Received: June 15, 2017
Revision received: May 22, 2018
Accepted: May 30, 2018

www.estp.com.tr
DOI 10.12738/estp.2018.5.015 ♦ October 2018 ♦ 18(5) ♦ 1138-1149

Research Article

Feature Extraction and Learning Effect Analysis for MOOCS Users Based on Data Mining

Biqin Yang[1]
Hainan Normal University

Zhi Qu [2]
Hainan Normal University

Abstract

Feature extraction is an important technology of data mining, it had been widely used in machine learning and pattern recognition. With the increasing number of enrolments in MOOCs, there was a large amount of learning behaviour data generated in MOOCs platforms. Through the data analytics of these learning behaviour data, some useful predictions and guidance information will be obtained. In this paper, two machine learning algorithms based on learner behavioural data is presented. According to pre-extracted features, our algorithms would take the history data into account and can detect changes in learner behaviour over time. In our experiments, we use our proposed algorithms to predict dropout rates of MOOCs. Compared with other existing algorithms in common use, our proposed algorithm can predict dropout accurately better.

Keywords

MOOCs • Feature Extraction • Machine Learning • Learning Behaviour Analytics

♦ Hainan Educational Science Planning Project (QJY1251519)

[1]**Correspondence to:** Biqin Yang (PhD), Hainan Normal University, Haikou 571158, China. Email: yangbiqin1987@163.com
[2] Hainan Normal University, Haikou China. Email: 82399897@qq.com

Citation: Yang, B. Q., Qu, Z., Feature Extraction and Learning Effect Analysis for MOOCS Users Based on Data Mining. *Educational Sciences: Theory & Practice*, *18*(5), 1138-1149. http://dx.doi.org/10.12738/estp.2018.5.015

Massive open online courses (MOOCs) are becoming an emerging research area for education analytics (Breslow, Pritchard & Deboer, 2013). Modeling MOOCs users' behaviors can understand user demands better, and the analytics results which are more conducive to learning can be offered. Data mining that users generated in MOOCs platforms provide the possibility to study learning effect, predict dropout rates, and designed targeted interventions for teaching guidance (Lockyer, Heathcote & Dawson, 2013).

Despite the increasing number of MOOC users, there were some common features of the behaviors of these users on the MOOCs platform. Through extracting these behavioral features from huge amounts of data, we can use machine learning algorithm for prediction.

In this paper, we firstly proposed some input features and learning behavior features which were extracted from the data of MOOCs platforms. Then, two machine learning algorithms based on support vector machines and artificial neural network for predicting dropout of MOOC course were proposed, respectively.

The rest of the paper is divided into the following sections. Related works regarding machine learning algorithms used to predict dropout and various types of features used in them are discussed in Section 2. Proposed machine learning algorithms based on artificial neural network is described in Section 3. Another machine learning algorithm based on support vector machines is proposed in Section 4. The experiments and analysis are given in Section 5. Finally, Section 6 is conclusions.

Related work

With the popularity of MOOCs, there had been many studies to predict future behavior of users in MOOCs through extracting a wide variety of features from user behavioral data and applying various machine learning algorithms.

Wen et al. used a linguistic algorithm to analyze the MOOC forum data and discovered valuable features for predicting dropout of users (Sinclair & Kalvala, 2016). Brinton, Buccapatnam & Chiang, (2015) depended on clickstream of video lecture in MOOC to capture users' behavioral patterns, which were used to construct information processing indicators of users. Hughes & Dobbins, (2015) proposed a hidden Markov model and some related features such as number of forum posts, percentage of lectures watched, and so on to predict attrition of users in MOOC. Sinha, Li and Jermann, (2014) used graph theory to capture sequence of active and passive user behavior and used graph metrics as features for predicting dropout. Amnueypornsakul, Bhat & Chinprutthiwong, (2014) proposed a new predicting model by some quiz related and behavior related features. Kloft, Stiehler & Zheng, (2014) extracted more than 15 features representing user behavior from clickstream log.

All above approaches use different machine learning algorithms including logistic regression, support vector machine, hidden Markov model and random forest. These algorithms have a common point, that is, they all need feature extraction.

