基于贝叶斯知识跟踪模型的慕课学生评价

王卓，张铭
（北京大学信息科学技术学院，北京 100871）

摘要：如何让学生能够更加高效地利用慕课（Massive Open Online Courses，简称 MOOC）上的资源历来就是一个重要的问题，实际上，MOOC 平台的高用户流失率正反应了这方面的不足。如果能够研究出一个适当的学习者模型，用于对学生学习方面进行建模和解释，将对进一步提高教学水平起到重要的帮助。而这正是现在 MOOC 所缺少的。贝叶斯知识跟踪模型（BKT）早在上个世纪九十年代便已有人提出，当时被应用在智能教育系统（ITS）上。这套模型很好地模拟了学生的学习过程，一直沿用至今。本文将贝叶斯知识跟踪模型应用到 Coursera，详细分析出 MOOC 的数据特点，据此对 BKT 变形，先后提出 Aspect-BKT 和 History-BKT 两套模型，并进行实验检验，根据实验结果分析 BKT 参数含义。实验结果表明，History-BKT 与前人提出的 IDEM-Count 模型有可比的预测效果。

关键词：大规模在线公开课程；贝叶斯知识跟踪模型；学生评价；隐马尔科夫模型
中图分类号：TP311

MOOC Student Assessment Based on Bayesian Knowledge Tracing Model

WANG zhuo, ZHANG Ming
(School of Electronics Engineering and Computer Science, Peking University, Beijing 100871)

Abstract: How let people use the resources on the MOOCs (Massive Open Online Course) efficiently becomes an important issue. In fact, a high churn rate on MOOC reactions on the deficiency. If we could come up with an appropriate learner model, aiming to model and explain the process of learning, it will be helpful to teaching and learning. However, this is now what MOOC lacks. We noticed that Bayesian knowledge tracing model proposed early in the nineties had been applied to the intelligent tutoring system. It models the students' learning process well, and still in use now. This paper introduces the Bayesian knowledge tracing model to Coursera, and proves the validity of this model on the MOOC. Besides, we analyze the characteristics of the data on MOOC in detail, and propose two novel models named as Aspect-BKT and History-BKT respectively. According to the analysis of experimental results, we try to explain the meaning of BKT parameters. Experimental results show that History-BKT has made a comparable prediction with the IDEM-Count model that was promoted before.

Key words: MOOC, Bayesian Knowledge Tracing Model, Student Assessment, Hidden Markov Model

0 引言

MOOC 平台 Coursera 的目标是：将人们和卓越的教育资源联系起来，让世界上任何一个人都可以没有障碍的去学习。然而，虽然这一平台提供了优质的教学资源，但是它也存在一些问题：缺少对学生的监督，助教和教师相对于注册学生过少等等，比如，北京大学 2013 年秋季开设的“数据结构与算法”课程总共有 13000 多人注册，但是教师和助教的教学团队只有十人左右。传统的课堂上，教师可能能够随时注意到每名学生的学习状况，并能进行有

基金项目：自然科学基金项目（项目编号：61472006）和博士点基金（项目编号：20130001110032）
作者简介：王卓（1991-），男，汉族，湖北，北京大学网络与信息系统研究所硕士
通信联系人：张铭（1966-），教授，博导，ACM Education Council 惟一的中国委员兼 ACM 中国教育专委会主席，科研方向为文本挖掘、社会网络分析、教育大数据挖掘，所主持的“数据结构与算法”课程被评为国家级和北京市级精品课程. E-mail: mzhang@net.pku.edu.cn

针对的辅导和帮助。而在大规模在线公开课程中，这一点显得十分困难。

实际上，MOOC 正面临着学生的高流失率的问题[1]。根据 Coursera 的统计数据[2]，在2012 年注册的所有用户中，仅有 50-60%的学生在注册课程后，回到课程中浏览了第一讲内容。而在需要编程和同伴评阅的课程中，提交作业的人数仅占浏览课程的总人数的 15-20%；在提交过作业的人中，仅有 45%的学生最终完成课程并获得了证书。也就是说在一门普通的 MOOC 课程中，最后完成课程的人数只占 5%。

当然，我们可以预计到一部分学生能够凭自己的努力，充分利用 MOOC 上的课程资源，达到自己的学习目的。而另一部分学生可能会因为自己一时的懈怠，最终无法实现自己的学习目标，不免十分遗憾。我们如果能够及时发现这一部分学生，给他们提供有针对性的帮助，从而激发学生的学习动力，使得 MOOC 上的教育资源得到更加有效的利用。另一方面，教师也可以根据学生们整体学习状况的信息反馈，发现自己教学过程中存在的问题，并进行适当调节，比如增加某一部分的习题讲解等。

