马尔科夫链模型在 GIS 数据预测中的应用

Application of Markov Chain Model to The Data Prediction of GIS

胡腾波　叶建栲　（浙江师范大学 数理与信息工程学院 浙江 金华 321004）

摘　要： 随着 GIS（地理信息系统）技术的不断发展，其应用领域越来越广泛。随之而来的数据量也逐渐增多。如何从大量的数据中挖掘出有用的信息，进而对数据进行分析，以便预测未来的趋势，这是一个亟待解决的问题。本文利用马尔科夫链模型对 GIS 数据进行预测，提出了几个与之相关的概念，最后通过一个例子来说明。实验结果表明，使用该模型进行预测效果良好。

关键词： GIS 数据挖掘 马尔科夫链模型 状态转移概率 状态转移概率矩阵 预测

1 引言

数据挖掘技术在许多领域都有着广泛的应用[1,2]。数据挖掘领域中有许多最新的研究成果，如关联规则、Web 挖掘、马尔科夫链模型等[3-5]。其中马尔科夫（Markov）链模型是近几年来在数据挖掘方面的一个研究热点。

马尔科夫链，是在数学领域中具有马尔科夫性质的离散时间随机过程。该过程中，在给定当前知识或信息的情况下，过去（即现在时期以前的历史状态）对于预测将来（即现在时期以后的未来状态）是无关的。如果 n 个连续变动事物，在变动过程中，其中任一次变动的结果都具有无后效性，那么，这 n 个连续变动事物的集合就叫做马尔科夫链，这类事物演变的过程称为马尔科夫过程[6]。

近年来，马尔科夫链模型在 GIS 中有着广泛的应用[7-10]。本文的贡献在于：利用马尔科夫链模型对 GIS 数据进行预测，最后通过一个实例进行说明。

2 马尔科夫预测法的基本原理

对事件的全面预测，不仅要能够指出事件发生的各种可能结果，而且还必须给出每一种结果出现的概率，说明被预测的事件在预测期内出现每一种结果的可能性程度。这就是关于事件发生的概率预测。

马尔科夫预测法，就是一种关于事件发生的概率预测方法。它是根据事件的目前状况来预测其将来各个时刻（或时期）变动状况的一种预测方法。马尔科夫预测法是 GIS 中重要的预测方法之一。

2.1 基本概念

为了介绍马尔科夫预测法在 GIS 中的应用，首先介绍几个基本概念。

2.1.1 状态

在马尔科夫预测中，"状态"是一个重要的术语。所谓状态，就是指某一事件在某个时刻出现的某种结果。一般而言，随着事件及其预测的目标不同，状态可以有不同的划分方式。譬如，在商品销售预测中，有"畅销"、"一般"、"滞销"等状态；在农业收成预测中，有"丰收"、"平收"、"欠收"等状态。

2.1.2 状态转移过程

在事件的发展过程中，从一种状态转变为另一种状态，就称为状态转移。譬如，天气变化从"晴天"转变为"阴天"、从"阴天"转变为"晴天"、从"晴天"转变为"晴天"、从"阴天"转变为"阴天"等都是状态转移。

事件的发展，随着时间的变化而所作的状态转移，就称为状态转移过程，简称过程。

2.1.3 马尔科夫过程

若每次状态的转移都只与前一时刻的状态有关而与过去的状态无关，或者说状态转移过程是无后效性的，则这样的状态转移过程就称为马尔科夫过程。

2.1.4 状态转移概率

在事件的变化过程中，从某一种状态出发，下一时刻转移到其它状态的可能性，称为状态转移概率。根据条件概率的定义，由状态 E_i 转为状态 E_j 的状态转移

概率 $P(E_i \to E_j)$ 就是条件概率 $P(Ej/Ei)$，即

$$P(E_i \to E_j) = P(E_j/E_i) = P_{ij} \tag{1}$$

2.1.5 状态转移概率矩阵

假定某一被预测的事件有 $E_1, E_2, \cdots, E_n$，共 n 个可能的状态。记 Pij 为从状态 E_i 转为状态 Ej 的状态转移概率，作矩阵

$$P = \begin{bmatrix} P_{11} & P_{12} & \cdots P_{1n} \\ P_{21} & P_{22} & \cdots P_{2n} \\ \vdots & \vdots & \vdots \\ P_{n1} & P_{n2} & \cdots P_{nn} \end{bmatrix} \tag{2}$$

