马氏链模型

Abstract: 应用马氏链模型可以处理一些随机影响的动态系统，该动态系统具有明显的特点：下个时期的状态只取决与这个时期的状态和转移概率。

Keywords: 马氏链的性质、如何应用及与实际情况的解释

1 Introduction

在考察随机因素影响的动态系统时，系统在每个时期所处的状态时随机的，从这个时期到下一个时期的状态会按照一定的概率进行转移。并且下一个时期的状态取决于这一时期的概率和转移概率，与以前的各时期的状态无关。这种性质称为无后效性或者马尔科夫性。通俗得来讲，已知现在、将来与历史无关。具有马氏性的，时间、状态均为离散的随机转移过程通常可以用马氏链模型。

在马氏链模型中，有两类非常重要的类型，正则链和吸收链。下面我们通过一个具体的实例来说明一下相关的概念：

某个商店每月考察一次经营情况，其结果用销售好和销售坏的两种状况中的一种表示。已知如果本月销售好，下月仍保持这种状况的概率为0.5,如果本月销售坏，下月转变为销售好的概率为0.4，试分析假若开始时商店处于销售好的状况，那么经过若干月后能保持销售好的概率是多少？如果开始时商店处于销售坏的情况呢？对该问题进行分析，商店的经营状况是随机的，每月会随机的转变。但是，也可以看出，下个月的销售和这个月的销售情况和转变概率有关。在初始状态已知的情况下，我们希望通过一个递推公式，推出将来销售好和坏的概率？

用随机变量X,表示第n个月的经营状况，$X_n = 1$表示销路好，$X_n = 2$表示销路坏。用$a_i(n)$表示第n个月处于状态i 的概率$a_i(n) = P(X_n = i)$，称之为状态概率, $p_{ij} = P(X_{n+1} = j|X_n = i)$表示本月处于状态$i$，下月转为状态$j$ 的概率，称之为转移概率。这里X_{i+1}只取决与X_n和p_{ij},而与前面的没有关系。

这样，就可以得到全概率公式：

$$\begin{cases} a_1(n+1) = a_1(n)p_{11} + a_2(n)p_{21} \\ a_2(n+1) = a_1(n)p_{12} + a_2(n)p_{22} \end{cases} \tag{1.1}$$

即

$$\begin{pmatrix} a_1(n+1) \\ a_2(n+1) \end{pmatrix} = \begin{pmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{pmatrix} \begin{pmatrix} a_1(n) \\ a_2(n) \end{pmatrix} = P \begin{pmatrix} a_1(n) \\ a_2(n) \end{pmatrix} \tag{1.2}$$

这里$p_{11} = 0.5, p_{21} = 0.4, p_{12} = 1 - p_{11} = 0.5, p_{22} = 1 - p_{21} = 0.6$. 若开始时销路好,则$a_1(0) = 1, a_2(0) = 0$,用上述公式可以计算出$a_1(n), a_2(n), n = 1, 2, ...$ 随着n值的增大，$a_1(n) = 4/9, a_2(n) = 5/9$. 而如果开始销路坏时，同样的处理方法在$n \to \infty$可以得到同样的结果。这也说明随着$n \to \infty$ 时，状态概率趋向于稳定值，而与初始状态无关。

还有另外一种情况，可以直接给出所对应的转移矩阵和初始的状态概率，其表达式为：

$$\begin{pmatrix} a_1(n+1) \\ a_2(n+1) \\ a_3(n+1) \end{pmatrix}' = \begin{pmatrix} a_1(n) \\ a_2(n) \\ a_3(n) \end{pmatrix}' \begin{pmatrix} p_{11} & p_{21} & p_{31} \\ p_{12} & p_{22} & p_{32} \\ p_{13} & p_{23} & p_{33} \end{pmatrix} = \begin{pmatrix} a_1(n) \\ a_2(n) \\ a_3(n) \end{pmatrix}' \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0 & 0 & 1 \end{pmatrix} \tag{1.3}$$

