文章编号:1671 - 6833(2018)05 - 0058 - 05

位置数据稀疏约束下的疑犯时空位置预测

段 炼[1,2], 党兰学[3], 李 铭[4], 高 超[5], 朱欣焰[6]

(1. 广西师范学院 地理科学与规划学院,广西 南宁 530001; 2. 广西师范学院 北部湾环境演变与资源利用教育部重点实验室,广西 南宁 530001; 3. 河南大学 计算机与信息工程学院,河南 开封 475001; 4. 南昌大学 空间科学与技术研究院,江西 南昌 330031; 5. 警用地理信息技术公安部重点实验室,江苏 常州 213000; 6. 武汉大学 测绘遥感信息工程国家重点实验室,湖北 武汉 430079)

摘 要: 低强度的社会活动监控方式,使警方难以准确掌握疑犯的社会时空移动模式,也限制了嫌疑人排查及拦截围堵等警务行动的有效性. 为此,本文基于张量联合分解位置(tensor collective decomposition location prediction, TCDLP)模型,在疑犯时空位置数据的稀疏约束下,估算疑犯个体在任意时段的空间分布概率. 该方法利用三维张量表达各疑犯在多个时空节点上的访问强度,基于张量分解算法,融合多源社会环境数据所刻画的区域间关联性,解算出该张量中的缺失值,进而获取各疑犯的潜在时空分布模式. 实验使用包含了 241 个疑犯、约 1.9 万个位置记录的真实疑犯位置数据集进行模型测试,结果表明,本方法优于其他位置预测方法.

关键词: 疑犯时空预测; 张量分解; 犯罪预测; 位置预测

中图分类号: P208 **文献标志码:** A **doi:** 10.13705/j.issn.1671 - 6833.2018.05.003

0 引言

疑犯位置预测对探明疑犯作案时空规律、评估案发位置与疑犯关联性等警务需求有重要的应用价值[1]. 但由于位置探测源(如旅店登记系统、进出港登记系统、ATM 机等)数量和类型有限,警方仅能获取到他们稀疏的位置数据[2],严重影响了疑犯位置预测的准确性. 在犯罪地理学中,已有研究基于犯罪个体的系列犯罪位置序列,基于平均作案距离[3]、路网结构[4],利用距离衰减函数[4]、贝叶斯公式[5]和动力学模型[6]等,估算锚点(住址或未来犯罪地点等)[7]在空间上的出现概率. 然而,这些研究既没有考虑数据稀疏性的影响[8],也极少考虑时间因素. 近年来,基于车辆定位数据[9]、Wi-Fi 信号[10]、公共交通数据[11]、人员轨迹数据[12]和地理社交网络 check-in 数据[13]等的位置预测成为研究热点. 然而,疑犯位置数据较这些数据更加稀疏,也不存在好友关系等数据以提高预测精度. 为应对以上挑战,笔者融合疑犯群体的统计先验知识和社会环境信息,基于张量联合分解方法来估算疑犯在所有时空节点上的驻留概率.

1 问题描述

疑犯位置数据集包括了 W 市 2012 年 1 月至 2012 年 6 月间 241 名疑犯的 18 754 个轨迹点. 将研究区域网格化,获得 $g \times g$ 格网,$G = \{p_1, p_2, \cdots, p_i, \cdots, p_{g\times g}\}$. 本文中 $g = 100$,每个网格覆盖的范围约为 256 m × 224 m[9]. 如图 1 所示.

利用各时段(笔者将一天划分为 12 个时段,每个时段为 2 小时)不同疑犯在各网格上的驻留次数,构建三维张量 $Q \in \mathbb{R}^{U\times G\times T}$,表达“疑犯 - 位置 - 时段”的相互关系,如图 2 所示. 其中,U 为疑犯数量;G 为网格数量;T 为时段数量. 由于疑犯位置数据的稀疏性,Q 中仅有 1% 的项才具有数值. 因此,需解决的问题是:估算 Q 内所有缺失项.

收稿日期: 2018 - 02 - 01; **修订日期:** 2018 - 03 - 27

基金项目: 国家自然科学基金资助项目(41401524);广西自然科学基金资助项目(2015GXNSFBA139191);警用地理信息技术公安部重点实验室开放课题资助项目(2016LPGIT03);北部湾环境演变与资源利用教育部重点实验室系统基金资助项目(2014BGERLXT14);广西高校科学技术研究项目(KY2015YB189、KY2016YB281);河南理工大学国家测绘局矿山重点实验室开放基金资助项目(KLM201409)

作者简介: 段 炼(1981—),男,湖南祁阳人,广西师范学院副教授,博士,主要研究方向为时空数据挖掘与犯罪时空预测,E-mail:wtusm@163.com.