在这方面有一个很好的例子就是普渡大学的 Course Signal 系统[1]，利用从最基础的考试分数、排名，到学习过程，甚至监测到学生与电子学习系统 Blackboard Vista 的互动情况，对学生的学习情况进行更全面的考量，并将结果按照红、黄、绿的分组返回给学生。通过普渡大学去年九月发表的报告来看，在两门或更多课程中使用了 Course Signal 的学生要比那些没有使用这套系统的学生毕业率提高了 21.48%（六年制）。目前，将近 24000 名学生参与这个项目，超过 145 名普渡大学的教授在他们的课程中使用这套算法。普渡大学的做法利用了“霍索恩效应”。所谓的“霍索恩效应”是指人们如果得知自己正在被研究或监测，行为表现就会有所增强。引申到学习领域，如果学生能够收到更多关于他们当前状态的反馈，他们分数就会更高。

本文主要研究如何更准确地推断学生是否掌握各个知识点，内容组织如下：第二部分介绍了相关研究工作，包括已有研究如何对 BKT 进行改造。第三部分介绍 MOOC 的数据特点以及如何对 BKT 进行改造使之能够合理的应用于 MOOC 之中，提出了基于知识点方面的 Aspect-BKT 和提交历史的 History-BKT。第四部分是实验结果和分析，比较了 BKT 和相关变形的预测效果，发现 BKT 的三个变形都比基本的 BKT 有显著提高，其中 Aspect-BKT 稍逊于前人提出的 IDEM-Count 模型，而 History-BKT 则与 IDEM-Count 的效果没有显著区别，在后面几次小测中 History-BKT 甚至表现更好。最后，对相关参数进行了进一步的实验和解释。第五部分对本文进行总结以及展望未来工作。

1 相关研究

1.1 贝叶斯知识跟踪模型的提出

贝叶斯知识跟踪模型(Bayesian Knowledge Tracing Model, 简称 BKT)，是模拟学生知识的一个很重要的模型，由 Corbett 和 Anderson 于 1995 年引入智能教育领域[3]，应用于智能教育系统(Intelligence Tutoring System, 简称 ITS)。在 ITS 中的一个重要问题是，什么时候能够判断这个学生掌握了这个知识点。一个比较简单的处理方式是要求学生连续对 N 个同一知识点相关的题目回答正确，当然这种方式现在仍然被某些系统利用。而 BKT 能够用一种更加直观而容易理解的方式解决这个问题。

[1] http://www.itap.purdue.edu/learning/tools/signals/

1.2 贝叶斯知识跟踪模型的基本原理

BKT 将学生所需要学习的知识体系划分为若干个知识点。而学生的知识状况则被表示为一组二元变量，每个二元变量表示其中一个知识点是否被掌握，即学生处于“知道这个知识点”和“不知道这个知识点”两种状态之一。这是一种将学生的知识状态作为一套隐含变量的表示方式。通过学生回答问题的正确与否来更新隐含变量的概率分布。也就是说，观测变量同样也是二元的。

表 1 BKT 的四个参数
Table 1 Four Parameters of BKT

参数	说明	解释
$p(L_0)$	Initializing Learning	学生最初掌握知识点的概率
p(T)	Acquisition	学生从不会到会的转移概率
p(G)	Guess	学生在不会的状态下，仍然猜对的概率
p(S)	Slip	学生在会的状态下，仍然做错的概率

具体来说，BKT 假设对于知识存在四个参数，见表 1。

其中 p(L0)和 p(T)是知识参数，主要用于表示学生的学习状态。p(L0)指的是学生在尚未接触这个学习系统时，这个知识点就已经被其掌握的概率。p(T)是指学习效率，即经过了一些学习机会之后，对于这个知识点从不懂到懂的转换概率。另外 BKT 假设学生不会遗忘，也就是说，对于一个知识点从懂到不懂的转换概率为 0.

