Discovering Learning Behavior Patterns to Predict Dropout in MOOC

Bowei Hong, Zhiqiang Wei, Yongquan Yang*
Department of Computer Science and Technology,
Ocean University of China
Qingdao, China
e-mail: i@hongbowei.com, weizhiqiang@ouc.edu.cn, i@yangyongquan.com

***Abstract*—High dropout rate of MOOC is criticized while a dramatically increasing number of learners are appealed to these online learning platforms. Various works have been done on analysis and prediction of dropout. Machine learning techniques are widely applied to this field. However, a single classifier may not always perform reliable for predictions. In this work, we study dropout prediction for MOOC. A technique is proposed to predict dropouts using learning activity information of learners. We applied a two-layer cascading classifier with a combination of three different machine learning classifiers — Random Forest (RF), Support Vector Machine (SVM), and MultiNomial Logistic Regression (MLR) for prediction. Experimental results indicate that the technique is promising in predicting dropouts with achieving 97% precision.**

***Keywords*—MOOC; machine learning; dropout prediction; ensemble of classifiers.**

I. INTRODUCTION

In a last few years, an increasing number of participants were attracted to Massive Open Online Course (MOOC) [1]–[4]. Taking advantage of the internet, individual learners are able to acquire knowledge in a flexible manner; in addition, MOOC also benefits institutions in providing more students with their educational services while reducing the overall expenditure[5]. However, despite the fact that MOOC appeals to a large amounts of learners, it also experiences a higher dropout rate (as high as 80-95%) with a large number of enrollments [6], [7].

Unlike in traditional classes, teachers cannot make proper adjustments according to in-class reactions of students in MOOC. Therefore, it would be useful to provide information for teachers about how likely a student would drop out during the course. With this information, teachers could take actions to encourage and maintain learning activities of students who are believed to be likely to drop out during the course[8]–[10].

The majority of MOOC platforms track interactions of learners among video lectures, assessments, and social networking[11], which can then be analyzed[12]–[14]. Generally, diverse decision rules or machine learning models are applied to various attributes of student learning activities for dropout predictions[13]–[17]. However, making a good decision on selection of learning attributes and machine learning models is vital.

In this paper, we present a dropout prediction technique to identify at-risk students based on their prior activities (e.g. watching course videos, working on course assignments, accessing the course wiki etc.). A student will be defined as dropout from the enrolled course if he or she leaves no records for the course in the log. Our technique takes student learning behavior logs in a period of time (e.g. one month) as inputs, then will classify students into two categories (whether drop out or not in next ten days).

The dataset we used to establish and test our prediction model is collected from XuetangX, one of the largest MOOC platforms in China. Our work can be applied for other MOOC platforms adopting Open Edx framework as XuetangX does. We applied a two-layer cascading classifier with a combination of three different machine learning classifiers — Random Forest (RF)[18], Support Vector Machine (SVM)[19], and MultiNomial Logistic Regression (MLR)[20] to achieve better prediction performance. In terms of evaluation metrics, accuracy and area under the ROC are adopted in reporting the performance of three individual classifiers as well as their combination.

The rest of this paper is organized as follows: Section 2 provides a comprehensive literature review of previous works as context for our proposed technique. Details of our dropout prediction framework are described in Section 3, including workflow of proposed system, data preprocessing, feature extraction and normalization. Section 4 reports metrics for system performance evaluation and experimental results. A conclusion is given in Section 5 about the main findings and future works.

II. RELATED WORK

Various works have investigated on the field of dropout prediction in online education. The majority of methodology these studies applied is rule-based and machine learning-based respectively.

Rule-based approaches conduct statistical analysis on extracted features, and then predict dropout via predefined rules. Ramesh et al. [21] applied probabilistic soft logic to identify two types of engagement, which is then used as a latent feature to help predict dropout. Taking advantage of standard social network analytic techniques, Yang et al. [22] explore factors related to student dropout behavior with a survival model they developed. In [17], four behavior features and one absent feature are extracted from interaction data,

comparing with predefined thresholds, decisions are made and combined for dropout prediction.

In [23], the experimental results show that simple classifiers (decision trees in this case) give a useful result with accuracies between 75 and 80% which is hard to beat with other more sophisticated models. This work suggests that student behaviors can be predicted from quantified information, which paves the way for significant machine learning work in this field.

