一种基于评价能力的概率模型同行评估技术

作者

许 嘉，杨攀原，吕 品

（作者单位）

（单位从大到小，省市 邮编；两个单位之间用“；”隔开，例如：

广西科学院,国家非粮生物质能源工程技术研究中心，非粮生物质酶解国家重点实验室，广西生物炼制重点实验室，广西南宁　530007；广西科学生命科学与技术学院，广西南宁　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：教育 同行互评 基数估计 概率模型 评价能力

中图分类号：

\*基金项目

**【作者简介】**

第一作者 (出生年—)，性别，职称，主要从事研究方向。（我刊不接受共同第一作者。）

【\*\*通信作者】

姓名 (出生年—)，性别，职称（需与第一作者同级或更高），主要从事研究方向，E-mail：

**0引言**

随着大规模的开放在线课程（MOOC）等在线教育平台逐渐兴起，越来越多的学生加入了这些平台，利用其提供的高质量免费教育资源完成学业。一门优秀的在线课程拥有数万参与者，这给平台的教师们带来了巨大的挑战：他们需要对大量的作业进行批改。尽管对于那些客观作业，例如选择题、判断题来说，自动化批改很容易实现，但是对于那些没有标准答案的开放式作业，例如简答题、作文题来说，目前还没有统一、有效的自动评分技术[1]。考虑到开放式作业对提高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Coursera1、edX2和ICourse3，都采用同行互评的方法来帮助教师评估学生提交的开放式作业。同行互评与众包类似，它要求所有提交作业的学生扮演同行评价者，根据教师提供的评分细则，以匿名的方式对所有提交的作业进行评价。最后将每份作业的所有评价结果进行汇总，得到每份作业最终评价。其中，使用最为广泛的同行互评技术是基数估计技术，即同行评价者必须给每份评价作业一个数值型的分数。除了能够减少教师的工作量之外，同行互评还可以带来其他的教育价值，可以激发学生的学习兴趣[3]，加强学生的课程参与[4]，提高学生的责任感[5]。

同行互评的最主要的研究问题之一是同行互评的分数汇总方式，即根据同行评价者的对作业评分集合，得到尽可能接近该作业的真实分数的结果。目前，大多数的MOOC平台（例如Coursera和中国大学MOOC）的汇总方式只是简单地计算作业所有评分的均值和中位数，作为该作业的最终分数。然而，同行评价者的评分质量乘次不齐，受到其可靠性、偏见等因素的影响[6]，其中，偏见意味着同行评价者的评分有放大或缩小的习惯性倾向，而可靠性表示了同行评价者的评分与真实分数之间的随机偏差。因此用均值和中位数来作为最终评分并不准确[7]。目前，一些研究者提出一系列的概率模型，将同行评价者的评分可靠性和偏见设置成了随机变量，通过变量之间的依赖关系来更加准确地估计每份作业的真实分数。然而现有的概率模型中，对同行评价者的可靠性建模只考虑了在本次作业上的最终分数（因为评价者也提交了作业）。虽然学生的作业分数在理想情况下可以反应一个学生的评价能力，但是他们并没有太强的依赖关系[20]。例如，一个成绩很好的学生不喜欢批改其他人的作业，因此在互评中批改的并不仔细，导致其同行评分与真实分数仍存在较大的差距。因此，将作业得分作为可靠性的建模标准并不准确。而真正可以直观的衡量评价者可靠性的，是学生的评价能力。评价能力可以由互评分数与真实分数之间的绝对差体现，绝对差越小，评价能力越高，可靠性也就越高[21]。

鉴于此，本文利用学生的评价能力，提出了两个新的概率模型*RPG*6和*RPG*7。这两个模型在现有模型的基础上，将可靠性的建模中的真实分数替换成了学生的评价能力，以提高模型在推导真实分数的准确性。*RPG*6和*RPG*7的区别在于，*RPG*6假设评价者可靠性服从伽马分布，而*RPG*7假设评价者可靠性服从正态分布。为了能够更准确的得到评价能力，我们还结合了教师抽查的方式，来更好量化学生的评价能力，最后通过真实数据集与传统的概率模型进行了对比，验证了该模型的有效性。