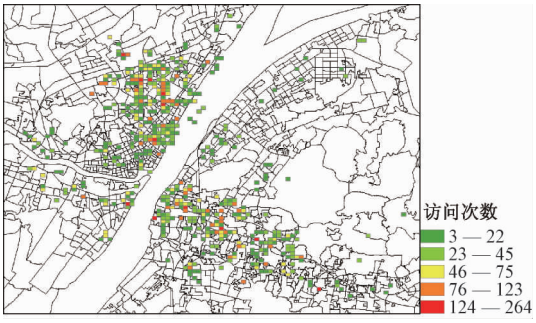

图 1　网格化后的疑犯空间分布强度

Fig. 1　Spatial distribution of suspects visiting density

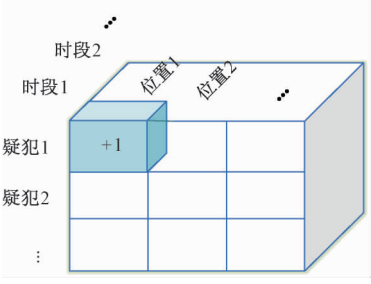

图 2　“疑犯-位置-时段”张量

Fig. 2　“Suspect-location-time” tensor

2　系统流程

本方法具体流程如图 3 所示. 首先,构建“疑犯-位置-时间”张量 Q. 其次,抽取所有疑犯在不同时空节点驻留的统计信息,构建“疑犯-位置”矩阵与“位置-时间”矩阵,表达疑犯对各时空节点的访问模式. 再将人口、路网和 POI 等信息按照网格尺度汇集,形成“位置-特征”矩阵,并利用出租车轨迹数据构建“位置-位置”矩阵,通过这两个矩阵描述位置间的关联性. 最终,对以上张量和矩阵进行协同分解,计算出“张量 Q 中的缺失值”,实现疑犯个体的时空预测.

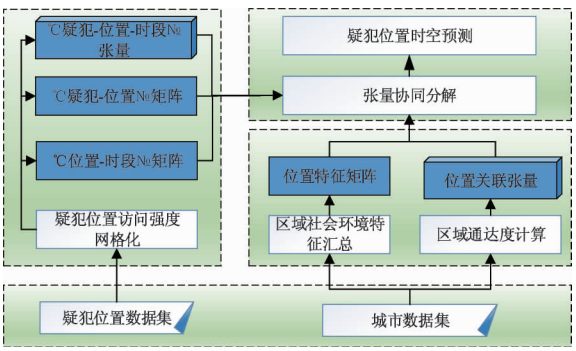

图 3　系统架构图

Fig. 3　System architecture

2.1　疑犯群体的位移特征抽取

基于疑犯位置数据,构建“疑犯-位置”矩阵 $\boldsymbol{E} \in \mathbb{R}^{U\times G}$,其中,$U$ 为疑犯总数;G 为网格总数. 该矩阵刻画各疑犯的全局空间分布模式.

为获得所有疑犯的全局时空分布模式,构建“位置-时间”矩阵 $\boldsymbol{D} \in \mathbb{R}^{G\times T}$,其中,$G$ 表示位置数量;T 表示一天内的所有时段数量. $\boldsymbol{D}$ 中第 i 行和第 j 列的项 $D(i,j)$ 表示所有疑犯在 j 时段访问 i 位置的次数.

2.2　位置特征抽取

2.2.1　位置-特征矩阵

具有类似社会经济环境的区域往往对疑犯具有类似的吸引力. 笔者涉及的社会经济环境信息包括 4 个部分:POI 特征集 ***Fp***、路网特征集 ***Fr***、房屋特征集 ***Fb*** 和人口统计特征集 ***Fc***. 据此,构建“位置-特征”矩阵 $\boldsymbol{C} \in \mathbb{R}^{G\times(p+r+b+c)}$,其中,$G$ 表示位置总数;p、r、b 和 c 分别表示 ***Fp***、***Fr***、***Fb*** 和 ***Fc*** 集的特征个数. 特别的,对于 category 类型的属性,将其转变为 1 和 0 表示的 one-hot 向量结构.

