网络出版时间：2014-06-09 10:06
网络出版地址：http://www.cnki.net/kcms/doi/10.13203/j.whugis20140060.html

DOI：10.13203/j.whugis20140060　　文章编号：1671-8860(2014)06-0695-06

利用多源领域知识迁移 CA 的城市建设用地模拟

刘轶伦[1,2]　黎　夏[1,2]

1　中山大学地理科学与规划学院，广东 广州，510275
2　中山大学广东省城市化与地理环境空间模拟重点实验室，广东 广州，510275

摘　要：建立传统元胞自动机(CA)模型时，如果样本数量不足，模拟效果往往非常不理想。提出了多源领域知识迁移 CA 模型，利用多个已有的旧样本数据集来帮助建立新的 CA 模型，并选取广东省深圳市作为试验区验证了其有效性。试验结果表明，该模型在新样本数量不足的情况下能够明显改善模拟效果，并且有效减小产生负迁移现象的风险。

关键词：知识迁移；元胞自动机模型(CA)；城市土地利用模拟；多源领域 TrAdaBoost 算法

中图法分类号：P208　　**文献标志码**：A

传统的用于城市用地增长模拟的元胞自动机模型(cellular automata，CA)在建立新模型时需要采集大量的样本数据来挖掘转换规则[1-4]。建立起来的转换规则只适用于特定区域或特定发展时期的城市用地模拟[5]。当想把已经建立起来的 CA 模型应用于不同的研究区域或者不同时期，往往得不到理想的模拟结果，不得不重新采集对应研究区域或时间区间的样本数据[4-5]。如果能够充分利用已有的旧样本数据来帮助建立新的 CA 模型，将会节省大量的采样成本，并且能够有效提高样本缺乏情况下的模型精度[6]。如何在相类似的研究问题之间共享知识和数据，迁移学习理论能够解决这个问题。

迁移学习是机器学习和数据挖掘领域近几年新兴的研究热点[7-9]。迁移学习是一系列方法的统称，目的是从已有相类似的模型或样本集中提取有价值的信息(或知识)，用于提高解决新问题的机器学习效率[7]。传统的数理统计或数据挖掘算法都基于一个假设，用于训练的标签数据和测试数据必须符合同一个数学分布[10]。当这个数学分布产生变化，大部分方法或模型就需要重新收集标签数据进行训练。迁移学习尝试从一个或多个相关的学习任务中迁移知识来突破这个同分布假设。

基于上述传统统计方法的缺点，Dai 等在集成学习理论的基础上提出了一种实例迁移学习算法(transfer AdaBoost，TrAdaBoost)较好地解决了这类问题[10]。TrAdaBoost 使用源领域(source domain)中的样本数据以及少量目标领域的样本数据，通过调整样本权重的策略，找出适合目标领域的样本数据建立更准确的预测模型。Li 等基于 TrAdaBoost 方法建立了基于单源领域知识迁移的 CA 土地利用模型，并且应用于珠江三角洲几个主要城市，证明了在模型训练样本较少的情况下，知识迁移 CA 能够明显改善模拟精度[6]。

基于单源领域的知识迁移对于提供辅助数据的源领域需要仔细甄别，要求辅助源领域和目标领域的数据分布尽量相似。但是在目标领域样本数据缺乏的情况下，衡量它的数据分布是很难做到的，很难找到与目标领域数据分布最相似的源领域。针对这些问题，Yi 等对 TrAdaBoost 进行了改进，通过引入多个源领域来提高迁移学习的效率，并且降低负迁移(加入源领域的知识后，训练出来的模型比不加源领域知识的模型精度更低)的风险[11]。本文在 TrAdaBoost 逻辑回归 CA 模型的基础上，根据多源领域知识迁移方法，提出了基于多源领域知识迁移的 CA 模型(multi-source TrAdaBoost CA，MSTra CA)。

1　基于多源领域知识迁移的 CA 模型

1.1　模型概述

在迁移学习理论中，待解决的新问题通常被

收稿日期：2014-01-17
项目来源：国家自然科学基金资助项目(41371376)。
第一作者：刘轶伦，博士生。主要研究方向包括空间数据挖掘、城市扩张模拟和空间智能体建模。E-mail：ealenliu@gmail.com

