

文章编号: 1000 - 0585(2002)06 - 0675 - 07

# 城市土地利用演变信息的数据挖掘 ——以上海市为例

王 铮<sup>1, 2</sup>, 吴健平<sup>1</sup>, 邓 悦<sup>1</sup>, 王凌云<sup>3</sup>, 熊云波<sup>1</sup>

(1. 华东师范大学教育部城市与环境开放实验室, 上海 200062;

2. 中国科学院政策与管理研究所, 北京 100080; 3. 阿拉斯加大学, 美国)

**摘要:** 城市土地利用变化, 具有非线性特征, 一般的数据挖掘方法基本上失效。本文研究了利用马尔可夫链和神经网络两种方法, 基于地理信息系统、遥感图像、电子地图, 预测了上海市中心城区、2002 年和 2005 年的土地利用总量和土地利用类型结构的变化, 从而研究了城市土地利用状况演变预测的地质数据挖掘技术。

**关 键 词:** 土地利用; 地质数据挖掘; 马尔可夫链; 人工神经网络

**中图分类号:** F293. 2; P208 **文献标识码:** A

## 1 导言

数据挖掘是随着计算机的普遍应用发展起来的<sup>[1]</sup>, 在地学及遥感领域, 存在大量数据挖掘问题。例如王雷等<sup>[2]</sup>用知识发现方法挖掘遥感影像的土地覆盖类型。数字城市的兴起, 提出了通过遥感图像挖掘城市空间结构及土地利用信息的问题。城市土地结构, 具有明显的复杂性<sup>[3]</sup>, 一般的回归模型、时间序列模型是一种线性模型, 对具有复杂性特性的城市空间结构可能失效。为了认识城市空间结构的动态发展, 需要发展新的面向数字城市问题的数据挖掘模型。为此, 我们尝试了在 GIS 基础上用马尔可夫链和神经网络两种方法挖掘空间城市中心城区变化信息的技术。同时讨论二者是因为在理论方面 De Wilde<sup>[4]</sup>强调了二者的联系; 虽然马尔可夫方法已经有人应用<sup>[5, 6]</sup>, 但是这种应用没有推广到复杂的城市土地利用结构中; 特别是二者的适应性需要同时比较。

本研究所用的原始数据是教育部城市与环境遥感开放实验室的梅安新、吴建平长期工作获得的上海市中心城区土地利用电子地图。这套地图依年份划分, 共八幅 (1947、1958、1964、1979、1984、1988、1993、1996 年)。数据挖掘的基本思想是将电子地图中各种土地利用类型 (或土地类型) 作为系统数据单元, 开展数据挖掘。上海的城市土地利用类型有: 城市工业用地、商业和居住用地、混合建设用地、农业用地、村镇建设用地、待建用地、道路用地和河流用地。商业用地没有与居住用地分开, 因为在遥感图像上商业和居住用地都表现为高楼, 无法区别, 因此不得不把两者合在一块称为“商居用地”。而

收稿日期: 2002 - 05 - 08; 修订日期: 2002 - 09 - 22

基金项目: 国家自然科学基金资助项目: 可计算人地关系协调模型 (49971008)

作者简介: 王铮 (1954 - ), 男, 云南陆良人, 研究员, 博士生导师。主要研究方向: 理论地理学、地理信息科学及区域管理。

待建和建设用地是正在转换的土地利用类型，是城市化过程中的混合建设用地。

## 2 马尔可夫链方法

### 2.1 转移矩阵挖掘

本文使用的第一种方法是马尔可夫链方法。马氏链主要应用于在无后效条件下时间和状态均为离散的随机转移问题。也就是说，其观察量第  $n$  次观察值仅和该观察量的第  $t-1$  次观察值及一个状态转移矩阵有关，而和  $t-1$  次观察值以前的状态情况无关。马氏链已经被广泛用在地理学中。但是马氏链方法如何作为面向数字城市信息提取的数据挖掘技术，需要探索。马氏链的原理可以由下列转换关系表示

$$S(t+1) = A(t+1, t) S(t) \quad (1)$$

式中： $S(t)$  是时刻  $t$  的系统状态， $A(t+1, t)$  状态是从时刻  $t$  转向时刻  $t+1$  状态的转移矩阵，在稳定的条件下，常取为常矩阵。因此，要对城市土地利用类型演变信息的提取和预测，我们必需知道观察值和相对应的状态转移矩阵。

