隐马尔可夫模型及其最新应用与发展①

朱 明 郭春生 (杭州电子科技大学 通信工程学院 浙江 杭州 310018)

摘 要： 隐马尔可夫模型是序列数据处理和统计学习的一种重要概率模型，已被成功应用于许多工程任务中。首先介绍了隐马尔可夫模型的基本原理，接着综述了其在人的行为分析、网络安全和信息抽取中的最新应用。最后对最近提出来的无限状态隐马尔可夫模型的原理及最新发展进行了总结。

关键词： 隐马尔可夫模型;行为分析;网络安全;信息抽取;无限状态隐马尔可夫模型

Hidden Markov Model and Its latest Application and Progress

ZHU Ming, GUO Chun-Sheng

(College of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: Hidden Markov Model (HMM) is an important probabilistic model of sequential data processing and statistical study. It has already been successfully applied in many projects in practice. Firstly, this paper introduces the basic principles of the Hidden Markov Model, and then gives a review to its latest application in the human activity analysis, network security and information extraction. Finally it summarizes the theory and latest progress of the recently proposed infinite Hidden Markov Model (iHMM).

Keywords: HMM；activity analysis；network security；information extraction；iHMM

1 引言

隐马尔可夫模型(Hidden Markov Model, HMM)作为一种统计分析模型，创立于 20 世纪 70 年代，80 年代得到了传播和发展并成功应用于声学信号的建模中，到目前为止，它仍然被认为是实现快速精确语音识别系统最成功的方法。作为信号处理的一个重要方向，HMM 广泛应用于图像处理，模式识别，语音人工合成和生物信号处理等领域的研究中，并取得了诸多重要的成果[1]。近年来，很多研究者把 HMM 应用于计算机视觉、金融市场的波动性分析和经济预算等新兴领域中，因此，结合实际应用，进一步研究各种新型 HMM 及其性质，具有重要的意义。文章首先介绍了 HMM 的基本理论，接着对其在人的行为分析、网络安全和信息抽取中的最新应用进行了综述。针对经典 HMM 应用中存在的两大问题，近年来提出了无限状态隐马尔可夫模型(infinite Hidden Markov Model ，iHMM)，文章的最后对其基本理论及最新发展进行了总结。

2 HMM的基本原理及结构

2.1 HMM 的基本原理

HMM 由两个随机过程组成，其中一个是状态转移序列，它是一个单纯的马尔可夫过程；另一个是与状态对应的观测序列，如图 1 为一状态数为 3 的 HMM 示意图，其中为状态序列，它们之间的转移是一个马尔可夫过程，为各状态下对应的观测值。在实际问题中，我们只能看到观测值，而不能直接看到状态，只能是通过观测序列去推断状态的存在及转移特征，即模型的状态掩盖在观测序列之中，因而称之为“隐”Markov 模型。

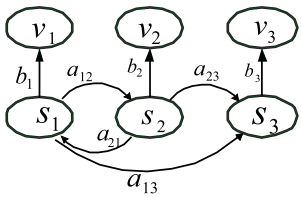

图 1 状态数为 3 的 HMM 示意图

设模型的状态数目为，可观测到的符号数目为，

① 收稿时间:2009-10-25;收到修改稿时间:2009-12-06

则可用三元组来表示一个 HMM[2]，其中为状态转移概率矩阵，为给定状态下的观测值概率矩阵或观测概率密度函数，为初始状态概率分布矢量。

可以用盒子与彩球模型来描述一个 HMM[3]：设某人在 3 个装有红白两种颜色的球的盒子中任取一个盒子，然后在此盒中每次抽取一个球，连续地在同一盒中按下面给定的方式抽取次，各盒中球的内容和抽取方式如表 1：

表 1　各盒中球的内容和抽取方式

	红球数	白球数	每次抽取方式
盒 1	90	10	随机取一球，记下颜色后不放回，而放进一个与它颜色不同的球
盒 2	50	50	随机取一球，记下颜色后放回
盒 3	40	60	随机取一球，记下颜色后不放回，并放进一个红球

现在如果某人用上述方法得到了一个观测序列(红，红，红，红，白)(即 T=5)，但并不告诉我们球出自哪个盒子。但我们通过概率计算，可以知道从第一盒中抽出样本(红，红，红，红，白)的概率要比从其它两盒中抽出该样本的概率大得多，从而推断球出自盒 1。此例中的不同盒的抽取方式可以抽象为不同的状态编码方式，这正启示了用 HMM 作为序列数据建模与分类的粗略梗概。

