

# 基于转移概率矩阵自学习的犯罪分布预测

魏新蕾， 颜金尧， 石拓， 张园

(中国传媒大学 信息工程学院, 北京 100024)

**摘要：**针对犯罪分布预测准确率低, 历史犯罪数据缺失严重的问题, 提出了基于历史犯罪数据, 融合所研究地区的社会环境因素的转移概率矩阵自学习的犯罪分布预测算法——TWcS。将包括距离信息、面积信息、人口信息在内的社会环境因素作为权重值引入到梯度下降策略中, 利用梯度下降实现 TWcS 算法的转移概率矩阵自学习。实验结果证明, TWcS 算法的性能明显优于包括当前最优基线算法(TPML-WMA)在内的其他预测算法(如 LR、AR、Lasso 回归算法、贝叶斯算法、决策树算法等), TWcS 算法的 MAE 值是其他算法 MAE 平均值的 33%。

**关键词：**犯罪分布预测；转移概率矩阵；梯度下降法

中图分类号：TP 399 文献标志码：A 文章编号：1001-0645(2020)01-0098-07

DOI：10.15918/j.tbit.1001-0645.2018.042

## Predicting Crime Distribution Based on Transition Probability Matrix Self-Learning Algorithm

WEI Xin-lei, YAN Jin-yao, SHI Tuo, ZHANG Yuan

(School of Information Engineering, Communication University of China, Beijing 100024, China)

**Abstract:** Aiming at the problem of low accuracy of crime distribution prediction and serious lack of historical crime data, a crime distribution prediction algorithm, TWcS, was proposed based on a transition probability matrix model, the historical crime data and integrating social environmental factors in the studied area. In this paper, the social environment factors including distance information, area information and population information were introduced as weights into the gradient descent strategy, and the transition probability matrix self-learning of TWcS algorithm was realized by gradient descent. The experimental results show that the performance of TWcS algorithm is superior to other prediction algorithms including TPML-WMA, LR, AR, Lasso regression algorithm, Bayesian algorithm, decision tree algorithm, etc. The MAE value of TWcS algorithm is only 33% of the average MAE value of the other algorithms.

**Key words:** crime distribution prediction; transition probability matrix; gradient descent

犯罪时空分析研究旨在根据历史犯罪数据来预测犯罪发生的时间、地点以及犯罪类型等信息, 对维护公共安全具有重要意义, 在学术界也引起越来越多的关注<sup>[1]</sup>。研究者们通常采用许多经典的机器学习和模式识别方法来解决这个问题<sup>[2-5]</sup>。但是, 以往的研究忽视了历史犯罪数据与社会环境因素(socio-

environmental factor)的充分结合。为了克服这一缺陷, 本文基于历史犯罪数据, 将相关社会环境因素(主要包括距离、人口和面积等)融入转移概率矩阵自学习, 来研究犯罪分布预测, 提出了一种全新的犯罪分布预测算法——TWcS(TPML-WMA considering socio-environmental factor) 算法, 作为 TPML-

收稿日期：2018-01-11

基金项目：国家自然科学基金面上项目(61971382); 中国传媒大学中央高校基本科研业务费专项资金资助

作者简介：魏新蕾(1981—), 女, 博士, E-mail: wei-xinlei@cuc.edu.cn。

通信作者：颜金尧(1973—), 男, 教授, 博士生导师, E-mail:jyan@cuc.edu.cn。

WMA (transition probability matrix learning and weighted moving average) 算法<sup>[6]</sup>的优化版本。TPML-WMA 算法是目前解决犯罪分布问题的基线算法,与传统的机器学习算法(例如线性回归算法等)相比,这个基于转移概率矩阵的算法在预测性能方面有显着的改善。然而,遗憾的是 TPML-WMA 算法并没有考虑人口分布、面积分布以及任意两个区域之间的距离等社会环境因素,导致其性能受限。

以往犯罪时空分析相关研究工作通常基于历史犯罪数据,采用经典的机器学习和模式识别方法来对该问题进行建模和预测,包括如线性回归(linear regression)算法、自回归(auto regression)算法、Lasso 回归(Lasso regression)算法、决策树(decision tree)算法和贝叶斯(Bayesian)算法等<sup>[1-6]</sup>。

还有一些研究侧重探索如何将机器学习算法引入到时空模型中,进而得到关于犯罪问题的一系列时空域结果<sup>[7-9]</sup>。文献[10-11]引入了自回归积分移动平均(auto-regressive integral moving average, ARIMA)模型,研究某城市在数个星期内的犯罪率预测。文献[6]提出基于转移概率矩阵学习的 TPML-WMA 算法用于分析犯罪率分布问题。

