Chen, W. et al. (Eds.) (2017). Proceedings of the 25th International Conference on Computers in Education. New Zealand: Asia-Pacific Society for Computers in Education

A Study on Prediction of Academic Performance based on Current Learning Records of a Language Class using Blended Learning

Byron SANCHEZ[a*], Xiumin ZHAO[b], Takashi MITSUISHI[c] & Terumasa AOKI[d]
[a] *Graduate School of Information Science, Tohoku University, Japan*
[b] *Center for Educational Informatics, Tohoku University, Japan*
[c] *Institute for Excellence in Higher Education, Tohoku University, Japan*
[d] *Graduate School of Information Science, Tohoku University, Japan*
*byronism@riec.tohoku.ac.jp

Abstract: In this paper, we describe a classification method that does not rely on historic data to predict changes in student academic performance, and therefore predict if a student will fail a class or not. By classifying students into groups given their grades, and extracting the common features in between them, it is possible to use those common features to predict if other students that share common characteristics will fall into the same classification groups. As well, those same common features can be used to help students improve their academic performance.

Keywords: K-Means clustering, feature selection, student performance prediction, unsupervised learning, learning analytics

1. Introduction

Learning assisting systems and educational models are ever growing areas in the world of education, and as so, topics of great interest. The following is part of a bigger 3-phased learning process designed to create an educational learning system that promotes continuous and sustainable learning. This educational environment includes a micro-learning environment and a smartphone application. Both designed to help struggling students in class by providing teacher to student feedback, and after-class learning exercises created to help with the processing of new lessons and information. One of the biggest challenges with this approach is being able to identify students in need of guidance, and even more so, predicting which ones are prone to having problems in a near future.

Nowadays, a large amount of research regarding learning analytics and the prediction of student performance using data mining is readily available. Ueno (2003) and Pardo et al. (2016) created predictive models using large amounts of historic data in order to predict student performance. Ade and Deshmukh (2016) demonstrated that by combining the expected outcomes of two well know algorithms using different voting strategies, a 3% increment to the predictive accuracy level of the next highest performing algorithm could be obtained. Bote-Lorenzo and Gómez-Sánchez (2017) and Hlosta et al. (2017) both recently made predictions using the information recollected up to the current lesson as model data to predict engagement and dropout risk, and therefore are key to this research.

In all these studies, student performance was successfully predicted at different accuracy rates using high volumes of data. Unfortunately, that is not the case in many educational scenarios, for which we apply a voting scheme combined with current data to perform predictions and make up for the low volume of data available.

2. Proposal of Prediction Method

The following method consists of a four-step linear process that utilizes K-Means clustering to create potential classification of students based on their academic performance, and then tries to predict

changes in those groups throughout the lessons using a decision tree algorithm (the summary of grades represents academic performance in the context of this paper).

First in the process, the data must be prepared to perform the analysis and clustering of such. Normalization was performed using the Z-Score for scaling, and new features were added to tell us how their values have changed throughout the lessons. The following 2 formulas were used for each feature:

$$PI = \frac{v_k - v_{(k-1)}}{\min(v_k, v_{k-1})} * 100, k > 1 \quad (1) \qquad API = \sum_{k=2}^{n} \left(\frac{v_k - v_{(k-1)}}{\min(v_k, v_{k-1})} * 100 \right) \quad (2)$$

In formulas (1) and (2), v_k and v_{k-1} are the current and last lesson's grade averages.

The second step consists of classifying students by their academic performance in class using only features that represent grade performance. By performing clusters of students based on their academic performance, it is possible to separate them by how similar their grades behaved throughout the lessons taken. We used a K-Means clustering algorithm, with four centroids and a K-Means++ initialization step, using the Euclidean Distance as a distance measurement.

In the third step, we use different combinations of the previously extracted features in order to create models that best identify the current data, and then use those models to predict the next classification of students. Models were created by combining a set of features, an amount of past lessons of information, and whether they include the previous cluster results or not. We rank the models by their accuracy score and choose the models that have the highest values as prediction models for the next lesson.

The final step of the process uses the before created prediction models, and applies the data from the last lesson as input in order to predict the classification groups of the next lesson. Given that there are many models possibly with the same accuracy value, we choose all models that have a value above the 3rd highest accuracy score, and created a voting scheme were the most voted group by all models is the decided group. We proceed to do this entire process for each lesson predicted.