Feature extraction is also called attribute selection. It is the process of selecting a subset of relevant features for the predictive problem in model construction (Ahmed, Qahwaji & Colak, 2013). Feature extraction can be used to identify relevant attributes from dataset that do really contribute to the accuracy of predictive model. Rossi et al. studied discussion threads in Coursera MOOC forums, and used machine learning techniques to analyze and classify by feature extraction (Rossi & Gnawali, 2014). Zhang, Yang & Huang, (2017) designed a personalized MOOC recommendation system based on feature extraction, and it can recommend the best suitable course for the users.

Compared with these existing literatures, the objective of our feature extraction is: improving the prediction performance of the predictors and providing faster and more cost-effective predictors. We extracted the most representative features from clickstream file and forum posts data in MOOC platform, then these features needed to be post-processed. After the post-processing, the prediction accuracy of the machine learning algorithm will be greatly improved, and it is useful for the instructor to take necessary measures to prevent or reduce user attrition during the course.

User dropout prediction algorithm based on artificial neural networks

The dataset which we extracted features from contains more than 3 million users click logs and over 5000 forum posts. The click streams logs consist of clicks made while watching video lectures and requests for viewing forums. There is time stamp with each click. The input features extracted from the dataset are shown in Table 1.

Table 1
Input Extraction Features List

Extraction features	Explanation
User ID	Unique numerical identity of the user in MOOC
Course Week	The number of weeks since course had begun
User week	The number of weeks since user has joined the course
Number of clicks	The number of clicks by the user in the current week
Number of study sessions	The number of study sessions by the user in the current week
Number of course pages viewed	The number of course pages viewed by the user in current week which include all pages except the video lectures
Number of forum pages viewed	The number of forum pages viewed by the user in current week
User sentiment	User sentiment of forum posts in the current week

All of the input features in Table 1. except User sentiment had been proved to be the most effective by previous researches. Next, we will introduce the role of user semantic analysis in prediction.

Sentiment analysis

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information (Zhao, Qin & Liu, 2010). Generally speaking, sentiment analysis aims to determine the attitude of a subject

with respect to some topic. There were mainly three kinds of approaches to sentiment analysis: knowledge-based methods, statistical methods, and hybrid methods (Cortes & Vapnik, 1995).

In this paper, a knowledge-based method to extract sentiment from forum posts of MOOC platform is used. Given the forum post, we used Stemmer to pass the stem of each word of forum post and used POS (Part-of-Speech) Tagger to pass its POS Tag to Senti Word Net 3.0 (Hamer, 1980), which can assign to each synset three sentiment scores: positivity, negativity, objectivity. The sentiment score of the forum post can be calculated by the total sentiment scores of all the words in the post.

The process of sentiment analysis of forum post can be shown as the block diagram in Figure 1.

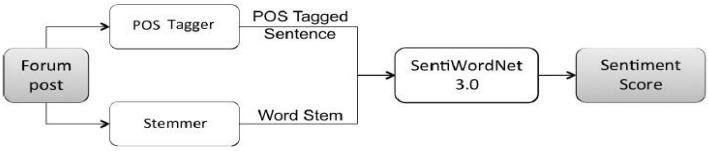

Figure 1. Block diagram of sentiment analysis of forum post

Artificial neural network

Because there was a large amount of inputs, especially, the specific relationships between inputs and outputs were unknown, it was very suitable to model of predicting user dropout in MOOC coursed by artificial neural networks. Different from other machine learning methods, artificial neural networks can model the output according to any arbitrary input function.

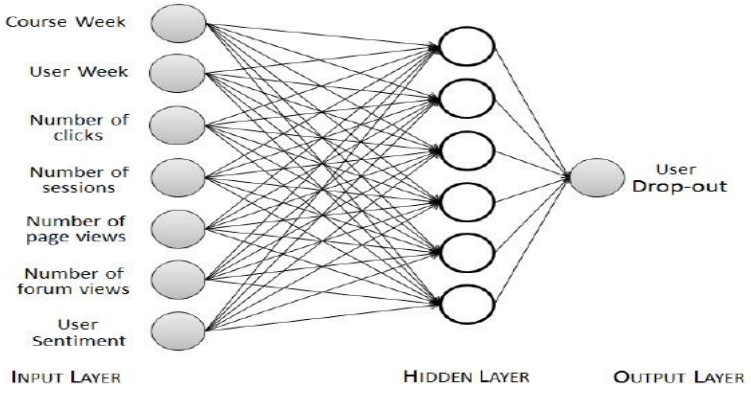

Figure 2. The structure of neural network used to predict user dropout

We selected seven extraction features in Table 1 except User ID as nodes in input layer of the artificial neural network which we constructed. Each input extraction feature should be normalized to between 0 and 1. There was only one node in output layer, it is responsible for predicting whether user would drop out in the next week. A hidden layer consisting of 6 neurons in the neural network between the input and output layer was

added. According to different demands, the number of neurons in the hidden layer can be adjusted to get the best results through experiment simulation.