此外,许多研究表明犯罪的发生往往伴随着相应的地理位置特征,而地理位置信息的引入展现出良好的应用效果<sup>[12-13]</sup>,也有研究尝试结合 GIS 系统与犯罪分析<sup>[14]</sup>。例如,着重考虑到地理位置,文献[15]提出了一个加权回归模型来生成犯罪的空间分布,并讨论地理因素之间的关系;区域性加权回归模型也被广泛用于实现犯罪分布预测<sup>[16]</sup>。

近年来,随着大数据采样和互联网信息处理技术的发展,犯罪行为与被研究地区的社会因素、经济因素、环境因素之间的关联关系也越来越受到重视<sup>[17]</sup>。但是上述社会环境因素尚未被引入犯罪分布预测研究中,导致相关研究的性能受限,不能对犯罪分布相关因素进行更全面的建模。

因此,基于 TPML-WMA 算法,本研究创新性地将距离信息、面积信息、人口信息建模为权重值引入到梯度下降策略中;在此基础上,将自学习策略引入到转移概率矩阵中,使其能够根据预测值与实际值之间的差异反馈实现自我调节,最终达到稳定。综上,本研究提出用于预测犯罪分布的 TWcS 算法。

在实验中,基于公安部门内部盗窃案例数据集(2006 年—2016 年),文本所提出的 TWcS 算法与目前最优基线算法 TPML-WMA 以及其他基线算

法(线性回归算法、自回归算法、Lasso 回归算法、贝叶斯算法和决策树算法等)进行对比。实验结果表明, TWcS 算法的性能明显优于其他对比算法。

## 1 融合社会环境因素的犯罪分布预测算法 TWcS

### 1.1 社会环境因素建模

针对以往研究在预测犯罪分布的时候忽略社会环境因素的不足,本研究探索如何将社会环境因素引入到转移概率矩阵自学习的过程中。

以往研究已经证明,居民人口因素、家庭因素、社会经济因素、住宅因素等都与犯罪率存在某种关联关系。本研究根据一线警务人员积累的工作经验,最终选择 3 个最常用、最相关的因素来完成本文的研究:距离(distance)、面积(area)和人口(population);并利用上述 3 个因素来调整转移概率矩阵中元素的值。选择上述因素的原因在于,相对于其他因素,这 3 个因素与犯罪率更加相关。下面对距离因素、面积因素和人口因素进行分别建模。

将整个研究地区按行政区域划分为若干个子块(administrative district),这些行政区块用  $\{S_1, S_2, \dots, S_n\}$  表示。相应行政区的犯罪频率分别表示为  $\{x^1, x^2, \dots, x^n\}$ 。因此,第  $i$  个年度相应行政区的犯罪率可以建模成一个向量  $x_i = [x_i^1 \ x_i^2 \ \dots \ x_i^n]^T$ ,  $i=1, 2, \dots, m$ 。综上,向量运动(vector motion, VM)模型可以表示为

$$\begin{bmatrix} a_i^{11} & \dots & a_i^{1n} \\ a_i^{21} & \dots & a_i^{2n} \\ \dots & & \dots \\ a_i^{n1} & \dots & a_i^{nn} \end{bmatrix} \begin{bmatrix} x_i^1 \\ x_i^2 \\ \dots \\ x_i^n \end{bmatrix} = \begin{bmatrix} x_{i+1}^1 \\ x_{i+1}^2 \\ \dots \\ x_{i+1}^n \end{bmatrix}. \quad (1)$$

其中,对于每个  $i \in [1, m]$ ,有如下结论

$$\sum_{j=1}^n x_i^j = 1. \quad (2)$$

式(1)中,所有行政区在相邻两年( $i$  年和  $i+1$  年)的犯罪率数据(向量  $x_i$  和向量  $x_{i+1}$ )之间的关系,被称为概率转移矩阵(transition probability matrix),记作  $A_i$ ,如下所示。

$$A_i = \begin{bmatrix} a_i^{11} & a_i^{12} & \dots & a_i^{1n} \\ a_i^{21} & a_i^{22} & \dots & a_i^{2n} \\ \dots & & & \dots \\ a_i^{n1} & a_i^{n2} & \dots & a_i^{nn} \end{bmatrix}, i = 1, 2, \dots, m-1.$$

同时,存在如下约束条件:

$$\sum_{j=1}^n a_i^{j,k} = 1. \quad (3)$$

两个区域  $S_s$  和  $S_t$  之间的距离定义如下.

