基于评价能力的同行互评概率图模型

**Probabilistic Graph Models for Peer Assessment Based on Grading Ability**

许 嘉1,2,3，杨攀原1，吕 品1,2,3\*\*

XU Jia1,2,3，YANG Pan-yuan1，LV Pin1,2,3\*\*

（1. 广西大学,计算机与电子信息学院，广西南宁　530004; 2. 广西大学, 广西高校并行与分布式计算技术重点实验室，广西南宁 530004; 3. 广西大学, 广西多媒体通信网络技术重点实验室，广西南宁 530004）

1. Guangxi University, School of Computer, Electronics and Information, Nanning 530004, Nanning, Guangxi, 530004; 2. Guangxi University, Guangxi Colleges and Universities Key Laboratory of Parallel and Distributed Computing Technology, Nanning, Guangxi, 530004, China; 3. Guangxi University, Guangxi Key Laboratory of Multimedia Communications and Network Technology, Nanning, Guangxi, 530004)

摘要：得益于MOOC平台的大规模兴起，开放式作业的同行互评分数汇总问题成为了教育领域研究的热点。为了提高同行互评汇总技术的准确性，本文提出了两种基于评价能力的同行互评汇总模型*RPG*6和*RPG*7，并且提出一种方法来量化评价者的评价能力，从而根据评价能力对评价者的可靠性进行建模，结合与评价者偏见的建模，我们估计出开放式作业的真实分数。最后在真实的课堂环境中将我们的模型与传统模型进行了比较，实验结果表明，我们的模型估计的分数更加接近真实的分数。

关键词：MOOC平台 同行互评 基数估计 概率图模型 评价能力

中图分类号：TP391.1 文献标志码：A 文章编号：

**Abstract:** Continuous aggregation query processing over data streams, which returns aggregated information of tuples within a certain time window, is one of vital operations for analyzing and mining data streams. It supports many important applications, including event tracing, financial analysis and network monitoring. Distributed parallel computing improves the efficiency of data stream query processing, but causes prominent out-of-order

**Key words：**MOOC platform; Peer assessment; Cardinal estimation; Probabilistic graph model; Grading ability

\*基金项目：国家自然科学基金资助项目（61402494），广西高等教育本科教学改革工程项目重点项目（2017JGZ10），广西大学科研基金资助项目（XGZ141182、XGZ150322），广西研究生教育创新计划资助项目（YCSW2018036）

**【作者简介】**

许嘉(1984—)，女，副教授，硕士生导师，主要从事教育大数据分析挖掘和数据管理技术等领域的研究。杨攀原(1998—)，男，硕士研究生，主要从事同行互评理论和系统的研究。吕品(1983—)，男，副研究员，硕士生导师，主要从事物联网和智慧教育领域的研究。

【\*\*通信作者】

吕品(1983—)，男，副研究员，硕士生导师，主要从事物联网和智慧教育领域的研究。E-mail：lvpin@gxu.edu.cn

**0引言**

近年来，随着大规模开放式在线课程（Massive Open Online Course，MOOC）平台的兴起，越来越多学生利用其提供的高质量教育资源完成了课程学习。一门优秀的MOOC课程可能拥有数万参与者，这无疑给平台教师批改主观题作业（例如简答题、作文题）带来了巨大挑战[1]，因为主观题作业通常没有标准答案。考虑到主观题作业对培养学生创新思维能力和语言组织能力至关重要[2]，目前主流的MOOC平台，包括Coursera[[1]](#footnote-0)、edX[[2]](#footnote-1)和ICourse[[3]](#footnote-2)，都采用同行互评（Peer Assessment）的方法来帮助教师评估学生提交的主观题作业。与众包类似，同行互评要求所有提交作业的学生同时担任该作业的同行评价者。具体而言，每名同行评价者首先被要求基于教师提供的评分指导并以匿名的方式对分配给其的作业进行评价，之后通过汇总每份作业的所有评价结果即可得到该作业的最终评价结果。当下使用最为广泛的同行互评技术是基数估计技术，即要求同行评价者针对每份待评价作业给出一个数值型的分数。除了能够减少教师的评判工作量之外，同行互评还具备许多教育价值，例如其可以激发学生的学习兴趣[3]、可以增强学生的课程参与度[4]、还可以提高学生的责任感[5]。

同行互评研究领域中的重要研究问题是如何汇总每份作业的所有同行评价分数从而使汇总结果尽可能接近该作业的真实分数。目前，大多数MOOC平台（例如Coursera和中国大学MOOC）只是简单以每份作业的所有同行评价分数的均值或中位数作为该作业的最终分数。然而，同行评价者的评分质量受到其评分可靠性和偏见等因素的影响[6]。其中，偏见是指同行评价者的评分相对于真实分数有放大或缩小的习惯性倾向，评分可靠性则表示同行评价者的作业评分在去除偏见影响后与作业真实分数之间的随机偏差。可见，基于均值或中位数的同行评价分数汇总方式会产生不精确的汇总结果[7]。为此，近年来研究人员提出了一系列汇总同行评价分数的概率图模型。这些模型将同行评价者的评分可靠性和偏见设置为随机变量，通过用概率图模型对变量之间的依赖关系进行建模从而可以更加准确地估计每份作业的真实分数。然而，现有概率图模型对同行评价者的评分可靠性进行建模时通常只考虑该生在本次作业上的得分对其评分可靠性的影响。虽然该生的作业得分会对其评分可靠性产生一定影响，但是二者间并没有绝对的正相关关系[20]。例如，一个作业得分较高的学生由于某种原因只是随便给出了待评价作业的分数，导致其评分与待评价作业的真实分数之间存在较大偏差；又如，一个作业得分较低的学生基于教师的指导非常认真地针对待评价作业给出了评价分数，其评分可能和待评价作业的真实分数非常接近。可见，依据评价者的作业得分来对其评分可靠性进行建模存在不精确性。而真正可以直接衡量评价者评分可靠性的是学生的评价能力，可以由作业互评分数与作业真实分数之间的绝对差进行量化：绝对差越小，学生的评价能力越强，评分可靠性也就越高[21]。

