JSiSE Research Report
vol.32,no.5(2018-1)

ブレンディッドラーニング型授業における学習パフォーマンス予測手法の検討

サンチェス バイロン[*1], 趙 秀敏[*2], 三石 大[*3], 青木 輝勝[*4]
*1 東北大学大学院情報科学研究科, *2 東北大学高度教養教育・学生支援機構,
*3 東北大学教育情報基盤センター, *3 東北大学未来科学技術共同研究センター

A Case Study on Prediction of Student Performance in a Blended Learning Class

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] New Industry Creation Hatchery Center, Tohoku University, Japan

We describe a prediction method that does not rely on long-term historic data to predict changes in student academic performance. By using a descriptive model of a previous period, and applying it to current data, we aim to predict the results of the current period, which has yet to pass, using the information of students who have already submitted their information into the system. We also incorporate the use of a voting method to increase accuracy.

Key Words: Learning analytics, Educational Data Mining, Supervised Learning, Student performance prediction

1. INTRODUCTION

The advancements in education have always been a high priority and a topic of much interest in learning analytics. This research is part of a bigger 3-phased blended learning process designed to create an educational learning system that promotes continuous and sustainable learning. After taking a lesson in a classroom, students then proceed to practice the previously learned contents online, and then evaluate their performance in the next class. This educational environment includes a micro-learning based smartphone application 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 likely to have problems in a near future. Students might be failing since the beginning of class, or their grades might be slightly decreasing as the lessons progress, and these are problems that are hard to control even in small classes of 30 students when our resources are limited. Even more so, students that have low academic performance are known to drop out of class because of this. Therefore, we find ourselves in need of way of identifying students that are having academic performance problems and are at risk of failing or even not finishing a class.

To explain a bit further, most of the existing research on Educational Data Mining (EDM) focuses on how to predict future results based on historic data, which is available from the records of past courses. However, past data does not always reflect upon new students, given that the contents of a class, the teacher's instruction, and students' way of thinking might change.

Another obstacle that we must tackle is that the amount of data available is not substantial and therefore it becomes quite hard to predict with a high level of accuracy which students are in need of assistance and which ones are not. As more data becomes available, the probability of discovering hidden patterns increases. But usual language classes are relatively small, so it is hard to adopt some methods with statistics.

The method we proposed in this paper uses recently obtained data as training data, in order to predict if a student's academic performance will decline. In addition, in order to make up for the small amount of data, it creates different features out of the existing data and different representations of it. Additionally, a sampling method and voting scheme are also used in order to increment the recall value of predictions.

The rest of the paper is organized the following way. Section 2 discusses related work, which served as basis for this research. Next, section 3 explains the overall functionality of the method proposed along with the ordered steps of execution. Section 4 presents the results of the experiment along with a description of the data used, as well as some of the problems encountered. Finally, section 5 includes the main conclusions of the paper and presents future works.

2. RELATED WORK

We are able to find a lot of research regarding prediction of student performance and drop out prediction using data mining. A large percentage of that research focuses mostly on the use of predictive algorithms and historic data to obtain the desired results.

On this topic, Ueno designed a learning system called Samurai, to help detect outliers using historic data and a Bayesian predictive distribution(1). The research focused on finding students with irregular e-learning processes using prior knowledge of the response time characteristics of each content, and the learner's ability parameters. Pardo et al. used recursive partitioning and a list of known available actions in an e-learning system to predict academic performance(2). By analyzing a large number of numeric features obtained from the interaction with the system, the method automatically selected the most robust according to their performance. Ade and Deshmukh applied an ensemble of different classifiers in order to further increment the accuracy of predictive models in a dataset of over 250 samples and 10 different attributes(3). By combining the expected outcomes of two well know algorithms using different voting strategies, a combination of Naive Bayes and k-Star algorithms showed promising results with a 3% increment to the predictive accuracy level of the next highest performing algorithm it was compared to. In these researches, high predictive accuracy levels were achieved in their respective contexts using a large amount of historic data and features. Now although these papers are not key to this research, they do provide many of the tools we used such as the implementation of a voting scheme and feature creation.