**定义 1:** 区域  $S_s$  和区域  $S_t$  之间的距离

$$d^{s,t} = \begin{cases} 0 & s = t \\ 1 & S_s \text{ 和 } S_t \text{ 相邻} \\ L & S_s \text{ 和 } S_t \text{ 相距 } L - 1 \text{ 个距离单位以上.} \end{cases} \quad (4)$$

此外,  $dd_s = \sum_{t \neq s} d^{s,t}$ , 表示所有区域到区域  $S_s$

的距离加和.

需要注意的是, 此处的距离并不需要数学上严格满足欧几里得空间中的距离定义的 3 个条件. 因为这里的距离只是为了定量描述每个区域的转移犯罪强度, 只作为权重来调整本文所提出算法中的转移分布比率. 进一步地, 对于所有  $n$  个行政区域, 有如下距离矩阵  $\mathbf{D}_n$ .

$$\mathbf{D}_n = \begin{bmatrix} d^{1,1} & d^{1,2} & \cdots & d^{1,n} \\ d^{2,1} & d^{2,2} & \cdots & d^{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ d^{n,1} & d^{n,2} & \cdots & d^{n,n} \end{bmatrix}, \quad (5)$$

不难发现, 上述  $D_n$  是一个对称矩阵, 即  $d^{s,t} = d^{t,s}$ .

在本研究中, 所研究城市的各区域的面积和人口分别表示为

$$\mathbf{s} = [s^1 \ s^2 \ \cdots \ s^n] \mathbf{x}^T, \quad (6)$$

$$\mathbf{u}_i = [u_i^1 \ u_i^2 \ \cdots \ u_i^n] \mathbf{x}^T. \quad (7)$$

同时,

$$\mathbf{s}\mathbf{s}^k = \sum_{t \neq k} s^t, \mathbf{u}\mathbf{u}_i^k = \sum_{t \neq k} u_i^t \quad (8)$$

通过相关性分析可知, 人口因素与犯罪率呈正相关, 而面积和犯罪率呈负相关. 综上, 本文将距离因素(式(4))、面积因素(式(6))和人口因素(式(7))引入犯罪分布预测中, 具体体现在算法 1 中的第 12 行.

## 1.2 TWcS 算法

基于对社会环境因素的建模, 本文创新性地采用以距离因素、人口因素和面积因素为权重值的梯度下降方法对转移概率矩阵  $\mathbf{A}_i$  进行自学习(self-learning), 并提出 TWcS 算法(如算法 1 所示). 算法概述如下.

将式(1)中矩阵任意行的乘积表示为

$$h_i^j [\mathbf{x}_i^1 \ \mathbf{x}_i^2 \ \cdots \ \mathbf{x}_i^n] = \sum_{k=1}^n a_i^{j,k} \mathbf{x}_i^k, \quad (9)$$

相应的损失函数被定义为

$$J_i^j [a_i^{j,1} \ a_i^{j,2} \ \cdots \ a_i^{j,n}] = \frac{1}{2} \sum_{m=1}^M [h_i^j(x^{(m)}) - x_{i+1}^{j,(m)}]^2, \quad (10)$$

其中,  $M$  代表数据集规模. 因此, 总的损失函数可以定义为

$$J_i = \sum_{j=1}^n J_i^j. \quad (11)$$

进一步有

$$\frac{\partial J_i}{\partial a_i^{jk}} = \sum_{j=1}^n \sum_{m=1}^M [h_i^j(x^{(m)}) - x_{i+1}^{j,(m)}] x_i^k, \quad (12)$$

$$a_i^{jk'} = a_i^{jk} - \alpha \sum_{j=1}^n \sum_{m=1}^M [h_i^j(x^{(m)}) - x_{i+1}^{j,(m)}] x_i^k. \quad (13)$$

本文所采用的随机梯度下降(stochastic gradient descent, SGD)算法, 如下所述:

$$a_i^{jk'} = a_i^{jk} - \alpha \sum_{j=1}^n [h_i^j(x) - x_{i+1}^j] x_i^k, \quad (14)$$

对任意的  $j$  和  $k$ , 存在约束条件如下:

$$\sum_{j=1}^n a_i^{jk} = 1. \quad (15)$$

在本文的实验中, 设置式(14)中参数  $\alpha$  的值随着迭代次数的增加而减小. 综上, 本文所提出的 TWcS 算法流程如下述算法 1 所示.

### 算法 1. TWcS 算法

输入: 向量对  $(\mathbf{x}_i, \mathbf{x}_{i+1}), i=1, 2, \dots, m-1$ .

输出:  $\mathbf{A}_i, i=1, 2, \dots, m-1$ .