鉴于此，本文基于评价能力提出了两种概率图模型，命名为*RPG*6和*RPG*7。具体而言，*RPG*6和*RPG*7在现有概率图模型的基础上将学生的评分可靠性建模为依赖于学生评价能力而非学生作业得分的变量，以提高模型所预测的作业分数的准确性。*RPG*6和*RPG*7的区别在于：*RPG*6假设评价者的评分可靠性服从伽马分布，而*RPG*7假设评价者的评分可靠性服从正态分布。为了能够更准确地量化学生评价能力，*RPG*6和*RPG*7还结合了教师抽查的方式，即要求教师抽查一定比率的学生作业并给出这些作业的真实分数。在真实数据集上进行的多组实验表明本文提出的*RPG*6和*RPG*7显著优于现有的概率图模型，能更准确地汇总每份作业的同行评价分数并预测其真实分数。

**1 相关工作**

根据同行评价者反馈的内容进行分类，可将面向MOOC平台的同行互评技术分为同行互评序数估计技术（Ordinal peer assessment）和同行互评基数估计技术（Cardinal peer assessment）。

**1.1 同行互评序数估计技术**

同行互评序数估计技术首先要求每名同行评价者对分配给其的一个作业子集中的作业进行的优劣排序，再将所有同行评价者给出的作业子集的排序进行汇总，最终得到所有作业的优劣排序。序数估计的研究问题是如何基于多个作业子集的排序来估计整个作业集的排序，进而估计每份作业的真实分数信息。近年来研究人员提出了一些有效的序数估计同行互评技术。例如，Shah等人在Bradley-Terry（BT）模型[8]的基础上，提出了RBTL模型[9]。该模型将每个学生反馈的有序作业对看做是成对的偏好分布，通过这些成对的偏好分布推导出作业全集的最终排序。而Raman[10]等人将一些传统的排序聚合模型（包括BT[8]、MAL[11]、Thurstone[12]和PL[13]）应用到序数估计模型中，并引入了同行评价者的评分可靠性变量，最终有效提升了序数估计模型对作业真实分数估计的准确性。Nicola等人则基于模糊数学理论定义了一个有序的同行评价模型FOPA[14]，与只有“大于”和“小于”比较操作符的传统序数估计不同，FOPA模型引入了更多的比较操作符，例如“远远大于”、“大于等于”和“约等于”等，以此提高序数估计的准确性。

序数估计技术将同行互评活动转化为一系列简单的作业两两排序任务，虽然降低了互评的难度，然而也存在一些问题：序数估计在估计质量相近的两份作业的得分时会扩大它们的得分差距[21]。

**1.2 同行互评基数估计技术**

同行互评基数估计技术首先要求每名同行评价者给待评价的作业打一个数值分数，之后再汇总每份作业同行评价分数并将汇总结果作为该作业最终分数。基数估计技术是当下主流MOOC平台采用的方法。相较于序数估计技术，基数估计技术能更好地体现作业两两间的得分差距。现有基数估计技术的汇总方法主要有两种：加权聚合模型和概率图模型。加权聚合模通常以每名同行评价者的评分准确性为权重对多名同行评价者针对同一份作业给出的评分进行加权聚合，并以聚合结果作为对作业真实分数的估计。Fang等人提出了RankwithTA模型[15]，通过教师抽查的方式确定每个同行评价者的评分准确性。在该模型中，教师对抽查的作业打分，基于学生评分和教师评分的差距确定该生的评分准确性权重，进而通过分析评分准确性权值已知的学生与其他学生的评价作业交集来确定其他学生的评分准确性权重。同时，Wang提出将学生的作业成绩等价于其评审能力是不全面的，于是提出了SSPA模型[16]。SSPA模型在设定同行评价者的评分准确性权重时同时参考了同行评价者与教师的评分差距和学生的作业得分，进而加权聚合得到对作业真实分数的估计。考虑到加权聚合模型不能很好量化同行评价者之间的个体差异，研究人员提出了基于概率图模型的基数估计技术。概率图模型以学生作业的真实分数、同行评价者的评分可靠性和偏见视为隐变量，然后通过可观测的变量，例如作业的评价分数等，来推演各个隐变量的取值，即可得到对作业真实分数的估计值。Piech等人最早提出三种同行互评概率图模型，即*PG*1，*PG*2和*PG*3[17]。其中，效果最好的*PG*3认为同行评价者的评分可靠性和其作业真实得分之间存在正相关线性关系。而Mi等人则认为线性关系不能准确概括评分可靠性和作业真实得分之间的关系，所以他们在*PG*3的基础上将同行评价者的可靠性建模为均值为评分者作业真实得分的概率分布模型，提出了*PG*4和*PG*5[18]，进一步提高了对作业真实分数预测的准确性。其中*PG*4假设评分可靠性取值服从伽马分布，而*PG*5则假设评分可靠性的取值服从高斯分布。然而，上诉模型中的可观测的变量只有互评分数，使得这些模型模型不能很好应对数据稀疏性问题。为此，Wang在*PG*4和*PG*5的基础上引入了相对分数（即同一评价者针对同一次作业的不同提交所给出的评分间的绝对值差异）作为观测变量，进而提出了*PG*6和*PG*7模型[19]，从而有效地解决了数据稀疏性带来的问题。可见，现有的同行互评概率图模型在对学生的评分可靠性进行建模时都假设其主要受学生作业的真实得分影响。然而，学生的评分可靠性其实和学生的作业得分之间没有直接联系：一个作业得分较高的学生也可能在同行互评活动中没有认真评价分配给其的作业；一个作业得分较低的学生也可能在同行互评活动中给出高质量的评价结果。因此，直接使用学生在作业上的真实得分作为其评分可靠性的建模依据必定会降低模型对作业真实分数的的预测精度。鉴于此，本文提出了多个基于评价能力的同行互评概率图模型，在模型中基于学生的评价能力来对其评分可靠性进行更为准确的建模。学生的评价能力直接体现学生评判的分数与作业真实分数之间的差距：差距越小，学生的评价能力越强，评分可靠性也就越高。

