MOOC 评价系统中同伴互评概率模型研究

孙 力 钟斯陶

(江南大学 人文学院,江苏无锡 214122)

[摘要] 2012 年起,基于网络、针对大众人群的大规模开放在线课程呈井喷式发展。目前的 MOOC 虽然能支持视频课程、论坛、测试和评价等功能,但对于学习者学习效果的评价和给予反馈的能力仍受到限制。在线学习的学习效果评价方法中,选择题和判断题等客观类试题可以通过机器进行评判反馈,但一些主观类试题,比如数学演算、设计问题和论文等一些复杂和开放性的作业任务就无法通过机器评判反馈。针对这一情况,一些 MOOC 平台正逐步引入同伴互评机制。虽然同伴互评机制的设立使得主观类试题得到有效评价,但学习者对同伴建议的准确性和权威性表示怀疑。调查发现,94% 学生更喜欢老师评语。如此,需要依据一定的理论或过程模型保证同伴互评的准确度、信服度和价值。本文构建了三种关联复杂度不同的同伴互评概率模型来提升 MOOC 评价系统中主观试题评分的客观性和准确性,并利用 Cousera 中"人机交互"课程的相关数据组来评测各同伴互评概率模型的准确度。评测方法采用了吉布斯采样法和期望最大化法。文章通过对使用三种概率模型得到的评测结果与通过 Cousera 平台同伴互评系统所得到的相应结果进行了比较,结果显示,准确度有显著提高。本文构建的模型可以提升同伴互评系统整体效果,且最高达到 30%。文章最后还对同伴互评概率模型的进一步改进方向和其在 MOOC 系统中的实际应用进行了探讨,包括增加新的关注参数,例如评分者在评分时投入的关注度等。

[关键词] MOOC;同伴互评;概率模型;评分者可靠度;评分者偏差

[中图分类号] G424.75 [文献标识码] A [文章编号] 1007-2179(2014)05-0083-08

一、引言

信息技术的迅猛发展推动了一个又一个传统行业的变革。教育领域也开始在互联网等新技术的影响下悄然发生着转变,例如网络教育、游戏化学习、虚拟社区与现实课堂有机结合的新型教育模式的涌现,数字化学校、数字化教师、网络课堂、远程学习、在线教育、云教育、云计算、大数据等虚拟化、扁平化的交互式学习平台,以及游戏化学习、因材施教、反转式课堂等新名词的诞生(周洪宇等,2013)。

2012 年,基于网络、针对大众人群的大规模开放在线课程(Massive Open Online Courses, MOOC)井喷式涌现。一般认为,MOOC 课程模式起源于基于互联网的开放课程,最早可追溯到 2007 年(陈肖庚等,2013)。通过这种互联网上的视频课程,学习者可以学习世界各国知名院校的一流课程,获得一流的教育,更好地迎合第三次工业革命对高素质劳动者和创新型人才的需求(郝丹,2013)。在国外,可汗学院的兴起引发了 MOOC 革命。随后,斯坦福大学、哈佛大学、麻省理工学院也开始跟进这种教学方式(周洪宇等,2013)。这些名校创办了有国际在线教育三驾马车之誉的 Coursera、Udacity 和 edX。Coursera 仅推出数月,注册学习者数就激增至 100 万。回望我国,2013 年也是 MOOC 兴起的重要一年,各大学和教育机构推出了相应的 MOOC 课程。例如,上海交通大学首次推出"南洋学堂"大规模开放在线课程。"南洋学堂"取名于上海交大的前身南洋公学,目前开设的课程有郑益慧教授的"模拟电子技术",刘西拉教授的"21 世纪的工程人才:知识,能力和素质",黄钢教授的"医学与绘画艺术"等。类似的的还有网易公开课推出的"中国大学视频公开课",集结了许多高校的优质课程。如南开

[收稿日期]2014-05-20 [修回日期]2014-08-26

[作者简介]孙力,博士,江南大学人文学院教授,继续教育与网络教育学院副院长,研究方向:网络教育系统的构建及开发(lisun@jiangnan.edu.cn);钟斯陶,江南大学人文学院硕士研究生。