而 p(G)和 p(S)作为用户的表现参数。p(G)是猜对的概率，即学生即使不知道这个知识点仍然正确回答的概率。P(S)是犯错的概率，即学生知道了这个知识点，但是仍然不小心回答错误的概率。当 p(G)和 p(S)为 0 时，学生回答问题的结果，将会 100%反应学生掌握这个知识点的情况，而当 p(G)和 p(S)为 0.5 时，学生回答问题的结果所反应的知识状况具有最大的不确定性。

不同的知识点在难度和所需要的练习上有很大的差别，所以各个知识点需要分别训练这样的一套参数。

根据以上对于参数的定义，我们很容易得到以下几个公式。

$$p(Correct_n)=p(L_n)p(\neg S)+p(\neg L_n)p(G) \quad (1)$$

即，学生答对题目的概率被解释为在知道知识点的情况下没有犯错，以及在不知道知识点的情况下猜对的概率之和。

另外，

$$p(L_n)=p(L_{n-1}|Evidence_{n-1})+(1-p(L_{n-1}|Evidence_{n-1}))p(T) \quad (2)$$

即，学生不会遗忘，而按照 p(T)的学习效率加强对知识点的理解。

另外，可以在训练好参数之后，根据数据对知识状况进行推断。利用如下公式：

当题目答对时：

$$p(L_n|Correct_n)=p(L_n)p(\neg S)+p(\neg L_n)p(G) \quad (3)$$

当题目答错时：

$$p(L_n|Incorrect_n)=p(L_n)p(S)+p(\neg L_n)p(\neg G) \quad (4)$$

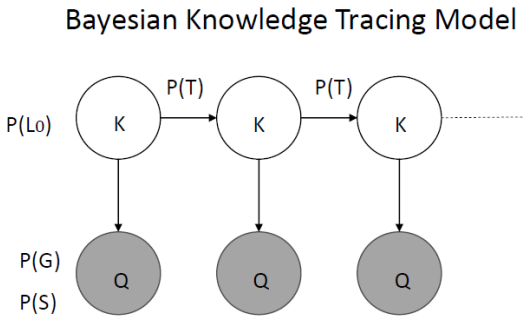

图 1 贝叶斯知识跟踪模型（BKT）
Fic.1 Bayesian Knowledge Tracing Model(BKT)

不难看出，BKT 其实就是一种特殊的隐马尔科夫模型(Hidden Markov Model, 简称 HMM)。相应的图示见图 1。用户在学习过程中的学习状态是随着时间变化的。

从 HMM 的角度来看，相应的参数可以用表 2 表示。

1.3 贝叶斯知识跟踪模型的变形

原始的 BKT 所有学生的共有相同的参数。其中与学生个人最相关的参数是学生初始知识状况 p(L0)和学习速率 p(T)。也就是说 BKT 假设所有学生的初始知识状况是相同的，同时学习速率也相同。这个假设让模型得到了很大的简化，但是难免让人惊讶，尤其在 MOOC 的环境之下。学习同一门课程的学生可能差别巨大，既有本科生，研究生，还有尚未参加过大学学习的学生。比如，Coursera 上对于“数据结构与算法”课程 600 多名学生的最高学历调查结果：6%为博士学位，29%硕士学位，41%的学士学位，8%高中学历。

表 2 HMM 相关参数
Table 2 Parameters of HMM

初始隐状态分布矩阵
The Initial Hidden State Distribution Matrix

不知道知识点	$p(L_0)$
知道知识点	$1-p(L_0)$

隐状态转移矩阵
The Hidden State Transition Matrix

	后来不知道知识点	后来知道知识点
原来不知道知识点	1-p(T)	p(T)
原来知道知识点	0	1

观测矩阵
The Observation Matrix

	回答错误	回答正确
不知道知识点	1-p(G)	p(G)
知道知识点	p(S)	1-p(S)

1.3.1 学生个体差别

Pardo 等人对学生的先验进行过详细研究，提出了各学生先验模型(Prior Per Student, 简称 PPS)[4]，对于如何设置各个学生的先验知识，他提出了许多启发式的方式。关键点是初始值设为多少，以及是否让其在训练的过程中进行调节。最后他发现，对于大部分的实验数据，PPS 的效果比基本的 BKT 要好。

Michael 等人针对个性化的先验知识和个性化的学习效率进行了比较[5]。他们以 0.5 作为

临界值，比较不同模型的预测结果的准确率，发现个性化的学习效率比个性化的先验知识要更好。

1.3.2 题目难度差别

也有研究将不同题目的难度引入模型，从而使得模型能够分析各个题目的难度，并提高预测的准确率，这也就是问题难度效果模型(Item Difficulty Effect Model，简称 IDEM)[6]，在这个模型中，不同的问题各自训练自己的 p(G)和 p(S)。p(G)高、p(S)低的题目被认为是容易的。而 p(G)低、p(S)高的题目被认为是困难的。实际上，在 MOOC 的小测中，不同题目之间的区别也确实是十分明显的。有的单项选择题，即使随机猜测也有 25%的准确率；而有的填空题，只有经过复杂而准确的计算，才能得到正确的结果。

