Some Scaling Laws for MOOC Assessments

Nihar B. Shah

Joseph Bradley

Sivaraman Balakrishnan

UC Berkeley UC Berkeley UC Berkeley

nihar@eecs.berkeley.edu joseph.kurata.bradley@gmail.com sbalakri@berkeley.edu

Abhay Parekh Kannan Ramchandran Martin J. Wainwright

UC Berkeley UC Berkeley UC Berkeley

parekh@berkeley.edu kannanr@eecs.berkeley.edu wainwrig@stat.berkeley.edu

ABSTRACT

A problem that arises with the increasing numbers of students in Massive Open Online Courses (MOOCs) is that of student evaluation. The large number of students makes it infeasible for instructors or teaching assistants to grade all assignments, while current auto-grading technology is not feasible for many topics. As a result, there has recently been an increase in the use of **peer-grading**, where students grade each other; in this way the number of graders automatically scales with the number of students. However, in practice**, peer-grading** has been observed to have high error rates and has come under criticism. In this paper, we take a statistical approach to assessing the feasibility of **peer-grading** for MOOCs. Our assessments lead to some good news and some bad news. Our study reveals that based on simple yet general models**, peer-grading** at scale is unfortunately unlikely to work well as a standalone option, i.e., s the number of students increases, the expected number of students mis-graded will grow proportionately. The good news however is that by considering a hybrid approach that combines **peer-grading with auto-grading** using dimensionality-reduction techniques, we can tackle the scaling challenge efficiently. Concretely, an automated approach is used for ‘dimensionality reduction’, a classical technique in statistics and machine learning, and **peer-grading** is used to evaluate this lower dimensional set of submissions.

Our contributions to the current MOOCs literature is twofold. First, we provide a principled analytical approach to a problem area that is predominantly empirically driven. In this context, we approach the problem at a high level to provide fundamental laws on scaling of assessments in MOOCs based on simple but general models. Secondly, we provide a constructive approach to tackling the scaling problem using tools from statistics. Our proposed hybrid approach can potentially inform the way next-generation MOOC-assessment algorithms evolve.

INTRODUCTION

We discuss the scalability of grading, with special attention given to **peer-grading** in Massive Open Online Courses (MOOCs). For budget-constrained OOCs, grading by the instructor(s) becomes infeasible as course sizes grow. Large courses require alternative approaches, such as auto-grading or **peer-grading**. We focus our discussion on MOOCs for which auto-grading using pre-trained models is difficult. For subjective topics and complex problems, it is often difficult to design machine grading systems which are accurate [2]. In these cases, it is important to include humans in the grading loop. **Peer-grading** is a system of grading where students taking a course re graded by fellow students in the same course. **Peer-grading** is a natural choice for MOOCs since the total number of graders in a **peer-graded** system self-scales in proportion to the number of students unrolled. For instance, Coursera employs **peer-grading** in its human-computer interaction (HCI) course. Since the students are not expert graders, in this **peer grading** system, the submission provided by each student is graded by 3 to 5 students. The final grade of a student is computed as the median of these individual grades [11]. Alternative algorithms for aggregating **peer-grades** are proposed in [19, 9, 23, 7, 16, 12].

Research has shown (e.g., [11, 12]) that current **auto-grading and peer-grading** systems make many mistakes. Qualitative observations about the inaccuracy of MOOC assessment have led to criticism of **auto-grading and peer-grading** [2, 20]. For MOOC course credits to gain increased acceptance, it is critical that these errors be reduced.

In this paper, we view the problem of assessments in MOOCs through the lens of statistical analysis. Our approach is orthogonal to and nicely complements the largely empirical nature of works in this field. We study the scaling behavior of **peer-grading,** i.e., the behavior when the number of students gets large. Our analysis reveals that under reasonable assumptions, these systems will *incorrectly grade a constant fraction of the students in expectation.* This constant fraction is not a problem for small courses, where an instructor can handle complaints from students who feel they were mis-graded. However, this is not a scalable solution for large courses since student complaints will overwhelm instructors. Our analysis gives some insight into why current **peer-grading** systems mis-grade any students.

Importantly, our proposed framework is flexible: it allows for diverse grade collection systems, applies to any arbitrary choice of (adaptive) assignments of graders to submissions and any arbitrary choice of aggregating the **peer-grades**.

Current efforts to improve grading have examined many aspects: improving auto-grading models [14, 24, 10], improving algorithms for aggregating **peer-grades** 19, 9, 23, 7, 16, 12], combining **auto-and-peer-grading** [12, 3], and dimensionality reduction [4, 5, 18, 21, 8, 15]. These efforts have shown useful improvements in practice. However, we argue that these methods do not change the scaling behavior of grading; at most, they decrease the constant fraction of mis-graded students.

A second contribution of this paper is to show that, on the upside, this scaling behavior may be improved via a combination of current methods used in MOOC assessments. Specifically, we present an algorithm for assessment that employs a combination of **auto-and-peer-grading** with ‘dimensionality reduction’. We prove that under very general error models, our algorithm has the potential to create vanishing error rates, i.e., to ensure that the expected fraction of students mis-graded goes to zero as OOC course sizes grow. This result provides a statistically sound, constructive path towards solving the scaling problem for assessments in MOOCs.

Any study involving MOOCs by definition must address scale. t is thus paramount when studying questions like grading for MOOCs that we tackle the most important high order bit” first. This motivates us therefore to address **peer-grading** for MOOCs with a view to first understanding the fundamental scaling behavior, namely to study scaling laws with respect to simple yet informative models. This will help inform more detailed subsequent studies involving specific algorithms and practical solutions needed to make this a reality.

SCALING ANALYSIS OF **PEER-GRADING**

We are interested in MOOCs which are massive and open. In other words, the courses are very large, and re offered for free or at low costs. This cost constraint prevents the number of instructors from scaling proportionally to the number of students. Thus, we may assume that instructors cannot hand-grade every student’s submission.1 Instead, grading is done via **peer-grading**, where students evaluate their peers’ submissions, and/or **auto-grading**, where a computer software evaluates the students’ submissions. Since grading requires considerable effort, we expect the average student to grade at most a few peers. e assume that most student graders perform better than random guessing but are imperfect, i.e., frequently make errors in the grading; empirical studies have shown this to be true (e.g., [11]). For our analysis, we assume that there is a “true” grade for each students’ submission. The grades provided by expert graders can be assumed to match the true grades, but students (i.e., **peer-graders**) provide only noisy measurements of the true grade.

Theorem 1 below analyses the scaling behavior of typical **peer-grading** systems described above. For very general settings, we show that in expectation, the grades of a constant fraction of students will be in error. In order to keep our claims independent of the model and inference algorithm, we consider a generic grading scenario and show that under this scenario, a constant fraction of students having a specific true grade are *statistically indistinguishable* from those for students having a different true grade.

1The term ‘submission’ will be used generically to refer to solutions to examinations, homeworks, or any ther material submitted for evaluation by the students.

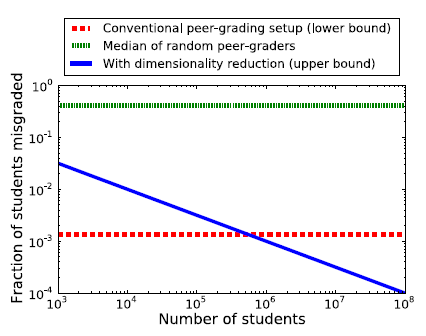


Figure 1. An illustration of the scaling of the expected fraction of students mis-graded in three settings: (a) lower bounds on error under conventional **peer-grading** setups, (b) error in the **peer-grading** algorithm that selects **peer-graders** uniformly at random and sets the final grade as the median of the **peer-grades**, and (c) an upper bound on the error under **peer-grading** with dimensionality reduction. The plot considers a setting with *q =0:25, k =3*, and no perfect graders or experts. In (c), the clustering algorithm is assumed to output clusters and wrongly clusters non-representative submissions. Observe how the fraction of students mis-graded emains constant in the conventional setups (a) and (b), but falls to zero with dimensionality reduction (c).

THEOREM .