大学的“六大名著导读”和“小词中的修养境界”,浙江大学的“王阳明心学”等。这些课程均以视频录像呈现,课后有相应讨论区,供师生进行互动讨论。除大学的大规模开放在线课程外,一些机构也推出了自己的MOOC,例如沪江网校的CCTALK。与高校MOOC不同的是,CCTALK在线直播课程分课程内容展示模块(多以PPT呈现)以及讨论模块,部分教师课后会发布与课程内容相关的练习题,但习题往往无法得到批阅因而不能对学习者的学习效果进行反馈。

当网络技术允许人们在MOOC系统中用扩展性的方式去发布视频教学内容,实现论坛交流和追踪学习者的进展时,对于学习者学习效果的评价和给予反馈的能力仍然受到限制。而缺乏有效的学习效果评价和反馈机制应该是制约MOOC快速发展和使其缺乏有效吸引力的主要瓶颈之一。当前,一些主要的MOOC平台正在尝试采用多元评价模式,即形式多样的形成性评价和终结性评价相结合的方式来建立有效的在线学习效果评价机制。比如,Coursera的“人机交互”(Human Computer Interaction)课程,每周要求学生提交作业,并已建立了实用的同伴互评系统评价作业。edX的“6.002x”课程的最终成绩由家庭作业(15%)、实验作业(15%)、期中考试(30%)和期末考试(40%)综合决定。杜克大学为了评估学生对课程内容的理解和应用能力,每周会进行两级测验,即A系列测验(多项选择题)和B系列测验(需要数值解的定量问题)。每周测验问题数量不一,学生可以重复进行所有的测验,按其所有测验中的最高分记录等级,此外还有期末考试和同伴互评。爱丁堡大学的每门课程至少有两次评价机会,不同课程的评价方式不同,有的利用视频测验,有的利用每周的作业测验,其他采用纯粹的同伴互评方式等(王海蓉等,2014)。

在线学习的学习效果评价方法中,客观类试题,诸如选择题和判断题,可以通过机器进行评判反馈。但一些主观类试题,如一些复杂的和开放性的作业任务,诸如数学演算、设计问题和论文就无法通过机器评判反馈。同伴互评是一个有效的交流和互动过程,它注重对学习过程的监控和反馈,是学生自我评价的一部分(Ashley & Goldin,2011)。在传统课堂取得良好的应用效果后,同伴互评机制正在逐步被各大MOOC平台引入(Goldin,2012)。例如,Cousera平台目前采用的同伴互评方式是先由教师制定评分规则,针对一个范例进行练习试评。在试评通过后,互评系统会随机抽取五份匿名作业给每名学生进行打分,老师最后以学生平均评分作为最终成绩。这种让学生参与评分的机制,一方面可以让学生更加深入地掌握课程内容,另一方面也能让作业得分相对教师单方面的评分显得更加客观公正。

由于MOOC学习者规模庞大,提交的作业数以万计,仅仅依靠教学团队,无法一一查阅与批改,同伴互评不仅弥补了这一不足,更让学习者参与其中,强化学习者之间的互动,因此,对于大规模在线教学,同伴互评无疑是一次很好的尝试。

二、同伴互评

(一)同伴互评的定义

同伴互评(peer review)是学生相互交换作业或测试并提出修改建议的写作教学活动,也被称作同伴反馈(peer feedback)、同伴反应(peer response)、同伴评论(peer critiquing)、同伴评估(peer evaluation)和同伴编辑(peer editing)等(莫俊华,2007)。美国亚利桑那大学刘汉森(Liu Hanson)在《第二语言写作课堂中的同伴反馈》一书中对“同伴反馈”的定义是:在写作过程中学习者互动并提供信息,承担通常情况下由训练有素的教师、助教或编辑担任的角色和责任,以书面或口头的方式评价、批改彼此的稿件(陈茂庆等,2013)。肖俊洪等(2008)认为同伴互评指的是由学生彼此对完成学习任务的情况进行互评,可以采取一对一形式,即结成学习对子,也可以采取一对多、多对一或多对多形式,应用于书面作业、口头陈述、表演和小组协同学习等活动中。

综合上述观点,我们将同伴互评定义为:学习者相互交换学习任务,以书面形式批阅并且点评彼此的学习任务。交换形式采取一个人批阅多份学习任务,或多人批阅一份学习任务等方式。

(二)同伴互评的意义

同伴互评能促使学生在学习中扮演更加积极的角色,能给学生创造交流、思考和合作的机会,提高学生的学习欲望和自主性(Sadler & Good,2006)。同伴互评能够很好地实现有数千甚至上万学生参与的课程中对学习者学习效果的评价和成绩评测,在