①POI 特征. POI 特征 ***Fp*** 包括:该位置内 POI 的空间密度以及 12 个类型的 POI 数量共 13 个特征. 为体现区域独有的社会经济环境特性. 借鉴 TF-IDF 方法,将位置 i 中类型为 j 的 POI 数量 q_{ij} 转换为 POI 类型重要度 Y_{ij},

$$Y_{ij} = \frac{q_{ij}}{\sum_{j}^{o} q_{ij}} \log \frac{|G|}{|\{q_i : q_{ij} > 0\}|}, \tag{1}$$

其中,o 为 POI 类型数量;$|G|$ 表示位置总数;$|\{q_i : q_{ij} > 0\}|$ 表示具有 POI 类型 j 的位置个数.

②路网特征. 路网特征 ***Fr*** 包括:该位置内的路口数量和 5 个等级(高速公路、一级公路、二级公路、三级公路及四级公路)的道路长度,共 6 个指标.

③建筑物特征. 笔者抽取的房屋特征 ***Fb*** 包括:楼房密度、5 类房屋(住宅型、商业性、行政型、工业型、其他)的数量分布、3 类高度(低层、多层、高层)房屋的数量分布,共 9 个指标.

④人口统计特征. 人口统计特征 ***Fc*** 涉及 10 个指标,分别是人口密度、4 个年龄段(18 岁以下、18 ~ 40 岁、40 ~ 60 岁、60 岁以上)的人口数量分布、5 类教育程度(文盲、初中、高中、大学、研究生)的人口分布.

2.2.2　位置可达性张量

位置间的空间邻近性和通勤强度体现了位置之间的疑犯转移倾向或流动的便捷程度. 下面利用出租车数据表达位置间的时态通勤强度,再结合空间邻近性,计算位置间的时空可达度.

设 v_{tij} 为 t 时段下从位置 i 到达位置 j 的出租车数量,v'_{tij} 为 t 时段下从位置 j 出发到达位置 i 的出租车数量,d_{ij} 为两位置的空间距离,则两位置在 t 时段下的关联度为:

$$p_{tij} = \frac{v_{tij} + v'_{tij}}{d_{ij}}. \tag{2}$$

基于上式,构建张量 $P \in \mathbb{R}^{T\times G\times G}$,将 p_{tij} 作为 P 中的项,得以刻画位置和位置之间的空间可达度.

3 多源数据融合下的张量分解

结合矩阵因子分解和张量因子分解方法计算出 Q 中的所有缺失项,以获取疑犯个体在任意时空节点的驻留概率. 张量 Q 可因此分解为:

$$Q \approx \boldsymbol{S} \times \boldsymbol{U} \times \boldsymbol{J} \times \boldsymbol{T}. \tag{3}$$

其中,核张量(core tensor) $S \in \mathbb{R}^{d^u\times d^g\times d^t}$,疑犯低阶潜在因子矩阵(low rank latent factors matrix) $\boldsymbol{U} \in \mathbb{R}^{U\times d^u}$、位置低阶潜在因子矩阵 $\boldsymbol{J} \in \mathbb{R}^{G\times d^l}$ 和时间低阶潜在因子矩阵 $\boldsymbol{T} \in \mathbb{R}^{T\times d^t}$,$d^u \leqslant u$,$d^l \leqslant g$,$d^t \leqslant t$(本文 $d^u = d^l = d^t$).

"疑犯-位置"矩阵 $\boldsymbol{E}$ 可因此分解为 $\boldsymbol{U}$ 和 $\boldsymbol{J}^{\mathrm{T}}$ 的乘积,即:

$$\boldsymbol{E} \approx \boldsymbol{U} \times \boldsymbol{J}^{\mathrm{T}}. \tag{4}$$

同理,"位置-时间"矩阵 $\boldsymbol{D} \approx \boldsymbol{J} \times \boldsymbol{T}^{\mathrm{T}}$,"位置-特征"矩阵 $\boldsymbol{C} \approx \boldsymbol{I} \times \boldsymbol{P}$ ($\boldsymbol{P} \in \mathbb{R}^{d^l\times(p+r+c)}$);位置可达性张量 $\boldsymbol{P} \approx \boldsymbol{W} \times \boldsymbol{J} \times \boldsymbol{J}^{\mathrm{T}}$,其中 $W \in \mathbb{R}^{d^l\times d^l\times d^t}$,$d^l \leqslant G$,$d^t \leqslant T$(本文中 $d^l = d^t$).