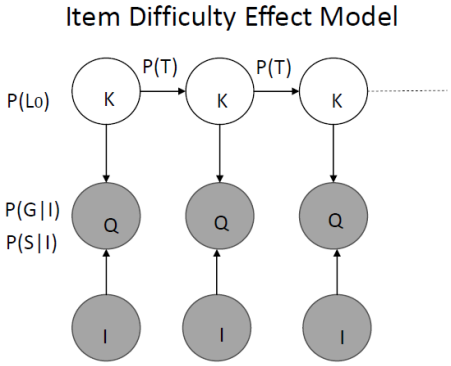

图 2 问题难度效果模型（IDEM）
Fig.2 Item Difficulty Effect Model (IDEM)

如图 2 所示，IDEM 实际上是通过增加一个问题节点 Item，训练不同问题下的猜测概率和犯错的概率，即 p(G|I=1),p(G|I=2)等等。根据 Pardos 等人的研究，如果数据不足，由于参数过多，可能效果反而不如基本的 KTM，而当每个问题的数据量足够多的时候，IDEM 的准确率显著更高。

2 MOOC 上的贝叶斯知识跟踪模型

2.1 课程背景介绍

表 3 北京大学 2013 秋季学期 Coursera 两门编程课程的数据对照
Table 3 Data Contrast of Two Programming Courses of Peking University in 2013 Fall Semester on Coursera

	数据结构与算法	计算概论
课程长度（周）	14	12
作业次数	14	15
平时小测次数	16	1
同伴互评作业	0	0
期中考试	有	有
期末考试	有	有
注册人数	13170	13890
获取证书人数	57	286
获取证书比例	0.004328	0.02059
帖子数	283	859
回帖数	824	3760
评论数	354	1934

2013 年秋季学期，北京大学在 Coursera 上开设了 6 门课程。其中，通过向 Coursera 申请和利用助教权限，本文作者能够很完整地获得“数据结构与算法”和“计算概论”这两门

课程的数据。这两门课的比较如表 3 所示。

“数据结构与算法”课程会每周定期发布教学视频和讲义，并布置作业和小测。课程小测利用的是 Coursera 提供的平台。助教事先设置好问题和答案，学生提交之后，系统自动把学生答案与正确答案进行匹配，并用绿色的勾代表正确和红色的叉代表错误反馈给学生，然后计算这次小测的得分。其中，有以下几点值得我们注意。第一，同一个小测可以提交多次，而取其中的最高分作为这次小测的得分。第二，系统的小测结果的反馈是即时的，所以不排除有学生尝试答案的行为。不过，大部分测验设置了 5 次或者 10 次的提交次数限制，另外两次尝试之间有 10 分钟的重试延迟，这在一定程度上缓解了这个问题。

2.2 数据预处理

Coursera 提供的数据包括三个部分。一个部分是以数据库导出文件的形式提供的，这部分数据包括作业相关的学生成绩，作业统计信息，小测相关的学生成绩，小测发布时间，论坛相关的学生发帖内容，论坛标签等。第二个部分是点击流数据。主要包括学生点击课程页面，点击课程视频的记录等[7]。

第三个部分也是本文主要利用的数据。BKT 需要利用详细到用户提交的每道题的正误情况，而这是前面数据中不具备的。作者利用助教权限直接导出学生提交小测的详细内容，以及评分用的答案，通过匹配，对所有学生进行打分。

2.3 BKT 针对 MOOC 的调整

广泛使用了 BKT 的 ITS 系统能够随时向学生提供帮助，而在 MOOC 系统中，学生需要自己从课程里这些丰富的学习资源中摸索。实际上，MOOC 上的许多特点为学习者建模提出了挑战。

2.3.1 缺少知识点的划分

智能教育系统的一个优势是，这个软件或者系统在设计之初，已经对知识点做了比较详细的规划。各个题目已经事先被标记与某一个知识点相关。这些是由相关学科的专家完成的。系统能够根据学生历史上对于问题的回答来推断其掌握知识点的程度。

但是 MOOC 并没有这样的专家，尤其是存在着这么多形式各样的课程，通过领域专家的详细划分将是一个浩大的工作。

Pardos 等人提出直接利用 MOOC 本身的结构来解决这个问题[8]。MOOC 课程每周都是一个时间单元，每周都有各自的学习内容，包括提供视频、课件，同时小测和作业也基本上是每周一次。所以一个简单的做法是直接将一周的内容作为一个知识点，假设每次小测的若干道题目都是围绕同一个知识点的。这样做的合理之处在于，授课教师确实会围绕某个问题设计一周的课程。但是这样显然粒度太粗，并不算好的做法。而且，对于一个知识点的研究也就只能是针对那一周的内容了，时间上跨度比较短。不过，这样做的好处是，对于 MOOC 上的不同领域的各个课程都能采用这种方式。

