Combining Periodic Feature and Behavioral Transfer With Ensemble of Models for MOOCs Dropout Prediction

Na Yan
College of Photonic and Electronic Engineering
Fujian Normal University
Fuzhou, China
yanna_rxzd@foxmail.com

Tianwu Lin
College of Photonic and Electronic Engineering
Fujian Normal University
Fuzhou, China
1228335170@qq.com

Yiheng Zheng
College of Photonic and Electronic Engineering
Fujian Normal University
Fuzhou, China
1379550594@qq.com

Zhexin Xu*
College of Photonic and Electronic Engineering
Fujian Normal University
Fuzhou, China
xuzhexin@fjnu.edu.cn

2021 IEEE International Conference on Engineering, Technology & Education (TALE) | 978-1-6654-3687-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/TALE52509.2021.9678819

***Abstract*—In recent years, massive open online courses (MOOCs) have developed rapidly and have attracted millions of online users. However, the high dropout rate is a persistent problem. In order to address this problem, the behavioral data of learners are analyzed to explore their tendency to dropout, which can provide a reliable basis for adopting corresponding strategies to decrease the dropout rate. Through in-depth data exploration and analysis, it is found that the periodicity in the learning behaviors of users, and different types of learning users also have obvious differences in their learning behavioral transfer patterns. Based on these characteristics, we propose a ensemble model based on periodic features and behavioral transfer feature(PBTF-EM). The experimental results of the KDDCUP dataset show that the proposed model has a good effect compared with several deep learning algorithms, and have better interpretabillity and less computing capability, and the support on small datasets is robust.**

Keywords*—*MOOC dropout prediction, periodic feature, behavioral transfer, ensemble model

I. Introduction

Online learning has gradually become the mainstream way of learning. MOOC (Massive Open Online Course) has begun to be known by more people, and with the rapid popularity of MOOCs. But online learning is different from traditional education.,learning users often participate in studying for different purposes [1], and there are no penalties for users who dropout. Although online learning has a high number of user traffic, it still faces the main problem of low passing rate. According to statistics, the completion rate of MOOCs is less than 10% [2],[3]. Therefore, predicting the user's dropout tendency in advance is conducive to the online learning platform to provide effective intervention measures.

The learning sequence information of online learning users contained in the MOOCs dataset is essentially time-series data, and the time interval between events is often different [4]. Not only that, the learning behaviors of users are relative for several consecutive days [5], and the dropout behavior is easily affected by the type of courses and the number of dropout friends [6]. The user's own cognitive thinking also have an impact on the final learning result [7]. Therefore, in the process of analyzing and mining MOOCs data, it is necessary to consider the factors that may affect the dropout of learning users, and apply these to construct dropout prediction models as much as possible.

Someone conducted statistics on the commonly used models in papers studying MOOC dropout prediction and found that logistic regression and SVM showed the highest frequency. Some researchers have also paid attention to Random Forest (RF), Decision Tree (DT), Recurrent Neural Network (RNN), Deep Neural Network (DL) [8], Convolution Neural Network (CNN) and Long Short Time Memory Network (LSTM) and other algorithms in the performance of dropout prediction. Wei et al. [1] built a Bayesian hierarchical Logistic regression model to study the completion gap between different motivations; Hong et al. [9] constructed RF and SVM into a two-level hierarchy The associative classifier is used to predict the user's dropout; the integrated algorithm GBDT [10],[11] can also achieve good results in the dropout prediction task; Chen et al. [12] mapped the results of the decision tree which based on the entropy theory to optimize ELM (Extreme Learning Machine) model; Zhou et al. [13] proposed a dropout prediction model based on multi-model superimposed integrated learning (MMSE), and established a two-layer integrated learning model. There are also some researchers who build predictive models on the basis of CNN. Qiu et al. [14] used CNN as a training model for feature sets; Wen et al. [5] used CNN to preserve the local correlation of learning behavior; Feng et al. [6] added an attention mechanism to CNN for calculation the weight of features; Zheng et al. [15] constructed FWTS-CNN which combines feature weighting and behavioral time series with CNN; Wei W et al. [16] combined CNN with RNN to construct dropout prediction model which can extracts the features automatically.

