基于一阶马尔可夫链的实验数据序列分类模型①

黄志成

(广东女子职业技术学院 信息资源中心, 广州 511450)

摘 要: 为了对在线实验系统产生的实验数据序列进行分析, 引入一阶马尔可夫链. 通过人工分类把实验数据分为学习积极和懒散作弊两类, 分别构建马尔可夫链模型. 根据输出概率判定测试数据来自哪一个模型的可能性较大. 最后讨论了状态的平稳分布情况. 实验结果表明, 基于马尔可夫链的分类模型具有较高的正确率.

关键词: 一阶马尔可夫链; 序列分类; 序列数据; 在线实验

Sequence Classifying Model of Experimental Data Based on First Order Markov Chain

HUANG Zhi-Cheng

(Information Resource Center, Guangdong Women's Polytechnic College, Guangzhou 511450, China)

Abstract: In order to analysis the sequence data generated by online experimental system, the first-order Markov chain is introduced. It artificially classifies the experimental data into the learning initiative and fraud, and constructs two Markov chain models. It determines by the larger possibility from which model the test data comes. At the end, it discusses the steady state distribution. Experimental results show that the model based on Markov chain has higher classification accuracy.

Key words: first-order Markov chain; sequence classification; sequence data; online experiment

在学习系统中, 学习者的学习过程随着时间的推移而发生改变, 产生一系列的数据. 不同的学习者产生不同的序列数据. 受经验、知识、人为等因素的交互影响, 学习者的思维活动复杂多变, 具有一定的随机性, 是一个复杂的非线性动力系统. 为了研究并区分不同的学习者的学习过程, 需要对学习序列数据进行分析. 传统教师人工监控学生学习过程存在着人力、时间、场地、效率的限制. 马尔可夫链法是以概率论和随机过程理论为基础, 利用随机数学模型来分析对象发展变化过程中数量关系的一种统计分析方法. 马尔可夫链理论在网页点击流分析[1]、学习效率预测[2]、计算机情感计算[3]、教学质量评价[4]等领域得到广泛应用, 但也存在应用不足, 如文献[5]使用了不稳定的转移概率构建马尔夫链来进行教学评价. 目前使用马尔可夫链进行学习过程监控的相关研究较少.

本文使用马尔可夫链, 对大量实验数据进行计算, 得出较为稳定的状态转移概率, 建立对学习序列数据进行分类的模型, 并研究了模型长时间转移后, 状态过程的平稳分布情况, 增强教师对学生学习过程的监控.

1 一阶马尔可夫链模型[6]

1.1 相关定义及定理

定义 1 假设有随机序列 $\{X_n, n \ge 0\}$, 若对任意的 $i_0, i_1, ..., i_n, ... \in S$, S 为有限状态空间, $P\{X_0 = i_0, X_1 = i_1, X_n = i_n\} > 0$, 且

$$P\{X_0 = i_0, X_1 = i_1, ..., X_n = i_n\} = P\{X_{n+1} = i_{n+1} \mid X_n = i_n\} \quad (1)$$

则称 $\{X_n, n \ge 0\}$ 为一阶马尔可夫链, 简称为马尔可夫链.

马尔可夫链表示一个随机序列的条件概率只与前一个时刻的系统状态有关, 而与以前的系统状态无关.

① 基金项目:广东省教育科学规划信息技术专项(12JXN036)

收稿时间:2013-09-22;收到修改稿时间:2013-10-18

马尔可夫的这种特性也称为无后效性.

定义 2　$\forall i,j\in S$，称 $P\{X_{n+1}=j\mid X_n=i\}=p_{ij}^{n,n+1}$ 表示如果在 n 时刻状态为 i，在 n+1 时刻单步转移到状态 j 的条件概率. 若对 $\forall i,j\in S$，有 $p_{ij}^{n,n+1}=p_{ij}$，即 p_{ij} 与 n 无关，则称 $\{X_n,n\ge 0\}$ 为齐次马尔可夫链. 记 $P=(p_{ij})$，称为 $\{X_n,n\ge 0\}$ 的单步转移概率矩阵. P 的阶等于状态的个数，且有

$$p_{ij}\ge 0,i,j\in S,\sum_{j\in S}p_{ij}=1,i\in S \tag{2}$$

条件概率 $P\{X_{n+m}=j\mid X_m=i\}$ 称为马尔可夫链 $\{X_n,n\ge 0\}$ 时刻 m 的 n 步转移概率，即状态 i 在 n 次转移后处于状态 j 的概率. 对于齐次马尔可夫链，与 m 无关，记为 $p_{ij}^{(n)}$.