表 4 Aspect-BKT 观测矩阵

Table 4 The Observation Matrix of Aspect-BKT

	回答错误	回答正确
不知道知识点，会应用	1-p(G)	p(G)
不知道知识点，不会应用	1-p(G)	p(G)
知道知识点，会应用	p(S)	1-p(S)
知道知识点，不会应用	1-p(G)	p(G)

针对这一点，本文提出了一个 Aspect-BKT 模型（知识点多方面的知识跟踪模型），如图 3 所示。基本的 BKT 用一个知识点概括一周的内容，也就是假设只要掌握了这个知识点，这一周知识的各个方面均已被掌握，这也是 BKT 模型的基础。不过，BKT 知识点的粒度太粗，这个假设需要弱化。也就是说，即使掌握了这个知识点，由于这个知识点存在多个方面(Aspect)，学生还需要一个将知识点运用到小测中各个题目的实际问题的能力，而这个主要受各个问题具体难度的影响。于是，各个题目结点可以由一个 Apply 结点来影响，这个结点同样取值为 0 和 1，反应的是不同题目之间的难度区别，p(A=1)越高，表示题目越简单，相应的观测矩阵见表 4。一个比较简单的方式是直接统计各个题目对于所有提交情况，总体的准确率作为 p(A=1)。这个方式，也是本文后面实验采用的。

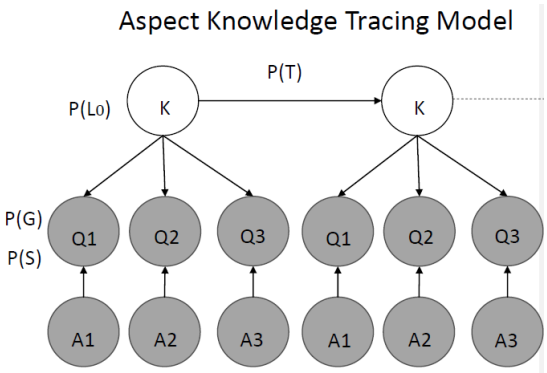

图 3 知识点多方面的知识跟踪模型（Aspect-BKT）

Fig.3 Bayesian Knowledge Tracing Model with Multi Aspects of Knowledge(Aspect-BKT)

2.3.2 多次重复提交

正如前面提到，同一次测验允许多次提交，而且取其中的最高分作为这次测验的成绩。实际上，本文正是把每次提交作为 BKT 的一个时间点，各个时间点各自有一个学生掌握知识点的概率 $p(L_n)$。而且一方面小测结果的反馈可能帮助学生的学习，另一方面，在小测的两次提交之间，学生还可能会有一些学习的行为，于是 $p(L_n)$会随时间点相应的变化。

针对多次重复提交的问题，Pardos 等人提出提交次数反应了这个题目的难度[8]。也就是说，难度大的题目可能提交次数会比较多。于是他提出了 Count 模型，也就是根据不同提交次数，分别训练 p(G)和 p(S)。之后，还将 Count 模型和 IDEM 结合起来，针对不同的题目和不同的提交次数分别训练 p(G)和 p(S)，得出 IDEM-Count 模型。当然这样的增加参数的方式，对数据量有更高的要求。

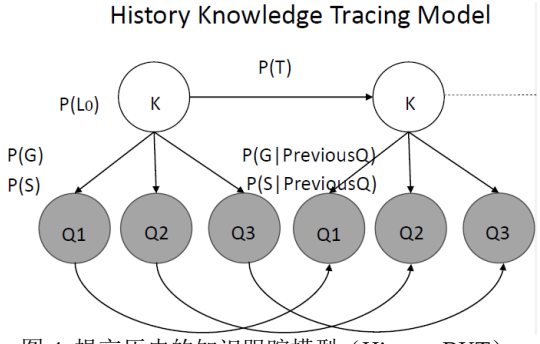

图 4 提交历史的知识跟踪模型（History-BKT）

Fig.4 Bayesian Knowledge Tracing Model with Submission History (History-BKT)

针对这一点，本文提出了一个基于提交历史的 BKT，即 History-BKT，如图 4 所示。基本想法是，虽然 Coursera 允许同一个题目在不同尝试之中有一些变化，比如，计算数据的