**2 基于评价能力的同行互评概率图模型**

本节首先介绍预备知识，其次给出学生评价能力的计算方法，最后详细介绍本文提出的两种基于学生评价能力的同行互评概率图模型。

**2.1 预备知识**

令符号*U*表示所有参与同行互评的学生集合，表示*U*中的第*i*个学生，且令符号*v*表示某同行评价者。通常情况下所有提交作业的学生都将作为同行评价者参与互评活动，即, *v*∈*U*。令表示分配给同行评价者评价的所有作业集合，表示评价学生作业的所有评价者集合，*V*表示所有评价者的集合，表示同行评价者*v*的评价能力，且*R*={*rv*|*v*∈*U*}表示所有学生的评价能力值的集合。设在同行互评活动中一名同行评价者被要求评价*l*份作业，即，且定义一份作业的满分为。下面给出与概率图模型相关的重要概念的名称和含义说明。

* **真实分数**：真实分数是模型中的隐变量，表示一份作业严格按照评分指导打分所获得的分数。在本文设置中，将教师对作业的评分视为该作业的真实分数。真实分数用符号*s*表示，而则表示学生所提交的作业的真实分数。
* **偏见**：偏见是模型中的隐变量，表示同行互评者在评价作业时多给分或少给分倾向，被认为是同行评价者的客观属性之一。本文用字母表示同行评价者*v*的偏见。
* **评分可靠性**：可靠性是模型中的隐变量，表示同行评价者的评分在经历去除偏见的校准后与作业真实分数之间的偏差。用符号表示同行评价者*v*的评分可靠性。
* **互评分数**：互评分数是模型中的可观测变量，是同行评价者给作业打出的评分，本文用符号表示评价者*v*针对学生的作业给出的评分。评价分数的集合*Z*表示为。
* **相对分数**：相对分数是模型中的可观测变量，表示同一评价者针对同一次作业的不同提交所给出的评分间的绝对值差异，用符号表示基于评价者*v*给学生作业的评分和给学生作业的评分得到的相对分数，其等于和之差的绝对值。相对分数的集合*D*表示为。

表1总结了本文涉及的重要符号及其含义。

表 1 符号及其描述

Table 1 Notations and their descriptions

|  |  |
| --- | --- |
| 符号 | 描述 |
|  | 参与同行互评的学生集合，表示第*i*个学生 |
| *v* | 某一同行评价者， |
|  | 分配给同行评价者评价的所有作业集合 |
|  | 评价学生作业的所有同行评价者集合 |
|  | 所有同行评价者的集合 |
|  | 同行评价者*v*的评价能力 |
| *l* | 给每名同行评价者分配的待评价作业数量 |
|  | 一份作业的满分值 |
|  | 学生作业的真实分数 |
|  | 同行评价者*v*的偏见 |
|  | 同行评价者*v*的评分可靠性 |
|  | 同行评价者*v*给学生作业的评分 |
|  | 基于同行评价者*v*对学生作业的评分和对学生作业的评分得到的相对分数 |

**2.2 评价能力**

在描述基于评价能力的概率图模型的实现细节之前，本节先给出学生评价能力的计算方法。学生的评价能力是由其给出的作业互评分数与作业真实分数之间的偏差决定的。本文基于学生的评价能力而不是学生的作业得分来对学生的评分可靠性进行建模，更符合评分可靠性的内涵。具体而言，同行评价者*v*的评价能力*rv*的计算公式如公式（1）所示。

|  |  |
| --- | --- |
|  |  |

其中，表示某学生本次作业的真实得分，表示同行评价者对学生作业的评分，是同行评价者*v*所评价的作业集合中具有真实分数的作业集合，表示一份作业的满分值。由公式（1）可知，要计算同行评价者*v*的评价能力则要求其所评价的部分作业已具有真是分数值。为此，要求教师抽查*n*=*N*×*x*份学生提交的作业并给出这些作业的真实分数，其中*N*为学生提交作业的总数、*x*为抽查比率。在本文中，这些因为被教师抽查而具有真实分数的作业被称为审查资源。在一次互评活动中，每个审查资源会随机分给*l*名同行评价者。若一名同行评价者*v*评价的作业集合中包含审查资源，则通过比较*v*对审查资源的评分与审查资源的真实分数即可基于公式（1）计算出*v*评价能力*rv*。而对于那些所评价的作业集合中不包含审查资源的同行评价者*v*，则将*rv*设定为其在上一轮同行互评活动中的评价能力值并参与概率图模型的推断过程。当本轮互评活动结束后，由概率图模型推断出的每份未被抽查作业的得分被视为该份作业的真实分数，该信息可以用于更新那些没有评价被抽查作业的评价者的评价能力值。值得注意的是，在模型冷启动的情况下，即所有同行评价者的评价能力值均无法确定，则将每名同行评价者的评价能力值设置为初始值0.5。图1展示了学生评价能力值的维护流程。

