Towards Adaptive Peer Assessment for MOOCs

Nicola Capuano
Dept. of Information Engineering, Electric Engineering and Applied Mathematics, University of Salerno
Fisciano (SA), Italy
ncapuano@unisa.it

Santi Caballé
Dept. of IT, Multimedia and Telecommunication
Open University of Catalonia
Barcelona, Spain
scaballe@uoc

Abstract—The increase in popularity of Massive Open Online Courses (MOOCs) requires the resolution of new issues related to the huge number of participants to such courses. Among the main challenges is the difficulty in students’ assessment, especially for complex assignments, such as essays or open-ended exercises, which is limited by the ability of teachers to evaluate and provide feedback at large scale. A feasible approach to tackle this problem is peer assessment, in which students also play the role of assessor for assignments submitted by others. Unfortunately, as students may have different expertise, peer assessment often does not deliver accurate results compared to human experts. In this paper, we describe and compare different methods aimed at mitigating this issue by adaptively combining peer grades on the basis of the detected expertise of the assessors. The possibility to improve these results through optimized techniques for assessors’ assignment is also discussed. Experimental results with synthetic data are presented and show better performances compared to standard aggregation operators (i.e. median or mean) as well as to similar existing approaches.

Keywords—Peer Assessment, MOOCs, e-Learning.

I. INTRODUCTION

A recent popular trend in open education is represented by Massive Open Online Courses (MOOCs) that, according to [1], is a continuation of the trend in innovation, experimentation and use of technology initiated by distance and on-line learning, to provide learning opportunities for large numbers of learners. Indeed MOOCs are intended for thousands of simultaneous participants, with some courses offered by Coursera1 and Udacity2 exceeding 100,000 registrants.

MOOCs have peculiar characteristics and challenges with respect to different educational resources, mostly due to their scale. In particular, in order to manage the delivery of courses to thousands of participants they require a complete re-adaptation of the traditional e-learning model based on tutor assistance. The use of automated tools, called to supplement the lack of human attention while maintaining acceptable quality standards, is also increasingly required.

In [2], a summary of the key challenges that MOOCs designers and providers are facing is reported. Among these challenges, assessment is one of the most prominent. Because of both the high numbers of students enrolled and the relatively small number of tutors, in fact, tutor involvement during delivery stages has to be limited to the most critical tasks. So it is not possible for human tutors to follow up with every student and review and grade assignments individually. According to [3], there is in MOOCs the impossibility of providing marks and feedback that is not either automated or peer assessed.

Unfortunately, both automated and peer-assessment grading strategies are problematic. On the one hand, automated grading is limited, disappointing and insufficient, with no partial marks and detailed explanations of answers. It may result particularly unsatisfactory when applied to complex tasks like the evaluation of the students’ ability of proving a mathematical statement, or expressing their critical thinking over an issue, or even demonstrating their creative writing skills [4].

On the other hand, peer assessment can also result unsatisfactory at some extent. In this approach, students are required to grade a small number of their peers’ assignments as part of their own assignment and the final grade of each student is obtained by averaging the received peer grades. Even if the literature reports on many learning benefits for the peer-assessor (e.g. exposure to different approaches, development of self-learning abilities, enhancement of reflective and critical thinking, etc.) [3], the use of grades provided by peers to evaluate a student may be seen as unprofessional and lacking the necessary expertise, both didactical and on the specific subject.

This paper aims at mitigating the issues connected to the application of peer assessment to MOOCs by considering alternative approaches to combine together peer grades coming from students, different from standard operators like median or mean. Described methods are capable of weighting grades provided by students according to assessors’ proficiency in the subject matter. In this way the opinion of high-skilled students has a greater impact on the final grade with respect to that of low-skilled ones.

Moreover, alternative methods with respect to random assignment are also discussed to associate each student to her assessors. Defined methods try to balance the number of reliable and unreliable assessors throughout the set of assignments. This way, situations where a student is evaluated only using grades proposed by unreliable assessors are avoided and the reliability is constant among all assessments.

