MOOC Dropout Prediction Based on Multi-view Learning

Feifan Li¹*, Zhengang Wei¹*
¹Department of Computer Science and Technology, Ocean University of China, Qingdao, Shandong, 266000, China
*E-mail: lifeifan@stu.ouc.edu.cn
*Corresponding author’s e-mail: wzgwzq@ouc.edu.cn

Abstract. Massive Open Online Courses (MOOCs) provides a promising way to support education for all. Nonetheless, one central challenge is the remarkably high dropout rate, with completion rates for MOOC recently reported to be below 5%, and high dropout rates limiting their effectiveness. Building on the analysis of dropout as closely related to user and course interaction patterns, this paper presents an end-to-end multi-view interactive framework (EMIF) to predict user dropout in MOOC. We extract learning activity information, user enrolment information, and course knowledge information through a set of 1D-CNNs, obtain mutual information of these and fuse the above information-view loss through a self-weighing method. Experimental evaluation of our solution on the KDDCUP 2015 and MOOCCube datasets showed better results (AUC & F1 scores). We have a better performance compared to several existing methods, without using different architectures for different datasets.

1. Introduction
Massive Open Online Courses (MOOC) has gained momentum in recent years as online learning course. It delivers educational resources in the form of videos, blogs, and documents to everyone via the Internet, making learning more accessible and more convenient, thus having a significant impact on higher education. Following the emergence of online learning sites, e.g., Udacity, Coursera and edX, MOOC has gained more and more global attention [14]. A recent survey conducted by Coursera shows that MOOC is really beneficial to learners who complete them, with 61% of respondents reporting educational benefits of MOOC, while 72% reported career benefits [1]. However, the high dropout rate has been a fundamental and long-standing issue in the development of MOOC [6]. Indeed, its average course completion rate is only 5% for the edX platform [8]. In the case of themselves, this occurs because: (1) The cost of online learning is low on account of the fact that most courses on MOOC platforms are free [7]. (2) Usually without a supervisor to monitor the learners and without penalties for dropping a course.

The effective and accurate dropout prediction will be of great help in the course management of MOOC, which can be taken to enhance learners' willingness to learn by corresponding measures. There is a wealth of learning behaviour information stored on the MOOC platform, e.g., assignment submissions, student clickstream data, grade information, forum discussion participation and more [5], [16]. It is common for researchers to analyse and capture the key factors affecting student learning outcomes from these learning activity data to further use the data in predicting dropout [10], [13]. Wang et al. used RNNs models based on CNNs to solve the dropout prediction problem, where CNNs achieve automatic feature extraction and RNNs take into account the influence of different temporal factors on dropout, and the method achieved good recognition results [12]. Haiyang L et al. presented an approach
to time series classification by constructing data based on learner behaviour as well as on activities in a number of online distance learning modules [4]. However, existing researches have generally only predicted dropout by features of learners' study habits, unable to exploit the relationship with information of the learners themselves and the MOOC course, all of which makes the model's effectiveness unstable and underperforming. Feng et al. have shown that user information and course attributes have an impact on dropout prediction, and Fig. 1 shows that [3].

Figure 1. Age is an essential factor, the tendency for young people to drop out; Gender also has an impact with males on the whole tending to drop out of non-science courses and female users likely to drop out of science courses; And educational background is also an important factor [3].

In this paper, we present an end-to-end multi-view interactive framework (EMIF) to predict user dropout in MOOC. In EMIF, we extract learning activity information, user enrolment information and course knowledge information by a set of 1D-CNNs, using multi-view learning to obtain mutual information about this information and combining the above information-view loss by a self-weighing fusion method. We evaluate the proposed EMIF on the KDDCUP and the MOOCCube. For the first dataset KDDCUP 2015 is used, and the second dataset is larger and extracted from the XuetangX system by [17]. The experiment conducted on both datasets shows that the method has better performance compared to several existing techniques.

2. Problem Statement
To convey this more accurately, we will first introduce the subsequent definitions.

(i) Enrolment Information: To a given set of unique enrolments $E = \{e_n\}_{n=1}^N$, in which $N$ is an aggregate number of unique enrolments. Let $U$ denote the set of users and $C$ the set of courses. For a definite $e_i$, consists of a tuple $\{u, c\}$, where $u \in U$ and, $c \in C$.