过程:

1: 为  $\mathbf{A}_i$  随机设置初值, 并且令  $\alpha=1$ .

2:  $\mathbf{x}_{i+1}^{\sim} = \mathbf{A}_i \mathbf{x}_i$

3:  $\Delta_{i+1} = [\Delta_{i+1}^1 \ \Delta_{i+1}^2 \ \cdots \ \Delta_{i+1}^n]^T = \mathbf{x}_{i+1}^{\sim} - \mathbf{x}_{i+1} =$

$$[\mathbf{x}_{i+1}^{1,\sim} \ \mathbf{x}_{i+1}^{2,\sim} \ \cdots \ \mathbf{x}_{i+1}^{n,\sim}]^T - [\mathbf{x}_{i+1}^1 \ \mathbf{x}_{i+1}^2 \ \cdots \ \mathbf{x}_{i+1}^n]^T$$

4:  $dd_s = \sum_{t \neq s} d^{s,t}$

5: repeat

6:     for  $i=1, 2, \dots, m-1$  do

7:         for  $k=1, 2, \dots, n$  do

8:             for  $j=1, 2, \dots, n$  do

9:                 if  $j=k$  then

$$10: \quad a_i^{j,k} = a_i^{j,k} - \alpha \sum_{j=1}^n \Delta_{i+1}^j x_i^k$$

11:                 else

$$12: \quad a_i^{j,k} = a_i^{j,k} + \frac{\alpha}{3} \left( \frac{d^{j,k}}{dd_k} + \frac{ss_k - s^j}{nss_k} + \frac{u_i^j}{uu_i^k} \right) \sum_{j=1}^n \Delta_{i+1}^j x_i^k$$

13:                 end if

```

14:            $\alpha = 1/(1+1/\alpha)$ 
15:       end for
16:   end for
17: end for
18: until 直到满足停止条件
19: repeat
20:   for  $i=1,2\cdots m-1$  do
21:     for  $k=1,2\cdots n$  do
22:       for  $j=1,2\cdots n$  do
23:          $a_i^{j,k} = 0.1a_i^{j,k} + 0.3a_{i+1}^{j,k} + 0.6a_{i+2}^{j,k}$ 
24:       end for
25:     end for
26:   end for
27: until 直到满足停止条件

```

在TWcS算法中,首先设置矩阵初值,然后按照式(1)进行迭代。式(1)可改写为如下形式:

$$\begin{bmatrix} a_i^{11} & a_i^{12} & \cdots & a_i^{1n} \\ a_i^{21} & a_i^{22} & \cdots & a_i^{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_i^{n1} & a_i^{n2} & \cdots & a_i^{nn} \end{bmatrix} \begin{bmatrix} x_i^1 \\ x_i^2 \\ \cdots \\ x_i^n \end{bmatrix} = \begin{bmatrix} x_{i+1}^{1\sim} \\ x_{i+1}^{2\sim} \\ \cdots \\ x_{i+1}^{n\sim} \end{bmatrix}. \quad (16)$$

计算下述公式:

$$\Delta_{i+1}^j = x_{i+1}^{j\sim} - x_{i+1}^j, j = 1, 2, \dots, n. \quad (17)$$

当  $\Delta_{i+1}^j > 0$  时,表明预测值大于实际值。因此,减小与  $x_{i+1}^{j\sim}$  相对应的矩阵第  $j$  行各个元素  $a_i^{jk}$  的值。这是因为

$$x_{i+1}^{j\sim} = \sum_{k=1}^n a_i^{jk} x_i^k. \quad (18)$$

当  $\Delta_{i+1}^j < 0$  时,表明预测值小于实际值,做类似反方向调整。根据式(17)(18),可以得到

$$x_{i+1}^j = \sum_{k=1}^n a_i^{jk} x_i^k - \Delta_{i+1}^j. \quad (19)$$

把差值  $\Delta_{i+1}^j$  按分布比率分到转移概率矩阵  $A_i$  的第  $j$  行各元素  $a_i^{jk}$  上(算法 1 第 10 行),过程描述如下:

$$\begin{aligned} x_{i+1}^j &= \sum_{k=1}^n a_i^{jk} x_i^k - \Delta_{i+1}^j = \\ &\sum_{k=1}^n (a_i^{jk} - \Delta_{i+1}^j) x_i^k. \end{aligned} \quad (19)$$

调整  $a_i^{jk}$  ( $j \neq k$ ) 来保证式(3)的成立,算法 1 的第 12 行有相应描述,这也是算法 1 的核心。直到满足迭代停止条件,算法停止迭代。