Based on previous research and analysis, it can be found that they did not consider the periodicity of users' learning behaviors and took it as a feature. They also didn't think of the transfer between user learning behaviors can often reflect the user's learning mode, and play a certain role in predicting the user's learning state. In addition, the network structure of the deep learning algorithms is inexplicable, and the training process requires a mass of data, which requires relatively higher computing capability. Therefore, in the research of this paper, we propose an ensemble model based on periodic feature and behavioral transfer feature, so that it can achieve good performance with better interpretabillity and less computing capability, and better support for small datasets.

The rest of this paper is organized as follows: Section II introduces the dataset and data exploration; Section III describes the algorithm and briefly introduces the construction methods of different features; Section IV performs simulation analysis, describing the entire experimental process and results; Finally, provides the final conclusion in Section V.

* denotes corresponding author.

II. DATASETS AND DATA EXPLORATION

A. Datasets

TABLE I. THE DATASETS OF KDDCUP2015

Filename	***columns***
enrollment_train.csv / enrollment_test.csv	enrollment_id
	username
	course_id
log_train.csv / log_test.csv	enrollment_id
	time
	source
	event
truth_train.csv	enrollment_id
	dropout
object.csv	course_id
	module_id
	category
	children
	start
date.csv	course_id
	from
	to

The content of this paper is based on the dataset from KDDCUP2015. KDD is short for Knowledge Discovery and Data Mining, is an annual competition organized by SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) of ACM (Association for Computing Machiner).This dataset provides learning behavior information of 39 courses and their enrolled students for nearly half a year. The dataset contains five files, all in CSV format, which record the behavior logs, information of courses and users on the platform. The purpose of the label is to predict whether the learner will continue to show the learning behavior in the next 10 days based on the learning behavior of the studying in the past 30 days. The detailed information is shown in Table I.

B. Data Exploration

The purpose of data exploration is to understand the overview of the data and insights of them, and provide evidences for data cleaning, feature extraction and selection of dropout predictive models.

Firstly, analyze the amount of data and the sample ratio. In Table II, it can be observed that there are 39 courses included in the dataset, and the opening period of each course is 30 days, so the subsequent behavior occurrence time is 30 days for exploratory analysis. And the number of training sample is about 2-3 times as many as test samples.

TABLE II. THE STATISTICS OF KDDCUP DATASETS

	Number of enrollment_id	***Number of records***	***Number of courses***	***Course duration(days)***
Train set	120,542	8,157,277	39	30
Test set	80,362	5,387,847		
Total	200,904	13,545,124		

