Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks

Devendra Singh Chaplot, Eunhee Rhim, and Jihie Kim

Samsung Electronics Co., Ltd.
Seoul, South Korea
{dev.chaplot,eunhee.rhim,jihie.kim}@samsung.com

Abstract. While there is increase in popularity of massive open online courses in recent years, high rates of drop-out in these courses makes predicting student attrition an important problem to solve. In this paper, we propose an algorithm based on artificial neural network for predicting student attrition in MOOCs using sentiment analysis and show the significance of student sentiments in this task. To the best of our knowledge, use of user sentiments and neural networks for this task is novel and our algorithm beats the state-of-the-art algorithm on this task in terms of Cohen's kappa.

Keywords: Student Attrition, MOOC, Educational Data Mining, Sentiment Analysis, Neural Network

1 Introduction

Massive Open Online Courses (MOOCs) have been gaining lot of interest in academia and industry in last few years. The key reasons in growing popularity of MOOCs include accessibility to every person in the world who has internet, scalability to handle any number of students with wide diversity of needs and expectations, and flexibility they provide to learners to study according to their routine. However, issues such as lack of instructor attention and absence of social learning environment, have led to high rates of attrition in MOOCs. With various unique benefits they offer over traditional classroom setting, online courses have the potential to transform future of education system, which brings out the importance of predicting student attrition in MOOCs.

With scalability, MOOCs also offer huge amounts of data of student activity, which can be utilized to train models for predicting attrition. The absence of physical learning environment makes the forums in MOOCs only medium of interaction with the instructor and peers. In this paper, we analyze the importance of sentiment analysis on these forum posts in predicting student attrition and study the effectiveness of neural network in modeling this problem.

The rest of the paper is divided into the following sections. Section 2 covers related work regarding machine learning techniques used to predict attrition and different kind of features used in them. Our algorithm is described in detail in Section 3. The experiments and results are presented in Section 4. Conclusions and future work are covered in Section 5.

2 Related Work

Recently, there have been many efforts to predict student attrition in MOOCs by extracting a wide variety of features from learner activity data and applying different machine learning approaches. [11] operationalize video lecture clickstream to capture behavioral patterns in student's activity, which is used to construct students' information processing index. [4] use feature such as number of threads viewed, number of forum posts, percentage of lectures watched, etc to predict student attrition. [12] construct a graph to capture sequence of active and passive learner activity, and use graph metrics as features for predicting attrition. [2] use quiz related (attempts and submissions) and activity related (length of action sequences, counts of various activities) features while [7] and [10] extract more than 15 features indicating learner activity and engagement from clickstream log. All these methods use variety of machine learning techniques including Logistic Regression, SVMs, Hidden Markov Models and random forest method.

There has not been much work on use of student sentiments in predicting attrition. [1] conclude that sentiment of students for assignments and course material has positive effects on successful completion of course. [14] also find correlation between sentiment expressed in the course forum posts and student drop out rate while they advice prudence against inconsistencies.

3 Proposed Algorithm

We have used click stream log and forum posts data from Coursera MOOC, 'Introduction to Psychology', which was prepared for MOOC Workshop at EMNLP 2014. The data consists of over 3 million student click logs and over 5000 forum posts. The click stream logs contain clicks made while watching video lectures and requests for viewing forums, threads, quiz, course wiki, etc. with time stamp of each click. More details about the dataset can be found in [7]. The following input features were extracted from the dataset:

- **User ID:** Unique numerical ID of the student.
- **Course Week:** Number of weeks since course has begun.
- **User week:** Number of weeks since student has joined the course.
- **Number of clicks** by the student in the current week.
- **Number of study sessions** by the student in the current week.
- **Number of course pages viewed** by the student in current week which include all pages except the video lectures.
- **Number of forum pages viewed** by the student in current week.
- **Student sentiment** of forum posts in the current week.

All the input features except Student Sentiments were indicated to be most effective by previous works mentioned in Section 2. The output of the algorithm is 1 indicating the user will drop out of the course in next week, and 0 otherwise. Note that we are predicting the exact week when the student is going to drop-out unlike [11] who predict whether student is going to finish the course or not. Our algorithm pinpoints the time when student is predicted to drop-out, which allows the course instructor and his team to take necessary steps to prevent or reduce student attrition during the course.

3.1 Sentiment Analysis

We follow a lexicon-based approach to extract sentiment from forum posts using SentiWordNet 3.0 [3] as the knowledge resource. It assigns a sentiment score to each synset in the WordNet [8]. Given the forum post, we pass the stem of each content word (using MIT JWI [6]) and its POS Tag (using Stanford POS Tagger [13]) to the SentiWordNet which returns a sentiment score. The sentiment score of the forum post is calculated by summing up the sentiment scores of all the words in the post. Fig. 1 shows a block diagram of this process.