The paper is organized as follows. The next section presents related work on peer assessment as well as some existing aggregation approaches proposed by recent literature. Section 3 presents the approaches we defined both for the aggregation of peer grades and for the assessors-assessee assignment. Defined

1 www.coursera.org
2 www.udacity.com
approaches are evaluated in Section 4 with synthetic data and obtained results are compared with results coming from other methods discussed in Section 3. Finally, Section 5 summarizes conclusions and outlines on-going work.

II. RELATED WORK

Even if some studies suggest a good correlation between peer assessment results and instructor ratings in conventional classrooms and online courses (at least for specific, high-structured domains) [5], there is still a general concern on the use of peer assessment in MOOCs as a reliable and accurate strategy to approximate instructor marking. Moreover, students themselves seem to distrust the results of peer assessment.

In order to address the issue of accuracy of peer assessment, several approaches, at various stages of development, have been proposed so far. For example, the Calibrated Peer Review (CPR) method [6] proposes a calibration step to be performed by students before starting to assess other students’ assignments. During the calibration, each student rates the same small set of assignments that have been already rated by the instructor. The discrepancy between grades provided by a student and the instructor measures her accuracy in assessment and is then used to weight subsequent assessments provided by that student. The more accurate is an assessor the more weight is given to her judgment on a peer assessment.

CPR has been experimented in several contexts demonstrating to be an effective instructional tool. Despite that, it requires additional work from those students who are asked to take part in the calibration step. Moreover, this method does not take into account the progresses that students make over time until a new calibration step is performed. For this reason, additional approaches have been defined able to automatically tune peer grades based on different parameters.

In [7], three probabilistic models for tuning grades obtained in peer assessment are presented. Such models estimate the reliability of each grader as well as her bias (i.e., a score reflecting the grader’s tendency to inflate or deflate her assessment) based on the analysis of grading performances on special “ground truth” submissions that are evaluated either by the instructor or by a big number of peers (hypothesising that the mean of many grades should tend toward the correct grade). Reliability and bias of each student are then used to tune the provided grades to other (non-ground-truth) submissions.

A similar approach has been applied in [8], where a Bayesian model is used to calculate the bias of each peer assessor in general, on each item of an assessment rubric and as a function of the assessor grade assigned by the instructor. As in the previous case, obtained biases are used to tune the grades provided during peer-assessment. Differently from the previous method, bias calculation is based on the results of a whole round of assessment rather than on just few “ground truth” submissions so, for the “calibration” round, all submissions are required to be rated also by the instructor.

The Vancouver algorithm [9], applied by CrowdGrader\(^1\), measures the grading accuracy of a student by comparing the grades given by her to each assignment with the average grade for that assignment. Differently from the other approaches, the assessor accuracy is used as a modifier of the assessor’s grade rather than of the assesse’s in order for the student’s grade to reflect not only the quality of her homework but also the quality of her work as a reviewer.

The PeerRank method [10] builds a grade for a given student by weighting the grades proposed by her assessors on the basis of the grades received by assessors themselves. In other words, the grade received by a student is considered as a measure of her ability to correctly rate other students. Given that students’ grades recursively depend on other student’s grades, an iterative algorithm based on an equation similar to that used in PageRank [11] is proposed for their calculation.

Differently from other methods, PeerRank does not require any instructor’s intervention. Indeed, there is no need to have a ground truth of professionally graded assignments. The same author has also proposed an improvement to the basic method that includes, as a component of the final grade of a student, the accuracy of proposed evaluations with respect to the average grades proposed by other peers. Such component is seen as an incentive for students to grade correctly.

Given the promising results shown in [10], we have implemented the PeerRank algorithm and have used it as the starting point for this study. In the next sections, we discuss the methods that we have defined for improving peer-assessment results as well as obtained experimental results.

III. THE DEFINED METHODS

In a typical peer assessment scenario an assignment is given to \(n\) different students. Each student elaborates her own solution (e.g. an essay, a set of answers to open-ended questions, etc.) generating a submission. Each student has then to grade \(m\) different submissions (with \(m < n\) coming from other students (maybe based on an assessment rubric).