Bote and Gómez recently predicted whether the engagement of students in a MOOC would increase or decrease by analyzing their behavior and the actions performed in the system(4). He used the data that became available during the course to create models for upcoming classwork. By creating different features focused on the actions performed on videos, exercises and assignments; using a CFS method for feature selection; and an SGD algorithm for classification, his team was able to detect disengagement of students at an early stage. Hlosta et al. also identified students at risk of failing a course using a model based not on legacy data, but on data recently obtained(5). Ouroboros (the method's name) is a self-learning approach that uses the patterns from student who have just submitted an assignment in order to predict if other students will also submit. Both these researches address the topic of prediction using non-legacy data (data recently collected) and do so successfully using information obtained from a VLE, and therefore are key to this research. Yet still, they both point out a particular issue that is, a lot of data is necessary.

Up to this point, many of the tools required to create predictions based on current data have been already introduced and proven successful in their respective contexts. The problem now is the lack of data available, or in other words, working with small data sets. Maharani et al. generated new synthetic data by considering k-nearest neighbors similarities between features(6). This way a class of 63 students was incremented to 225 instances, where the artificial data behaved as neighbors of the existing data. This solves the class imbalance phenomenon, where more students with regular academic performance exists than those with high or low performance do.

The uniqueness of the method proposed in this paper is that it searches to address the issue of predicting student performance when no historic data and a small data set are available.

3. PROPOSAL OF PREDICTION METHOD

3.1 Definition of academic performance

The general purpose of the method is to determine if a student's AP (academic performance) shall decline or not. In order to do this, we first define what AP is in this context. We have defined AP as the combination of the likelihood of a student's grades declining and his usage of the learning material (the mobile application) also declining. A student's grade is considered to be declining if his deviation from the average score/grade for a unit has declined over time. A student's usage of the application is also considered to be declining if he is not performing all of the exercises per unit, as for each one of the exercises has been designed to cover a specific aspect of the unit and they should therefore all be attempted.

Using these concepts, we define three types of student prediction results as shown in **Table 1.**

Table 1: Types of student predictions

Score declining	App usage declining	Result
YES	YES	DANGER
YES	NO	CAUTION
NO	YES	CAUTION
NO	NO	SAFE

The aim of the method is to increment the recall value for the students that are in either DANGER or CAUTION, with special attention to the former.

3.2 Prediction Method Overview

The method is based on two main concepts: prediction using machine learning & recent data, and the expansion of small data. The use of current data and machine learning, serve the purpose of predicting the result of the current week, under the premise that student behavior gives the same results in the form of a pattern. The expansion of small data serves the purpose of finding hidden relationships in the limited amount of data available. By combining them, we seek to predict future student actions by finding hidden patterns in the expanded small amount of data that we have, and use these patterns to infer on a student's next action. **Figure 1** is a general representation of the entire method.

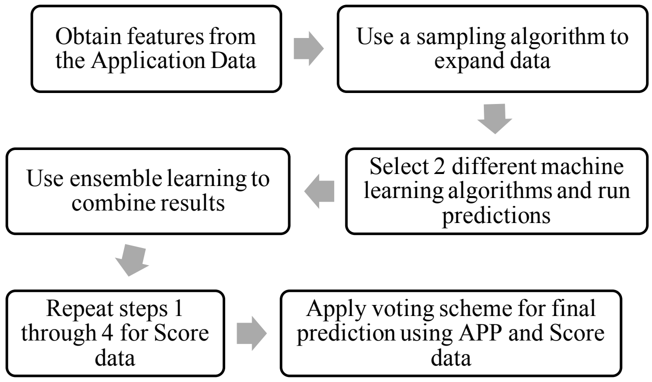

Figure 1: Overview of process

3.3 Data Analysis and Sampling

First in the process, the data must be prepared to perform the analysis and clustering of such. In order to do this, data must be normalized, and new features that best represent the current data must be created. Normalization was performed using the Z-Score for scaling using the mean average and standard deviation per number of exercises performed of each lesson.

The data used for prediction is obtained from two different sources: application data and score data.

3.3.1 Application data

The application data was obtained from the use of the mobile application (KOTOTOMO) for Chinese language learning. The data here is divided into units, and each unit is divided into 4 different exercises. **Table 2** has a list of the features per exercise.