2.2 在应用中需要解决的三个基本问题

(1)学习问题。就是从大量的已知观测序列出发，估计模型参数组，须用动态规化的方法解决该问题，常用的为 Baum-Welch 算法，也叫 EM 算法。

(2)分类问题。即对于一个特定的观测序列，要计算其在已知模型下出现的概率，常用前向变量法求解。

(3)解码问题。从一段观测序列及已知模型出发，估计状态序列的最佳值，可用 Viterbi 算法计算。

通过解决以上全部或部分问题，可以实现很多复杂的工程任务。如通过解决前两个问题，HMM 可方便地应用于模式识别中，可再用上述的盒子与彩球模型来描述这一过程：各盒中球的内容和球的抽取方式可以抽象为不同的模式，在上例中即有 3 种模式，在一般的模式识别过程中，开始并不知道各盒中球的内容和球的抽取方式，但我们可以通过大量的已知出自各盒的观测样本来推断各盒中球的内容和球的抽取方式，并用 3 个 HMM 来分别表示，即 HMM 的学习问题；而在识别阶段，得到的未知观测序列，已知它由 3 个学习好的 HMM 之一所得，通过计算其在各 HMM 下出现的概率，可以知道其最有可能出自哪个盒子，从而完成对模式的识别，因此，识别阶段就是一个 HMM 的分类问题。

3　HMM的最新应用

HMM 作为序列数据处理和统计学习的一种重要概率模型，具有建模简单，物理意义明确等优点，且已经有很多成熟的算法，是一种精确的匹配时变数据的技术，已经广泛应用于如语音识别、生物信号分析、模式(如人脸，步态，表情等)识别、故障诊断等的研究中，并取得了丰硕的成果，文章将不再对 HMM 在这些方面的应用进行赘述。而是对 HMM 在人的行为分析、网络安全和信息抽取中的最新应用进行综述。

3.1 HMM 在人的行为分析中的应用

人的行为分析在视频会议、人机交互、智能监控、基于行为的视频检索以及医疗诊断等方面有着广泛的应用前景和潜在的经济价值，是当前计算机视觉领域的一个研究热点。它要解决的问题是根据来自摄像机的原始图像数据，通过提取图像中的运动目标，并计算其速度、轨迹、灰度等特征信息来识别人体的动作，并结合上下境信息，来分析人体动作的目的，理解其要传递的语义信息。

行为分析首先对人的运动模式进行分析与描述，然后根据描述进行行为识别。行为识别可以看作是时变序列数据的分类问题，即将未知序列与经过学习得到的代表典型行为的已知序列进行匹配。HMM 的结构可以很好地和这一匹配过程相对应，如人跑步和行走可看作为 HMM 的两状态，而速度可以当作是状态下的一观测值，通过观测值速度的大小可以判断人处在何种状态(跑步或行走)下。

Yamato 等人[4]于 1992 年首次将 HMM 引入到人的行为分析中，开始了行为分析的各种状态空间算法研究。该文利用二维小区域块人运动的网格特征(速度、色彩、纹理等)作为观测序列进行行为的学习和识别；学习是利用 Baum-Welch 算法来优化各行为 HMM 的参数，识别是通过判断未知图像序列在各 HMM 下前向变量的概率计算结果来完成，实验结果表明，HMM 建模能较好地对网球比赛中不同运动员的不同动作(正反手拦网，正反手击球，大力扣球，发球等)进行分类识别。

根据不同环境下人行为的特征，很多文献对 HMM 结构进行了扩展，大部分通过利用 HMM 的分层结构

来建模行为的不同层次。Bregler 等人[5]跟据人体动力学系统中行为的层次性提出了一个综合性的网络来识别人的运动，识别过程分为三个阶段，在低级和中级处理阶段，通过检测与跟踪提取运动特征（速度，轨迹等）并匹配为动力学中的简单运动；高级阶段，HMM 被用来建模由这些简单运动组合而成的复杂行为，识别是通过判断行为在各 HMM 下后验概率的计算结果来确定，实验结果表明，该分层 HMM 能准确地识别出人正常行走和滑动行走等差别很小的行为。

在行为分析中，训练数据为大量的图像系列，并且对于每一类行为都要建立一个 HMM，学习时所需的运算量是巨大的。针对这个问题，很多文献提出了改进的学习方法。如李和平等人[6]提出了一种半监督学习的行为建模方法，实验结果表明，该方法能够在小样本的情况下快速地通过较小的运算量学习好 HMM，进而更实时地检测人的异常行为。

对行为分析的研究已有近 40 年的历史，但只有最近 10 年才成为研究的热点，HMM 的优良特性使之成为行为分析的有力工具。总体而言，行为分析的研究仍处在初级阶段，还有很多问题需要解决。