*Consider the following setting. There are d students enrolled in the course. Each student submits one submission, and these d submissions must be graded. Each student has a true underlying grade which is equally likely to be either ‘pass’ or ‘fail’, and the goal is to infer these true grades. Each* ***(peer-)grader*** *grades k submissions. Each student is an imperfect grader with probability γ (0, 1], and is a perfect rader ith probability (1- γ) . When asked to grade a submission, an imperfect grader grades incorrectly with probability q, for some q (0, 0:5), while a perfect grader always grades correctly. A correlation between rading and answering skills may also be assumed: a randomly selected student with a true grade of ‘pass’ is a perfect grader with a probability p [0, 1). A constant fraction α [0, 1) of submissions may also be raded by the instructors who grade perfectly. The parameters k, γ, q, p and α are all independent of the total number of enrolled students d. The values of these parameters may be known to the* ***peer-grading*** *algorithm.*

*A* ***peer-grading*** *algorithm must first assign (****peer or instructor) graders*** *to each submission. We allow this assignment to be executed in a completely adaptive manner, i.e., the algorithm is allowed to wait for the output of one grader before making the next submission-grader assignment. There is no constraint on the number of people who may grade any individual submission. Upon receiving all the* ***(peer and instructor)*** *grades, the algorithm must then assign final grades to each submission.*

*Under any* ***peer-grading algorithm****, in expectation, the final grades of a constant fraction of students will be in error.*

The scaling results derived here imply that if the quality of **peer-graders** does not improve such ith an increase in the number of enrolled students, then as the number of students in the course increases, the number of mis-graded students will increase proportionately. The student experience will suffer in such a scenario, and instructors may be faced with an overwhelming number of student complaints. This scaling analysis is illustrated in Figure 1 along with a comparison to the dimensionality-reduction based **peer-grading** algorithms to be presented in the next section.

The remainder of this section discusses the assumptions and some extensions of Theorem 1. The assumption of having the parameters *γ, p* and q independent of the total number of enrolled students d means that the average abilities of the students do not vary (significantly) when the number of students enrolled in the course increases. We assume that the number of submissions *α*d graded by the instructors increases at most linearly with the total students d: a reasonable assumption for freely offered courses. Finally, we assume k to be a constant (and not increasing in d) since grading is a task that requires significant time and effort n the part of the grader, and we must impose a limit on the number of submissions that any student needs to **peer-grade.** Note that this restriction does *not* restrict each submission to be graded by at most k **peer-graders**.

The result of Theorem 1 may be extended to cover many alternative settings. A similar argument holds for the case when the grades collected from the graders are on a finer scale (with more than two possible values of the grades), when the true underlying grade may not be uniformly distributed, or when the **peer-grading** is ordinal where students are asked to compare two or more submissions instead of assigning numeric scores to the submissions [22]. **Peer-grading** may be augmented to also include grading y people not taking the course, e.g., by people who have previously taken the course. If the number of such ‘outside-graders’ remains linear in the number of students, then arguments similar to those of theorem 1 continue to apply. In order to ensure fairness, the use of extraneous information about the students’ skills is generally avoided in the grading process [19], as assumed in the theorem. One could potentially obtain some information about the students’ grading abilities from their performance in previous homeworks or tests, but this information will not change the scaling laws as long as the number of homeworks/tests conducted remains independent of the number of students.

BREAKING THE BARRIER VIA DIMENSIONALITY REDUCTION

In this section, we provide one possible means of breaking the barrier of a constant fraction of submissions being graded erroneously: dimensionality reduction. We discuss two methods for dimensionality reduction: clustering and featurization. Clustering is the more intuitive method, but featurization is more general. These methods combine auto-grading with **peer-grading** in a manner to be discussed in the sequel. The **peer-grading** interface remains the same as before, where students grade submissions submitted by their peers.

Clustering

For simplicity of exposition, the following discussion will assume a submission to be the answer to a single question. The algorithm may be applied separately to ach of the questions. Suppose we use a computer program to cluster the collection of all d submissions provided by the d students with respect to the similarity of their content. The clustering algorithm is such that ideally, within each cluster, all submissions have the same true grade. Note that multiple clusters could have the same true grade. Given this assumption, grading any one submission in the cluster effectively grades all submissions in the cluster. The total number of submissions to grade is effectively reduced.

A cluster can be graded by collecting **peer-grades** for any submission within the cluster. These grades may be aggregated via any reasonable algorithm, for instance, taking a median of the received grades. Finally, the submission of any student is assigned the grade that its cluster receives. This approach is formalized in Algorithm 1, and illustrated pictorially in Figure 2.

Algorithm 1 A grading algorithm with rigorous guarantees: combining dimensionality reduction and **peer-grading**

• Cluster submissions into Δ clusters based on ‘similarity’

• Select one representative submission for each cluster

• Choose **peer-graders** for each cluster uniformly at random, while ensuring that no peer-grader is chosen for more than k clusters

• For each cluster, assign the representative submission to its corresponding chosen **peer-graders**

• Upon receipt of **peer-grades**, for each cluster, assign median of the **peer-grades** as the final grade of all submissions in that cluster.

The following theorem shows that given a good clustering algorithm, the grading algorithm proposed in algorithm 1 can reduce the number of erroneously graded students to a vanishing fraction, i.e., ensure that the expected fraction of students mis-graded goes to zero. The theorem considers **peer-grading** conditions worse than that considered in Theorem 1, namely, does not assume the existence of any perfect graders or instructor graders. In what follows, D(p1 || p2) denotes the Kullback-Liebler divergence [13] between two Bernoulli distributions having parameters p1 and p2 respectively, i.e.,

D(p1 || p2)= p1 log +(1 - p1) log

We use the standard Bachmann-Landau [1] notations of o(∙),O(∙) and Θ(∙) to represent the limiting behavior of functions. Informally, f(x)= o(g(x)) means that f is dominated by g asymptotically, f(x)= O(g(x)) means that f is upper bounded by g asymptotically, and f(x) = Θ (g(x)) means that f is both upper and lower bounded by g(x) with different constant pre-factors.

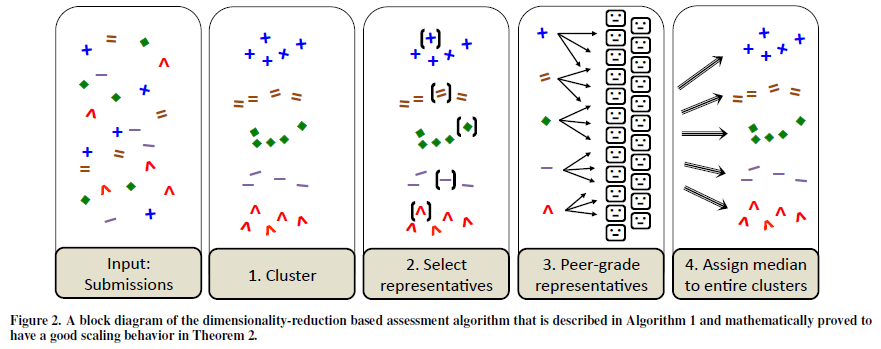


Figure 2. A block diagram of the dimensionality-reduction based assessment algorithm that is described in Algorithm 1 and mathematically proved to have a good scaling behavior in Theorem 2.

THEOREM 2.

*Consider the following setting. There are d students enrolled in the course. Each student submits one submission, and these d submissions must be graded. Each student has a true underlying grade which is either ‘pass’ or ‘fail’, and the goal is to infer these true grades. Each (****peer-)grader*** *grades k submissions. When asked to grade a submission, a grader grades incorrectly with probability q, for some q (0, 0:5). Here, k and q are independent of the number of students d.*

*Suppose grading is performed as per Algorithm 1. Then, if the number of clusters satisfies* Δ ≤ *kD(0.5 || 1- q), and if the clustering algorithm employed in algorithm 1 erroneously clusters at most o(d) non-representative submissions and at most o(*Δ*) representative submissions at random, then the fraction of submissions that are mis-graded goes to zero as the total number of submissions d grows.*

The theorem says that if the submissions can be clustered into clusters of average size of order *log d* or higher, then even if each student can grade only a constant number of submissions, the number of submissions graded incorrectly will be independent of the number of students d. The system can now scale to accommodate arbitrarily large numbers of students without the worry of the number of grading errors blowing up. This scaling analysis is illustrated in Figure 1. The theorem assumes a good performance of the clustering algorithm. This may seem to be at odds with our earlier discussion on the performance of auto-grading algorithms suggesting that they do not perform very well for topics that are subjective in nature. To this end, we note that while auto-grading requires the algorithm to understand the semantics of each submission in order to grade, c*lustering only requires understanding the features on which similarity testing must be performed.* The job of clustering is thus a subset of the harder task of auto-grading. Indeed, designing algorithms for clustering submissions for educational assessment is an area of active research in the community, e.g., [4, 5, 18, 21, 8, 15]. Moreover, research has shown that students tend to answer in similar ways [15, 4] allowing the number of clusters to be small.

Let us now discuss the assumption on the performance of the clustering algorithm from a statistical perspective. We have assumed that the number of submissions clustered erroneously rows sub-linearly in d. In other words, we assume that the fraction of submissions clustered incorrectly educes with an increase in the total number of submissions. An intuitive justification of this assumption s that clustering is often performed by comparing submissions with each other, and as the total number f submissions d grows, the number of submissions (in each cluster) available for comparison also grows, hereby allowing the clustering algorithm to improve its performance as d increases. Using literature on statistical guarantees for clustering problems (e.g., [6]), we make this intuition mathematically rigorous.