The assignment of submissions to assessor students is performed in accordance to an assessment grid i.e. a Boolean \(n \times n\) matrix \(A\) where \(A_{ij} = 1\) iff student \(j\) has to grade the submission of the student \(i\). The matrix \(A\) has the following properties:

1. the sum of the elements in each row and column is equal to \(m\) (i.e. each student grades and is graded by \(m\) other students);
2. the sum of the elements in the main diagonal is equal to 0 (i.e. none evaluates himself).

The easiest way to build the assessment grid is by filling it at random with an algorithm that preserves the above properties. A feasible algorithm starts with a null matrix and initialises its elements according to the following equation:

\[ A_{\text{mod}(i+j-1,n)+1,j} = 1 \quad \forall \text{ } 1 \leq i \leq n; \ 1 \leq j \leq m. \] (1)

The obtained matrix is then randomly shuffled in several iterations by randomly selecting a couple of rows (or columns) \(i\) and \(j\) so that \(A_{ij} = A_{ji} = 0\) and swapping them.

---

\(^1\)www.crowdgrader.org
The grades proposed by students are then collected in the grades matrix $G$ where $G_{ij}$ is the grade proposed by the student $j$ for the student $i$ and $0 \leq G_{ij} \leq 10$. In standard peer-assessment settings the final grade $g_i$ of each student $i$ is obtained from the matrix $G$ by averaging all the grades obtained by peers (a matrix row) with the following equation:

$$g_i = \frac{1}{m} \sum_{j=1}^{n} G_{i,j} \quad \forall 1 \leq i \leq n. \quad (2)$$

Some authors propose to average all obtained grades apart the best and the worst [9]. Other authors use the median in place of the average. Despite that, the average is the most used aggregator and is the baseline against which we compare other aggregators proposed in next subsections.

A. PeerRank

In [10] the author proposed to weight the grade that an assessor student proposes for another student by her own grade i.e. to use the grade of a student as a measure of her ability to grade correctly. Given that the grades of all assessor students are themselves weighted averages of grades obtained by their own assessors, an iterative process, named PeerRank, was proposed to calculate the final grade of each student.

Let be $g_i^0$ the grade of the student $i$ at the $t$-th iteration, the grade of $i$ at the iteration $t + 1$ is defined as follows:

$$g_i^{t+1} = (1 - \alpha)g_i^t + \alpha \frac{\sum_{j=1}^{n} G_{i,j}g_j^t}{\sum_{f=1}^{n} g_f^t} \quad (3)$$

where $0 \leq \alpha \leq 1$ is a constant affecting the convergence speed and $g_i^0 = g_i$ (i.e. it is initialised according to equation 2). Both summations are performed over all students $j$ that have evaluated the student $i$ (indicated with $j \rightarrow i$) i.e. so that $A_{ij} = 1$.

Equation 3 takes into account that every student only evaluates $m$ peers according to the assessment grid. This is a more realistic setting with respect to the one described in [10] where each student was assumed to evaluate any other student. The author has also demonstrated some useful properties for the defined grade updating rule as well as that after a limited number of iterations, it converges to stable values.

Given that the proposed equation does not incentivize students to evaluate their peers accurately, the same author of [10] defined an updated version of Equation 3 as follows:

$$g_i^{t+1} = (1 - \alpha - \beta)g_i^t + \alpha \frac{\sum_{j=1}^{n} A_{i,j}g_j^t}{\sum_{f=1}^{n} g_f^t} + \beta \frac{\sum_{j=1}^{n} \left| g_{i,j} - g_i^t \right|}{m} \quad (4)$$

where $0 \leq \beta \leq 1$ is a constant, so that $\alpha + \beta \leq 1$, that weights the reward given to a student according to the inverse normalised absolute error in the grades given by her to other students.

If $\beta = 0$ then the equation 4 degenerates to the Equation 3. For $\beta > 0$, if $G_{ij} = g_j^t$ for all $j$, then the grades assigned by $i$ are all exact and the contribution of the third addendum is $10\beta$. At the opposite, if $G_{ij} - g_j^t = 10$ for all $j$ then the grades assigned by $i$ are completely wrong and the contribution of the third addendum is 0.

