论文概述

第三章

理想同行分级系统：

(1)提供高度可靠和准确的评估

(2)分配系统分配给学生和staff的工作负载非常均衡

(3)可伸缩班级规模数万或数十万学生

(4)广泛应用于不同的收集问题的设置

模型存在假设变量：

真实分数(True scores)：我们假设每一个提交的u都与一个真实的潜在分数(表示为s_u)相关联，s_u是未观测到的，需要进行估计。

评分者偏见（Grader biases）：每一个评分者v都有一个偏差，$b_v \in R$。这些偏差变量反映了评分者在一定百分比内夸大或缩小其评估结果的趋势。

评分者可靠性（Grader reliabilities）：对评分者的可靠性进行建模，$t_v \in R^+$，反映了评分员的同行评估在校正偏倚后与提交的真实分数的平均接近程度。在下面的模型中，t_v总是指正态分布的精度或逆方差。

观察到的实际分数（Observed grades）：z_v^u表示评分者v对于提交的u的评估分数。

三种模型

第一种模型PG_1

PG_1将先验分布置于潜在变量上，并假设，例如，当一个年级学生的偏差可能是非零，许多年级学生的平均偏差是零。

$$
\begin{aligned}
\text{(Reliability) } \tau_v &\sim \mathcal{G}(\alpha_0, \beta_0) \text{ for every grader } v, \\
\text{(Bias) } b_v &\sim \mathcal{N}(0, 1/\eta_0) \text{ for every grader } v, \\
\text{(True score) } s_u &\sim \mathcal{N}(\mu_0, 1/\gamma_0) \text{ for every user } u, \text{ and} \\
\text{(Observed score) } z_u^v &\sim \mathcal{N}(s_u + b_v, 1/\tau_v), \\
&\qquad \text{for every observed peer grade.}
\end{aligned}
$$

在我们的实验中，我们也考虑了一个简化版本的PG1模型，其中每个年级的信度被固定为相同的值。我们引用这个更简单的模型，在这个模型中，只允许评分者的偏差作为pg1偏差变化。

第二种模型PG_2

这个模型引入了时间的相关性，通过某评分者过去的作业与评分数据来进一步确定评分者的偏执和可行度。但是我们需要证明过去的评分记录是否可以作为现在评分者的参考。

$$
\begin{aligned}
\tau_v^{(T)} &\sim \mathcal{G}(\alpha_0, \beta_0) \text{ for every grader } v, \\
b_v^{(T)} &\sim \mathcal{N}(b_v^{(T-1)}, 1/\omega_0) \text{ for every grader } v, \\
s_u^{(T)} &\sim \mathcal{N}(\mu_0, 1/\gamma_0) \text{ for every user } u, \text{ and} \\
z_u^{v,(T)} &\sim \mathcal{N}(s_u^{(T)} + b_v^{(T)}, 1/\tau_v^{(T)}), \\
&\qquad \text{for every observed peer grade.}
\end{aligned}
$$

这个模型将b_v使用过去得到的偏执信息。模型pg2要求我们将不同家庭作业的成绩标准化为一个一致的尺度。例如，在我们的实验中，我们注意到在不同的家庭作业中，评分者的偏差集有不同的方差。然而，使用标准化分数(z-score)允许我们传播学生的潜在偏见，同时保持对作业工件的稳健。请注意，虽然可以类似地想象出一个模型，它捕捉真实分数和分配的可靠性的动态，但我们只关注了这项工作的偏差动态(这对提高准确性贡献最大，同时仍然是公平的)。

第三种模型PG_3

由于评分者自己也可以在互评中获得分数，我们可以考虑，一个得分越高的评分者是否更具有可信度。

$$
\begin{aligned}
b_v &\sim \mathcal{N}(0,\, 1/\eta_0) \text{ for every grader } v, \\
s_u &\sim \mathcal{N}(\mu_0,\, 1/\gamma_0) \text{ for every user } u, \text{ and} \\
z_u^v &\sim \mathcal{N}\left(s_u + b_v,\, \frac{1}{\theta_1 s_v + \theta_0}\right), \\
&\qquad \text{for every observed peer grade.}
\end{aligned}
$$

于是在第三个模型中，我们得知用户的可信度可以由该用户所得的评价分数得到。

模型pg3的约束更大，迫使分级器的可靠性依赖于单个参数，而不允许任意变化，从而防止我们的模型过拟合。

注意：评分模型的准确性并不是模型追求的最终标准，而是公平与公正性。例如一个黑人可能在实际的互评中可以获得更高的分数，但是这并不是我们的模型想要改进的目标。并且对于学生的历史作业成绩，即使这些分数展示出了很强的历史相关性（皮尔森系数为0.46），我们也决定不让它参与模型，因为这样会影响公平性。

模型pg3允许提交的真实分数依赖于评分者的分数，这看起来可能有争议，但这种依赖是弱的，只影响特定评分者对最终预测的影响，这是可取的。我们发现，当一个学生的成绩会受到她作为评分员的评估与其他评分员的评估是否一致的影响，他们会更加认真的进行评估。因此，pg3模型允许学生的成绩取决于他们作为评分者的表现，作为评分机制，可以激励学生在评分上投入更多的努力。

模型的改进（Inference and evaluation)