*THEOREM 3. (Adaption of [6, Theorem 2.2]) Consider a clustering algorithm that operates in the following manner. The algorithm has a black-box comparator that, given any pair of submissions, correctly identifies whether they belong to the same cluster or not with a probability at least + ε (independent f all other pairs), for some ε >0. The value of ε is fixed, independent of d, and is unknown to the algorithm. Suppose there are Δ ≤ cε2 clusters, of equal size, and log d the value of Δ is known to the algorithm. Here, c is a specific universal constant. Then there exists a clustering algorithm such that the expected number f submissions clustered incorrectly is upper bounded by a universal constant (independent of d).*

REMARK 1.

*The setting of [6, Theorem 2.2] differs from the setting of* ***peer-assessment*** *considered in this section in some respects. We assume the average cluster-size to be at least  log d (for some constant  > 0) and that the clustering algorithm does not know the sizes of these clusters, while [6, Theorem .2] assumes that the sizes of all clusters are identical, lower bounded by  log d, and the size of each cluster is known to the algorithm. On the other hand, the results of [6, Theorem 2.2] provide a very strong guarantee in that only O(1) submissions are in wrongly clustered in expectation, whereas for our requirements, a weaker guarantee of o(d) errors would suffice.*

The scaling analysis presented in this section suggests that reducing the dimension of the submissions by a logarithmic factor may suffice for designing a scalable grading system. Previous works on clustering which we reference used clustering to id instructors in grading. This approach can lessen the burden on instructors, who can assign grades to groups of submissions. Our analysis motivates as well as provides a theoretical justification for the use of clustering tools [4, 5, 18, 21, 8, 15] for massive open online courses, where in conjunction with **peer-grading**, we show that they can help achieve scalability in the grading process.

Featurization

We briefly discuss a more general type of dimensionality reduction in this section. The clustering method discussed earlier assumes that many submissions are similar enough to be declared equivalent by a clustering algorithm. One could generalize to assume that *parts* or *aspects* of many submissions are similar and can e compared or clustered; we describe this as featurization, where the content of a submission is summarized by a set of features.

Suppose that submissions may be described by Δ features. Assume that he grade of each submission may be computed as a function of these features; e.g., a simple such unction for pass/fail grades would be thresholding a weighted sum of the features. Then we have reduced the problem of grading to a traditional regression setting: submissions are examples, features are computed algorithmically for each example, and **peer-grades** provide noisy labels for the examples.

The regression model is simply another aggregation method for **peer-grades**. Assume for now that we use a generalized linear model y (i;t) ~ f(w T x (i) )+ *ε* (i;t), where x(i) is the feature vector for submission i, y(i;t) is he tth **peer-grade** for submission i, f(∙) is the inverse link function, and *ε* (i;t) is the noise added by the **peer-grader**. This model generalizes simple **peer-grading** systems which treat all submissions independently: d boolean features are indicators corresponding to the d submissions, so the feature vector x(i) for submission i has a “1” for feature i and “0” elsewhere. The model also generalizes clustering: we have Δ boolean features corresponding to the Δ clusters.

This setup could allow grading systems to raw n extensive research on feature engineering and modeling. The success of previous work on feature-based clustering of submissions in MOOCs indicates that useful features can be found. The fact that current aggregation methods for **peer-grades** can be generalized by simple regression models indicates that such models are reasonable. Depending on the choice of the regression model, one can achieve scaling results similar to those for clustering. E.g., consider a logistic regression model (with a logit link function) to classify submissions as pass/fail. Then previous work [17] on logistic regression has shown hat, with Δ features and Θ(d) samples (**peer-grades**), it suffices to have Δ = o(d).

Previous work has considered combining auto-grading with **peer-grading** [12, 16]. Both these works employ an adaptive process for grading, with a separate auto-grading processes used to supplement the **peer-grading** process. However, none of these works have rigorous guarantees on the performance. Moreover, these works **separate peer-grading and auto-grading** into multiple stages, and combine the **auto-grades and peer-grades** in a certain manner to compute the final grade. This approach leaves the scaling laws unchanged: either auto-grading makes a constant fraction of grading errors (in which case a constant fraction of errors remain after **peer-grading as well), or auto-grading** makes a vanishing fraction of errors (in which case **peer-grading** becomes unnecessary for the purpose of obtaining vanishing error rates).

CONCLUSIONS

In this paper, we gave a statistical analysis of the scaling properties of various grading mechanisms in OOCs. We saw that under very simple and general models, the kinds of **peer-grading** systems employed today will not scale. We then showed that combining (auto-) dimensionality reduction and **peer-grading** as the potential to scale. Dimensionality reduction is already an active topic of research [4, 5, 18, 21, 8, 15], and the proposal of combining it with **peer-grading** falls under the more general paradigm of combining machine and human intelligence [12, 3].

While most current research on assessment in MOOCs is empirical, this paper provided a more analytical approach, helping understand the basic principles behind **peer-assessment** in current grading systems, and identifying a path for future research to overcome assessment errors. Accurate, reliable, and scalable assessment will help to pave the way for MOOCs to democratize education.

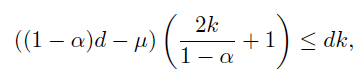
APPENDIX: PROOFS

This section presents proofs for the three theorems stated in the paper. The proofs focus on demonstrating the scaling laws and do not attempt to optimize the constants (The plots of Figure 1 have optimized constants). Since we are interested in only the scaling laws, without loss of generality, we assume d to be large enough to ignore any floor/ceiling effects of integers.

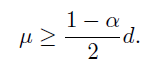
PROOF OF THEOREM 1.

Consider any values of the parameters d, k, q, α, γ, p as defined in the statement of the theorem. Consider any **peer-grading** algorithm Α that minimizes the expected fraction of students mis-graded. The high-level idea of the proof is as follows. We will select two students uniformly at random from the set of d students. We will show that with a probability lower bounded by a positive value independent of d, the performance of the two students appear statistically indistinguishable to algorithm A, irrespective of their true underlying grades. Conditioned on being statistically indistinguishable, algorithm A will mis-grade at least one of these two students in expectation. Aggregating over all students, we will show that he fraction of students mis-graded is lower bounded by a positive value that is independent of d. Given the formulation of Theorem 1 and this high-level idea, the rest of the math behind the proof is fairly simple.

Under the setting considered in the theorem, at least (1-α)d submissions will not be graded by instructors. These submissions must be evaluated based on the **peer-grades** assigned to them and the accuracy of these (1-α)d students in the **peer-grading** that they performed. The total number of **peer-grades** available to algorithm A is upper bounded by dk. Let µ denote the number of submissions from this collection of (1-α)d submissions that have received no more than **peer-grades.** The pigeonhole principle implies that

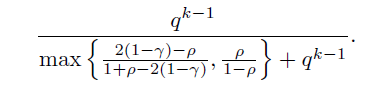


and some algebraic manipulations lead to

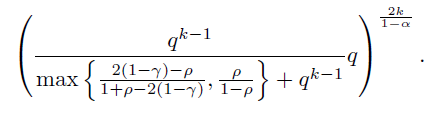


Let us momentarily focus our attention only on these µ submissions. Suppose you choose two submissions uniformly at random from this set of µ submissions, say the submissions of students s1 and s2. We derive a lower bound on the probability that the performance of student s2 is statistically indistinguishable from that of student s1, irrespective of the true grades of students s1 and s2. Consider he **peer-grading** process of student s2 under algorithm A. In order to obtain a **peer-grade** for s2’s submission, Algorithm A may wish to enroll a perfect or an imperfect grader. The algorithm has the following information available at its disposal for inferring whether a certain **peer-grader** is perfect or imperfect: the performance of this **peer-grader** on the (k - 1) or fewer submissions hat she has already graded, and the evaluation of this **peer-grader’s** own submission coupled with the knowledge of the correlation between good graders and good grades.

