Success Prediction in MOOCS based on a transfer learning approach

Antoine Pigeau, J Wu, N Ma

Success Prediction in MOOCS based on a transfer learning approach

A. Pigeau, J. WU and N. MA

LS2N, University of Nantes
2 rue de la Houssinière, Nantes, France
firstName.LastName@univ-nantes.fr

ABSTRACT

Massive Open Online Courses (MOOCs) typically present a high rate of non completing learners. Studying the characteristics of pass and fail learners should help to provide assistance and identify root causes for dropouts. We propose in this work to improve a success prediction task on a specific course with a solution based on a transfer learning approach. Our experiments are validated on two datasets with different trace properties.

1. INTRODUCTION

Massive Open Online Courses (MOOCs) are now available on numerous platforms, attracting thousands of learners eager to learn new skills in various areas of competencies. Improvement of the learner experience is an active research area in the field of Educational Data Mining with numerous studies focusing on solutions to increase the course completion. The success prediction falls within this research area.

Our work proposes an improvement of a success prediction task based on a transfer learning approach [1, 2]. This field aims to transfer the knowledge across domains. In our context, the domains are the online courses. Practically, our problematic is: *is there a common behavior of the learners among different courses, that can be exploited to learn a classification model for a prediction task on another course ?*

Our contribution consists to predict a learner's success on a specific course while learning a classification model from learners' traces of this course, plus learners' traces of other courses. A transfer learning method is applied to build the training datasets for each prediction task, based on a mapping approach and feature normalization respectively for the categorical and numerical attributes.

The motivation of this work is the common lack of training data, especially the data related to the success learner traces, due to the high rate of failure/abortion in the MOOC environment [3].

Success prediction [3] is still tackled in numerous works [4, 5] that study which features, models, and metrics to use. In the literature, classification methods rely on shallow methods (logistic regression, Adaboost, ...), neural networks, or sequence-based methods. Because this work is a sequel of our previous work [6], we decided here to only assess the best models obtained in our previous experiments.

To our knowledge, few works uses a transfer learning approach in the elearning domain. Boyer et al. [7] proposes a stopout prediction model based on transfer learning. The training set are built from different version of a same course (a version is defined as the same course taught at several moments all along the years). Ding et al. [8]...

In our proposal, our training sets are built from the current course, on which is carried out the prediction, plus different

courses obtained from the same elearning platform.

The validation of our proposal is carried out on two datasets, the first provided by OpenClassrooms, a major french company of online courses, and the second provided by the Open University, the Open University Learning Analytics dataset.

The remainder of this paper is organized as follows: section 2 presents the datasets and details how learner's traces of different courses are merged. Section 3 quickly depicts the features obtained from the literature and the classification models. Section 4 presents the results of our prediction task. Finally, we summarize our work and draw perspectives in section 5.

2. INPUT DATA

This section presents the Openclassrooms dataset (OCR) and the Open University Learning Analytics dataset (OULAD). We describe first the datasets and next the algorithm used to merge the learner's traces among different courses.

2.1 Datasets

Name	# users	# pass	# fail
OCR			
Java	7761	34 (0%)	7727 (100%)
XML	855	10 (1%)	845 (99%)
Ionic	960	46 (1%)	914 (99%)
Rubys	149	5 (3%)	144 (97%)
Node JS	2227	81 (4%)	2146 (96%)
Arduino	2487	115 (5%)	2372 (95%)
Bootstrap	8402	727 (8%)	7675 (92%)
Audace Entr.	225	26 (12%)	199 (88%)
JavaScript	8105	1803 (22%)	6302 (78%)
Gestion Projet	1808	666 (36%)	1142 (64%))
Twitter	817	328 (40%)	489 (60%)
Web	7947	3502 (44%)	4445 (56%)
OULAD			
CCC-2014B	1681	663 (39%)	1018 (60%)
DDD-2013B	1217	510 (41%)	707 (58%)
DDD-2014B	1117	479 (42%)	638 (57%)
CCC-2014J	2305	1014 (43%)	1291 (56%)
DDD-2013J	1772	829 (46%)	943 (53%)
FFF-2014B	1363	654 (47%)	709 (52%)
DDD-2014J	1651	792 (47%)	859 (52%)
FFF-2013B	1510	782 (51%)	728 (48%)
BBB-2013B	1538	803 (52%)	735 (47%)
FFF-2013J	2099	1095 (52%)	1004 (47%)
FFF-2014J	2127	1117 (52%)	1010 (47%)
BBB-2014B	1299	727 (55%)	572 (44%)
BBB-2013J	1883	1072 (56%)	811 (43%)
BBB-2014J	1925	1152 (59%)	773 (40%)
GGG-2014B	773	478 (61%)	295 (38%)
GGG-2014J	698	444 (63%)	254 (36%)
GGG-2013J	898	592 (65%)	306 (34%)
AAA-2014J	357	253 (70%)	104 (29%)
AAA-2013J	378	278 (73%)	100 (26%)

Table 1: Number of learners per group for each course of the OCR and OULAD datasets. For both datasets, the first column contains the name of the courses, the second column presents the total number of learners and the two last ones details the number of pass and fail. The courses are ordered by the proportion of the passing learners (the percentages are rounded to the nearest integer).

