

# 一种改进的 Markov 预测模型方法

张友志, 胡国胜, 程玉胜

(安庆师范学院 计算机与信息学院, 安徽 安庆 246133)

**摘要:** 马尔可夫(Markov)模型的链式结构简便易行, 适合作为一个预测模型来预测用户的页面访问模式。针对 Markov 原始预测模型算法时间和空间高开销的缺点, 引入聚类方法对模型进行改进, 以有效降低原始 Markov 预测模型计算开销。改进的 Markov 模型虽较好地克服了原始 Markov 模型的缺陷, 但在提高效率的同时, 模型的预测准确度有所降低。不过由于是将高阶 Markov 模型类别预测和低阶 Markov 模型页面预测相结合, 和原始低阶 Markov 模型页面预测相比, 准确性具有一定优势。

**关键词:** Markov 模型; 访问模式; 聚类方法

**中图分类号:** TP182

**文献标识码:** A

**文章编号:** 1673-629X(2008)12-0078-03

## An Improved Method of Markov Forecasting Model

ZHANG You-zhi, HU Guo-sheng, CHENG Yu-sheng

(Department of Computer and Information, Anqing Teachers' College, Anqing 246133, China)

**Abstract:** The chain structure of Markov model is convenient and easy for forecasting user's browsing patterns. For the original Markov forecasting model's disadvantage of high spending in time and space, improve on Markov model through introducing clustering approach in order to reducing calculational spending of original Markov forecasting model effectively. Improved Markov model overcome the shortcoming of original Markov preferably, but reduce the forecasting veracity at the same time of improving efficiency. Owing to uniting the high-level Markov model sort forecasting and low-level Markov model page layout forecasting, the improved Markov has more advantage than original low-level Markov page layout forecasting in the veracity.

**Key words:** Markov model; browsing pattern; clustering approach

## 0 引言

在对马尔可夫<sup>[1]</sup>(Markov)模型的实际研究与应用中, 人们发现原始的尤其是高阶的 Markov 预测模型空间时间计算开销非常的大, 影响了系统的性能<sup>[2,3]</sup>。文中就此缺陷引入聚类这一数据挖掘中重要的研究分支方法, 以有效降低 Markov 预测模型的时间和空间开销。

## 1 聚类方法

聚类(clustering)是一个将资料集划分为若干组(class)或类(cluster)的过程, 并使得同一个组内的数据对象具有较高的相似度; 而不同组中的数据对象是不相似的<sup>[4]</sup>。相似或不相似的描述是基于数据描述属性的取值来确定的, 通常就是利用各对象间距离来进行

表示的。通常聚类算法可以分为以下几类:

### (1)划分的方法。

给定一个包含  $n$  个对象或数据行, 划分方法将数据集划分为  $k$  个子集(划分)。其中每个子集均代表一个聚类( $k < n$ )。也就是说, 将数据分为  $k$  组, 这些组应满足以下要求: 一是每组至少应包含一个对象; 二是每个对象必须只能属于某一组。需要注意的是后一个要求在一些模糊划分方法中可以放宽。给定需要划分的个数  $k$ , 一个划分方法创建一个初始划分; 然后利用循环再定位技术, 即通过移动不同划分(组)中的对象来改变划分内容。一个好的划分衡量标准通常就是同一个组中的对象“相近”或彼此相关; 而不同组中的对象“较远”或彼此不同。当然还有许多其它判断划分质量的衡量标准。

为获得基于划分聚类分析的全局最优结果就需要穷举所有可能的对象划分。为此大多数应用采用一至二种常用启发方法: 一种是  $k$ -means 算法, 该算法中的每一个聚类均用相应聚类中对象的均值来表示; 一种是  $k$ -medoids 算法, 该算法中的每一个聚类均用相

收稿日期: 2008-03-21

基金项目: 安徽省自然科学基金(070412061)

作者简介: 张友志(1977-), 男, 安徽桐城人, 讲师, 研究方向为数据挖掘等。

应聚类中离聚类中心最近的对象来表示。这些启发聚类方法在分析中小规模数据集以发现圆形或球状聚类时工作的很好。但为了使划分算法能够分析处理大规模数据集或复杂数据类型,就需要对其进行扩展。