作为数字城市技术，状态转移矩阵的获取需要与遥感信息分析结合。我们的做法是根据实际资料估计从某一时间点转移到下一时间点的各种土地利用类型相互转化的概率矩阵。在上海试验区我们引用了梅安新等从遥感图像得到的上海土地利用电子地图，该地图在 GIS 流行软件 ArcView3.1 平台下实现，数据形式是土地利用状况矢量图和相关的表格。该数据表格以矢量图中的图斑的 ID（系列号）作为惟一识别标志。每一块图斑对应 8 种土地利用类型中的一种，分别用 1, 2, 3, 4, 5, 6, 7, 8 表示。

### 2.2 地学的处理

我们知道城市的结构是有序的，建立中心商务区（CBD）不同，城市的土地利用性质不同。早在 1925 年伯吉斯就发现城市的空间结构呈有序环状模式<sup>[7]</sup>，各圈层土地利用状态转移的概率是不同的。因此对整个上海土地利用情况总结应用 2.1 节的方法是不恰当的，为此，我们将上海城市划分为 5 个圈层，求各层土地利用类型转化情况。5 个圈层第一个为上海传统的 CBD，它以人民广场为中心，到黄浦江为止，第二层大致为最近 20 年的扩张 CBD，包括浦东商业区。在第四带中可以观察到工业扇区的特征，第五圈层包括郊区。为了数学运算的等价，我们取每个圈层半径一致。5 个圈层逐渐扩大，这样的划分后面的圈层始终包含城市内部状况，这是考虑城市土地利用转移存在向城市中心用地单向转移的特点选择的。

确定圈层后，我们用一段 Avenue 程序分别统计出这 5 幅图对应的 5 个状态转移概率矩阵和圈层相减得到 5 个环，其中的土地利用状态对应 5 个状态向量。在计算第  $j$  环土地将 8 个类型数据作为状态数据，决定状态之间转移是  $j$  圈对应的状态转移矩阵。

应用上述方法，我们估计了状态转化的转移矩阵，它就成为了数据挖掘的技术工具。利用计算了环带的状态转化，最后对各环带状态求和，求出城市土地利用状态变化状况。在表 1 中，我们给出了利用本文发展的方法预测的 2002 年和 2005 年土地利用类型的结构，同时列出 1996 年的土地利用情况以作对比。从表 1 可以看出，马氏链方法得到的结果具有与城市演化一致的特征：2000 年后城市内的农业用地将继续减少，而待建用地将上升。这表明城市中的农业用地将不断被开发，由于土地利用类型的转换有一定的时滞，不及时利用的土地就成为城市建设用地和待建用地。同时，居住和商业用地和道路用地将

有略微增长，也反映了上海的现实情况。

表 1 马尔可夫链预测 2002 年及 2005 年上海城市土地利用类型结构 (单位: km<sup>2</sup>)

Tab. 1 The forecasting of land - use type on Markov Chain (unit : km<sup>2</sup>)

土地利用类型	商居用地	城市工业用地	混合建设用地	农业用地	村镇建设用地	待建用地	道路用地	河流
1996 年 (已知)	61. 4050	33. 0700	31. 9950	12. 1650	5. 7950	10. 0570	11. 9170	11. 5930
2002 年 (预测)	61. 6704	33. 0123	32. 0614	9. 9891	5. 6631	11. 9169	11. 8262	11. 5931
2005 年 (预测)	61. 8055	32. 9993	32. 1599	9. 0504	5. 5946	12. 7138	11. 9175	11. 5931

3 神经网络算法

在本质上，马氏链模型是线性的，为了认识复杂性，我们需要非线性模型。神经网络算法是分析非线性现象的数据挖掘工具，与 GIS 一起被成功地应用在自然地理现象的分析和预报中<sup>[8,9]</sup>。关于它的内容细节见参考文献 [4] 和 [10]，这里不再赘述。它的基本原理如图 1 所示。开始的状态作为输入，拟预报的状态作为输出，中间过程的处理类似神经网络，一个输入点受到刺激，中间的隐层点按照各自一定的权重响应，隐层可能有几层。权重不同，响应结果不一样。预测估计结果与实际结果的差被用以调整权重和隐层特点的构成单元数，这个过程称为训练或学习。这个算法是一个非线性过程。