定义为属于目标领域，而帮助解决目标领域问题的知识或样本则来源于源领域[7-9]。假设需要模拟某个时间区间某个区域的城市用地变化，可将这个研究问题视为目标领域，目标领域的样本集称为 D_T。其他研究区域或同一区域不同时间区间的城市用地变化模拟则称为源领域，对应的源领域样本集为 D_S。由于 D_T 的样本数量太少，不足以建立一个准确的模拟模型，可利用迁移学习算法从多个源领域的样本集中迁移有效的样本到目标领域，帮助建立一个准确的 CA 模型。

应用多源领域 TrAdaBoost 算法的一个直观示例如图 1 所示。当只有少量目标领域样本数据时(图 1(a))，要得到一个准确的模型是很困难的，可以尝试使用已有的大量源领域样本帮助建立模型，但是通常源领域的数据分布与目标领域的数据分布不完全一样(图 1(b))，直接使用源领域样本建立的模型用于目标领域往往效果欠佳，可以利用多源领域 TrAdaBoost 算法从大量的源领域样本中找出符合目标领域样本分布的数据(图 1(c))，建立一个更加准确的模型。

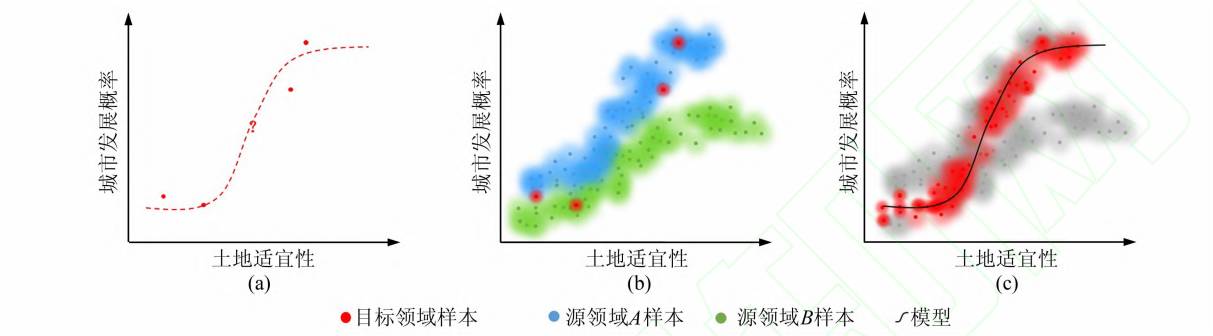

图 1 关于多源领域 TrAdaBoost 算法的一个直观示例

Fig. 1 An Example of Multi-source TrAdaBoost Algorithm for Dealing with Training Data of Different Distributions

1.2 MSTra CA 基础模型：逻辑回归 CA

MSTra 本质上是一个集成学习框架，需要有一个基础模型作为集成的基础。本研究选择逻辑回归 CA[2]作为 MSTra 的基础模型。

逻辑回归 CA 通过逻辑回归模型对样本数据进行拟合(或训练)，得到城市用地的发展适宜性 $1/(1+\exp(-z_{ij}^t))$。再结合邻域发展概率、规划因子等其他影响城市发展的因素，得到综合城市发展概率 p_{ij}^h[2, 12]：

$$p_{ij}^h = (1+(-\ln\gamma)^\alpha)\frac{1}{1+\exp(-z_{ij}^h)}\times f(\Omega_{ij}^h)\times \mathrm{con}_{ij} \quad (1)$$

其中，p_{ij}^h 是 h 时刻元胞 ij 的发展概率；z_{ij}^h 是发展适宜性的综合得分，$z_{ij}^h = a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m + \cdots + a_Mx_M$，$a$ 是回归系数；x 是空间影响因子；γ 是随机影响因子，取值 0 到 1；α 是随机因子的控制系数；$f(\Omega_{ij}^h)$ 是邻域影响因子；con_{ij} 是土地利用规划因子，取值 0 或 1。

