**中图分类号：**

**论文编号：**



硕士学位论文

面向程序设计语言课程的个性化学习效果建模研究

作者姓名 王珊珊

学科专业 软件工程

指导教师 吴文峻 教授

培养院系 计算机学院

**The Research of Modeling Student Personalized Learning Outcomes for Programming Language Course**

A Dissertation Submitted for the Degree of Master

**Candidate：Shanshan Wang**

**Supervisor：Prof.Wenjun Wu**

School of Computer Science and Engineering

Beihang University, Beijing, China

**中图分类号：**

**论文编号：**

硕 士 学 位 论 文

面向程序设计语言课程的个性化学习

效果建模的研究

作者姓名 王 珊 珊 申请学位级别 工学硕士

指导教师姓名 吴 文 峻 职 称 教授

学科专业 软件工程 研究方向 软件工程

学习时间自 2014年 09月 15日 起至 2017年 03月 日止

论文提交日期 2017年 03月 22日 论文答辩日期 2017年 03月 日

学位授予单位 北京航空航天大学 学位授予日期 2017年 03月 日

关于学位论文的独创性声明

本人郑重声明：所呈交的论文是本人在导师指导下独立进行研究工作所取得的成果，论文中有关资料和数据是实事求是的。尽我所知，除文中已经加以标注和致谢外，本论文不包含其他人已经发表或撰写的研究成果，也不包含本人或他人为获得北京航空航天大学或其它教育机构的学位或证书等而使用过的材料。与我一同工作的同志对研究所做的任何贡献均已在论文中作出了明确的说明。

若有不实之处，本人愿意承担相关法律责任。

学位论文作者签名： 日期： 年 月 日

关于学位论文使用授权的声明

本人完全同意北京航空航天大学有权使用本学位论文（包括但不限于其印刷版和电子版），使用方式包括但不限于：保留学位论文，按规定向国家有关部门（机构）送交学位论文，以学术交流为目的赠送和交换学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存学位论文。

保密学位论文在解密后的使用授权同上。

学位论文作者签名： 日期： 年 月 日

指导教师签名： 日期： 年 月 日

摘 要

**关键词：**程序设计语言课程，学习效果评估，知识跟踪模型，贝叶斯网络

**Abstract**

**Key words:** Programming Language Course, Student Outcomes Modeling, Knowledge Tracing Model, Probabilistic Graphical Model

**目 录**

[第一章 绪论 1](#_Toc438217673)

[1.1 研究背景与意义 1](#_Toc438217674)

[1.2 国内外研究现状 1](#_Toc438217675)

[1.2.1 学习效果评估研究现状 **错误!未定义书签。**](#_Toc438217676)

[1.2.2 同伴互评研究现状 **错误!未定义书签。**](#_Toc438217677)

[1.3 研究目标与内容 1](#_Toc438217678)

[1.4 论文组织结构 1](#_Toc438217679)

[第二章 相关技术介绍 2](#_Toc438217680)

[2.1 分类模型介绍 **错误!未定义书签。**](#_Toc438217681)

[2.1.1 线性判别分析 **错误!未定义书签。**](#_Toc438217682)

[2.1.2 逻辑斯特回归 **错误!未定义书签。**](#_Toc438217683)

[2.1.3 支持向量机 **错误!未定义书签。**](#_Toc438217684)

[2.2 隐马尔科夫模型 **错误!未定义书签。**](#_Toc438217685)

[2.2.1 模型表示 **错误!未定义书签。**](#_Toc438217686)

[2.2.2 基本问题 **错误!未定义书签。**](#_Toc438217687)

[2.2.3 参数学习 **错误!未定义书签。**](#_Toc438217688)

[2.3 马尔科夫链蒙特卡洛模拟 **错误!未定义书签。**](#_Toc438217689)

[2.3.1 共轭分布 **错误!未定义书签。**](#_Toc438217690)

[2.3.2 MCMC方法 **错误!未定义书签。**](#_Toc438217691)

[2.3.3 吉布斯采样 **错误!未定义书签。**](#_Toc438217692)

[2.4 本章小结 **错误!未定义书签。**](#_Toc438217693)

[第三章 学习效果评估 **错误!未定义书签。**](#_Toc438217694)

[3.1 基于学习行为进行初步预估 **错误!未定义书签。**](#_Toc438217695)

[3.1.1 问题描述 **错误!未定义书签。**](#_Toc438217696)

[3.1.2 行为特征设计 **错误!未定义书签。**](#_Toc438217697)

[3.1.3 行为特点分析 **错误!未定义书签。**](#_Toc438217698)