可见,$\boldsymbol{Q}$ 与 $\boldsymbol{E}$、$\boldsymbol{D}$、$\boldsymbol{C}$ 及 $\boldsymbol{P}$ 共享了潜在因子矩阵 $\boldsymbol{U}$、$\boldsymbol{J}$ 和 $\boldsymbol{T}$;$\boldsymbol{P}$ 也与 $\boldsymbol{E}$、$\boldsymbol{D}$ 以及 $\boldsymbol{C}$ 共享了潜在因子矩阵 $\boldsymbol{J}$ 和 $\boldsymbol{T}$. 依据这些信息交互关系,得到融合疑犯位移、社会经济环境和位置可达性数据的张量因子分解目标函数:

$$\begin{aligned} L(\boldsymbol{Q},\boldsymbol{S},\boldsymbol{W},\boldsymbol{U},\boldsymbol{J},\boldsymbol{T},\boldsymbol{P}) = {} & \frac{1}{2}\|\boldsymbol{Q} - \boldsymbol{S}\times\boldsymbol{U}\times\boldsymbol{J}\times\boldsymbol{T}\|^2 + \\ & \frac{\lambda_1}{2}\|\boldsymbol{P} - \boldsymbol{W}\times\boldsymbol{J}\times\boldsymbol{J}^T\times\boldsymbol{T}\|^2 + \frac{\lambda_2}{2}\|\boldsymbol{E} - \boldsymbol{U}\times\boldsymbol{J}^{\mathrm{T}}\|^2 + \\ & \frac{\lambda_3}{2}\|\boldsymbol{D} - \boldsymbol{J}\times\boldsymbol{T}^{\mathrm{T}}\|^2 + \frac{\lambda_4}{2}\|\boldsymbol{C} - \boldsymbol{I}\times\boldsymbol{P}\|^2 + \\ & \frac{\lambda_5}{2}(\|\boldsymbol{S}\|^2 + \|\boldsymbol{W}\|^2 + \|\boldsymbol{U}\|^2 + \|\boldsymbol{J}\|^2 + \|\boldsymbol{T}\|^2 + \|\boldsymbol{P}\|^2). \end{aligned} \tag{5}$$

其中,$\|\cdot\|$ 为 Frobenius 范数(norm);($\|\boldsymbol{S}\|^2 + \|\boldsymbol{W}\|^2 + \|\boldsymbol{U}\|^2 + \|\boldsymbol{J}\|^2 + \|\boldsymbol{T}\|^2 + \|\boldsymbol{P}\|^2$)作为正则惩罚项以防止模型过拟合;$\lambda_1$、$\lambda_2$、$\lambda_3$、$\lambda_4$ 和 λ_5 分别为目标函数中相应项的权重值,以表达各项在目标函数中的重要程度,当它们都为 0 时,目标函数退化成普通的 tucker 分解形式(tuckerdecomposition). 由于没有数值解析方法(closed-formsolution)能计算出该目标函数的全局最优解,我们基于 PARAFAC-style 张量分解方法[14]找出该目标函数的最优解.

4 试验与分析

试验硬件配置为 Intel (R) Core (TM) i777003.6 GHz (4 核),16 GB 内存的计算机,操作系统为 Windows 7,软件采用 MATLAB2 016 a 和 TensorToolbox 包[17]. 采用均方根误差和 top-k 最近距离作为模型性能的评价指标,其中:均方根误差(*RMSE*)为预测值与真实值之间的误差累加均方根,

$$RMSE = \sqrt{\frac{\sum_{t=1}^{G}(\hat{y_t} - y_t)^2}{G}}, \tag{6}$$

式中:$\hat{y_t}$ 为 Q 第 t 个项的预测值;y_t 为真实值. 由于 baseline 方法中一些模型的输出结果为概率值,因此这些模型不采用该指标进行比较.

Top-k 最近距离(*SED*@k):目标位置与前 top-k 个预测结果的最小距离.

$$SED@k = \min[dis(\hat{y_t}, y_t)], t = 1,\cdots,k. \tag{7}$$

该指标越小越好,本文中 $k = 10$. 两网格间的距离为它们的中心间距.

4.1 比较方法

笔者所提方法称为 TCDLP. Baseline 方法.

①时态约束下的 Kriging 克吕格插值法(TK):基于每个时间槽内空间邻近位置的访问次数作为目标位置的访问次数.

②层次 Pitman-Yorprocess 语言统计模型(HPHD):描述用户在各位置上的语义时间访问强度. 该方法无法对未知位置建模.

③HOSVD[15]:仅对"疑犯-位置-时间"张量进行因子分解来获取其缺失值.

