Detecting Cheaters in MOOCs Using Item Response Theory and Learning Analytics

Giora Alexandron
Massachusetts Institute of Technology
giora@mit.edu

Sunbok Lee
Massachusetts Institute of Technology
sunbok@mit.edu

Zhongzhou Chen
Massachusetts Institute of Technology
zchen22@mit.edu

David E. Pritchard
Massachusetts Institute of Technology
dpritch@mit.edu

ABSTRACT

The focus of this work is on developing a general method for identifying cheaters in MOOCs in a way that does not assume a particular method of cheating. For that, we develop a classification model that takes as input a set of features that operationalize performance and behavioral parameters that are known to be associated with cheating. These include students' ability, the level of interaction with the course resources, solving time, and Item Response Theory (IRT) person fit parameters. We start with a list of six candidate features, and after a feature selection process, remain with four. We use these to build a probabilistic classifier (logistic regression) that yields an Area Under the Curve (AUC) of 0.826. Our data is based on an Introductory Physics MOOC. The features are computed using data-mining and standard IRT packages. We consider only the users who received a certificate in the course. Each of these users is considered as an example for the classifier. The positive examples are the set of users who were detected as "using multiple accounts to harvest solutions" by a different algorithm that was reported in a previous publication.

CCS Concepts

•**Applied computing** → **Interactive learning environments; E-learning;** •**Information systems** → *Data mining;*

Keywords

Academic dishonesty; MOOCs; learning analytics; Item Response Theory

1. INTRODUCTION

Academic dishonesty is a serious problem, with studies reporting that up to 95% of college students are engaged in academic dishonesty of some form [3, 8, 9, 15, 18]. In online setting, Palazzo et al. [16] found that between 3 and 11% of the submissions in an interactive online learning system were copied.

Massive Open Online Courses (MOOCs) are a relatively new domain, with certificates that currently do not have formal value (except for few pilot programs). However, several studies already reported the non-surprising findings that cheating exists also in MOOCS. According to [1, 2, 14, 17], between 1 to 10% of the students are using multiple accounts to harvest solutions. This cheating method was dubbed CAMEO (Copying Answers using Multiple Existence Online [14]; We refer to a person who uses this method as *CAMEO user*).

The amount of cheating that has been reported so far for MOOCs involves only the use of CAMEO and requires that the master and the harvester accounts can be linked by IP. Since there are certainly other forms of cheating in MOOCs (including performing CAMEO using accounts not linked by IP), the above is only a lower bound to the actual size of this phenomenon.

The main risk posed by cheating is decreasing the perceived value of the MOOC certificates, since a significant amount of cheating reduces the confidence that the certificate truly reflects students' ability. For example, we found that students who used CAMEO gained almost half of standard deviation in their IRT ability by using this method.

Another risk of cheating is affecting the results of educational research [1, 17]. We found that CAMEO users had better performance, both in terms of success rate and response time, than the rest of the certificate earners in the course. In addition, the CAMEO users that we observed tended to do a lot of questions, but not to interact a lot with the instructional materials. This might lead to a false conclusion that in our course it is better (or even suffice) to spend time on doing questions, rather than learning from the instructional materials.

MOOC providers acknowledge the fact that cheating is a problem that they need to address, and use various proctoring systems. Currently these systems are mainly designed against impersonating. They are not effective , for example, against CAMEO, and probably also against other methods that are still unknown.

The goal of this work is to bypass the possibility of students designing methods to specifically thwart the CAMEO detectors by developing a general detection method that is not tailored to a specific form of cheating, but rather relies on measuring aspects of behavior that are either associated with or affected by cheating. The aspects that we currently consider include the amount of interaction with the course resources, time to answer, student's ability, and two person-fit parameters obtained from IRT – Guttman error [10], and the standard error of ability estimates.

The rationale for using the amount of interaction with the resources is based on the assumption that cheaters will have less interaction with the instructional resources, as they do not need them to solve the questions on which they cheat.

Very fast time to answer was identified by [16] as a strong signal for cheating.

The rationale for using student's ability is that cheaters tend to have a relatively high performance comparing to the rest of the certificate earners [17].

The rationale for using person-fit parameters is based on the assumption that cheaters have a relatively 'noisy' performance, as their performance depends not only on their ability, but also on whether they cheat or not. Following this rationale, researchers in the psychometrics community developed various person-fit indexes to measure unusual response patterns, including cheating [6, 11]. Among them, *Guttman error*, which measures the number of item pairs in which an easier item is answered incorrectly and a more difficult item is answered correctly, was shown by Meijer [10] to be a simple and effective person-fit index for identifying cheating. Thus, we use this parameter.