[3.2 利用答题表现进行准确评判 **错误!未定义书签。**](#_Toc438217699)

[3.2.1 动机 **错误!未定义书签。**](#_Toc438217700)

[3.2.2 知识跟踪模型 **错误!未定义书签。**](#_Toc438217701)

[3.2.3 个性化的知识跟踪模型 **错误!未定义书签。**](#_Toc438217702)

[3.3 本章小结 **错误!未定义书签。**](#_Toc438217703)

[第四章 同伴互评机制设计 **错误!未定义书签。**](#_Toc438217704)

[4.1 基于学习效果的任务分配算法 **错误!未定义书签。**](#_Toc438217705)

[4.1.1 问题描述 **错误!未定义书签。**](#_Toc438217706)

[4.1.2 LPT算法 **错误!未定义书签。**](#_Toc438217707)

[4.1.3 MLPT算法 **错误!未定义书签。**](#_Toc438217708)

[4.2 结果聚合概率模型 **错误!未定义书签。**](#_Toc438217709)

[4.2.1 变量设定 **错误!未定义书签。**](#_Toc438217710)

[4.2.2 模型构建 **错误!未定义书签。**](#_Toc438217711)

[4.2.3 参数学习 **错误!未定义书签。**](#_Toc438217712)

[4.3 本章小结 **错误!未定义书签。**](#_Toc438217713)

[第五章 数据收集与实验分析 8](#_Toc438217714)

[5.1 数据收集 **错误!未定义书签。**](#_Toc438217715)

[5.1.1 数据来源 8](#_Toc438217716)

[5.1.2 数据规模 **错误!未定义书签。**](#_Toc438217717)

[5.2 两阶段学习效果评估方法实验 **错误!未定义书签。**](#_Toc438217718)

[5.2.1 阶段一 行为预测模型 **错误!未定义书签。**](#_Toc438217719)

[5.2.2 阶段二 知识跟踪模型 **错误!未定义书签。**](#_Toc438217720)

[5.3 同伴互评模型实验 **错误!未定义书签。**](#_Toc438217721)

[5.3.1 任务分配MLPT算法 **错误!未定义书签。**](#_Toc438217722)

[5.3.2 结果聚合概率模型 **错误!未定义书签。**](#_Toc438217723)

[5.4 本章小结 **错误!未定义书签。**](#_Toc438217724)

[第六章 算法应用 8](#_Toc438217725)

[6.1系统总体逻辑架构 8](#_Toc438217726)

[6.2学习效果评价模型与系统的集成 **错误!未定义书签。**](#_Toc438217727)

[6.3同伴互评机制与系统的集成 **错误!未定义书签。**](#_Toc438217728)

[6.4本章小结 **错误!未定义书签。**](#_Toc438217729)

[总结与展望 9](#_Toc438217730)

[参考文献 9](#_Toc438217733)

[攻读硕士学位期间取得的学术成果 11](#_Toc438217734)

[致 谢 12](#_Toc438217735)

**图 录**

**表 录**

# 绪论

## 研究背景与意义

## 国内外研究现状

## 研究目标与内容

## 论文组织结构

本文共分为七章，具体结构组织如下：

第一章，绪论。首先阐述本文的研究背景与意义，从学生的等方面阐述国内外的研究现状，紧接着介绍本文工作的研究目标与内容，最后梳理本文的组织结构。

第二章，相关技术介绍。本章介绍本文涉及到的主要技术，包括项目反应理论、贝叶斯网络、隐马尔可夫模型和采样方法。

第三章，学习效果评估。本章主要介绍了学习效果评价模型。首先介绍了基于学习行为初步预估学习效果的方法，其中重点分析了学习行为特征的设计和抽取，然后介绍了利用答题表现准确评判学习效果的方法，通过结合学习行为得到的初始知识掌握情况建立个性化的知识跟踪模型，对学生的知识水平进行准确评判。

第四章，同伴互评机制设计。本章着重介绍了同伴互评机制中的任务分配和结果聚合问题，详细阐述了基于学生学习效果设计的MLPT任务分配算法和综合考虑评分者水平和偏差的结果聚合概率模型。