The structure of the neural network used to predict user dropout is shown in Figure 2.

The neural network model can be trained by back propagation. The specific definition of the model can be shown as follows:

$$\hat{r}_{u,a} = b_u + p_u^t W f_{ua} \quad (1)$$

where $\hat{r}_{u,a}$ represents the possibility of dropout for a user u in the MOOCs course a, b_u represents bias term for user u, f_{ua} is the extracted feature vector, W is the coefficient matrix, p_u^t represents a vector which holds the relationships of user u within the different extracted features.

User dropout prediction algorithm based on support vector machines

In machine learning, support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for prediction and regression analysis.

Our proposed algorithm is used to predict the users' drop-out during the next week through using the data of the past weeks. Consequently, all features should be computed for each user and for each week. Some important properties of the data were visualized in Figure 3.

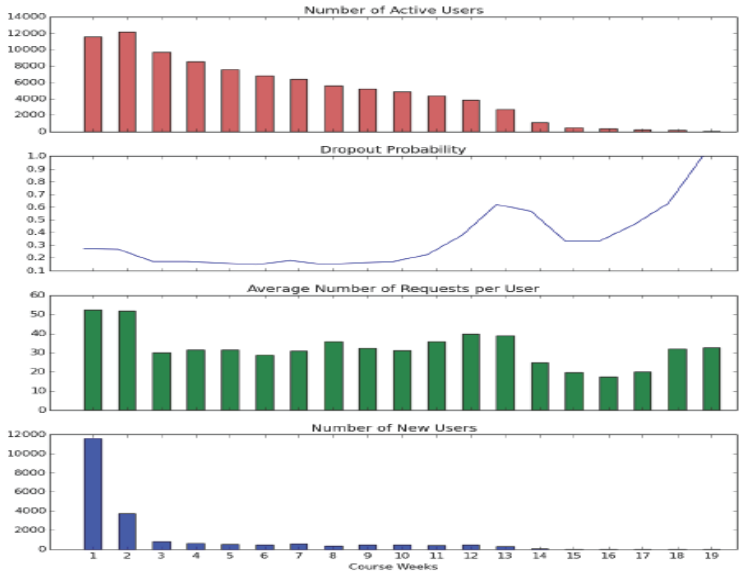

Figure 3. Some important properties of the analyzed dataset

Table 2
Extraction Features List

Extraction features	Explanation
Number of requests	Total number of requests
Number of sessions	a reflection of how often logging into the MOOCs platform
Number of active days	an active day means that the user had at least one session on that day
Number of page views	the page views consist of lecture pages, wiki pages, homework pages and forum pages
Number of page views per session	the average number of pages viewed by each user per session
Number of video views	total number of videos click actions
Number of video views per session	average number of videos click actions per session
Number of forum views	number of course discussion forum views
Number of wiki views	number of course wiki page views
Number of homework page views	number of course of homework page views
Number of straight-through video plays	straight-trough playing video means that the users played video without any jump
Number of start-stop during video plays	start-stop during video plays stands for a lecture video being paused and resumed
Number of skip-ahead during video plays	skip-ahead means that the user played a video with a forward jump
Number of re-listens during video plays	re-listen means that a backward jump was made as the user was playing a video
Number of slow play rate use	the feature means language difficulties or a lack of relevant background knowledge of the user
Most common request time	the feature is to separate day time learning from night time learning
Number of requests from outside of Coursera	the feature is to discover how many requests from third-party tools
Number of screen pixels	the screen pixels are an indicator of the device that the user used
Most active day	the feature shows that starting late or early could have an impact on dropout
Country	the feature could reflect geographical differences
Operating System	the feature reflects the operating system which the user used
Browser	the feature reflects the browser which the user used

We can see that the number of active users quickly decreases over time. The ratio of dropout is a little high in the beginning of the MOOC course and rising from around week 12 to the end of the MOOC course.

The all features extracted from the dataset are shown in Table 2.

There are three kinds of extracted features in Table 2. The first kind of features is that need to be summed up, the second kind is that need to be averaged and the last one is that need to be decided by majority vote.