城市发展概率大于转换阈值的元胞将会由非城市用地转变为城市用地：

$$S_{ij}^{h+1} = \begin{cases} \text{Converted}, & p_{ij}^h \geqslant Q_{\text{land}} \\ \text{NonConverted}, & p_{ij}^h < Q_{\text{land}} \end{cases} \quad (2)$$

其中，S_{ij}^{h+1} 是 ij 元胞在 $h+1$ 时刻的状态；Q_{land} 是元胞状态的转换阈值，阈值是通过预测城市建设用地需求量得到的。

1.3 基于多源领域的知识迁移

MSTra CA 建模首先需要输入数据和参数以及初始化样本权重。在 $t=1,\cdots,L$ 次迭代中，D_T 分别和 k 个不同的 D_S 建立逻辑回归 CA 模型，并且分别评价这 k 个模型的误差，选择误差最小的模型作为第 t 次迭代得到的有效基础模型[11]。随后根据模型误差调整样本权重，接着进行新一轮的迭代。迭代完成后，根据每个模型的误差输出最终的模拟结果。

1) 数据和参数输入

模型需要的输入包括 k 个源领域的标签数据集 $D_{S1},\cdots,D_{Sk}$；一个目标领域的标签数据集 D_T；基础模型逻辑回归 CA，逻辑回归 CA 模型需要用到的研究区模拟起始年份的城市用地数据、空间变量数据以及需要集成的基础模型数量 L。

2) 样本权重初始化

$D_{S1},\cdots,D_{Sk}$ 和 D_T 分别来源于不同的研究区，因而这些样本集有不完全相同的数据分布特征。用样本权重(w)来表示不同样本集中每个样本对于建立目标领域模型的重要性。

在初始状态，分别对于各个样本集中的样本赋予相同的权重。用 $w^{Sk} = (w_1^{Sk},\cdots,w_{n_{Sk}}^{Sk})$ 表示源领域样本权重，用 $w^T = (w_1^T,\cdots,w_{n_T}^T)$ 表示目标领域的样本权重。

3) 在 $t=1,\cdots,L$ 次迭代中，训练基础模型

首先标准化样本权重,让所有样本权重(w^{S1},…,w^{Sk},w^T)的总和为 1。为了提高逻辑回归 CA 模型的训练效率和减小误差,设置阈值 β 剔除权重较小的源领域样本[6]。公式如下:

$$\beta_t^k = \mathrm{mean}(w_t^{Sk}, w_t^T) \tag{3}$$

根据 β 筛选出一个新的训练样本集,输入逻辑回归 CA 得到一个模拟模型 f_t。计算 f_t 在 D_T 上的误差 ε:

$$\varepsilon_t = \sum_{i=1}^{n} w_t^T(i) I(f_t(x_i^T) \neq y_i^T) / \sum_{i=1}^{n} w_t^T(i) \tag{4}$$

其中,n 是 D_T 的样本数量;x 为 D_T 第 i 个样本的因变量;y 为 D_T 第 i 个样本的样本标签;$I(f_t(x_i) \neq y_i)$ 是一个判别函数,当 $f_t(x_i) \neq y_i$ 时,函数取值为 1,否则为 0。

在 MSTra CA 模型中,当一个目标领域训练样本被预测错误后,它认为这是一个比较难预测的训练样本。于是增加这个训练样本的权重,用来强调这个样本。下一次迭代,这个样本被误判的概率就会减少。对于源领域的训练样本,则采取另一种权重变化策略,当它们被错误预测后,可认为这些数据与目标领域数据相差较大,因而减少这些数据的权重,以降低它们在下一次分类训练中的影响。根据式(4)对数据集 D_S 和 D_T 进行权重调整:

$$\begin{cases} w_i^{Sk} \leftarrow w_i^{Sk} \mathrm{e}^{-\alpha_S I(f_t(x_i^{Sk}) \neq y_i^{Sk})} \\ w_i^T \leftarrow w_i^T \mathrm{e}^{\alpha_t I(f_t(x_i^T) \neq y_i^T)} \end{cases} \tag{5}$$