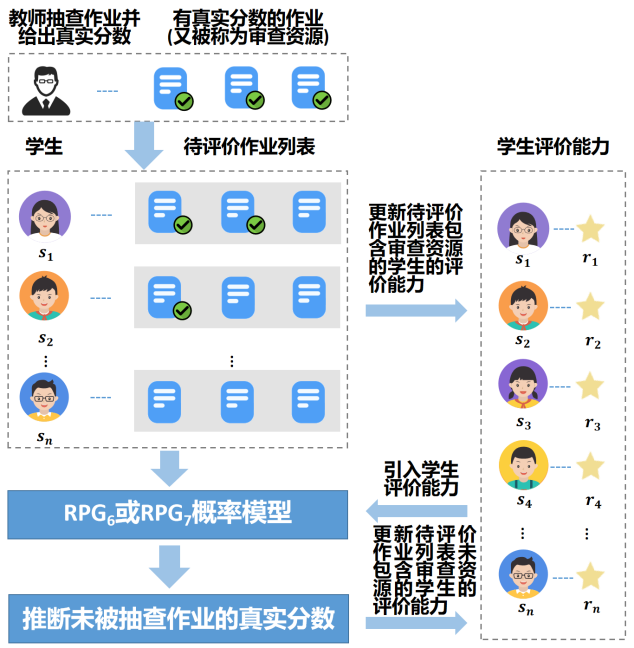


图1 学生评价能力的维护流程图

Fig.1 Flow chart for maintaining grading ability of students

基于以上的符号表征和学生评价能力的计算方法，本文研究问题可描述为：已知同行评价分数集合*Z*、相对分数集合*D*和所有学生的评价能力值集合*R*，构建概率图模型推导出每名同行评价者*v*的评分可靠性、偏见以及每份作业的真实分数。下节将详细介绍本文针对该研究问题提出的两种基于评价能力的同行互评概率图模型。

**2.3 概率图模型设计**

本节提出了两种基于评价能力的同行互评概率图模型，命名为*RPG*6和*RPG*7。这两个模型中的随机变量间的依赖关系如下图2所示。

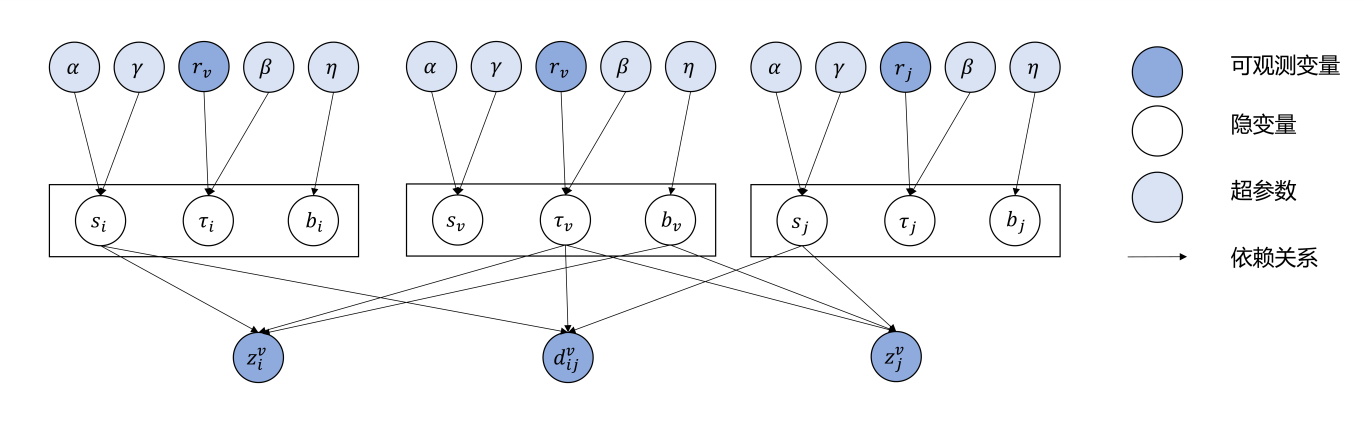


图2 *RPG*6和*RPG*7模型中变量的依赖关系

Fig.2 Dependency of variables in *RPG*6 and *RPG*7 models

如图2所示，同行评价分数*z*、相对分数和学生评价能力*r*是模型中的可观测变量，而学生所提交作业的真实得分*s*，同行评价者的评分可靠性、偏见则是模型中需要估计的隐变量。这些隐变量的先验分布与超参数的分布是彼此联系的。

本文提出的*RPG*6模型改进了现有的*PG*6模型[19]，其模型定义如公式（2）所示：

|  |  |
| --- | --- |
|  | （2） |

*RPG*7模型和*RPG*6的模型的区别在于，*RPG*6模型假设同行评价者的评分可靠性满足伽马分布，而*RPG*7则假设同行评价者的评分可靠性满足高斯分布。具体而言，本文提出的*RPG*7模型改进了现有的*PG*7模型[19]，其模型定义如公式（3）所示：

|  |  |
| --- | --- |
|  | （3） |

由于概率图模型*RPG*6和*RPG*7中的隐变量没有闭式解，因而采用近似离散推断的策略来推导该隐变量的近似后验分布。需要说明的是*RPG*6和*RPG*7对隐变量*s*和*b*后验分布的推断结果是一致的。具体而言，隐变量*s*的后验分布的推断结果如下：

|  |  |
| --- | --- |
|  | （4） |

隐变量*b*的后验分布的推断结果如下：

|  |  |
| --- | --- |
|  | （5） |

*RPG*6模型中隐变量的后验分布的推断结果如下：

|  |  |
| --- | --- |
|  | （6） |

*RPG*7模型中的后验分布的推断结果如下：

|  |  |
| --- | --- |
|  | （7） |

为了推断这些隐变量的值，*RPG*6和*RPG*7模型采用*Gibbs*采样技术对每个隐变量的后验分布进行了采样。然后对于每个隐变量，通过可观测变量和其它隐变量的近似后验分布进行采样。当模型迭代计算若干次之后，将所得到的若干隐变量组的取值的均值作为这些隐变量的最终的推断值。特别地，对于每一组隐变量，模型都进行*T*次*Gibbs*采样，且在计算它们均值时模型去除了前*B*次的老化采样结果，使得最终得到的隐变量推断值更加准确。