B. ExpPeerRank and BestPeer

The PeerRank rule, described by Equation 4, prescribes that the influence of the grade of an assessor student on any grade she proposes is linear. In order to improve the quality of the final grades and to minimize the contribution of low skilled students while maximising those of highly skilled ones, we propose an updated rule named ExpPeerRank that applies a super-linear modifier according to the following equation:

$$g_i^{t+1} = (1 - \alpha - \beta)g_i^t + \alpha \frac{\sum_{j=1}^{n} G_{i,j} \exp(g_j^t)}{\sum_{f=1}^{n} \exp(g_f^t)} + \beta \frac{\sum_{j=1}^{n} \left| g_{i,j} - g_i^t \right|}{m} \quad (5)$$

where $\exp(x)$ stands for the exponential function $e^x$ while the other parameters have the same meanings as in equation 4.

Bringing this reasoning to the extreme, we can imagine to assign the maximum influence only to the best grader for each student and no influence at all to any other proposed grade. This is the case of another rule we propose, named BestPeer. It calculates the grade $g_i$ for any student $i$ with one of the previous methods and then assigns to each student the final grade $g_i^t$ according to the following rule:

$$g_i^t = G_{i, \text{arg max}_j \exp(A_{i,j} g_j)} \quad (6)$$

The latter method, intuitively, is capable of performing particularly well when, for each student, at least one good grader is available. Unfortunately, this condition cannot be granted with the random assessor-assessee assignment proposed by equation 1. For this reason, in the next sub-section we discuss an alternative assignment method that, under certain conditions, can overcome the limitations of the random assignment.

C. Smart Assignment

The randomized assessor-assessee assignment can generate settings in which some students are assessed by only non reliable graders (i.e. students with a low grade). In this case, even weighting the grades, the overall peer-assessment performance may be poor. Balancing reliable graders among students is a feasible approach to overcome this issue but, unfortunately, we have no information about the grades when the assessment grid is built.

A first option to overcome this issue is to initialize the assessment grid based on grades coming from previous assessments. To do that, a feasible algorithm starts with a null matrix and initialises its elements according to the following equation:

$$A_{\text{mod}(m(i-1)+j-1), n} \cdot \text{rank}(i) = 1 \quad \forall 1 \leq j \leq n; 1 \leq i \leq m. \quad (7)$$

where rank($i$) is the position of the $i$-th student in the student ranking i.e. the list of the students ordered decreasingly on the average grade obtained in previous assessments.

Unfortunately, equation 6 does not ensure the fulfilment of the second property of assessment grids. For this reason an additional check is needed and if $A_{ij} = 1$ for some $1 \leq i \leq n$ then the most close column $j$ so that $A_{ij} = 0$ and:

$$\exists z \mid A_{z,i} = 0 \text{ and } A_{z,j} = 1 \quad (8)$$
is selected and the values of $A_{ij}$ and $A_{ij}$ are swapped as well as values of $A_{ij}$ and $A_{ij}$. In other words, the assessor $i$ does not assess himself anymore but the student $z$ assigned to the closest performer $j$ that, in return, takes care of evaluating $i$.

A second option for optimizing the assessor-asseeess assignment is to proceed incrementally (i.e., to perform the assessment session in $m$ rounds). At the first round, each student is randomly assigned just one student to grade. At each subsequent round students are ranked in two lists:

1. list 1 orders students, decreasingly, on the average grade obtained in the preceding rounds (it so ranks the students basing on their quality as graders);
2. list 2 orders students, increasingly, based on the average grades obtained by their graders in the preceding rounds (it so ranks the students based on the quality of obtained grades).

Then, for the subsequent round, each student from list 1 has to grade the student from the list 2 with the same rank. This ensures that, in each step, the best graders are assigned to the students that, in the previous steps, have obtained grades from the worst ones. Some additional checks must be made to ensure that no student evaluates herself and that no student evaluates another student more than once.

This method has the advantage that it does not need any information about past assessments. Conversely, its incremental nature requires that every grade is assigned for a given round before starting the next one. This constraint can be very expensive, especially in massive contexts, when some student may be late in providing grades ore may not provide grades at all. For these reasons we decided, in this work, to focus our attention just on the first method, postponing the exploration of the second one for future works.