改变。但是，不得不承认，它与前一次提交的题目存在紧密的联系。前一次回答正确的题目，这一次很有可能回答正确；前一次答错的题目，则这次可能因为新的学习，而将其回答正确。简单来说，针对前一次答题情况分别训练 p(G)和 p(S)。共有三类，第一次提交，前一次答对和前一次答错。不同情况，回答正确的概率分布会有所不同。

2.4 算法过程

本文利用 EM 算法实现了 BKT 及相关变形模型的参数估计。其中在 E-Step，需要使用前向-后向算法做精确推理，而 M-Step 则通过求导重新估计参数值。具体推到参考[9]关于 HMM 的相关章节。

此外，参数的初始化很大程度上受到具体小测内容和学生群体状态的影响，并没有特别有效的方式，由于这部分也不是本文的研究重点，本文仅仅是根据之前参数训练情况，人为的定义 p(G)=0.1; p(S)=0.15; p(T)=0.1; p(L_0)=0.1;

Algorithm 1 EM 算法训练参数

Input 学生答题正误情况

Output 参数 L_o, G, S, T

开始

1. 初始化参数 L_o, G, S, T
2. 读入用户数据，将用户随机分为 5 组，进行交叉验证
3. While 没有达到迭代上限：
4. 根据公式依次计算前向变量 α 和后向变量 β，进而计算 L_i (E-step)
5. 根据公式更新参数 L_o, G, S, T (M-step)
6. 计算似然值 p(X)
7. If 与前一次的似然值之差小于阈值
8. break

结束

图 5 EM 算法伪代码
Fig.5 the Pseudo Code of EM Algorithm

另外，终止条件是，似然值不再变化，或者超过一定的迭代次数。实际上，后面的实验都是使用了后者作为终止条件。

3 实验及评价

3.1 实验数据集

如前所述，本文使用的数据是 2013 年秋北大开始的“数据结构与算法”和“计算概论”两门课程的数据，见表 5。主要利用所有小测的学生提交的正误情况。其中“数据结构与算法”共有 16 次小测，不过最后两次小测是选做内容，数据量过少，不作考虑。而“计算概论”只有一次小测，是关于指针方面的考察，不过这次小测题目比较多，共有 15 题，提交人数也比较多。整体上说，“数据结构与算法”课程的小测的用户提交量，呈现明显的递减趋势。第一次小测有 1028 名学生提交，总共提交 1839 次；第十四次小测有 58 名学生提交，总共提交 294 次。一方面是因为部分学生没有坚持学习这门课程而中途流失了，另一方面，相对于前面几次小测，后面的小测难度有所加大，题目有所增多。

表 5 数据量描述
Table 5 the Description of Data Amout

Quiz id	#Submission	#Students	#Questions
0	886	351	15
1	1839	1028	6
2	1128	651	6
3	708	361	6
4	603	278	5
5	505	205	8
6	391	164	5
7	345	116	12
8	236	90	10
9	171	85	9
10	208	82	8
11	202	72	7
12	156	69	7
13	282	65	7
14	294	58	8

3.2 评价标准

本文将参加小测的学生随机分为 5 组，进行 5 折交叉验证。依次取其中一组作为测试数据，而其他四组作为训练数据，训练参数。最后将 5 次测试的结果取平均值。相对于直接选取测试集和训练集，通过这样的方式增加结果的可信度。

对于每一次实验，作者将学生最后一次提交结果的数据作为待预测数据，利用训练集训练参数，并用测试集的学生前面几次提交结果推测最后一次提交的时候掌握知识点的概率。

度量标准是 Area Under the Curve(简称 AUC)，这种评价标准对于二分类问题是一个十分有效的方式，而且对于我们这种离散的情况，能够精确的计算结果。对于所有题目的最后一次提交的结果，把回答正确看作正例，错误回答看作负例。对于各个题目，可以分别枚举所有的正例-负例对。对于这样一对，我们分别有一个预测值，检查对于这样的预测值，是否满足正例的预测值高于负例的。并计算满足的比例，作为 AUC 值。所以，当 AUC 为 1 时，表示负例的预测最大值小于正例预测的最小值，分类效果最好，当 AUC 为 0 时，表示正好将所有结果预测错误，当 AUC 为 0.5 时，表示预测效果和随机分类一样。