**1 相关工作**

根据同行评价者反馈的内容进行分类，可以将开放式作业的同行互评的汇总估计技术分为序数估计技术和基数估计技术。

**1.1 序数估计技术**

序数估计技术要求每个同行评价者给出其评价作业的优劣排序，最后将所有同行评价者的部分作业排序进行汇总，得到所有作业的最终排名。序数估计的主要研究点在于将多个部分的排序序列整合，推断出整体序列。Shah等人在Bradley-Terry(BT)模型[8]的基础上，提出了RBTL模型[9]，将每个从学生那里收集的有序比较对，看做是两两比较的成对偏好分布，通过两两间的关系推导出最终的排名。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]、PL[13]）应用到了序数估计模型中，并且引入了同行评价者的评分可靠性变量，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同行评价模型FOPA[14]，与只有“大于”、“小于”传统的序数估计不同，FOPA模型映入更多的关系符号，例如“远远大于”、“大于等于”、“约等于”等，以此提高序数估计的准确性。

序数估计技术将评审活动转化为了一项简单的比较任务，降低了互评的难度。但是，同行互评还存在一些的问题：对于一些水平相近的作业，序数估计在对其量化时会扩大它们的实际差距。此外，学生提交的比较对也可以通过直接打分的方式排序得到，这说明和直接打分相比，序数估计包含了更少的信息。

**1.2 基数估计技术**

基数估计技术要求每个同行评价者根据评分准则，给评价的作业一个数字评分。之后通过一些方式汇总每份作业获得的所有评分，得到最终分数。目前基数估计是主流MOOC平台采用的同行互评的评判方法。对比序数估计，基数估计更能体现作业之间的绝对分数差，可以更加准确的的量化作业之间的分数差距。基数估计的汇总方法主要有两种：加权聚合模型和生成概率模型。加权聚合模型主要通过计算每个同行评价者的准确性作为权重，然后将同行评价者的评分加权聚合，得到作业的真实分数。Fang等人提出了RankwithTA模型[15]，通过教师半监督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，教师将抽查部分学生的作业打分，以教师的评分作为学生的权重，并通过这部分学生与其他学生间的评价交集得到所有学生的权重。而Wang认为将学生的作业成绩等价为评审能力是不全面的，于是在此基础上提出了SSPA模型[16]。SSPA在构建同行评价者的准确性时，将同行评价者与教师评分的差距也作为其影响因素，结合学生的作业成绩来共同决定学生的准确性，进而加权聚合得到学生的真实分数。虽然加权聚合模型比起均值、中位数更加准确，但是它还是没有很好的量化同行评价者之间的个体差异。而生成概率模型则是将学生作业的真实分数、同行评价者的可靠性和偏见建模为隐变量概率分布模型，然后通过可观测的互评分数推演隐变量，即真实分数的值。同行互评的生成概率模型最早由Piech等人提出，他们一共提出了三种概率模型*PG*1，*PG*2，*PG*3[17]。其中效果最好的*PG*3认为同行评价者的可靠性与其作业的真实分数之间存在联系，于是在两者之间建立了线性关系进行转化，在实验的数据中，*PG*3的预测分数的准确性在三个模型中是最好的。而Mi等人认为，线性关系不能准确的概括真实分数与可靠性的联系，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真实分数的概率分布模型，提出了*PG*4和*PG*5[18]，以此提高了预测最终分数的准确性。其中*PG*4假设可靠性的概率服从伽马分布，而*PG*5假设可靠性的概率服从高斯分布。然而，上诉模型都只依靠互评分数这一个可观测的变量，使得预测模型存在数据稀疏的缺陷。因此，Wang在*PG*4和*PG*5上加入了相对分数等级，即同一评分者不同提交成绩的绝对同伴等级的差异，构成了*PG*6和*PG*7[19]。两者在原有的模型基础上，加入了一个相对分数这一可观测变量，有效地解决了数据稀疏性带来的问题。我们在之前的工作中，将学生在互评时的行为数据引入了可靠性建模中，提出了*BPG*6和*BPG*7[20]，取得了更加准确的估计分数。然而，现有的概率生成模型在对学生的可靠性建模时，将学生作业的真实分数作为其主要的影响因素。然而学生的评价能力和作业成绩没有直接联系。例如一个具有高水平学术能力的学生在没有认真评价的情况下，依然会给出一个低质量的评价结果。所以，直接使用真实分数作为可靠性的建模标准会产生一定的误差。对此我们提出了定义学生的评价能力的方法，将其作为可靠性的参数。评价能力可以直接体现在学生评分与真实分数的差距上，差距越小，说明学生评价的可靠性越高。