IV. EXPERIMENTAL EVALUATION

In order to evaluate the performances of the defined methods, we made five experiments with synthetic data. In all experiments 100 students are supposed to have submitted a solution to an assignment composed of 10 questions. For a correct answer a student gains 1 point while for a wrong answer she gains 0 points. The real grade of each student is then an integer belonging to $[0, 10]$.

Each student has then to evaluate the submissions of 4 other peers. In our model, we suppose that each student $i$ with a real grade $g_i$ has probability $g_i/10$ of marking correctly each answer of a peer submission. So if the student $i$ grades the submission of a student $j$ (with real grade $g_j$), then the proposed grade $G_{ij}$ is a random variable so that:

$$G_{ij} \sim B(g_j, g_i/10) + B(1 - g_j, 1 - g_i/10)$$  \hspace{1cm} (9)

where $B(m, p)$ represents a binomial distribution of $m$ trials with probability $p$.

Each experiment is made of several iterations. For each iteration, real grades are randomly assigned (with different probability distributions). Then, the assessment grid is built (according to different methods) and the grades matrix is randomly filled according to the probability distribution given in Equation 9. The final grades are then calculated (according to different methods) and compared to real grades by calculating the Root Mean Square Error (RMSE). The details and the results of each experiment are discussed in the next sub-sections.

A. Binomial distribution of real grades

In this experiment, the real grades are assigned according to a binomial distribution: each student, for each of the 10 questions of her assignment, has a probability $p$ to answer correctly and $1 - p$ to answer wrongly. In other words, the real grade of a student $i$ is assigned according to:

$$g_i \sim B(10, p).$$  \hspace{1cm} (10)

In each step of the experiment a probability $p$ is chosen and 1000 iterations are performed. For each iteration, the real grades are assigned as described above (with probability $p$). Then a 100x100 assessment grid is randomly generated according to Equation 1 and a grades matrix, including all proposed grades, is randomly generated basing on the distribution given in Equation 9.

For each iteration, the final grade of each student is calculated as the Average of grades proposed by peers (Equation 2), with the PeerRank rule (Equation 4), with the ExpPeerRank rule (Equation 5) and with the BestPeer method (Equation 6). For each iteration, the RMSE between final and real grades is calculated over the 100 students. The obtained values are then mediated over all iterations.

Figure 1 plots the performances obtained by applying the four methods to the defined marking model in terms of mean RMSE against the probability $p$ used to generate the real grades. As it can be seen both PeerRank and ExpPeerRank outperform the Average method for $p > 0.6$. Conversely, the performances of all methods are quite similar for $0.5 \leq p \leq 0.6$ while, for $p < 0.5$ the best method remains the Average.

Obtained results show that both PeerRank and ExpPeerRank need $p > 0.5$ to get any useful signal out of the data. It is worth noting that $p = 0.5$ means that students are answering (or
Best performances are obtained when the first experiment with a RMSE lower or equal to most all conditions. Only for \(B\) forms a little better than \(D\).

Fig. 2. Performance in terms of RMSE of the described methods on a uniform distribution of grades with different values for \(m\) (minimum grade for an assignment).

marking) questions just as well by tossing a coin. This means that, in real contexts, assuming that \(p > 0.5\) is not so restrictive constraint. Moreover, as it can be seen, \(ExpPeerRank\) performs a little better than \(PeerRank\) while the \(BestPeer\) method is better than other methods only for \(p > 0.9\).

The best choice for this distribution of grades seems to be \(ExpPeerRank\) that ensures, in best cases, a decrease in RMSE of about 1 grade with respect to the baseline \(Average\) method. This means that, on average, each student will have a final grade closer to the real one of approximately 1 point over 10.

B. Uniform distribution of real grades

In this experiment, the real grades are assigned according to a uniform distribution rather than a binomial one: each student receives an integer random grade \(m\) to \(10\). Hence the real grade of a student \(i\) is assigned according to:

\[ g_i \sim U(\{m, \ldots, 10\}) \]  

(11)

where \(U(S)\) defines a discrete uniform distribution of the set \(S\).

Figure 2 plots the performances obtained by applying the four methods to the defined marking model in terms of mean RMSE against the minimum grade \(m\). Also in this case \(ExpPeerRank\) outperform both \(Average\) and \(PeerRank\) in almost all conditions. Only for \(m = 0\) the performances of all methods are quite the same.

