Toward Replicable Predictive Model Evaluation in MOOCs

Josh Gardner, Christopher Brooks
School of Information
University of Michigan
{jpgard, brooksch}@umich.edu

ABSTRACT

In this paper, we present and apply a procedure for evaluating predictive models in MOOCs. First, we expand upon a procedure to statistically test hypotheses about model performance which goes beyond the state-of-the-practice in the community and covers the full scope of predictive model-building in MOOCs. Second, we apply this method to a series of algorithms and feature sets derived from a large and diverse sample of MOOCs ($N = 31$), concluding that several models built with simple clickstream-based feature extraction methods outperform those built from forum- and assignment-based feature extraction methods.

1. INTRODUCTION AND RELATED WORK

Building predictive models of student success has emerged as a core task in the fields of learning analytics and educational data mining.[1] The process of building such models in MOOCs involves at least three key stages: (1) extracting structured data and informative features from raw platform data (clickstream server logs, database tables, etc.); (2) selecting algorithms and models; and (3) tuning hyperparameters. Together, these stages profoundly influence the performance of predictive models. We identify at least two methodological gaps in current educational data mining research as it relates to this task: (1) current research typically isolates these steps, e.g., evaluating different approaches to feature extraction or algorithm selection separately without considering their relation to each other; and (2) procedures for rigorous and reproducible statistical inference about the relative performance of these models, and accounting for the many model specifications considered in the course of an experiment, are often not followed.

Previous predictive modeling research in MOOCs has evaluated features derived from clickstreams, discussion fora, assignments, and surveys, among other sources. In addition, this research has applied a variety of algorithms to such data for dropout prediction, including linear and logistic regression, support vector machines, tree-based methods, ensemble methods, neural networks, and deep learning. However, a literature survey by the authors indicated that accepted statistical practices for evaluating these models are often neglected by such research[2] In particular, more than half of surveyed research did not utilize any statistical testing for evaluating model performance, despite obtaining estimates directly on the training set through cross-validation for multiple models. These methods are susceptible to spurious results and low replicability due to multiple comparisons, biased performance estimates, and random variation from resampling schemes [3, 4, 7, 11]. Recent research has provided evidence that some MOOC research may not be replicable when applied to new or different courses [1]; at the very least, this highlights the importance of adopting reproducible and statistically valid methods for model evaluation in MOOCs [8]. An extensive literature exists on statistically reliable methods for model evaluation [4, 6, 11].

2. METHODOLOGY

We implement a testing and inference procedure from [3] for selecting the best of $k > 2$ models across $N > 1$ datasets (in this experiment, a *model* is a *feature set-algorithm-hyperparameter combination*), which consists of two steps. First, a Friedman test is used to test the null hypothesis that the performance of all models is equivalent [5]. The Friedman statistic

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right] \tag{1}$$

where R_i^j is the rank of the jth of k algorithms on N datasets and the statistic is χ_{k-1}^2 distributed, is compared to a critical value at the selected significance level ($\alpha = 0.05$ in this experiment). If H_0 is rejected, then we proceed to the second stage, the post-hoc Nemenyi test, where

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{2}$$

is used to determine whether the performance between any two classifiers is significantly different, where q_α is based on the Studentized range statistic divided by $\sqrt{2}$.

This two-stage procedure allows us to conduct comparisons across multiple models and datasets to draw inferences about

[1]The current work evaluates models of student dropout in MOOCs, but this methodology applies to any supervised predictive modeling task.

[2]This survey reviewed the 2014-2016 International Society for Educational Data Mining (EDM) and the International Learning Analytics and Knowledge (LAK) conference proceedings, and included research which attempted to predict completion or performance using behavioral or academic features with features derived from MOOC platform data; a full survey is forthcoming in a future work.

whether true performance differences exist, accounting for the number of comparisons k and datasets N. Unlike using simple average cross-validated training performance, this procedure uses statistical testing to evaluate whether the observed difference is statistically significant or may be merely spurious, based on the available data. In applying this method to a *feature set + algorithm + hyperparameter* combination, we can (1) evaluate feature extraction as a testable modeling component; (2) capture and evaluate the synergy between feature extraction, algorithm, and hyperparameters; and (3) draw inferences which fully account for the number of comparisons across all of these elements. [3]

3. EXPERIMENT AND RESULTS

As an illustrative example, we compare a series of models using three feature sets and two predictive algorithms on a set of 31 offerings of 5 unique courses offered by the University of Michigan on Coursera, with 298,909 total learners. From the raw clickstream files and database tables, we extracted a series of features intended to replicate (with some additions) features shown to be effective dropout predictors, with each utilizing information from a different raw data source: *clickstream* [10], *assignment* [9], and *forum* features [1].