(2)层次的方法。

层次方法就是通过分解所给定的数据对象集来创建一个层次。根据层次分解形成的方式,可以将层次方法分为自下而上和自上而下两种类型。自下而上的层次方法从每个对象均为一个单独的组开始,逐步将这些(对象)组进行合并,直到组合并在层次顶端或满足终止条件为止;自上而下层次方法从所有均属于一个组开始,每一次循环将其(组)分解为更小的组:直到每个对象构成一组或满足终止条件为止。

层次方法存在的缺陷就是在进行(组)分解或合并之后,无法回溯。这一特点也是有用的,因为在进行分解或合并时无需考虑不同选择所造成的组合爆炸问题。但这一特点也使得这种方法无法纠正自己的错误决策。将循环再定位与层次方法结合起来使用常常是有效的,即首先利用自下而上层次方法;然后再利用循环再定位技术对结果进行调整。一些具有可扩展性的聚类算法,如:BIRCH 和 CURE,就是基于这种组合方法设计的。

(3)基于密度的方法。

大多数划分方法是基于对象间距离进行聚类的,这类方法仅能发现圆形或球状的聚类而较难发现具有任意形状的聚类。而基于密度概念的聚类方法实际上就是不断增长所获得的聚类直到“邻近”(数据对象或点)密度超过一定阈值(如:一个聚类中的点数,或一个给定半径内必须包含至少的点数)为止。这种方法可以用于消除数据中的噪声(异常数据),以及帮助发现任意形状的聚类。DBSCAN 就是一个典型的基于密度方法,该方法根据密度阈值不断增长聚类;OPTICS 也是一个基于密度方法,该方法提供聚类增长顺序以便进行自动或交互式数据分析。

(4)基于网格的方法。

基于网格方法是将对对象空间划分为有限数目的单元以形成网格结构,所有聚类操作均是在这一网格结构上进行的。这种方法主要优点就是处理时间由于与数据对象个数无关而仅与划分对象空间的网格数相关,从而显得相对较快。STING 就是一个典型的基于网格的方法;CLIQUE 和 Wave-Cluster 是两个基于网格和基于密度的聚类方法。

(5)基于模型的方法。

基于模型方法就是为每个聚类假设一个模型,然后再去发现符合相应模型的数据对象。一个基于模型

的算法可以通过构造一个描述数据点空间分布的密度函数来确定具体聚类。它根据标准统计方法并考虑到“噪声”或异常数据,可以自动确定聚类个数,因而它可以产生很强壮的聚类方法。基于模型的方法主要是统计学方法和神经网络方法两类。

2 聚类方法基础上的 Markov 预测模型

基于聚类的 Markov 预测模型方法主要分三个步骤:

(1)首先是采用既定聚类方法对页面集合进行聚类,从而得到聚类结果集。

(2)然后用聚类结果用 Markov 预测模型进行预测,结果为某个聚类。

(3)最后在第一次预测的结果类范围内进行低阶的 Markov 预测,得到最终预测结果页面。

下面通过实例来描述基于聚类的 Markov 预测模型过程第二步,现有一用户会话序列:  $US1\{P3, P2, P1\}$ 、 $US2\{P3, P5, P2, P1, P5, P4\}$ 、 $US3\{P4, P5, P2, P1, P5, P4\}$ 、 $US4\{P3, P4, P5, P2, P1\}$ 、 $US5\{P1, P4, P2\}$ ;采用既定聚类方法,可得到聚类结果集: $C1 = \{P1, P2\}$ ,  $C2 = \{P3, P4\}$ ,  $C3 = \{P5\}$ 。将原始会话根据聚类进行转换可得到: $US1\{C2, C1, C1\}$ 、 $US2\{C2, C3, C1, C1, C3, C2\}$ 、 $US3\{C2, C3, C1, C1, C3, C2\}$ 、 $US4\{C2, C2, C3, C1, C1\}$ 、 $US5\{C1, C2, C1, C3, C2\}$ ,根据 Markov 模型算法,可以得到一阶和二阶转移概率矩阵,如表 1 和表 2 所示。

