Yang, J. C. et al. (Eds.) (2018). Proceedings of the 26th International Conference on Computers in Education. Philippines: Asia-Pacific Society for Computers in Education

Prediction of Students' Academic Performance based on Tracking logs

Anna Y. Q. Hung[a], Jian-Xuan Weng[b], Jeff C. H. Huang[c], Owen H. T. Lu[d], Bin-Shyan Jong[e] & Stephen J. H. Yang[f*]

[abdf*]*Computer Science & Information Engineering, National Central University, Taoyuan City, Taiwan*

[c]*Computer Science & Information Engineering, Hwa Hsia University of Technology, New Taipei City, Taiwan*

[e]*Information & Computer Engineering, Chung Yuan Christian University, Taoyuan City, Taiwan*

*jhyang@csie.ncu.edu.tw

Abstract: In this paper, we predict students' academic performance based on tracking log of students' learning activities. We compare the prediction of six datasets from Kyoto University (KU), National Central University (NCU), and Chung Yuan Christian University (CYCU) by eight classification models. We use the evaluators of accuracy, recall, precision, F1-score, and Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC). According to the prediction results, we found that sample size and feature category influence the prediction performance of classification. We also found that the significant features based on Pearson correlation analysis have greatly influence on the prediction performance of classification.

Keywords: classification, academic performance

1. Introduction

The mechanism of predicting and classifying students' performance is very important for promoting students' success in learning (Lu, Huang, Huang, & Yang, 2017; Lu, Huang, Huang, Lin, Ogata, Yang, 2018; Romero, López, Luna, & Ventura, 2013). Finding at-risk students through predicting students' performance can help teachers give timely interventions to students to improve their success. From previous studies (Asif, Merceron, & Pathan, 2014; Oladokun, Adebanjo, & Charles-Owaba, 2008; Lu et al., 2018; Yoo & Kim, 2014; Romero et al., 2013), machine learning methods such as Naive Bayes (NB), Decision Tree (DT) and Neural Network (NN), Support Vector Classification (SVC), Logistic Regression (LR) and Random Forest (RF) are the common used classification algorithm to predict students' learning performance. Therefore, this study applied Gaussian Naive Bayes (GaNB), SVC, Linear-SVC, LR, DT, RF, NN, and Extreme-Gradient Boosting (XGBoost) algorithms to construct student classification for academic performance.

The goal of this paper is to build students' academic performance prediction model by using various classification methods for different datasets which were recorded students tracking logs. We have compared six datasets from Kyoto University (KU), National Central University (NCU), and Chung Yuan Christian University (CYCU). To measure the prediction performance of the applied eight classifications, this study uses the evaluators of accuracy, recall, precision, and F1-score. To improve the prediction performance of classifications, this study also discuss the factors influence on prediction performance for the six datasets. Therefore, the research questions in this study are proposed as following.

- RQ1: Can we predict students' academic performance based on different categories of students' tracking logs?
- RQ2: Which classification methods are suitable for predicting students' academic performance?

2. Literature Review

2.1 Classification methods for predicting students' academic performance

Finding at-risk students in education, classification algorithm is one of the most frequently used methods in machine learning. Classification methods can be generally divided into four types which consisted of statistical classification, NN (McCulloch, & Pitts,1943), probabilistic classification, and vector space based classification. LR (Cox, 1958) is a statistical classification for constructing binary classification to deal with linear or nonlinear data. GaNB and NB (John, & Langley, 1995) are statistical classification. The vector space based classification generally includes SVC and Linear-SVC algorithms. SVC is a Support Vector Machine (SVM) for classification (Cortes & Vapnik, 1995). DT (Quinlan, 1983), RF (Breiman, L., 2001) and XGBoost (Chen, & Guestrin, 2016) are tree based classifications.

In general, the evaluators of classification performance include accuracy, recall, precision, F1-score, and Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC). The evaluators of accuracy, recall, precision and F1-score are derived from confusion matrix. The other evaluator, AUC, is derived from ROC curve. AUC is the area under the ROC curve. The value of AUC is range from 0 to 1. The value of AUC near to 0.5 indicated that the classification similar to random guess. The classification with higher value of AUC implied the better prediction performance.