语言学、师范类专业中效果尤为明显(Russell,2004)。同伴之间的相互启发能够促进学生的深度理解,培养高阶思维。在文章撰写方面,同伴互评能够使学生相互借鉴,相互纠正语法等方面的错误,从而提高写作水平(陈茂庆等,2013)。同时,这种方法使得学习者可以接触到他人作品,增进同伴之间情感和思想交流。互动和交流才是智慧最大层面上的流通,知识的流动不再是单向而是双向的。它是学习者彼此智慧和文化的"接纳"过程。这种教学模式凸显个性,符合培养创新人才的理念。同伴互评过程也使得学习者提前感受到作为一名教师是如何进行作业审阅,一定程度上也是一种教师职前训练,为师范类学校的学生提供了实践锻炼的机会。

在MOOC系统中,如果要将主观类试题作为对学生学习效果的形成性和终结性评价的主要手段,同伴互评可以作为一种非常有效的实现方法。同伴互评机制的设立不仅使得主观类试题得到有效的评估,同时也有助于培养学生批判性思维能力,激发学生学习兴趣,降低学习焦虑感,增强其纠错能力,进而提高学习效率。匿名同伴互评反馈活动能够对评测结果作出更直接、更真实的反馈。

(三)同伴互评的研究现状

为了了解国外研究状况,渭南师范学院的杨阿朋查阅了国外十多种语言学期刊,包括《第二语言写作杂志》(Journal of Second Language Writing)和《英语教学杂志》(ELT Journal)等;还以"peer review"和"second language writing"为关键词在中国硕博士论文库中进行检索,查阅的时间范围是1989至2009年。研究发现,同伴互评主要用于写作教学(杨阿朋,2012)。通过查阅CNKI收录期刊1986-2014年间刊载的关于同伴互评的文献,我们发现,对于同伴互评的研究也主要集中在写作、外国语言文学(如专业英语翻译)等方面。写作教学、外国语言文学的学习都属于复杂的、开放性的主观性学习任务。通过对这些研究现状的分析可以确定,同伴互评的评价方式能为主观性试题的评价提供一定的方法支持。

虽然通过批阅他人习作的确可以提高学生的学习能力,但学习者对同伴互评中,同伴给予的作业建议的准确性和权威性表示怀疑。通过对美国两所大学外语学习者在写作课上对自评、教师评和同伴互评的认可度的调查发现,94%学生更喜欢老师评语(Zhang,2012)。同伴互评机制在评价过程中需要依据一定的理论或过程模型来保证同伴互评的准确度、信服度和价值。

从Coursera的"人机交互"课程采用同伴互评方式实现学生作业评价的效果看,同伴互评的作业成绩整体上与教师给出的成绩取向基本一致。但近40%学生作业的同伴互评评分与教师评分有超过10%的差距,最大的差距达到70%。因此,构建合理的同伴互评过程,采用符合实际的同伴互评评分计算方法以获得可靠和精确的同伴互评得分,将有很大的研究和改进空间(Goldin & Ashley,2011)。

本文将通过建立和优化同伴互评概率模型来提升MOOC评价系统中对主观试题评分的客观性和准确性。

三、同伴互评概率模型的构建

本文构建了三种关联复杂度不同的同伴互评概率模型,用于推断对同伴互评生成评价成绩起重要作用的相关参数的数据,以保证同伴互评的客观性和准确性。

(一)同伴互评系统的特性

适用于MOOC的理想的同伴互评系统应有以下特性:1)能提供可靠和精确的评价;2)给学生、课程教师和系统管理员安排稳定的工作量;3)随着课程学习人数的增长,系统具有可扩展性;4)能适用于各种不同主观题型的课程作业和考试题目。

(二)相关参数的设定

在模型建立过程中,我们使用特定符号将模型中主要的相关参数予以设定。首先是基本参数的设定。对于课程中某一作业题目或考核项目,学生提交的所有作业设定为U,其中某份作业设定为u,则$u \in U$,且设定每个学生对应一份作业。所有评分者设定为G,其中某个评分者为v,则$v \in G$。用$v \rightarrow u$表示评分者v评判了作业u。集合$\{u : v \rightarrow u\}$表示由学生v评判的所有作业的集合。其他参数设定如下:

