

基于MOOC平台的互评质量控制的研究

张 忠¹, 宋继红¹, 付笑晗²

(1.沈阳工业大学, 辽宁 沈阳 110870;

2.东北大学, 辽宁 沈阳 110819)

摘 要:近年来,随着众包的发展,对工作者的准确率估计越来越受到关注。而作为一种特殊的众包形式,学生互评也被MOOC平台(大规模开放式在线课程)广泛应用。本篇文章研究MOOC课程中的在线互评机制,通过结合学生互评质量与学生的学习能力对其互评能力做出估计,以此设计互评分配算法。通过更精准地分配互评任务,得到更符合学生作业质量情况的互评结果。

关键词: MOOC; 推荐互评; 众包; 质量控制; 分配算法

中图分类号: TP391 **文献标识码:** A

Research on Quality Control of Peer Grading in MOOCs

ZHANG Zhong¹, SONG Jihong¹, FU Xiaohan²

(1. Shenyang University of Technology, Shenyang 110870, China;

2. Northeastern University, Shenyang 110819, China)

Abstract: In recent years, with the rapid development of crowd-sourcing, estimating the precision ratio of crowd workers has attracted more and more attention. As a special form of crowd-sourcing, peer grading has been used by most of Massive Open Online Courses (MOOCs). This paper studies peer grading mechanism in MOOCs, and estimates students' peer grading ability through a combination of their peer assessment and study ability, so as to design an allocation algorithm and achieve more accurate results in compliance with students' works by means of more accurate assignments of peer assessment.

Keywords: MOOC; peer grading; crowd-sourcing; quality control; allocation algorithm

1 引言(Introduction)

近年来,随着大型在线开放课程MOOC平台的流行,使得越来越多的人可以通过观看视频的形式学习具有大学水平的在线课程。然而,新型互联网科技虽然可以让学生获取视频资源,同时也使得对学生给出的复杂的、开放式的作业进行评估和反馈的能力大大受限,例如数学证明、设计问题和文章等形式的作业^[1]。相比较于学校中常规的课程,MOOC课程的规模是非常大的,每个课程通常超过20000名学生,因此,学生直接互评作业机制的引入势在必行^[2]。

同学间的互评之所以受到质疑和挑战,是因为学生的知识和能力与教师存在很大的差距,因此把单纯而随机地通过给学生分配互评任务得到的互评评价结果作为课程的反馈在一定程度上会遭到学生的质疑。而在大规模的在线课程中,学生的能力、语言又不尽相同,这可能给学生的互评带来新的挑战^[3]。因此,本文的主要思路是在没有黄金标准^[4]的条件下,结合学生以往的互评工作经验对学生的评价能力进行估

计,并设计互评分配算法。

2 互评质量控制的简述(Overview of quality control of peer grading)

学生互评本质上是一种众包,而众包是一种将任务通过互联网外包给个人的方法。这种方法将具有特殊需求的任务分布式地发放出去,其目的是结合群众的智慧。然而众包之所以受到很多争议,正是由于参与众包任务的工作者大多不是专业人士。如果任务涉及了专业的知识,那么并不能保证工作者们给出的结果的准确性^[5]。

目前,三大MOOC平台之一的Coursera已经将互评机制引入课程学习的过程中。学生的作业全部需要上传到服务器,每个作业包含一个评估准则,它描述了该作业的打分标准^[6]。在学生评阅之前,工作人员大约需要完成12份作业的评阅工作,其中8份用来对每个学生进行训练,剩下的4份用来估计学生评阅的准确率。学生在每个作业的评阅过程都包含两个阶段:校正和评价。

在学生完成校正过程后,便开始正式的评阅过程。在此过程中,学生会被分配五份作业进行评阅。但学生不知道这五份作业中的一份是工作人员所评阅过的作业,这份作业用来对学生评阅的准确度进行评估。综上所述,每个学生会收到四份随机分配的其他同学的作业,以及一份与工作人员共同评阅的作业。本算法应用于在线互评系统中,将学生评阅质量估计与学习能力结合起来进行学生评阅作业的分配,实现以更小的评阅数目得到更准确的评阅结果。