3. Method

3.1 Datasets Description

This study aims to build students' classification based on tracking logs of learning. The learning environments of KU, NCU, and CYCU datasets are ebook reading in BookRoll, online learning in Open edX, and online learning in iLearning, respectively. For exploring the students' ebook reading behavior, Ogata, Yin, Oi, Okubo, Shimada, Kojima, & Yamada (2015) and Flanagan & Ogata (2018) have briefly describe the students' reading actions in BookRoll. For KU, the datasets (https://lab.let.media.kyoto-u.ac.jp/icce2018la/) of case 1.1 and 1.2 collect students' ebook reading actions in BookRoll for two different courses. For NCU, the datasets of case 2.1 and 2.2 collect students' online learning actions in Open edX for the courses of university Calculus (case 2.1), and high school Calculus (case 2.2). For CYCU, the datasets of case 3.1 and 3.2 were collected students' online learning actions in iLearning for the courses of System Programming (case 3.1), and Operation System (case 3.2).

Table 1 shows the briefly description of six datasets. For extracting ebook reading features from students' clickstream, Yamada, Oi & Konomi (2016) have extracted 19 features to represent students' reading actions. Based on these features (Yamada et al., 2017), this study extracted 15 features for KU datasets of case 1.1 and 1.2. The objective of this manuscript aims to identify students' class which consist of high achievement and low achievement classes. In general, student got score lower than 60 should be belonged into low achievement class. But in KU datasets, there are only 10 and 8 students got the score lower than 60 for case 1.1 and 1.2, respectively. To balance the number of students for high achievement and low achievement classes, the label of student was set as high and low when students' score higher than or equal to 80 and score lower than 80, respectively. Two datasets collected from NCU have highest number of features. There are more than 100 students in the three courses from NCU and CYCU. Among six courses, the two courses in KU have the lowest number of features and students. For the descriptive statistics of academic performance (score) in the six cases, the mean score of case 2.2 extracted from NCU were near to 30 which is the lowest value. This is because that there are too many students got lower score. In contrast, the mean score of case 1.1 and 1.2 extracted from KU were near to 70 which are the highest value in the six datasets. It seems that the score of most students are more than 60. For the case 3.1 and 3.2 extracted from CYCU, the mean scores of two cases are near to 60 which fall in the middle area for the six cases.

Table 1

Descriptive statistics of academic performance (score) for six datasets

University	Case	Numbers of Students	Numbers of Features	Mean of score	Std. of score	Number of high / low
KU	1.1	53	15	78.01	25.12	35 / 18 (score >=80 / <80)
	1.2	55	15	77.64	18.88	32 / 23 (score>=80 / <80)
NCU	2.1	59	55	66.64	16.12	38 / 21 (score>=60 / <60)
	2.2	128	55	36.05	27.57	28 / 100 (score>=60 / <60)
CYCU	3.1	135	16	59.68	23.57	84 / 51 (score>=60 / <60)
	3.2	125	16	62.37	24.89	88 / 37 (score>=60 / <60)

3.2 Procedure of students' academic performance classification

For the students' academic performance classification, the label of student can be considered as high achieve and low achieve classes, the features were extracted from students' tracking logs during learning. This study aims to construct students' academic performance classification which consisting of data pre-processing, constructing classification, and evaluation phases. The main tasks of data pre-processing phase include of data integration and data normalization. Data integration focuses on integrating various learning environments to construct learning datasets. Take NCU datasets as an example, the learning environments include of Open edX, MapleTA, and traditional classroom. Data integration aims to integrate the students' learning data derived from Open edX, MapleTA, homework and quiz scores. This study extracted 55 features to represent students' learning actions in NCU two datasets. Data normalization aims to redefine or transform the range of data value in a smaller and specific range. This study applied z-score normalization to the proposed features for students' academic performance classification.

