Peer Grading

Mert Can ÇIKLA, Emre BEKTAŞ

December 16, 2014

Abstract

Massive open-access online courses serve the purpose of providing free, high level education in vast scales. Peer assessment is an important part of these courses where the number of peers can scale up to tens of thousands. However peer assesment is still far perfect from getting on level with instructor grading. Bias of the peers and their accuracy in grading another peer are matters that require work to make peer assessment more usable in practice. In the case where the assignments are open to interpretation and can not be graded in an automated manner, various peer grading methods are needed. In this study we delve into some highly popular peer grading methods. We identify and differentiate them considering their way of approaching the problem. We experiment on various forms of synthetically generated data in order to measure their performance in a simulated environment. We propose three ideas to improve accuracy and reduce error in peer grading.

1 Introduction

In recent years, massive open-access online courses (MOOCs) have gotten immensely popular and being free and requiring only an internet connection are the main reasons why. It allows people from all across the world to enroll and study at their own pace. Although this means online education can be provided in huge scales, there is however a problem that arises if the students need to be examined. In a case where the exam can not be assessed by a computer whether it is due to its type —such as an essay— or the nature of the field of study. Most popular solution to this problem is peer assesment where every peer grades certain amount of other peer's exams which in the end will build up to be that peer's grade for that exam. Peer grading is the most efficient way to evaluate when the peer count reaches thousands but it has it's flaws. Peers with no incentive to spend time on grading is a huge problem in peer assessment. Various studies [**?**, **?**, **?**]have tried to incentivize this effort by giving a portion of the final grade to the grader depending on how accurately they have graded compared to others. Due to the vast scale of the peers and their different backgrounds, this does not always suffice to reduce error. It might be that one does not want to spend time on grading and grades randomly or one might not be knowledgeble about a certain subject and may not grade others precisely. There are a lot more similar challenges to be overcome in this case where you have next to no information about how skilled a peer is in evaluating so there are a lot of unknown variables in the equation.

In this study we identify the problem of peer grading and review the solution methods that have been proposed and implemented in the literature. In the implementation details section we carefully explain our methodology and our algorithms we use when testing on the current methods. In experiments section we give detailed graphs of the experimentations we held and analyze the methods on various forms of synthetic data. Finally we identify some of the issues with peer grading and try to point out possible improvements that are to be implemented and tested in the future.

2 Literature Review

There have been various studies to increase the use of Peer Grading by reducing the error induced in evaluation. Walsh's [**?**] adapts PageRank —the website ranking algorithm used by Google— [**?**] to peer grading. **PeerRank** method constructs a matrix composed of grades and computes peer grades depending on their own grade and their bias as a grader. The GeneralizedPeerRank adds the notion of reliability that couples grader's accuracy in grading with their own grading to incentivize correct grading and reduce error. Another method that incorporates bias and grader performance is **CrowdGrader** [**?**]. It combines the grades provided by the students into a consensus grade for each submission students make by utilizing an algorithm that relies on a reputation system. The algorithm computes a resulting grade for each submission students make by weighing the students input grade by their accuracy which are then used to update each students estimated accuracy of grading.This algorithm seems to outperform median and average reliabilty computation techniques but shows mixed results over real life data. The authors also questioned whether it is best to use solely ranking or grading.

Piech et al. [**?**] defined 3 statistical models for peer grading. The first one **PG1** attempts to detect grader's bias and compensate accordingly. Authors try to make use of any possible coherence between a peer's performancwe on two different assignments at different times with **PG2**. **PG3** relates a peer's evaluation performance and grade like the method in [**?**]. They have experimented with these methods on a dataset obtained from a sizeable MOOC and report significant improvements.

Another approach used in the literature is to work on rankings instead of a cardinal grade. This approach eases the load of students by asking for pairwise comparisons or a set of rankings but lacks in the value of information retrieved. The method in [**?**] compares ranking methods to those who only grade and show cases where making pairwise comparisons reduce error in comparison to cardinal evaluation. Their work extends the Bradley-Terry-Luce Model [**?**, **?**] to peer grading and report ordinal evaluation to be more robust to lack of grader expertise.

The work of [**?**] tackles the Peer Grading as a rank aggregation and extends works of [**?**, **?**, **?**, **?**]. They evaluated those methods on data gathered from a University course and show that Ordinal methods are in some cases superior to PG1 of [**?**].

3 Implementation Details

We have prepared a Java program that enables us to run experiments with varying number of participants. In order to generate this synthetic

we have used the following algorithm.

Algorithm 1 Grade Assignment Algorithm

```
gen(s)                                   ▷ s: size of class, w: workload
S_i ← G_{i+1}..G_{i+5}                   ▷ students i + 1..i + 5 grade student i
for i ← 1, s do
    rg_i ← N(70, 30)                     ▷ real grade of the student
    if rg_i < 0 then
        g_i ← 0
    else if rg_i ≥ 100 then
        g_i ← 100
    else
        g_i ← rg_i
    end if
end for
```

Algorithm 1 creates a classroom of s students and assignes them w graders and a grade sampled from a normal distribution $\mathcal{N}$ with $\mu = 70$ and $\sigma = 30$. Also note that we have also experimented with various other distributions.

