

改进最短距离法自动获取元胞自动机转换规则^{*}

刘小平¹, 黎夏¹, 陈蕾²

(1 中山大学遥感与 GIS 工程系, 广东 广州 510275)

2 国家海洋局南海海洋工程勘察与环境研究院, 广东 广州 510300)

摘 要: 提出了一种利用改进的最短距离算法自动获取地理元胞自动机转换规则的方法。CA 的核心是如何定义转换规则, 但目前主要是采用启发式的方法来定义转换规则, 受主观因素影响较大。该模型通过熵化空间变量特征权重, 对最短距离算法进行改进, 自动获取 CA 的转换规则和模型参数值。并与一般的最短距离算法进行对比分析, 结果表明, 改进后的最短距离算法所提取的转换规则在模拟城市发展时具有更高的精度, 并且具有清晰的物理意义。

关键词: 最短距离; 熵; CA; 转换规则

中图分类号: U945.23 文献标识码: A 文章编号: 0529-6579 (2006) 04-0103-04

元胞自动机具有强大的空间运算能力, 常用于自组织系统演变过程的研究。近年来, 随着元胞自动机理论的深入, 许多学者开展了元胞自动机在城市模拟方面的研究^[1-5]。这些研究表明, 通过简单的局部转换规则可以模拟出复杂的城市空间结构。

转换规则的定义是 CA 模型的核心。目前, 有关 CA 模型转换规则定义的方法存在一定的缺陷。Clarke^[6]提出了利用肉眼判断的方法来获取模型参数值。该方法受主观因素影响很大, 当空间变量较多时, 如何选择合适的组合方案是一个棘手的问题。Wu^[7]曾提出利用层次分析法 (AHP) 来确定模型参数值, 但这种方法需要专家打分, 也具有一定的主观性。刘小平等^[8]提出利用核学习机提取 CA 非线性转换规则的方法, 但是, 该方法获取的转换规则其物理意义并不清晰。黎夏等^[9]提出了利用神经网络训练的方法自动获取 CA 模型的转换规则, 不足的是, 该方法属于黑箱结构, 对模型参数的物理意义难以解释。Li 等^[10]随后又提出了利用决策树的方法来获取 CA 的转换规则, 但这种方法产生的决策规则较多, 选择合适的决策规则较为费时。

本文提出了利用改进的最短距离算法来自动获取 CA 模型参数值的新方法, 该方法的特点是简单实用, 无需人为主观的确定 CA 模型参数值。利用熵方法获取的各空间变量的权重值, 具有较强的物

理意义, 能够清晰地揭示城市发展的演化模式。

1 基于熵化最短距离的 CA 模型

最短距离算法是遥感分类常用的一种方法^[11], 该方法的特点是简单实用。最短距离算法的基本思想是计算训练数据每一类别的均值向量, 然后以均值向量作为该类在特征空间中的中心位置, 计算待分类对象到各类中心的距离, 把待分类对象归入到距离最小的一类中去, 距离是这种方法的判别准则。常用的距离函数采用欧式距离函数:

$$d_m^2 = \sum_{n=1}^N (x_n - M_{mn})^2 \tag{1}$$

其中, d_m^2 为待分类对象到第 m 类中心欧式距离的平方, x_n 为待分类对象的第 n 个特征值, M_{mn} 为第类第 n 个属性特征的均值。但是欧式距离函数有明显的缺陷^[11], 因为不同类别的属性特征方差大小是不同的, 简单地用对象到类中心的距离来划分对象的归属较容易发生差错 (图 1)。在图 1 中, 若只简单的根据到类中心的距离, 则会将待分类对象归入到类 B, 而实际应该归入到类 A。