其中,D_S 和 D_T 的权重调整因子分别为[11]:

$$\alpha_S = \frac{1}{2}\ln\left(1 + \sqrt{2\ln\frac{n_S}{L}}\right),\ \alpha_t = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}$$

4) 输出结果

经过 L 次迭代后,训练得到 L 的 CA 模型。由于样本在开始时都被赋予了相同的权重,有效的和无效的源领域样本都一起被选入训练集,因而在最初几次迭代训练得到的 CA 模型误差较大,随着迭代次数的增加,CA 模型的误差会慢慢缩小,最终输出的结果由后 $L/2$ 次训练得到的 CA 模型加权得到,如式(6)[11]所示:

$$\mathrm{argmax}\left(\sum_{t=L/2}^{L} \alpha_t I(f_t(x_i) = y_i)\right) \tag{6}$$

2　试验与分析

2.1　试验区与数据

选择深圳市作为试验区验证模型的有效性。目标领域为 2000～2008 年深圳市城市用地变化。模型要求源领域与目标领域具有相类似的城市增长特征(即 CA 模拟考虑相同的城市发展影响因子),并且源领域的样本分布应与目标领域的样本分布尽量接近,若样本分布差异较大,则容易产生负迁移(即迁移模拟结果比未迁移结果较差)。因而选择与实验区处于同一个城市群的广州市区 2000～2008 年城市用地变化样本(D_{S1})以及东莞市 2000～2008 年城市用地变化样本(D_{S2})分别作为两个源领域。三个城市在研究时间段内均经历了非常快速的城市增长,并且城市用地扩张具有类似的特征[13]。

CA 模型主要考虑空间区位因子,需要输入的空间因子包括距市中心距离、距区镇中心距离、距一般道路距离、距高速路距离以及距铁路距离 5 个因子,如图 2 所示,数据分辨率为 30 m。

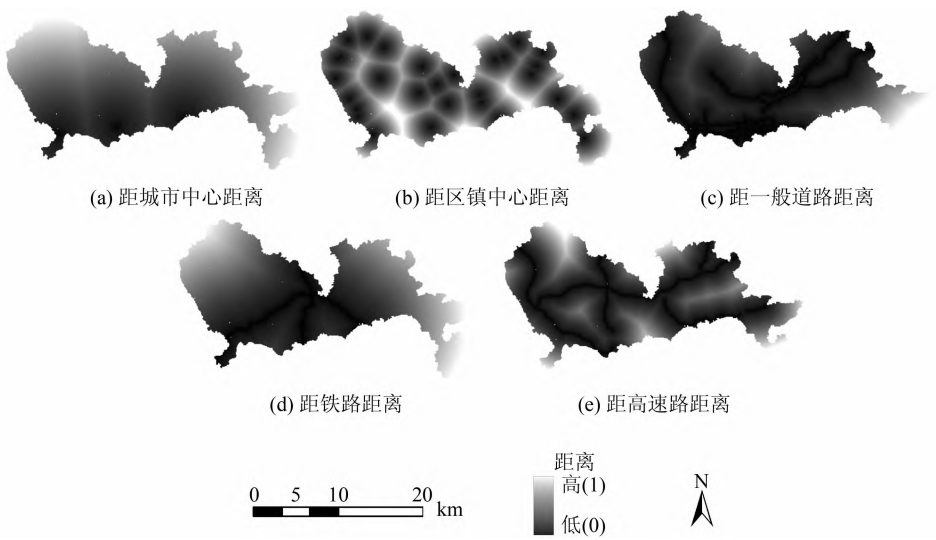

图 2　模型的空间因子数据

Fig. 2　Various Proximity Variables Related to Shenzhen Urban Dynamics

目标领域数据与源领域数据按不同比例分为10组(如表1所示)。为了减少采样偏差的影响,每个组合将会从样本库中随机抽取数据10次,并分别进行模拟,模拟精度取平均值。

表1 10组不同比例的训练数据集

Tab.1 Different Combinations (Ratios) of D_T/D_S for Urban Simulation

组合	D_T数量	D_S数量	比例/%	组合	D_T数量	D_S数量	比例/%
1	5	500	1	6	30	500	6
2	10	500	2	7	35	500	7
3	15	500	3	8	40	500	8
4	20	500	4	9	45	500	9
5	25	500	5	10	50	500	10