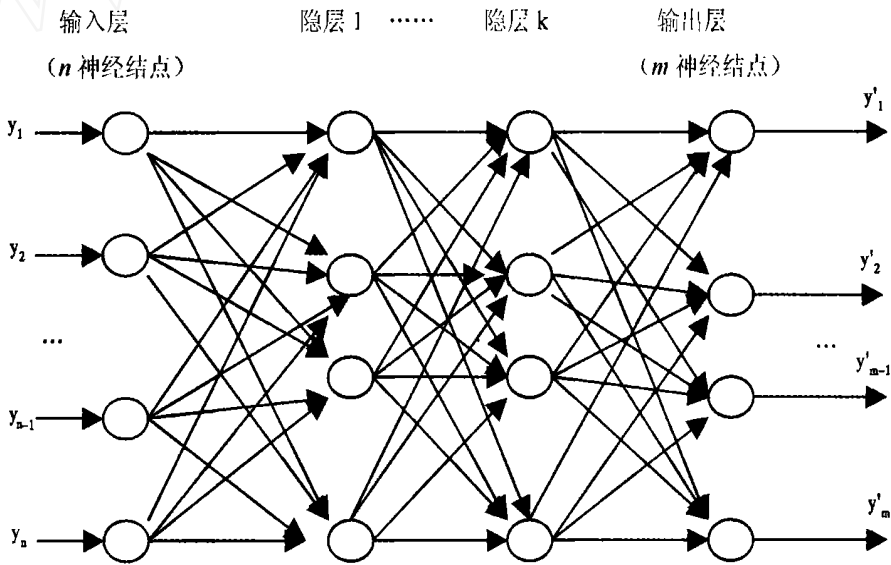


图 1 神经网络算法模型

Fig. 1 Model of an arithmetic of ANN (Artificial Neural Network)

根据上海土地利用类型特征，我们的第一个模型采用了一个 8 输入、8 输出的神经网络，我们称它为方法 1，以区别后面将要使用的方法。方法 1 以相邻年份的土地面积作为输入和输出的目标值，训练网络，常用的具体训练方法是常见的 BP 算法。预测将第  $t$  年的 8 种土地利用类型作为输入，预报第  $t + d$  年的土地利用类型面积，这里  $d$  是预报步长。实际工作中，我们用前 7 个年份资料训练样本，预报 1996 年值与实测值比较，得到

误差, 训练神经网络。当满足一定的误差要求时, 即固定神经网络的权值。输入待预测年份的面积数据, 输出即为需要挖掘出的数据信息。其中中间隐含层的神经接点个数的确定需要经过不断地反复调整, 使得误差图中显示的误差最小; 在上海例子中经过反复调整, 发现隐层为 18 个接点时最为合适。

在训练神经网络的时候, 考虑到 BP 神经网络对接近 0 或者 1 的数据学习不是很理想, 故我们对原始数据按下式对其进行归一化, 使其值处于 0 和 1 的中间。在计算完成后, 为了消除系统误差, 我们对计算得到的数据作了重整化处理, 使得各土地利用类型预测的结果之和等于土地的总面积。用方法 1 建立的神经网络模型预报的土地利用类型误差情况给出在图 2 中。可以看出, 在经过 3000 步训练后, 误差稳定在 10 % 附近。稳定性表明人工神经网络方法可以用于城市土地利用预测, 问题在于如何进一步提高预报精度。