因此, 考虑到各类别的标准差向量, 对欧式距离函数进行改进^[12], 则式 (1) 可变为:

$$d_m^2 = \sum_{n=1}^N (x_n - M_{mn})^2 / \sigma_{mn}^2 \tag{2}$$

式中 σ_{mn}^2 为第 m 类第 n 个属性特征的方差。在式

^{*} 收稿日期: 2005-10-17

基金项目: 国家杰出青年基金资助项目 (40525002); 国家自然科学基金资助项目 (40471105); “985 工程”GIS 与遥感的地学应用科技创新平台资助项目 (105203200400006)

作者简介: 刘小平 (1978 年生), 男, 博士生; 通讯联系人: 黎夏; E-mail: lixia@mail.sysu.edu.cn

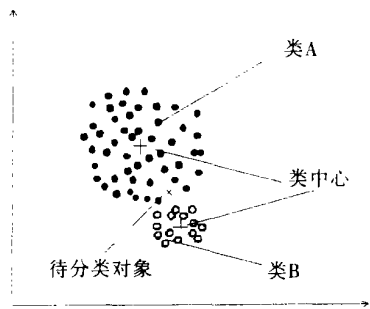


图 1 最小距离可能导致的误分

Fig 1 Possible error classification induced by Minimum distance

(2) 中，对于第 m 类的每一个属性特征对欧式距离的贡献都采用一样的权重，但是在这些属性特征中，有些属性特征与分类强相关；有些属性特征与分类弱相关；还有些属性特征与分类不相关。如果存在较多的属性特征与分类不相关，则欧式距离会被大量不相关特征所支配，这样会导致分类出现误差。因此，需要确定每个属性特征对欧式距离的“贡献”度，即权重，式 (2) 可以修正为：

$$d_m^2 = \sum_{n=1}^N w_n (x_n - M_m)^2 \rho_{nm}^2 \tag{3}$$

其中， w_n 为第 n 个属性特征的权重，我们利用熵的方法来确定各属性特征的权重。熵是信息论中关于不“确定性”的度量，信息量越大，不确定性就越小，熵也越小。反之，信息量越小，不确定性越大，熵也越大。根据熵的特性，可以通过计算熵值来判断一个事件的随机性及无序程度。我们也可以利用熵值来判断对象某个属性特征的离散程度。对于某个属性特征，若各个对象的值没有太大区别，则该属性特征离散程度较小，在综合分析中所起的作用不大，对欧式距离的“贡献”较小；反之，若对某个属性特征而言，各个对象的值有很大的波

动，即该特征的离散程度很大，则这个特征对综合分析有很重要的“贡献”。因此，把熵引入最短距离算法中，用熵值来衡量属性特征的重要性。根据熵的定义^[13]，则第 n 个属性特征信息熵为：

$$E_n = -K \sum_{h=1}^H (f_{hn} \lg f_{hn}) \tag{4}$$

其中， H 为总的样本数， $K = 1 / \lg H$ ， $f_{hn} = |a'_{hn}| / \sum_{h=1}^H |a'_{hn}|$ ，并假定：当 $f_{hn} = 0$ 时， $f_{hn} \lg f_{hn} = 0$ 。 a'_{hn} 为消除量纲影响后的第 n 个样本第 n 个属性特征值，则第 n 个属性特征信息熵权表示为：

$$w_n = (1 - E_n) / (N - \sum_{h=1}^H E_n) \tag{5}$$

从式 (5)、(6) 可以看出，当某个属性特征在各样点上的值完全相同时，熵值达到最大，熵权为零，这意味着该属性特征对欧式距离的“贡献”为零，应该剔除该空间距离变量。反之，当某个属性特征在各样点上的值相差较大时，熵值较小，熵权较大，这说明该空间距离变量对欧式距离的“贡献”较大。