假设初始状态的概率为$(a_1(0), a_2(0), a_3(0)) = (0.5, 0.3, 0.2)$, 当$n \to \infty$时，可以得到$(a_1(n), a_2(n), a_3(n)) = (0, 0, 1)$. 而且不管初始状态如何，最后的结果都是一样的。但是如果选择初始状态为$(0, 0, 1)$,则对于任意的n, 都有$a_1(n) = a_2(n) = 0, a_3(n) = 1$. 这说明对于这种转移矩阵，并非所有的状态都可以通过有限次转移达到另外的任意状态.也就是说，一旦进入状态3，就永远不会转移到其他状态。

在这里涉及到两个概念：转移概率和状态概率。如果X_{n+1}的取值只取决于X_n取值及转移概率，而与$X_{n-1}, X_{n-2}, ...$的取值无关，这种离散状态按照离散时间的随机过程称为马氏链。而且根据状态转移的无后效性和全概率公式可以写出马氏链的基本方程为：

$$a_i(n+1) = \sum_{i=1}^{k} a_j(n)p_{ji}, i = 1, 2..., k \tag{1.4}$$

而且$a_i(n)$和p_{ij}应满足

$$\sum_{i=1}^{k} a_i(n) = 1, n = 0, 1, 2, ..., \sum_{j=1}^{k} p_{ij} = 1, i = 1, 2, \cdots, k, p_{ij} \geq 0 \tag{1.5}$$

这样一个对于马氏链的基本方程可以表示为：

$$a(n+1) = a(n)P = a(0)P^n \tag{1.6}$$

这里P为随机矩阵，其行和为1.

这两个马氏链有较大的区别，它们属于马氏链的两个重要类型：正则链和吸收链。

从任意状态出发经过有限次转移都能达到另外的任意状态：

一个有k个状态的马氏链如果**存在**正整数N,使从任意状态i经过N次转移都以大于零的概率到达状态j,则称之为正则链。应用下面的定理很容易证明是否是正则链：

定理：若马氏链的转移矩阵为P，则它是正则链的充要条件是存在正整数N,使得$P^N > 0$(指矩阵的每个元素都大于零).

定理：正则链存在唯一的极限状态概率$w = (w_1, w_2, ..., w_k)$,当$n \to \infty$时，状态概率$a(n) \to w$,w与初始状态概率$a(0)$ 无关，称之为稳态概率。并且满足

$$wP = w, \sum_{i=1}^{k} w_k = 1 \tag{1.7}$$

而从状态i 出发经过n次转移，第一次到达状态j的概率称为i到j的首达概率，记作：$f_{ij}(n)$ 。从我们刚才所讲的例题也可以看出，从一个状态第一次到达另外一个状态的概率会发生变化。定义$\sum_{i=1}^{\infty} nf_{ij}(n)$ 为状态i 出发第一次到达状态j的平均转移次数。

如果转移概率$p_{ii} = 1$ 的状态i 称为吸收概率。如果马氏链中至少包含一个吸收状态，并且从每一个非吸收状态出发，能以正的概率有限次转移到达某个吸收状态，称这个马氏链为吸收链。若已经进入吸收状态，则不会再逃离这一状态。

那么这是如何来计算非吸收状态到达吸收状态的次数？

假设在转移矩阵中有r 个吸收状态，$k-r$个非吸收状态，对矩阵进行交换位置，得到标准形式。

$$\begin{pmatrix} I_{rr} & 0 \\ R & Q \end{pmatrix} \tag{1.8}$$

而且若要满足从任一非吸收状态出发经过有限次转移可到达某吸收状态的条件，对矩阵Q就有一定的要求，这里Q不能是随机矩阵，它至少存在一个小于1的行和，且满足如下定理：

定理3：$M = (I-Q)^{-1} = \sum_{s=0}^{\infty} Q^s, e = (1, 1, ..., 1), y = Me$ 的第i个分量是从第i个非吸收状态出发，被某个吸收状态吸收的平均转移次数。而f_{ij} 为从非吸收状态i出发终将被吸收状态j吸收的概率。记$F = \{f_{ij}\}_{(k-r)\times r} = MR.$