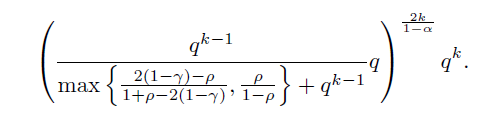
Recall that in the setting under consideration, a randomly selected student is an imperfect grader with probability γ and a perfect grader with probability (1-γ). A randomly selected student has a ‘pass’ or a ‘fail’ grade on her own submission with a probability 0.5 each. Given a pass grade, the student is a perfect grader with a probability p. In order for the algorithm to confuse the imperfect grader for a perfect grader, the imperfect grader should have correctly graded all (at most (k - 1)) submissions assigned to her, and should have a true underlying grade that is most common among perfect graders. Some algebraic manipulations lead to a lower bound on the probability f the algorithm mistaking an imperfect grader for a perfect grader as



When an imperfect grader happens to be selected, there is a q ∊(0, 0.5) chance that she will mis-grade and a (1 - q) chance that she will grade correctly. It follows that for any arbitrary sequence of or fewer gradings for student s2, the probability of observing this sequence is lower bounded by

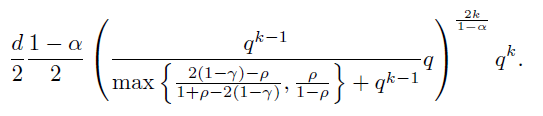


The grades of the two students s1 and s2 may also depend on their performance as **peer-graders**. The probability that student s2 **peer-grades** in a manner identical to s1 is lower bounded by qk. It follows that he probability that these two randomly chosen students appear statistically indistinguishable to algorithm A is lower bounded by



Conditioned on students s1 and s2 appearing statistically indistinguishable to algorithm A irrespective of their true grades, the algorithm will mis-grade at least one of them in expectation.

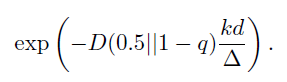
We now return to the set of all d students. Aggregating all arguments from above, we get that the algorithm will mis-grade at least



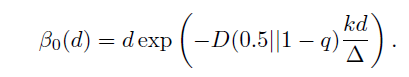
students in expectation. This quantity is linear in d thereby proving our claim.

PROOF OF THEOREM .

First consider the setting of a perfect clustering algorithm. The total number of grades obtained from the students is kd. Let β0(d) denote the expected number of students mis-graded when the total number of enrolled students is d. We show that in the setting described in the statement of the theorem, β0(d)=O(1). Since there are Δ clusters, each cluster receives grades. Each grade is correct with a probability (1 - q) ∊(0.5, 1], and hence, applying the Chernoff bound, we get that the probability of mis-grading any individual cluster is upper bounded by



Thus, in expectation, the total number of mis-graded students is

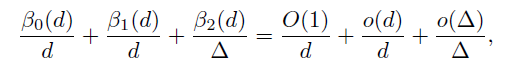


Substituting 4-11.pngas assumed in the statement of the theorem, we get that only O(1) students are mis-graded in expectation.

Now suppose the clustering algorithm incorrectly clusters β1(d) non-representative submissions and β2(d) representative submissions randomly, with β1(d) = o(d) and β2(d) = o(Δ). Then the expected total number of is-graded students is

4-12.png

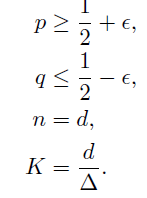
The expected fraction of students mis-graded is



which goes to 0 as d gets large.

PROOF OF THEOREM 3.

The parameters p, q, n and K in [6, Theorem 2.2], when translated to our setting, have the following relations with the parameters of our setting:



The assumption 4-15.pngwith c being large enough, log d ensures that the condition [6, Equation (7)] required by [6, Theorem 2.2] s satisfied. While the statement of [6, Theorem 2.2] guarantees correct clustering of all submissions with a probability at least (1-c2d-c3 ) for some positive constants c2 and c3 that are independent of d, the proof f the theorem establishes the value of the constant c3 as 1. It follows that the number of submissions lustered incorrectly is upper bounded by

4-16.png

REFERENCES

1. Family of Bachmann-Landau notations. http://en.wikipedia.org/wiki/Big\_O\_notation# Family\_of\_Bachmann.E2.80.93Landau\_notations. Retrieved October 29, 2014.

2. Professionals against machine scoring of student essays in high-stakes assessment. http://humanreaders.org/petition/index.php. Retrieved June 1, 2013.

3. Aggarwal, V., Minds, A., Srikant, S., and Shashidhar, V. Principles for using machine learning in the assessment of open response items: Programming assessment as a case study. In NIPS Workshop on Data Driven Education (Dec. 2013).

4. Basu, S., Jacobs, C., and Vanderwende, L. Powergrading: a clustering approach to amplify human effort for short answer grading. TACL 1 (2013), 391–402.

5. Brooks, M., Basu, S., Jacobs, C., and Vanderwende, L. Divide and correct: Using clusters to grade short answers at scale. In Learning at Scale (2014).

6. Chen, Y., and Xu, J. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. arXiv preprint arXiv:1402.1267 (2014).

7. D.ez, J., Luaces, O., Alonso-Betanzos, A., Troncoso, A., and Bahamonde, A. Peer assessment in MOOCs using preference learning via matrix factorization. In NIPS Workshop on Data Driven Education (2013).

8. Glassman, E. L., Scott, J., Singh, R., and Miller, R. C. Overcode: visualizing variation in student solutions to programming problems at scale. In Proceedings of the adjunct publication of the 27th annual ACM symposium on User interface software and technology, 129–130.

9. Gutierrez, P., Osman, N., and Sierra, C. Collaborative assessment. In EDM Workshop on Feedback from Multimodal Interactions in Learning Management Systems (July 2014).

10. Kakkonen, T., Myller, N., Timonen, J., and Sutinen, E. Automatic essay grading with probabilistic latent semantic analysis. In Proceedings of the second workshop on Building Educational Applications Using NLP, Association for Computational Linguistics (2005), 29–36.

11. Kulkarni, C., Pang-Wei, K., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. R. Peer and self assessment in massive open online classes. ACM Transactions on Computer-Human Interaction 9, 4 (2013).

12. Kulkarni, C. E., Socher, R., Bernstein, M. S., and Klemmer, S. R. Scaling short-answer grading by combining peer assessment with algorithmic scoring. In Proceedings of the first ACM conference on Learning@ scale conference, ACM (2014), 99–108.

13. Kullback, S., and Leibler, R. A. On information and sufficiency. The Annals of Mathematical Statistics (1951), 79–86.

14. Larkey, L. S. Automatic essay grading using text categorization techniques. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (1998), 90–95.

15. Luxton-Reilly, A., Denny, P., Kirk, D., Tempero, E., and Yu, S.-Y. On the differences between correct student solutions. In Proceedings of the 18th ACM conference on Innovation and technology in computer science education, ACM (2013), 177–182.

16. Mitros, P., Paruchuri, V., Rogosic, J., and Huang, D. An integrated framework for the grading of freeform responses, 2013.

17. Ng, A., and Jordan, M. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In NIPS (2002).

18. Nguyen, A., Piech, C., Huang, J., and Guibas, L. Codewebs: scalable homework search for massive open online programming courses. In Proceedings of the 23rd international conference on World wide web (2014), 491–502.

19. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. Tuned models of peer assessment in MOOCs. In International Conference on Educational Data Mining (2013).

20. Rees, J. Peer grading can’t work. http://www.insidehighered.com/views/2013/03/05/ essays-flaws-peer-grading-moocs. March 5, 2013.

21. Rogers, S., Garcia, D., Canny, J. F., Tang, S., and Kang, D. ACES: Automatic evaluation of coding style. Master’s thesis, EECS Department, University of California, Berkeley, May 2014.

22. Shah, N. B., Bradley, J. K., Parekh, A., Wainwright, M., and Ramchandran, K. A case for ordinal peer-evaluation in MOOCs. In NIPS Workshop on Data Driven Education (Dec. 2013).

23. Walsh, T. The PeerRank method for peer assessment. arXiv preprint arXiv:1405.7192 (2014).

24. Zhenming, Y., Liang, Z., and Guohua, Z. A novel web-based online examination system for computer science education. In IEEE Frontiers in Education Conference, vol. 3 (2003).

Некоторые законы подобия для MOOC оценок

Проблема, которая возникает с увеличением числа студентов в массовые открытых онлайн-курсов (MOOCs) является оценка студентов. Большое количество студентов делает невозможным для инструкторов или учебных ассистентов оценить все задания, в то время как современные технологии автоматического оценивания не представляются возможными для многих тем. В результате, в последнее время наблюдается рост в использовании коллегиального оценивания, где студенты оценивают друг друга; таким образом, число оценщиков автоматически масштабируется по числу студентов. **Тем не менее, на практике, коллегиальное оценивание показывает высокие коэффициенты ошибок и подвергается критике**. В этой статье, мы берем статистический подход к оценке целесообразности коллегиального оценивания для MOOCs. Наши оценки приводят к некоторым хорошим новостям и плохим новостям. Наше исследование показывает, что на основе простых, но общих моделей, коллегиального оценивания в масштабе, к сожалению, вряд ли хорошо работать в качестве автономного варианта, т.е. S количество студентов увеличивается, **ожидаемое количество студентов неправильно оцененных будет расти пропорционально.** Хорошая новость, однако, что с учетом гибридного подхода, сочетающего коллегиальное оценивание с автоматическим оцениванием с использованием методов по сокращению размерности, мы можем решить задачу масштабирования эффективно. Конкретно, автоматизированный подход используется для ‘уменьшения размерности’, классической техники в области статистики и машинного обучения, и коллегиальное оценивание используется для оценки этого множество представлений низкой размерности.