**2 同行互评概率模型**

在本节中，我们会首先介绍使用到的符号，然后介绍我们自定义的互评能力的计算方法，最后介绍我们的同行互评概率模型。

**2.1 符号意义**

我们使用*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。用*v*表示同行评价者，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打分。评价者和被评价者之间的关系用和表示，其中表示评价者评价的所有作业的集合，表示被评价者作业的所有评审人员的集合，表示评价者*v*的评价能力。在互评活动中，一名同行评价者被要求为*l*份作业评分，即，而作业的满分定义为。

下面是有关概率模型重要概念的定义。

* **真实分数**：真实分数是隐变量，表示一份作业严格按照评分准则获得的分数。在实验中，我们将教师对作业的评分视为该作业的真实分数。真实分数用字母*s*表示，是学生作业的真实分数。
* **偏见**：偏见是隐变量，表示同行互评者在评价作业时夸大或缩小其打分的倾向,其被认为是同行评价者的存在的客观属性。用字母表示同行评价者*v*的偏见。
* **可靠性**：可靠性是隐变量，表示评价者的评分在经历了偏见的校准后，与作业真实分数存在的差距。同行评价者*v*的可靠性通过表示。传统的概率模型中，评价者的可靠性主要依赖于其提交作业的真实分数，在我们的概率模型中，评价者的可靠性将取决于其对作业的打分与该作业的真实分数之差。
* **互评分数**：互评分数是可观测变量，表示同行评价者给作业打出的评分，用表示评价者*v*给学生的作业打出的评分。评价分数的集合*Z*表示为。
* **相对分数**：相对分数是可观测变量，表示同一个同行评价者的互评分数们之间的差距，用表示评价者*v*对学生作业的评分和学生作业的评分的互评分数之间的绝对差，即和之间的绝对差。

表一总结了以上变量。

表 1 主要符号及其含义

Table 1 Main 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* | 一位同行评价者， |
|  | 评价者评价的所有作业的集合 |
|  | 评价了学生的所有评价者的集合 |
|  | 评价者*v*的评价能力 |
| *l* | 每个学生在同行评审阶段评价的作业数量 |
|  | 作业的满分 |
|  | 学生作业的真实分数 |
|  | 同行评价者*v*的偏见 |
|  | 同行评价者*v*的可靠性 |
|  | 评价者*v*给学生的作业打出的评分 |
|  | 评价者*v*对学生作业的评分和学生作业的评分的互评分数之间的绝对差 |

**2.2 评价能力**

现有的同行概率模型中，对同行评价者的可靠性建模由学生的真实分数决定。然而作业分数与学生评价能力之间并没有绝对的关系[16]，同行评价者的准确性可以由其互评分数与真实分数之间的差距决定。因此我们将同行评价者的互评分数与真实分数作为一个新的变量，称为同行互评者的评分能力，重新对可靠性进行建模。计算评分能力的公式如下：

(1)

表示学生作业的真实得分，表示评价者对学生作业的评分。表示同行评价者*v*评价的作业集合中，具有真实分数的作业集合，表示作业的满分。根据公式（1），为了计算学生的评价能力，需要提前知道其评价作业的真实分数。通过教师提前评价部分学生作业的方式，可以获取部分作业的真实分数。这些具有真实分数的作业被称为审查资源，在互评流程中，通过对比评价者打分和审查资源的分数，就可以计算出他们的评价能力。对于那些没有被评价审查资源的学生，将会沿用历史留下的评价能力。并且在互评结束后，我们会将每位学生的最终得分视为真实分数，与其打分进行对比，从而更新这部分没有被抽查的评价者的评价能力，用于下一次的同行互评。图1是我们系统更新评价能力的流程图如下。