下面我们通过一个具体的实例来分析马氏链在完全优势基因遗传理论中的应用。基因分为优势基因和劣势基因两种，分别用d和r表示，每种外部表现特征都可以由体内的两个基因决定，而每个基因是d或r中的一个。这样就有三种基因类型dd, dr, rr，分别称为优种、混种和劣种。用D, H, R 表示。含有优种D和混种H基因类型的个体，外表呈现出优势的特征。生物繁殖时，一个后代随机地继承父亲和母亲两个基因中的一个，形成它的两个基因。一般情况下，哪一个基因遗传下去是等概率的。父母基因类型的组合有全是优种DD,全是劣

种RR，全是混种HH,一优种一混种DH,一优种一劣种DR,一混种一劣种HR,简单计算就可以得到每种组其后代各种基因类型的概率。

Table 1: 父母基因确定后代各种基因类型的概率

父母基因类型的组合		DD	RR	DH	DR	HH	HR
后代各种	D	1	0	1/2	0	1/4	0
基因类型	H	0	0	1/2	1	1/2	1/2
的概率	R	0	1	0	0	1/4	1/2

下面我们就以马氏链作为工具讨论两个具体的基因遗传模型。所对应的一种是随机交配模型，一种是近亲繁殖模型。所谓的随机交配模型。是指一个群体中雄性和雌性的比例相等，并且两者具有相同的基因类型分布，假设初始一代的群体中，三种基因类型的数量比例是$D:H:R=a:2b:c$，满足$a+2b+c=1$,并且记$p=a+b,q=b+c$。则在整个群体中优势基因d和劣势基因r的数量之比为$d:r=p:q,p+q=1$。

设$X_n=1,2,3$ 分别表示第n代的一个体属于D,H,R 基因状态。$a_i(n)$ 表示个体属于第i种基因类型的概率。转移概率p_{ij}可以这样来解释，当母亲的基因类型i确定，后代的类型的基因类型为j的概率. 而其在母亲基因类型已知的情况下，就由父亲的基因类型确定。可以直接考查从雄性个体中$p:q$的比例获得优势基因和劣势基因。

我们分三种情况进行讨论，也就是母亲的基因类型分别为D 的时候，后代的基因类型为D 的概率。(根据上表进行计算)

若母亲的基因类型为D ，为使后代是D则只需从雄性群体中以概率p获得d. 则$p_{11}=p$

若母亲的基因类型为D ，则后代的基因类型为H的概率为q,则$p_{12}=q$

若母亲的基因类型为D ，则后代的基因类型为R的概率为0,则$p_{13}=0$

若母亲的基因为H,后代为D的概率为p_{21}:后代要以1/2的概率从母体获得d,同时以p的概率从雄性群体中获得d, 则$p_{21}=p/2$.

若母体味H,则后代为H 的概率为$\frac{p}{2}+\frac{q}{2}=\frac{1}{2}$.即后代需以1/2的概率从母体获得$d$,同时需要以$q$的概率获得$r$,同样也可以若1/2的概率从母体获得$r$,同时可以从$p$的概率获得$d$.

类似可以计算$p_{23}=\frac{q}{2}$.同样的方法可以算出p_{31},p_{32},p_{33},得到转移矩阵：

$$P=\begin{pmatrix} p & q & 0 \\ p/2 & 1/2 & q/2 \\ 0 & p & q \end{pmatrix} \tag{1.9}$$

根据初始的基因类型的分布状态$a(0)=(a,2b,c)$,则可以得到如下的表达式：

$$\begin{aligned} a(1)&=a(0)P=(p^2,2pq,q^2) \\ a(2)&=a(1)P=(p^2,2pq,q^2) \end{aligned} \tag{1.10}$$