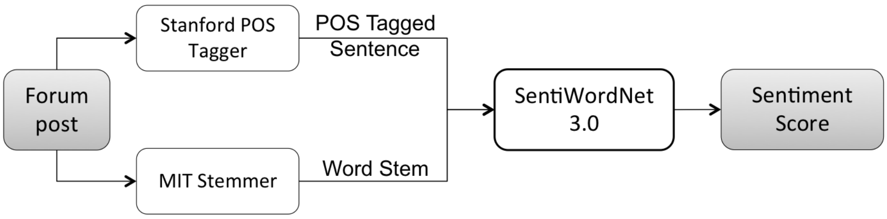

Fig. 1. Block Diagram of lexicon-based sentiment analysis using SentiWordNet 3.0.

3.2 Neural Network

Artificial neural networks are suitable to model the problem of predicting student attrition as there are a large number of inputs, and any mathematical relationship between input and output is unknown. Unlike many other machine learning techniques, neural networks are able to model the output as any arbitrary function of inputs and considered extremely robust if network structure, cost function and learning algorithm are selected appropriately through experiments. Downside of neural networks is inability to interpret the model.

We construct an artificial neural network consisting of 7 nodes in input layer: Course Week, User week, Number of clicks, Number of sessions, Number of page views, Number of forum views and Student sentiment as described above. Output layer consists of single node predicting whether student is going to drop-out in the next week. Each input feature is normalized to take values between 0 and 1. We add a hidden layer of 6 neurons in the neural network between the input and output layer. The number of neurons in the hidden layer were experimentally determined to get best possible results. Fig. 2 shows the structure of the neural network used to predict student attrition. To train the neural network, we use resilient propagation heuristic. It gave best results in our experiments among back propagation, Manhattan propagation and quick propagation.

4 Experiments & Results

In predicting student attrition, our focus is to capture all students who are going to drop-out and thus, minimizing false negative rate is important. False negative rate is the ratio of students who are predicted to stay in the course (predicted negative) in next week but actually drop out in the next week. While minimizing false negative rates, its also necessary to maintain overall accuracy so as to not produce too many false positives for the course instructor to handle.

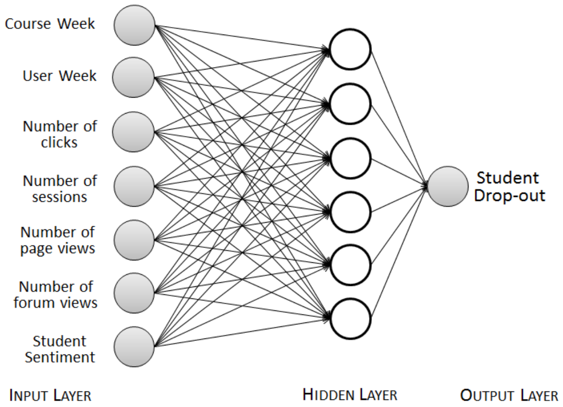

Fig. 2. The structure of neural network used to predict student attrition.

Since we are predicting whether student will drop-out in next week, our data set is highly imbalanced towards negative (will not drop-out) class. This is because a student who joins the course in 1st week, and drops out in 11th week, will have 9 negative class data points (week 1 to 9) and 1 positive class data point (week 10). Since the data set consists of student logs over 19 weeks, it is highly imbalanced with only 22.56% positive data points. Due to high imbalance in data set, we believe comparison of Cohen's kappa [5] is more suitable than comparing total accuracy directly. [9] show that Cohen's Kappa provides a unbiased estimate of performance of a classifier, and is thus much more meaningful than Recall, Precision, Accuracy, and their biased derivatives. It is more robust than total accuracy as it excludes proportion of correct predictions occurring by chance which is important in case of imbalanced data set, as a simple majority classifier would get 77.44% accuracy in this task.

In Table 1 we report our results with and without using student sentiments using 5-fold cross validation and compare them with some other approaches mentioned in Section 2. The proposed algorithm provides the best Cohen's Kappa values as compared to previous algorithms. Fall in accuracy and false negative rate when our algorithm doesn't use student sentiments indicates its importance in predicting attrition. Note that the algorithm which provides the best accuracy [10] also has the highest number of false negatives and the algorithm with best false negative rate has the lowest accuracy (Sinha-14 Baseline + Graph). This is due to imbalance in data which is explained in the following subsection. Note that the proposed algorithm has either better accuracy or better false negative rates than each of the previous algorithms, and this is reason behind better Kappa values. Since the dataset is from a MOOC which had free enrollment, there are many initial lurkers in the first week of the course who just want to browse the contents of the course. Thus, we believe predicting student attrition in first week is not very useful. Substantial improvement in performance of our algorithm without using first week's data is also shown in Table 1.