We train two classifiers – standard classification trees and adaptive boosted trees – on various combinations of the three feature sets, performing no hyperparameter tuning (to limit the number of comparisons, k). Figure 1 presents the results of our analysis.

Results from dropout prediction after course week 2 are shown in Figure 1, but our findings were consistent across all four weeks examined. We find that models utilizing clickstream features consistently outperform those using forum and quiz features. This difference was statistically significant for all model configurations tested. Changing the classification algorithm had little effect on the performance of quiz- and forum-featured models, which were statistically indistinguishable from each other in every week evaluated. When the clickstream features are combined with forum and quiz features to form a "full" model, this model achieves better performance than the clickstream features alone, but this improvement is never statistically significant over the best clickstream-only model. This suggests that the forum and quiz features contain useful structure which may require powerful, flexible classification algorithms to capture. Our conclusion – that the highest-performing model is statistically indistinguishable from other models in this analysis – stands in contrast to the practice of much of the prior research surveyed, which often concludes that the best average performance is the "best" model; this is intended to serve as an example for inferential language in future research.

4. FUTURE RESEARCH

Future research should utilize this or other methods for statistically evaluating performance comparisons of predictive models. In particular, it should explore Bayesian methods for model evaluation, which allow the direct estimation of probabilities of hypotheses, avoid concerns about multiple comparisons, and have other additional advantages [2].

[3]There are clear advantages to adopting this specific procedure over other testing approaches such as ANOVA, or other nonparametric approaches; see §3.2.1 of [3] for detailed discussion of these benefits.

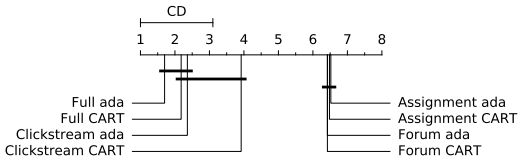

Figure 1: Critical Difference (CD) diagram of week 2 dropout prediction models. Models are plotted by average rank, with bold CD lines indicating statistically indistinguishable models (at $\alpha = 0.05$). We reject H_0 of equivalent performance for models not connected by CD lines. These results show a statistically significant performance gap between clickstream features and assignment or forum features.

5. REFERENCES

[1] J. M. L. Andres, R. S. Baker, G. Siemens, D. GAŠEVIĆ, and C. A. Spann. Replicating 21 findings on student success in online learning.

[2] A. Benavoli, G. Corani, J. Demsar, and M. Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. 14 June 2016.

[3] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7(Jan):1–30, 2006.

[4] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, 10(7):1895–1923, 15 Sept. 1998.

[5] M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Stat.*, 11(1):86–92, 1940.

[6] S. Garcia and F. Herrera. An extension on"statistical comparisons of classifiers over multiple data sets"for all pairwise comparisons. *J. Mach. Learn. Res.*, 9(Dec):2677–2694, 2008.

[7] C. Nadeau and Y. Bengio. Inference for the generalization error. *Mach. Learn.*, 52(3):239–281, 2003.

[8] F. van der Sluis, T. van der Zee, and J. Ginn. Learning about learning at scale: Methodological challenges and recommendations. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, L@S '17, pages 131–140, New York, NY, USA, 2017. ACM.

[9] K. Veeramachaneni, U.-M. O'Reilly, and C. Taylor. Towards feature engineering at scale for data from massive open online courses. 20 July 2014.

[10] W. Xing, X. Chen, J. Stein, and M. Marcinkowski. Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Comput. Human Behav.*, 58:119–129, 2016.

[11] O. T. Yildiz, E. Alpaydin, and Senior Member. Ordering and finding the best of k>2 supervised learning algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(3), 2006.