表 1 聚类上的一阶 Markov 转移矩阵

	C1	C2	C3	SUM
$S1 = \{C1\}$	0.5	0.125	0.375	8
$S2 = \{C2\}$	0.33	0.17	0.5	6
$S3 = \{C3\}$	0.5	0.5	0	6

表 2 聚类上的二阶 Markov 转移矩阵

	C1	C2	C3	SUM
$S1 = \{C2, C1\}$	0.5	0	0.5	2
$S2 = \{C1, C3\}$	0	1	0	3
$S3 = \{C1, C1\}$	0	0	1	2
$S4 = \{C3, C1\}$	1	0	0	3
$S5 = \{C2, C3\}$	1	0	0	3
$S6 = \{C2, C2\}$	0	0	1	1
$S7 = \{C1, C2\}$	1	0	0	1

3 基于聚类方法的 Markov 预测方法分析

网页的聚类可以从多个角度、多个属性去考察,对于成熟网站上的固定页面和相同的考察点来说,它的

聚类隶属度的变化率应该很小<sup>[5]</sup>。在不考虑增加新页面的情况下,页面聚类没有必要频繁进行。所以在文中没有把聚类的开销计算到 Markov 预测模型当中。参照上面提出的方法并结合对 Markov 原始预测模型方法的分析,可得到基于聚类方法的 Markov 预测模型算法的空间开销约为  $C^{k+1}$ ,时间开销约为  $SC \times SL \times c \times k + C^{k+1}$ ,其中  $c$  为聚类类别数, $k$  为 Markov 预测模型阶数, $SC$  为会话集大小, $SL$  为会话长度。

而对于有  $n$  个状态的  $k$  阶 Markov 原始模型,首选是两个矩阵的空间开销,根据算法需要,每个矩阵的空间开销为  $n^k \times n$ ,其中  $n^k$  为序列  $S$  的组合数。因为两个矩阵的操作不是同步的,所以可以对状态转移概率矩阵稍作修改,将两矩阵的数据存在一个矩阵里。以第 2 部分所举实例为例,经过两次循环:在第一个大循环,存储完每个状态转移计数值后得到矩阵(如表 3 所示);在第二个大循环中首先对每个序列  $S_i$  的所有状态进行遍历,求得状态转移计数总和,然后再遍历第二遍求得每个状态转移的概率。这里给出第一次遍历求和和结束时状态转移概率矩阵的状态(如表 4 所示)。

表 3 第一个循环后的一阶 Markov  
模型状态转移概率矩阵

	P1	P2	P3	P4	P5	SUM
S1 = {P1}	0	0	0	1	2	0
S2 = {P2}	4	0	0	0	1	0
S3 = {P3}	0	1	0	1	1	0
S4 = {P4}	0	1	0	0	2	0
S5 = {P5}	0	3	0	3	0	0

表 4 第一次遍历完后的一阶 Markov  
模型状态转移概率矩阵

	P1	P2	P3	P4	P5	SUM
S1 = {P1}	0	0	0	0.33	0.67	3
S2 = {P2}	4	0	0	0	1	5
S3 = {P3}	0	1	0	1	1	0
S4 = {P4}	0	1	0	0	2	0
S5 = {P5}	0	3	0	3	0	0

这样,矩阵所需要的空间约为  $n^{k+1}$ 。然后是算法的时间开销,第一个单次循环的时间开销主要消耗在序列  $S$  在矩阵中的定位上。如果采用遍历的方法,时间开销是  $n^k$ ,而通过某些方法,可以将该复杂度降低。比如采用树结构存储矩阵行的入口,该复杂度可降为  $n \times k$ 。第一个循环的时间开销计算还包括会话集的大小和会话的长度,但由于作为学习的对象,它们相对独立于算法之外,基本上任何预测算法均需要遍历每个会话中的页面访问序列,所以这里的计算没有将它们考虑在内。然而第二个循环要遍历整个矩阵,时间开销仍