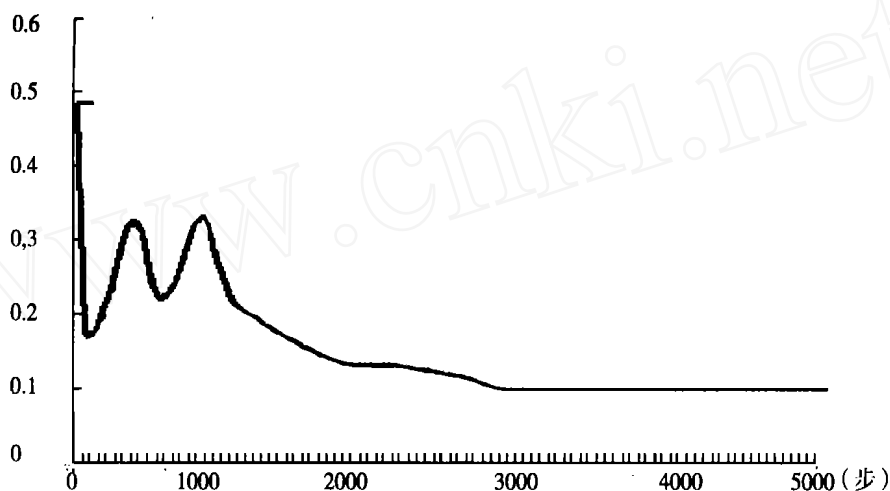


图 2 神经网络算法的总计算误差变化

Fig. 2 The change of computation error on ANN

表 2 给出了类型估计的误差情况。可以看出, 应用模型 1 对各种典型城市用地类型 (城镇商居用地, 城市工业用地和城市待建用地) 估计精度偏低。考察这个问题的原因是有意义的。模型 1 中活跃函数采取了相同的权重因子, 这意味着土地转化是各向同性的, 但是, 城市土地利用具有方向性, 基本上只有农田和 (小) 城镇用地会转化为城市用地, 其他土地转化为城市用地的机会小; 其次, 在上海的城市进程中商居用地几乎不向其它类型用地转化, 所以它是一个吸收壁, 神经网络方法的 BP 算法特别是模型 1 需要改进。

根据城市土地转化的方向性特征, 为了克服误差, 我们探讨了模型 2。模型 2 将给定的土地利用类型的 1 到  $n-1$  年的该土地利用类型面积看作一个输入矩阵, 输出 2 到  $n$  年的土地利用类型面积, 真正预报的是第  $n$  年的土地利用类型面积, 这一方法考虑了不同土地利用类型转化时活跃函数权重有差异。计算中我们估计我们的数据量把输入取作对该种土地利用类型逼近, 模型的隐层 1 和隐层 2 均为 9 个神经结点, 如图 3 所示。计算结果, 城镇用地的误差明显降低。而且, 实际中变化不大的道路用地和河流面积, 其预测结果也只有较小的变化。在技术上, 由于 BP 神经网络对于单点逼近是很快的, 方法 2 很快得到满意的结果。



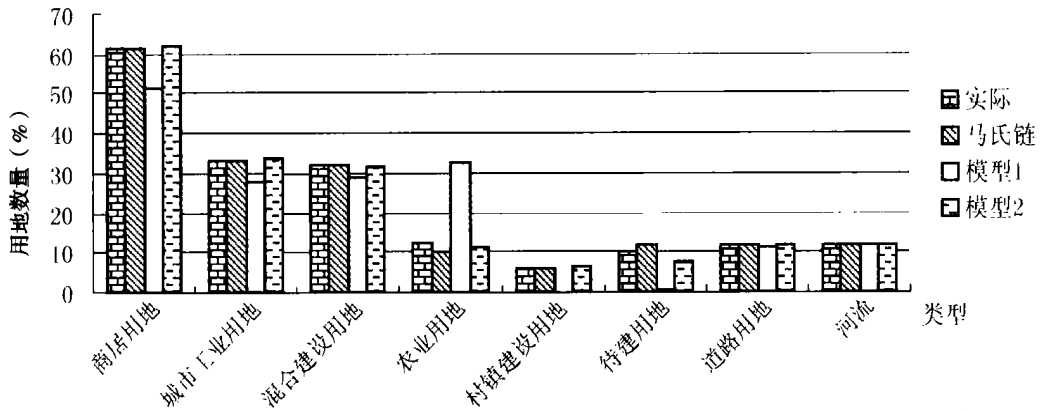


图 4 3 种方法的估计结果比较

Fig. 4 The comparison of estimated results of 3 methods

样难以保证等间距。上海是富裕的城市，目前尚未得到等间距采样，其它城市更难。另外由于经济发展的不均匀，采样时间间隔应该与 GDP 增长量联系起来，可能完成一个增长量，进行一次采样，或者完成一个增长百分点，进行一次采样，这是需要另外研究的。按照第一种观点，由于经济发展的基数加大，后面的年代，采样间隔要缩短。上海采样是这样做的，政府领导人根据感觉发现土地利用有明显变化了，就布置一次采样。