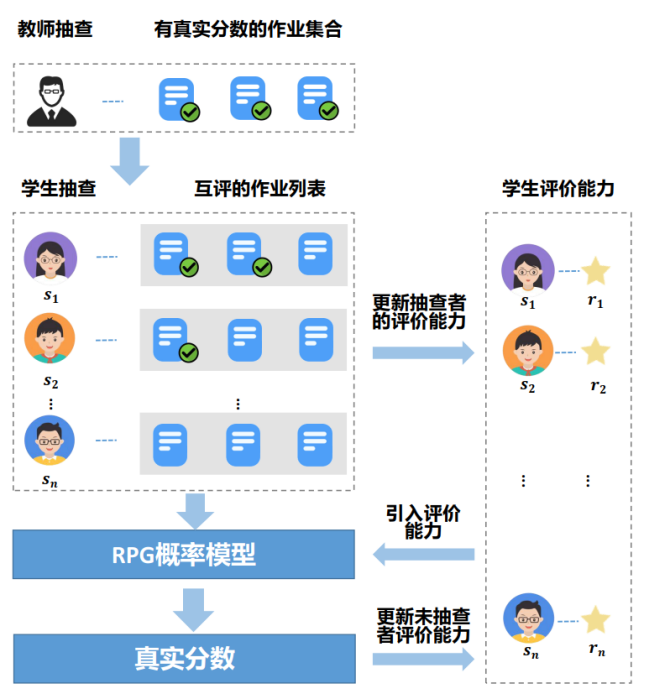


图1 更新评价能力的流程图

Fig.1 Flow chart for updating evaluation capacity

在拥有了评价能力之后，我们就可以对概率模型进行建模了。基于以上的符号表征，我们的研究问题是：给定同行分数*Z*、相对分数*D*和评价能力*R*，需要推导出所有评价者、*b*以及所有的。下面，我们将介绍我们提出的概率模型，以及如何通过概率模型和可观测变量，推导隐变量。

**2.3 概率模型**

我们一共提出了*RPG*6和*RPG*7两个概率模型，其随机变量之间的条件依赖结构如图2中的图形模型所示。

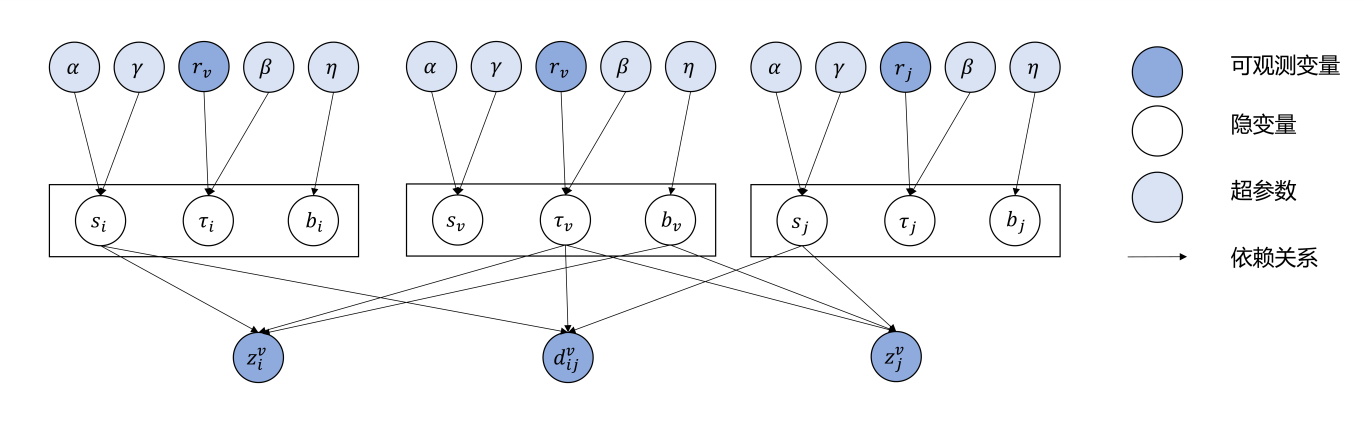


图2 *RPG*6和*RPG*7的概率图模型

Fig.2 Probabilistic graphical model for *RPG*6 and *RPG*7

如图2所示，互评分数*z*、相对分数*d*和同伴的评价能力*r*是模型中可观测的随机变量。学生提交的真实得分，评分者的可靠性偏见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。