In general, approaches based on various machine learning algorithms make predictions with the classifiers trained from history data. In [14], Kloft et al. extract features from weekly history of student clickstream data and use SVM to predict dropout. Hu et al. [24] adopted time-dependent variables in their early warning system and they trained three single classifiers — Decision Tree (C4.5), Regression Tree (CART), and Logistic Regression (LGR) for dropout prediction. In [13], three kinds of classification features are observed with four single supervised classification models — SVM, Logistics Regression (LR), Random Forest and Gradient Boosting Decision Tree (GBDT).

In [25], applying several modeling techniques on time-dependent features, Hidden Markov Model (HMM) performs best in prediction with an ROC AUC (Receiver Operating Characteristic Area Under the Curve score) of 0.710. Taylor et al. [26] applied logistic regression on crowd-sourced engineering of over 25 predictive features, achieving an AUC 0.88 in average.

However, there are certain weaknesses with the existing techniques. For rule-based techniques, the definition of rules and the selection of acceptance/rejection thresholds are vital to interpret features. Machine-learning-based techniques tend to yield the best performance by choosing one classifiers from multiple classifiers[13], [25], [26]. Though this strategy works well, it has been observed that the sets of patterns misclassified by different classifiers would not necessarily overlap[27]. Therefore, a single classifier may not always perform reliably for predictions. For both rule-based and machine learning based techniques, it is critical to select reliable features as well.

In the proposed system, instead of a single classifier, we applied a two-layer cascading classifier with a combination of three different machine learning classifiers to achieve better performance. Thirteen features are used for dropout prediction.

III. Proposed System for Dropout Prediction

A. Overview

Fig. 1. shows the system flow of our processing framework. The database and corresponding ground truth labels are prepared following instruction introduced in experiment section. Each sample in the database extracted from three dataset: course content, student enrollment, learning access log, which are combined together in the preprocess procedure.

The diagram of the cascading classifier is described in Fig. 2. We applied a two-layer cascading classifier with a combination of three different machine learning classifiers. These classifiers are namely Random Forest (RF)[18], Support Vector Machine (SVM)[19], and MultiNomial Logistic Regression (MLR)[20]. During the training phase, each classifier is presented with a set of sample data pairs (X, Y), where X, Y represents the input feature and the corresponding output respectively. In this study, Y will receive a real-valued probability of dropout in the first layer. And in the second layer, we use outputs from the first layer as the inputs for the second layer; Y can receive one of the following values: 1 for a dropout event and 0 for continuing study.

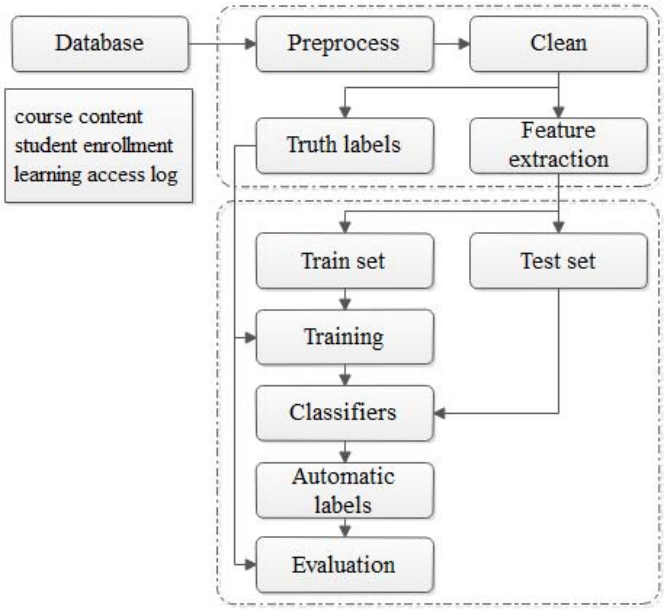

Fig. 1. Flowchart of Our System

Suppose, the feature number is n, f_{1n} represents nth feature in the first layer. We use three different classifiers in layer one, they adjusts their internal parameters to infer the mapping implied by the training data provided respectively. In the second layer, the estimate probabilities (f_{21} for RF, f_{22} for SVM, f_{23} for MLR) calculated from three trained classifier in first layer are combined as inputs for dropout prediction. Final prediction is made by the classifier trained in second layer. We applied three classifiers to complete this work separately, and their performances are given in experimental section. The following sub-section gives the attributes used for dropout prediction in the proposed technique.