将改进的最短距离算法应用在地理元胞自动机中，自动从训练数据获取 CA 的转换规则。在利用 CA 模拟城市系统时，往往需要选取系列空间变量及局部变量 (表 1)。在城市 CA 模型中，单元的状态只有两种：城市或非城市，因此训练数据只分为两类，即城市单元和非城市单元。式 (3) 中的 x_n 对应为单元的各空间距离变量，当 $m = 1$ 时，单元状态为城市， m_n 和 σ_{n1}^2 分别为城市单元各空间距离变量的均值和方差；当 $m = 2$ 时，单元状态为城市， m_n 和 σ_{n1}^2 为非城市单元各空间距离变量的均值和方差。根据式 (3) 分别计算待判别单元到城市单元中心和城市单元中心的改进欧式距离，根据最短距离判别准则判断该单元是否发展为城市用地。

表 1 CA 模型变量

Tab. 1 Variables of the CA model

空间距离变量						局部变量	
离市中心 的距离	离镇中心 的距离	离村委 的距离	离国道、 省道的距离	离乡道 的距离	离铁路 的距离	离高速 公路的距离	3×3 邻域内 单元约 束条件

最短距离判断准则是通过布尔规则来确定单元是否发展为城市。许多研究表明^[14]，地理复杂现象具有一定的不确定性，利用布尔规则来判断复杂的地理现象，结果并不理想。为此，我们有必要把模糊概念引入到地理现象的判断过程中，用概率的形式来确定所属类别。结合模糊数学，则单元转化

为城市的发展概率可由下式表示：

$$P_{dev}(ij) = \frac{d_1^2}{d_1^2 + d_2^2} = \left(\sum_{n=1}^N w_n (x_n - M_{n1})^2 \rho_{n1}^2 \right) / \left(\sum_{n=1}^N w_n (x_n - M_{n1})^2 \rho_{n1}^2 + \sum_{n=1}^N w_n (x_n - M_{n2})^2 \rho_{n2}^2 \right) \tag{6}$$

由式 (6) 所计算的单元发展概率只考虑到空间距离变量对其转化的影响，而 CA 模型的邻域影响是一

个非常重要的因素，因此，还需要重新考虑邻域对中心单元的影响。在前面进行判别分析时，没有考虑邻域影响，是因为邻域内的城市单元数目是动态变化的，需要单独考虑其对中心单元的影响，选择作为邻域窗口，定义 $t-1$ 时刻的邻域影响值为：

$$\Omega_{t-1}(ij) = \frac{\sum_{3 \times 3} N(\text{urban}(ij))}{3 \times 3 - 1} \tag{7}$$

其中， $\sum_{3 \times 3} N(\text{urban}(ij))$ 为 3×3 邻域窗口内城市化像元数。

同时，还需要考虑外部的约束条件，如道路、河流、陡峭的山地、优质的农田等发展为城市用地的概率较小。综合考虑局部邻域范围、单元约束条件及空间距离的影响，某单元在时刻发展为城市用地的概率可由下式表达：

$$P_t(ij) = AP_{\text{dev}}(ij) \text{con}(\text{suit}(ij)) \Omega_{t-1}(ij) \tag{8}$$

其中， $\text{con}(\text{suit}(ij))$ 为单元约束条件总值，其值为 $0 \sim 1$ ， A 为模型调整参数， $\Omega_{t-1}(ij)$ 为 $t-1$ 时刻的邻域影响值。

3 模型应用及检验

本模型选择东莞市作为试验区，首先利用 1988 年和 1993 年的 TM 卫星遥感图象自动分类，提取模型所需要的训练数据。从分类数据中选择训练数据。本文运用随机分层取样的方法，从转换为城市用地的单元和可能转换为城市单元而尚未转换的单元中分别选择 20% 的样点，获取这些样点的空间坐标，再运用 ARC/INFO 的 Sample 功能分层读取对应所需的模型空间变量。分别计算城市单元和非城市单元的均值和方差，根据式 (4) 和 (5) 计算空间变量的熵值权重，经计算，市中心、镇中心、村委、国道省道、乡道、铁路、高速公路等各空间距离的熵值权重分别为：0.050、0.136、0.094、0.401、0.197、0.079、0.043。各空间距离变量熵值权重具有非常清晰的物理意义，熵值权重越大，则该空间距离变量对欧式距离的贡献也就越大。计算结果表明：公路和镇中心的距离变量熵值权重较大，尤其是国道和省道，由此可以得知：东莞市 1988 - 1993 年间主要是沿国道、省道、乡道及城镇中心发展，沿路网发展模式尤为明显。