Наши взносы в текущую литературу MOOCs дважды. Во-первых, мы предоставляем принципиальный аналитический подход к проблемной области, которые преимущественно эмпирически приводом. В этом контексте, мы подходим к проблеме на высоком уровне, чтобы обеспечить фундаментальные законы о масштабировании оценок в MOOCs на основе простых, но общих моделей. Во-вторых, мы предлагаем конструктивный подход к решению проблемы масштабирования, используя инструменты из статистики. Предлагаемый гибридный подход может потенциально сообщить путь развития следующего поколения алгоритмов MOOC -оценки.

**ВВЕДЕНИЕ**

Мы обсудим масштабируемость оценивания, при этом особое внимание уделяется равному-оценивании в массивных открытых онлайн-курсов (MOOCs). Для бюджетных ограниченными OOCs, сортировка инструктором (ы) становится невозможным, как размеры курс расти. Большие курсы требуют альтернативных подходов, таких как авто-оценка или экспертной оценивании. Мы сосредоточим наше обсуждение MOOCs, для которых авто-оценки с использованием заранее подготовленных моделей трудно. Для субъективных темы и сложных проблем, часто трудно разработать машина профилирования системы, которые являются точными [2]. В этих случаях, важно включают людей в оценивании петли. Peer-оценивании является система оценивании, где студенты, принимающие курс повторно классифицированы по сокурсников в тот же курс. Peer-оценивании является естественным выбором для MOOCs так как общее число автогрейдеров в реферируемых оцениваются системе самообслуживания весов пропорционально к количеству студентов развернул. Например, Coursera использует **коллегиальное оценивание** в взаимодействие человека с компьютером (HCI), конечно. Поскольку студенты не эксперт грейдеры, в этом сверстников системы оценивании, представление предусмотрено каждого студента оценивается по 3 до 5 студентов. Итоговая оценка студента вычисляется как среднее из этих отдельных оценок [11]. Альтернативные алгоритмы агрегирования Peer-классов предлагается в [19, 9, 23, 7, 16, 12].

Исследования показали, (например, [11, 12]), что нынешние авто-оценки и сверстников-системы аттестации делают много ошибок. Качественные наблюдения о неточности оценки MOOC привели к критике авто-оценки и коллегиального оценивания [2, 20]. Для MOOC курс кредитов, чтобы получить повышенную признание, важно, что эти ошибки быть снижена.

В этой статье мы рассматриваем проблему оценок в MOOCs через призму статистического анализа. Наш подход ортогональна и красиво дополняет основном эмпирический характер работ в этой области. Изучается поведение масштабирования экспертной оценивании, т.е. поведения, когда число студентов получает большой. Наш анализ показывает, что при разумных предположениях, эти системы будут неправильно сорта постоянную долю студентов в ожидании. Эта постоянная фракция не является проблемой для малых курсов, где инструктор может обрабатывать жалобы от студентов, которые чувствуют, что они были неправильно оценены. Тем не менее, это не масштабируемое решение для крупных курсов со студенческих жалоб будет подавлять инструкторов. Наш анализ дает некоторое понимание, почему текущих сверстников-сортировочных систем MIS-класса либо студенты.  
Важно отметить, что предлагаемая нами структура является гибкой: она позволяет систем сбора разнообразной класса, относится к любой произвольного выбора (адаптивных) заданий классов в представлениях и любом выборе агрегирования Peer-классов.

*В настоящее время усилия по улучшению оценивании изучили многие аспекты: повышение авто-сортировочные модели [14, 24, 10], улучшение алгоритмов агрегирования Peer-классов 19, 9, 23, 7, 16, 12], сочетая авто- и peer-- оценивании [12, 3], а также снижение размерности [4, 5, 18, 21, 8, 15]. Эти усилия показали полезные улучшения в практике. Тем не менее, мы считаем, что эти методы не изменить поведение масштабирования оценивании; в крайнем случае, они уменьшают постоянную долю в МИС-градуированных студентов.*

Второй вклад данной работы является показать, что на верху, такое поведение масштабирования может быть улучшена с помощью комбинации современных методов, используемых в оценках MOOC. В частности, мы представляем алгоритм для оценки, которая использует комбинацию авто- и-экспертной оценивании с "снижения размерности». Мы докажем, что при весьма общих моделей ошибок, наш алгоритм имеет потенциал для создания исчезающих количество ошибок, т.е. для того, чтобы ожидаемая доля студентов неправильно градуированных стремится к нулю размеры МНК курс расти. Этот результат дает статистически звук, конструктивный путь к решению проблемы масштабирования для оценок в MOOCs.  
Любое исследование с участием MOOCs по определению должны обратиться масштаб. т, таким образом, первостепенное значение при изучении вопросов, как оценивании для MOOCs, что мы будем решать наиболее важные немного высокого порядка "в первую очередь. Это мотивирует нас, поэтому для решения сверстников-градуировку для MOOCs с целью сначала понять фундаментальное поведение масштабирования, а именно для изучения законов масштабирования по отношению к простым, но информативным моделей. Это поможет информировать более подробные последующие исследования с участием конкретных алгоритмов и практические решения, необходимые, чтобы сделать это реальностью

РАСШИРЕНИЕ АНАЛИЗ Peer-выбраковки  
Мы заинтересованы в MOOCs, которые массивные и открытым. Другими словами, курсы очень большие, и повторно предлагаются бесплатно или по низким ценам. Эта стоимость ограничение предотвращает число инструкторов из масштабирования, пропорционально числу студентов. Таким образом, можно считать, что преподаватели не могут ручной сорт каждого студента submission.1 Вместо оценивании делается с помощью сверстников-градуировкой, где студенты оценки своих сверстников Доводы, и / или автоматической оценивании, где компьютерная программа оценивает студентов представлений , Так оценивании требует значительных усилий, мы ожидаем, что средний студент в классе на большинстве несколько сверстников. е предположить, что большинство студентов грейдеры работают лучше, чем гадать, но случайная несовершенны, то есть, часто делают ошибки в оценивании; эмпирические исследования показали, что это, чтобы быть правдой (например, [11]). Для нашего анализа, мы предполагаем, что есть "правда" класса для представления друг студентов. Сорта, предлагаемые специалистами классов можно считать соответствовать реальным оценки, но студенты (например, равный-классники) обеспечивают только шумные измерения истинного класса.  
Теорема 1 ниже анализирует поведение масштабирования типичных сверстников-оценивании систем, описанных выше. Для очень общих настройках, мы показываем, что в ожидании, что сорта с постоянной долей студентов будет ошибкой. Для того, чтобы сохранить наши претензии зависит от модели и алгоритма вывода, мы рассмотрим общую ситуацию оценивании и показать, что в этом случае, постоянная доля студентов, имеющих конкретную истинную оценку статистически неотличимы от тех, для студентов, имеющих различный истинную оценку.  
  
1В термин «представление» будет использоваться для ссылки на решения для экзаменов, домашних заданий, или любой þér материалы, представленные для оценки студентами.

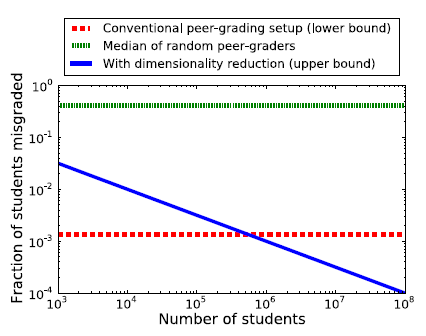


Рисунок 1. Иллюстрация масштабирования ожидаемого фракции студентов MIS-градуированных в трех настроек: (а) нижние границы погрешности при обычных сверстников-сортировочных установок, (б) ошибка в алгоритме экспертной оценивании, который выбирает Peer-классников равномерно наугад и устанавливает окончательную оценку в качестве медианы сверстников-классах, и (в) Верхняя граница погрешности при экспертной оценивании с сокращением размерности. Сюжет считает установку с Q = 0: 25, K = 3, и не совершенных классов или экспертов. В (С), алгоритм кластеризации предполагается кластеров выходной; D / (2 г) журнал и ошибочно кластеров √D нерепрезентативных представлений. Обратите внимание, как доля студентов MIS-градуированных emains постоянной в обычных установках (а) и (б), но падает до нуля с уменьшением размерности (с).  
  