则称 P 为状态转移概率矩阵。

如果被预测的某一事件目前处于状态 E_i，那么在下一时刻，它可能由状态 E_i 转向 $E_1, E_2, \cdots E_i \cdots E_n$ 中的任一个状态。所以 P_{ij} 满足条件：

$$\begin{cases} 0 \le P_{ij} \le 1 (i, j = 1,2,\cdots,n) \\ \sum_{j=1}^{n} P_{ij} = 1 (i = 1,2,\cdots,n) \end{cases} \tag{3}$$

一般地，我们将满足条件(3)的任何矩阵都称为概率矩阵。不难证明，如果 P 为概率矩阵，则对任整数 m > 0，矩阵 Pm 都是概率矩阵。

如果 P 为概率矩阵，而且存在整数 m > 0，使得概率矩阵 Pm 中诸元素皆非零，则称 P 为标准概率矩阵。可以证明，如果 P 为标准概率矩阵，则存在非零向量 $\alpha = [x_1, x_2, \cdots, x_n]$，而且 x_i 满足 $0 \leqslant x_i \leqslant 1$

及

$$\sum_{i=1}^{n} x_i = 1 \text{，使得 } \alpha P = \alpha \tag{4}$$

这样的向量 α 称为平衡向量或终极向量。

计算状态转移概率矩阵 P，就是要求出每个状态转移到其它任何一个状态的转移概率 Pij(i,j = 1,2,…,n)。为了求出每一个 Pij，我们采用频率近似概率的思想来计算。

2.1.6 马尔科夫性

马尔科夫性:过程在时刻 t 所处的状态为已知的条件下，过程在时刻 t+1 所处状态的条件分布与过程在时刻 t 之前所处的状态无关。

2.2 马尔科夫预测法

为了运用马尔科夫预测法对事件发展过程中状态出现的概率进行预测，还需要再介绍一个名词:状态概率 $\pi_j(k)$。$\pi_j(k)$ 表示事件在初始(k = 0)状态为已知的条件下，经过 k 次状态转移后，第 k 个时刻处于状态 Ej 的概率。根据概率的性质，显然有：

$$\sum_{j=1}^{n} \pi_j(k) = 1 \tag{5}$$

从初始状态开始，经过 k 次状态转移后到达状态 Ej 这一状态转移过程，可以看作是首先经过(k - 1)次状态转移后到达状态 $E_i (i = 1,2,\cdots,n)$，然后再由 Ei 经过一次状态转移到达状态 Ej。根据马尔科夫过程的无后效性及 Bayes 条件概率公式，有

$$\pi_j(k) = \sum_{i=1}^{n} \pi_i(k-1) P_{ij} (j = 1,2,\cdots,n) \tag{6}$$

若记行向量 $\pi(k) = [\pi_1(k), \pi_2(k), \cdots, \pi_n(k)]$，则由(5)式可得逐次计算状态概率的递推公式：

$$\begin{cases} \pi(1) = \pi(0)P \\ \pi(2) = \pi(1)P = \pi(0)P^2 \\ \quad\vdots \qquad\quad \vdots \qquad\quad \vdots \\ \pi(k) = \pi(k-1)P = \cdots = \pi(0)P^k \end{cases} \tag{7}$$

(7)式中，$\pi(0) = [\pi_1(0), \pi_2(0), \cdots, \pi_n(0)]$ 为初始状态概率向量。

2.2.1 第 k 个时刻的状态概率预测

由上述分析可知，如果某一事件在第 0 个时刻的初始状态已知(即 π(0) 已知)，则利用(7)式，就可以求得它经过 k 次状态转移后，在第 k 个时刻处于各种可能状态的概率(即 π(k))，从而得到该事件在第 k 个时刻的状态概率预测。