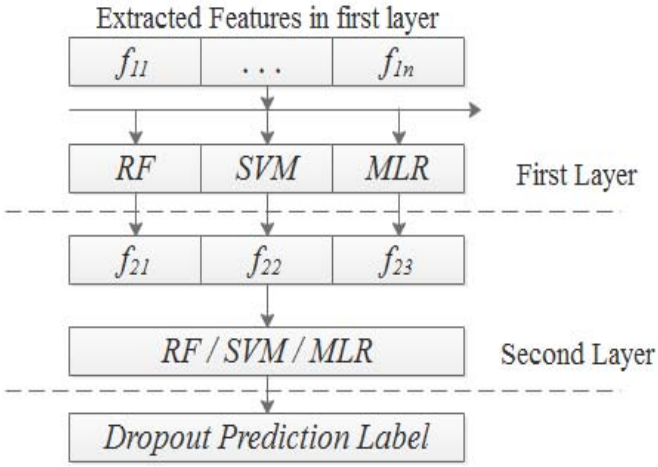

Fig. 2 Diagram of the cascading classifier

B. Feature Description

From datasets of course content, student enrollment, and learning access logs, we obtained thirteen features which contains information of learning behavior for one student in a course. Actually, with the study progress records increasing over time, the accuracy of classification will increase because classifier could gain more knowledge. The details of feature description were characterized in Table I.

TABLE I. DETAIL DESCRIPTION OF LEARNING BEHAVIOR FEATURES

	Name		Name
f_1	ActiveNumber	f_8	NavigateVisit
f_2	ActiveRate	f_9	PageCloseVisit
f_3	ProblemVisit	f_{10}	CorseNumber
f_4	VideoVisit	f_{11}	UserNumber
f_5	AccessVisit	f_{12}	UserDropoutRate
f_6	WikiVisit	f_{13}	ClassDropoutRate
f_7	DiscusssionVisit		

All of the features are described as follow:

- **ActiveNumber(f_1)**- Number of user activity.
- **ActiveRate(f_2)**- Rate of user activity, including the number of active days, number of active per day and percent of average number of active days.
- **{X}Visit(f_3- f_9)**- The frequency and percent of user visit the X module.
- **CorseNumber(f_{10})**- The number of course selected by user.
- **UserNumber(f_{11})**- The number of all users of the course selected by user.
- **UserDropoutRate(f_{12})**- The rate of user dropout, inlcuding the number of courses the user has droped out, the rate of user dropout and the score of dropout class.
- **ClassDropoutRate(f_{13})**- The rate of class dropout, inlcuding the number of dropout users, the rate of user dropout and the score of dropout users.

After extracting learning behavior features from the dataset, we applied Min-Max Normalization by (1) for feature normalization.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Where $x = (x_1, ..., x_n)$ and z_i is i^{th} normalized sample.

IV. EXPERIMENT AND ANALYSIS

A. Dataset

The dataset we used to establish and test our prediction model is provided by XuetangX, a MOOC platform based on Open Edx framework in China. This dataset contains information of course, student enrollment, and student learning access logs with a time span of 40 days (one month plus 10 days). In this paper, a student will be defined as dropout from the enrolled course if he or she leaves no records for the course in the log for 10 consecutive days. Therefore, learning access logs of the first 30 days would be used for feature extraction, model training, and performance evaluation while we obtain the ground truth from the rest ten-day logs for our dataset. As defined in Section 2, the ground truth receives following values: 1 for a dropout event and 0 for continuing study. After feature extraction, we have got a dataset contains 120542 samples from 39 courses. All user data has been privacy cleaned. Using 80% of the available data, we trained classifiers, and the remaining 20% of the data was used to test the performance of them. Table II shows sample distribution of training set for each classification category.

TABLE II. SAMPLE DISTRIBUTION PER CLASSIFICATION CATEGORY

	Sample per Classification Category	
	Dropout	***Continuing Study***
Sample Number	76470	20059
Percentage of Sample	79.22	20.78

As shown in Table II, the sample distribution is unbalanced. Thus, to achieve better classifier performance, we applied synthetic minority oversampling technique (SMOTE)[28] on the training set. The sample distribution after applying SMOTE is given in Table III.

TABLE III. SAMPLE DISTRIBUTION AFTER APPLYING SMOTE

	Sample per Classification Category	
	Dropout	***Continuing Study***
Sample Number	76470	60177
Percentage of Sample	55.96	44.03

Moreover, we used five primary metrics in reporting the performance of the proposed system: precision, recall, F1-score, accuracy and AUC (area under the ROC). Our metrics may be interpreted as follow:

- **Precision** – proportion of dropout events that are correctly classified
- **Recall** – proportion of dropout events that are correctly classified versus the total number of actual dropout events
- **F1-score** –a single scalar value combined by P and R representing overall performance of the proposed classifier
- **AUC** – A single scalar value representing the overall performance of the classifier.
- **Accuracy** – proportion of all events that are correctly classified.