3.2 HMM 在网络安全中的应用

随着计算机技术的飞速发展，信息网络已经成为社会发展的重要保证，网络安全越来越受到关注。在网络安全研究中，入侵检测是其中重要的一个方面，是对入侵行为进行处理以保证网络安全的前提与基础。传统的入侵检测方法是 Forrest 等人[7]提出的时延嵌入序列(TIDE)方法。

Warrender 等人[8]在 1999 年首次在入侵检测中引入 HMM，随后，HMM 被广泛地应用到入侵检测中，并逐渐展现了其优越性。HMM 应用于入侵检测，主要是基于这样两个现象：(1) 在正常操作时，程序执行的系统调用是局部稳定的；(2)当入侵发生时，程序将会执行大量的异常系统调用。检测通常分为两步，首先利用正常操作程序执行的系统调用作为观测序列来学习 HMM 参数，建立正常操作 HMM，即 HMM 的学习问题；然后将未知程序执行的系统调用观测序列输入到该正常操作 HMM 中，即 HMM 的分类问题，当计算出的前向变量概率低于一定值时，则认为该程序执行的调用不符合正常操作 HMM，进而判断入侵的发生。

在国内，闫巧等人[9]的研究具有重要的参考价值，她通过实验证明：(1)HMM 方法建立的库比 TIDE 方法建立的库要小，检测时速度更快(2)HMM 方法在学习数据不充分时也能得到近似完备的正常轮廓数据库(3)HMM 方法的检测精度比 TIDE 方法更高。然而，HMM 学习和工作中所需要的计算量很大，检测效率和实时性较差，这在一定程序上限制了它在实际系统中的应用。针对这个问题，邬书跃[10]等人提出了一种运算量较小的序列匹配算法来学习 HMM，利用状态序列出现的概率对被监测用户的操作进行分类，实验表明，该方法在保证高检测准确度下同时具有较高的效率。陶龙明[11]等人将 HMM 用于检测隐蔽性强、持续时间长且分步完成的复杂网络攻击(如网络钓鱼攻击、大规模 DDOS 攻击等)，该文通过关联分析不同网络监视器的报警事件，产生用于 HMM 模型学习及检测的报警序列，实验结果表明，HMM 不仅能较好地检测出这些复杂的网络攻击，而且还能对它们进行分类。

在网络安全领域，HMM 除了应用于入侵检测，还应用于数据库异常检测[12]、Web 用户异常访问[13]等的检测中。

3.3 HMM 在信息抽取中的应用

WWW 的广泛应用使得 Internet 成为了信息的海洋，信息抽取(Information Extraction, IE)是处理海量文本信息的重要环节，旨在帮助人们从海量联机文本中快速、准确地抽取自己真正需要的信息，抽取出来的信息以一定的方式存储在数据库中，为情报分析和检测、比价购物、自动文摘、文本分类等各种应用提供服务。

信息抽取中目标信息在网页、文章等中的具体位置，称为抽取域，而要抽取的信息，即抽取域中的内容称为语义项，信息抽取的过程首先要确定抽取域，然后提取相应的语义项。信息的这种抽取方式正好和 HMM 的结构相吻合，即各抽取域和 HMM 的隐状态序列相对应，而语义项和各状态的观测相对应。HMM 应用于信息抽取，具有易于建立、适应性好、抽取精度高等优点，近年来得到了广泛的关注和研究。

于江德[14]等人将 HMM 应用于中文科研论文头部信息的抽取中，HMM 的每个状态和要抽取的一个域(标题区域、作者区域等)相对应，而观测和每一个语义项(标题的内容、作者的姓名等)相对应，该文首先依据论文头部段落中的回车和逗号、分号等标点符号对各

语义项进行切分，然后将大量已知论文的语义项作为观测序列来学习 HMM，在抽取阶段，应用 Viterbi 算法来解码未知观测序列在学习好的 HMM 下的状态序列，即找出语义项相关的抽取域，进而抽取各语义项，实验结果表明，HMM 应用于中文科研论文信息抽取中具有很高的准确率，对提高论文检索效率、方便各种论文统计分析等具有重要的潜在参考价值。

王宇宁[15]利用 HMM 来设计网上汽车服务系统，研究了其中的信息抽取模块的需求和设计方案，该文将汽车车型的排量、指导价、车身重量、油耗等信息作为抽取域，利用网页 HTML 语言语法结构特点对各语义项进行切分，生成信息节点树作为信息抽取 HMM 的观测序列，进而学习 HMM 并抽取相关信息，实验结果证实了应用 HMM 抽取网页信息的有效性。