For each week of the MOOC course (i = 1, 19), the dropout possibility of each user being active in that week can be computed. The vector $y_i \in \{-1,1\}^{n_i}$ is defined for the prediction results, where +1 represents dropout and -1 represents no dropout.

For the 22 extracted features described in Table 2, the first 19 features can be represented by a real number, other features had to be represented by a multidimensional space. For real number features, a matrix $X_i \in \mathbb{R}^{19\times n_i}$ is defined related to each week i of the MOOC course. The rows and columns of the matrix correspond to the features and user ids, respectively. Then we appended all the features of the previous weeks to the matrix. The process can be represented by $X_i \in \mathbb{R}^{19i\times n_i}$. Therefore, X_i can be defined as follows:

$$X_i = (x_1, \ldots, x_{n_i}) \tag{2}$$

where x_j represents the feature vector of the *j*th user.

Subsequently, simple *t-tests* for each extracted feature were performed, and the *Fisher score* defined in Eq. (3) was computed.

$$f_j = \sqrt{\frac{\mu_+ - \mu_-}{\sigma_+^2 + \sigma_-^2}} \tag{3}$$

where μ_+ and μ_- represents the mean of the positive and negative class, σ_+ and σ_- represents the variance of the positive and negative class.

Though *t-tests* and *Fisher scores* can get comparable results, superior approach with the Fisher score was made. The achieved scores of several video features, the most common request time, and the most active day feature were close to zero, so they should be removed from features of prediction.

The support vector machine (SVM) model which was used to predict user dropout is defined as follows:

$$f(x) := \langle w, x \rangle + b \tag{4}$$

The following equation can ensure the model to maximize the margin between positive and negative.

$$(w, b) := argmin_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} max(0, 1 - y_i(\langle w, x_i \rangle + b) \tag{5}$$

Experiments and analysis

Datasets

Our approaches were evaluated on three MOOCs: StMed (Statistics in Medicine), StLearn (Statistical Learning) and IntroCN (Introduction to Computer Networking).

There were lots of log files about a user viewing video lectures, checking forum posts, attempting quizzes and homeworks in all these datasets. There were also the number of enrolled users, the number of assessments, quizzes, homeworks, and corresponding scores. The distribution of users attempting the different assessments in the three MOOCs courses are shown in Figure 4.

Experiments based on artificial neural networks

In the experiments of predicting user dropout, our focus is to capture all users who are going to drop-out, so it is very important to minimizing false negative rate. False negative rate is the ratio of users who were predicted to go on studying in the MOOC platform (predicted negative) in next week but actually drop out in the next week.

In Table 3 the experiment results with and without using user sentiments by 5-fold cross validation are presented.