1)真值分数:假设每份作业u都有一个真值分数,设定为s_u。真值分数无法精确地观察到,但可以推理。

2)评分者偏差:每位评分者v的评判都有一个

评价偏差 $b_v, b_v \in R$。b_v 体现了评分者评分对应真值分数上下偏差的趋势。

3)评分者可靠度:是指评分者 v 评判作业 u 时,给出的分数经过偏差纠正后和真值分数的相近程度,表示为 $\tau_v, \tau_v \in R^+$。在本文模型中,τ_v 代表实际分数与正常分布的匹配度或反向差异。

4)显性分数:z_u^v 代表评分者 v 评判作业 u 的实际分数。所有同伴互评的显性分数集表示为 $Z=\{z_u^v\}$。

(三)概率模型的构建

1.模型一(PG1)

模型一仅考虑评分偏差和评分者可靠度两个参数。同时假定:如果某个评分者的偏差不为零,则所有评分者的平均偏差为零。模型一对相关参数给出如下先验分布。

1)评分者可靠度:$\tau_v \sim G(\alpha_0, \beta_0)$,对应于每个评分者 v;

2)评分者偏差:$b_v \sim N(0, 1/\eta_0)$,对应于每个评分者 v;

3)真值分数:$s_u \sim N(\mu_0, 1/\gamma_0)$,对应于每个被评价作业 u;

4)显性分数:$z_v^u \sim N(S_u+b_v, 1/\tau_v)$,对应于每个同伴互评分数。

其中,G 是有固定超参数 α_0 和 β_0 的 γ 分布,表示正态分布,η_0 和 γ_0 是分别对应于评分者偏差和真值分数的固定超参数,假定评分者偏差的正态分布值为 0。

模型一实现过程中,在选择评分者 v 时,其评分者可靠度 τ_v 要作为主要筛选条件之一。因为,从模型一给定的先验分布可知,评分者可靠度越大,显性分数正态分布曲线就越窄,则显性分数越稳定。由于学生对某门课程的学习水平趋于固定,显性分数越稳定,则表明评分者给出的显性分数越精确。

此外,我们还考虑了一种简单情况,即所有评分者可靠度都为相同值,只有评分者偏差是可变的,记为模型一简化版(PG1-bias)。

2.模型二(PG2)

经过模型一的验证,考虑评分者可靠度和评分者偏差,对于提升同伴互评的准确性有着较好的实际效果。但如果只考虑评分者 v 在一次同伴互评过程中对少量提交作业的评价,其 b_v 和 τ_v 数据的可靠性就会受到一定限制。要获得评分者 v 更为可靠的 b_v 和 τ_v 数据,有效的方法是观测该评分者更多的同伴互评行为,即同时观测其在其他同伴互评任务中对其他提交作业的评价结果。要做到这点,必须要了解评分者 v 对不同作业的同伴互评行为,其评分者偏差和评分者可靠度之间是否有关联,它们是相同的还是随时间变化的。

通过在模型一中应用 Cousera 提供的"人机交互"课程相关数据,我们发现,某个评分者对于两个连续的同伴互评任务的评分者偏差 b_v 数据之间满足皮尔逊相关系数,数值为 0.33,表明了不同时间段具有稳定的时间相关性。而评分者对于两个连续的同伴互评任务的评分者可靠度 τ_v 数据间基本上无变化。

因此,在模型二中,我们进一步考虑评分者偏差的时间相关性。我们假定评分者 v 对作业 T 的评分者偏差 $b_v^{(T)}$ 依据其对作业 T-1 的 $b_v^{(T-1)}$ 而定。模型二中相关参数的先验分布设定如下:

1)评分者可靠度:$\tau_v^{(T)} \sim G(\alpha_0, \beta_0)$,对应于每个评分者 v;

2)评分者偏差:$s_v^{(T)} \sim N(b_v^{(T-1)}, 1/\omega_0)$,对应于每个评分者 v;

3)真值分数:$s_u^{(T)} \sim N(\mu_0, 1/\gamma_0)$,对应于每个被评价作业 u;