Table 2: APP features per exercise

Description of features
Number of attempts to do an exercise
Date of first attempt
Date of last attempt
Time difference between first and last attempt
Average duration of attempts in minutes
Number of attempts completed
Whether the exercises were attempted or not

Features that describe the entire unit are in **Table 3.**

Table 3: APP features per unit

Description of features
Whether all exercises were attempted or not
The amount of the addition of all attempts

3.3.2 Score data

The score data was obtained from the course grades for short quizzes provided by the teacher, not from the mobile application. The features that describe the data provided by the teacher are in **Table 4** (for each unit):

Table 4: Score features per unit

Description of features
Short quiz grade
Deviation from class average grade
Whether or not the score deviation increased

3.3.1 Sampling

Sampling was performed using the SMOTETomek algorithm for minority over-sampling (SMOTE[7]) and cleaning (Tomek Links), balancing the amount of positive and negative samples available to train the learning algorithm. Sampling is performed in order to make up for the small amount of data available from the sources.

3.4 Ensemble learning

We then proceed to predict if the score and app usage will decline, individually. In order to increment the accuracy of prediction, we used an ensemble learning method with two known prediction algorithms (Multi-Layer Perceptron and Random Forest) and a majority voting scheme in order to determine a student's prediction type.

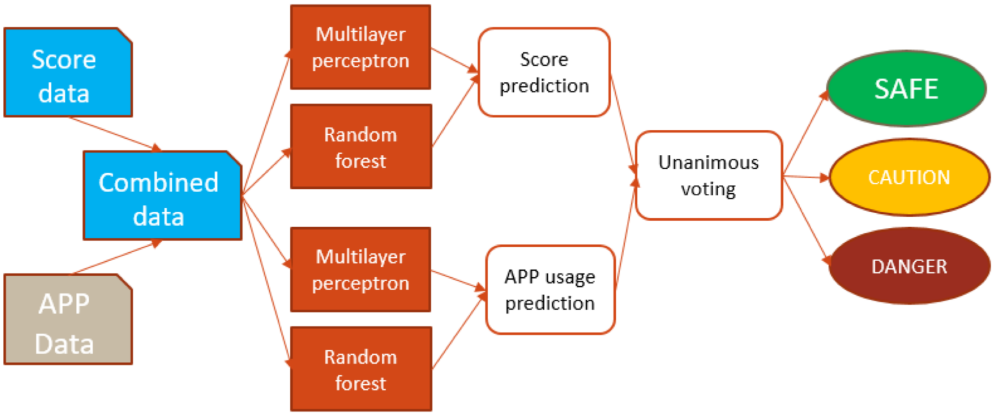

Figure 2: Overview of prediction model

As seen in **Figure 2**, the data provided from both source is combined into "combined data" to perform the predictions, but the results are predicted individually ("score prediction" and "app usage prediction").

3.5 Calculating the results

Finally, we proceed to combine the before obtained results in order to predict the final AP for each student, determining if he is classified as a student in DANGER, CAUTION or SAFE. In order to prove our results, we used a 10-Fold cross validation method using the information for the next unit as our ground truth source

3.6 Training the prediction algorithm

In order to train the prediction algorithms, we defined a prediction model based on the information from previous units, using the results of the current unit of students that have already submitted their information. A prediction model is defined as a descriptor of a mathematical relation between inputs and outputs.

$$M_w = MLA\langle C_w | S_{w-1 \ldots w-i}, A_{w \ldots w-i} \rangle, i = \{1..3\} \quad (1)$$

In equation 1, MLA is a Machine Learning Algorithm, M_w is the model, C_w is the result, and S is the score information, A is the app usage information and w is the unit number. Therefore, if we want to explain the attempt results C_{w+1}, given that we have the previous information, we can do:

$$C_{w+1} = MLA\langle M_w(S_{w \ldots w-i+1}, A_{w+1 \ldots w-i+1}) \rangle \quad (2)$$

4. EXPERIMENTATION

4.1 Data

In this study, we ran our proposed method with the data belonging to Japanese-Chinese language course. 7 classes with a total of 280 students, and 4 units were used for this case study. A student may attempt the same exercise many times without limit on the mobile application. Students are required by the teacher to use the application and have a score penalization for not using it at all. The score penalization is not considered for the prediction, but does serve as motivation and to explain behavior patterns.

4.2 Results and Discussion

The initial validation of the need of a machine-learning algorithm is to rule out the manual analysis of the data, since special behaviors of the data are not easily visible, and large amounts of data are not easily readable. The proposed method has an overall high accuracy and even higher recall for students in need of assistance because of decaying AP. **Table 5** shows the resulting values for prediction.