获取各空间距离变量熵值权重后，根据式 (8) 即可求得每个单元发展为城市用地的概率，求出单元发展概率后，还需要判断该单元是否发展，传统的做法是给定一个阈值 ($0 \sim 1$)，比较单元发展概率和阈值的大小。公式表达如下：

$$S_{t+1}(ij) = \begin{cases} \text{Development } P_t(ij) > P_{\text{threshold}} \\ \text{UnDevelopment } P_t(ij) \leq P_{\text{threshold}} \end{cases} \tag{9}$$

模拟时，初始状态为 1988 年 TM 影像分类所获取的城市用地，模型进行 200 次迭代运算，模拟结果见图 2。从图 2 可以看出，模拟过程中城市主要沿国道和省道等轴线发展，或者是以城镇为中心发展。这与实际情况非常相符，在 20 世纪 90 年代初，东莞的经济飞速发展，大批工业进驻该地区，需要大量的土地。这些开发用地主要是沿公路进行发展，因此城市发展没有形成紧凑式的形态，这种松散式的城市发展形态大大地增加了能源的消耗，并导致土地资源的浪费。

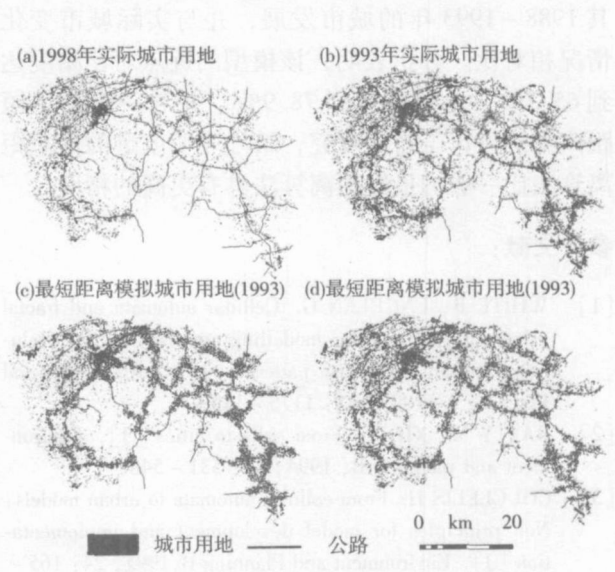


图 2 模拟结果与实际城市发展对比图
Fig.2 Comparison between the simulated patterns and the actual development patterns

我们也用一般的最短距离法获取的 CA 转换规则模拟了东莞市 1988 - 1993 年城市的发展 (图 2)，以更好的检验改进最短距离算法的优点。从图 2 可以看出，改进后最短距离算法模拟的结果跟实际情况非常相似，而一般最短距离法模拟的结果跟实际情况有较明显的差异，虽然大致形态也是沿国道和省道等轴线发展，但显得过于分散。为了进一步检验模型，分别计算了一般最短距离法和改进最短距离法模拟的逐点对比精度，一般最短距离法的转变精度为 59.6%，不转变精度为 78.4%，总精度为 75.3%。改进最短距离法的转变精度为 65.7%，不转变精度为 83.2%，总精度为 78.9%。可见，改进最短距离法模拟的逐点对比精度要比一般最短距离法好得多。

4 结 论

城市是一个异常复杂的巨系统, 元胞自动机则是一种进行复杂系统分析与模拟的重要手段, 它有着强大的空间自组织能力, 在模拟城市发展时有着很大的优势。CA 模型的关键是如何定义转换规则和确定模型参数值。本研究通过改进的最短距离算法自动获取 CA 模型转换规则, 所获取的转换规则物理意义清晰, 简单实用。

