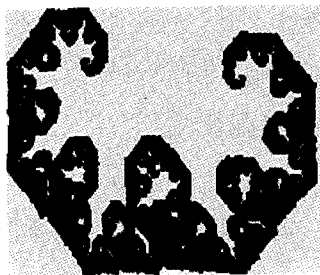


图3 GEP 重构图像

6 结束语

从基因表达式编程具有快速逼近和快速搜索能力出发, 构建 GEP 的分形二值图像压缩编码算法. 通过 GEP 的快速搜索和进化收敛能力求解二值图像最优解, 进化收敛速度比基于 GA 算法快. 而基于 GEP 的彩色图像分形图像压缩编码算法的研究是

我们下一步的研究工作.

参考文献:

[1] Dipankar Dasgupta, German Hernandez, Fernando Niño. An evolutionary algorithm for fractal coding of binary images [J]. IEEE Transactions on Evolutionary Computation, 2000, 4 (2): 172-181.

[2] 刘冠蓉, 郭京蕾, 何华. 基于遗传算法的二值图像压缩 [J]. 武汉理工大学学报: 信息与管理工程版, 2001, 23 (4): 9-13.

[3] Candida Ferreira. Gene expression programming: a new adaptive algorithm for solving problems [J]. Complex Systems, 2001, 13(2): 87-129.

[4] 安建成, 靳海娟. 自适应策略在实数编码遗传算法中的应用研究 [J]. 微电子学与计算机, 2011, 28(4): 140-142.

[5] 元昌安, 彭昱忠, 覃晓, 等. 基因表达式编程算法原理与应用 [M]. 北京: 科学出版社, 2010: 13-95.

作者简介:

李文敬 男, (1964-), 教授. 研究方向为演化算法、图像处理.