第五章，

第六章，数据收集与实验分析。本章首先介绍了实验数据来源和规模，接着利用收集到的数据设计实验对本文提出的学习效果评价模型和同伴互评机制进行了有效性分析和验证。

第七章，算法应用。本章首先介绍了系统的总体逻辑框架，然后重点介绍了如何将本文提出的学习效果评价模型和同伴互评机制集成到当前的MOOC平台中。

最后对本论文的研究结果进行总结与展望，并提出相关需要改进的工作。

# 相关技术介绍

本章介绍了本文涉及到的相关理论和技术，主要介绍了项目反应理论、贝叶斯网络、隐马尔科夫模型、以及采样方法。

## 项目反映理论

### 理论概念

### 模型表示

### 参数估计

## 贝叶斯网络

### 数学定义

### 动态贝叶斯网络

### 模型推理

## 隐马尔科夫模型

隐马尔科夫模型 ( Hidden Markov Model, HMM )[1]是一种时序概率统计分析模型，描述由一个隐藏变量的马尔可夫链生成一个不可观测的随机状态序列，再由状态序列生成一个随机观测序列的过程。每个状态生成一个观测，序列中的每个位置可以看作为一个时刻。隐马尔科夫模型是双重随机过程，分为两部分：描述隐藏状态之间转移关系的马尔科夫链，用状态转移概率描述，此转移具有马尔科夫性，下一时刻的状态只与当前的状态有关，与过去状态无关；描述状态与观测序列关系的一般随机过程，其用观测概率描述。该模型的双重随机过程如图\*\*所示：



图1 隐马尔科夫模型的双重随机过程

### 基本理论

隐马尔科夫模型由一个五元组表示，下面我们介绍这五个基本要素：

*Q*为所有状态的集合，*V*为所有观测的集合。, 其中，*N*表示模型中的状态数目，*M*表示模型中观测值的数目；

*I*为长度T的状态序列，*O*为观测序列。其中

为初始状态概率矩阵，表示初始时刻t=1处于状态的概率。满足约束 ；

为状态转移概率矩阵， 表示由时刻t-1状态为转移到时刻t状态为的概率，与具体的时刻无关。*A*满足约束 ；

为观测值概率矩阵（发射矩阵），表示在时刻t隐含状态为的条件下观测值为的概率，也与具体时刻无关，*B*满足约束；

如图\*\*所示为隐马尔科夫模型的状态转移示意图，共包含三个隐藏状态和三个观测值。



图 隐马尔科夫模型的状态转移示意图

### 问题表述

其实对于隐马尔可夫模型来说，如果已经知道所有隐含状态之间的转移概率和所有隐含状态到观测值之间的观测概率，做模拟是比较容易的。但是一般应用该模型时候，往往是缺失了一部分信息，我们需要运用相关算法去估计这些缺失值，总结起来隐马尔科夫模型可以分为三类基本问题：

1. 评估问题。给定观测序列和模型参数，计算从此模型生成该观测序列的概率，可用来用来评估一个模型和已定观测序列的匹配程度，一般采用Forward-backward算法来解决评估问题。
2. 解码问题。已知观测序列和模型参数，对给定的观测序列*O*解码出最佳隐藏状态序列*I*，使最大。本文采用Viterbi算法来解决解码问题。
3. 学习问题。给定观测序列，估计模型参数，利用极大似然估计，使在该模型参数条件下给定观测序列出现的概率最大，可以采用Baum-Welch算法解决学习问题。

### 基本算法

* 1. Forward-backward算法

1. 前向算法：给定隐马尔可夫模型，定义到时刻t的状态为而且观测序列为，记：

可以递归求得前向概率及观测序列出现概率。具体过程如下：

初始化：

递归公式：

终止：

1. 后向算法：性质上跟前向算法是相同的，只是递推的方向不同。给定隐马尔可夫模型且时刻t的状态为的条件下，从时刻t+1到T的观测序列为的概率为后向概率，记：

可以递归求得后向概率及观测序列出现的概率。具体过程如下：

初始化：

递归公式：

终止：

* 1. Viterbi算法

Viterbi算法采用动态规划算法，首先定义两个变量，定义在时刻t状态为的所有单个路径中概率最大值为：

变量递推公式：

定义在时刻t状态为的所有单个路径中概率最大的路径的第个结点为：

初始化：

递推公式：

终止：

最优路径回溯：对，求得最优路径

* 1. Baum-Welch算法

隐马尔可夫模型的学习问题 属于贝叶斯网络中 结构已知、数据 不完整的参数学习，主要采 用Baum-Welch算法，它是一种 期望最大化(Expectation-Maximization, EM)算法。EM算法是一种 迭代计算方法，用于计算带隐含节点 的贝叶斯网络模型参数的最大似然(MLE)或 最大后验概率(MAP)。在EM算法中，交替使用Expectation步骤 和Maximization步骤：在Expectation阶段，利用观测样本值和 当前参数值计算隐含 节点的期望值；在Maximization阶段，调整参数值最大化 观测节点值和隐含节点期望值 出现的似然概率。