3 评阅质量估计方法及实验分析(Estimation method and experiment analysis on Quality of Peer Grading)

3.1 问题定义

每次发布的作业会有 k 个评分标准,而每个评分标准又分为 m 个不同的评分。因此每个学生对某一作业给出的反馈为一组 k 维的向量 S ,而每份作业又会交给随机选择的 n 个人进行评阅,那么作业 t_i 得到的反馈为一个由 n 个向量组成的集合,用 G_{ti} 表示该集合:

$$G_{ti} = \{g_{ti,s1}, g_{ti,s2}, \dots, g_{ti,sn}\}$$

$g_{ti,s1}$ 为作业 t_i 得到的学生 s_1 的评分。 $g_{ti,s2}[k]$ 为 $g_{ti,s2}$ 中的第 k 个元素,则 $G_{ti}[k]$ 表示集合 G_{ti} 中每个向量的第 k 个元素组成的集合:

$$G_{ti}[k] = \{g_{ti,s1}[k], g_{ti,s2}[k], \dots, g_{ti,sn}[k]\}$$

3.2 学生能力估计

在所有学生都完成评阅之后,我们可以得到由学生的评阅数据所产生的集合 G_{ti} 。若想对学生的评阅质量进行估计,那么需要得到每个作业的各标准分数。由于互评的特点,自身能力较强的同学倾向于给出较平均分更低的结果,而能力较弱的同学又有很大的几率给出较平均分更高的结果^[1]。每份作业的评阅者都会包含能力较强和较弱的学生,因此我们需要在 n 个评阅结果中将最高分和最低分去除,以将偏差较大的数据点去除。

定义1 $\bar{g}_{ti}$ 表示作业 t_i 的标准分数,则有:

$$\bar{g}_{ti}[k] = \frac{\sum_{j=1}^n g_{ti,j}[k] - \max(G_{ti}[k]) - \min(G_{ti}[k])}{n-2} \quad (1)$$

其中 $\bar{g}_{ti}[k]$ 为 $\bar{g}_{ti}$ 的第 k 个标准所得分数,是由集合 G_{ti} 所有向量的第 k 个元素的集合去除最大值和最小值后取均值得出。对所有 $k \in (1, n)$ 进行计算后可以得到一个 n 维向量 $\bar{g}_{ti}$,则将该向量作为作业 t_i 的标准分数,我们以此分数向量各元素的和来代表上传了作业 t_i 的学生 s 的能力估计值 a_s ,既:

$$a_s = \sum_{k=1}^n \bar{g}_{ti}[k] \quad (2)$$

3.3 评价质量估计

在前面介绍了对学生能力估计的方法,所得到的估计值

结果 $\bar{g}_{ti}$ 质上是作业 t_i 的标准分数,那么只要计算某学生给出的作业 t_i 的评分与标准分数之间的偏差,便可以对学生的评价质量进行估计。

而计算该偏差需要选取适当的距离函数。欧氏距离和余弦相似度被广泛应用于计算用户评分的相似度。但是余弦相似度函数只能判断出两个评分方向上的相似程度,而无法识别距离上的相似程度。因此,我们选用欧氏距离来计算学生评分与标准得分之间的相似度。设两个向量 $\alpha(a_1, \dots, a_n)$ 、 $\beta(b_1, \dots, b_n)$ 的相似度 $\text{dist}(\alpha, \beta)$ 为:

$$\text{dist}(\alpha, \beta) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2} \quad (3)$$

则学生 s_i 对作业 t_i 评价的偏差为 $\text{dist}(g_{ti,si}, \bar{g}_{ti})$ 。由于每份作业会被 n 名同学评阅,为了保证平均分配,那么每个学生同样需要评阅 n 份作业。即只要综合学生评阅的 n 份作业的偏差均值便可以求得学生评阅的总体偏差。

定义2 设 d_s 为学生 s 的总体偏差,则有:

$$d_s = \frac{\sum_{i=1}^n \text{dist}(g_{ti,si}, \bar{g}_{ti})}{n} \quad (4)$$

3.4 归一化和排序

依据3.2节和3.3节中所叙述的方法,我们可以分别计算出学生 s 的评价质量估计值 d_s 和学习能力估计值 a_s 。但所得 d_s 和 a_s 的标量却并不相同,因此并不能将两个估计值简单相加和比较,而是需要依据各自的范围区间分别对两个值进行归一化,将两个标量映射到一个相同的范围,以产生相互可比较的结果。