2.2.2 终极状态概率预测

经过无穷多次状态转移后所得到的状态概率称为终极状态概率，或称平衡状态概率。如果记终极状态概率向量为 $\pi = [\pi_1, \pi_2, \cdots, \pi_n]$，则

$$\pi_i = \lim_{k \to \infty} \pi_i(k) (i = 1,2,\cdots,n) \tag{8}$$

即：

$$\pi = \left[\lim_{k \to \infty} \pi_1(k), \cdots \lim_{k \to \infty} \pi_n(k)\right] = \lim_{k \to \infty} \pi(k) \tag{9}$$

按照极限的定义可知：

$$\lim_{k \to \infty} \pi(k) = \lim_{k \to \infty} \pi(k+1) = \pi \tag{10}$$

将(10)式代入(7)式得

$$\lim_{k \to \infty} \pi(k+1) = \lim_{k \to \infty} \pi(k) P$$

$$\text{即：} \pi = \pi P \tag{11}$$

这样，就得到了终极状态概率应满足的条件：

① $\pi = \pi P$

② $0 \leqslant \pi_i \leqslant 1 (i = 1,2,\cdots,n)$

③ $\sum_{i=1}^{n}\pi_i = 1$

②与③是状态概率的要求，其中②表示，在无穷多次状态转移后，事件必处在 n 个状态中的任意一个；①就是用来计算终极状态概率的公式。终极状态概率是用来预测马尔科夫过程在遥远的未来会出现什么趋势的重要信息。

3　一个实例

本节通过一个具体的实例（某地区农业收成情况）来说明马尔科夫模型的应用。

考虑某地区农业收成变化的三个状态，即"丰收"、"平收"和"欠收"。记 E_2 为"丰收" 状态，E_2 为"平收"状态，E_3 为"欠收"状态。

表 1 给出了该地区 1950 – 1989 年期间农业收成的状态变化情况。

表 1　某地区农业收成变化的状态转移情况

年份	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
序号	1	2	3	4	5	6	7	8	9	10
状态	E_1	E_1	E_2	E_3	E_2	E_1	E_3	E_2	E_1	E_2
年份	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
序号	11	12	13	14	15	16	17	18	19	20
状态	E_3	E_1	E_2	E_3	E_1	E_2	E_1	E_3	E_3	E_1
年份	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
序号	21	22	23	24	25	26	27	28	29	30
状态	E_3	E_3	E_2	E_1	E_1	E_3	E_2	E_2	E_1	E_2
年份	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
序号	31	32	33	34	35	36	37	38	39	40
状态	E_1	E_3	E_2	E_1	E_1	E_2	E_2	E_3	E_1	E_2

3.1 计算该地区农业收成变化的状态转移概率矩阵

从表 1 可知，在 15 个从 E_1 出发（转移出去）的状态转移中，有 3 个是从 E_1 转移到 E_1 的（即 1→2，24→25，34→35），有 7 个是从 E_1 转移到 E_2 的（即 2→3，9→10，12→13，15→16，29→30，35→36，39→40），有 5 个是从 E_1 转移到 E_3 的（即 6→7，17→18，20→21，25→26，31→32）。

$$P_{11} = P(E_1 \to E_1) = P(E_1 \mid E_1) = \frac{3}{15} = 0.2000$$

$$P_{12} = P(E_1 \to E_2) = P(E_2 \mid E_1) = \frac{7}{15} = 0.4667$$

按照同样的办法可以得到其他的概率值。所以，该地区农业收成变化的状态转移概率矩阵为

$$P = \begin{bmatrix} 0.2000 & 0.4667 & 0.3333 \\ 0.5385 & 0.1538 & 0.3077 \\ 0.3636 & 0.4545 & 0.1818 \end{bmatrix} \tag{12}$$

如果将 1989 年的农业收成状态记为 π(0) = [0，1，0]（因为 1989 年处于"平收"状态），将（12）式及 π(0) 代入（7）式，就可求得 1990 – 1993 年可能出现的各种状态的概率（见表 2）。