4)显性分数:$z_u^{v,(T)} \sim N(s_u^{(T)}+b_v^{(T)}, 1/\tau_v^{(T)})$,对应于每个同伴互评分数。

在模型二的实现过程中,考虑到评分者对不同作业的评判尺度不同,我们通过固定的评判尺度对不同作业的同伴互评分数进行了标准化处理。从相关参数先验分布可以看出,显性分数是真值分数与前一次得出的评分者偏差之和。每一次的评分者偏差总与前一次的数据相关。这样 b_v^T 会趋于稳定,待 b_v^T 趋于稳定后就能得出评分者 v 的 b_v 数据。

3.模型三(PG3)

同伴互评中的评分者是学生,其作业也同时被其他学生评分。因此,对评分者的作业得分和其评分能力之间的关系进行分析,并在建模过程中运用是很有价值的。通过对 Cousera 的"人机交互"课程相关数据进行分析,我们发现,某学生自身的作业得分越高,其评分者可靠度也就越高,反之亦然。

根据这一结果,我们假定评分者 v 的同伴互评

分数与其作业得分相关,而不考虑评分者可靠度。相关参数的先验分布设定如下:

1)评分者偏差:$b_v \sim N(0,1/\eta_0)$,对应于每个评分者 v;

2)真值分数:$s_u \sim N(\mu_0,1/\gamma_0)$,对应于每个被评价作业 u;

3)显性分数:$z_u^v \sim N\left(s_u+b_v,\frac{1}{\theta_1 s_v+\theta_0}\right)$,对应于每个同伴互评分数;$s_v$ 是评分者 v 作业的实际得分,θ_1 和 θ_0 是相应的固定系数。

从模型三设定的相关参数先验分布可知,评分者作业的成绩越高,其给出的显性分数正态分布曲线越窄,表明其给出的显性分数越稳定。从模型一可知,评分者给出的显性分数越精确。由此在模型三的实现过程中,在选择评分者时,我们将他们自身的作业成绩也作为筛选的重要条件。

与模型一相比,模型三增加了新的参数因子。即可以通过一个学生的作业得分来预测其评分能力。同时,模型三的限定更严格,前两个模型中的评分者可靠度只依据其自身作业得分一个参数而定,而不是设定了一个有边界的 γ 分布,这样可以防止模型过度拟合。

四、同伴互评概率模型的评测

(一)评测数据来源

本文用于评测同伴互评概率模型的数据来自 Cousera 的“人机交互”课程,该课程由美国斯坦福大学斯科特·克莱默(Scott Klemmer)教授讲授。通过 Cousera 的同伴互评系统,“人机交互”课程每周评价学生提交的作业。我们通过与斯坦福大学计算机科学专业博士生辛梅·库尔卡尼(Chinmay Kulkarni)合作,并获得克莱默教授首肯,获得了该课程同伴互评系统大量的相关数据。

Cousera 系统要求学生在评判其他学生的作业之前,首先要能够正确评判用来训练的样品作业。每个学生会评判 5 份作业,同时自己的作业也会被其他 5 名同学评判。同学评判成绩的中间值就是这份作业的最后成绩。同伴互评采用匿名制。该类同伴互评的数据组记为 HCI1。在进行首次同伴互评后,该系统又采用了若干方法对其同伴互评机制进行优化。比如,该课程将进行同伴互评的学生分成不同的语言组(例如,英语和西班牙语),从而解决作业被非母语者批改以及“爱国主义倾向批改效果”等问题。优化之后的同伴互评数据组记为 HCI2。我们收集了来自世界各地 7240 名提交的 13972 份作业,共有 63199 个同伴互评分数。经统计,总共有 3607 名学生属于 HCI1,3633 名学生属于 HCI2,这些学生大多来自美国以外的国家。我们所使用的模型评测数据集如表一所示。

表一 模型评测数据集

	HCI1	HCI2
学生数	3 607	3 633
布置作业数	5	5
提交作业数	6 702	7 270
同伴互评分数总量	31 067	32 132

(二)评测过程

本文中所构建的同伴互评概率模型,目的是推理和计算同伴互评过程中一些主要参数的数据,而这些数据是无法直接得到的,例如某学生的评分者偏差、评分者可靠度、提交作业的真值分数等。其实质就是推理和计算这些参数在所有给定同伴互评结果条件下的后验数据分布(概率函数)。但由于这些参数之间是相互关联的,因此准确计算后验数据分布十分困难。例如,准确地推理所有评分者对某作业 u 的评分者偏差,能更好地计算 u 的真值分数;推理一个评分者的评分者偏差,还需要很好地计算评分者评价的所有作业的真值分数。这是一个“鸡和蛋问题”。