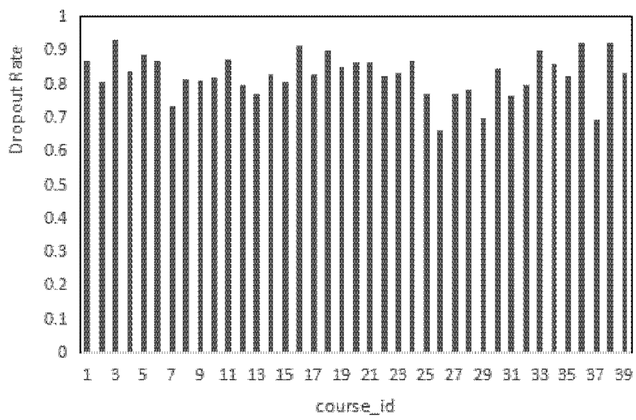


Fig. 1. Distribution of dropout rates in different courses

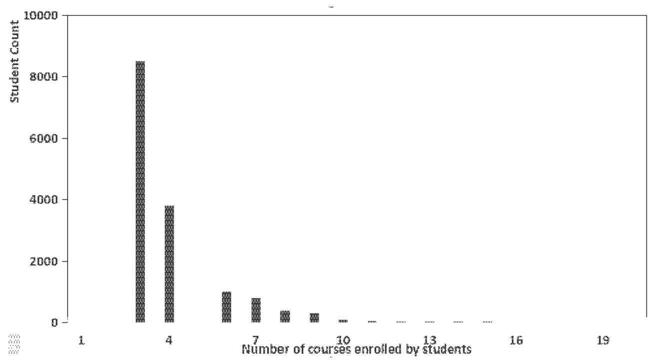


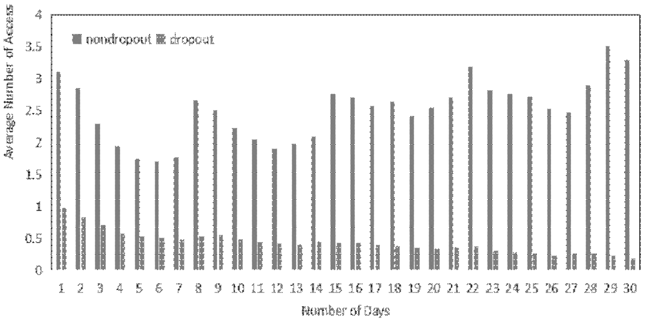
Fig. 2. Distribution of dropout rates in different courses

It can be seen from Fig.1 that the dropout rate of each course is relatively high. By observing the distribution of the number of courses enrolled by users in Fig.2, it can be found that there are more than 50% users who have enrolled for less than two courses, which will affect the calculation of the personal dropout rate. If we use the number of dropout courses to the total amount of courses enrolled by users as the personal dropout rate. Like that users with fewer courses, the personal dropout rate is more likely to be affected, so that the dropout rate cannot truthfully reflect the user's status. Therefore, a threshold for the number of courses should be set for the calculation of the personal dropout rate, and users below the threshold should use means as filling to reduce uncertain interference.

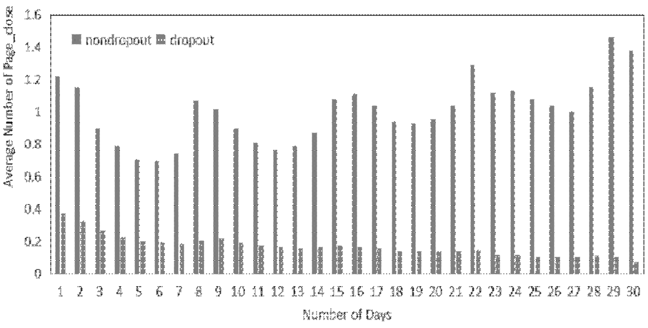
TABLE III. THE STATISTICS OF EVENT BEHAVIORS

The type of Event behaviors	***Number of Train set***	***Number of Test set***	***Ratio_Train***	***Ratio_Test***
Access	3,777,130	2,240,915	0.3626	0.2151
Discussion	1,057,218	610,205	0.1015	0.0577
Navigate	1,589,302	901,814	0.1526	0.0866
Problem	1,322,061	816,909	0.1269	0.0784
Video	1,013,788	586,670	0.0973	0.0563
Wiki	172,986	96,030	0.0166	0.0092
Page_close	1,485,056	889,872	0.1426	0.0854

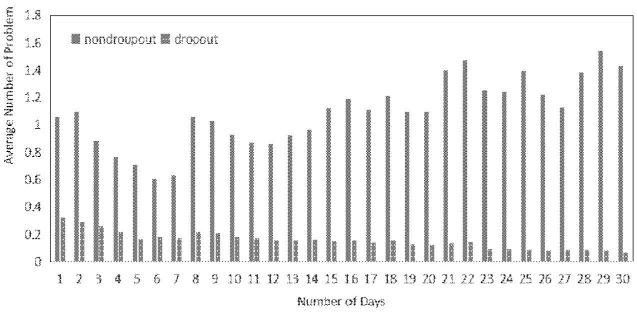
Periodicity of Behaviors. The data of various behaviors contained in the event have been listed in Table III. We conducted these statistical data according to the number of days and found that some behaviors showed a certain periodicity as shown in Fig.3, the trends of access, page_close, video, and problem behaviors have clear boundaries between two different users who are non-dropout and users who are dropout.