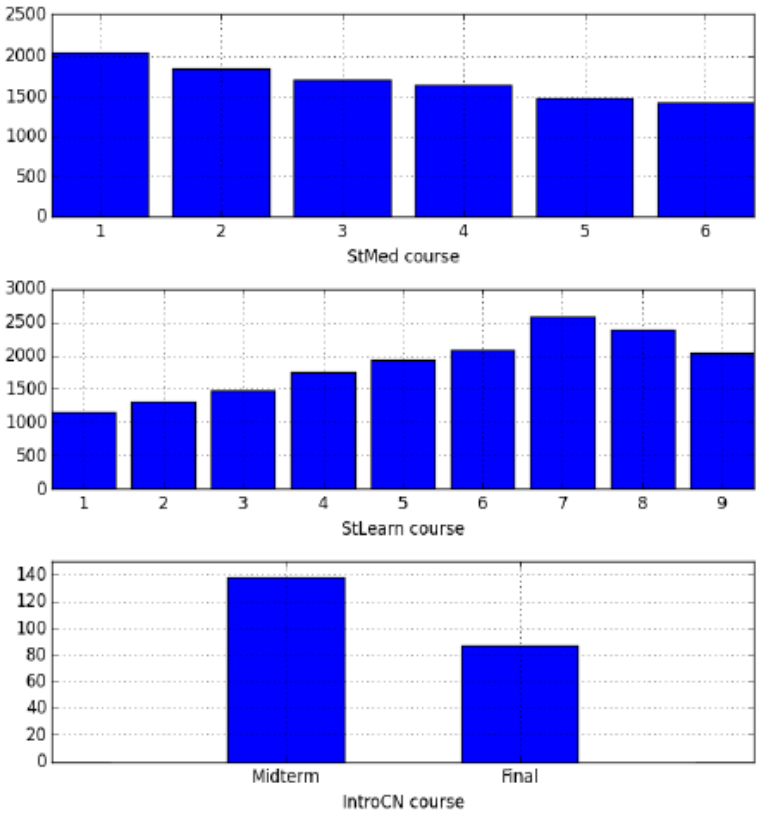

Figure 4. Distribution of users attempting each assessment in three datasets

Table 3
Comparison of Different Prediction Algorithms

Algorithms	Accuracy	False Neg.	Kappa
Balakrishnan-13 Stacking	80.1%	0.352	-
Sharkey-14	86.2%	0.462	-
Neural Network (NN)	70.6%	0.197	0.366
NN with Sentiment Analysis (SA)	75.8%	0.137	0.401
NN with SA & without Week 1	77.3%	0.131	0.433

From Table 3 we can see that our proposed algorithm provides the best Cohen's Kappa values compared with usually used prediction algorithms. If our algorithm uses user sentiments, the accuracy and false negative rate all promote greatly, it indicates the importance of user sentiments in predicting dropout. In particular, it is necessary to pay special attention to that Sharkey-14 algorithm provides the best accuracy, at the same time, it has the highest number of false negatives. According to better Kappa values, we can know that our proposed algorithm has either better accuracy or better false negative rates than other algorithms.

Because the dataset is derived from MOOCs platform which can enroll freely, there may be many tryers who just want to browse the contents of the course in the first week. So we believe that the first week is not very useful to predicting user dropout. The experimental results also prove that our algorithm without using first week's data improve greatly in various indicators.

In order to verify the importance of these extracted features, we test the influence of extracted features in predicting user dropout under the absence of each feature. The comparison of prediction influence with removal of different feature for StLearn dataset is shown in Figure 5.

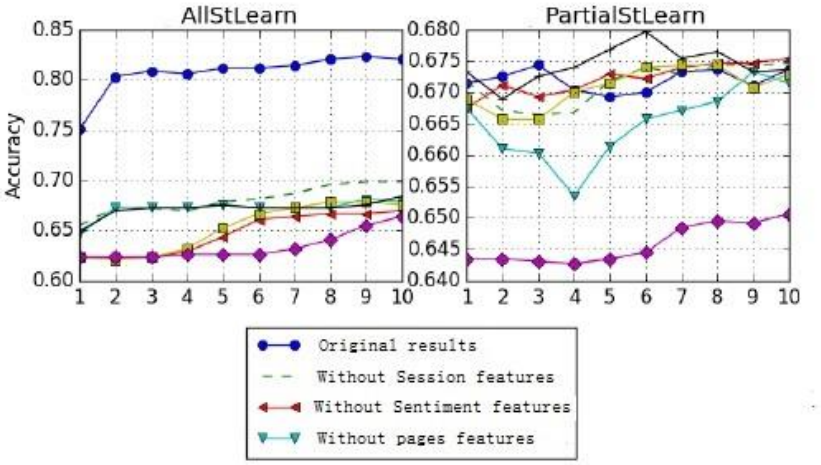

Figure 5. Prediction influence with removal of different feature

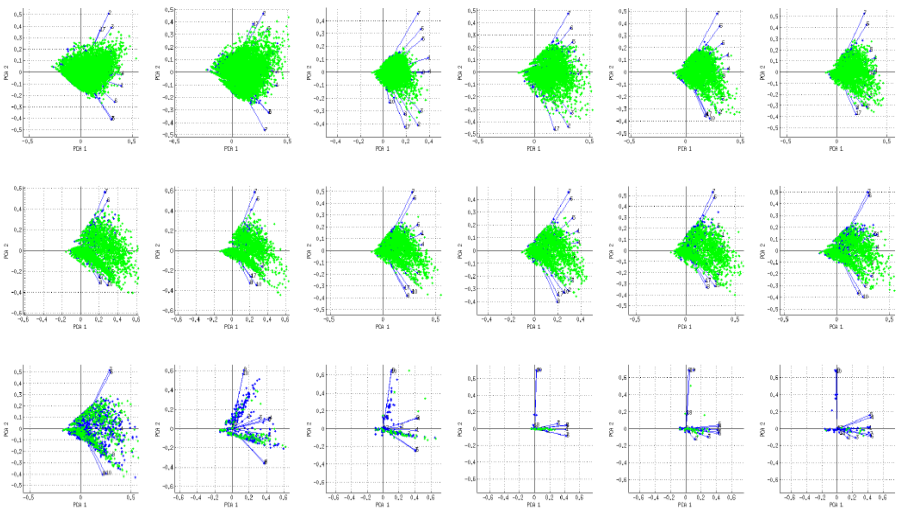

Figure 6. Experimental result of principal component analysis

From Figure 5 we can see that sentiment features have the greatest influence on the accuracy of the prediction.