我们可以将评分者的偏执、真实分数和可信度作为后验概率，但是，由于所有的变量都是相互关系的，因此在推理中就出现了一个循环。例如，如果一个只有知道用户的真实分数，可以更好地推测出这个用户的偏差；但是为了获得这个好的偏差，我们需要了解该用户的真实分数。

真实分数的标准

一开始我们将staff的评分作为真实分数，但是我们发现，用学生们的共识评价分数作为真实分数有更加良好的效果。因为当staff的评分与学生的共识分数不一致时，学生的评分更加偏向于共识分数。

接下来我们进行了模型的对比，在对照组模型中，给学生的分数是他们收到的四个同辈分数的中位数。但这些对照组模型没有考虑个别评分者的偏差和可靠性，也不包含关于真实成绩分布的先验知识。

第四章 实验

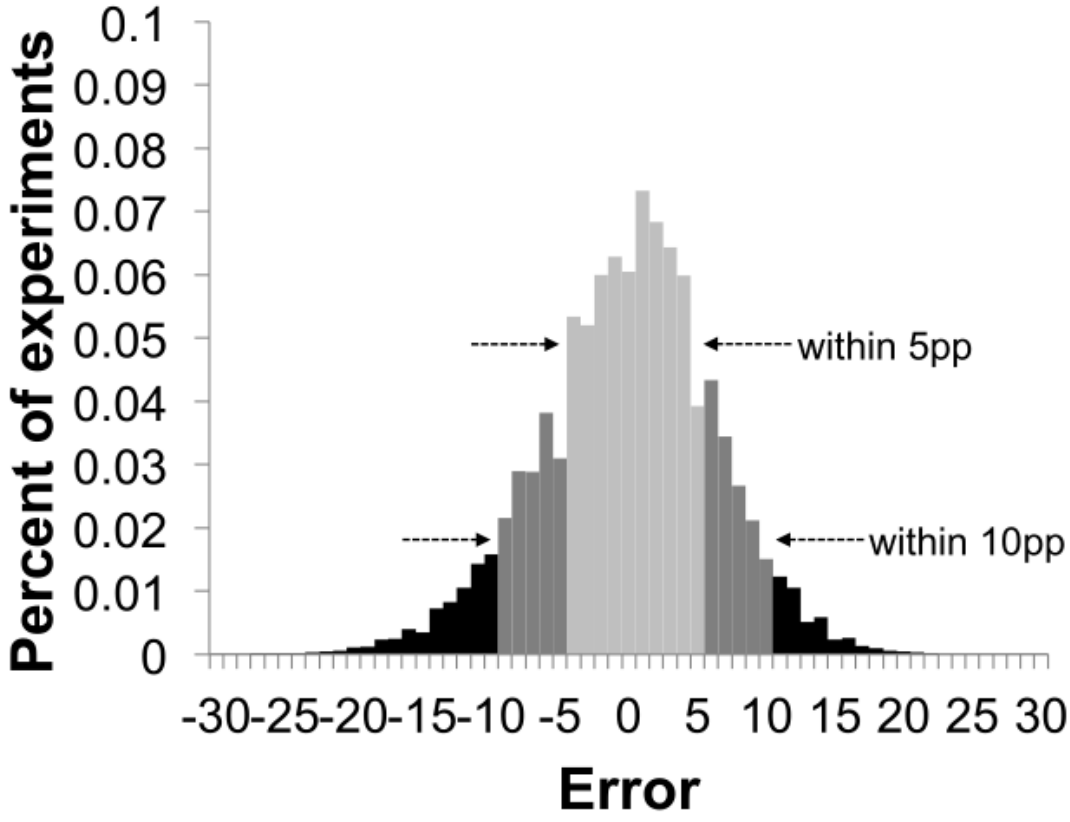
对照效果如表二：

Table 2: Comparison of models on the two HCI courses

	HCI 1					HCI 2				
	Baseline	$\mathbf{PG_1}$**-bias**	$\mathbf{PG_1}$	$\mathbf{PG_2}$	$\mathbf{PG_3}$	Baseline	$\mathbf{PG_1}$**-bias**	$\mathbf{PG_1}$	$\mathbf{PG_2}$	$\mathbf{PG_3}$
RMSE	7.95	5.42	5.40	5.40	**5.30**	6.43	4.84	4.81	4.75	**4.73**
% Within 5pp	51	69	69	**71**	70	59	72	73	73	**74**
% Within 10pp	81	92	94	94	**95**	88	96	96	**97**	**97**
Mean Std	7.23	5.00	4.96	4.92	**4.77**	6.19	4.57	**4.52**	4.53	**4.52**
Worst Grade	-43	-34	**-30**	-32	**-30**	-36	-26	-26	**-25**	-26

baseline方法是直接取中位数。比起对照方法，RMSE降低了，最终分数与实际误差在5%和10%的比例也降低了。

由于课程改进，我们观察到HCI2组的学生与HCI1组的学生相比，在评分时表现得更加一致。注意到在HCI1上运行的每个模型在每个指标上都优于在HCI2上运行的基准评分系统，这表明同级评分的最佳收益可能来自改进的类设计和统计建模。


Percent of experiments
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
within 5pp
within 10pp
-30 -25 -20 -15 -10 -5 0 5 10 15 20 25 30
Error

(a)