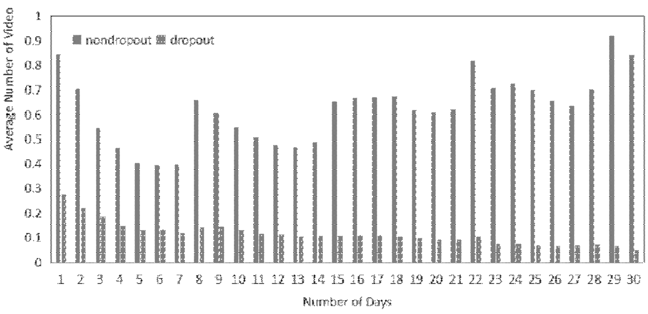
(a) Trend of Access



(b) Trend of Page_close



(c) Trend of Problem



(d) Trend of Video

Fig. 3. Trends of different event behaviors

In addition, it can also be found that users who are non-dropout exhibit a certain periodicity in the distribution of access, page_close, problem and video, while the average number of behaviors of dropout users is generally small and shows a downward trend with time.

Trough the data exploration, it can be found that access, page_close, problem, and video have obvious periodicity (weekly) in event, and the average value of the number of behaviors shows a periodic increase trend. Therefore, it would be a good choice to use the periodicity of the behavioral data of these events as features. The specific construction of periodic features will be explained in the algorithm description section.

Behavioral Transfer. We did another analysis to explore how distinct of the behavior pattern between two different users. We extract user learning patterns from the perspective of transfer frequency. According to the different behavior patterns, the probability distribution of the next behavior selected by the user is also different. In order to quantify the behavior transition probability, we designed a behavior transition matrix based on the adjacent matrix of the graph structure[17]. Regard each type of learning behavior as a node in the graph structure, the relationship between the nodes is stored in an adjacent matrix, and the weight of the edge between the nodes represented by the amount of such behavioral transfer's occurrences.

In order to further quantify the difference between the two different user behavior patterns, we conducted a statistical comparison of behavior transfer. There is also an obvious distinction between the amount of behavioral transfers of these two different users, as shown in Fig.4. The figure shows the comparison of some average data of two kinds of user. It can be found that these behavior transition pairs have a clear degree of discrimination e.g. the number of behavioral transfer about access to page_close pair is higher than the dropout users.

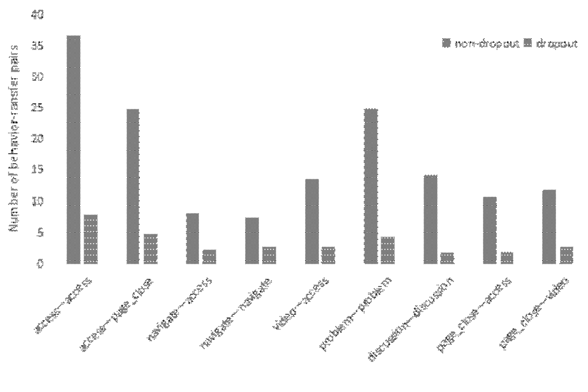


Fig. 4. Comparison of behavioral transfer

III. ALGORITHM DESCRIPTION

A. *Model Diagram*

The proposed method are showed in Fig.5. The architecture of the PBTF-EM can be divided into three parts: feature processing and selection, basic learners construction and the construct of ensemble models. In feature processing and selection, we traverse the entire log file, classify each line of the log according to the enrollment_id in that row, and store it in the dictionary data structure using the enrollment_id as an index., and a dictionary structure is used to store the mapping relationship between character values and integer values. In the process of traversing the log file, the unique mapping relationship between "course_id (C), enrollment_id(E) and username (U)" is obtained and stored in a dictionary format. And then through the data exploration and selection we extract the features of datasets.

The construction of base learners is mainly to train the four algorithms of logistic regression, SVM, random forest, and GBDT (Gradient Boosting Decision Tree) using the feature sets fused by the constructed periodic feature and behavioral transfer feature with the basic feature respectively. The ensemble models are constructed by training the three ensemble algorithms of XGBoost, LightGBM, and CatBoost from different prediction subsets of base learners with outstanding effects, and finally use weighted average to combine predictions from three ensemble models so that we obtain the final results.