The constructing classification phase aims to construct students' academic performance classification. This study has applied GaNB, SVC, Linear-SVC, LR, DT, RF, NN, XGBoost classifiers to build the students' academic performance classification. The applied 8 classification methods were brief description in section 2.1.

The evaluation phase focuses on measuring the classification performance of the proposed classifiers. We have applied the evaluators which include of accuracy, precision, recall, F1-measure and AUC to evaluate the performance of the students' academic performance classification. The cross-validation mechanism proposed by Golub, Heath, and Wahba (1979) aims to evaluate the prediction performance. According to the requirements of the 5th ICCE workshop on Learning Analytics (LA) & Joint Activity on predicting student performance (https://lab.let.media.kyoto-u.ac.jp/icce2018la/), we applied the average of 3-fold cross-validation that have been run 10 times to evaluate prediction performance.

4. Results and Discussion

4.1 Performance results of students' academic performance classifications

As showed in Table 1, the KU datasets consisted of case 1.1 and case 1.2 for two different courses. Each dataset has collected students' score and students' clickstreams during reading ebook by using BookRoll system. Students' clickstream file records the logged activity data from students' interactions with the BookRoll system. Students' score file records the final score for each student. Two NCU datasets have collected students' online learning actions in Open edX, students' online practice actions in Maple TA, and scores of homework and quiz. The CYCU datasets consisted of case 3.1 and 3.2 for System Programming and Operation System courses at CYCU. Two CYCU datasets have collected students' online learning actions in iLearning, and scores of homework, quiz, and project. Table 2 shows the performance of students' academic performance classification for KU datasets (case 1.1 and case 1.2), NCU datasets (case 2.1, case 2.2), and CYCU datasets (case 3.1 and case 3.2). The NCU datasets consisted of case 2.1 and 2.2 for one university Calculus course and one higher school Calculus course.

To reply RQ1 (Can we predict students' academic performance based on different categories of students' tracking logs?) from Table 2, the best values of accuracy were range from 0.65 for case 2.1 to 0.96 for case 2.2 in the six datasets. From Table 2, the case 2.2 of NCU datasets achieved the highest classification performance of 0.96. But the case 2.1 in NCU datasets obtained the lowest classification performance of 0.65. According to Table 1, the number of students in case 2.1 is relatively small in the comparing six datasets. In contrast, more students were collected in the case 2.2. This maybe the reason for the classification of case 2.1 obtained the lowest performance, but the classification of case 2.2 achieved the highest performance. Similarly, due to both number of case 3.1 and 3.2 in CYCU datasets are more than 100 students, the classification performance of case 3.1 and 3.2 have obtained high performance of .86 and .92, respectively. For the KU datasets from Table 2, case 1.1 and case 1.2 have obtained lower classification performance of .66 and .67, respectively, due to the number of students in KU datasets were smaller in the comparing six datasets. In education field, the accuracy of students' academic performance classifications ranged from 0.75 (Villagrá-Arnedo, Gallego-Durán, Compañ, Llorens Largo, & Molina-Carmona et al., 2016) to 0.95 (Hu, Lo, & Shih, 2014). According to the results of prediction accuracy, the prediction performance of six datasets were similar with the recent studies (Hu et al., 2014; Villagrá-Arnedo et al., 2016). Therefore, the students' academic performance can be predicted based on different categories of students' tracking logs.

In Table 2, the evaluators of classification performance are included of accuracy, precision, recall, F1-measure, and AUC. To reply RQ2 (Which method is the suitable classification algorithm for predicting students' academic performance?) from Table 2, LR, DT, and XGBoost are the suitable classification algorithms in comparing six datasets. The XGBoost can obtain the best classification performances in cases of 1.2 and 2.1. The LR can achieve the highest classification prediction performance in cases of 1.1, 3.1, and 3.2. For case 2.2, the best classification performance was obtained by using DT. From above description, the suitable classification algorithms include LR, DT, and XGBoost for comparing six datasets.