Experiments based on support vector machines

In experiment, a principal component analysis (PCA) for each week where was performed, the experimental results are shown in Figure 6.

From Figure 6 we can see that the users who had dropped out could be better identified from the users who didn't drop out when the week id increases.

Then we compared our SVM algorithm to the trivial baseline algorithm (Reilly, 1997) whose prediction value was either -1 or 1. p_i represents the probability of dropout in week *I*. So, the classification accuracy of the trivial baseline algorithm can be expressed as follows:

$$acc_{trivial} := max(p_i, 1 - p_i) \quad (6)$$

The comparison of our machine learning algorithm and the trivial baseline algorithm is shown in Figure 7.

We can observe from the Figure 7 that the dropout can't be predicted very well before week 8, and then the accuracy of prediction would steadily increase. It is because that with the increment of weeks more and more features being available for the later weeks.

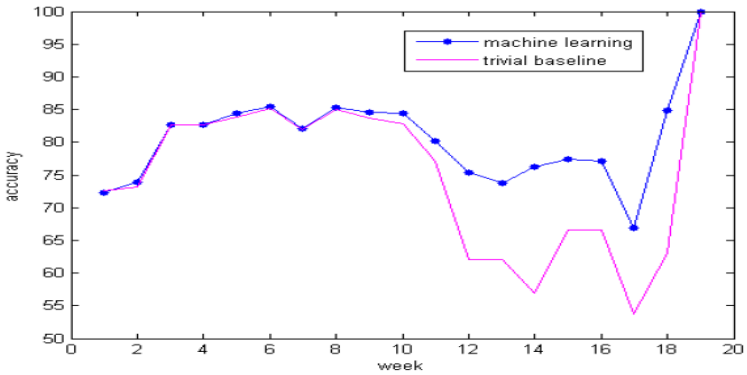

Figure 7. Experimental result of principal component analysis

Another method to analyze the importance of feature is regression model. The evaluation equation of the importance of the *i*th feature can be defined as follows:

$$I_i = \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{d=1}^{l} p_{n_S,d} f_{n_S,i} w_{d,i}}{\sum_{d=1}^{l}\left|p_{n_S,d}\sum_{k=1}^{n_F} f_{n_S,k} w_{d,k}\right|} \quad (8)$$

where N represents the number of test samples, n_S represents the number of users corresponding to the nth test sample, $f_{n_S,i}$ represents the feature value of relationship between user n_S and activity i, n_F the number of features. $p_{n_S,d}$ represents the relationship of user n_S with the dth regression model.

The importance of feature on StMed and PartialStMed dataset based on regression model were shown in Figure 8.

From Figure 8 we can see that the feature importance was completely different in the two datasets. It shows that feature extraction is closely related to datasets.