针对 HMM 在信息抽取中具体应用所遇到的问题，很多文献提出了改进的 HMM 算法。周顺先等人[16]提出了一种基于二阶 HMM 的文本信息抽取算法，该二阶 HMM 合理地考虑了概率和模型历史状态的关联性，对错误有更强的识别能力，通过在科研论文信息抽取的仿真实验表明，该算法比一阶 HMM 的算法具有更高的抽取精确度。

除了在人的行为分析、网络安全和信息抽取中的应用外，近年来，还有人将 HMM 用于金融、管理和心理情绪等的建模中，但这方面的文章仍然较少。随着时代的发展，作为时变数据匹配的重要工具，HMM 必将有更广泛的应用。

4　无限状态隐马尔可夫模型

虽然 HMM 是一种常用的概率统计模型，但在实际应用中，却受到很大的限制，主要体现在：(1)HMM 学习使用的经典 EM 算法中 M 步骤估计时没有考虑到模型的复杂度，不能解决模型的过适应或欠适应问题；(2)在使用前，必须确定 HMM 的结构，即模型中的状态数目，然而，由于实际数据的复杂性以及数据的动态更新性，人为地指定状态数目通常不能最佳地描述数据，这样，模型只能从给定的数据集中得到有限的信息。针对这些问题，Beal 等人[17]于 2002 年提出了无限状态隐马尔可夫模型(infinite Hidden Markov Model, iHMM)，iHMM 不再需要人为地指定状态数目，而是让数据自己说话，跟据数据自身的特性智能地挑选最优的状态数目，从而解决上述问题。不同于其它算法对 HMM 的改进，iHMM 是一种全新的 HMM，是对 HMM 基础数学理论和经典算法的全面改变，但它体现的数学建模思想（隐状态的转移和与状态对应的观测等概念）是一样的。

4.1 iHMM 原理

iHMM 中状态的转移并不是一个马尔可夫过程，而是一个狄利克雷过程(Dirichlet Process, DP)。狄利克雷过程可用一个中国餐馆过程来形象地描述：假设在一个餐桌数量无限的中国餐馆中来了 n 个顾客，他们坐在前面的 k 张餐桌，且第 i 张餐桌有 n_i 个顾客，当第 n+1 个顾客到达时，他将：

(1)以 $\frac{n_i}{\alpha+n}$ 的概率坐在已有人的第张餐桌

(2)以 $\frac{\alpha}{\alpha+n}$ 的概率新开一张餐桌坐下

其中是一个正的常量。从中我们可以看出，在一个中国餐馆过程中，新来的顾客被分配到有人坐的第 i 张餐桌的概率主要和 n_i 相关，这样将导致“人越多的餐桌，再有人入坐的可能性越大”，而顾客开一张新餐桌的概率由 α 控制。

在 iHMM 中，模型的状态序列和中国餐馆过程中的餐桌对应，而观测序列和每个餐桌的顾客对应。从中国餐馆过程可以看到，如果有必要，模型中的状态数(即餐桌数)可以为无穷大，而又因为“人越多的餐桌，再有人入坐的可能性越大”，这将导致餐桌的最终数量趋于一有限的固定值。即由 DP 控制的 iHMM 中状态的数目将跟据数据的实际情况以一定的概率增加或减少，从而可以自动地选择描述数据的最佳状态数目，进而自适应地优化 iHMM 的结构。

4.2 iHMM 的最新发展

解决 HMM 三个基本问题的经典算法都是在马尔可夫过程原理的基础上推导得到的，因此不适应于状态转移为 DP 过程的 iHMM 的分析计算中，要应用 iHMM，必须为其设计新的算法。

Beal 的文章提出并详细介绍了 iHMM，也分析了解决 iHMM 三个基本问题的方法，如用 Gibbs 采样来解码隐状态序列和学习模型的参数，用无限状态粒子滤波器来解决评估问题等。然而在应用中，却发现这些算法运算量太大，效率低，并不实用。

Teh 等人的文章[18]系统地介绍了 DP 及其在 iHMM 中的应用，使得 iHMM 有了坚实的数学理论基

础，文章还介绍了一种相对有效的 Gibbs 采样方法，但这仍然不能满足 iHMM 的应用，并且研究者逐渐认识到 Gibbs 采样本身并不适合应用在时间序列模型的计算中。Fox 等人[19]在 Teh 的文章基础上提出了一种 iHMM 处理时变数据的方法，并深入地研究了一种基于截断近似的采样方法，文中给出的实验结果表明，该方法在某些应用(如说话人检索)上性能有显著的提高。