Table 2

The prediction(classifiaction) performance of students' academic performance based on six datasets

Method	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
	KU datasets: Case 1.1 / Case 1.2				
GaNB	.62/.58	.60/.65	.62/.58	.60/.57	.55/.61
Linear-SVC	.59/.56	.56/.56	.59/.56	.56/.56	.50/.55
SVC	.60/.55	.53/.56	.60/.55	.53/.55	.48/.55
LR	**.66**/.57	**.63**/.58	**.66**/.57	**.63**/.57	**.57**/.57
DT	.58/.61	.57/.61	.58/.61	.57/.61	.52/.60
RF	.56/.58	.55/.58	.56/.58	.56/.57	.50/.57
NN	.62/.59	.60/.59	.62/.59	.60/.59	.54/.58
XGBoost	.56/**.67**	.53/**.67**	.56/**.67**	.54/**.66**	.48/**.65**
	NCU datasets: Case 2.1 / Case 2.2				
GaNB	.59/.90	.61/.90	.59/.90	.59/.90	.58/.85
Linear-SVC	.58/.93	.59/.93	.58/.93	.58/.93	.56/.90
SVC	.59/.93	.59/.93	.59/.93	.59/.93	.55/.90
LR	.60/.94	.60/.94	.60/.94	.60/.94	.57/.92
DT	.62/**.96**	.62/**.96**	.62/**.96**	.62/**.96**	.59/**.94**
RF	.58/.93	.57/.93	.58/.93	.57/.93	.53/.87
NN	.60/.91	.59/.91	.60/.91	.60/.91	.56/.87
XGBoost	**.65**/.95	**.64**/.95	**.65**/.95	**.64**/.95	**.60**/.92
	CYCU datasets: Case 3.1 / Case 3.2				

GaNB	.81/.78	.83/.82	.81/.78	.81/.79	.83/.80
Linear-SVC	.82/.90	.82/.91	.82/.90	.82/.91	.81/.89
SVC	.84/.92	.84/.92	.84/.92	.84/.92	.82/.90
LR	**.86/.92**	**.86/.92**	**.86/.92**	**.86/.92**	**.84/.90**
DT	.80/.88	.80/.88	.80/.88	.80/.88	.78/.86
RF	.83/.87	.83/.87	.83/.87	.83/.87	.82/.84
NN	.86/.91	.86/.91	.86/.91	.86/.91	.84/.88
XGBoost	.85/.89	.85/.89	.85/.89	.85/.89	.84/.86

4.2 Exploring the influence factors on classification performance

This study applied Pearson correlation to explore the relationships among the learning actions and learning outcome. The Pearson correlation coefficient is a test statistics based on the covariance to measure the statistical relationship or association between two variables. For representing students' learning actions, this study has extracted 15, 55 and 16 features from learning environments in KU, NCU and CYCU datasets, respectively. Table 3, 4 and 5 show the Pearson correlation coefficient between the extracted features and learning outcome for KU, NCU, and CYCU datasets, respectively.

Table 3 shows the features category and Pearson correlation coefficient of KU datasets. Among the extracted 15 features, the numbers of significant features are 1 and 7 for the case 1.1 and 1.2, respectively. The feature category of KU datasets only extracted the category of online ebooks reading.

Table 3

The feature category and Pearson correlation coefficient of KU datasets

Feature name	**Category**	**Yamada et al. (2017)**	**KU**	**Grade**	
				Case1.1	**Case1.2**
Add Bookmark	Online ebook reading	√	√	-.108	.072
Add Marker	Online ebook reading	√	√	-.216	.307*
Add Memo	Online ebook reading	√	√	-.136	.289*
Change Memo	Online ebook reading	√	√	.134	.234
Close	Online ebook reading	√	√	.153	.083
Delete Bookmark	Online ebook reading	√	√	-.108	.030
Delete Marker	Online ebook reading	√	√	-.207	.301*
Delete Memo	Online ebook reading	√	√		
Jump	Online ebook reading	√	√	.059	.019
Next	Online ebook reading	√	√	.208	.371**
Open	Online ebook reading	√	√	.156	.191
Prev	Online ebook reading	√	√	.271*	.334*
Search	Online ebook reading	√	√	.049	.031
Marker	Online ebook reading	√	√	-.235	.335*
Memo	Online ebook reading	√	√	-.102	.315*
Change marker	Online ebook reading	√			
Landscape	Online ebook reading	√			
Portrait	Online ebook reading	√			
Zoom	Online ebook reading	√			