为使上述模型的主要参数推理过程简单化,我们在评测过程中采用近似推理方法,主要采用“吉布斯采样”和“期望最大化”两种方法,从收集的评测数据集中合理地推算出本文所述模型在实现过程中的相关参数数据。

1. 吉布斯采样方法

吉布斯采样是一种从多个随机变量的联合分布中抽取样本的方法,广泛使用在贝叶斯推断和机器学习中。该方法适用于在符合某种条件分布的数据集合中抽取变量,它依赖于所有其他变量的当前值,对其中每个变量进行迭代采样,因而采样具有较好的敏感性和快速收敛性。

我们采用 MATLAB 工具箱中相应的吉布斯采样实现工具,从评测数据集中抽取相关数据。这些

数据被用来通过统计平均的方法推理得到以上模型实现过程中一些主要参数的数值。比如,通过从数据集中抽取到的作业u的若干个同伴互评组合的真值分数$s_u^1, s_u^2, \cdots, s_u^T$,推理出作业u的真值分数为:$\hat{s}_u = \frac{1}{T}\sum_{t=1}^{T} s_u^t$。我们共经过了800次迭代过程,同时摒弃了最初作为训练数据集的80个相关数据。

2.期望最大化方法

期望最大化方法被广泛地应用在模型参数估计领域,是解决对不可观察变量进行似然估计的一种方法。该方法的核心思想是:根据已有数据,借助隐藏变量,通过期望值之间的迭代,估计似然函数。该方法经过两个步骤交替进行计算,第一步计算期望值,利用对隐藏变量的现有估计值,计算其最大似然估计值;第二步最大化,即通过第一步求得的最大似然值计算参数的值,第二步找到的参数估计值被用于下一个第一步计算中。这个过程不断交替进行,逐步改进模型参数,使参数和训练样本的似然概率逐渐增大,最后终止于极大点。该方法的最大优点是简单和稳定。

我们采用MATLAB工具箱中相应的期望最大化算法实现工具,选择真值分数和评分者偏差作为参数,采用迭代坐标下降方法获取参数的估计值。

在评测过程中我们发现,两种方法的结果是相似的。只是期望最大化方法实现过程更快速,而通过吉布斯采样获得的参数结果更自然。

对于某个作业真值分数的确定,是取教师的评分还是取成千上百同学评分的均值?考虑到教师和学生评价习惯的差异,我们取学生对某个作业评分的均值作为其真值分数。一个有趣的现象是,在评测数据集里,某份作业的显性分数趋向于学生评分的平均值。其中,得分低的作业其分布区间窄,而得分高的作业其分布区间宽。我们将教师评分作为模型的训练数据,这样可以使得模型更信任评分习惯与教师相近的学生的评分数据。

我们分别用HCI1和HCI2组的数据,通过以下两个步骤的迭代,生成本文各模型的评价结果,与Cousera系统生成的评价结果进行比较,评测本文各模型的优劣程度。

1)为了获得更为可靠的评分者偏差和评分者可靠度数据,用除相关作业同伴互评数据以外的其他数据来推理计算评分者的这两个参数。

2)从HCI1和HCI2数据组中,某作业的同伴互评数据中任意选用4位学生的评分,来模拟形成该作业的同伴互评成绩,并记录其与真值分数的差。对于每份作业,分别用不同的模型算法做1500次此类模拟,然后计算与模拟结果相关的四个数据:方均根误差(RMSE),与真值分数分别偏差5%以内、10%以内的数量,以及平均偏差。

(三)评测结果

如表二所示,我们将通过模型一、模型一简化版、模型二和模型三计算所得的四个相关数据项,与通过Coursera平台的同伴互评系统所得到的相应结果进行比较。可以看出,在同伴互评准确率的提高方面,与Coursera平台系统相比,模型三和模型二的效果较好。其中,由于考虑了更多的相关参数,模型三的改进效果最好。对于HCI1和HCI2数据组,其方均根误差(RMSE)分别降低33%和31%;与真值分数的偏差在5%以内的数量分别提高19%和15%;与真值分数的偏差10%以内的数量分别提高14%和9%,平均偏差也大大减小。