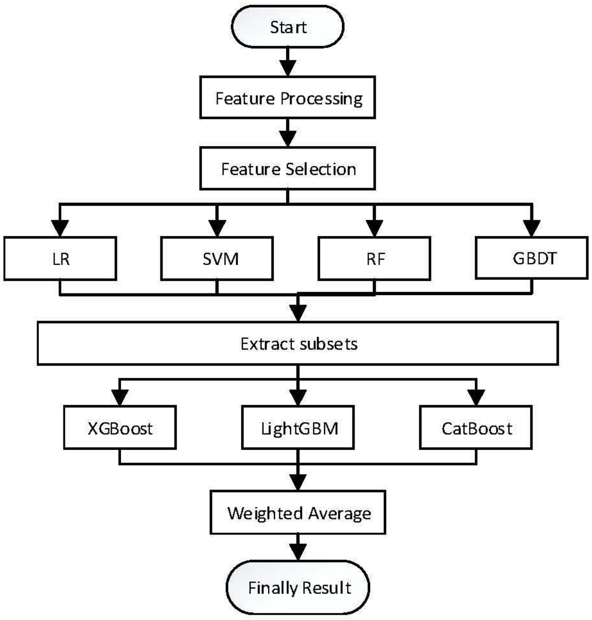
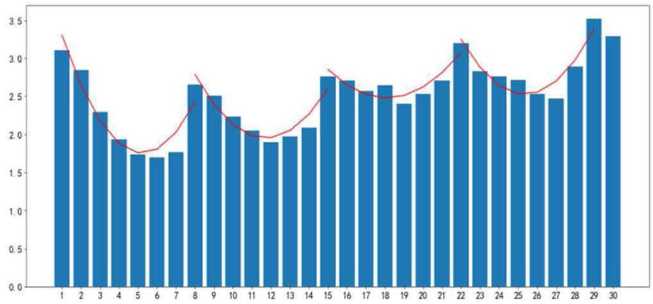


Fig. 5. Framework of PBTF-EM

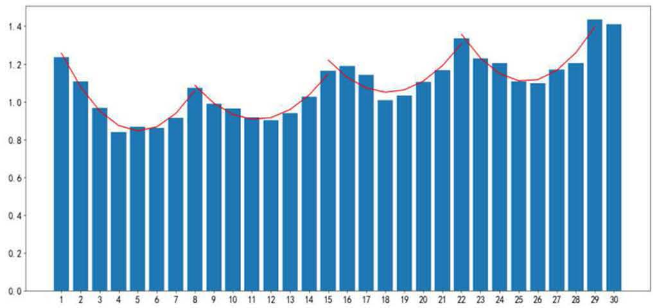
B. *Construction of Feature Modules*

- Periodic feature

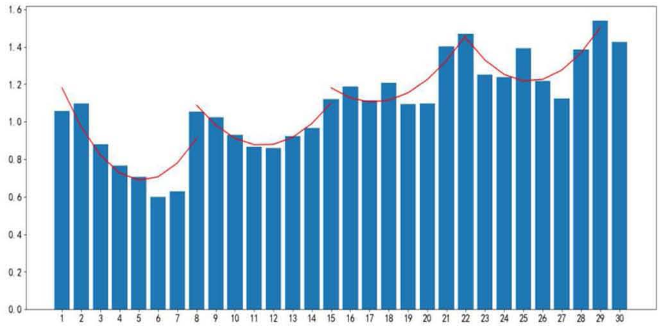
The purpose of establishing periodic features is because the learning behaviors of learning users are locally correlated, which indicates that the learning state of users tends to show a certain regularity. Through the exploration and analysis of the dataset in the second section, it can be found that the learning states of the two types of users in the continuous time are obviously different, and the non-dropout users have obvious periodicity in some learning behaviors while dropout uses are decreasing gradually. In order to express this periodicity, this paper constructs three periodic features: quadratic fitting coefficient, slope and cosine similarity. We use quadratic linear function to fit the periodic behaviors of non-dropout users as shown in Fig.6. And then, we extract their slopes ,quadratic fitting coefficients from the fitting equation as periodic features. The formulas for the quadratic linear function and slope are shown in (1) and (2) respectively. Among them, α, β, γ are the parameters of equation.