Table 5: Model accuracy results

Prediction	Type of acc.	Accuracy value
Score	General	**0.83**
	Negative (Recall)	0.775
APP Usage	General	0.74551971
	Negative (Recall)	0.84
Ensemble	General	0.79928315
	Negative (Recall)	**0.88038278**

The ensemble recall value is higher than any of the individual results for either score or APP usage, which is the ultimate goal of this prediction. **Table 6** shows the actual and predicted number of students in each status (This is the prediction for unit 4).

Table 6: Model predicted and actual results

	PREDICTED	ACTUAL
DANGER	65	62
CAUTION	150	147
SAFE	64	70

This model maximizes the amount of DANGER and CAUTION successful predictions, and minimized the amount of wrongly predicted DANGER and SAFE students. The following table shows the counts per error type:

Table 7: Error type count

	Type of error	Count
E1	DANGER student predicted as SAFE	0
E2	SAFE predicted as CAUTION, OR CAUTION predicted as SAFE	25
E3	SAFE predicted as DANGER	0
	TOTAL amount of errors	25

The results of this method were compared to other prediction methods in order to prove its efficacy for prediction. The information in the following chart shoes the accuracy value and the amount of errors of each approach. We seek to increment the recall value while minimizing the amount of E1 and E2 errors (Error types in **Table 7**).

Table 8: Model results comparison

Method	Type of accuracy	Accuracy	E1	E2	E3
Naïve Bayes	General	0.76	4	43	3
	N. Recall	0.77			
Neural network	General	0.77	3	33	3
	N. Recall	0.82			
Random forest	General	**0.80**	2	**23**	6
	N. Recall	**0.88**			
Our method	General	0.79	**0**	25	**0**
	N. Recall	**0.88**			

The method is also necessary because of its ability to find unusual learning behaviors in the data. The following graphs show an example of students with similar learning patterns that have different results.

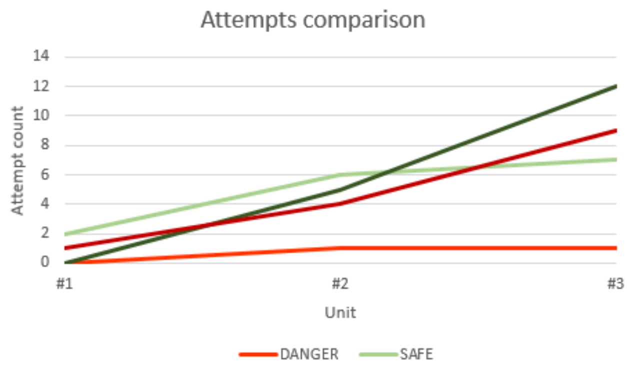

Figure 3: Usage comparison similar patterns

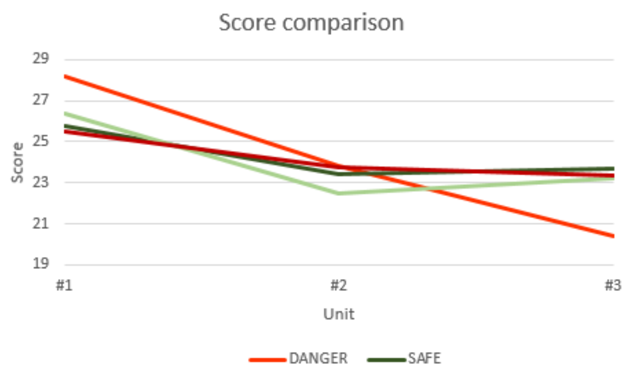

Figure 4: Score comparison similar patterns

In **Figures 3 and 4** both red lines are students that will be predicted successfully as in DANGER in unit #4, and both green lines are students that will be predicted as SAFE in unit #4. By just looking at the current context, it seems as if the dark red and dark green lines behave similar, but they are in fact predicted differently.

Figures 5 and 6 show the actual result of unit #4.