Algorithm 2 Sum of Binomials Marking Model

```
review(s)                                ▷ s: size of class, w: workload
for all G_{i,j} do                       ▷ jth grader of the student i
    α = B(rg_i, rg_{G_{i,j}})            ▷ B(n, p) n trials and p probability
    β = B(100 − rg_i, 1 − rg_{G_{i,j}})
    g_{i,j} = α + β;
end for
```

2 was proposed by Walsh in [**?**] and models a very intuitive way of assigning grades. Aside from the α component which is straight-forward and represents a student grading correctly a question that is correct. The β component of the mark is added by getting his wrong solution incorrectly graded as correct. However an issue arrises with this model that, when a student with a real grader lower than 50 grades a similar grade student they grade themselves much higher than possibly anticipated. For instance, assume the grade of every student in the classroom is 20, this yields on average a marking of 80 $bin(20, 0.2) = \alpha \approx 16 + bin(80, 0.8) = \beta \approx 64$ even though one would expect an average grade of 20. We have implemented **Algorithm 1** and **2** to measure performance of different peer grading algorithms in the literature. The PeerRank and GeneralizedPeerRank methods [**?**] were implemented by us. To measure the algorithm in [**?**] we used their implementation. The ranking methods described in [**?**] were also measured using the author's implementation after verification.

4 Experiments

Throughout the experiments PR refers to Peer Rank [**?**] and GPR refers to Generalized Peer Rank from the same paper. CG refers to CrowdGrader of [**?**]. MAL refers to Mallow's model originally proposed by Mallow [**?**] and adapted to peer assessment by [**?**]. MALS is the score-weighted Mallow's model and MALBC is Mallow's model with Borda count approximation also proposed by Raman and Joachims [**?**]. BT which stands for Bradley-Terry Model, THUR abbrevation of Thurstone model and PL the Plackett-Luce Model are also explained by Raman and Joachims in [**?**]. NCS is the PG1 model proposed by Piech et al. [**?**] but with a maximum likelihood estimator instead of Gibbs Sampler.

Our first set of experiments figures 1 to 4 are based on the experiments conducted by Walsh in [**?**]. We have created a synthetic dataset using 1 and 2. The dataset has 500 students with each student assigned to grade 5 other. Note that the original experiment by Walsh used 10 students all grading eachother, with that said our results verify those of Walsh in [**?**] and show almost exact results.

Figure 1: Binomial distribution with varying probability.

Figure 2: Normal distribution with increasing stdev.

Figure 3: Uniform distribution with increasing lower bound.

Figure 4: Normal distribution with varying grader bias.

In order to compare ordinal methods' performance with cardinals' we have conducted the following experiments using grades coming from Binomial, Normal and Uniform distributions and we have measured their Kendall-τ distances. Kendall-τ is the total miss ordering of a method compared to the realgrade(instructor) ordering.

Figure 5: Binomial distribution with varying probability.

Figure 6: Normal distribution with increasing stdev.

Figure 7: Uniform distribution with increasing lower bound.

The experiments show the Generalized PeerRank method proposed by [?] outperforms others when the grades are Normally or Uniformly distributed. However the method seems to be average when the grades are Binomial.

5 Future Work

First possible improvement we have theorised is to improve the Generalized PeerRank by looking at the cause of sub-par performance when the grades are Binomially distributed as it is odd for the results to be vastly different although the Binomial and Normal cases map a very similar distribution of grades overall.

Our second idea for improvement is to gather new information about students by splitting the exam into parts according to a rubric or simply grading each question seperately. Consider the case where a student has gotten a full-mark from the first question but zero marks from the second. In this case the student can be a much more effective as a grader if we can assign the student to only grade the first question for this exam instead of the exam as a whole. This methodology is not limited to exams that can be divided into questions. Any exam with a rubric is an applicable case since for example the grammar used in an English essay can be graded by a grader with grammar expertise and another can grade its cohesion.

Lastly there is a decent chance of extracting valuable information by finding a coherence between a peer's performance on different times. The likely coherence and identification of this should help better estimate a grader's reliability.

6 Conclusion

MOOCs are essential to take the education to the next era and Peer Grading has a crucial part in this. However Peer Grading methods still have a long way to go to reduce the error induced due to malicious grading or lack of grading performance.

We have experimented using synthetically generated data and distinguished Generalized PeerRank method proposed by [?] to be the method with the least error induced and we are planning to work on ways of improving this particular method.

We also propose an idea that has great potential to improve grading performance by assigning graders things that they are relatively more knowledged about.

References

[1] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs. *Biometrika*, 39:324–345, 1952.

[2] Luca de Alfaro and Michael Shavlovsky. Crowdgrader: Crowdsourcing the evaluation of homework assignments. *CoRR*, abs/1308.5273, 2013.

[3] R.. Ducan Luce. *Individual Choice Behavior a Theoretical Analysis*. John Wiley, 1959.

[4] C. L. Mallows. Non-null ranking models. i. *Biometrika*, 44(1-2):114–130, June 1957.

[5] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[6] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong B. Do, Andrew Y. Ng, and Daphne Koller. Tuned models of peer assessment in moocs. *CoRR*, abs/1307.2579, 2013.

[7] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):pp. 193–202, 1975.

[8] Karthik Raman and Thorsten Joachims. Methods for ordinal peer grading. *CoRR*, abs/1404.3656, 2014.

[9] Nihar B. Shah, Joseph Bradley, Abhay Parekh, Martin J. Wainwright, and Kannan Ramchandran. A case for ordinal peer-evaluation in moocs. *Neural Information Processing Systems (NIPS): Workshop on Data Driven Education*, 2013.

[10] Leon L Thurstone. The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, 21(4):384, 1927.

[11] Toby Walsh. The peerrank method for peer assessment. *CoRR*, abs/1405.7192, 2014.