*Note: *p<.05, **p<.01, ***p<.001*

Table 4 shows the features category and Pearson correlation coefficient of NCU datasets. Among the extracted 55 features of NCU datasets, the numbers of significant features are 11 and 55 for the case 2.1 and 2.2, respectively. The feature category of NCU datasets include of online self-learning, online exercise, online discussion, online video viewing, online quiz, offline exercise, and offline quiz categories.

Table 4

The feature category and Pearson correlation coefficient of NCU datasets

Feature name	**Category**	**Grade**	
		Case2.1	**Case2.2**
active_num_days	Online self-learning	.311*	.799**
active_avg_count	Online self-learning	.269*	.509**
active_sum_count	Online self-learning	.328*	.765**
problem_num_days	Online exercise	.171	.761**
problem_avg_count	Online exercise	.254	.548**
problem_sum_count	Online exercise	.289*	.758**
video_num_days	Online video viewing	.235	.731**
video_avg_count	Online video viewing	.116	.376**
video_sum_count	Online video viewing	.146	.506**
forum_num_days	Online discussion		.203*
forum_avg_count	Online discussion		.190*
forum_sum_count	Online discussion		.207*
num_watched	Online video viewing	.232	.709**
num_complete	Online video viewing	.237	.671**
num_incomplete	Online video viewing	.114	.472**
complete_rate	Online video viewing	.205	.677**
incomplete_rate	Online video viewing	.090	.224*
watched_time_hour	Online video viewing	.164	.592**
watched_time_weekday	Online video viewing	.188	.420**
seek_video_sum	Online video viewing	.213	.591**
seek_video_avg	Online video viewing	.178	.522**
pause_video_sum	Online video viewing	.250	.728**
pause_video_avg	Online video viewing	.218	.706**
stop_video_sum	Online video viewing	.186	.531**
stop_video_avg	Online video viewing	.154	.412**
video_forward_seek_sum	Online video viewing	.188	.319**
video_forward_seek_avg	Online video viewing	.156	.291**
video_backward_seek_sum	Online video viewing	.227	.379**
video_backward_seek_avg	Online video viewing	.107	.277**
video_pause_forward_seek_sum	Online video viewing	.074	.465**
video_pause_forward_seek_avg	Online video viewing	-.004	.357**
video_pause_backward_seek_sum	Online video viewing	.265*	.564**
video_pause_backward_seek_avg	Online video viewing	.182	.506**
video_stop_backward_seek_sum	Online video viewing	.168	.471**
video_stop_backward_seek_avg	Online video viewing	.161	.278**
all_type_video_forward_seek_sum	Online video viewing	.188	.466**
all_type_video_forward_seek_avg	Online video viewing	.153	.410**
all_type_video_backward_seek_sum	Online video viewing	.247	.452**
all_type_video_backward_seek_avg	Online video viewing	.129	.370**
all_type_video_seek_sum	Online video viewing	.220	.481**
all_type_video_seek_avg	Online video viewing	.160	.428**
video_pause_sum	Online video viewing	.246	.544**
video_pause_avg	Online video viewing	.160	.438**
video_stop_sum	Online video viewing	.171	.527**
video_stop_avg	Online video viewing	.145	.403**
video_play_sum	Online video viewing	.253	.508**
video_play_avg	Online video viewing	.183	.457**
video_events_sum	Online video viewing	.255	.520**
video_events_avg	Online video viewing	.183	.455**

mt_practice_sum	Online quiz	.341**	.577**
mt_unit_sum	Online quiz	.454**	.720**
mt_online_num_day	Online quiz	.378**	.413**
mt_online_practice_num_day	Online quiz	.445**	.692**
hw_mean	Offline exercise	.349**	.789**
qz_mean	Offline quiz	.610**	.944**

*Note: *p<.05, **p<.01, ***p<.001*

Table 5 shows the features category and Pearson correlation coefficient of CYCU datasets. Among the extracted 16 features of CYCU datasets, both case 3.1 and 3.2 have 16 significant features. The feature category of CYCU datasets include of online exercise, online discussion, online video viewing, online quiz, offline exercise, and offline quiz categories.