表 2　该地区 1990 – 1993 年农业收成状态概率预测值

年份	1990		
状态概率	E_1	E_2	E_3
	0.5385	0.1528	0.3077
年份	1991		
状态概率	E_1	E_2	E_3
	0.3024	0.4148	0.2837
年份	1992		
状态概率	E_1	E_2	E_3
	0.3867	0.3334	0.2799
年份	1993		
状态概率	E_1	E_2	E_3
	0.3857	0.3589	0.2779

在关于该地区农业收成状态概率的预测中，设终极状态的概率为 $\pi = [\pi_1, \pi_2, \pi_3]$，则

$$[\pi_1, \pi_2, \pi_3] = [\pi_1, \pi_2, \pi_3]\begin{bmatrix} 0.2000 & 0.4667 & 0.3333 \\ 0.5385 & 0.1538 & 0.3077 \\ 0.3636 & 0.4545 & 0.1818 \end{bmatrix}$$

即
$$\begin{cases}\pi_1 = 0.2000\pi_1 + 0.5385\pi_2 + 0.3636\pi_3 \\ \pi_2 = 0.4667\pi_1 + 0.1538\pi_2 + 0.4545\pi_3 \\ \pi_3 = 0.3333\pi_1 + 0.3077\pi_2 + 0.1818\pi_3\end{cases} \quad (13)$$

求解(13)式得：$\pi_1 = 0.3653$，$\pi_2 = 0.3525$，$\pi_3 = 0.2799$。这说明，该地区农业收成的变化，在无穷多次状态转移后，"丰收"和"平收"状态出现的概率都将大于"欠收"状态出现的概率。

3.2 结论

在 GIS 预测中，被预测对象所经历的各个状态和状态之间的转移概率是最为关键的。马尔科夫预测法就是利用状态之间的转移概率矩阵预测事件发生的状态及其发展变化趋势的。马尔科夫预测法的基本要求是状态转移概率矩阵必须具有一定的稳定性。因此，必须具有足够多的统计数据，才能保证预测的精度与准确性。即马尔科夫链模型必须建立在大量的统计数据的基础之上。

4 结束语

本文应用马尔科夫链模型预测基于 GIS 的各种具体的应用系统的未来状态，并通过一个具体的例子来说明。实验结果表明，采用该模型来预测，效果良好。马尔科夫链模型将会在地理统计学方面，尤其是在 GIS 领域中，有着更加广阔的应用前景。

参考文献

1 Jiawei Han，Micheline Kamber. 数据挖掘概念与技术. 范明，孟小峰，等译. 北京：机械工业出版社，2004：301－304.
2 吴艳. 数据挖掘技术在药物配方中的研究及应用. 计算机系统应用，2008，17(3)：73－76.
3 吕志芳，王怀阳，贾吉庆. 基于 MFP－Miner 算法的图书借阅数据关联规则挖掘. 计算机系统应用，2008，17(2)：90－93.
4 黄浩，王建军. WEB 使用挖掘研究. 计算机系统应用，2008，17(1)：125－128，124.
5 林文龙，刘业政，姜元春. Web 浏览预测的 Markov 模型综述. 计算机科学，2008，35(1)：9－14.
6 SARUKKAI R R. Link Prediction and Path Analysis Using Markov Chains. Proc of the 9 th International World Wide Web Conference，2000，33 (1－6)：377－386.
7 杜修平. 基于数据挖掘的证券态势估计系统［博士学位论文］. 天津：天津大学，2006.
8 沈永梅. 基于统计试验的马氏链点值预测方法和时间序列分析预测方法的比较分析［硕士学位论文］. 南京：河海大学，2006.
9 孙才志，林学钰. 降水预测的模糊权马尔可夫模型及应用. 系统工程学报，2003，18(4)：294－299.
10 王实，高文，黄铁军，马继勇，李锦涛. 基于隐马尔可夫模型的在线零售站点的自适应. 软件学报，2001，12(4)：599－606.