*RPG*6模型的推断算法的伪代码如算法1所示。*RPG*7模型的推断算法流程与*RPG*6模型类似。

|  |
| --- |
| 算法1：*RPG*6模型 |
| 输入：*Z*, *D*, *R*, *V*, *T*, *B*, ,,, |
| 输出：() for all |
|  |
|  |
|  |
|  |
|  |
| **for** **do** |
| **for** each with **do** |
| 基于公式（4）对 进行采样 |
|  |
| **for** each with **do** |
| 基于公式（5）对 进行采样 |
|  |
| **for** each with **do** |
| 基于公式（6）对 进行采样 |
|  |
|  |
|  |
| **return** |

**3实验评价**

通过收集的真实同行互评数据集，本节对本文中提出的基于评价能力的同行互评模型*RPG*6、*RPG*7和相关的同行互评技术进行了实验比较。

**3.1 数据集**

实验中使用的同行互评数据集收集于自主研发的教学服务系统。这些数据来自三门真实教学的课程上教师给学生布置的同行互评作业，这些课程分别是“数据库原理”、“数据结构”和“计算机网络”。每次布置的作业中，只包含一道开放式的作业题，其满分均是10分。表中给出了课程中的同行互评数据集的统计信息。

表2 同行互评数据集的统计信息

Table 2 Statistical information on the peer review dataset

|  |  |  |  |
| --- | --- | --- | --- |
|  | 数据库原理 | 数据结构 | 计算机网络 |
| 互评次数 | 4 | 4 | 2 |
| 互评记录 | 711 | 771 | 396 |
| 作业总分 | 10 | 10 | 10 |
| 每次平均提交数 | 177 | 192 | 198 |

**3.2 对比方法**

为了评估*RPG*6和*RPG*7的效果，将它与以下4种方法进行了比较。

* 均值：将互评分数的平均值作为最终分数。
* 中位数：将互评分数的中位数作为最终分数。
* *PG*6：目前准确性最好的概率图模型之一，引入了相对同伴的分数，假设同行评价者可靠性满足伽马分布，*RPG*6模型扩展于该模型。
* *PG*7：与*PG*6类似，但假设同行评价者可靠性满足高斯分布，引入了相对同伴的分数，假设同行评价者可靠性满足伽马分布，*RPG*7模型扩展于该模型。

**3.3 实验设置**

在所有的同行互评作业中，每位提交作业的学生都作为同行评价者参与了互评活动，根据教师提供的互评准则来评估其他三份提交的作业，以确保每份提交的作业都有三个互评分数。所有的同行互评活动都是双盲的，同行评价者和被评价者都不知道彼此的身份。为了获得数据集中所有互评记录的真实分数，两位具有6年丰富教学经验的教师对所有互评记录严格按照评分准则进行了打分，这两位教师的打分视为该记录的真实分数。

本文提出的概率图模型*RPG*6和*RPG*7在建模时用到了一些超参数，为这些超参数设置合理的值对概率图模型的准确性来说非常重要。由于所提出的模型是*PG*6和*PG*7模型的扩展，实验中对其和*RPG*6和*RPG*7模型中共享的超参数设置了相同的值。对于*T*和*B*，取600和60，即实验中对于每一组隐变量，模型都进行600次*Gibbs*采样，且在计算结果时去除了前60次的老化采样结果。对于概率图模型中的真实分数*s*的超参数，即均值和方差, 分别设置为当前作业互评分数的均值和方差。对于偏见*b*的超参数，设置为1。对于可靠性的超参数，在其它参数取值固定的前提下，以100为步长尝试在[100, 600]范围中的不同取值，然后以其中所得到的对真实分数最准确的估计值为该模型的最终估计值。由于概率图模型的估计具有一定的随机性，对于所有概率图模型，超参数的每种取值取得的是10次真实分数的均值。

**3.4 实验结果**

3.4.1 教师抽查比例对模型准确性的影响

RMSE常常在概率图模型研究中被用来衡量预测分数与真实分数的差距，其计算公式如下：

|  |  |
| --- | --- |
|  | （8） |

其中是所有学生的集合，表示学生作业的真实分数，表示学生作业根据概率图模型得到的分数。

在*RPG*6和*RPG*7模型中，教师的抽查比例将影响评价能力的准确性。而教师抽查这一过程决定了初始具有真实分数的作业数量。所以在拥有了全部互评记录的真实分数之后，通过设置初始具有真实分数的作业数量，就可以虚拟教师抽查的这一过程，从而找出教师抽查作业的比例，对模型准确性的影响。

表3 不同教师抽查比例的RMSE（粗体表示最优结果）

Table 3 RMSE for different teacher sampling rates (best results are in bold)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 教师抽查比例 | 数据库原理 | | 数据结构 | | 计算机网络 | |
| *RPG*6 | *RPG*7 | *RPG*6 | *RPG*7 | *RPG*6 | *RPG*7 |
| 10% | 3.404 | 3.443 | 3.591 | 3.443 | 5.212 | 5.1886 |
|
| 15% | 3.182 | 3.157 | 3.357 | 3.284 | 5.172 | 5.032 |
|
| 20% | 3.187 | 3.146 | **3.262** | **3.257** | 5.044 | 4.974 |
|
| 25% | **3.146** | **3.141** | 3.265 | 3.263 | **4.941** | **4.909** |
|
| 30% | 3.153 | 3.143 | 3.264 | 3.261 | 4.989 | 4.945 |
|