OpenClassrooms is a MOOC platform that provides courses in various areas, from art and culture to computer science. The Open University is an online university providing paid access to modules, where a module is a set of courses. Based on the properties proposed in [3], both the datasets are defined as massive, online, and heterogeneous. The differences are that the OCR dataset is open, without stakes, and asynchronous while the OULAD dataset is payable, with stakes, and synchronous.

The tables 1 presents the courses available for the OCR and OULAD datasets. The OCR dataset covers 12 courses while the

OULAD dataset contains 6 courses, with several sessions for each one (the name is defined as N-YM where N is the name, Y the year, and M the starting month for the course).

The format of both datasets is a clickstream. Practically, each learner's trace is defined as a sequence of events, where each event is one resource access. A resource can be either a document (section, chapter, document, ...) or an assignment. One important difference is the granularity of the time dimension: for OCR, each event is timestamped with a date and time (with a second precision) while the OULAD just provides the number of access per resource on each day.

Two groups are defined for our goal of success prediction: the passing group, defined as the set of learners that validates the course, and the failing group, defined as the set of learners that fails the course. The two rightmost columns of table 1 present the number of learners, pass and fail in each group. Due to the open context, the groups in the OCR dataset are unbalanced in favor of the fail learners, while the paid access of the Open University entails a more balanced dataset.

2.2 Transfer Learning

The clickstream datasets are composed of a sequence of events for each learner, where each event is defined with the followings attributes: learner id, resource id, resource type, timestamp, intensity (time spent on the resource), mark (if the resource type is an assignment).

To learn a classification model from traces of several courses, we rely on the transfer learning approach based on a mapping [1]. Our solution consists to build mappings of attributes between a *base* course and the other ones, called the *merge* courses: the values of each attribute of the *merge* courses are mapped to the values of the *base* course. Of course, the problem is that each course contains specific resources (specific documents and specific assignments), different numbers of resources, and relative duration spent on a specific resource.

Our merge process is defined as follows :

1. detect the *base* course: the course with the minimum number of documents is selected. This course is named the *base* course and the others are named the *merge* courses;
2. for each course in the *merge* set, map each document id to a document id of the *base* course. This mapping relies on the availability of an order for the documents. Each document of the *merge* course is mapped to its closer document of the *base* course based on their relative order;
3. for each course in the *merge* set, map each assignment id to an assignment id of the *base* course. This mapping relies on the availability of an order for the assignments. Each assignment of the *merge* course is mapped to its closer assignment of the *base* course based on their relative order;
4. the intensity (duration spent on each resource) and the marks of the *base* course are normalized. These attributes of the *merge* courses are also normalized with the standard variations obtained from the *base* course.

The choice of the *base* course with the minimum number of documents involves a straightforward mapping where each document is associated with another one (the domain of the output is smaller than the input one), but also involves that the more courses are merged, the more the training dataset is degraded.

3. PREDICTION TASK

In this section, we present first very shortly the features and then the classification methods applied for our success prediction task.

3.1 Features

As stated in [9], feature extraction is a critical task, possibly of even more importance than the choice of the learning algorithm. Our previous work [6] already assess a subset of the best features identified in [9, 10, 11, 12, 13, 14]. In this work, we extend our proposal to 40 features and adapt them for the OULAD datasets (our previous work [6] focuses only on the OCR dataset with 30 features).

Our features are regrouped in three different categories:

- activity features: related to the time passed on the features, the number of events, ...;
- inter-activity periods: related to the time between the access to the resource, the time before each evaluation, ...;
- assignment: the marks obtained on the exercises, time spent between each assignment, ...

Course	25%	50%	75%	100%
OCR - default				
Random Forest	0.92	0.94	0.97	1.0
Adaboost	0.85	0.93	0.96	1.0
Logistic regression	0.94	0.96	0.98	1.0
Neural network	0.87	0.88	0.96	0.99
Oulad - default				
Random forest	0.91	0.94	0.96	0.97
Adaboost	0.91	0.94	0.96	0.98
Logistic regression	0.92	0.95	0.97	0.98
Neural network	0.90	0.92	0.95	0.96

Table 2: AUC scores of the different models tested for the OCR and Oulad datasets.