(ii) Learning habits: We extract a multi-dimensional vector $H(u, c)$ related to learning habits for $e$ through user learning activity logs, in which each element $h_i(u, c) \in H(u, c)$ is a continuous features value of habits associated with users in learning activities, i.e., video activities, forum activities, assignment activities, and session activities.

(iii) Learner Information: We extract statistical information about the user, i.e., gender, age, location, education level to represent the learner information $I_l(u, c)$, with category information (e.g., education) represented by 1-hot vector and continuous information (e.g., age) represented as the value itself.

(iv) Course Information: For course information, the attribute common to both datasets is the course category, and for MOOCCube we add the course concept attribute, jointly denoted as $I_c(u, c)$ which embedded by word2vector.

With these definitions, the dropout prediction issue is defined as follows, given the learning activity of user $H(u, c)$ on course $c$ during the historical record cycle and user information $I_l(u, c)$ and course information $I_c(u, c)$, predict whether user $u$ would have dropped out of $c$ during the forecast. More formally, we define the dropout prediction as $Y(u, c) \in \{0, 1\}$. The issue can be expressed as a function,
\[
f: (H(u, c), I_l(u, c), I_c(u, c)) \rightarrow Y(u,c)
\]  

3. Framework and Methodologies
In this section, we firstly introduce the framework of EMIF. Then we will explain in detail how to build a multi-view model with an adaptive-weighting loss fusion strategy. The model structure is shown in Fig. 2.

3.1. The View Feature Extraction
The view feature extraction consists of these steps: embedding and feature extraction. For each instance \( x \) we collect the feature vectors \( \{x_v\}_{v=1}^V \) from \( V \) views to ensure diversity and complementarity of MOOC-learner information, i.e., learning habits \( x_1 \), learner information \( x_2 \), course information \( x_3 \). Then each \( x_v \) is transformed into a dense vector by means of an embedding layer. Since \( x_1 \) is such a continuous variable, the corresponding embedding vector is obtained by simply multiplying it with a vector of parameters \( a \in \mathbb{R}^{d_e} \),

\[
e = x_1 \cdot a
\]  

We use \( E_x^1 \in \mathbb{R}^{d_x \times d_e} \) to denote the embedding matrix of \( x_1 \). For the embedding of \( x_2 \), the processing is similar to that above, except that 1-hot encoding is used for category information, which we represent using \( E_x^2 \) for clarity of exposition. In the \( x_3 \) existence of knowledge concepts (e.g., “c++,” “binary tree”), which comprises a wealth of semantic information. We use Word2vector \( E_x^3 \) [9] to generate the word embedding. Note that, for the KDDCUP, an embedding matrix can be used. Define a set of one-dimensional convolutional neural networks \( \{f_v\}_{v=1}^V \), where \( f_v \) captures high-level specific information about the \( k \)-th view and converts \( E_x^v \) to \( y_v \), from \( \mathbb{R}^{d_x \times d_e} \) into \( \mathbb{R}^d \), that is,

\[
y_v = \sigma(W_{\text{conv}} \phi(E_x^v) + b_{\text{conv}})
\]  

where \( \phi(E) \) is flattening matrix \( E \) to a vector, \( \sigma \) denotes active function and \( W_{\text{conv}} \in \mathbb{R}^{d_f \times d_e} \) is the convolution kernel, \( b_{\text{conv}} \in \mathbb{R}^{d_f} \) is the bias term. \( \{f_v\}_{v=1}^V \) are concurrently trained through a minimisation of the final loss, in which the parameters of each \( f_v \) are learned in parallel.
3.2. Interaction Information Calculation

Interaction information is often used to explore potentially consistent information about the same instance, and to establish relationships between views, by calculating the cross-correlation matrix. Let's define a set $C_v$ that includes different view pairs about the $v$-th view, $C_v = \{(v, \overline{v})\}_{v \neq \overline{v}}$. We then compute the cross-correlation matrix among each element of $\tilde{y}_v$ and each element of $C_v$, which results in a set of interactive maps, denoted as $\tilde{C}_{ccm}^{v}$. Then we compressed $\tilde{C}_{ccm}^{v}$ using a compression network $\phi$ formed by the fully connected layer in order to learn the deep interaction information $\tilde{y}_v^*$ from $\mathbb{R}^{d \times d}$ to $\mathbb{R}^d$, and concatenate it with $y_v$ in the appropriate proportions,

$$y_v = [\text{vec}(\tilde{y}_v^*), y_v]$$ (4)