由于MSTra CA模型是一个集成学习模型,需要在开始运算前输入需要集成的模型数量L。图3给出了模型在第10组试验数据集上回归误差随着L增大的变化曲线。当集成模型数较少($L<40$)时,回归误差较大,而且结果较不稳定,随着L的增大,结果逐渐趋于稳定。而且MSTra CA模型的时间复杂度为$O(\sqrt{\ln(n/N)})$,效率随着集成模型数量的增加而降低,因而为了让模型获得一个稳定的结果,并考虑到计算效率问题,本文L值取50。

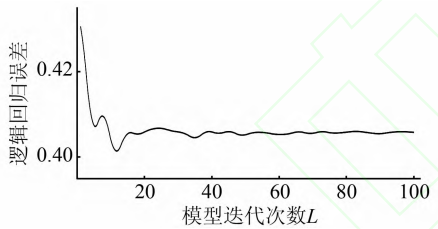

图3 D_T数量为50时,回归误差随着迭代次数L的变化曲线

Fig.3 Variation of Regression Error with the Increase of the Iteration (L) Using 50 D_T Samples

2.2 模拟结果与分析

采用Figure of Merit(FoM)[14]作为模拟结果的评价指标。由于不同研究区间新增城市的数量不一致,FoM可以排除模型模拟的起始年份已有的城市用地,比较直观地评价模拟结果。通过模拟结果与验证数据进行叠置分析后,FoM可由式(7)计算得到。当FoM=0%时,表明模拟结果与验证数据没有重叠部分;当FoM=100%时,表明模拟结果与验证数据完全重叠。

$$\mathrm{FoM}=\mathrm{CC}/(\mathrm{PC}+\mathrm{CC}+\mathrm{CP}) \tag{7}$$

其中,PC是模拟结果为城市用地而验证数据为非城市用地;CC是模拟结果和验证数据均为城市用地;CP是模拟结果为非城市用地而验证数据为城市用地。

除了本研究提出的MSTra CA模型,还使用其他三个模型结果作为对比,分别是基于D_T训练得到的传统逻辑回归CA;基于D_T和D_{S1}训练得到的单源领域知识迁移CA(TraCA(D_T+D_{S1}))以及基于D_T和D_{S2}训练得到的TraCA(D_T+D_{S2})。

图4为四个模型在各组试验数据模拟结果的FoM值。当D_T数量较少(D_T数量小于30)时,基于知识迁移算法的CA都能得到优于传统逻辑回归CA的模拟结果。在D_T数量极少(D_T数量等于5)时,传统逻辑回归CA甚至无法训练模型,而知识迁移CA仍表现良好。随着D_T数量逐渐增多,逻辑回归CA的模拟精度迅速上升,当D_T数量大于30时,由于D_{S1}与D_T数据分布相差较大,TraCA(D_T+D_{S1})出现了负迁移现象,模拟结果反而比传统CA要差。但是同样使用了D_{S1}的MSTra CA则避免了负迁移现象,模拟结果优于逻辑回归CA。MSTra CA在绝大多数组别的试验数据中均取得最优的模拟结果,最高达到了37.9%;在D_T较少时,相对于逻辑回归CA,最大改善为11.7%。

在实际应用中,往往很难预先得知目标领域的数据分布,因而不能找到分布最接近的源领域数据,很容易出现负迁移现象。MSTra CA可以通过输入多个源领域来减小由于某些源领域数据分布差异较大而产生的负迁移风险。