Теорема.  
Рассмотрим следующую настройку. Есть D студентов, обучающихся в курсе. Каждый студент представляет одно представление, и эти D материалы должны быть классифицированы. Каждый студент имеет истинную основную оценку, которая в равной степени вероятно, будет либо "передача" или "провал", и цель вывести эти истинные оценки. Каждый (peer-) грейдер сортов K представлений. Каждый студент является несовершенным грейдер с вероятностью у £ (0, 1], и является идеальным Рейдер Ith вероятность (1 γ). Когда его спросили, в класс представление, несовершенной оценки бульдозерным неправильно с вероятностью д, в течение некоторого д е ( 0, 0: 5), в то время как идеальным грейдер всегда сортов правильно корреляция между Радинг и отвечать навыков может быть также принято:. случайно выбранный студент с истинной степени "проход" является идеальным грейдер с вероятностью р е [0 , 1). Постоянная доля альфа ε [0, 1) представления также может быть raded инструкторами, которые прекрасно сорта. Параметры к, γ, д, р и α являются независимыми от общего числа и контингент D. Значения этих параметров может быть известно с алгоритмом экспертную оценивании.  
Алгоритм экспертной оценивании необходимо сначала назначить (или сверстников инструктор) автогрейдеры каждому представления. Мы позволяют это назначение будет выполняться в полностью адаптивной образом, то есть, алгоритм позволил ждать выхода одного грейдера, прежде чем сделать следующее задание подачи грейдер. Там нет ограничений на количество людей, которые могут оценка любой индивидуальный представление. После получения всех (Peer и инструктор) оценки, алгоритм должен затем назначить итоговые оценки в каждом представлении.  
При любом алгоритма экспертной оценивании, в ожидании, окончательные сортов постоянной доли студентов будет ошибкой.  
Результаты масштабирования, полученные здесь, означает, что, если качество сверстников-классников не улучшится, такие Ith увеличению числа зачисленных студентов, то, как число студентов в увеличении курса, количество неправильно градуированных студентов увеличится пропорционально. Опыт студент будет страдать в таком сценарии, и преподаватели могут столкнуться с подавляющим числом студентов жалоб. Этот анализ масштабирования показано на рисунке 1 наряду с по сравнению с на основе алгоритмов экспертной оценивании размерность сокращения, который будет представлен в следующем разделе.  
Остальная часть этого раздела посвящена предположения и некоторые обобщения теоремы 1. Предположение того, параметры γ, р и д зависит от общего числа зачисленных студентов D означает, что средняя способности студентов не меняются (значительно), когда количество студентов, обучающихся в увеличении курса. Мы предполагаем, что количество представленных ар оцениваются инструкторами увеличивается на наиболее линейно с общего числа студентов D: разумное предположение, для свободно, предлагаемых курсов. Наконец, предположим, к быть постоянным (и не увеличивается в D), так как классификация является задачей, которая требует значительного времени и усилий н часть грейдера, и мы должны наложить ограничение на количество представленных что любой студент должен Peer-класс. Обратите внимание, что это ограничение не ограничивает каждый подчинение быть классифицированы по более к сверстников-х классов.  
Результат теоремы 1 может быть распространена на множество альтернативных настроек. Аналогичное рассуждение справедливо для случая, когда сорта, собранные из классов находятся на более мелком масштабе (с более чем двух возможных значений классов), когда истинное основной сорт не может быть равномерно распределенных, или когда собеседник-градуировка порядковый где студенты попросили сравнить два или несколько представлений вместо назначения числовые показатели в представлениях [22]. Peer-оценивании может быть увеличена также включать оценивании у людей не принимать курс, например, люди, которые ранее принятых курс. Если число таких «внешних первоклассников» остается линейный числа студентов, то рассуждения, аналогичные теореме 1 продолжают применяться. Для того, чтобы обеспечить справедливость, использование посторонней информации о навыков учащихся, как правило, избежать в процессе оценивании [19], как это предполагается в теореме. Один потенциально могли бы получить некоторую информацию о студенческих оценивании способностей от их производительности в предыдущих домашних заданий или тестов, но эта информация не будет менять законы масштабирования до тех пор, как количество домашних заданий / испытания, проведенные остается независимым от количества студентов.  
НАША барьер VIA размерности СОКРАЩЕНИЕ  
В этом разделе, мы предоставляем один из возможных способов взлома барьер постоянной доли представленных будучи градуированных ошибочно: снижение размерности. Мы обсудим два метода снижения размерности для: кластеризации и featurization. Кластеризация является более интуитивный метод, но featurization является более общим. Эти методы объединяют авто-оценки с экспертной оценивании в порядке, которые будут обсуждаться в дальнейшем. Интерфейс сверстников-оценивании остается такой же, как и раньше, где студенты ранга материалы, представленные своих сверстников.  
Кластеризация  
Для простоты изложения, последующее обсуждение будет предположить, представление, ответ на один вопрос. Алгоритм может быть применен отдельно АХ вопросов. Предположим, мы используем компьютерную программу для кластера совокупность всех д материалов, представленных на д студентов по сходству их содержания. Алгоритм кластеризации такова, что в идеале, в каждом кластере, все представленные имеют одинаковую истинную оценку. Обратите внимание, что множество кластеров могут иметь одинаковый класс истинную. Учитывая это предположение, оценки любого одно представление в кластере эффективно чины все материалы в кластере. Общее количество представленных в классе эффективно уменьшается.  
Кластер может быть классифицированы по сбору Peer-классов для любого представления в кластере. Эти сорта могут быть объединены с помощью любого разумного алгоритма, например, принимая медиану полученных сортов. Наконец, представление любого студента назначается класс, который получает его кластера. Этот подход оформляется в алгоритме 1, и показано наглядно на рисунке 2.  
Алгоритм 1. оценивании алгоритм с строгих гарантий: Объединение снижения размерности и экспертной оценивании  
• Кластерные материалы в б кластеров на основе сходства ""  
• Выберите одно представительство представления для каждого кластера  
• Выберите DK / А-классников сверстников для каждого кластера равномерно наугад, обеспечивая отсутствие сверстников-классник не выбрали для более к кластеров  
• Для каждого кластера, назначить представительный подчинение его соответствующих выбранных сверстников-х классов  
• После получения экспертных оценок-для каждого кластера, назначить медиану сверстников-классах в качестве окончательной оценки всех представленных в этом кластере.  
Следующая теорема показывает, что, учитывая хороший алгоритм кластеризации, градация алгоритм, предложенный в алгоритме 1, может уменьшить количество ошибочно отнесенных студентов исчезающей фракции, то есть, убедитесь, что ожидаемая доля студентов MIS-градуированных стремится к нулю. Теорема считает сверстников-оценивании условия хуже, чем считается в теореме 1, а именно, не предполагать существование каких-либо совершенных классов или инструктора классов. В дальнейшем, D (р1 || p2) обозначает дивергенцию Кульбака-Либлер [13] между двух распределений Бернулли с параметрами P1 и P2 соответственно, то есть,  
D (P1 || р2) = p1 войти p\_1 / p\_2 + (1 - р1) войти 〖1-P〗 \_1 / (1-P\_2)  
  
Мы используем стандартный Бахман-Ландау [1] обозначения О (∙), O (∙) и Θ (∙) представлять предельное поведение функций. Неофициально, F (X) = O (г (х)) означает, что F преобладают г асимптотически, F (х) = О (г (х)) означает, что F ограничена сверху по г асимптотически и F (X) = Θ (г (х)) означает, что F является одновременно верхней и нижней ограничена г (х) с различными постоянными предварительных факторов.

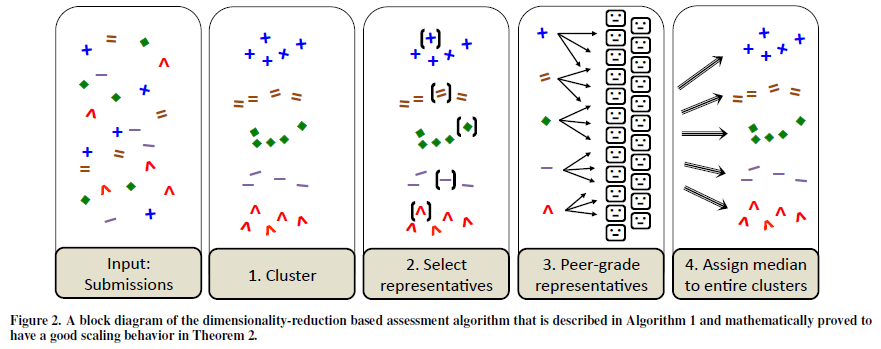
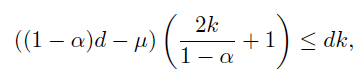
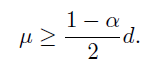


Рисунок 2. Блок-схема алгоритма оценки размерность восстановительной основе, что описано в алгоритме 1 и математически доказано, чтобы иметь хорошее поведение масштабирования в теореме 2.  
  