3.3. Loss Fusion Strategy

An adaptive weighted loss fusion strategy is used to allow multiple neural networks, which make joint decisions, to achieve multi-view classification [15], and can be described as follows,

$$\min_{\alpha} \sum_{v=1}^{V} \alpha_v^T L_v(\psi(y_v), \text{label}) \text{ s.t. } \alpha^T 1 = 1, \alpha \geq 0$$ (5)

where $\alpha \in \mathbb{R}^V$, and define the process of generating logits and the softmax layer as the $\psi$. The $\alpha$ is a value that is continuously updated and optimized, and the updating equation as,

$$\alpha_v = \frac{1}{\sum_{m=1}^{V} L_v^{1-\gamma}}$$ (6)

The $\gamma$ is the exponential of the power of the weight $\alpha_v$ of the $v$-th view, which allows for flexible adjustment to the weight distribution for the different views and avoids trivial solutions.

4. Experiments

We conducted various experiments on two datasets, KDDCUP and MOOCCube, to assess the effectiveness of EMIF. On the MOOCCube dataset, there are 55,203 users and 706 courses. The student's registration behaviour data mainly occurred between 2016-2019, and in order to filter noisy data, we use a part of data, 12,486 students, for experiment [11]. Another dataset is a subset of the dataset previously used for KDDCUP 2015. In the original dataset there were 39 courses with 79,186 users, with the most engaged course involving 12,004 users. We used the portion of these most users. For each pair in both datasets, a label is assigned in which 1 indicates dropout and 0 indicates no dropout. The information for each user is made up of a series of user behaviours, as detailed in Table 1.

| Time          | Source | Event | Object                         |
|--------------|--------|-------|--------------------------------|
| 2014-08-16 T08:26:27 | server | navigate | 2Z6eSRzdq…… |
| 2014-08-16 T08:26:37 | server | access | 3T6XwoiM……       |

With the MOOCCube dataset, a 35-day history period is set and a 10-day prediction period is set. With the KDDCUP dataset, the competition organisers set 30 and 10 days respectively. Please note that the learner information is not used, as it is not provided in the dataset. We aim to predict these labels and see if our model can outperform in the prediction period. We give the evaluation standard and the comparative models:

(i) Evaluation Standard: For the binary classification problem, there are four cases when an instance is classified as positive or negative. TP (True Positive): the number of correctly classified positive classes; TN (True Negative): the number of correctly classified negative classes; FP (False Positive): the number of incorrectly classified positive classes; and FN (false negative): the number of incorrectly
classified negative classes. Consequently, it is possible to calculate the evaluation metrics for the model, with the ROC, AUC and F1 metrics considered to be the most effective ones, as discussed in [2].

(ii) Comparison Methods: Comparative experiments with the following methods have been carried out: logistic regression model (LR); the support vector machine with linear kernel (SVM); random Forest model, which trained by tuning parameters consisting of the number of trees (RF); 3-layer deep neural network (DNN); the previous representative work (ConRec) [12].

Table 2 and Table 3 present the results on the KDDCUP dataset and MOOCCube dataset for all comparison methods. Compared to LR and SVM, EMIF achieves 4.23–4.52% and 5.63–6.28% AUC score improvements on KDDCUP and MOOCCube, respectively. Moreover, compared to the RF and DNN, EMIF also demonstrated better performance on these datasets, while also achieving better performance than previously reported techniques. These results demonstrate the effectiveness of EMIF, which does not rely on hand-crafted features and is presented as a dataset-independent architecture. We follow the experimental settings in [12] to produce *.