In addition, the standard errors of ability estimates in IRT model could also be used as a measure of unusual response patterns. The rationale behind using this measure is that an aberrant response provides inconsistent psychometric information, and thus leads to an increase in the standard error of the ability estimates [7].

Using these parameters, we train a probabilistic classifier (logistic regression) on data that contain $\tilde{1}0\%$ cheaters who used CAMEO, and were identified by algorithms that their description and verification process are described in detail in [1, 17]. On this data, the classifier achieves an AUC of 0.826.

To the best of our knowledge, this is the first study that suggests a general method for detecting cheating in MOOCs. It does so by combining machine learning, psychometrics, and learning analytics. Thus, we believe that the results are of interest for the educational data science research community, though these results are still preliminary.

The rest of this paper is arranged as follows. In Section 2 we present in detail the data and the method. In Section 3 we present the results. Discussion, limitations and future work are presented in Section 4

2. DATA AND METHODS

2.1 Data

We use the data from the Introductory Physics MOOC 8MReVx given by the third and fourth listed authors in summer 2014 through edX.org. The course covers the standard topics of a college introductory mechanics course. It contains 273 e-text pages, 69 videos, and about 1000 problems (checkpoints problems embedded within the e-text and videos, and homework and quiz questions which are given at the end of the units). About 13500 students registered to the course, and from them, 502 earned a certificate. For this research, we considered 495 out of the 502 certificate earners (7 were omitted due to technical reasons). Among these 495 certificate earners, 65 were detected as CAMEO users (namely, users who harvested answers using multiple accounts) by the algorithm reported in [1], which is a modification of the algorithm presented in [17]. Both algorithms were verified using manual and statistical inspection methods (a full description of the algorithms and the verification process can be found in [1, 17]).

2.2 Feature selection

2.2.1 Predictors

We start with an initial set of predictors that divides into two groups:

Behavioral parameters:

i. Fraction of videos watched.

ii. Fraction of correct answers that were submitted in less than 30 seconds (the cutoff considered by [16]).

iii. Mean time for submitting a correct answer.

(For ii and iii, the submission time is operationalized as the gap between the time of entering the page in which the problem resides, and the time the correct answer is submitted.)

The rationale for using these parameters is described in Section 1. These parameters were mined from the logs using standard scripts.

Ability and person fit parameters:

iv. Student's ability, computed by a two parameter logistic (2PL) model in IRT using the BILOG software package. The input to the IRT algorithm is a binary response matrix computed from students logs. The response matrix contained only the certificated students (accounts), and items that were answered by at least 50% of these students.

v. Guttman error – the number of item pairs in which an easier item is answered incorrectly and a more difficult item is answered correctly.

vi. Standard error of student's ability estimate from IRT.

The two person-fit parameters – Gutmman and standard error, where computed using the output of the 2PL model used to estimate students' ability.

2.2.2 Dependent variable

It is a binary variable that indicates whether the student is a CAMEO user, namely, used multiple accounts for harvesting solution in our course. The positive examples are the accounts that were identified as CAMEO users by the algorithms described in [1, 17].

2.2.3 Initial feature set

For each user, we build an example vector containing the values computed for this student for parameters i-vi, and the cheater/non-cheater tag. Together these form the feature set.

2.2.4 Standardizing the data

The independent variables were standardized using *z-scores*, so that we can compare the relative importance of features based on standardized logistic regression coefficients [12].

2.2.5 Removing redundant features

To remove redundant features, we use a $L1$ regularized logistic regression and pick the features that have a non-zero coefficient [5]. This is implemented using R's glmnet package [4]. The features that are found to be redundant are the *mean-submission-time* (iii), and the *ability parameter* (iv).

2.2.6 Final feature set

After removing the redundant features, we remain with a feature set containing four predictors: Standard error for IRT student ability parameter, Guttman error, fraction of videos watched, and fraction of questions answered in less than 30 seconds.

2.3 Classification model

We use this set to build probabilistic classifier using a logistic regression. The classifier is evaluated by examining the area under the ROC curve, using k-fold cross-validation. The results are presented below.

3. RESULTS

Below we present, per feature, the difference in the distribution of the values among cheaters and non-cheaters, and the results of the classifier that is built on these features.

3.1 Difference between cheaters and non-cheaters – individual parameters

For each of the four features, there is a statistically significant difference between the cheaters and the non-cheaters (*p-value* < 0.001 for all the features). Table 1 shows, per feature, the standardized mean value for each group.