图5为D_T数量等于10时,逻辑回归CA和MSTra CA模型得到的模拟结果,绿色部分为模型模拟正确的新增城市用地斑块。由于采样偏差的影响,逻辑回归CA模拟的新增城市用地偏差非常大(图5(a)中模拟正确的元胞主要集中在研究区西北部和东北部),而且多次随机采样模拟得到的结果相差非常大,这种结果往往是没有实际应用价值的。而MSTra CA得到的模拟结果误差较小,在空间上分布均匀(图5(b)),而且多次随机采样得到的模拟结果比较接近,证明在缺乏足够数量(样本数小于30个)目标领域的样本数据时,知识迁移CA能够从相似的源领域迁移有用的旧样本,与少量的目标领域样本一起进行训练,得到足够准确的模型。

3 结 语

本文提出了MSTra CA模型,通过输入多个

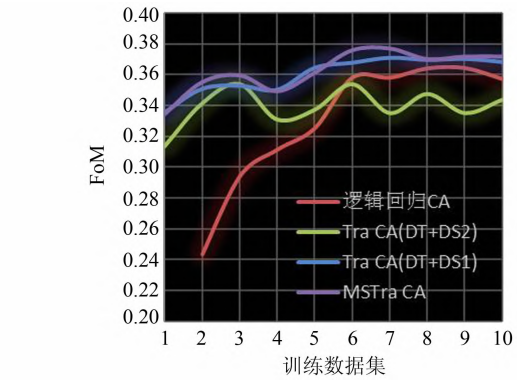

图 4　四个模型的模拟结果评价

Fig. 4　Figure of Merit of the Simulation from Four Methods

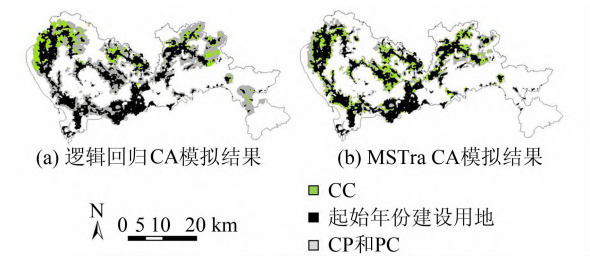

图 5　D_T 数量等于 10 时得到的模拟结果对比

Fig. 5　Simulation of Urban Growth Using 10 D_T Samples

源领域样本数据,帮助建立只有少量样本的目标领域模型。试验结果表明,相对于传统逻辑回归 CA,MSTra CA 在大多数情况下都能获得最佳的模拟结果,并且在目标领域样本数较少时,模拟精度最大增加了 11.7%。而相对于单源领域 TraCA,能够有效地减小产生负迁移的风险,这在知识迁移 CA 的实际应用中非常有意义。由于目标领域数据需要与各个源领域数据集分别做计算,模型的运行时间与 TraCA 相比成倍增加,如何提高 MSTra CA 的运行效率是后续研究需要考虑的问题。另外,模型要求输入的源领域与目标领域具有相类似的城市增长特性,而影响城市增长特征的因素比较复杂,如何定量地评价不同城市扩张的相似程度也是后续研究的重点。

参　考　文　献

[1] Batty M, Xie Y. From Cells to Cities[J]. *Environment and Planning B*, 1994,21:31-48

[2] Wu F. An Experiment on the Generic Polycentricity of Urban Growth in a Cellular Automatic City[J]. *Environment and Planning B: Planning and Design* 1998,25:731-752

[3] Li X. Emergence of Bottom-up Models as a Tool for Landscape Simulation and Planning[J]. *Landscape and Urban Planning*, 2011,100:393-395

[4] ISanté I, García A M, Miranda D, et al. Cellular Automata Models for the Simulation of Real-world Urban Processes: A Review and Analysis[J]. *Landscape and Urban Planning*, 2010,96:108-122

[5] Li X, Yang Q, Liu X. Discovering and Evaluating Urban Signatures for Simulating Compact Development Using Cellular Automata[J]. *Landscape and Urban Planning*,2008,86: 177-186

[6] Li X, Liu Y, Liu X, et al. Knowledge Transfer and Adaptation for Land-use Simulation with a Logistic Cellular Automaton[J]. *International Journal of Geographical Information Science*, 2013, 27: 1 829-1 848

[7] Pan S J, Yang Q. A Survey on Transfer Learning [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010,22:1 345-1 359