由于所有学生完成某次课程作业的互评工作后,评阅数据已经不再变动,即不会再有新生成的数据集加入,因此我们可以采用离差标准化的方法的同时将 d_s 和 a_s 归一化。用 D 代表所有学生的评价质量估计值 d 所产生的集合,则对学生 s 的归一化方法如式(5)所示:

$$d_s = \frac{d_s - \min(D)}{\max(D) - \min(D)} \quad (5)$$

同理,对 a_s 采用相同的方法归一化后,便可通过赋予相应的线性系数加权产生本次作业的评阅估计值 e_s :

$$e_s = k_1 d_s + k_2 a_s \quad (6)$$

在MOOC中,由于在一门课程的学习周期中教师会布置多次作业,若只考虑学生单次的评阅偏差,并不能准确地反映学生的真实能力。因此本文采用跟踪评阅质量的方法对学生的互评能力进行估计。将本次的评阅结果估计值和以往作业的估计值赋予一定的权值后结合起来,形成该学生的总体评阅估计值。用 S_k 代表学生 s 在完成第 k 次作业的评阅后得到的估计值 e_s 。

3.5 分层评阅

根据求得的总体估计值 e_s ，我们可以得到学生的评阅结果排名 R 。设某次作业参与评阅的学生总数为 n 。首先，在 R 中选取排名靠前的 k 个学生作为上层学生，其他的学生作为下层学生，则我们的目的是尽可能地让上层学生和下层学生按比例分配作为每份作业的评阅者，即假设每份作业的评阅人数为 m ，那么对于任意提交的作业，我们要选取 $\left\lfloor \frac{m \times k}{n} \right\rfloor$ 或 $\left\lfloor \frac{m \times k}{n} \right\rfloor + 1$ 名上层学生来评阅。例如每份作业要求10人来评阅，而我们从排名中选取前60%的学生作为上层学生，则每份作业应该由6名上层学生和4名下层学生评阅。

在整个分配的过程中，我们采用了同班级回避的方法。因为如果不采取这样的策略，那么可能会存在同一个班级甚至同一个寝室的学生相互评阅。若学生在完成作业的过程中存在相互讨论的情况，那么会影响最终评阅的准确性。

(1) 实验数据集描述

在本实验中，我们共计选取789名学生参加此次实验。这些学生共同参与大学计算机相关课程，共有48课时，经历12周完成教学任务。其中在课程中期依次布置三次作业：数据处理作业、图片处理作业和科技论文写作。数据集统计见表1。

表1 数据集统计

Tab.1 Statistics of data set

	数据处理作业	图片处理作业	科技论文写作
学生人数	677	789	533
作业评阅数	10	10	10
评价标准	10	8	7
打分总数	20144	23016	15776

(2) 实验过程

每次作业的评阅过程中，对学生提交的作业会分配10名学生进行评阅。而每次作业会由教师根据课程的需要制定7—10个评分标准，学生会根据作业的完成质量在评阅时对每个评分标准给出0—5的打分。

在课程的第一次作业中，老师会将课程的作业发布在网上。然后通过班级回避的随机分配算法分配学生之间的评阅关系。最后，学生登录互评系统并完成在线评阅。整个过程是单盲实验，学生既无法知道他所评阅的作业来自哪一位同学，又无法知道自己的作业被哪些同学评阅。这会尽量使得学生给出的评阅分数不存在作弊的行为。第一次作业所得到的结果作为下一次作业分配算法的输入数据。

在第二次作业中，为了直观地比较实验结果，我们通过学号把这些学生分为四组，其中第一组不做任何特殊的处理，只是依据班级回避的策略进行随机分配。而第二、三、四组学生使用第一次作业的评阅结果作为上述分层评阅的