Table 1: Mean values (standardized).

Feature	Cheaters	Non-cheaters
Standard error	0.36	-0.05
Guttman error	0.94	-0.14
Very quick answers	0.99	-0.14
Videos watched	-0.77	0.11

The distribution of these values is presented in Figure 1. The upper left figure shows the fraction of questions that were solved (correctly) in less than 30 seconds. For non-cheaters, this percentage is very small (the blue sharp curve), while for cheaters, this is much higher. The right upper figure shows the distribution of the fraction of videos watched. The red sharp curve shows that most of the cheaters watched a very small fraction of the videos, relative to the non-cheaters. The bottom left curve shows the standard error for the ability parameter. As can be seen, non-cheaters tend to have a smaller error. However, for this feature the distinction is less clear. The bottom right figure shows the distribution of the Guttman error. Again, non-cheaters tend to have lower Guttman error.

In all the figures, the relative smoothness of the cheaters' curve might be related to the differences in the amount of cheating among the cheating students (ranging from 1% to more than 50% of the correct answers).

3.2 Performance of the classifier

We used the feature set to build a logistic regression classifier, using R's e1071 package [13]. The performance of the classifier was evaluated by examining the area under the ROC curve (AUC), using a 3-fold cross validation. We pick $k = 3$ because the data is biased (about 10% cheaters and 90% non-cheaters), and we want to ensure that in each iteration, with high probability both the training and the test sets will include sufficient number of positive examples (cheaters).

AUC. Overall, the AUC of the model was 0.826, with a variance of 0.0016. These results are for 3-fold cross-validation, ran for 500 times.

ROC curve. Next, we divide the data at random to 2/3 training set and 1/3 test set. The AUC on the test set of a model built on the training set was 0.852. Figure 2 shows the ROC curve for this model.

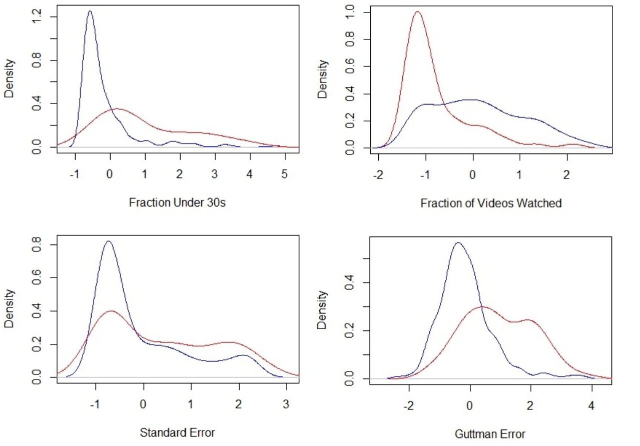

Figure 1: Distribution of the parameters among cheaters (red) and non-cheaters (blue)

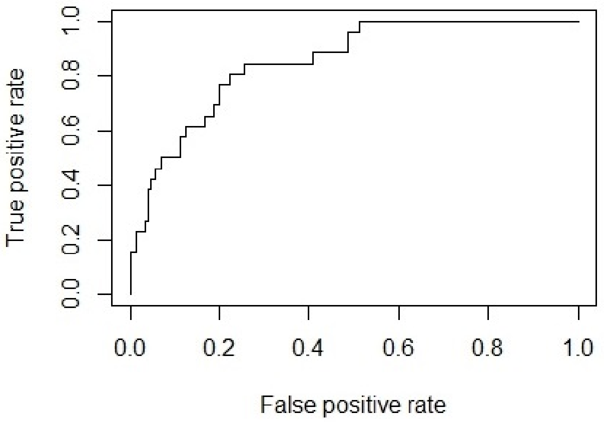

Figure 2: The ROC curve.

To compute the optimal cutoff, we look for the optimal point on two metrics. One is the cutoff that minimizes the distance from the ROC curve to the optimal classification point (0,1). The other is the cutoff the maximizes the sum of the true-negative and the true-positive rates. The cutoffs are 0.148 and 0.149, respectively.

4. DISCUSSION

Our results show that a probabilistic classifier that uses four features – two IRT person fit parameters (Guttman error and standard error), and two simple learning analytics parameters (fraction of videos watched and fraction of correct answer in less than 30 seconds), can detect users who used multiple accounts to collect correct answers ('CAMEO users') in the 2014 run of 8.MReVx MOOC with a good level of accuracy (AUC of 0.826).