表3展示了教师的不同抽查比例下，*RPG*6和*RPG*7模型的RMSE的的大小。可以观察到，在抽查比例达到25%时，数据库原理和计算机网络两门课程中模型的表现最好。而数据结果课程中模型的表现在抽查比例达到20%时最好。另外值得注意的是，抽查比例与模型的准确性并不是线性相关的。当抽查比例达到了20%以上，模型的RMSE并不会发生太大的变动。所以为了让模型具有更好的效果，需要保证将教师的抽查比例设置在20%左右。

3.4.2 同行互评汇总技术的评估准确性

与之前的相关工作一样，本文继续采用RMSE来评估不同互评汇总技术的准确性差异。

表4 同行互评汇总技术的RMSE（粗体表示最优结果）

Table 4 RMSE for peer assessment aggregation techniques (best results are in bold)

|  |  |  |  |
| --- | --- | --- | --- |
| RMSE | 数据库原理 | 数据结构 | 计算机网络 |
| 均值 | 3.746 | 4.168 | 5.615 |
| 中位数 | 3.744 | 4.213 | 5.785 |
| *PG*6 | 3.637 | 3.832 | 5.576 |
| *PG*7 | 3.587 | 3.771 | 5.504 |
| *RPG*6 | 3.146 | 3.362 | 4.941 |
| *RPG*7 | **3.141** | **3.257** | **4.909** |

表4展示了不同同行互评汇总技术的在数据集中的表现。由表4可知，本文提出的基于评价能力的同行互评技术*RPG*6和*RPG*7在3门课程中RMSE均小于其他技术。由于使用了比真实分数更好的概括学生可靠性的评价能力为可靠性建模，*RPG*6模型在3门课程的RMSE比*PG*6平均降低了11.76%，*RPG*7模型在3门课程的RMSE比*PG*7平均降低了11.74%。实验结果证实了结合学生的评价能力建模可靠性比学生的真实分数的建模要更加准确。

3.4.3 同行互评汇总技术的最大偏差

除了RMSE，本文还通过真实分数与汇总分数之间的最大偏差来分析同行互评汇总技术的准确性。

表5 各个同行互评技术的最大偏差（粗体表示最优结果）

Table 5 Maximum bias of each peer assessment

technique (best results are in bold)

|  |  |  |  |
| --- | --- | --- | --- |
| 最大偏差 | 数据库原理 | 数据结构 | 计算机网络 |
| 均值 | 10 | 9 | 10 |
| 中位数 | 9 | 9 | 10 |
| *PG*6 | 5.171 | 5.518 | 8.047 |
| *PG*7 | 6.766 | 8.447 | 7.841 |
| *RPG*6 | 4.261 | **5.445** | **7.457** |
| *RPG*7 | **4.151** | 5.651 | 7.889 |

如表5所示。从表中可看出，基于评价能力的*RPG*6和*RPG*7模型在3门课程中的最大评分偏差是最小的，说明在学生的可靠性建模中，评价能力确实比真实分数具有更好的效果。

3.4.4 同行互评汇总技术的超参数敏感性

为了展示*RPG*6和*RPG*7模型中的超参数取值对汇总结果的影响，本文对超参数进行了离散型取值分析。将两个模型中的超参数设置在[100, 600]范围内，以100为步长进行变化，得到的实验结果如图3和图4所示。结果表明：这两种模型能够使得超参数在合理的取值范围内保持一定的鲁棒性， 其不同的取值对模型的估计结果误差控制在可接受的范围内。