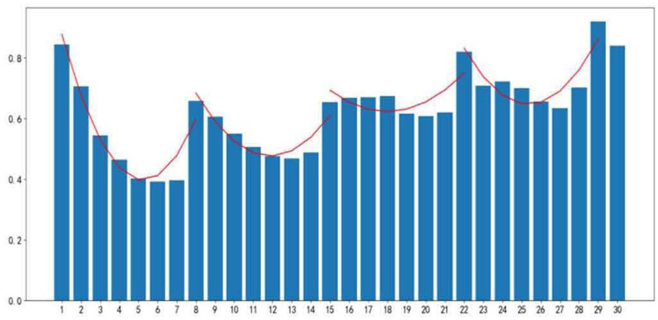
(a) Fitting of access



(b) Fitting of page_close



(c) Fitting of problem



(d) Fitting of video

Fig. 6. Fitting diagram of each periodic behavior

$$\psi = \alpha\xi^2 + \beta\xi + \gamma \tag{1}$$

$$k = \frac{\phi(\xi + \Delta\xi) - \phi(\xi)}{\Delta\xi} \tag{2}$$

In addition, in order to better reflect the periodicity, we also calculate the similarity of each type of users and consider similarity as a feature. The process is as follows: the 30-day behavior of each user is divided into four periods, and statistic each behavior number of the user occurs per day. As shown in (3), Ψ^i is the entire behavioral transfer matrix, and ψ_1^i is the situation in week 1, $c_{1,2}^i$ is the status on the 2nd day of the first week, and so on. Construct each period into the form of a vector, and finally each behavior obtains its corresponding period matrix in this way, and finally uses (4) to calculate the similarity between each period.

$$\Psi^i = \begin{pmatrix} \psi_1^i \\ \psi_2^i \\ \psi_3^i \\ \psi_4^i \end{pmatrix} = \begin{pmatrix} c_{1,1}^i & c_{1,2}^i & \cdots & c_{1,7}^i \\ c_{2,1}^i & c_{2,2}^i & \cdots & c_{2,7}^i \\ c_{3,1}^i & c_{3,2}^i & \cdots & c_{3,7}^i \\ c_{4,1}^i & c_{4,2}^i & \cdots & c_{4,7}^i \end{pmatrix} \tag{3}$$

$$\cos(\xi_1, \xi_2) = \frac{\xi_1 \cdot \xi_2}{|\xi_1||\xi_2|} \tag{4}$$

- Behavior Transfer Feature

The purpose of establishing the features of learning behavioral transfer is to extract the user's learning pattern from the perspective of transfer frequency. In order to quantify the behavior transition probability, first, we need to extract seven behaviors from the event as nodes in the graph structure, and then calculate the number of transfer in pairs for each enrollment_id per day based on logs, and finally construct a behavioral transfer matrix as the demo shown in Fig.7.

Through previous studies, it is found that the time interval of learning behavior also has a certain impact on the future state of users. Therefore, this paper will also extract the duration sequence of each user's learning behaviors for model training, as shown in Fig.8.

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \Leftarrow \left\{ \begin{array}{l} \textit{access} \\ \textit{video} \\ \textit{discussion} \\ \textit{video} \\ \textit{discussion} \\ \textit{problem} \\ \textit{wiki} \\ \textit{access} \end{array} \right.$$

Fig. 7. The construction of behavioral transfer matrix

$$\text{Time interval sequence:} \begin{bmatrix} 2 \\ 4 \\ 591 \\ 400 \\ 35 \\ 26 \\ 2 \\ \cdots \end{bmatrix} \left\{ \begin{array}{l} T0{:}103\,\textit{access} \\ T1{:}105\,\textit{navigate} \\ T2{:}109\,\textit{wiki} \\ T3{:}700\,\textit{video} \\ T4{:}1100\,\textit{discussion} \\ T5{:}1135\,\textit{problem} \\ T6{:}1161\,\textit{page\_close} \\ \cdots\cdots \end{array} \right.$$