经验证,如果将选用学生评分的个数从4个增加到5个,对于HCI1和HCI2,模型三的方均根误差可以分别减小到4.36和4.19。

五、讨论与展望

本文构建的模型通过重点关注与评分者相关的主要参数,包括评分者偏差、评分者可靠度及其相关特性,可以使较大学生规模同伴互评系统得到的结果更加可信、精确和有效。与已有的同伴互评系统相比,本文构建的模型可以提升同伴互评系统整体效果,方均根误差最多可降低30%。

从上述概率模型相关参数的先验分布设定和实际的评测效果可以看出,评分者偏差对模型实际效果的提高非常有效,而评分者可靠度对模型的实际效果影响甚微。为了研究这一现象,我们又进行了进一步的实验。我们运用上述评测数据集通过模型一生成一组合成数据集。当每位学生只选用其对4个同伴的作业评分时,要准确推算出该学生的评分者可靠度是非常困难的。要使得评分者可靠度数据较为准确并对本文所述模型产生较为有效的提升效果,学生需对10个以上同伴的作业进行评分,并且

表二 模型计算结果与 Coursera 平台互评结果比较

	HCI1					HCI2				
	Cousera	PG1-bias	PG1	PG2	PG3	Cousera	PG1-bias	PG1	PG2	PG3
RMSE	7.95	5.42	5.40	5.40	5.30	6.43	4.84	4.81	4.75	4.73
偏差≤5%	51	69	69	71	70	59	72	73	73	74
偏差≤10%	81	92	94	94	95	88	96	96	97	97
平均偏差	7.23	5.00	4.96	4.92	4.77	6.19	4.57	4.52	4.53	4.52

评分数据都要进入实验数据集,这无疑将大大增加模型实现的计算复杂度。这个实验结果也对模型三的效果要好于模型一给出了解释。尽管模型一设定的相关参数先验分布比模型三更易理解,但模型三对评分者可靠度考虑两个系数(θ_0 和 θ_1),比模型一要精确推算出 τ_v,从统计学的角度看更为合理,同时也更易操作。

本文相关模型的进一步优化工作包括增加新的关注参数。比如,对于全球开放的课程,还需考虑评分者的语言和文化背景、写作习惯等个性化信息,以及评分者在评分时投入的关注度高低,制定公正而又具有激励效果的评分机制,来促使学生全身心地投入到同伴互评中。本文模型三中考虑学生自身作业的得分也是一种合理的激励机制。

本文所建立的同伴互评概率模型中,模型三的提升效果最为明显。如要将其应用到实际的 MOOC 系统中,可以将整个过程分为学习者评分、评分者筛选和生成最终评分三个步骤。其中,学习者评分是某份作业的完成者对其他学习者提交的作业依据一定的评分标准进行评价;评分者筛选依据初步的评分合成规则,比如平均分原则,筛选出得分较高的学习者作为合格评分者;选取合格评分者的评分数据,依据模型三定义的评分合成原则生成最终得分。以上步骤可以通过数据库技术、Web 页面呈现技术、PHP 技术和 Java 脚本编程等生成同伴互评模块,组合到 MOOC 系统中,实现对学生提交主观题作业的同伴互评过程。

本文对 MOOC 评价机制的研究只是 MOOC 研究领域的一个小分支,其更多有益于教学的特性功能还有待进一步探讨。MOOC 以其开放性、共享性和交互性等特征呈现出强有力的发展势头,它不仅是信息社会的一个数字化学习构件,更为教育终身化提供了有力保证,同时能为学有余力的学习者提供拓展知识的平台。

[参考文献]

[1] Ashley, K., & Goldin, I. (2011). Toward an enhanced computer-supported peer review in legal education[C]. Proceedings of the 24th International Conference on Legal Knowledge and Information Systems:77-88.

[2]陈肖庚,王顶明(2013). MOOC 的发展历程与主要特征分析[J]. 现代教育技术,(11):5-10.

[3]陈茂庆,李宏鸿,高惠蓉(2013). 名著阅读与同伴互评[J]. 外语教学理论与实践,(1):71-78.

[4] Goldin, I. (2012). Accounting for peer reviewer bias with Bayesian models[A]. Proceedings of the Workshop on Intelligent Support for Learning Groups of the 11th International Conference on Intelligent Tutoring Systems[C]. 122-140.

[5] Goldin, I., & Ashley, K. (2011). Peering inside peer review with Bayesian models[C]. Proceedings of the 15th International Conference on Artificial Intelligence in Education, AIED'11:90-97.