下面用 λ 表示马尔可夫链模型. 给定长度为 l 的序列 $O=o_1,o_2,...,o_l$，模型输出 O 的概率可表示为

$$\begin{aligned}P(O\mid\lambda)&=P(o_l\mid o_{l-1})P(o_{l-1}\mid o_{l-2})...P(o_2\mid o_1)P(o_1)\\&=P(o_1)\prod_{i=2}^{l}P(o_i\mid o_{i-1})\\&=P(o_1)\prod_{i=2}^{l}p_{i-1,i}\end{aligned} \tag{3}$$

一个 3 状态的马尔可夫链模型如图 1 所示.

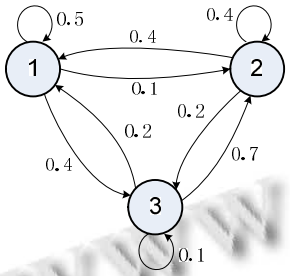

图 1　一个 3 状态的马尔可夫链模型

定理 2　对于一个不可约的可遍历的马尔可夫链，$\lim_{n\to\infty}p_{ij}^{(n)}$ 存在且与 i 独立. 令

$$\pi_j=\lim_{n\to\infty}p_{ij}^{(n)},\quad j\ge 0$$

那么

$$\pi_j=\sum_{i=0}^{\infty}\pi_i p_{ij},\quad j\ge 0,\quad \sum_{j=0}^{\infty}\pi_j=1 \tag{4}$$

的唯一非负解. 称 π_j 为马尔可夫链处在状态 j 的平稳分布，也称为极限概率.

1.2 马尔可夫过程检验

使用马尔可夫链建立模型，需要检验系统 $\{X_t,t\in T\}$ 是否具有马尔可夫性. 马尔可夫性的稳定与否也影响着系统的准确度. 通常采用 χ^2 统计量检验.

记 $(c_{ij})_{n\times n}$ 为状态转移次数的矩阵，令 $\tilde{p}_{\cdot j}$ 表示 $(c_{ij})_{n\times n}$ 第 j 列的和与各行和的比值，p_{ij} 为状态 i 转移到状态 j 的概率. 有

$$\begin{aligned}\tilde{p}_{\cdot j}&=\sum_{i=1}^{n}c_{ij}/\sum_{i=1}^{n}\sum_{k=1}^{n}c_{ik}\\p_{ij}&=c_{ij}/\sum_{k}^{n}c_{ik}\end{aligned} \tag{6}$$

当 n 较大时，统计量服从自由度为 $(n-1)^2$ 的 χ^2 分布. 选择置信度 α，查表得 $\chi_\alpha^2((n-1))^2$. 令

$$\hat{\chi}^2=2\sum_{i=1}^{n}\sum_{j=1}^{n}c_{ij}\,|\log(p_{ij}/\tilde{p}_{\cdot j})| \tag{7}$$

如果 $\hat{\chi}^2>\chi_\alpha^2((n-1))^2$，就认为系统 $\{X_t,t\in T\}$ 具有马尔可夫性.

2　构造马尔可夫链模型

2.1 实验数据介绍

本文以作者开发的研究项目“SQL 在线评测实验系统”为例. 学生登录系统，在实验题库中选择练习题，按照要求编写 SQL 语句，并在线提交. 系统对 SQL 语句进行智能评估，反馈 6 种结果：正确、SQL 错误、记录数错误、字段数错误、数值错误和字段名错误. 可把系统反馈的结果看作是系统状态. 学生实验过程中，由于对知识点处于未掌握状态，思维处于一种无序、混乱、反复的状态，实验反馈结果具有一定的不确性. 也就是说，除去一些偶然的人为事件(如蓄意行为、恶作剧等)，系统状态之间的转移具有一定的随机性. 系统状态的转移概率是建立在对大量学生的实验过程数据统计之上，具有较为全面的稳定性. 学生个别实验数据的偶然偏差对系统状态的转移概率的影响可以忽略. 总的来说，应用本实验数据建立的马尔可夫链模型具有较为稳定的转移概率.

2.2 计算模型参数

为了区分学习积极与学习懒散、作弊学生，通过人工方式分别提取典型学生的学习积极实验数据集

(记为 Set^+)和学习懒散、作弊实验数据集(记为 Set^-),剩下部分数据留作测试集(记为 Set^*).