Table 2. KDD Metrics Comparison.

| Type          | Algorithm | AUC   | F1    |
|---------------|-----------|-------|-------|
| Baseline      | LR        | 0.8497| 0.9124|
|               | SVM       | 0.8468| 0.9078|
|               | RF        | 0.8632| 0.9187|
|               | DNN       | 0.8692| 0.9193|
| Prior Work [12]| ConRec   | 0.8742| 0.9227|
| Proposed      | EMIF      | 0.8920| 0.9231|

Table 3. MOOCCube Metrics Comparison.

| Type          | Algorithm | AUC   | F1    |
|---------------|-----------|-------|-------|
| Baseline      | LR        | 0.8208| 0.8813|
|               | SVM       | 0.8143| 0.8802|
|               | RF        | 0.8435| 0.8976|
|               | DNN       | 0.8503| 0.8988|
| Prior Work [12]| ConRec *| 0.8538| 0.8996|
| Proposed      | EMIF      | 0.8771| 0.9045|

5. Conclusions
In this paper, we have proposed an end-to-end multi-view interactive framework called EMIF for learner dropout prediction. The various aspects of learner interaction are analysed in detail. We extract learning activity information, user enrolment information, and course knowledge information through a set of 1D-CNNs and use multi-view learning to obtain mutual information of these, and fuse the above information-view loss through a self-weighing method. This framework removes hand-crafted features and achieves better performance metrics on both datasets. For the future, we aim to extend this work into the development of dropout prevention interventions.

Acknowledgments
We express our gratitude to the staff of KDDCUP 2015 for providing the dataset. Sincere thanks are also extended to the Knowledge Engineering Group of Tsinghua University for collecting the MOOCCube for further research.

References
[1] Z. Chen et al. “Who's Benefiting from MOOCs, and Why”. In: Harvard Business Review Digital Articles (2015).
[2] Jesse Davis and Mark Goadrich. “The relationship between Precision-Recall and ROC curves”. In: Proceedings of the 23rd international conference on Machine learning. 2006, pp. 233-240.

[3] Wenzheng Feng, Jie Tang, and Tracy Xiao Liu. “Understanding dropouts in MOOCs”. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 01. 2019, pp. 517-524.

[4] Liu Haiyang et al. “A time series classification method for behaviour-based dropout prediction”. In: 2018 IEEE 18th international conference on advanced learning technologies (ICALT). IEEE. 2018, pp. 191-195.

[5] Sherif Halawa, Daniel Greene, and John Mitchell. “Dropout prediction in MOOCs using learner activity features”. In: Proceedings of the second European MOOC stakeholder summit 37.1 (2014), pp. 58-65.

[6] Jiazhen He et al. “Identifying at-risk students in massive open online courses”. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 29. 1. 2015.

[7] Kate S Hone and Ghada R El Said. “Exploring the factors affecting MOOC retention: A survey study”. In: Computers & Education 98 (2016), pp. 157-168.

[8] Rene F Kizilcec, Chris Piech, and Emily Schneider. “Deconstructing disengagement: analyzing learner subpopulations in massive open online courses”. In: Proceedings of the third international conference on learning analytics and knowledge. 2013, pp. 170-179.

[9] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: arXiv preprint arXiv:1301.3781 (2013).

[10] Pedro Manuel Moreno-Marcos et al. “Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs”. In: Computers & Education 145 (2020), p. 103728.

[11] Jindan Tan et al. “Attentional Autoencoder for Course Recommendation in MOOC with Course Relevance”. In: 2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). IEEE. 2020, pp. 190-196.

[12] Wei Wang, Han Yu, and Chunyan Miao. “Deep model for dropout prediction in MOOCs”. In: Proceedings of the 2nd International Conference on Crowd Science and Engineering. 2017, pp. 26-32.

[13] Tang Xiaolin et al. “Research on Face Recognition Algorithm Based on Improved Residual Neural Network”. In: Automation, Control and Intelligent Systems 9.1 (2021), p. 46.

[14] Wanli Xing et al. “Temporal predicaton of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization”. In: Computers in human behavior 58 (2016), pp. 119-129.

[15] Jinglin Xu et al. “Deep embedded complementary and interactive information for multiview classification”. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. 04. 2020, pp. 6494-6501.

[16] Diyi Yang et al. “Towards an integration of text and graph clustering methods as a lens for studying social interaction in MOOCs”. In: International Review of Research in Open and Distributed Learning 15.5 (2014), pp. 215-234.

[17] Jifan Yu et al. “MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs”. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, pp. 3135-3142.