为  $n^k \times n$ ,两个循环综合考虑下,时间开销约为  $SC \times SL \times n \times k + n^{k+1}$ ,其中  $SC$  为会话集大小, $SL$  为会话长度。

由以上分析可以看出,在页面状态数  $n$  和 Markov 模型阶数  $k$  比较大时,Markov 原始预测模型算法的空间和时间开销都非常惊人。而对于基于聚类方法的 Markov 预测模型算法,因为在一般的聚类方法中,聚类数  $c$  要远小于页面总数  $n$ ,所以无论是空间开销和时间开销,基于聚类的方法都大大降低了。以  $c:n = 1:10$  为例,空间开销就是原来  $10^k$  分之一。当  $SC$  远大于  $c$  时,时间开销可以近似地用  $SC \times SL \times c \times k$  来计算,是原来的十分之一。另外由于在聚类方法中,聚类数  $c$  作为一个参数是人为选择的。因此,基本上 Markov 预测模型的空间规模变得可以控制,大大提高了 Markov 模型的应用范围。而在新添少量网页的时候,先做聚类分析,如果聚类没有变化的话,高阶 Markov 预测模型不需要重新构建,只需要重建低阶的预测模型。这在一定程度上提高了预测模型的可扩展性。

#### 4 结束语

文中提出了一种改进的 Markov 预测模型方法。通过实例看到改进后的 Markov 模型较好地克服了原始 Markov 模型的缺陷。但在同时,也看到改进的 Markov 预测模型并不是完美的,在效率提高的结果里,付出了降低预测准确率的代价。这里的原因主要来自两个方面:

(1)使用类别转移来预测下一步访问页面所属的类别,而对于历史访问记录的考察也是到类的级别,和原始模型直接考察页面级的访问相比,有一定的模糊性。

(2)预测过程分两步执行,在第一步的预测时就可能产生失败,接而在第二步的预测当中将这个误差传递了下去。所以和原始模型直接预测页面相比,准确性有所下降。

不过由于是将高阶 Markov 模型类别预测和低阶 Markov 模型页面预测相结合,和原始低阶 Markov 模型页面预测相比,准确性还是有优势的,毕竟改进的模型考察了页面类别的访问历史。在今后的研究工作中,将考虑从页面权重和个体用户访问习惯、偏好等角度进一步提高 Markov 模型预测结果的准确度。

#### 参考文献:

- [1] 张延安. 试论马尔可夫模型及应用[J]. 沈阳大学学报, 2001(2): 44-46.

(下转第 83 页)

编码提供了一个良好条件,因此采用三维小波编码中的帧间小波编码方法,如图 3 所示,将运动补偿技术和时间一维变换结合起来,同时采用运动估计补偿和时间一维小波分解两种技术去除时间冗余信息。对于视频编码,一般采取帧间运动估计运动补偿和帧内小波变换相结合的方法,只要在静止图像编码基础之上考虑运动补偿和内插就可以实现动态图像的编码。因为对于视频序列来说,相邻两幅图像之间的变化很小,只需对两幅图像的差值进行编码,在解码后与第一幅图像叠加便可得到后一幅的图像。

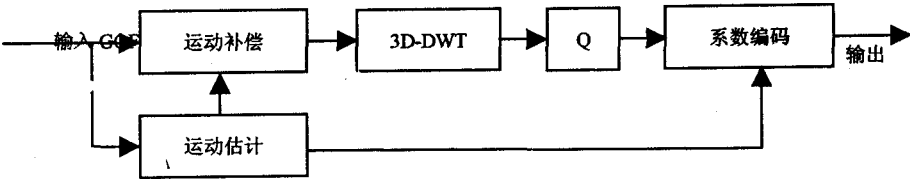


图 3 帧间小波编码框图

3 实验结果

为了验证所提出的改进方法的有效性,采用 QCIF 格式的 Carphone、Akiyo、Foreman、Claire 视频序列,选取 4 级形状自适应 5/3 整数小波变换,对所提方法的性能做了测试。如图 4 所示,截取 Carphone 视频序列的第 161 帧为例,分别选取不同的感兴趣阈值进行对比观察,实验结果表明,所改进的方法是有效的,它能