*RPG*6模型拓展于现有的*PG*6模型，*RPG*6模型的所有随机变量的分布如下所示：

*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可靠性满足伽马分布，而*RPG*7则假设同行评价者的可靠性满足高斯分布。*RPG*7模型拓展了现有的*PG*7模型，其所有随机变量的分布如下所示：

由于概率模型*RPG*6和*RPG*7中的隐含变量没有闭式解（*closeform solution*），因而采用近似离散推断的策略得到该隐含变量的近似后验分布。对于*s*和*b*，*RPG*6和*RPG*7中的后验分布推断结果是一样的。其中*s*结果的隐含变量的后验分布推断如下：

(2)

*b*结果的隐含变量的后验分布推断如下：

(3)

*RPG*6中的隐含变量的后验分布推断结果如下：

(4)

*RPG*7中的隐含变量的后验分布推断结果如下：

(5)

为了确定隐变量的值，我们对每个隐变量的后验分布信息进行了采样。我们使用了*Gibbs*采样技术。对于每个隐变量，通过可观测变量和其他隐变量的后验分布进行计算。在计算若干次之后，我们将生成的若干组隐变量的均值作为其最终值。对于一组隐变量。我们进行了600次吉布斯采样，在计算均值时，我们去除了前60次的老化采样结果，使得最终的结果更加准确。

*RPG*6具体的算法流程如下,而*RPG*7的算法流程类似：

|  |
| --- |
| 算法：*RPG*6模型 |
| 输入：Z,D,R,V,T,B,,,,, |
| 输出：() for all |
|  |
|  |
|  |
|  |
|  |
| **for** **do** |
| **for** each with **do** |
| Sample according to (2) |
|  |
| **end for** |
| **for** each with **do** |
| Sample according to (4) |
|  |
| **end for** |
| **for** each with **do** |
| Sample according to (3) |
|  |
| **end for** |
|  |
| **end for** |
|  |
| **return** |

**3实验结果**

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。这些数据来自三门课程上举行的同行互评任务，这些课程分别是“数据库原理”、“数据结构”和“计算机网络”。我们邀请了两位具有6年丰富教学经验的教师，对于数据集中所有的互评记录进行评审，给出了分数，在本次实验中，我们将其看做是互评的真实分数。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 |

**3.2 对比方法**

为了评估*RPG*6和*RPG*7的效果，我们将它与以下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**

数据集中同行互评任务的载体是教学平台上布置的作业，所有的作业内容只包含一道开放式的作业题。每位参与作业的学生都将作为同行评价者，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。

本文提出的概率模型*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，为了更好的验证在离散变量中的取值，我们对其和*RPG*6和*RPG*7模型中共享的超参数设置了相同的值。对于概率模型中的真实分数*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，我们设置为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［100,700］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。由于概率模型的估计具有一定的随机性，对于所有概率模型，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 同行互评汇总技术的评估准确性

我们采用RMSE来评估不同互评汇总技术的准确性差异。RMSE的计算公式如下：

(6)

其中是所有学生的集合，表示学生作业的真实分数，表示学生作业根据汇总技术得到的分数。

表3 同行互评汇总技术的RMSE（粗体表示最优结果）

Table 3 RMSE for peer assessment aggregation techniques (best results 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 | 3.151 | 3.357 | 5.044 |
| *RPG*7 | **3.146** | **3.284** | **4.951** |

表3展示了不同同行互评汇总技术的在数据集中的表现。由表3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。

3.4.2 同行互评汇总技术的最大偏差

除了RMSE，我们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4 同行互评汇总技术的最大偏差（粗体表示最优结果）

Table 4 Maximum bias in peer assessment

aggregation techniques (best results 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 | **4.331** | **5.07** | **7.366** |
| *RPG*7 | 4.349 | 5.423 | 7.589 |

如表4 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。

3.4.3 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，我们超参数进行了离散型取值分析。我们将两个模型中的超参数设置在［100，700］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。