正如上述讨论,  $\Delta_i$  是第  $i$  年预测值和实际值之间的差值向量。向量的元素可能有正值有负值,所

以本文按照如下方式定义准确度(accuracy):

$$\theta_i = \frac{1}{n} \sum_{j=1}^n |\Delta_i^j|. \quad (20)$$

基于警务工作实际应用的需要,本研究把上述准确度  $\theta_i$  作为阈值来控制迭代的停止时间。

### 1.3 评价标准

本文选择平均绝对误差(mean absolute error, MAE)来对预测结果进行性能分析。平均绝对误差是最常用的一种评价预测结果的方法,被广泛用作衡量预测任务准确率的评估标准,其定义如下:

$$\sigma_i = \frac{1}{n} \sum_{j=1}^n |(x_i^{j\sim} - x_i^j)|. \quad (21)$$

式中: $\sigma_i$  代表第  $i$  年的平均绝对误差; $x_i^{j\sim}$  代表第  $i$  年的第  $j$  个分量; $n$  代表向量维数,即行政区域数。

## 2 实验及结果分析

本研究共计设置 5 组实验,从不同角度探索犯罪分布预测算法 TWcS 的性能,并对实验结果进行分析,包括:①初始值选择对实验结果的影响;②系数设置对实验结果的影响;③TWcS 算法与基线算法(TPML-WMA 算法、线性回归、自回归等算法)的性能比较;④同样引入距离因素的情况下,算法 TWcS 与基线算法(Lasso 回归、贝叶斯、决策树等算法)的性能比较;⑤评价标准 MAE 下各算法的性能比较。

### 2.1 数据集

本研究所使用的历史犯罪数据为 2006—2016 年中国某一线城市 18 个行政区域的盗窃案例数据。该数据集包括盗窃案件发生时间和发生地点等信息。本文所提出 TWcS 算法用于处理前 10 年(2006—2015 年)的数据,然后预测最后一年(2016 年)的盗窃案例分布情况。最后,使用评价标准 MAE 来比较预测值和实际值,进而评估算法的有效性。

本研究所使用的辅助数据集来自中华人民共和国国家统计局门户网站,其中包括 2010—2016 年上述 18 个行政区域的 5 个因素,分别为户籍人口、常住人口、GDP、收入、财务支出。以某一特定行政区域为例,表 1 展示出了上述 5 个因素的指标值(受限于篇幅,表 1 仅以其中一个行政区为例,展示了 2010—2016 年的数据)。这些因素的相关数据将会用于训练对比算法(例如 Lasso 回归、贝叶斯和决策树等算法)。

**表 1 某一区域的 5 个因素的取值示例(2010~2016 年)**  
**Tab. 1 An example of the value of 5 factors in a certain area (2010~2016)**

年份	犯罪率	户籍登记人口/万	常住人口/万	GDP/(10 亿人民币)	财政收入/(10 亿人民币)	财政支出/(10 亿人民币)
2010	0.243	171.1	280.2	130.17	8.530	6.044
2011	0.187	174.5	291.8	155.91	10.659	8.384
2012	0.213	178.4	303.0	190.49	14.155	13.063
2013	0.181	181.8	327.1	214.40	16.832	15.078
2014	0.193	185.3	336.4	238.04	19.066	16.866
2015	0.167	188.6	354.5	280.42	23.426	22.803
2016	0.162	193.2	365.8	327.22	31.683	43.833

## 2.2 初始值选择对实验结果的影响

本实验中,距离矩阵  $D_n$  元素的初始值都被设置为固定值  $1/18$ ,其他因素保持不变;此外,迭代次数分别设为 10、50 和 100。表 2 显示,当迭代仅 10 次时,预测值就达到了  $10^{19}$  数量级。这表明,固定初始值会导致实验结果非常不理想,并且随迭代次数的增加而大量积累。

**表 2 固定初始值情况下,分别迭代 10/50/100 次后的结果对比**  
**Tab. 2 The results of 10/50/100 iteration are compared with fixed initial values**

行政区域	实际值	10 次迭代/	50 次迭代/	100 次迭代/
		$10^{19}$	$10^{123}$	$10^{252}$
1	0.012	-4.141	-2.10	-8.93
2	0.016	-4.141	-2.10	-8.93
3	0.006	-4.141	-2.10	-8.93
4	0.017	-4.141	-2.10	-8.93
5	0.192	-4.141	-2.10	-8.93
6	0.170	-4.141	-2.10	-8.93
7	0.020	-4.141	-2.10	-8.93
8	0.156	-4.141	-2.10	-8.93
9	0.012	-4.141	-2.10	-8.93
10	0.043	-4.141	-2.10	-8.93
11	0.052	-4.141	-2.10	-8.93
12	0.058	-4.141	-2.10	-8.93
13	0.090	-4.141	-2.10	-8.93
14	0.084	-4.141	-2.10	-8.93
15	0.016	-4.141	-2.10	-8.93
16	0.015	-4.141	-2.10	-8.93
17	0.018	70.400	35.70	152.00
18	0.014	-4.141	-2.10	-8.93