Fig. 8. Time interval sequence

IV. EXPERIMENT AND RESULTS

Basic features are fused with he three periodic features of quadratic fitting coefficient, slope and cosine similarity and the behavioral transfer feature. A variety of different machine learning algorithms are used to train the prediction models, and we tune the parameters based on 5-fold cross validation with the grid search. Among them, the regularization strength is set to 0.1 and use L2 regularization term in LR (Logistic Regression), and the number of classifiers in RF (Random Forest) is 300. At same, in ensemble models, learning rate is set to 0.1, and there are 10 classifiers in XGBoost (eXtreme Gradient Boosting) and CatBoost (Categorical Boosting) and lambda is set to 0.5, besides there are 100 classifiers in LightGBM (Light Gradient Machine). The adjustment of these parameters reinforces the performance of the model to a certain extent. And we use precision, recall, f1 score, area under the ROC curve (AUC) and accuracy to evaluate the predict models.

Table IV is the best result of base learners we trained in the first layer, among them, the basic features fuse with cosine similarity in the LR, the basic features fuse with the quadratic fitting coefficient in the RF, the basic features fused with the quadratic fitting coefficient and the cosine similarity in the SVM (Support Vector Machine), and the basic features fuse with behavioral transfer feature in the GBDT (Gradient Boosting Decision Tree) perform best. And the different prediction subsets of the best-performing base learners are used to train the ensemble learners: XGBoost, LightGBM and CatBoost. The results of them are shown in Table V.

TABLE IV. BEST PERFORMANCE OF BASE LEARNERS

	Precision	*Recall*	*F1*	*AUC*	*Accuracy*
LR	0.8832	0.9685	0.9204	0.8815	0.8657
SVM	0.8828	0.9657	0.9201	0.8802	0.8654
RF	0.9012	0.9380	0.9134	0.8746	0.8821
GBDT	0.9047	0.9473	0.9191	0.8870	0.8758

TABLE V. BEST PERFORMANCE OF ENSEMBLE LEARNERS

	Precision	*Recall*	*F1*	*AUC*	*Accuracy*
XGBoost	0.8968	0.9547	0.9218	0.8853	0.8840
LightGBM	0.8981	0.9606	0.9275	0.8964	0.8921
CatBoost	0.8976	0.9617	0.9274	0.9006	0.8934

TABLE VI. PERFORMANCE OF PBTF-EM VS. DIFFERENT MODELS

	F1	*AUC*
CNN+RNN [16]	0.9214	0.8742
DP-CNN [14]	0.8421	0.8785
DT+ELM [12]	0.9184	0.8586
Attention+CNN [6]	0.9287	0.9003
MMSE [13]	0.9260	0.8355
PBTF-EM	0.9201	0.8907

TABLE VII. ERROR OF PBTF-EM VS. DIFFERENT MODELS

	F1	*AUC*
PBTF-EM vs. CNN+RNN [16]	-0.0013	0.0165
PBTF-EM vs. DP-CNN [14]	0.0780	0.0122
PBTF-EM vs. DT+ELM [12]	0.0017	0.0321
PBTF-EM vs. Attention+CNN [6]	-0.0086	-0.0096
PBTF-EM vs. MMSE [13]	-0.0059	0.0552

Combining the results of Table IV and Table V, it can be found that the performance of the tree-based ensemble model of F1, AUC, and Accuracy are greater than that of the single models. This is because the ensemble algorithm optimizes the training scheme under the premise of ensuring interpretability and to reduce over-fitting.

Finally, we use weighted average to attain the final result. Due to the imbalance of positive and negative samples in this dataset, precision, recall and accuracy cannot well reflect the performance of the model, so in the end we use F1-score and AUC as the final evaluation indicators and compare the proposed model with the models in other references. Combining Table VI and Table VII, it can be found that the F1 and AUC values of the proposed model are exceed some deep learning algorithms, especially in AUC, it's only less than the model combined attention mechanism with CNN[6]. This is precisely because we have a more detailed analysis of user behaviors and constructed features that can distinguish two different types of users, so that they can achieve a good effect without the aid of neural networks. And to compare with neural network, the proposed model is more interpretable and requires less computing capability, as well, the support is robuster on small datasets.