More details on these features can be obtained in [6]. By providing the implementation of a high number of features, our objective is to continue the identification of the best features for a success prediction task. This identification task is not the point of this work.

3.2 Classification models

Based on our previous work [6], we decided to rely on the following classification models for our experiments: random forest, AdaBoost, logistic regression, and dense neural network. Except for the latter, these models provided the best results on the OCR dataset in our previous experiments. The neural network presented lower performance, but it is due, to our opinion, to a lack of training data. Because our proposal enables us to improve this drawback, we decided to keep it.

For the logistic regression model, the first step is the selection of features. We rely here on a wrapper method, applied on each course independently, with a forward selection to emphasize the best features. A subset of features is iteratively built, starting from an empty set and adding one by one the features that best improve our model accuracy. The process is stopped when the accuracy does not increase anymore.

Our prediction task is carried out on each course of the two datasets. For a specific course, two predictions are tested:

- the *default* experiment consists to learn a model only from the traces associated with the *base* course. The training set is composed of 80% of the data and the prediction task is evaluated on the remaining 20%;
- the *transfer* experiment consists to learn the model from 80% of the traces of the *base* course, plus all the traces of the other courses of the same dataset. The prediction task is evaluated on the remaining 20% traces of *base* course.

For each prediction on a course, a cross-validation 80% train - 20% test is carried out 10 times (80% + the traces of the other courses of the same dataset for the *transfer* experiment).

Finally, to assess our prediction at different time steps of the learning process, classification tasks are tested on truncated versions of the traces. Experiments provide results for 25%, 50%, 75%, and 100% of learner's traces length. 100% means that all the events of each learner are used to build the classification model. 25% means that only the first 25% of the events are kept in the traces of the learners.

4. EXPERIMENTS

Table 2 summarizes the *area under the curves* (AUC) scores for the *default* and the *transfer* experiments. The rows present the models tested for each dataset for both experiments. The two top horizontal blocks present the *default* experiment and the two bottom ones the *transfer* experiment. The columns represent the learner's trace length used for the prediction task. In the two last blocks, the colors red and blue indicate respectively an increase and a decrease between the *default* and the *transfer* experiments.

Compare to our experiments presented in [6], two changes were introduced: new features were added, so that the input of our model is different, and the prediction task of the Logistic Regression method was done on each course with their associated best features (and not the average best features on all the courses). So that a fair comparison with our previous experiments is not relevant.

On the *default* experiment, the logistic regression model outperforms the other models on both datasets. It is followed by the random forest and Adaboost models that present very similar performances (except for the 25% on OCR where Adaboost is

Course	25%			50%			75%			100%		
OCR-transfer	AUC	Accuracy	#	AUC	Accuracy	#	AUC	Accuracy	#	AUC	Accuracy	#
AdaBoost	0.11	2%	12	0.03	3%	6	0.02	2%	6	0.0	1%	1
RandomForest	0.07	6%	7	0.01	8%	1	0.01	2%	1	0.01	3%	1
Perceptron	0.1	6%	11	0.13	5%	9	0.02	4%	5	0.01	1%	1
LogisticRegression	0.08	3%	10	0.0	0%	0	0.02	12%	2	0.0	0%	0
OULAD-transfer	AUC	Accuracy	#	AUC	Accuracy	#	AUC	Accuracy	#	AUC	Accuracy	#
AdaBoost	0.04	7%	4	0.02	6%	3	0.03	0%	5	0.0	2%	4
RandomForest	0.0	0%	0	0.02	7%	1	0.04	0%	6	0.0	0%	0
Perceptron	0.05	7%	5	0.04	6%	5	0.05	0%	6	0.03	4%	4
LogisticRegression	0.03	7%	3	0.02	6%	1	0.03	0%	5	0.0	3%	2

Table 3: AUC scores of the different models tested for the OCR and Oulad datasets.

lower). The Neural Network model is the weakest on the 25% and 50% tests but catches up with the other models on the 75% and 100% tests.

On the *transfer* experiment, the best models are now the neural network model and the Adaboost model respectively for the OCR and OULAD datasets. For the OCR dataset, we obtain a balanced number of improvements and decays. Nevertheless, the decay scores are very close to the ones obtained in the *default* experiment (-0.1 or -02) while the improved ones present a real gap, particularly for the neural network model. For the OULAD dataset, we observe a sharp improvement of the scores on all the models except the logistic regression one.