配算法的输入数据，对学生进行排序和分配。其中第二组学生依据第一次作业的评阅结果，选取60%的学生作为上层学生；第三组选取70%的学生作为上层学生；第四组选取80%的学生作为上层学生。而代表学生能力估计和评价能力所占比重权值 k_1 和 k_2 ，则各设置为0.5。

而在第三次作业中，仍然沿用第二次作业的分组方法进行评阅分配，但与其不同的是将前两次作业的评阅估计结果作为输入数据，实现迭代的计算过程。由于相比较来说，学生当次作业的评阅结果要比之前的评阅结果更为重要，因此我们设置权值 $a_1=0.25, a_2=0.75$ 。

(3) 实验结果

得到学生的评阅结果后，我们通过计算每个学生各个评分标准所得所有分数的均值作为该标准所得的最终分数。设学生 s_1 对学生 s_2 的某一评分标准 t_i 的评分为 g_1 ， s_2 的 t_i 标准的最终得分为 g ，用 ε 表示该分数的偏差，则 $\varepsilon = g_1 - g$ 。假设某学生的某个标准的最终得分为3.8分，那么无论是4或5分都是正确的分数，即 $|\varepsilon| < 1$ 。因此，我们可以根据所有学生评阅的打分情况统计出其中给出正确结果的学生票数的比例。更高的正确比例代表着这组学生有着更为一致的评阅结果，其评阅准确程度也会更高。实验结果详见表2和表3。

表2 第二次作业实验结果

Tab.2 Results of the second operation experiment

实验组	$ \varepsilon < 1$	$1 \leq \varepsilon < 2$	$2 \leq \varepsilon < 3$	$3 \leq \varepsilon < 4$	$ \varepsilon < 5$
普通组	77.3%	17.7%	2.5%	2.1%	0.4%
分层前60%	81.5%	14.7%	2.2%	1.5%	0.1%
分层后40%	75.6%	17.9%	3.2%	2.9%	0.4%
分层前70%	78.8%	17.6%	1.7%	1.6%	0.3%
分层后30%	74.9%	19.9%	2.5%	2.3%	0.4%
分层前80%	80.9%	15.8%	1.3%	1.7%	0.3%
分层后20%	77.3%	19.5%	1.7%	1.2%	0.3%

表3 第三次作业实验结果

Tab.3 Results of the third operation experiment

实验组	$ \varepsilon < 1$	$1 \leq \varepsilon < 2$	$2 \leq \varepsilon < 3$	$3 \leq \varepsilon < 4$	$ \varepsilon < 5$
普通组	76.8%	18.2%	3.3%	1.5%	0.2%
分层前60%	81.5%	13.4%	2.8%	2.1%	0.2%
分层后40%	63.8%	25.3%	6.0%	4.7%	0.2%
分层前70%	78.1%	18.3%	2.3%	1.2%	0.1%
分层后30%	69.1%	24.5%	3.7%	2.1%	0.6%
分层前80%	76.3%	18.1%	3.1%	2.2%	0.3%
分层后20%	63.2%	26.8%	4.0%	4.9%	1.1%

可以看出在根据第一次作业计算排名后所分出的排名靠

前的学生相比较靠后的学生普遍具有更高的一致性。其中将上层学生的百分比设置为60%时取得了较好的结果, $|\varepsilon| < 1$ 的比例相差5.9%, 而当选取80%的学生作为上层学生时, 该比例相差3.6%。

而在第三次作业的评阅中, 由于结合了第二次的评阅结果估计值以及当前的评阅情况, 因此产生了好的结果。可以看出, 设置60%的上层学生依然产生了最好的结果, 上层学生和底层学生 $|\varepsilon| < 1$ 的比例相差17.7%, 而相差最小的是按照7:3分配上下层学生的9%, 大于第二次作业中的相差的最大值5.9%。这说明结合以往的估计值会大大提升算法的鲁棒性, 可以更好地识别出评阅更为准确的学生。而将上层学生和底层学生的比例设置为6:4时得到了最好的结果, 因而可以看出算法对评阅

4 结论(Conclusion)