## 4 讨论与结论

本研究最后得到了上海市土地利用趋势数据，并为上海市城市发展战略规划所采用。

通过引用马氏链方法和神经网络算法对上海市空间结构的演化预测分析，我们得到了地学数据挖掘的几个认识。

1. 马氏链方法和神经网络算法对城市土地利用类型变化的预报都有较好的效果，本研究表明，马氏链方法以考虑环带内变化为较好，神经网络算法在考虑土地利用的方向性方面为好。

2. 研究表明，对于仅仅转出的土地利用类型，标准的神经网络 BP 方法给出的误差较大，一种解释是像村镇这种土地转化，不能采用与其它土地相同的权重因子，土地利用类型的转化不是各向同性的，这就使得神经网络算法不如能够描写具有单向性转化性质的马氏链方法。在考虑土地转化的方向性后，神经网络算法的估计精度明显提高。

3. 在提高预测能力方面，神经网络方法还有许多潜力可以发挥，它必然会在地学数据挖掘方面发挥重要作用。

### 参考文献：

- [ 1 ] Han J, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Inc. 中译本, 范明等译. 北京: 机械工业出版社, 2001.
- [ 2 ] 王雷, 冯学智, 都金康. 遥感影像分类与地学知识发现的集成研究, 地理研究, 2001, 20(5): 637 ~ 643.
- [ 3 ] 王铮, 邓悦, 等. 上海城市空间结构的复杂性分析. 地理科学进展, 2001, 20(2): 331 ~ 340.
- [ 4 ] de Wilde P. Neural Network Models: Theory and Projects. London & Belin: Springer, 1997.

- [ 5 ] 徐建华. 地理学中的现代数学方法. 北京: 高等教育出版社, 1995.
- [ 6 ] 贾华, 祝国瑞. 土地利用规划预测中农作物单产预测的灰色 - 马尔可夫链方法. 武汉测绘大学学报, 1998, 23(2): 149 ~ 152.
- [ 7 ] Burgess E W. The growth of the city. In: Park R E, Burgess E W, McKenzie R D (eds). The City, Chicago. University of Chicago Press. 1925.
- [ 8 ] Gangopadhyay S, Gautam T B, Gupta A D. Subsurface characterization using artificial neural network and GIS. *Journal Computing in Civil Engineering*, 1999, 13(3): 153 ~ 161.
- [ 9 ] Liong S Y. River stage forecasting in Bangladesh: neural network approach. *Journal Computing in Civil Engineering*, 2000, 14: 1 ~ 8.
- [ 10 ] 罗发龙, 李衍达. 神经网络信号处理. 北京: 电子工业出版社, 1993.

## Data mining for geo - information of urban land utility :the case of Shanghai

WANG Zheng<sup>1,2</sup>, WU Jian - ping, DENG Yue<sup>1</sup>, WANG Ling - yun<sup>3</sup>, XIONG Yun - bo<sup>1</sup>

(1. Urban & Environmental Dynamics and Geocomputation Laboratory,  
East China Normal University, Shanghai 200062, China;

2. Policy and Management Institute, CAS, Beijing 100080, China;

3. Alaska University, USA)

**Abstract:** This paper focuses on GIScience - techniques with the involvement of two techniques of data mining, the Markov Chain method and Artificial Neural Network (ANN) method. The results were used to the strategic planning and development of the city of Shanghai. In order to estimate state transition matrix, firstly the city was divided into 5 zones based on GIS, remote sensing images and electronic maps, and then the total amount of land use in Shanghai in 2002 and 2005 was estimated by Markov Chain method. Secondly, changes in land use types in core areas of Shanghai were forecasted with ANN method which shows the ANN Model 2 (Fig. 3) is better than Model 1 (Fig. 1). This indicates that each one has its strong point, for the ANN method can be better used to forecast directional aspect of land use, and the Markov Chain method can be better used to forecast changes of land use within the zones. Here Markov Chain method was found that is wreath to take the inside variety, ANN considered fit the land that make use in class direction is good. It is further found out that merely for the converted land use types, the standard ANN - BP method gives greater error. This can be explained as the transform process of land use types of Shanghai city is directional, a general ANN - BP arithmetic gives each land type conversion with the same weight function. Although it is self - contradictory in directionality of the transformation, the modified model can still overcome this difficulty.

**Key words:** land utility; geo - data mining; Markov Chain; Artificial Neural Network