[8] Qin Jiangwei. Research of Transfer Learning and Its Application in Classifying Cross-domain Data [D]. Guangzhou: South China University of Technology, 2011(覃姜维. 迁移学习方法研究及其在跨领域数据分类中的应用[D]. 广州:华南理工大学, 2011)

[9] Torrey L, Shavlik J. Transfer Learning[J]. *Handbook of Research on Machine Learning Applications*, *IGI Global*, 2009,3:17-35

[10] Dai W, Yang Q, Xue G R, et al. Boosting for Transfer Learning[C]. The 24th International Conference on Machine Learning, Corvallis, OR,2007

[11] Yao Y, Doretto G. Boosting for Transfer Learning with Multiple Sources[C] . 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Niskayuna, NY, 2010

[12] Li X, Chen Y, Liu X, et al. Concepts, Methodologies, and Tools of an Integrated Geographical Simulation and Optimization System[J]. *International Journal of Geographical Information Science*, 2011,25: 633-655

[13] Ye Yuyao, Zhang Hong'ou, Liu Kai, et al. Impact of Site Factors on Expansion of Construction Land: A Case Study in the Pearl River Delta[J]. *Progress in Geography*, 2010, 29:1 433-1 441(叶玉瑶, 张虹鸥, 刘凯,等. 地理区位因子对建设用地扩展的影响分析——以珠江三角洲为例[J]. 地理科学进展, 2010,29:1 433-1 441)

[14] Pontius Jr R G, Millones M. Death to Kappa: Birth of Quantity Disagreement and Allocation Disagreement for Accuracy Assessment[J]. *International Journal of Remote Sensing*, 2011,32:4 407-4 429

Knowledge Transfer and Adaptation for Urban Simulation Cellular Automata Model Base on Multi-source TrAdaBoost Algorithm

LIU Yilun[1,2] *LI Xia*[1,2]

1 School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China

2 Key Laboratory of Urbanization and Geographical Spatial Simulation, Sun Yat-sen University, Guangzhou 510275, China

Abstract: Traditional cellular automata (CA) cannot adequately simulate urban dynamics and land-use changes when there are insufficient training samples. To address this problem, we propose a multi-source knowledge transfer CA model. This model utilizes several existing label data sets to help train a new model. This proposed model, MSTra CA, is employed to urban simulation in Shenzhen City in Guangdong Province of China. Experiments have demonstrated that the proposed method can alleviate the sparse data problem using knowledge transfer thus reducing the negative transfer learning risk.

Key words: knowledge transfer; cellular automata; urban simulation; multi-source TrAdaBoost

First author: LIU Yilun, PhD candidate. His research interests include spatial data mining, urban simulation and agent based model. E-mail: ealenliu@gmail.com

Foundation support: The National Natural Science Foundation of China, No. 41371376.

(上接第694页)

Experimental Geography Based on Virtual Geographic Environments (VGEs)

LIN Hui[1,2,3] *CHEN Min*[1,2]

1 Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Shatin, Hong Kong

2 Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen 518057, China

3 Department of Geography and Resource Management, The Chinese University of Hong Kong, Shatin, Hong Kong

Abstract: In recent years, the structure and functions of Virtual Geographic Environments (VGEs) have become clearer and how they could be used to support geographic analysis and geographic experiment has received serious attention. Based on the analysis of the characteristics of geographic experiments, this paper discusses the potential contributions of VGEs to traditional geographic experimentation, thus to encourage researchers to perform geographic experiments based on VGEs in a fused reality-virtuality and collaborative way.

Key words: virtual geographic environments; geographic experiment; reality-virtuality fusion; collaboration

First author: LIN Hui, PhD, professor, Academician of Euro-Asia International Academy of Sciences. He specializes in Virtual Geographic Environments(VGEs) and remote sensing applications. E-mail: huilin@cuhk.edu.hk

Corresponding author: CHEN Min, PhD, associate researcher. E-mail: chenmin0902@cuhk.edu.hk

Foundation support: The National Natural Science Foundation of China, Nos. 41171146, 41101439, 41371424.