Algorithm	Accuracy	False Neg.	Kappa
Balakrishnan-13 Stacking [4]	80.5%	0.353	-
Balakrishnan-13 Cross-Product [4]	80.1%	0.442	-
Sharkey-14 [10]	**88.0%**	0.460	-
Sinha-14 Baseline + Graph [12]	62.4%	**0.095**	0.277
Sinha-14 Graph [12]	69.2%	0.157	0.365
Neural Network (NN)	70.7%	0.199	0.365
NN with Sentiment Analysis (SA)	72.1%	0.141	0.403
NN with SA & without Week 1	74.1%	0.132	**0.432**

Table 1. Comparison of accuracy and false negative rates with and without using student sentiments. The best results in each column is marked in **bold**.

4.1 Problem of data imbalance

The high data imbalance leads to biasing of the classifier towards the majority class. The problem of data imbalance in the same task is also addressed by [2] who try to solve it by oversampling the minority class, but were unsuccessful. We counter this problem by setting the boundary for classification to the ratio of drop out data points to total number of data points in the training set. This means that if the value of output neuron is greater than this ratio, then student is predicted to drop out in the next week, and vice-versa otherwise. If complete data set is used as training set, then this boundary would be 0.2256, meaning student is predicted to drop-out if value of output neuron is greater than 0.2256, rather than 0.5 by default. This adjustment to the boundary allows us to train the neural network on highly unbalanced dataset and still achieve very good recall over minority class while maintaining the overall accuracy.

The boundary is essentially a trade-off between accuracy and false negative rate. It can be adjusted to get better accuracy or false negative rates depending upon the application. This boundary can also be calculated using receiver operating characteristic (ROC) Curve.

5 Conclusion & Future Work

We propose an algorithm to predict student attrition using an artificial neural network. Sentiment analysis of forum posts is shown to be an important feature to predict student attrition in MOOCs. We also provide an approach to tackle the problem of data imbalance which can be extended to wide variety of applications in many other domains. This approach allows to find a good middle ground between accuracy and false negative rates and leads our algorithm to beat the previous algorithms in terms of Cohen's Kappa.

Most methods provide analysis of MOOC data which indicate factors responsible for attrition. In contrast, we provide a method to pin-point students who are likely to drop-out during in the following week. Since our algorithm has a very low false negative rate, it can be used in MOOCs to capture most students who are likely to drop-out in near future and take necessary actions specific to the student to prevent them from dropping out. Apart from MOOCs, the proposed algorithm can also used in smart schools using digital methods for learning and interaction, which are becoming increasingly popular in recent years.

References

1. Adamopoulos, P.: What makes a great MOOC? An interdisciplinary analysis of student retention in online courses. In: Proceedings of the International Conference on Information Systems, ICIS 2013, Milano, Italy (2013)
2. Amnueypornsakul, B., Bhat, S., Chinprutthiwong, P.: Predicting Attrition Along the Way: The UIUC Model. In: Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs. pp. 55–59. Association for Computational Linguistics, Doha, Qatar (October 2014)
3. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (2010)
4. Balakrishnan, G.: Predicting Student Retention in Massive Open Online Courses using Hidden Markov Models. Master's thesis, EECS Department, University of California, Berkeley (May 2013)
5. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20(1), 37 (1960)
6. Finlayson, M.: Java libraries for accessing the princeton wordnet: Comparison and evaluation. In: Orav, H., Fellbaum, C., Vossen, P. (eds.) Proceedings of the Seventh Global Wordnet Conference. pp. 78–85. Tartu, Estonia (2014)
7. Kloft, M., Stiehler, F., Zheng, Z., Pinkwart, N.: Predicting MOOC Dropout over Weeks Using Machine Learning Methods. In: Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs. pp. 60–65. Association for Computational Linguistics, Doha, Qatar (October 2014)
8. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM 38(11), 39–41 (Nov 1995)
9. Powers, D.M.W.: Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Tech. Rep. SIE-07-001, School of Informatics and Engineering, Flinders University, Adelaide, Australia (2007)
10. Sharkey, M., Sanders, R.: A Process for Predicting MOOC Attrition. In: Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs. pp. 50–54. Association for Computational Linguistics, Doha, Qatar (October 2014)
11. Sinha, T., Jermann, P., Li, N., Dillenbourg, P.: Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions. ArXiv e-prints (Jul 2014)
12. Sinha, T., Li, N., Jermann, P., Dillenbourg, P.: Capturing "attrition intensifying" structural traits from didactic interaction sequences of MOOC learners. CoRR abs/1409.5887 (2014)
13. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. pp. 173–180. NAACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA (2003)
14. Wen, M., Yang, D., Rosé, C.P.: Sentiment analysis in mooc discussion forums: What does it tell us. In: Proceedings of Educational Data Mining (2014)