给定一个训练序列 $O=o_1,o_2,...,o_l$,马尔可夫链的转移概率可通过下式进行计算[7]

$$p_{ij}=c_{ij}/\sum_{h=1}^{n}c_{ih} \tag{8}$$

其中,c_{ij} 是 O 中符号 j 跟随符号 i 的次数. 对于 K 个训练序列 $O^{(k)}=o_1^{(k)},o_2^{(k)},...,o_l^{(k)}$ 的情况,式(3)可引申为

$$p_{ij}=\sum_{k=1}^{K}c_{ij}^{(k)}/\sum_{k=1}^{K}\sum_{h=1}^{n}c_{ih}^{(k)} \tag{9}$$

分别使用 Set^+ 和 Set^- 数据集生成马尔可夫链模型 λ^+ 和 λ^-. 为了确定给定的序列 O 来自于哪一个模型,定义模型输出概率比的对数 ϑ

$$\begin{aligned}\vartheta&=\log\frac{P(O|\lambda^+)}{P(O|\lambda^-)}=\log\frac{P(o_1)\prod_{i=2}^{n}p_{i-1,i}^{+}}{P(o_1)\prod_{i=2}^{n}p_{i-1,i}^{-}}\\&=\sum_{i=2}^{n}\log\frac{p_{i-1,i}^{+}}{p_{i-1,i}^{-}}\end{aligned} \tag{10}$$

记 ε 为一定的阀值,如果 $\vartheta>\varepsilon$,我们认为 O 来自模型 λ^+;如果 $\vartheta<-\varepsilon$,认为 O 来自模型 λ^-;如果 $|\vartheta|\le\varepsilon$,说明 O 和两个模型都较接近,很难准确区分来自哪一个模型.

3 序列数据分类

模型 λ^+ 和 λ^- 计算得出的单次转移概率 P^+ 和 P^- 为

$$P^+=\begin{bmatrix}0.28&0.24&0.26&0.03&0.11&0.08\\0.21&0.53&0.12&0.01&0.06&0.07\\0.29&0.16&0.39&0.03&0.06&0.06\\0.15&0.12&0.08&0.42&0.12&0.12\\0.39&0.06&0.08&0.01&0.44&0.02\\0.27&0.16&0.11&0.01&0.05&0.39\end{bmatrix}$$

$$P^-=\begin{bmatrix}0.68&0.06&0.12&0.03&0.06&0.06\\0.30&0.35&0.05&0.08&0.20&0.03\\0.44&0.09&0.33&0.09&0.03&0.02\\0.44&0.08&0.28&0.08&0.04&0.08\\0.25&0.21&0.07&0.21&0.11&0.14\\0.36&0.05&0.32&0.09&0.05&0.14\end{bmatrix}$$

取 α =0.05,根据式(7)分别对模型 λ^+ 和 λ^- 计算 $\hat{\chi}^2$,结果如表 1 所示.

表 1 两模型的统计量值

统计量	模型 λ^+	模型 λ^-
$\hat{\chi}^2$	759.659	331.502
$\chi_\alpha^2((n-1))^2$	16.919	16.919

显然,两模型的 $\hat{\chi}^2$ 都大于 $\chi_\alpha^2((n-1))^2$,都是马尔可夫过程.

对数据集 Set^* 的序列数据应用式(10)计算 ϑ 的值,部分结果如表 2 所示.

表 2 数据集 Set^* 的部分结果

序列	序列长度	ϑ 值
1	147	-3.97
2	94	0.94
3	47	-21.52
4	44	-20.38
5	58	18.84
6	37	20.51
7	68	11.07
8	71	25.89
9	89	6.84
10	52	44.97

经过试验,取 ε 为 1~2.5 较为合适. 本模型对 10 个序列的分类结果如图 2 所示,人工分类结果如图 3 所示.

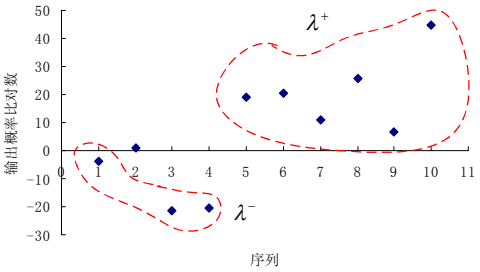

图 2 模型自动分类结果

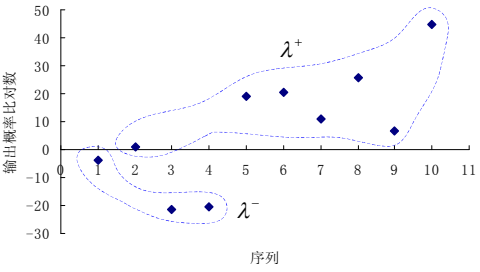

图 3 人工分类结果

按本文方法计算,序列 2 由于 $\vartheta<\varepsilon$,说明和模型 λ^+、λ^- 都非常接近. 但实际情况是,序列 2 大部分数据和 Set^+ 相似,极小部分来自 Set^- 的数据具有较高的

状态转移概率. 序列 2 的学生学习较为认真, 都是通过自身练习得出正确的实验结果. 人工分类时把序列2归入模型 λ^+. 由图2和图3可知, 除了个别数据有偏差外, 模型自动分类结果基本上和人工分类结果一致.