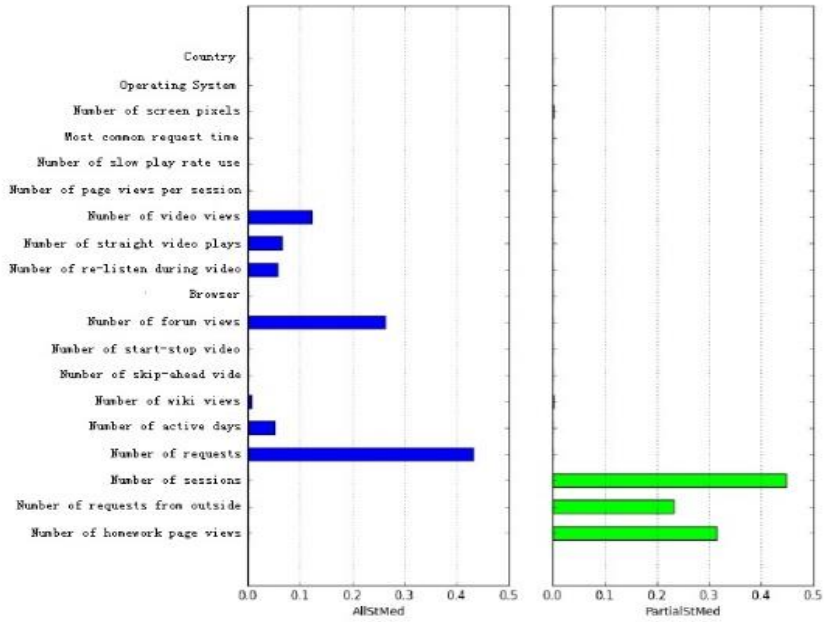

Figure 8. Experimental result of principal component analysis

Conclusion

In this paper, two machine learning algorithms for the prediction of dropout in Massive Open Online Courses from clickstream file and forum posts data. At the heart of our approaches lies in the extraction of features capturing learning behavior of users. Most existing methods provide analysis of MOOC data which indicate factors responsible for user dropout. However, our algorithms can predict users who are likely to drop-out during in the following some weeks and detect significant features from the data and achieve an increase in prediction accuracy. Through experiment, we found the prediction accuracy of our algorithms is better at the end of the course than at the beginning of the course. How to improve the prediction accuracy in the initial period of the course would be tackled in future work.

References

Ahmed, O. W., Qahwaji, R., & Colak, T. (2013). Solar flare prediction using advanced feature extraction, machine learning, and feature selection. *Solar Physics*, *283*(1), 157-175. https://doi.org/10.1007/s11207-011-9896-1

Amnueypornsakul, B., Bhat, S., & Chinprutthiwong, P. (2014). Predicting attrition along the way: The UIUC model. *EMNLP Workshop on Analysis of Large-Scale Social Interaction in Moocs*, 55-59.

Breslow, L., Pritchard, D. E., & Deboer, J. (2013). Studying learning in the worldwide classroom research into edX's first MOOC, *Research & Practice in Assessment*, *11*(8), 13-25.

Brinton, C. G., Buccapatnam, S., & Chiang, M. (2015). Mining MOOC clickstreams: On the relationship between learner behavior and performance. *Computer Science*, *64*(14), 1-7.

Cambria, E., Schuller, B., & Xia, Y. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2), 15-21. https://doi.org/10.1109/MIS.2013.30

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Hamer, R. M. (1980). The test: Fisher's test of the difference score population symmetry about zero. *Educational & Psychological Measurement*, *40*(1), 157-159.

Hughes, G., & Dobbins, C. (2015). The utilization of data analysis techniques in predicting student performance in massive open online courses (MOOCs). *Research & Practice in Technology Enhanced Learning*, *10*(1), 10-16. https://doi.org/10.1186/s41039-015-0007-z

Kloft, M., Stiehler, F., & Zheng, Z. (2014). Predicting MOOC dropout over weeks using machine learning methods. *EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in Moocs*, 60-65.

Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing pedagogical action: aligning learning analytics with learning design. *American Behavioral Scientist*, *57*(10), 1439-1459. https://doi.org/10.1177/0002764213479367

Reilly, J. R. (1997). Trivial and nontrivial baselines. *Point of Beginning*, *6*, 630-650

Rossi, L. A., & Gnawali, O. (2014). Language independent analysis and classification of discussion threads in Coursera MOOC forums, *International Conference on Information Reuse and Integration*. *IEEE*, 654-661. https://doi.org/10.1109/IRI.2014.7051952

Sinclair, J., Kalvala, S. (2016). Student engagement in massive open online courses. *International Journal of Learning Technology*, *11*(3), 218-226. https://doi.org/10.1504/IJLT.2016.079035

Sinha, T., Li, N., & Jermann, P. (2014). Capturing "attrition intensifying" structural traits from didactic interaction sequences of MOOC learners. *EMNLP Workshop on Modelling Large Scale Social Interaction in Massive Open Online Courses*, 595-599.

Zhang, H., Yang, H., & Huang, H. (2017). DBNCF: Personalized courses recommendation system based on DBN in MOOC Environment. *International Symposium on Educational Technology, IEEE*, 106-108. https://doi.org/10.1109/ISET.2017.33

Zhao, Y. Y., Qin, B., & Liu, T. (2010). Sentiment analysis. *Journal of Software*, *21*(8), 1834-1848. https://doi.org/10.3724/SP.J.1001.2010.03832