Even one percent improvement in some of these metrics can be substantial. As a reference, for dropout prediction in KDD Cup 2015 there was only 2% improvement in AUC from the 182nd to the best score on the leaderboard.

B. Result

Table IV lists the classification performance of six classifiers using our features on testing set. Among them, SVM, RF, MLR are single classifiers; C-SVM, C-RF, C-MLR are proposed cascading classifiers with different classifier in second layer. All classifiers get decent performance, which indicates that feature selection is successful. Comparing to single classifiers, cascading classifiers all get improvements in terms of precision, F1-score, and AUC, but a decrease of recall, which means that the proposed approach is accurate in correctly identifying dropouts but not as accurate as single classifiers in continuing learner misclassifications. In particular, we note that C-RF appears to work quite well achieving the highest score of all metrics respectively.

TABLE IV. PERFORMANCE OF EACH CLASSIFIERS

	SVM	RF	MLR	C-SVM	C-RF	C-MLR
Precision	0.877	0.885	0.880	0.957	0.979	0.971
Recall	0.959	0.952	0.955	0.866	0.889	0.865
F1-score	0.916	0.917	0.916	0.910	0.932	0.915
AUC	0.795	0.852	0.855	0.909	0.932	0.916
Accuracy	0.861	0.865	0.861	0.904	0.927	0.910

A compare between cascading classifier and its corresponding single classifier are given in Table V. Though recall of C-RF descents to 0.889033 compare to RF, it improves precision 10.58%, F1-score 1.5%, 9.4% AUC than RF respectively. Meanwhile, C-MLR improves precision 10.33% than MLR, but worsens recall as high as 9.34%. Moreover, it is interesting that, compare to SVM, recall of C-SVM descends to 0.866831 while making AUC 14.33% improved with only a 4.93% increase of accuracy.

TABLE V. A COMPARE BETWEEN CASCADING CLASSIFIER AND ITS CORRESPONDING SINGLE CLASSIFIER

	C-SVM vs SVM	C-RF vs RF	C-MLR vs MLR
Precision	9.14%	10.58%	10.33%
Recall	-9.63%	-6.67%	-9.34%
F1-score	-0.71%	1.54%	-0.08%
AUC	14.33%	9.46%	7.23%
Accuracy	4.93%	7.22%	5.66%

It is obvious that the proposed model suffers a decrease in terms of recall. However, in practical applications, we are more concerned about finding potential dropouts and giving support to them rather than avoiding misclassifying learners who will not leave the course as dropouts. Moreover, the misclassification will not drive learners giving up courses, namely increase dropouts, though it might consume some additional energy of teachers. Therefore, we believe that the proposed technique is effective and feasible in the practical application.

V. CONCLUSION

This paper presented a dropout prediction technique for MOOC platform. The technique is designed to use information of learning activities from the MOOC platform. We applied a two-layer cascading classifier with a combination of three different machine learning classifiers — Random Forest (RF), Support Vector Machine (SVM), and MultiNomial Logistic Regression (MLR) for dropout prediction. Experimental results show that the proposed technique do achieve better performance in terms of precision, F1-score, AUC and accuracy. In addition, compare to individual techniques, the proposed technique suffers a descent to recall, which is acceptable. In order to obtain better performance without sacrificing recall, we will keep improving and optimizing the proposed technique in future research.

ACKNOWLEDGMENT

We acknowledge Dr. Lei Huang, Dr. Zhen Li and Dr. Hao Liu for their instructive suggestions and valuable advice. This work is supported by the CERNET Innovation Project under Grand No.NGII20150201; the Shandong Province key research and development plan under Grand No.2016ZDJS09A01; the Shandong Province Science development plan under Grand No.2014GGX101005.

REFERENCES

[1] L. Harasim, “Shift happens: online education as a new paradigm in learning,” Internet High. Educ., vol. 3, no. 1–2, pp. 41–61, 2000.

[2] K. Swan, “Building Learning Communities in Online Courses: the importance of interaction,” Educ. Commun. Inf., vol. 2, no. 1, pp. 23–49, 2002.

[3] T. Väljataga, H. Põldoja, and M. Laanpere, “Open online courses: Responding to design challenges,” Stanford Univ. H-STAR Institute, USA; to Assoc. Profr. Jukka M. Laitam{ä}ki, from New York Univ. USA, to Profr. Yngve Troye Nord. from Lillehammer Univ., p. 68, 2011.

[4] C. Tekin and M. van der Schaar, “eTutor: Online Learning for Personalized Education,” arXiv Prepr. arXiv1410.3617, 2014.