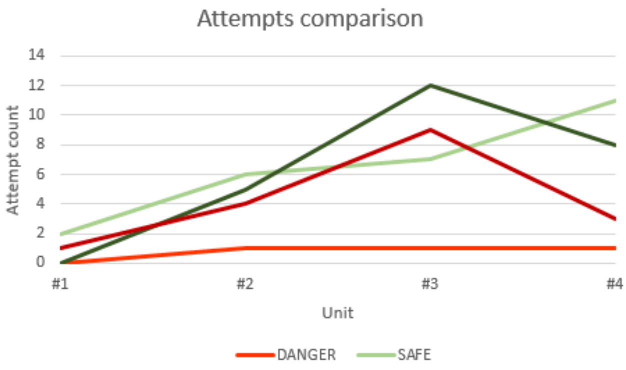

Figure 5: Usage comparison unit #4

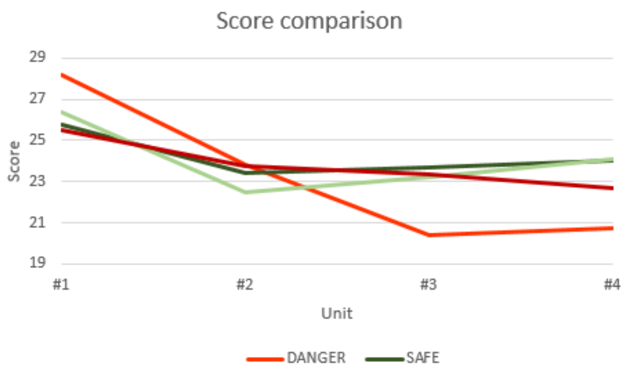

Figure 6: Score comparison unit #4

The student represented by the dark green line, in fact raises his score grade, and the dark red line decreases, even though they appeared to have the same pattern. This serves as an example of a pattern difficult to interpret by plain analysis.

5. CONCLUSIONS

The results found in this paper indicate that by applying this method to a set of features created from a limited amount of data, we are able to predict with better accuracy than other simpler methods, the AP of a student. Given that the main objective is to identify student in risk of decreasing AP, we increment our recall value at the expense of precision, without affecting the latter value significantly.

We proved that the method is successful at predicting results when similar patterns are present, given that the algorithms uses all the available features for prediction, and not just actual scores and attempts at an exercise. The patterns are difficult to see by simple visualizations and are therefore critical for successfully assisting students in need.

We classified the prediction of students in 3 types, and focus on predicting DANGER and CAUTION students in order to help as many students in need as possible. An important factor considered is the minimization of E1 and E3 errors when predicting.

As a future work, we are currently working on clustering the resulting prediction groups in order to separate them by common features, and then identify those features in order to discover unusual learning patterns in the students. We also need to address the problems of outliers and detect them in order to exclude them from the analysis and prediction. This is important because given the small amount of data available, a single exception can distort the predictive patterns found by the machine-learning algorithm.

ACKNOWLEDGMENTS

This work is supported by the JSPS KAKENHI grant, numbers JP15K02709 and JP15K01012.

REFERENCES

(1) Ueno, M. (2003). On-Line Statistical Outlier Detection of irregular learning processes for e-learning. In D. Lassner & C. McNaught (Eds.), Proceedings of EdMedia: World Conference on Educational Media and Technology 2003 (pp. 227-

234). Association for the Advancement of Computing in Education (AACE).

(2) Abelardo Pardo, Negin Mirriahi, Roberto Martinez-Maldonado, et al. 2016. Generating actionable predictive models of academic performance. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16). ACM, New York, NY.

(3) R. Ade and P. R. Deshmukh, "Classification of students by using an incremental ensemble of classifiers," Proceedings of 3rd International Conference on Reliability, Infocom Technologies and Optimization, Noida, 2014, pp. 1-5.

(4) Miguel L. Bote-Lorenzo and Eduardo Gómez-Sánchez. 2017. Predicting the decrease of engagement indicators in a MOOC. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17). ACM, New York, NY, USA.

(5) Martin Hlosta, Zdenek Zdrahal, and Jaroslav Zendulka. 2017. Ouroboros: early identification of at-risk students without models based on legacy data. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17). ACM, New York, NY, USA.

(6) K. Maharani, T. B. Adji, N. A. Setiawan, et al. "Comparison analysis of data mining methodology and student performance improvement influence factors in small data set," 2015 International Conference on Science in Information Technology (ICSITech), Yogyakarta, 2015, pp. 169-174.

(7) Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, et al. 2002. SMOTE: synthetic minority over-sampling technique. *J. Artif. Int. Res.* 16, 1 (June 2002), 321-357.