试验采用交叉验证,随机从疑犯位置数据集抽取 70% 为训练数据,20% 位验证数据,10% 作为测试数据.

4.2 模型性能比较

TCDLP 的参数 $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = 0.05$,各潜在因子数量 $k = 10$. 表 1 为各模型在 *RMSE* 和 *SED*@10 上的性能. 笔者提出的模型在这 3 个指标上都优于其他 3 种方法,说明融合多源城市社会经济环境数据对疑犯时空节点估算是有效的. TK 的

各项指标性能值均为最差,说明在数据稀疏情况下,空间邻近性还无法充分刻画疑犯位置分布的时空模式. 基于矩阵/张量分解的方法(如 TCDM 和 HOSVD)的各项性能指标均超过了 TK,这表明,位置间的环境相似性能为疑犯时空分布模式的挖掘提供有效信息. 由于 *HPHD* 给出的结果为概率形式,因此无法对其进行 *RMSE* 指标测试.

表 1　各模型的预测性能

Tab. 1　Models performances, by "mean ± std".

模型	*RMSE* 指标	*SED*@10 指标
TK	4.64 ± 1.43	124 ± 65
HPHD	—	108 ± 62
HOSVD	3.53 ± 0.63	92 ± 59
TCDLP	2.04 ± 0.82	81 ± 58

4.3　TCDLP 参数影响分析

让 $\lambda_1 \sim \lambda_5$ 在 0 ~ 10 变化,观察 TCDLP 方法在 *RMSE* 和 *SED*@10 两个指标的变化,如图 4 所示. 验证各外部环境信息 ***E***、***D***、***C*** 和 ***P*** 对疑犯位置预测性能的影响. 由图 4 可知,集成了外部环境信息后,模型预测性能有了较大提升,*RMSE* 和 *SED*@10 的变化较大;但随着各参数的增加,相对于 *RMSE*、*SED*@10 的变化幅度不大,这再次验证了疑犯的社会活动趋向于集聚性. 随着 λ_3 的增加,模型的 *RMSE* 和 *SED*@10 都有明显提升,说明位置间的社会环境相似性对疑犯社会移动具有显著的影响. 然而,一旦 λ_4 和 λ_5 增加到一定数值,模型的 *RMSE* 急速下降,*SED*@10 也有一定的上升,这可能是疑犯位置关联性数据中存在噪声,λ_4 和 λ_5 的增加放大了这样的噪声,造成模型性能降低.

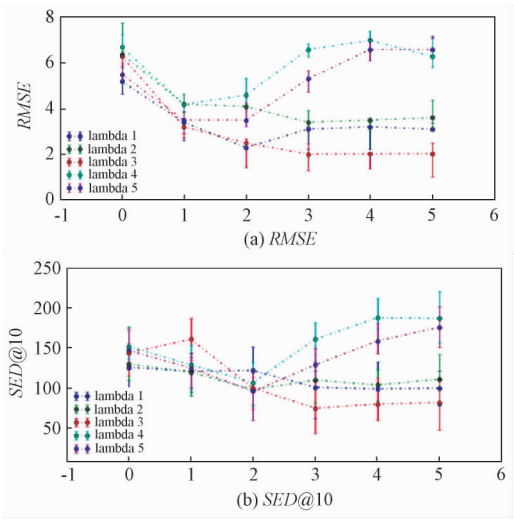

图 4　$\lambda_1 \sim \lambda_5$ 对 *RMSE* 和 *SED*@10 的影响

Fig. 4　Impact of $\lambda_1 \sim \lambda_5$ on *RMSE* and *SED*@10

5　结论

提出基于张量协同分解模型估算疑犯的潜在时空分布概率算法. 该算法引入社会环境信息,通过张量和矩阵的联合分解估算疑犯位置时空分布,缓解了疑犯位置数据的稀疏性. 基于真实疑犯位置跟踪数据的实验结果表明,笔者所提算法在 *RMSE* 和 *SED*@10 两个指标上分别平均高于其他 baseline 方法 50% 和 18%. 今后的工作将对疑犯进行分类,如盗窃类、抢劫类等,针对不同犯罪类型特点设计算法,进一步提高算法的精度.

参考文献:

[1] Office of the privacy commissioner of canada [EO/BL]. Available online: https://www.priv.gc.ca/en/ (accessed on 10th Mar, 2016).

[2] 孙楠. 警用多源数据轨迹分析系统设计与实现[J]. 测绘科学, 2013, 38(5): 51 - 53.