本文通过结合学生评阅能力和学习能力来对互评中学生的评阅准确性做出估计, 准确地识别出评阅工作中的上层学生和底层学生。在后续的工作中, 我们将研究如何通过算法所产生的估计值转化为对学生评阅准确率的估计, 进而将此算法应用于更为一般的众包系统中, 实现对众包工作者的工作能力估计, 以及实现使用更少的上层众包工作者来实现较为准确的结果, 以此来降低发布众包任务的成本。

参考文献(References)

[1] Kulkarni.C, Wei K.P, Le H. Peer and self assessment in massive

(上接第13页)

算法。该算法与适用性高的基于Voronoi图LOD算法相比, 时间复杂度低, 减少数据量的同时, 缩短了服务器响应时间, 从而提高了GPS轨迹数据动态可视化效率。

参考文献(References)

- [1] Mao Y, Zhong H, Qi H, et al. An Adaptive Trajectory Clustering Method Based on Grid and Density in Mobile Pattern Analysis[J]. Sensors, 2017, 17(9): 2013.
- [2] Cai L, Zhou Y, Liang Y, et al. Research and Application of GPS Trajectory Data Visualization[J]. Annals of Data Science, 2017(4): 1-15.
- [3] Liu Y, Kang C, Gao S, et al. Understanding intra-urban trip patterns from taxi trajectory data[J]. Journal of Geographical Systems, 2012, 14(4): 463-483.
- [4] 贾奋励, 宋国民. 电子地图显示中点状要素LOD模型的建立[J]. 测绘学院学报, 2002(01): 62-64.
- [5] 李佳田, 康顺, 罗富丽. 利用层次Voronoi图进行点群综合[J]. 测绘学报, 2014(12): 1300-1306.
- [6] 毋河海. 凸壳原理在点群目标综合中的应用[J]. 测绘工程, 1997(01): 1-6.

online classes[J]. ACM Transactions on Computer-Human Interaction, 2013(39): 1-32.

- [2] Joglekar.M, Garcia-Molina.H, Parameswaran.A. Evaluating the crowd with confidence[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2013: 686-694.
- [3] Guo.S, Parameswaran.A, Garcia-Molina.H So who won?: dynamic max discovery with the crowd[C]. ACM SIGMOD International Conference on Management of Data. ACM, 2012: 385-396.
- [4] Welinder.P, Perona.P. Online crowdsourcing: Rating annotators and obtaining cost-effective labels[C]. Computer Vision and Pattern Recognition Workshops. IEEE, 2010: 25-32.
- [5] Bellare.K, Iyengar.S, Parameswaran.A, et al. Active Sampling for Entity Matching with Guarantees[J]. ACM Transactions on Knowledge Discovery from Data, 2013, 7(3): 12.

作者简介:

张 忠(1990-), 男, 硕士生. 研究领域: 推荐系统.

宋继红(1963-), 女, 硕士, 副教授. 研究领域: 计算机网络通信, 计算机网络远程控制, 嵌入式技术.

付笑吟(1990-), 男, 硕士, 工程师. 研究领域: 大数据信息安全.

[7] Hans-Uli Feldmann. Cartographic Generalisation-Topographic Maps[M]. Swiss: Swiss Society of Cartography, 2005.

[8] 闫浩文, 王家耀. 基于Voronoi图的点群目标普适综合算法[J]. 中国图象图形学报, 2005(05): 633-636.

[9] Kreveld M J V, Oostrum R W V, Snoeyink J. Efficient Settlement Selection for Interactive Display[J]. In Proc. Auto-Carto 13: ACSM/ASPRS Annual Convention Technical Papers, 1997: 287-296.

[10] Sadahiro Y. Cluster Perception in the Distribution of Point Objects[J]. Cartographica the International Journal for Geographic Information & Geovisualization, 1997(03): 49-62.

[11] 李雯静, 李少宁, 龙毅, 等. 利用重力模型进行GIS点群选取[J]. 武汉大学学报(信息科学版), 2013, 38(8): 945-949.

[12] 艾廷华, 刘耀林. 保持空间分布特征的群点化简方法[J]. 测绘学报, 2002, 31(2): 175-181.

作者简介:

孙泽昌(1990-), 男, 硕士, 助理工程师. 研究领域: GIS与铁路选线.