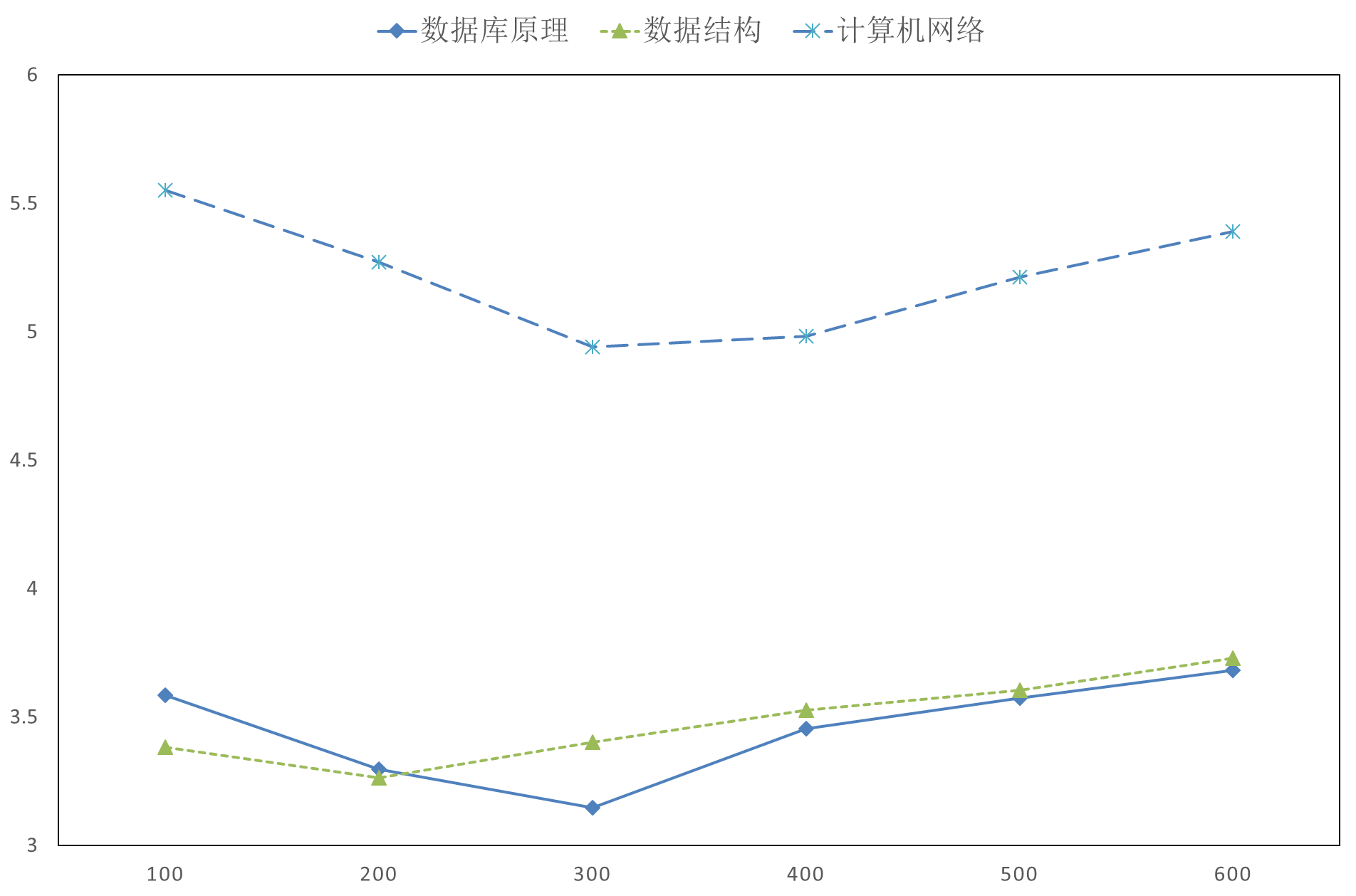


图3 *RPG*6的超参数敏感性分析

Fig.3 Sensitivity analysis of hyper-parameters in *RPG*6

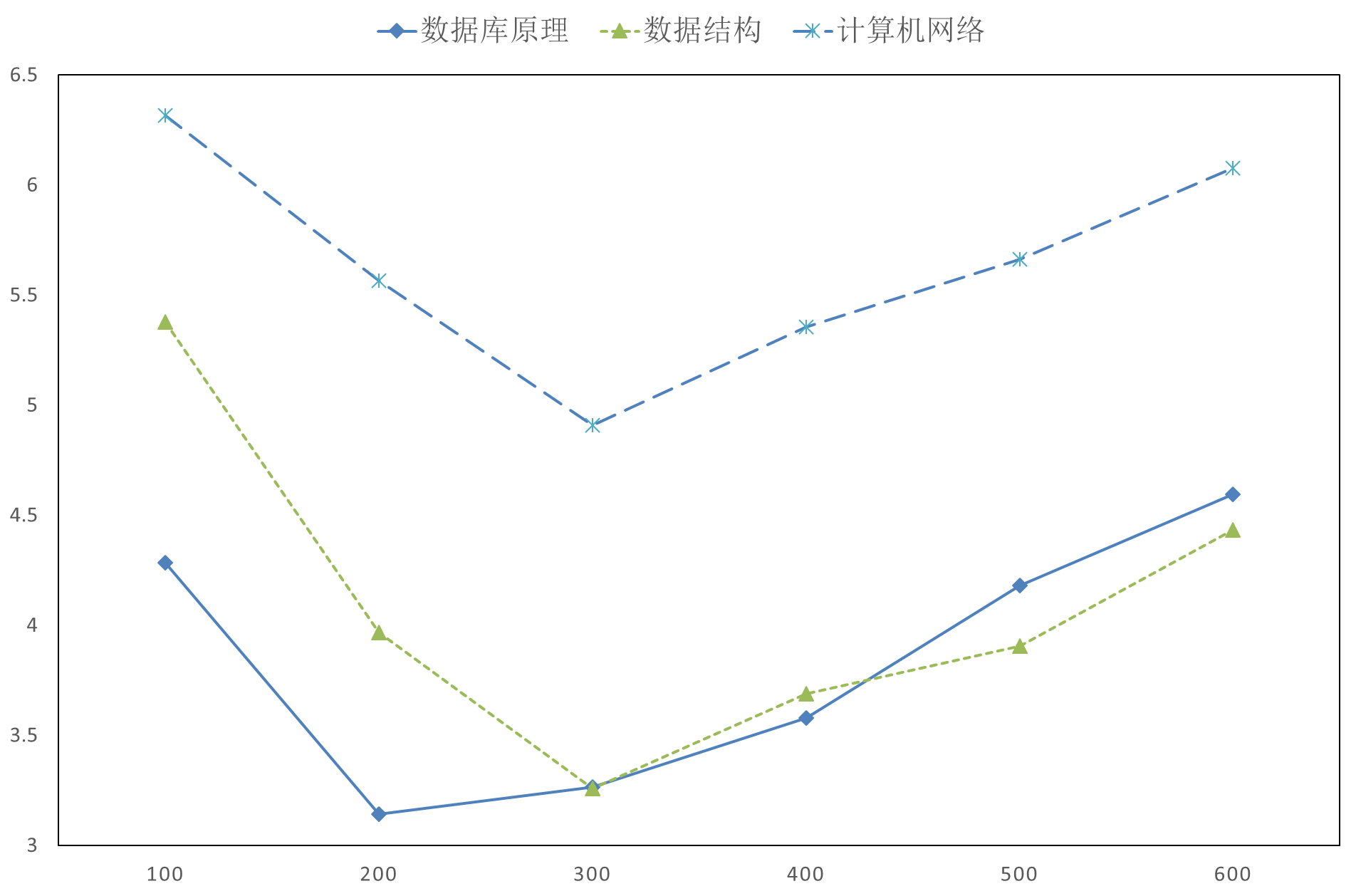


图4 *RPG*7的超参数敏感性分析

Fig.4 Sensitivity analysis of hyper-parameters in *RPG*7

**4结论**

本文提出了两个新的同行互评概率图模型，通过学生的互评能力取代学生的真实分数来重新对评价者可靠性进行建模。我们提出了一种方法来量化学生的评价能力，并且根据学生在每次作业中的表现来更新评价能力，然后利用这种形式来优化评分者的可靠性建模，并且根据新的观测变量和隐变量关系推到了概率图模型生成公式。基于真实的同伴评估数据集的实验结果表明，这两个模型提高了对同伴评估真实分数的估计的准确性，这说明评价能力确实有助于提高估计同行评估的真实分数的准确性。除了教育同行互评领域，我们计划将提出的模型应用于众包领域，并且教师抽查这种形式也可以泛化为少量的专家评审。未来我们将尝试引入其他影响评分者的可靠性和偏差的因素，以进一步改善同行评估的估计。