[6]郝丹(2013). 国内 MOOC 研究现状的文献分析[J]. 中国远程教育,(11):42-50.

[7]莫俊华(2007). 同伴互评:提高大学生写作自主性[J]. 解放军外国语学院学报,(3):35-39.

[8] Russell, A. A. (2004). Calibrated peer review: A writing and critical-thinking instructional tool[J]. Teaching Tips: Innovations in Undergraduate Science Instruction: 54-70.

[9] Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning[J]. Educational Assessment, 11(1): 1-31.

[10]王海荣,王静(2014). 国外 MOOC 评估报告对我国高校教学改革的启示[J]. 中国远程教育,(3):37-41.

[11]肖俊洪,张永胜,彭一为,肖哲英(2008). 同伴互评——远程英语教学的有机组成部分[J]. 中国远程教育,(12):41-46+79.

[12]杨阿朋(2012). 国外同伴互评实证研究的现状及发展趋势[J]. 语文学刊(外语教育教学),(8):172-173.

[13]周洪宇,鲍成中(2013). 第三次工业革命与人才培养模式变革[J]. 教育研究,(10):4-9.

[14] Zhang, S. (1995). Reexamining the affective advantage of peer feedback in the ESL writing class[J]. Journal of Second Language Writing,4(3): 35-52.

(编辑:魏志慧)

Probabilistic Models of Peer Assessment in MOOC System

SUN Li & ZHONG Sitao

(*School of Humanities*, *Jiangnan University*, *Jiangsu*214000, *China*)

Abstract: *With the rapid development of information technology, more and more traditional industries have been transformed, and the field of education has also been changed under the influence of new technologies, such as the Internet. Since* 2012, *there has been an explosive growth of MOOCs* (*Massive Open Online Courses*). *Currently, the MOOC supports video lectures, forum, testing and other instructional functions. However, the learning evaluation and feedback functions are still very limited. Among the methods of evaluating learning, multiple choice questions could be accomplished by computers. However, for subjective questions, like complex and open-ended tasks, it would be very difficult. In order to alleviate such issues, a peer review mechanism has been introduced into many MOOC platforms. The establishment of a peer evaluation mechanism both reduces the subjectivity of questions and also improves the learner's abilities of learning through reading other assignments. However, learners are still skeptical about the accuracy and authority of the grades acquired through peer review. A surveys found that* 94% *of students preferred teachers' grading.*

*Consequently, it is beneficial to add theoretic and procedural models into the peer review to maintain its accuracy, credibility and value. In this paper, we constructed three probability models to improve the objectivity and accuracy of scores to subjective questions through peer review in MOOC system. The records was used to evaluate the peer assessment models from Coursera's HCI course. They were divided into two groups, HCI*1 *and HCI*2, *among which those records from Coursera's peer assessment system were named HCI*1. *Then the peer grading system was refined in several ways. For example, graders were divided into different language groups, such as English and Spanish, to address concerns of assignments being graded by non-native speakers as well as the observed patriotic grading effect. After these optimizations, Coursera's peer assessment system was worked again. Those new records gained were named HCI*2. *We used these two record groups to evaluate the accuracies of three probability models. There were two evaluation methods in this paper, Gibbs Sampling and Expectation Maximization* (*EM*). *We found that the results of two methods were close to each other, while EM was quicker and Gibbs Sampling was more natural. Those results from the evaluation process were compared to the Coursera's peer assessment results. The accuracies were found to be greatly improved. Most of all, the improvement effect of Model* 3 (*PG*3) *was the best owing to considering more related parameters. Through PG*3, *the root-mean-square error* (*RMSE*) *was cut by* 33% *and* 31%, *with regard to HCI*1 *and HCI*2 *successively, the quantity with deviation from the true value less than* 5% *was increased by* 19% *and* 15%, *and the quantity with deviation less than* 10% *was increased by* 14% *and* 9% *successively. Meanwhile, the mean deviation was greatly reduced.*

The three models, by focusing on those principal parameters associated with graders like grader bias and grader reliability, would make the peer assessment results more reliable, accurate and efficient. Compared to the existing peer assessment system, those models in this paper would enhance the overall effect up to 30%. *Finally, we discussed the further optimizations and practical applications of those models, including increasing the parameters, such as the attention rates of graders.*

Key words: *MOOC; peer assessment; grader reliability; grader bias*