将该模型应用在珠江三角洲地区发展最快的东莞市, 利用不同年份的卫星遥感图像作为主要观察数据。根据所提出的方法来获取转换规则, 模拟了其 1988-1993 年的城市发展, 并与实际城市变化情况相对比。分析表明, 该模型的逐点对比精度达到 65.7%, 总精度达到 78.9%。并与一般的最短距离算法进行了对比研究, 结果表明改进的最短距离算法比一般的最短距离算法具有更高的精度。

参考文献:

- [1] WHITE R, ENGEL G. Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land-use patterns [J]. *Environment and Planning A*, 1993, 25: 1175-1199.
- [2] BATTY M, XIE Y. From cells to cities [J]. *Environment and planning B*, 1994, 21: 531-548.
- [3] COUCLELIS H. From cellular automata to urban models: New principles for model development and implementation [J]. *Environment and Planning B*, 1997, 24: 165-174.
- [4] LI X, YEH A G O. Modelling sustainable urban development by the integration of constrained cellular

- automata and GIS [J]. *International Journal of Geographical Information Science*, 2000, 14(2): 131-152.
- [5] WU F L. Calibration of stochastic cellular automata: the application to rural-urban land conversions [J]. *International Journal of Geographical Information Science*, 2002, 16(8): 795-818.
- [6] CLARKE K C, HOPPEN S, GAYDOS L. A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area [J]. *Environment and Planning B: Planning and Design*, 1997, 24: 247-261.
- [7] WU F L. SimLand: a prototype to simulate land conversion through the integrated GIS and CA with AHP-derived transition rules [J]. *International Journal of Geographical Information Science*, 1998, 12(1): 63-82.
- [8] 刘小平, 黎夏. 从高维特征空间中获取元胞自动机的非线性转换规则 [J]. *地理学报*, 2006, 61(16): 663-672.
- [9] 黎夏, 叶嘉安. 基于神经网络的单元自动机 CA 及真实和优化的城市模拟 [J]. *地理学报*, 2002, 57(2): 159-166.
- [10] LI X, YEH A G O. Data mining of cellular automata's transition rules [J]. *International Journal of Geographical Information Science*, 2004, 18(8): 723-744.
- [11] 戴昌达, 姜小光, 唐伶俐, 等. 遥感图像应用处理与分析 [M]. 北京: 清华大学出版社, 2002.
- [12] WANG F J. Fuzzy Supervised Classification of Remote Sensing Images [J]. *IEEE Transactions on GIS*, 1990, 28(2): 194-201.
- [13] THEIL H. *Economics and Information Theory* [M]. North-Holland: Amsterdam, 1967: 488.
- [14] 黎夏, 叶嘉安, 廖其芳, 等. 利用案例推理 CBR 方法对雷达图像进行土地利用分类 [J]. *遥感学报*, 2004, 8(3): 246-253.

An Improved Minimum-Distance Method for Mining Transition Rules of Cellular Automata

LIU Xiao-ping, LIXia, CHEN Lei

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

Abstract A new method was presented to retrieve transition rules of cellular automata (CA) using an improved minimum-distance method. The core of CA is how to define transition rules in a consistent way. In most situations, transition rules of CA are defined by using heuristic methods which may be subject to some uncertainties. For example, the multicriteria evaluation (MCE) may be used to determine the parameters of transition rules. However, this method can be influenced by expert preferences. The proposed model is to define transition rules by using an improved minimum-distance method, in which the weights are determined by entropy. This method has been applied to the simulation of a fast growing city, Dongguan, in the Pearl River Delta. Comparison indicates that this method can produce better simulation accuracies than the general minimum-distance method. Its model structure is transparent and can be easily understood by users.

Key words minimum distance; entropy; CA; transition rule