**参考文献：**

1. Caragiannis I, Krimpas G A, Voudouris A A. Aggregating partial rankings with applications to peer grading in massive online open courses[J]. arXiv preprint arXiv:1411.4619, 2014.
2. Paré D E, Joordens S. Peering into large lectures: examining peer and expert mark agreement using peerScholar, an online peer assessment tool[J]. Journal of Computer Assisted Learning, 2008, 24(6): 526-540.
3. Chang S C, Hsu T C, Jong M S Y. Integration of the peer assessment approach with a virtual reality design system for learning earth science[J]. Computers & Education, 2020, 146: 103758.
4. Hovardas T, Tsivitanidou O E, Zacharia Z C. Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students[J]. Computers & Education, 2014, 71: 133-152.
5. Li L, Liu X, Steckelberg A L. Assessor or assessee: How student learning improves by giving and receiving peer feedback[J]. British journal of educational technology, 2010, 41(3): 525-536.
6. de Alfaro L, Shavlovsky M. Dynamics of peer grading: An empirical study[J]. 2021.
7. Capuano N, Caballé S. Towards adaptive peer assessment for MOOCs[C]//2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC). IEEE, 2015: 64-69.
8. Bradley R A, Terry M E. Rank analysis of incomplete block designs: I. The method of paired comparisons[J]. Biometrika, 1952, 39(3/4): 324-345.
9. Shah N B, Bradley J K, Parekh A, et al. A case for ordinal peer-evaluation in MOOCs[C]//NIPS workshop on data driven education. 2013, 15: 67.
10. Plackett R L. The analysis of permutations[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1975, 24(2): 193-202.
11. Bradley R A, Terry M E. Rank analysis of incomplete block designs: I. The method of paired comparisons[J]. Biometrika, 1952, 39(3/4): 324-345.
12. Thurstone L L. The method of paired comparisons for social values[J]. The Journal of Abnormal and Social Psychology, 1927, 21(4): 384.
13. Plackett R L. The analysis of permutations[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1975, 24(2): 193-202.
14. Capuano N, Loia V, Orciuoli F. A fuzzy group decision making model for ordinal peer assessment[J]. IEEE Transactions on Learning Technologies, 2016, 10(2): 247-259.
15. Fang H, Wang Y, Jin Q, et al. RankwithTA: A robust and accurate peer grading mechanism for MOOCs[C]//2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE). IEEE, 2017: 497-502.
16. Wang Y, Fang H, Jin Q, et al. SSPA: An effective semi-supervised peer assessment method for large scale MOOCs[J]. Interactive Learning Environments, 2022, 30(1): 158-176.
17. Piech C, Huang J, Chen Z, et al. Tuned models of peer assessment in MOOCs[J]. arXiv preprint arXiv:1307.2579, 2013.
18. Mi F, Yeung D Y. Probabilistic graphical models for boosting cardinal and ordinal peer grading in MOOCs[C]//Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.
19. Wang T, Jing X, Li Q, et al. Improving Peer Assessment Accuracy by Incorporating Relative Peer Grades[J]. International Educational Data Mining Society, 2019.
20. 方慧. 众包系统中基于参与者互评的数据质量控制研究[D].南京邮电大学, 2019.DOI:10.27251/d.cnki.gnjdc.2019.000458.
21. De Alfaro L, Shavlovsky M. CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments[C]//Proceedings of the 45th ACM technical symposium on Computer science education. 2014: 415-420.

**附录**

本节将详细介绍*RPG*6和*RPG*7概率图模型的推导过程。*RPG*6和*RPG*7的各个隐变量（,,）的联合后验分布如下：

假设有一个随机变量集X，其中的每一个目标变量x∈X在马尔可夫毯MB(x)条件下，和其非马尔可夫毯变量条件独立。因此对于x而言，所有非马尔可夫毯变量都是冗余的， 关于目标变量x的所有信息都包含在x的马尔可夫毯中。所以为了得到x的分布情况，只需要其马尔可夫毯的信息即可，而不需要对整个数据集求解。因此，对于各个隐变量的后验分布的推断实际是在其马尔可夫毯条件下进行求解。

下面将分别介绍*RPG*6模型中各个隐变量（,,）的推导采样过程。其中*RPG*7模型的,两个隐变量的推导过程和*RPG*6相似，因此只给出*RPG*7中的可靠性的推导过程。

***RPG*6模型中真实分数的推导**

由于概率模型*RPG*6和*RPG*7中的隐含变量没有闭式解，因而采用近似离散推断的策略在[0, 20]区间内以0.1为间隔求解该隐含变量的近似后验分布。

***RPG*6模型中偏见的推导**

***RPG*6模型中可靠性的推导**

***RPG*7模型中可靠性的推导**

1. https://www.coursera.org/ [↑](#footnote-ref-0)
2. https://www.edx.org/ [↑](#footnote-ref-1)
3. https://www.icourse163.org/ [↑](#footnote-ref-2)