The crux of our approach is using 'circumstantial evidence' that is associated with cheating, but is not specific to a certain method (e.g. CAMEO). Thus, we believe that such a model can identify students who use other forms of cheating. One of the steps that we take to evaluate the feasibility of this approach is examining the logs of the 'false positives' – accounts that are identified by the algorithm as cheaters, but were not detected as CAMEO users by the CAMEO algorithm.

Our analysis indicates that at least some of these 'false positives' are students who are identified by the CAMEO algorithm as 'suspicious users' but are filtered because they also behave as *harvesters* (the accounts that are used to collect the correct answers). We believe that these are actually users who collaborate with each other, and for example divide some of the work between them (and thus their account sometimes appears as the account that collects the answers, and sometimes as the accounts that uses the answers collected by another account).

We regard it as likely that our methods will generalize to other courses, as we see no reason to believe that the Introductory Physics course that we studied is specifically attractive to cheaters. Thus we expect to see cheating in other MOOCs as well, and we believe that this will be also associated with performance and behavioral patterns that could be used to distinguish between cheaters and non-cheaters.

4.1 Limitations

Generalizing the results. The main limitation for generalizing our results to other courses is the fact that our data is based on one course. Generalizing to (identifying) other methods of cheating is limited by the fact that we trained our model on data that includes cheaters who used a specific method.

Post-factum analysis. The approach that we present in this paper is post-factum in nature, and is less suitable for identifying cheating events as they occur. Because it relies on 'circumstantial evidence', rather than on a direct evidence for a specific kind of cheating, it is required to accumulate a considerable amount of evidences in order to achieve a sufficient level of confidence. However, it seems reasonable to assume that this can be done during the course (for example, at the end of each chapter).

4.2 Future work

Main directions for future research include studying additional features (e.g., person-fit and behavioral parameters) that can be used to improve the classification, and extending the study to more courses and other forms of cheating.

5. ACKNOWLEDGMENTS

The authors would like to thank Jose Antonio Ruiperez Valiente for his help in conducting this research.

6. REFERENCES

[1] G. Alexandron, J. A. Ruiperez-Valiente, Z. Chen, P. J. MuÃśoz-Merino, and D. E. Pritchard. Copying@scale: Using harvesting accounts for collecting correct answers in a mooc. Manuscript under review.

[2] G. Alexandron, J. A. Ruiperez-Valiente, and D. E. Pritchard. Evidence of mooc students using multiple accounts to harvest correct answers, 2015. Presentation given at Learning With MOOCs II, NY, October 2015.

[3] S. Davis. Academic dishonesty in the 1990s. *The Public Perspective*, 1993.

[4] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[6] G. Karabatsos. Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16(4):277–298, 2003.

[7] F. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum Associates, 1980.

[8] D. L. McCabe and L. K. Trevino. Academic dishonesty: Honor codes and other contextual influences. *The Journal of Higher Education*, 64(5):522–538, 1993.

[9] D. L. McCabe, L. K. Trevino, and K. D. Butterfield. Cheating in Academic Institutions: A Decade of Research. *Ethics & Behavior*, 11(3):219–232, July 2001.

[10] R. R. Meijer. The number of guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, 18(4):311–314, 1994.

[11] R. R. Meijer. Person-fit research: An introduction. *Applied Measurement in Education*, 9(1):3–8, 1996.

[12] S. Menard. Six approaches to calculating standardized logistic regression coefficients. *The American Statistician*, 58(3):218–223, 2004.

[13] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2015. R package version 1.6-7.

[14] C. Northcutt, A. D. Ho, and I. L. Chuang. Detecting and preventing "multiple-account" cheating in massive open online courses. *CoRR*, abs/1508.05699, 2015.

[15] D. J. Palazzo. Detection, patterns, consequences, and remediation of electronic homework copying, 2006. Masters Thesis.

[16] D. J. Palazzo, Y.-J. Lee, R. Warnakulasooriya, and D. E. Pritchard. Patterns, correlates, and reduction of homework copying. *Phys. Rev. ST Phys. Educ. Res.*, 6:010104, Mar 2010.

[17] J. A. Ruiperez-Valiente, G. Alexandron, Z. Chen, and D. E. Pritchard. Using multiple accounts for harvesting solutions in moocs. In *Proceedings of the Third ACM Conference on Learning @ Scale*. ACM, 2016.

[18] A. C. Singhal. Factors in students' dishonesty. *Psychological Reports*, 51(3):775–780, 1982.