Our proposal to learn a model from different courses seems conclusive: the neural network model and the Adaboost model, respectively on the OCR and OULAD datasets, succeed to outperform the scores obtained in the *default* experiment. These models benefit from the enrichment of the training dataset since their scores on the different trace lengths are all improved on both datasets. The random forest model only presents a sharp improvement on the OULAD dataset and the logistic regression model does not benefit at all from the enrichment of the training dataset.

The improvement is more emphasized on the OULAD dataset than the OCR one. This difference is nevertheless difficult to explain since these datasets present different properties (for instance, different time granularities of the events) and address different profiles of learners.

5. CONCLUSION

The goal of our work is to assess a new approach for a success prediction task in the context of Massive Online Open Courses. We hypothesize that a learner presents a common underlying behavior when they are engaged on different online courses. Our proposal is then to learn a model with a training set obtained by merging learner's traces of different courses. Our solution is assessed on the OCR and OULAD datasets with 4 different models.

The results obtained are encouraging since we succeeded to improve our success prediction task. We conclude that we can exploit the common behavior of the learners among different courses to train a classification model. A short-term perspective work is now to confirm our proposal on other MOOC datasets.

6. REFERENCES

[1] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[2] Z. Fuzhen, Q. Zhiyuan, D. Keyu, D. Dongbo, Z. Yongchun, Z. Hengshu, X. Hui, and H. Qing, "A comprehensive survey on transfer learning," *Computing Research Repository (CoRR)*, 2019.

[3] J. Gardner and C. Brooks, "Student success prediction in moocs," *User Modeling and User-Adapted Interaction*, vol. 28, pp. 127–203, June 2018.

[4] D. Sanyal, N. Bosch, and L. Paquette, "Feature selection metrics: Similarities, differences, and characteristics of the selected models," in *Proceedings of the 13th International Conference on Educational Data Mining, EDM*, pp. 212–223, 2020.

[5] R. Yu, Q. Li, C. Fischer, S. Doroudi, and D. Xu, "Towards accurate and fair prediction of college success: Evaluating different sources of student data," in *Proceedings of the 12th International Conference on Educational Data Mining, EDM*, pp. 292–301, Jul. 2020.

[6] A. Pigeau, O. Aubert, and Y. Prié, "Success prediction in moocs: A case study," in *Proceedings of the 12th International Conference on Educational Data Mining, EDM* (M. C. Desmarais, C. F. Lynch, A. Merceron, and R. Nkambou, eds.),

pp. 390–395, Jul. 2019.

[7] S. Boyer, , and K. Veeramachaneni, "Transfer learning for predictive models in massive open online courses," in *International conference on artificial intelligence in education*, pp. 54–63, Springer, 2015.

[8] M. Ding, Y. Wang, E. Hemberg, and U. O'Reilly, "Transfer learning using representation learning in massive open online courses," in *Proceedings of the 9th international conference on learning analytics & knowledge*, pp. 145–154, 2019.

[9] C. Taylor, K. Veeramachaneni, and U. O'Reilly, "Likely to stop? predicting stopout in massive open online courses," *CoRR*, vol. abs/1408.3382, 2014.

[10] Z. Ren, H. Rangwala, and A. Johri, "Predicting performance on MOOC assessments using multi-regression models," in *Proceedings of the 9th International Conference on Educational Data Mining, EDM*, pp. 484–489, Jun. 2016.

[11] J. Whitehill, J. J. Williams, G. Lopez, C. A. Coleman, and J. Reich, "Beyond prediction: First steps toward automatic intervention in mooc student stopout," in *Proceedings of the 8th International Conference on Educational Data Mining, EDM*, pp. 171–196, 2015.

[12] Y. Chen, Q. Chen, M. Zhao, S. Boyer, K. Veeramachaneni, and H. Qu, "Dropoutseer: Visualizing learning patterns in massive open online courses for dropout reasoning and prediction," in *Conference on Visual Analytics Science and Technology, VAST*, pp. 111–120, Oct. 2016.

[13] S. Nagrecha, J. Z. Dillon, and N. V. Chawla, "Mooc dropout prediction: Lessons learned from making pipelines interpretable," in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 351–359, Apr. 2017.

[14] M. S. Boroujeni, K. Sharma, L. Kidzinski, L. Lucignano, and P. Dillenbourg, "How to quantify student's regularity?," *Proceedings of the 11th European Conference on Technology Enhanced Learning*, pp. 15. 277–291, 2016.