## 2.3 权重系数设置对实验结果的影响

对于算法 TWcS(算法 1)中的第 23 行中的权

重系数,本文选取了三组权重系数  $\{1/3\}$ 、 $\{0.1\}$ 、 $\{0.3\}$ 、 $\{0.6\}$ 、 $\{0.6\}$ 、 $\{0.3\}$ 、 $\{0.1\}$ ,然后比较它们对实验结果的影响。由表 3 可知(\* 代表负值),当平均地设定权重系数(即权重系数组合  $\{1/3, 1/3, 1/3\}$ )时,实验结果出现了一个负值项(行政区域 10),此外存在 6 个预测结果比其他两种系数设置更好。如果将系数设置为  $\{0.6, 0.3, 0.1\}$ ,则出现两个负项(行政区 10 和 15),而且仅有 4 项预测结果优于其他两个系数设置方式。当将系数设置为  $\{0.1, 0.3, 0.6\}$  时,有 7 个预测项比其他两种设置方法更好,并且没有负值项。因此,本文所提出的的 TWcS 算法使用的参数设置为  $\{0.1, 0.3, 0.6\}$ 。

**表 3 不同权重系数设置对实验结果的影响**

**Tab. 3 The influence of different weight coefficient settings on the experimental results**

行政区域	实际值	权重系数		
		$\{1/3, 1/3, 1/3\}$	$\{0.1, 0.3, 0.6\}$	$\{0.6, 0.3, 0.1\}$
1	0.012	0.088	0.089	0.101
2	0.016	0.057	0.059	0.036
3	0.006	0.095	0.072	0.084
4	0.017	0.064	0.089	0.056
5	0.192	0.141	0.130	0.149
6	0.170	0.182	0.130	0.241
7	0.020	0.077	0.047	0.064
8	0.156	0.076	0.090	0.057
9	0.012	0.049	0.035	0.040
10	0.043	-0.003*	0.011	-0.011*
11	0.052	0.020	0.021	0.022
12	0.058	0.027	0.013	0.027
13	0.090	0.015	0.044	0.001
14	0.084	0.021	0.031	0.021
15	0.016	0.019	0.020	-0.008*
16	0.015	0.035	0.050	0.078
17	0.018	0.016	0.045	0.024
18	0.014	0.014	0.022	0.019

## 2.4 TWcS 算法与 TPML-WMA/线性回归/自回归算法的对比

本实验将本文所提出的 TWcS 算法与犯罪分布预测领域常用基线算法进行性能对比,包括目前最优基线算法 TPML-WMA(记为 TPML-WMA)、线性回归(记为 LR)算法和自回归(记为 AR)算法。所有的对比算法均使用相同的历史数据来预测和分析犯罪率。相对于其他算法,算法 TWcS 的特殊之处在于,它将区域性的社会和地理因素(人口因素、面积因素以及距离因素)引入到历史数据中来解决单纯使用历史数据来预测犯罪率所面临的问题。

图1展示了算法 TWcS 与最优基线算法 TPML-WMA 的预测结果曲线。算法 TPML-WMA 算法在 1,3,8 以及 10~17 区的预测值均不理想,而算法 TWcS 则显着提高了这些区域的预测性能(即预测值与实际值更加拟合)。此外,算法 TWcS 成功地避免了算法 TPML-WMA 在 6,7,18 区中引起的过拟合。由此可以得出如下结论:犯罪的发生不仅与两个区域的距离存在关系,而且与人口和面积等因素存在关系。换言之,把人口、经济等诸多社会环境因素结合起来协同考虑,是研究犯罪率预测的一个很好的途径,也为从犯罪分布反向研究地区经济和社会发展状况提供了一种可行性。

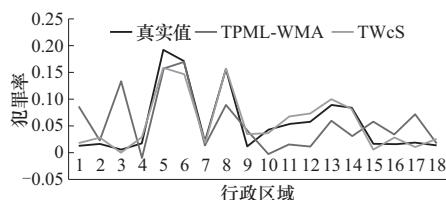


图1 TWcS 算法和最优基线算法 TPML-WMA 的对比  
Fig. 1 Comparison between algorithm TWcS and optimal baseline algorithm TPML-WMA