[5] M. A. A. Dewan, F. Lin, D. Wen, and Kinshuk, “Predicting Dropout-Prone Students in E-Learning Education System,” in Uic-Atc-Scalcom-Cbdcom-Iop, 2015, pp. 1735–1740.

[6] C. G. Brinton and M. Chiang, “Social learning networks: A brief survey,” 2014 48th Annu. Conf. Inf. Sci. Syst., pp. 1–6, 2014.

[7] L. Yuan and S. Powell, “MOOCs and Open Education: Implications for Higher Education,” Cent. Educ. Technol. Interoperability Stand., vol. 4, no. 4, pp. 206–207, 2013.

[8] C. E. Hmelosilver and H. S. Barrows, “Goals and Strategies of a Problem-Based Learning Facilitator.,” vol. 1, no. 1, pp. 21–39, 2006.

[9] C. E. Hmelo-Silver, C. P. Rosé, and J. Levy, “Fostering a learning community in MOOCs,” 2014.

[10] D. Koller and A. Ng, “The Online Revolution: Education at Scale,” L Educ., 2012.

[11] C. G. Brinton, R. Rill, S. Ha, M. Chiang, R. Smith, and W. Ju, “Individualization for Education at Scale: MIIC Design and Preliminary Evaluation,” Learn. Technol. IEEE Trans., vol. 8, no. 1, pp. 136–148, 2015.

[12] C. Zhao, J. Yang, J. Liang, C. Li, L. Hduqlqj, H. Wr, P. Kdrff, X. Hgx, F. Q. Dqjmldq, and V. X. Hgx, “Discover learning behavior patterns to predict certification,” in 2016 11th International Conference on Computer Science Education (ICCSE), 2016, no. Iccse, pp. 69–73.

[13] J. Liang, C. Li, and L. Zheng, “Machine learning application in MOOCs: Dropout prediction,” in 2016 11th International Conference on Computer Science Education (ICCSE), 2016, no. Iccse, pp. 52–57.

[14] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart, “Predicting MOOC Dropout over Weeks Using Machine Learning Methods,” in EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in Moocs, 2014, pp. 60–65.

[15] C. Rose, G. Siemens, and C. P. Rosé, “Shared Task on Prediction of Dropout Over Time in Massively Open Online Courses,” in EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in Moocs, 2014, pp. 39–41.

[16] A. Cohen and U. Shimony, “Dropout prediction in a massive open online course using learning analytics,” in E-Learn, 2016.

[17] S. Halawa, D. Greene, and J. Mitchell, “Dropout prediction in moocs using learner activity features,” 2014.

[18] A. Liaw and M. Wiener, “Classification and Regression by randomForest,” R News, vol. 2, no. 3, pp. 18–22, 2002.

[19] C.-C. Chang and C.-J. Lin, “{LIBSVM}: A library for support vector machines,” ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, p. 27:1--27:27, 2011.

[20] A. J. Dobson, “An introduction to generalized linear models,” Technometrics, vol. 98, no. 464, pp. 1086–1087, 2001.

[21] A. Ramesh, D. Goldwasser, B. Huang, H. Daum, and L. Getoor, “Modeling Learner Engagement in MOOCs using Probabilistic Soft Logic,” 2013.

[22] D. Yang, T. Sinha, D. Adamson, and C. P. Rose, “‘Turn on, Tune in, Drop out’: Anticipating student dropouts in Massive Open Online Courses,” in NIPS Workshop on Data Driven Education, 2013.

[23] J. M. . M. Vleeshouwers, G. W. . Dekker, M. . Pechenizkiy, and J. M. . M. Vleeshouwers, “Predicting students drop out : a case study,” Int. Work. Gr. Educ. Data Min., no. March 2017, pp. 41–50, 2009.

[24] Y. H. Hu, C. L. Lo, and S. P. Shih, “Developing early warning systems to predict students’ online learning performance,” Comput. Human Behav., vol. 36, pp. 469–478, 2014.

[25] Balakrishnan and G. Eecs, “Predicting Student Retention in Massive Open Online Courses using Hidden Markov Models,” 2013.

[26] C. Taylor, K. Veeramachaneni, and U. M. O’Reilly, “Likely to stop? Predicting Stopout in Massive Open Online Courses,” Comput. Sci., 2014.

[27] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, “On combining classifiers,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 3, pp. 226–239, 1998.

[28] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, no. 1, pp. 321–357, 2011.

[29] F. M. Harper, D. Moy, and J. A. Konstan, “Facts or Friends?: Distinguishing Informational and Conversational Questions in Social Q&A Sites,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2009, pp. 759–768.