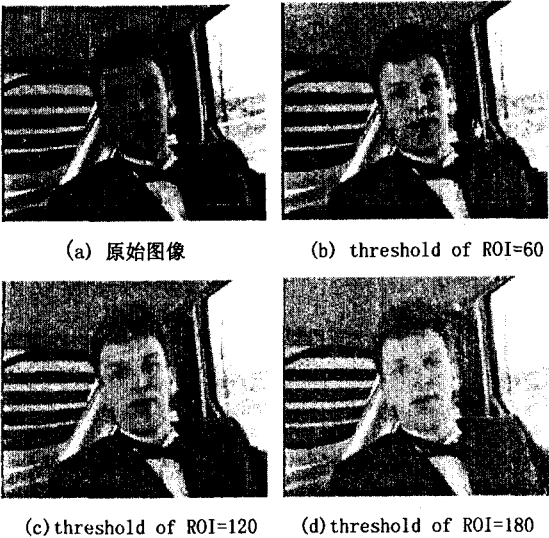


图 4 小波变换可分级编码效果图

够在一定程度上提高编码效率,并且随着感兴趣区域阈值的提高,增强层的编码码率随之降低,而全分辨率图像的信噪比也会随之降低。

表 1 Carphone 等视频序列感兴趣区域小波变换可分级编码信噪比与码率比较

threshold value of ROI	60		120		180	
客观标准 单位: dB/(bit/frame)	PSNR/BitRate		PSNR/BitRate		PSNR/BitRate	
Carphone	32.2917	92748.3	30.1073	59383.4	25.5039	28594.5
Akiyo	33.5990	139604	29.4083	106493	26.8937	84728.9
Foreman	31.7385	79383.5	28.8402	67385.3	25.9373	49375.3
Claire	32.7385	110483	29.7395	85934.5	27.3792	67395.7

4 结束语

视频编码作为热门的研究方向之一,受到人们的广泛关注。文中在可分级编码方法的基础上对其进行改进,实验收到了预期的效果。同时其也有一些不完善的地方,比如在感兴趣区域与基本层图像对应区域进行匹配方面还有待研究与改善。随着感兴趣区域检测技术与基于小波的视频可分级编码技术的完善,视频编码效率及重建视频质量会进一步提高。

参考文献:

[1] 沈兰荪,卓力.小波编码与网络视频传输[M].北京:科学出版社,2005:137-139.

[2] Christopoulos C, Askelof J, Larsson M. Efficient methods for encoding region of interest in the upcoming JPEG2000 still image coding standard[J]. IEEE Signal Processing Letters, 2000,7(9):247-249.

[3] 蒋鹏.基于小波变换的感兴趣区域压缩编码技术研究[D].长春:吉林大学,2007:13-16.

[4] 曾啸天.基于感兴趣区域可分级视频编码研究[D].大连:大连理工大学,2007:58-64.

[5] 王艳娟,陈晓红,黄晓欣.图像感兴趣区域检测技术[J].计算机与数字工程,2007,35(5):138-139.

[6] 王艳娟,陈晓红,邹丽.图像感兴趣区域自动提取算法[J].科学技术与工程,2007,7(12):2867-2871.

[7] 贾冬顺,张正炳,邓慧萍.基于小波变换的视频编码算法分类研究[J].电子与电脑,2007(3):100-102.

(上接第 80 页)

[2] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition [J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[3] Park H S, LEE S W. A truly 2-D hidden Markov model for off-line handwritten character recognition [J]. Pattern Recognition, 1998, 31(2): 1864-1894.

[4] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范明, 孟小峰译. 北京:机械工业出版社, 2001.

[5] 王实, 高文, 李锦涛, 等. 基于隐马尔可夫模型的兴趣迁移模式发现[J]. 计算机学报, 2001(2): 152-156.