A. 定义变量，表示在观测序列为条件下，时刻t状态处于并且时刻t+1处于状态的概率，可表示为：

|  |  |
| --- | --- |
|  | (2.11) |
|  | (2.12) |

其中为前向概率；为后向概率。

B. 定义变量，表示在观测序列的条件下时刻t处于状态的概率，可表示为：

|  |  |
| --- | --- |
|  | (2.13) |
|  | (2.14) |

1. 用(2.14)式表示隐藏节点值，极大化似然函数求得模型参数，可得：

|  |  |
| --- | --- |
|  | (2.15) |
|  | (2.16) |
|  | (2.17) |

反复迭代执行上述步骤，直至似然值的变化小于 某一个固定的阈值，训练结束，输出模型参数，即解决了学习问题。

## 采样方法

对于⼤多数实际应用过程中的概率模型来说，精确推断常常是不可⾏的，因此我们需要使用一些采样方法然后进行近似推断，其基本思想就是要把不能显示可解的问题转化为通过一些采样方法模拟解决的问题，随机模拟方法的核心就是如何获得一个概率分布的样本，即采样。接下来我们将介绍常用的采样方法：

### 基本采样算法

1. 简单采样方法

最简单的采样是均匀采样，也就是均匀产生[0, 1]之间的随机数，这些样本服从均匀分布。随机数的产生，在计算机中经常采用线性同余的方法来实现；正态分布也可以在均匀采样的基础上加拒绝采样实现。

1. 接受-拒绝采样

很多实际问题中，p(x)概率分布太过于复杂很难直接采样，我们可以引入易抽样的分布q(x)比如高斯分布，通过一些方法拒绝某些不合适的样本，使最终的采样效果接近p(x)分布。我们可以设一个易抽样的函数q(x)，一个常量k，使p(x)始终小于等于kq(x)，如图\*所示。

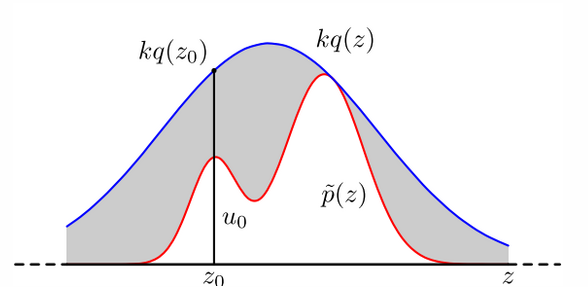


图 拒绝采样示意图

x轴方向，从 q(x) 分布中采样得到 a；y 轴方向，从均匀分布[0, kq(a)]中采样得到 u。如果u刚好落到图\*中的灰色区域即u > p(a)，拒绝本次采样，否则接受这次抽样。重复以上过程，直到采样样本达到要求为止。

1. 重要性抽样

重要性采样主要是用来计算数学期望，当原分布f(x)难以求积分且难以采样，这时候可以引入一个辅助的分布q(x)来采样，然后相应的f(x)/q(x)就作为采样点的重要性权重，也是该采样点对数学期望的贡献程度，所以叫做重要性采样。

### 马尔科夫链蒙特卡洛

无论是拒绝抽样还是重要性采样，他们都是独立采样，即上个样本与下个样本之间是无关的。这种采样方式效率比较低，比如拒绝采样所采取的样本中很可能有一些是无效的工作。马尔科夫链蒙特卡洛（Markov Chain Monte Carlo, MCMC）方法是关联采样，下一个样本与以前样本有关系，从而使得采样效率高。

马氏链收敛定理指出，对于一个非周期马氏链，具有状态转移矩阵P，而且它的任意两个状态均可以联通，状态概率分布将会收敛到一个平稳分布，与转移矩阵P满足公式。MCMC方法基于上述定理，我们可以构造一个状态转移矩阵为的马氏链，使对于给定的概率分布，该马氏链的平稳分布恰好收敛到。随机从初始状态出发，沿着马氏链状态转移矩阵P做跳转，采样得到一个状态序列，假设第n步时马氏链已经收敛，则可以选取作为的样本。

MCMC中的Metropolis-Hastings算法[43]基于马氏链的细致平稳条件，即如果非周期马氏链的转移矩阵和其状态分布，满足：

则就是该马氏链的平稳分布，上述等式称为细致平稳条件。对于给定的概率分布，具有转移矩阵Q的马氏链表示从状态转移到的概率与并不一定满足细致平稳方程，但通过引入函数满足,使得