将算法 TWcS 与算法 LR 进行对比,结果如图 2 所示。算法 LR 的预测值与实际值相差甚远。例如,在 5,10,11,13 和 14 区,算法 LR 的预测效果都非常不好(即预测值与实际值的拟合程度欠佳)。此外,通过分析图像不难发现,算法 LR 的曲线波动非常大,例如在 5 区达到非常高的点,在 10 区又达到非常低的点;而相对地,本文所提出的算法 TWcS 性能明显优于算法 LR,其产生的预测值与实际值拟合较好。

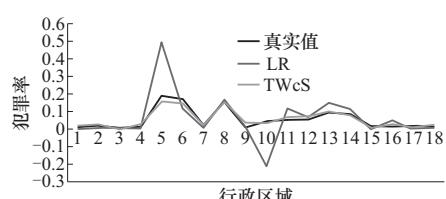


图2 TWcS 算法和 LR 的对比  
Fig. 2 Comparison between algorithm TWcS and LR

算法 TWcS 与算法 AR 的对比结果如图 3 所示。与图 2 中的算法 LR 相比,算法 AR 的性能更好,所产生的预测值能够很好地拟合实际值。例如,算法 AR 在 5,10,11,13 和 14 区的表现均优于算法 LR,而在 1 区、9 区和 12 区的表现不如算法 LR。而与算法 AR 相比,本文所提出的算法 TWcS 在 1,5,6,8,13 和 16 区表现更好。相比较于算法 TWcS,算法 AR 的劣势在于其无法适用于所有的数据集:因

为算法 AR 只在时间序列匹配某种特定模式的情况下才能展现出良好的性能,否则,如果数据不存在某些特定时间变化规律时,算法 AR 的准确性将大大降低,限制了算法 AR 的普适性。

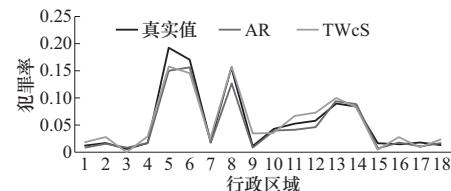


图3 TWcS 算法和 AR 的对比  
Fig. 3 Comparison between algorithm TWcS and AR

## 2.5 TWcS 算法与 Lasso 回归/贝叶斯/决策树算法的对比

本文所提出的算法 TWcS 将距离因素引入到犯罪预测研究中。为了更好地验证本文算法的性能,本实验将距离因素同样融入到 Lasso 回归算法、贝叶斯算法和决策树算法中来预测犯罪分布,然后与本文算法 TWcS 进行对比。图 4 展示了实际值与 Lasso 回归算法、贝叶斯算法、决策树算法和本文所提出的算法 TWcS 的预测值之间的差异。从图中可以看出:贝叶斯算法在所有对比算法中表现最差,特别是在 1 区、2 区、8 区和 10 区;从 6 区到 10 区,Lasso 回归算法、贝叶斯算法和决策树算法出现大幅波动,而算法 TWcS 在这些地区走势相对温和,拟合实际值的效果也优于其他算法,体现除了比其他算法更优秀的性能。

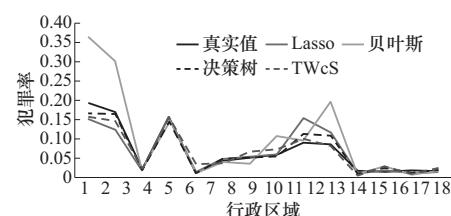


图4 TWcS 算法和 Lasso 回归算法、贝叶斯算法以及决策树算法的对比  
Fig. 4 Comparison between algorithm TWcS and Lasso regression algorithm, bias algorithm and decision tree algorithm

## 2.6 在评价标准 MAE 下对比各算法的性能

表 4 展示了本文前述的各算法在评价标准 MAE 下的实验结果。显而易见,算法 TPML-WMA 和算法 TWcS 预测的犯罪率总和均接近 1,而算法 LR 和算法 AR 的犯罪率总和较偏离 1.000(分别为 1.109 和 0.885)。这表明,算法 LR 和算法 AR 所产生的预测结果是不合理的。由于 Lasso 回归算法、贝