Теорема 2.  
Рассмотрим следующую настройку. Есть D студентов, обучающихся в курсе. Каждый студент представляет одно представление, и эти D материалы должны быть классифицированы. Каждый студент имеет истинную основной класс, который является либо «проход» или «неудачу», а целью является вывод этих истинных оценки. Каждый (peer-) грейдер сортов K представлений. Когда его спросили, в класс представление, грейдер оценки неправильно с вероятностью д, для некоторых д е (0, 0: 5). Вот, к и д не зависит от количества студентов D.  
Пусть оценивании осуществляется в соответствии с алгоритмом 1. Тогда, если число кластеров удовлетворяет Δ ≤d / log⁡d кД (0,5 || 1- д), и если алгоритм кластеризации используется в алгоритме 1 ошибочно кластеров в большей O (г ) нерепрезентативных материалы и самое о (Δ) репрезентативные материалы наугад, то доля представленных, которые неправильно классифицированы стремится к нулю общего числа представленных д растет.  
  
 Теорема утверждает, что если представленные материалы могут быть сгруппированы в кластеры среднего размера журнала порядок D или выше, то, даже если каждый студент может сорта только постоянное число представленных, количество представленных градуированных неправильно будет зависеть от количества студентов д. Теперь система может масштабироваться для размещения сколь угодно большие числа студентов не беспокоясь числа сортировочных ошибок дует вверх. Этот анализ масштабирования показано на рисунке 1. Теорема предполагает хорошую производительность алгоритма кластеризации. Это может показаться не в ладах с нашей предыдущей дискуссии о выполнении автоматической оценивании алгоритмов, предполагающих, что они не очень хорошо работают по темам, которые носят субъективный характер. Для этого заметим, что в то время как авто-сортировка требует алгоритм, чтобы понять семантику каждого представления для того, чтобы класс, кластеризация только требует понимания особенностей, на которых должны быть выполнены испытания сходство. Работа кластеризации, таким образом, подмножество сложнее задачи автоматического оценивании. Действительно, разработке алгоритмов для кластеризации представления для образовательных оценки является областью активных исследований в сообществе, например, [4, 5, 18, 21, 8, 15]. Кроме того, исследования показали, что студенты, как правило, чтобы ответить аналогичным образом [15, 4], позволяющих число кластеров будет мало.  
Давайте теперь обсудим предположение о выполнении алгоритма кластеризации с точки зрения статистики. Мы предположили, что количество представленных сгруппированы ошибочно строк югу линейно в д. Другими словами, мы предполагаем, что доля представленных сгруппированы неправильно Развивается при увеличении общего числа заявок. Интуитивно понятный обоснование этого предположения с, что кластеризация часто выполняется путем сравнения представлений друг с другом, и, как общее количество F материалы д растет, количество представленных (в каждом кластере) для сравнения также растет, тем самым позволяя алгоритм кластеризации для улучшить свою эффективность в качестве D увеличивается. Использование литературы на статистических гарантий для задач кластеризации (например, [6]), мы можем сделать этот интуиция математически строгим.  
Теорема 3. (Адаптация [6, теорема 2.2]) Рассмотрим алгоритм кластеризации, который работает следующим образом. Алгоритм имеет черный ящик компаратора, что, учитывая любую пару материалов, правильно определяет, принадлежат ли они к одному кластеру или не с вероятностью не менее 1/2 + ε (независимый е всех других пар), для некоторых ε> 0 , Значение е фиксировано, зависит от D, и неизвестно алгоритму. Предположим есть б ≤ cε2d / log⁡d кластеры, одинакового размера, и войти д значение А, как известно, алгоритм. Здесь C является специфическим универсальной постоянной. Тогда существует алгоритм кластеризации, такие, что ожидаемое число е представленных кластерных неправильно ограничена сверху универсальной постоянной (независимо от г).  
Замечание 1.  
Установка [6, теорема 2.2] отличается от настройки экспертной оценки рассматриваемого в этом разделе в некоторых отношениях. Мы предполагаем, что в среднем размер кластера по крайней мере с журнала D (для некоторой постоянной C> 0), и что алгоритм кластеризации не знать размеры этих кластеров, в то время как [6, теорема 0,2] предполагается, что размеры из все кластеры являются идентичными, нижняя ограничена с лог д, и размер каждого кластера, как известно, алгоритм. С другой стороны, результаты [6, теорема 2.2] обеспечивают очень надежную гарантию в том, что только O (1) материалы находятся в неправильно сосредоточены в ожидании, в то время как для наших требований, слабее гарантия о (г) ошибки будет достаточно ,  
Анализ масштабирование представлены в этом разделе показывает, что уменьшение размера из материалов, представленных в логарифмической фактор может быть достаточно для разработки масштабируемой системы оценивании. Предыдущие работы по кластеризации, которые мы справка б кластеризации идентификаторов инструкторов в оценивании. Этот подход может уменьшить нагрузку на преподавателей, которые можно назначить классы для групп материалов. Наш анализ мотивирует, а также обеспечивает теоретическое обоснование использования кластеризации инструментов [4, 5, 18, 21, 8, 15] для массивных открытых онлайн-курсов, где в сочетании с экспертной оценивании, мы показываем, что они могут помочь в достижении масштабируемость в процессе оценки.  
Featurization  
Мы кратко обсудим более общий тип сокращения размерности в этом разделе. Метод кластеризации обсуждалось ранее предполагает, что многие материалы достаточно схожи быть объявлены эквивалентными алгоритма кластеризации. Можно обобщить, чтобы предположить, что части или аспекты многих представлениях похожи и могут по электронной сравнению или сгруппированы; мы описываем это как featurization, где содержание представления, кратко изложена набором функций.  
Предположим, что представления могут быть описаны А функций. Предположим, что он сорт каждого представления может быть вычислен как функция этих функций; например, простой, такие соборование для прохода / не сорта будут пороговой взвешенную сумму функций. Тогда мы свели задачу к сортности традиционной обстановке регрессии: материалы, представленные примеры, особенности вычисляются алгоритмически для каждого примера, и со сверстниками-классы обеспечивают шумные этикетки для примеров.  
Регресс модель просто еще один способ агрегирования экспертных оценок-. Предположим теперь, что мы используем обобщенную линейную модель у (я, т) ~ F (W T х (я)) + е (я; т), где х (I) является функция вектор для представления я, у (я ; т) он TTH сверстников-класс для представления я, е (∙) является функцией обратного ссылку, и ε (я; т) является шум, от сверстников-классника. Эта модель обобщает простые сверстников-сортировочные системы, которые лечат все материалы самостоятельно: D логические функции показатели, соответствующие д материалов, поэтому данная функция вектор х (я) для представления я есть "1" для функции я и "0" в другом месте. Модель также обобщает кластеризации: у нас есть А логические функции, соответствующие кластеры б.  
Эта установка может позволить системы аттестации в сырой п обширных исследований по инженерно художественного и моделирования. Успех предыдущей работы по художественного основе кластеризации представленных в MOOCs указывает, что полезные функции могут быть найдены. Тот факт, что нынешние методы агрегации для сверстников-классах могут быть обобщены с помощью простых моделей регрессии показывает, что такие модели являются обоснованными. В зависимости от выбора модели регрессии, можно добиться масштабирования результаты, аналогичные для кластеризации. Например, рассмотрим модели логистической регрессии (с функцией связи логит), чтобы классифицировать представления, как прошел / не прошел. Тогда предыдущая работа [17] на логистической регрессии показал шляпу, с особенностями А и Θ (d) образцов (Peer-классы), достаточно иметь Д = O (D).  
Предыдущая работа считается объединения автоматического профилирования с экспертной оценивании [12, 16]. Оба эти произведения используют адаптивную процесс для оценки, с отдельным автоматическим оценивании процессов, используемых в дополнение к процессу экспертной оценивании. Тем не менее, ни одна из этих работ не имеют строгие гарантии по производительности. Кроме того, эти работы отдельных сверстников-оценивании и авто-оценивании в несколько этапов, и сочетают в себе авто-классы и Peer-классов в определенном порядке, чтобы вычислить окончательную оценку. Этот подход оставляет законы масштабирования без изменений: либо авто-сортировка делает постоянную долю оценивании ошибок (в этом случае постоянная доля ошибок остаются после экспертной оценивании, а), или авто-оценки дает исчезающий часть ошибок (в котором Дело сверстников-оценивании становится ненужным для целей получения исчезающих количество ошибок).