[3] CHEN N C, WEI S, SONG D W. Prediction of series criminals: An approach based on modeling[C]// International Conference on Computational and Information sciences, 2010: 72 - 75.

[4] KENT J D, LEITNER M. Incorporating land cover within bayesian journey-to-crime estimation models [J]. International journal of psychological studies, 2012, 4(2): 120 - 140.

[5] BAUMGARTNER K C, FERRARI S, SALFATI C G. Bayesian network modeling of offender behavior for criminal profiling[C]// IEEE Conference on Decision and Control, 2005:2702 - 2709.

[6] MOHLER G O, SHORT M B. Geographic profiling from kinetic models of criminal behavior[J]. Siam journal on applied mathematics, 2012, 72 (1):163 - 180.

[7] MARTINEAU M, ERIC B. Journey to murder: Examining the correlates of criminal mobility in sexual homicide[J]. Police practice and research. 2016, 17 (1):68 - 83.

[8] YANG A M, WU R J, WU H M, et al. The Research of Tree Topology Model for Growth of Natural Selection and Application in Geographical Profile for Criminal [C]// Information Computing and Applications International Conference, 2010, 106: 383 - 390.

[9] BRÉBISSON A D, SIMON É, AUVOLAT A, et al. Artificial neural networks applied to taxi destination prediction[C]// European Conference on Principles of Data Mining and Knowledge Discovery, 2015.

[10] SONG L, KOTZ D, JAIN R, et al. Evaluating location predictors with extensive wifi mobility data[C]//

Joint Conference of the IEEE Computer and Communications Societies, IEEE, 2004, 2 (4): 1414 – 1424.

[11] YUAN N J, WANG Y, ZHANG F, et al. Reconstructing individual mobility from smart card transactions: A space alignment approach[C]// International Conference on Data Mining, IEEE, 2014, 44 (2): 877 – 886.

[12] 胡燕, 朱晓瑛, 马刚. 基于 K-Means 和时间匹配的位置预测模型[J]. 郑州大学学报(工学版), 2017, 38(2): 17 – 20.

[13] JURGENS D, FINETHY T, MCCORRISTON J, et al. Geolocation prediction in twitter using social networks: a critical analysis and review of current practice[C]// AAAI, 2015:129 – 141.

[14] CICHOCKI A, ZDUNEK R, PHAN A H, et al. Nonnegative matrix and tensor factorizations: applications to exploratory multiway data analysis and blind source separation [J]. Wiley publishing, 2009, 25 (Q2): 1 – 3.

[15] VERVLIET N, DEBALS O, SORBER L, et al. Tensorlab 3.0[CP/OL]. https://www.tensorlab.net/.

Spatiotemporal Prediction of Suspect under Location Data Sparsity Constraint

DUAN Lian[1,2], DANG Lanxue[3], LI Ming[4], GAO Chao[5], ZHU Xinyan[6]

(1. School of Geographical Sciences and Planning, Guangxi Teachers Education University, Nanning 530001, China; 2. Education Ministry Key Laboratory of Environment Evolution and Resources Utilization in Beibu Bay, Guangxi Teachers Education University, Nanning 530001, China; 3. College of Computer and Information Engineering Henan University, Kaifeng 475001, China; 4. Institute of Space Science and Technology, Nanchang University, Nanchang 330031, China; 5. Key Laboratory of Police Geographic Information Technology, Ministry of Public Security, Changzhou 213000, China; 6. State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China)

Abstract: Due to the low monitoring intensities on key tracking persons(suspects), the police suffered from the very small amounts of suspect social location data, which was hard to effectively reveal the social mobility patterns of suspects, and restrict the police action validity for suspects filtering and crime blockading etc. Facing this data sparsity challenge, a novel Tensor Collective Decomposition Location Prediction (TCDLP) model was proposed, to estimate the latent visiting intensity at an arbitrary spatiotemporal node. Specifically, it modeled the visiting intensities of suspects with 3D tensor, where the three dimensions stood for suspects, locations, and time slots respectively. Then, the missing entries in the tensor would be filled through a multi-data fusion tensor decomposition approach, which integrated the correlations of locations and suspects relying on multiple social environment data. So by supplementing the visiting intensities in this tensor, the social spatiotemporal distribution pattern for each suspect could uncovered. TCDLP was evalvated by using a real-world suspect dataset collected from 241 suspects over 6 months with about 19 thousands location records, showing our model outperformed state-of-the-art approaches to the problem.

Key words: suspect spatiotemporal prediction; tensor decomposition; crime prediction; location prediction