3.3 实验结果

如图 6 及表 6 所示，IDEM 相对于 Basic 有 0.09375 的提高，而 Aspect 相对于 Basic 有 0.07382 的提高。Aspect 模型比 IDEM 要差约 0.02（p-value=0.0257）。History 和 IDEM-Count 没有显著区别，p-value=0.5739。不过，值得注意的是对于后面几次数据不足的小测，History 的表现相对更好，去掉前面 8 次小测的数据，History 的均值要显著高 0.04（p-value=0.048）。而对于前面 8 次小测的数据，两者没有明显区别，p-value=0.4138。

也就是说，通过知识点在不同类型上的运用难度不同而建立的 Aspect 虽然从直观上比较容易理解，恐怕直接使用题目的全局正确率的方式有待商榷。而 History 表现不错，尤其是当数据较少的时候，仍然能够保持不错的准确率。

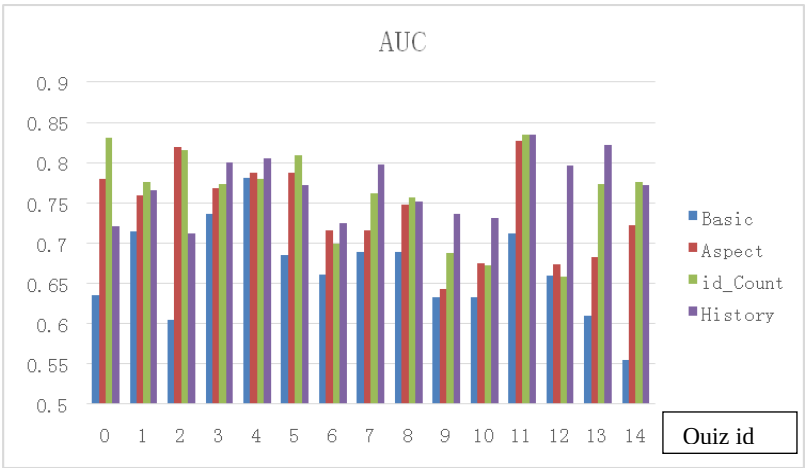

图 6 多模型 AUC 比较

Fig.6 AUC Comparison of Multiple Models

表 6 多模型 AUC 均值比较

Table 6 Average AUC Comparison of Multiple Models

模型	Basic-BKT	Aspect-BKT	IDEM-Count	History-BKT
AUC 均值	0.666611	0.740434	0.76036	0.769537
相对于 Basic 的 p-value	-	0.0004	0.0002	<0.0001

3.4 实验结果分析

3.4.1 关于 p(G)和 p(S)的解释

这里，作者进一步分析题目的全局正确率和 IDEM 训练出来的不同 GS 之间的关系。

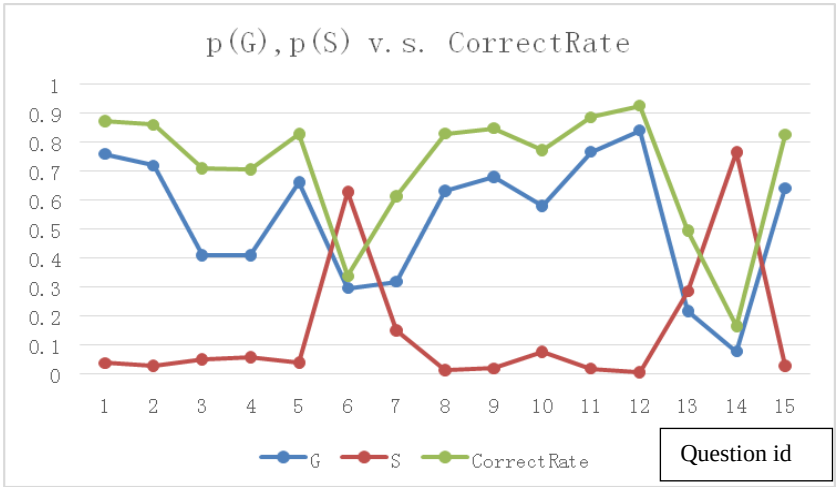

图 7 p(G)和 p(S)特点

Fig.7 the Features of p(G) and p(S)

作者拿“计算概论”的小测的 15 道题目进行实验，如图 7，这里，横坐标是 15 个题目的编号。数据来自于其中一次利用训练数据训练出了的 G 和 S，以及相应的每道题目的学生正确率。可以看出明显同步的变化趋势，也就是说题目越容易，则越容易猜对，越难犯错，即 G 越大，S 越小，同时，学生回答的准确率越高。Aspect 利用准确率 Correct Rate 的合理性就在这里，IDEM 直接让机器训练不同的 G 和 S 参数的方式更加合理，效果也更好。