It is interesting to note that \(BestPeer\) behaves better than in the first experiment with a RMSE lower or equal to \(PeerRank\). Best performances are obtained when \(m \leq 5\) (high variance of real grades) and with \(m \geq 8\) (high average real grade).

This can be explained by the fact that, when there is a high variance in student levels, there is a high probability that a peer is evaluated also by non reliable graders that affect the quality of the final grade in all methods (at different levels) apart from \(BestPeer\) where only the best grade is selected. This advantage disappears when \(m\) increases because in that case, proposed grades increase their average quality.

C. Smart assignment and binomial grades distribution

This experiment replicates the Experiment 1 with the difference that the assessment grid is generated according to equations 7 and 8 rather than to Equation 1. In the model, we assume that the average grade obtained in previous assessments (needed to generate the student ranking) is equal to the assigned real grade. This is a simplification that supposes that students maintain constant performances across several assignments. Given this simplification, the results of this experiment can be considered as an upper bound of the results obtainable with smart assignment in real contexts.

Figure 3 plots the performances obtained by applying the four methods to the defined marking model with random (dashed lines) and smart (plain lines) assignment methods. As it can be seen, with a binomial distribution of real grades \(Average\), \(PeerRank\) and \(ExpPeerRank\) are quite insensitive to smart assignment. Instead, as it might be supposed, \(BestPeer\) have a substantial improvement because the smart assignment ensures that each student is assessed by at least one good grader whose proposed grade is selected as the final one.

D. Smart assignment and uniform grades distribution

This experiment replicates the Experiment 2 with the difference that the assessment grid is generated according to Equations 7 and 8 rather than to Equation 1, with the same assumptions made in Experiment 3 with respect to the average grade obtained in previous assessments.

Figure 4 plots the performances obtained by applying the four methods to the defined marking model with random (dashed lines) and smart (plain lines) assignment methods in case of uniform distribution of real grades. In this case, while \(Average\) and \(PeerRank\) result again quite insensitive to smart assignment, \(ExpPeerRank\) and (at a greater extent) \(BestPeer\), show a good improvement.

In particular, \(BestPeer\) outperforms all the other methods, especially in configurations with high grades variance (\(m < 5\)) and high average real grade (\(m > 6\)). Only for \(m \leq 1\) its performances are equal or worse than those of other methods. Hence in this case, the best choice seems to be \(BestPeer\), whose per-
Performances in contexts with a high variance of student levels are boosted by the smart assignment.

E. BestPeer and support methods

As described in Section III.B the BestPeer method calculates the final grade for any student with one of the other methods, then assigns to each student the grade coming from the assessor with the best final grade. In the previous experiments we used ExpPeerRank as support method for BestPeer. In this last experiment we wonder if ExpPeerRank is the best possible choice, at least in the configuration of Experiment 4.

We have so repeated Experiment 4 only with BestPeer adopting different support methods. Obtained results are plot in Figure 5 against the standard Average method. As it might be supposed, ExpPeerRank (the method with the best performance in the majority of configurations) seems to be the best choice.

V. FINAL REMARKS

In this paper, we have proposed different aggregation methods to be used in peer-assessment as well as a smart assessor- assessee assignment method in MOOCs aimed at balancing good graders among students. Experimental results with simulated data show that the ExpPeerRank method outperforms other methods in most configurations (with random or smart assignment) while, in particular circumstances (uniform distribution of grades) the BestPeer method with smart assignment performs better.

The assumption of this work, confirmed by other studies, is that the grade obtained by a student is not only a measure of proficiency in a given subject but also a measure of her ability to grade correctly. A limitation of this approach is that other factors that may affect the student’s ability to grade (e.g. the grader’s tendency to inflate or deflate proposed grades, the general attitude to review other’s work) are disregarded. Such parameters will be considered in a future work that will also arrange an experiment with real MOOC users.

ACKNOWLEDGMENTS

This research was partially supported by the Spanish Government through the project: TIN2013-45303-P “ICT-FLAG: Enhancing ICT education through Formative assessment, Learning Analytics and Gamification”.

REFERENCES