其中，记，这样就得到了新的以为转移矩阵的马氏链，且其平稳部分恰好为。综上，MCMC算法可以描述为表2所示：

表 2 MCMC采样方法

|  |
| --- |
| **算法：MCMC采样方法** |
| 1：初始化马尔科夫链的初始状态  2：对 循环以下过程：   1. 第个时刻马尔科夫链的状态为，进行采样 2. 从均匀分布进行采样 3. 如果，则接受转移，令 4. 否则不接受转移，令 |

### 吉布斯采样

对于是高维变量的情形，由于接受率的存在（），通常M-H算法效率较低。Gibbs采样法[44]基于条件概率给出了一种接受率等于1的高效算法。首先考虑二维情形，给定概率分布，对坐标相同的两点和，有：

所以可得：

由此可见，对于直线上任意两点，若利用条件概率做转移概率，则满足细致平稳方程。同理对坐标相同的两点也有类似等式。因此可以构造如下二维的转移概率矩阵：

如果

如果

其他

上述转移概率矩阵对平面上任意两点，满足细致平稳方程：

因此，该马氏链将收敛到平稳分布。以上过程可以很容易推广到高维情形，高维情形下利用做转移概率。综上，Gibbs采样算法可以描述为表3所示：

表 3吉布斯采样方法

|  |
| --- |
| **算法：n维Gibbs Sampling方法** |
| 1: 随机初始化  2: 对 循环采样 |

## 本章小结

本章介绍了本文用到的主要技术，首先介绍了项目反应理论，主要从理论概念和模型表示这两方面概述了该理论，紧接着给出关于该模型的参数估计方法；然后介绍贝叶斯网络，包括其数学定义和模型推理以及动态贝叶斯网络问题；第三小节介绍了隐马尔可夫模型，简介其基本理论，并描述了隐马尔可夫模型中的三个基本问题：评估问题、解码问题、学习问题，接着介绍Forward-backward算法、Viterbi算法、Baum-Welch算法依次解决上述问题；最后介绍了采样方法，从基本的采样算法介绍到马尔科夫链蒙特卡洛方法，以及超高接受率（接受率为1）的吉布斯采样方法。

# 

# 数据收集与实验分析

## 6.1数据收集

### 6.1.1数据来源

### 6.1.2数据规模

## 6.2本章小结

# 算法应用

## 7.1本章小结

# 总结与展望

## 总 结

## 展 望

# 参考文献

1. 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.

# 攻读硕士学位期间取得的学术成果

1. Shanshan Wang, Wenjun Wu. App Store Analysis: Using Regression Model for App Downloads Prediction[C]. International Conference of Young Computer Scientists, Engineers and Educators. Springer Singapore, 2016: 206-220.

# 致 谢

转眼间两年半的研究生生活飞速而过，这段时间我学习到了很多，获得了导师以及师兄师姐同学的无私帮助，这里要感谢的人很多，没有他们的指导和支持与鼓励，没有我今天的成绩。

首先，我最感谢的人就是我的导师吴文峻教授，是他将我带入到了北航计算机学院的大家庭里，给我一个优秀的学习氛围与资源平台。吴老师严谨的治学态度与渊博的计算机学术知识，无不使我受益匪浅。在做毕设途中，吴老师总是抽出宝贵的时间给我梳理研究方向和研究方法，帮助我查找相关学术论文，引导创新与实现一个个阶段目标。从选题到毕业设计完成这段过程中，吴老师对我的帮助太多。在此，谨向吴老师表达我最大的敬意与谢意！！！

其次感谢我实验室我的师兄楼奕华、郭帅、韩勇，他们对我研究生学习与生活中，无不起到带头与指导的作用，不仅给我学习与毕设上莫大的帮助与引导，在就业与生活上也有许多的信息共享。再次感谢我的实验室的其他成员胡征慧、王鑫、周萱、黄剑琨同学，在我对生活产生迷茫不安的时候及时带我走出思想混沌的阴霾，以及我的室友江丽丽、樊鑫鑫、刘美媛，感谢她们对我学习生活上的帮助与包容，感谢那段寝室晚上探讨计算机学术之路以及生活方方面面的“卧谈会”的美丽时光，感谢她们陪伴我度过这段美好的研究生生涯。

最后，我要感谢我的家人，没有他们的经济上的支持，就没有我的研究生求学之路，感谢我的父母对我的二十余年的培养，他们是我一生感恩的人。再多的感谢也表达不了我对父母的养育之恩，唯有不辜负他们的期许，做一个对社会有用的人才。祝愿我的父母身体健康，事事顺心！！！

向所有帮助我的人，献上最真挚的祝福！