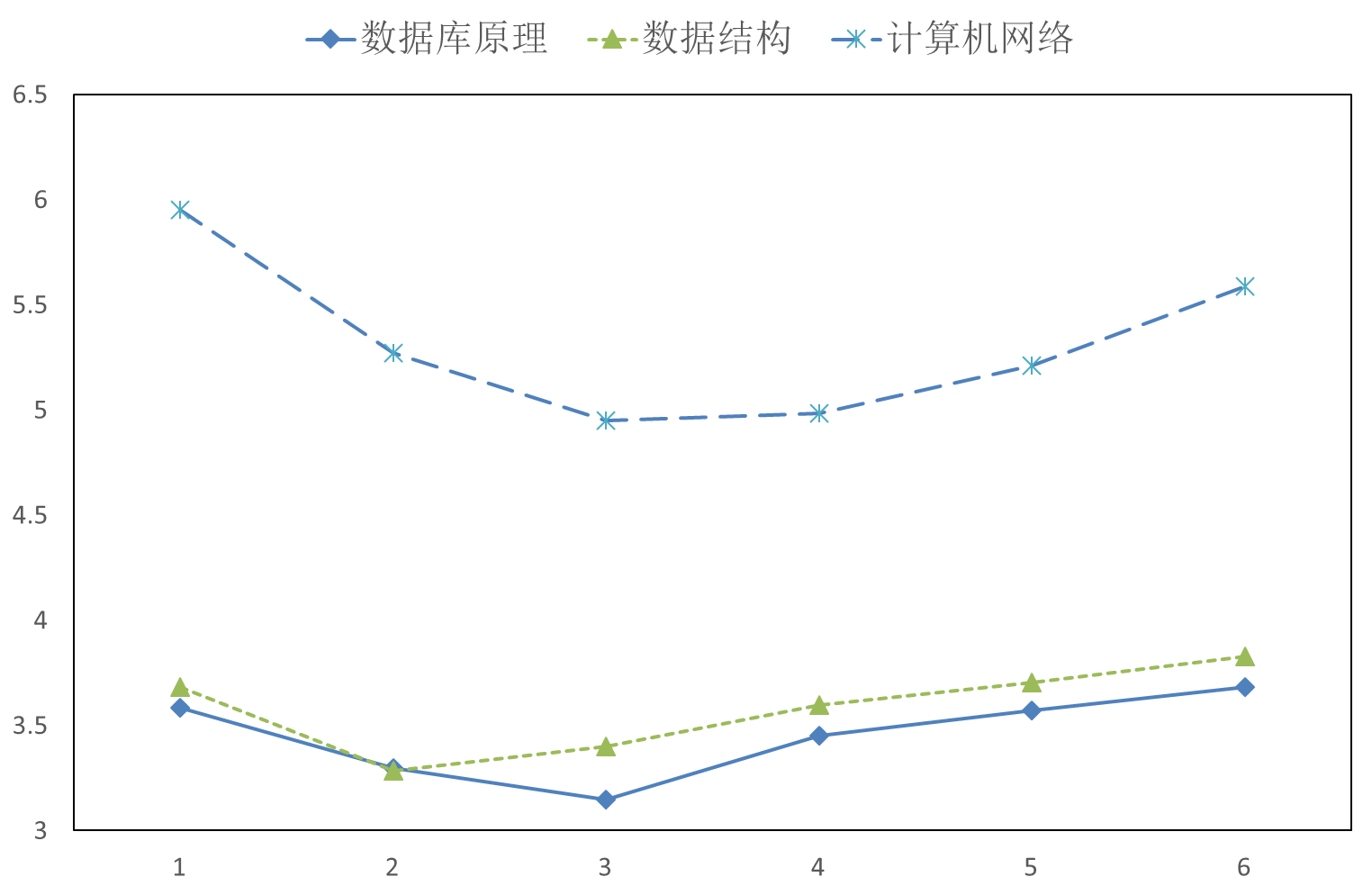


图3 *RPG*6的超参数敏感分析

Fig.3 Hyperparametric sensitivity analysis of *RPG*6

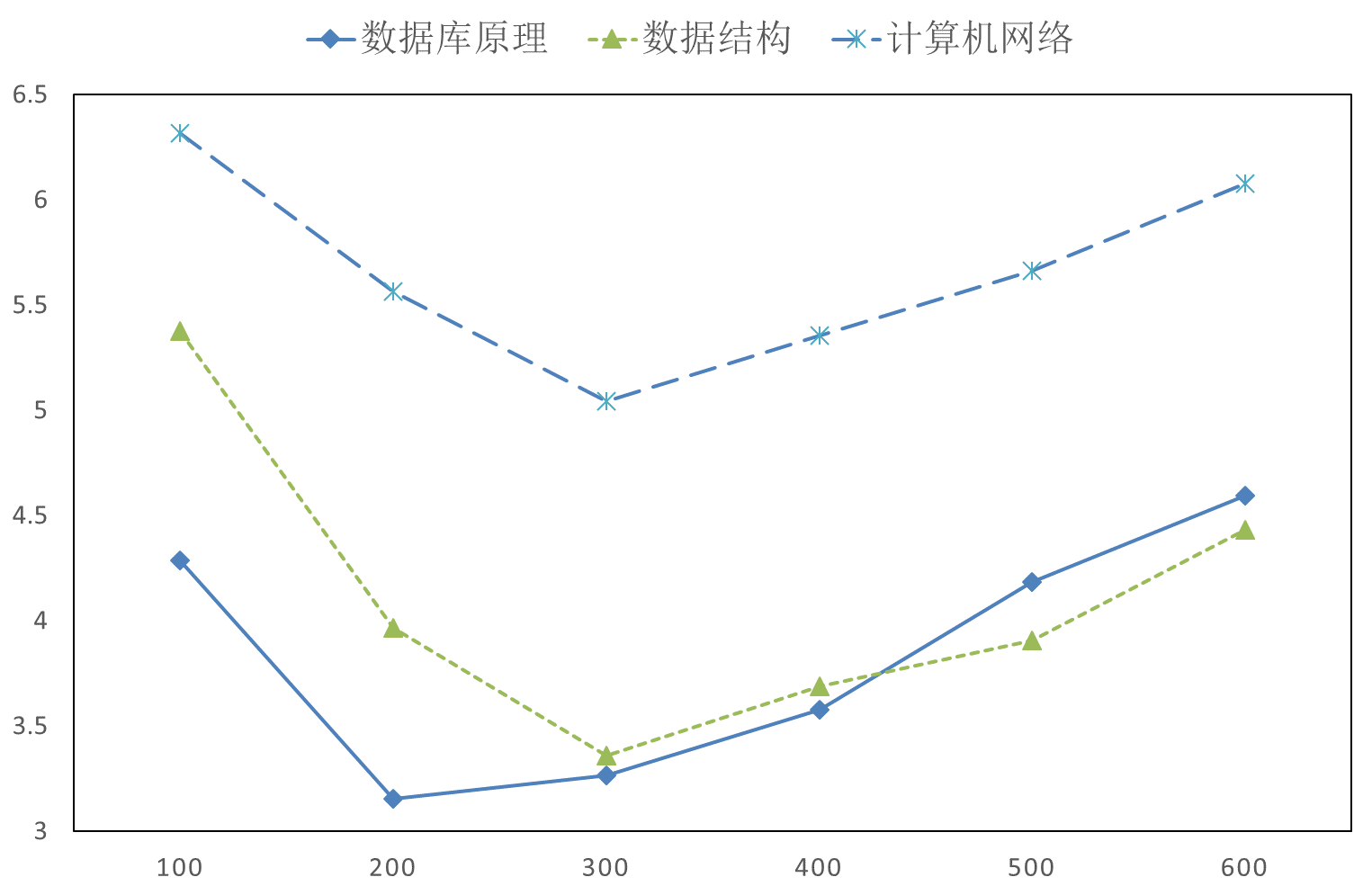


图4 更新评价能力的流程图

Fig.4 Hyperparametric sensitivity analysis of *RPG*7

**4结论**

本文提出了两个新的同行互评概率模型，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. Xu J, Liu J, Lv P, et al. Improving Peer Assessment Accuracy by Incorporating Grading Behaviors[C]//2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2021: 1162-1169.
21. 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.