3.4.2 History-BKT

History-BKT 表现较好，这与模型的建立方式本身有关。我们把每次提交一次小测作为一个时间点，虽然 Coursera 允许同一道题目在不同的尝试中有变化，但是一般至多修改计算用的数据等。所以前一次回答正确的题目，下一次提交确实很容易正确。

实际上，针对“计算概论”的 5 组交叉验证的小测的数据结果，如表 7 所示。仍然正确表示所有学生所有题目里面前一次提交的答案正确，这次仍然正确的频数，突然犯错表示前一次回答正确，但是这一次答错。仍然犯错和突然正确类似。通过计算比例，可以看出对于某一次提交正确的同学来说，下一次对于同一道题目有高达 92%以上可能性会仍然正确。正是由于不同次提交之间存在这么显著的前后关系，History-BKT 的效果才会那么好。

如表 8 所示，同样利用“计算概论”的小测的数据进行训练得到的一组实验结果。可以具体分析训练出来的参数，G 的区别相对不大，但是原来回答正确的 S 明显更低，也就是说，原来回答错误的题目，下一次仍然很容易做错，但是原来正确，下一次很难犯错。正是因为这种数据特点的有效性，才能使得 History-BKT 有效。而且，这个特点与数据量大小无关。

表 7 History-BKT 分析
Table 7 Analysis of History-BKT

编号	仍然正确	突然犯错	仍然犯错	突然正确	突然犯错比例	突然正确比例
1	4126	354	1188	1112	0.079018	0.483478
2	3545	298	880	947	0.077544	0.518336
3	3960	323	1089	1033	0.075414	0.486805
4	4037	340	1150	1088	0.077679	0.486148
5	4072	345	1097	1116	0.078107	0.504293
平均	3948	332	1081	1059	0.077552	0.495812

表 8 History-BKT 参数分析
Table 8 Parameter Analysis of History-BKT

	原来回答错误	原来回答正确	第一次提交
G	0.2393166	0.291944	0.388292
S	0.450365	0.049532	0.319725

4 总结及未来工作

本文将原本广泛应用于 ITS 系统的 BKT 应用到 MOOC 上，并根据 MOOC 的数据特点进行相应调整，分别提出 Aspect 和 History 模型，并与现有模型进行比较。

实验表明，Aspect 模型还是过于粗糙，比不上现有的 IDEM-Count 模型。而 History 模型取得了不错的效果，这也是 MOOC 的数据特点所致。

总的来说，前面将一周的知识点作为一个整体知识点的定义方式仍然不够理想，如何更细致的模拟学生的知识状况值得我们思考。而且，后续工作可以考虑针对 p(T)参数的变化。而这个可以利用丰富的学生行为数据，通过分析学生在两次小测之间的行为，来影响相应的 p(T)值。

[参考文献] (References)

[1] Anderson A, Huttenlocher D, Kleinberg J, et al. Engaging with massive online courses[A]. In: Proceedings of the 23rd international conference on World wide web[C]. International World Wide Web Conferences Steering Committee, 2014: 687-698.
[2] Koller D, Ng A, Do C, et al. Retention and intention in massive open online courses: In depth [J]. Educause Review, 2013, 48(3).
[3] Corbett A T, Anderson J R. Knowledge tracing: Modeling the acquisition of procedural knowledge[J]. User

modeling and user-adapted interaction, 1994, 4(4): 253-278.
[4] Pardos Z A, Heffernan N T. Modeling individualization in a bayesian networks implementation of knowledge tracing[A]. In: User Modeling, Adaptation, and Personalization[C]. Springer Berlin Heidelberg, 2010: 255-266.
[5] Yudelson M V, Koedinger K R, Gordon G J. Individualized bayesian knowledge tracing models[A]. In: Artificial Intelligence in Education[C]. Springer Berlin Heidelberg, 2013: 171-180.
[6] Pardos Z A, Heffernan N T. KT-IDEM: Introducing item difficulty to the knowledge tracing modelA]. User Modeling, Adaption and Personalization[C]. Springer Berlin Heidelberg, 2011: 243-254.
[7] Veeramachaneni K, Dernoncourt F, Taylor C, et al. Moocdb: Developing data standards for mooc data science[A]. In: AIED 2013 Workshops Proceedings Volume[C]. 2013: 17.
[8] Pardos Z A, Bergner Y, Seaton D T, et al. Adapting Bayesian Knowledge Tracing to a Massive Open Online Course in edX[A]. In: EDM[C]. International Educational Data Mining Society 2013: 137-144
[9] Bishop C M. Pattern recognition and machine learning[M]. New York: springer, 2006.