V. CONCLUSION AND FUTURE WORK

This paper we conduct data exploration and analysis of the dropout in MOOCs, and we construct the periodic features and behavior transfer features and proposed PBTF-EM. Through experimental analysis, it is concluded that the proposed model without neural network has improved in some evaluation indicators compared with several deep learning algorithms with less computing capability. And the interpretabillity and applicability on small datasets are better than deep learning algorithms. However, in order to fully study and compare the differences and commonalities between ensemble algorithms in machine learning and neural networks in the construction of MOOC dropout prediction models, this paper focuses on the application of ensemble algorithm in dropout prediction at first, and in the next, we will construct a network model based on CNN and the proposed model in this paper, continue to improve and optimize, so as to better understand the machine learning and enhance MOOCs dropout prediction.

ACKNOWLEDGMENT

This work was supported partly by Natural Science Foundation of Fujian Province (No. 2020J01157)

REFERENCES

[1] L. Wei and F. N. Kondo, "Exploring important motivation drivers for completion of massive open online courses by Bayesian Hierarchical Logistic Regression model," 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI), 2019, pp. 196-202.

[2] A. Ramesh, D. Goldwasser, B. Huang, H. DaumÈ III, and L.Getoor, "Learning latent engagement patterns of students in online courses," in Proc. the Twenty-Eighth AAAI Conference on Artificial Intelligence, Qu´ebec, Canada, 2014, pp. 1–7.

[3] J. He, J. Bailey, B. I. P. Rubinstein, and R. Zhang,"Identifying at-risk students in massive open online courses,"in Proc. the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 2015, pp. 1749–1755.

[4] Wang, L, and H. Wang, "Learning behavior analysis and dropout rate prediction based on MOOCs data." 2019 10th International Conference on Information Technology in Medicine and Education (ITME) 2019.

[5] Wen Y, Tian Y, Wen B, Zhou Q, Cai. G, and Liu. S, "Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs." Tsinghua Science and Technology 25.3(2020): 336-347.

[6] Feng W, J. Tang, and T. X. Liu, "Understanding dropouts in MOOCs." Proceedings of the AAAI Conference on Artificial Intelligence 33(2019): 517-524.

[7] Sinha T, Jermann P, Nan L , and Dillenbourg P, "Your click decides your fate: Inferring information processing and attrition behavior from MOOC video clickstream interactions." EMNLP Workshop on Modelling Large Scale Social Interaction in Massive Open Online Courses 2014.

[8] Tamada M. M, J. Netto, and D. Lima, "Predicting and reducing dropout in virtual learning using machine learning techniques: A systematic review." 2019 IEEE Frontiers in Education Conference (FIE) IEEE, 2019.

[9] Hong B, Z. Wei, and Y. Yang, "Discovering learning behavior patterns to predict dropout in MOOC." International Conference on Computer Science & Education 2017: 700-704.

[10] Ardchir S, Ouassit Y, Ounacer S, Jihal H, and Azouazi M, "Improving Prediction of MOOCs Student Dropout Using a Feature Engineering Approach." 2020.

[11] Liang, J. , L. Chao, and Z. Li, "Machine learning application in MOOCs: Dropout prediction." International Conference on Computer Science & Education IEEE, 2016.

[12] Chen J, Feng J, Sun X, Wu N, and Chen S, "MOOC dropout prediction using a hybrid algorithm based on decision tree and extreme learning machine." Mathematical Problems in Engineering 2019.1(2019): 1-11.

[13] Zhou Y, and Z. Xu, "Multi-Model stacking ensemble learning for dropout prediction in MOOCs." Journal of Physics Conference Series 1607(2020): 012004.

[14] Qiu L, Y Liu, Hu Q, and Y Liu, "Student dropout prediction in massive open online courses by convolutional neural networks." Soft Computing 23(2018).

[15] Zheng Y. F, Gao Z. H, Wang Y. H, and Fu Q, "MOOC dropout prediction using FWTS-CNN model based on fused feature weighting and time series." IEEE Access.

[16] Wei. W, Y. Han , and C. Miao, "Deep model for dropout prediction in MOOCs." the 2nd International Conference 2017.

[17] Wang X, Huang T, Wang D, et al, "Learning intents behind interactions with knowledge graph for recommendation," arXiv preprint arXiv: 2102.07057, 2021.