第三章 模型

基数评估模型

PG_4和PG_5是PG_3的变种，他们将模型中评分者的可信性和评分者的真实分数建立了一个概率模型。

其中PG_4是伽马分布模型，PG_5是高斯分布模型。

如果评分者没有在互评中提交作业，则评分者的可信性将会被看成是最低。

数据集

模型的性能评估

模型使用root-mean-square error (RMSE)均方根误差来评估性能，计算方法是通过教师的打分和模型的预测分数。

表三显示了模型运行了十次后的结果，Median表示取中位数的方法。

	Assignment 1		**Assignment 2**		**Assignment 3**	
	Mean	*Std*	*Mean*	*Std*	*Mean*	*Std*
Median	4.94		5.54		4.12	
PG_1	3.77 (23%)	0.02	4.93 (11%)	0.03	3.66 (11%)	0.01
PG_3	**3.22 (35%)**	0.02	5.24 (5%)	0.04	3.15 (23%)	0.02
PG_4	3.35 (32%)	0.05	4.75 (14%)	0.06	2.83 (31%)	0.09
PG_5	3.31 (33%)	0.05	**4.69** (15%)	0.05	**2.76 (33%)**	0.09

Table 3: Experimental results for cardinal models. Median represents taking the medium of the peer grades. PG_1 and PG_3 are models proposed in (Piech et al. 2013) and PG_4 and PG_5 are our models described above.

从实验结果可以看出：在Assignment1中PG_3的性能要优于PG_4、PG_5，但是在Assignment2、Assignment3中PG_4和PG_5的性能更好。

这说明了将评分者的真实分数与可信性的关系转化为概率模型的效果要优于线性模型。

模型的最坏情况与敏感度

我们将每个模型预测的分数与教师的评估分数进行对比，找出其中的最大值，结果如表4.

	Assignment 1	Assignment 2	Assignment 3
PG_3	6.52	11.10	6.77
PG_4	5.84	9.86	6.70
PG_5	**5.81**	**9.85**	**5.79**

Table 4: Maximum prediction deviation from the ground truth for the optimal settings in Table 3.

从表看出：PG_3的偏置要高于PG_4和PG_5，其中PG_5的效果最优秀。

但是在PG_3中，评分者的观测分数将会受到评分者的真实分数和被评者的真实分数的双重影响，尽管PG_3的表现在Assignment1中的表现要优于PG_4和PG_5.但是在PG_5和PG_4中，评分者的真实分数只会影响到他的可信度。因此在新的模型中，真实分数对评估分数的敏感性要小于PG_4。

序数评估模型

序数评估模型使用了Bradley-Terry、RBTL、BT+G.

序数模型和**基数模型**的结合公式如下：

$$\mathcal{L} = \frac{\lambda}{2\sigma^2} \sum_{u \in U} (s_u - \mu_u)^2 - \sum_{v \in V} \sum_{s_{u_i} \succ_{\rho(v)} s_{u_j}} \log(\text{hypothesis})$$

其中，预测分数μ_u由基数模型得出，最终求得预测准确率。

假设由BLT模型生成：

$$\text{hypothesis} = P(u_i \succ_{\rho(v)} u_j) = \frac{1}{1 + \exp(-(u_i - u_j))}$$

最终，根据预测的准确性，我们给出了测试结果：

	Assignment 1	Assignment 2	Assignment 3
Cardinal Models			
PG_3	0.7526	0.6155	0.7775
PG_4	0.6928	0.6552	0.7854
PG_5	0.6979	0.6616	0.7889
“Cardinal + Ordinal” Models			
PG_3+BT	0.7577	0.6110	0.7892
PG_4+BT	0.7221	0.6484	0.7931
PG_5+BT	0.7191	0.6646	0.8000
PG_3+BT+G	0.7645	0.6587	0.7879
PG_4+BT+G	0.7145	0.7032	0.7896
PG_5+BT+G	0.7170	**0.7065**	**0.8013**
PG_3+RBTL	**0.7660**	0.6494	0.7979
PG_4+RBTL	0.7064	0.6745	0.7835
PG_5+RBTL	0.7201	0.6845	0.8009
Pure Ordinal Models			
BT (or BTL)	0.6536	0.6329	0.6896
RBTL	0.6583	0.6432	0.6996
BT+G	0.6547	0.6535	0.7009
BT Same Initial	0.6387	0.6194	0.6407
BT Random Initial	0.6381	0.6416	0.6667
Baseline Method			
Median	0.6043	0.6610	0.6753

尽管纯序数模型会忽略评分信息，但是最终效果与中位数的准确性相差不大。

消融实验

基数模型的实验结果始终好于序数模型，这是因为在基数模型中，使用了更加细粒度的数值，比序数模型的二进制比较效果更加优秀。

我们可以将基数模型的分数信息解释为用户的偏好信息，再借助于序数模型，确定这些偏好，从而得出更好的结果。

为了防止这样的结果被解释为基数模型自身的优势，我们又使用基数模型的评估方法测试了基数+序数模型，测试结果如下：

	Assignment 1	**Assignment 2**	**Assignment 3**
PG_3+BT	3.04	5.30	3.18
PG_3+BT+G	3.01	**4.95**	**3.10**
PG_3+RBTL	**3.00**	5.04	3.15
PG_4+BT	3.47	4.87	3.03
PG_4+BT+G	**3.31**	**4.52**	2.91
PG_4+RBTL	3.44	4.70	**2.77**
PG_5+BT	3.30	4.77	2.93
PG_5+BT+G	3.35	**4.50**	2.74
PG_5+RBTL	**3.24**	4.62	**2.70**

Table 6: Cardinal evaluation (RMSE) results for "Cardinal+Ordinal" models

从图中可以看出，两个模型的结合可以比纯基数模型获得更好的结果。