4 状态平稳分布分析

学生解决一道实验题平均要花费 20-40 分钟, 如果连续解决多道实验题的话, 将耗费数个小时, 甚至更长. 我们关心长时间连续实验的情况下, 6 个状态的平稳分布情况. 考虑模型 λ^+, 由 P^+ 的数据可知, 模型 λ^+ 是一个有限状态的马尔可夫链, 所有的状态都相通且可达, 并且可遍历, 因此是不可约链, 且从任一状态出发经一步就可返回自身状态, 所以是非周期的. 由定理 2 知有限状态的不可约非周期马尔可夫链存在唯一的平稳分布, 即极限分布. 根据式(4), 平稳分布 π_j 可由下面的线性方程组得到

$$\begin{cases} \sum_{j=1}^{6}\pi_j = 1 \\ \sum_{i=1}^{6}\pi_j p_{ij} = \pi_j \end{cases} \quad i, j = 1,2,3,4,5,6$$

解得 π_1 =0.27, π_2 =0.27, π_3 =0.21, π_4 =0.03, π_5 =0.12, π_6 =0.10. 由结果可知, 在长时间学习中, 学生实验过程中状态会趋于平稳分布, 且最有可能出现的是状态 1 和状态 2. 也就是说, 学生学验过程中编制的SQL语句普遍存在错误, 学生对SQL的语法还未熟练运用. 虽然如此, 但经过长时间实验练习, 最终还是得出正确结果.

5 结论

本文对马尔可夫链进行了深入研究, 使用大量实验数据构建稳定的状态转移概率, 构建模型对学习系统产生的学习序列数据进行分类, 并讨论了长时间学习下的系统状态的平稳分布情况. 实验结果表明, 使用马尔可夫链对实验序列进行分类, 相比较人工分类, 前者更智能, 具有更高的效率, 而且分类精度可与人工媲美. 同时, 本文研究也存在一些不足, 通过人工方式对数据进行预筛选存在局限性, 尤其是大数据量的时候. 这时可以通过聚类等方式对数据进行预分类[8].

马尔可夫链法作为一种定量分析方法, 应用在学生实验过程中存在着一定的局限性, 与其它应用场合有所不同. 例如, 由实验过程反馈数据构成的系统是否具有马尔可夫性, 状态转移概率是否稳定, 这是使用马尔可夫链的前提. 除此之外, 学生实验过程反馈结果的准确性、实验题的难度、学生是否独立实验、实验过程中偶然事件的出现等, 这些因素都会直接或间接影响系统状态的转移概率. 有些文献仅使用前后两次测验成绩构建马尔可夫链概率转移矩阵进行教学评价, 状态的转移未必真实反映学生水平的提高或降低. 因此, 实际工作中, 可以配合其它方法, 处理实验数据的收集、清洗、噪声消除等, 尽可能降低对模型的影响.

参考文献

1 郑明秀,杨明根.一阶马尔可夫链在点击流分析中的应用.西南民族大学学报(自然科学版),2007,33(1):174–177.

2 邹杨,许安见.基于马尔可夫链的一类学习效率预测模型.重庆教育学院学报,2012,25(6):8–11.

3 滕少冬,王志良,王莉等.基于马尔可夫链的情感计算建模方法.计算机工程,2005,31(5):17–19.

4 沈晋会.马尔可夫链法在教学质量评估中的应用.内蒙古师范大学学报(教育科学版),2013,26(6):10–13.

5 张雅清.马尔可夫链在教学质量评价中的应用.湖北广播电视大学学报,2010,30(3):128–129.

6 卢瑞瑞.基于 K-means 聚类的马尔可夫过程在股价趋势预测中的应用[学位论文].武汉:华中科技大学,2009:14–21.

7 Han JW, Kamber M. 范明,孟小峰,译.数据挖掘:概念与技术(第二版).北京:机械工业出版社,2007:339– 345.

8 刘应东.有约束的半监督聚类方法.计算机工程与应用,2009,45(2):100–102.