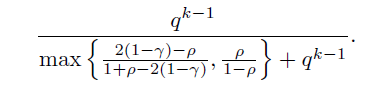
ВЫВОДЫ  
 В этой статье, мы дали статистический анализ масштабных свойств различных механизмов профилирования в MOOCs. Мы видели, что при очень простых и общих моделей, виды экспертной оценивании систем, используемые сегодня не будет масштабироваться. Затем мы показали, что сочетание (Авто) уменьшение размерности и сверстников-оценивании как потенциал для расширения. Снижение размерности уже активно темой исследований [4, 5, 18, 21, 8, 15], а также предложение объединения его с экспертной оценивании попадает под более общей парадигмы объединения машину и человеческий интеллект [12, 3] ,  
В то время как большинство современных исследований по оценке в MOOCs эмпирические, этот документ при условии, более аналитический подход, помогая понять основные принципы, лежащие в экспертной оценке в современных системах оценивании и идентификации путь для дальнейших исследований, чтобы преодолеть ошибки оценки. Точные, надежные, масштабируемые и оценка поможет проложить путь для MOOCs демократизировать образование.  
ПРИЛОЖЕНИЕ: Доказательства  
В этом разделе представлены доказательства для трех теорем, изложенных в статье. Доказательства сосредоточиться на демонстрации законы масштабирования и не пытаться оптимизировать константы (участки рисунке 1 оптимизировали константы). Так как мы заинтересованы только в законах масштабирования без потери общности, мы предполагаем, д быть достаточно большой, чтобы игнорировать любые пол / потолок эффекты целых чисел.  
Доказательство теоремы 1.  
Рассмотрим любые значения параметров D, K, Q, А, Г, р, как определено в формулировке теоремы. Рассмотрим любые сверстников-оценивании алгоритм А, которая минимизирует ожидаемая доля студентов неправильно градуированных. Высокая уровне идея доказательства заключается в следующем. Мы будем выбирать двух студентов случайно равномерно из множества д студентов. Мы покажем, что с вероятностью меньше ограниченной положительное значение, независимо от D, производительность двух студентов появляются статистически неразличимы алгоритм А, независимо от их истинных, лежащих в основе оценки. С кондиционером на статистически неразличимы, алгоритм А будет неправильно класса, по крайней мере один из этих двух студентов в ожидании. Объединение по всем студентам, мы покажем, что он часть студентов MIS-градуированного ограничены снизу положительным значением, что не зависит от г. Учитывая формулировка теоремы 1, и это на высоком уровне идея, остальная часть математике за доказательства достаточно прост.  
Под обстановке, рассмотренного в теореме, по крайней мере (1-α) D Доводы оцениваться не будут инструкторами. Эти материалы должны быть оценены на основе экспертных оценок-возложенных на них и точности этих (1-a) D студентов в коллегиальных оценивании, что они выполняли. Общее количество сверстников-классах, доступных для алгоритма А ограничена сверху Д.К.. Пусть μ обозначает число представлений из этой коллекции (1-α) D материалов, которые получили не более 2k / (1-α) Peer-классов. Принцип предполагает, что откладывать в долгий ящик



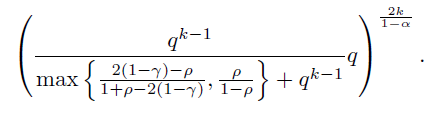
и некоторые алгебраические манипуляции приводят к



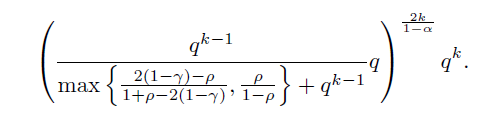
Давайте на мгновение сосредоточить наше внимание только на этих представлениях М. Предположим, вы выбираете два представления равномерно случайным образом из этого набора М представленных, скажем, представлений студентов S1 и S2. Выводится нижняя граница вероятности того, что производительность студент s2 является статистически неотличимы от студенческой s1, независимо от истинных сортов студентов S1 и S2. Рассмотрим он Peer-оценивании процесс студенческой s2 под алгоритма A. Для того, чтобы получить экспертную оценку-S2 для представления, в Алгоритм А может хотите зарегистрировать идеальным или несовершенный грейдер. Алгоритм имеет следующую информацию в своем распоряжении для выведения ли определенный сверстников-грейдер совершенным или несовершенным: производительность этого сверстников-классника на (к - 1) или меньше материалы шляпе она уже дифференцированного и оценку собственного представления Этот узел грейдер вкупе со знанием корреляции между хорошими грейдеров и хорошими оценками.  
Напомним, что в условиях рассматриваемого случайно выбранный студент является несовершенным грейдер с вероятностью у и совершенным грейдер с вероятностью (1-Г). Случайно выбранный студент имеет "пропуск" или "провал" класс на собственном представлении с вероятностью 0,5 каждый. Учитывая класс частот, студент является идеальным грейдер с вероятностью р. Для того, чтобы алгоритм, чтобы запутать несовершенную грейдер для идеального грейдер, несовершенство грейдер должен правильно классифицированы все (в большинстве (к - 1)) Доводы, назначенные ей, и должны иметь истинную основной класс, который является наиболее распространенным среди идеально грейдеры. Некоторые алгебраические преобразования приводят к нижняя граница вероятности ф алгоритма приняв несовершенный грейдер для идеального грейдер как



При несовершенной грейдер случается выбран, есть водный ε (0, 0.5) вероятность того, что она будет неправильно сорта и (1 - д) вероятность того, что она будет правильно сорт. Отсюда следует, что для любого 2k / (1-alpha) произвольной последовательности или меньше градуировки для студентов s2, вероятность наблюдения этой последовательности ниже ограничена

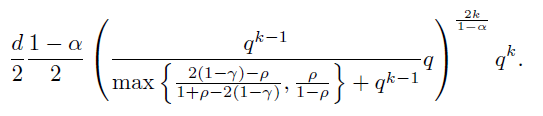


Сорта двух студентов S1 и S2 может также зависеть от их деятельности в качестве Peer-классов. Вероятность того, что студента s2-аналоги сорта в способом, аналогичным способу s1 меньше, ограничена Qk. Из этого следует, что вероятность того, что эти два случайно выбранных студентов появляются статистически неразличимы алгоритму А, меньшим, ограничена

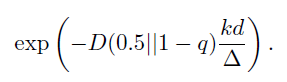


С кондиционером на студентов S1 и S2 появляться статистически неразличимы алгоритме независимо от их истинных оценок, алгоритм будет неправильно класса, по крайней мере один из них в ожидании.

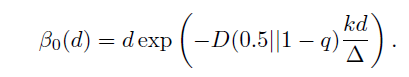
Вернемся теперь к множеству всех д студентов. Объединение все аргументы сверху, мы получаем, что алгоритм будет неправильно сорт, по крайней мере



Студенты в ожидании. Эта величина является линейной в д доказывая тем самым наше утверждение.  
Доказательство теоремы.  
Сначала рассмотрим установку совершенного алгоритма кластеризации. Общее количество классов, полученных из студентов кд. Пусть β0 (г) обозначим ожидаемое количество студентов MIS-градуированных когда общее число зачисленных студентов д. Мы покажем, что в условиях описанного в формулировке теоремы, В0 (г) = O (1). Поскольку существует А кластеры, каждый кластер получает оценки. Каждый сорт правильно с вероятностью (1 - д) е (0,5, 1], и следовательно, применяя Чернова связаны, мы получаем, что вероятность неправильной градуировкой любого отдельного кластера ограничена сверху по



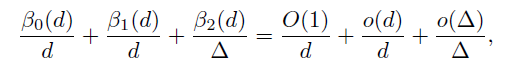
Таким образом, в ожидании, общее число неправильно классифицированы студентов



Подставляя 4-11.pngкак предполагалось в формулировке теоремы, мы получаем, что только O (1) студенты неправильно градуированные в ожидании.  
Теперь предположим, что алгоритм кластеризации неправильно кластеров В1 (D), не представительных доклада и В2 (г) представитель материалы случайно, с & beta; 1 (г) = O (г) и & beta; 2 (г) = O (Δ). Затем ожидается общее число является-градуированных студентов

4-12.png

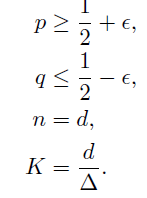
Ожидаемый доля студентов MIS-градуированного



который идет к 0 при г становится большим.

Доказательство теоремы 3.

Параметры P, Q, N и К в [6, теорема 2.2], при переводе в нашей ситуации, имеют следующие отношения с параметрами наших условиях:



Предположение 4-15.pngс с достаточно крупным, необходимо войти D гарантирует, что состояние [6, уравнение (7)] требуется [6, теорема 2.2] S удовлетворены. В то время как заявление [6, теорема 2.2] гарантирует правильную кластеризацию всех представленных с вероятностью по крайней мере (1-C2D-с3) для некоторых положительных констант c2 и c3, независимыми от D, доказательство е теоремы устанавливает значение постоянная с3 как 1. Отсюда следует, что количество представленных lustered неправильно ограничена сверху по

4-16.png