3. Experiment and Results

The data used belongs to a Chinese - Japanese language course. The class contains 32 students, and 12 lessons. Grades are scored over 100 points, and students may attempt a same exercise an unlimited amount of times. The features obtained from the data are: "Score", "Finished attempts", "Abandoned attempts", "Time between exercise release and first attempt", "Avg. time to complete an attempt", "Time between first and last attempt".

Table 1: Prediction Accuracy comparison

	Top scoring model	**Voting scheme**
Lesson 6	0.8125	0.90625
Lesson 7	0.78125	0.96875
Lesson 8	0.875	0.90625
Lesson 9	0.8125	0.90625
Lesson 10	0.6875	0.9375
Lesson 11	0.875	0.875
Lesson 12	0.8125	0.9375
Average total	**0.8203**	**0.9178**

When using the newly added features in combination with the existing data for student grades, the clustering algorithm correctly identified four groups of students. Figure 1. shows the academic performance of each student as a line, and to what classification group the student belongs to as color (data was used up to lesson 5 as an example). Following steps 3 and 4, we obtained a list of the highest scoring attribute/parameter combination and used those to create prediction, and then used the models to predict on what group to classify each student on each lesson. Table 1 shows the result comparison

between using the voting scheme and just the top scoring model (the first lessons haven been omitted because the values do not change significantly).

The results improve when using a voting scheme in most cases. We also found that using the top 3 highest accuracy values, gives us the best results. Models used after the 3rd do not increment the accuracy significantly. These results were omitted.

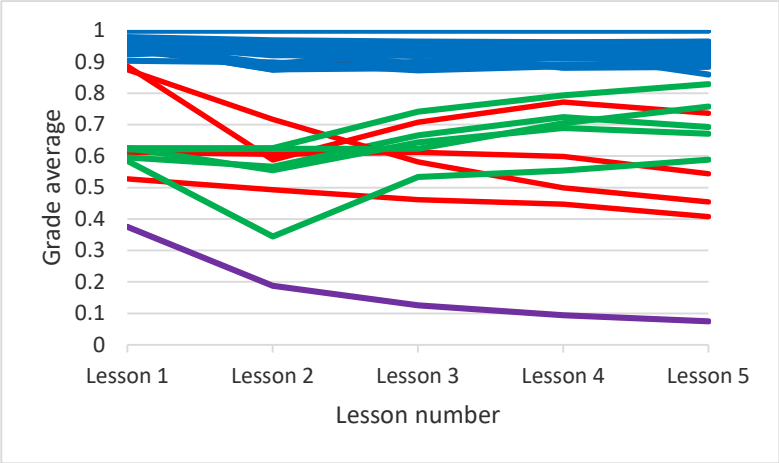

Figure 1. Student academic performance groups

4. Conclusions

The results found in this paper indicate that by applying this method to a set of features, we are able to predict with a high accuracy the changes to student performance in a class with limited amount of data. By identifying/predicting which groups have declining or improving grades, and helping students in those groups, overall academic performance can be monitored, and most of all improved with proper management. As future work, we shall try different prediction algorithms and different data from another year of the course, with the purpose of improving accuracy and validating the method.

Acknowledgments

This work was supported by the JSPS KAKENHI grant, numbers JP15K02709 and JP15K01012. We would also like to thank the members of both Mitsuishi Lab and Aoki Lab at Tohoku University for their insight and input into the topic since the beginning.

References

Ueno, M. (2003). On-Line Statistical Outlier Detection of irregular learning processes for e-learning. In D. Lassner & C. McNaught (Eds.), Proc. World Conference on Educational Media and Technology 2003 (pp. 227-234). Association for the Advancement of Computing in Education (AACE).

Pardo, A., Mirriahi, N., Martinez-Maldonado, R., Jovanovic, J., Dawson, S. & Gašević, D. (2016). Generating actionable predictive models of academic performance. In *Proc. 6th Int. Learning Analytics & Knowledge Conference* (LAK '16). ACM, New York, NY.

Ade, R. & Deshmukh, P.R. (2014). Classification of students by using an incremental ensemble of classifiers. *Proc. 3rd Int. Conf. Reliability, Infocom Technologies and Optimization*, Noida, pp. 1-5.

Bote-Lorenzo, M.L., & Gómez-Sánchez, E. (2017). Predicting the decrease of engagement indicators in a MOOC. In *Proc. 7th Int. Learning Analytics & Knowledge Conference* (LAK '17). ACM, New York, NY, USA

Hlosta, M., Zdrahal, Z., & Zendulka, J. (2017). Ouroboros: early identification of at-risk students without models based on legacy data. In *Proc. 7th Int. Learning Analytics & Knowledge Conference* (LAK '17). ACM, New York, NY, USA.