叶斯算法和决策树算法只预测了14个区域的犯罪率,所以无法计算出上述犯罪率的加和。在评估标准MAE下,本文所提出的算法TWcS、算法AR、Lasso回归算法和决策树算法优于其他算法。其中,算法TWcS和算法AR的表现是最好的,算法AR甚至优于算法TWcS。然而因为算法AR预测的犯罪率之和偏离了约12%,所以预测结果仍然不可信。综合SUM和MAE指标,算法TWcS较其他算法性能表现最好,MAE值是其他算法MAE平均值的33%。

表4 各算法的性能对比

Tab. 4 Performance comparison between algorithm TWcS and other algorithms

指标	TWcS (本文)	TPML-WMA	LR	AR	Lasso 回归	贝叶斯	决策树
SUM	1.001	0.999	1.109	0.885			
MAE	0.010	0.038	0.050	0.009	0.015	0.059	0.011

### 3 结 论

本文通过将距离信息、人口信息和面积信息引入转移概率矩阵自学习,提出了一种全新的犯罪分布预测算法(TWcS)。多组对比实验的结果证明,通过融合距离、面积、人口等社会环境因素的TWcS算法,能够合理建模犯罪分布并达到更高的预测准确率,能够有效描述封闭系统内所有行政区域的犯罪率分布和转移关系。

#### 参考文献:

- [1] Helbich M, Arsanjani J J. Spatial eigenvector filtering for spatio temporal crime mapping and spatial crime analysis [J]. *Cartography & Geographic Information Science*, 2015, 42(2): 134–148.
- [2] Leong K, Sung A. A review of spatio-temporal pattern analysis approaches on crime analysis [J]. *International e-Journal of Criminal Sciences*, 2015, 9: 1–33.
- [3] Almanie T, Mirza R, Lor E. Crime prediction based on crime types and using spatial and temporal criminal hotspots [J]. *Computer Science*, 2015, 5: 1–19.
- [4] Vlek C, Prakken H, Renooij S, et al. Modeling crime scenarios in a Bayesian network [C] // Fourteenth International Conference on Artificial Intelligence & Law. New York: ACM Press, 2013: 150–159.
- [5] Liao R, Wang X, Li L, et al. A novel serial crime prediction model based on Bayesian learning theory [C] // International Conference on Machine Learning & Cybernetics (ICMLC). Piscataway, NJ: IEEE, 2010: 1757–1762.
- [6] Wei X, Yan J, Chen Z, et al. Analysis of crime rate distribution based on TPML-WMA [C] // International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. [S. l.]: IEEE, 2017: 157–160.
- [7] Nath S. Crime pattern detection using data mining [C] // IEEE/WIC/ACM International Conference on Web Intelligence & Intelligent Agent Technology Workshops. Piscataway, NJ: IEEE, 2007: 41–44.
- [8] Malleson N, Birkin M. Analysis of crime patterns through the integration of an agent-based model and a population microsimulation [J]. *Computers Environment & Urban Systems*, 2012, 36(6): 551–561.
- [9] Yu C H, Ward M W, Morabito M, et al. Crime forecasting using data mining techniques [C] // IEEE International Conference on Data Mining Workshops. Piscataway, NJ: IEEE, 2012: 779–786.
- [10] Chen P, Yuan H, Shu X. Forecasting crime using the ARIMA model [C] // Fifth International Conference on Fuzzy Systems and Knowledge Discovery. FSKD Piscataway, NJ: IEEE, 2008: 627–630.
- [11] Noor N M M, Retnowardhani A, Abd M L, et al. Crime forecasting using ARIMA model and fuzzy alpha-cut [J]. *Journal of Applied Sciences*, 2013, 13(1): 167–172.
- [12] Zhang J L, Wang X Y, Ma L Y, et al. Traffic flow state identification method based on dynamic Bayesian network [J]. *Transactions of Beijing Institute of Technology*, 2014, 1: 45–49.
- [13] Ratcliffe J. The hotspot matrix: a framework for the spatio-temporal targeting of crime reduction [J]. *Police Practice & Research*, 2004, 5(1): 5–23.
- [14] Chainey S, Ratcliffe J. GIS and crime mapping [M]. London: Wiley, 2013.
- [15] Brantingham P J, Brantingham P L. Crime pattern theory [M]. Cullompton Brantingham: Willan Publishing, 2008: 78–93.
- [16] Brunsdon C, Fotheringham A S, Charlton M E. Geographically weighted regression: a method for exploring spatial nonstationarity [J]. *Geographical Analysis*, 1996, 28(4): 281–298.
- [17] Bogomolov A, Lepri B, Staiano J, et al. Once upon a crime: towards crime prediction from demographics and mobile data [C] // Proceedings of the 16th International Conference on Multimodal Interaction. Now York: ACM Press, 2014: 427–434.

(责任编辑:刘芳)