这说明这个分布将保持下去，这个结果称为Hardy-Weinberg平稳定律。可以计算在随机交配下，三种基因类型的首次返回平均转移次数,即平均经过多少代每种基因类型首次回到原来的类型

$$\mu_{11}=1/p^2, \mu_{22}=1/2pq, \mu_{33}=1/q^2 \tag{1.11}$$

若$p=q=1/2$，则D,H,R的平均换代数目分别为$4,2,4$代。

与前面讨论的不同，前面的讨论的是随机模型讨论的是后代群体中的基因类型，这里是从同一对父母的后代中随机地选取一雄一雌进行交配，产生后代，持续下去，而这里分析的是后代配对中基因类型的变化。

近亲繁殖：近亲繁殖是指这样的一种繁殖方式，从同一对父母的大量后代中，随机的选取雌雄进行配对，产生后代。我们来分析这样的基因演变。与前面不同的是，前面的这样状态应该取的是配对的基因类型组合。这样共有六种情况$X_n=1,2,3,4,5,6$,分别定义为$DD,RR,DH,DR,HH,HR.$

实际上需要考虑的是当父母的基因类型为上述六种状态之一的时候，子女自由配对组合对应的这六种基因类型的概率。

父母基因类型的组合DD,则子女的状态为DD,RR,DH,DR,HH,HR的概率为$(p_{11},p_{12},p_{13},p_{14},p_{15},p_{16})=(1,0,0,0,0,0)$

父母基因类型的组合RR,则子女的状态为DD,RR,DH,DR,HH,HR的概率为$(p_{21},p_{22},p_{23},p_{24},p_{25},p_{26})=(0,1,0,0,0,0)$

父母基因类型的组合DH,则子女的状态为DD,RR,DH,DR,HH,HR的概率为$(p_{31},p_{32},p_{33},p_{34},p_{35},p_{36})=(1/4,0,1/2,0,1/4,0)$

因为DH的后代中为D,H的概率分别为$1/2$,则其随机选取的配对为DD的概率是$1/4$，同样其随机选取的配对为HH的概率是$1/4$. 后代为DH的情况分为两种:雄性为D雌性为H和雄性为H雌性为D的概率为$\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{2}$

父母基因类型的组合HH,则后代各种基因类型的概率为D,H,R分别为$1/4,1/2,1/4$. 则子女的状态为DD,RR,DH,DR,HH,HR的概率为$(p_{51},p_{52},p_{53},p_{54},p_{55},p_{56})=(1/16,1/16,1/4,1/8,1/4,1/4)$

父母基因类型的组合HR,则后代各种基因类型的概率为D,H,R分别为$0,1/2,1/2$.则子女的状态为DD,RR,DH,DR,HH,HR 的概率为$(p_{61},p_{62},p_{63},p_{64},p_{65},p_{66})=(0,1/4,0,0,1/4,1/2)$ 则对应的转移矩阵为

$$P=\begin{pmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\1/4&0&1/2&0&1/4&0\\0&0&0&0&1&0\\1/16&1/16&1/4&1/8&1/4&1/4\\0&1/4&0&0&1/4&1/2\end{pmatrix}\tag{1.12}$$

这个矩阵里面包含有DD,RR两种吸收状态，这是一个吸收链，说明不论最初选取的配对是哪种基因类型组合，经过若干代近亲繁殖，最终都会变成这两种状态。那么从任一个非吸收状态出发，平均经过多少代才能被吸收状态吸收。首先把原问题化成标准型，得到对应的Q,R

$$Q=\begin{pmatrix}1/2&0&1/4&0\\0&0&1&0\\1/4&1/8&1/4&1/4\\0&0&1/4&1/2\end{pmatrix},R=\begin{pmatrix}1/4&0\\0&0\\1/16&1/16\\0&1/4\end{pmatrix}\tag{1.13}$$

计算$M=(I-Q)^{-1},y=Me=(4\frac{5}{6},6\frac{2}{3},5\frac{2}{3},4\frac{5}{6})$. 对上式进行解释，从近亲$DH$配对的情况出发，经过$4\frac{5}{6}$ 代就会被状态1，2吸收，全部变为优种和劣种。