Gael 等人[20]于 2008 年提出了一种束采样(Beam Sampling)方法，该方法用层采样将无限模型自适应地截断成有限模型进行处理，这样就能同时进行动态规化与重采样整个状态系列，从而提高运算效率。实验表明，比起 Gibbs 采样，该方法大大提高了 iHMM 的学习效率，把运算量降到了应用中可以接受的范围内，使 iHMM 具有了实用价值。

以上 4 篇文献为 iHMM 的应用提供了坚实的平台，在这个基础上，研究者可以将 iHMM 方便地应用于各个领域的建模中。

5 总结

文章介绍了 HMM 的基本原理，并综述了其在人的行为分析、网络安全和信息抽取中的最新应用。最后重点介绍了 iHMM，iHMM 解决了经典 HMM 应用中存在的两大限制条件的约束问题，能跟据实际数据自动地学习模型的结构。目前，在国内的期刊上，还没有关于 iHMM 应用的文章，在国外期刊上，这类文章也较少，随着 iHMM 理论的成熟，iHMM 必将大量应用于各种序列信号的建模与分析中。

参考文献

1 Bilmes JA. What HMMs can do. IEICE TRANSACTIONS on Information and Systems, 2006,89(3):1－24.

2 Rabiner LR A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of IEEE, 1989,77(2):257－286.

3 龚光鲁,钱敏平.应用随机过程教程及在算法和智能计算中的随机模型.北京:清华大学出版社, 2004.247－249.

4 Yamato J, Ohya J, Ishii K. Recognition human action in time sequential images using hidden Markov model. Proc. of IEEE Conference on Computer Vision and Pattern Recognition. LosAlamitos, CA: IEEE, 1992:379－385.

5 Bregler C. Learning and recognizing human dynamics in video sequences. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE, 1997:568－574 .

6 李和平,胡占义,吴毅红,吴福朝.基于半监督学习的行为建模与异常检测.软件学报, 2007,18(3):527－537.

7 Forrest S, Hofmeyr SA, Somayaji A, Longstaff TA. A sense of self for UNIX processes. Proc. of the 1996 IEEE Symposium on Security and Privacy. Los Alamitos, CA: IEEE Computer Society Press, 1996:120－128.

8 Warrender C, Forrest SI, Pearlmuter B. Detecting Intrusions Using System Calls: Alternative Data Models. 1999 IEEE Symposium on Security and Privacy.Los Alamitos, CA: IEEE Computer Society Press, 1999: 133－145.

9 闫巧,谢维信,宋歌,喻建平.基于 HMM 的系统调用异常检测.电子学报, 2003,31(10):486－1490.

10 邬书跃,田新广.基于隐马尔可夫模型的用户行为异常检测新方法.通信学报, 2007,28(4):38－43.

11 陶龙明,史志才,彭丹,马武.HMM 模型在检测复杂网络攻击中的应用.计算机工程与应用, 2008,44(7): 136－138.

12 王岳斌,阳国贵,邝祝芳.基于 HMM 的数据库异常检测系统设计与实现.计算机应用与软件, 2009,26 (1):96－99.

13 谢逸,余顺争.基于 Web 用户浏览行为的统计异常检测.软件学报, 2007,18(4):967－977.

14 于江德,樊孝忠,尹继豪,顾益.基于隐马尔可夫模型的中文科研论文信息抽取.计算机工程, 2007,33 (19):190－192.

15 王宇宁.隐马尔可夫模型在信息抽取中的应用研究[硕士学位论文].大连:大连理工大学, 2007.

16 周顺先,林亚平,王耀南,易叶青.基于二阶隐马尔可夫模型的文本信息抽取.电子学报, 2007,35(11): 2226－2231.

17 Beal MJ, Ghahramani Z, Rasmussen CE. The Infinite

(下接第 216 页)

(上接第 259 页)

Hidden Markov Model. Dietterich ed. Advances in Neural Information Processing Systems. Cambridge, MA:MIT Press, 2002:577－584.

18 Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical dirichlet processes. Journal of the American Statistical Association, 2006,101:1566－1581.

19 Fox EB, Sudderth EB, Jordan MI, Willsky AS. An HDP-HMM for Systems with State Persistence. ICML 2008 － Proc. of the 25th International Conference on Machine Learning. New York: ACM Press, 2008:312－319.

20 Gael JV, Saatci Y, Teh YW, Ghahramani Z. Beam sam pling for the infinite hidden markov model. ICML 2008- Proc. of the 25th International Conference on Machine Learning. New York: ACM Press, 2008:1088－1095.