Table 5

The feature category and Pearson correlation coefficient of CYCU datasets

Feature name	**Category**	**Grade**	
		Case3.1	**Case3.2**
video_watching_days	Online video viewing	.651**	.520**
video_watching_total_time	Online video viewing	.558**	.499**
video_watching_times	Online video viewing	.400**	.366**
bbs_days	Online discussion	.498**	.322**
bbs_num	Online discussion	.507**	.482**
video_watching_num	Online video viewing	.633**	.395**
time_between_start_first_watching	Online video viewing	.732**	.549**
pre_watching_num	Online video viewing	.181*	.204*
in_watching_num	Online video viewing	.474**	.336**
re_watching_num	Online video viewing	.425**	.336**
no_watching_num	Online video viewing	-.474**	-.336**
post_watching_num	Online video viewing	.552**	.387**
online_hw	Online exercise	.648**	.654**
online_quiz	Online quiz	.808**	.617**
pgm	Offline exercise	.603**	.745**
quiz	Offline quiz	.835**	.834**

*Note: *p<.05, **p<.01, ***p<.001*

Table 3, 4 and 5 show the feature category and Pearson correlation coefficient for the comparing six datasets. For the comparing six datasets, Table 6 summarizes the number of extracted features and significant features in each feature category according to the results of Pearson correlation coefficient. The KU, NCU and CYCU datasets have extracted 15, 55, and 16 features, respectively. The significant rate (sig. rate) can be defined as the number of significant features divided by the number of extracted features. From Table 6, significant rate of case 1.1, 1.2, and 2.1 were lower than 0.5 and were the three lowest values in the six datasets, and their classification performance is range from 0.65 to 0.67. The number of significant features is not sufficient in 1.1, 1.2, and 2.1 cases resulting in their lower classification performance.

Table 6

The number of extracted features and significant features in each feature category

Categories	KU		NCU		CYCU	
	Case1.1	Case1.2	Case2.1	Case2.2	Case3.1	Case3.2
Online Self-learning	0/0	0/0	3/3	3/3	0/0	0/0

Online video viewing	0/0	0/0	1/40	40/40	10/10	10/10
Online reading	1/15	7/15	0/0	0/0	0/0	0/0
Online Discussion	0/0	0/0	0/3	3/3	2/2	2/2
Online exercise	0/0	0/0	1/3	3/3	1/1	1/1
Online quiz	0/0	0/0	4/4	4/4	1/1	1/1
Offline exercise	0/0	0/0	1/1	1/1	1/1	1/1
Offline quiz	0/0	0/0	1/1	1/1	1/1	1/1
Total	1/15	7/15	11/55	55/55	16/16	16/16
Sig. rate	.06	.46	.2	1.0	1.0	1.0

Note: number of significant features / number of features

For exploring the factors influence on the performance of classifications, Table 7 summaries the situation of influence factors for each dataset. For six datasets, this study has investigated the three influence factors affected on the performance of classifications which include of sample size, feature category, and significate features. This study aims to establish students' academic performance classification; the sample size can be set as the number of students which shows in Table 1. For case 1.1 and 1.2 of KU datasets, the classification performances are very low due to the small sample size, small feature categories, and small significant rate. The reason for the low classification performance of case 2.1 is that both the sample size and significant rate are too small. In contrast, the three cases of 2.2, 3.1 and 3.2 with sufficiently larger of samples, feature categories and significant rate to obtained higher classification performance.

Table 7

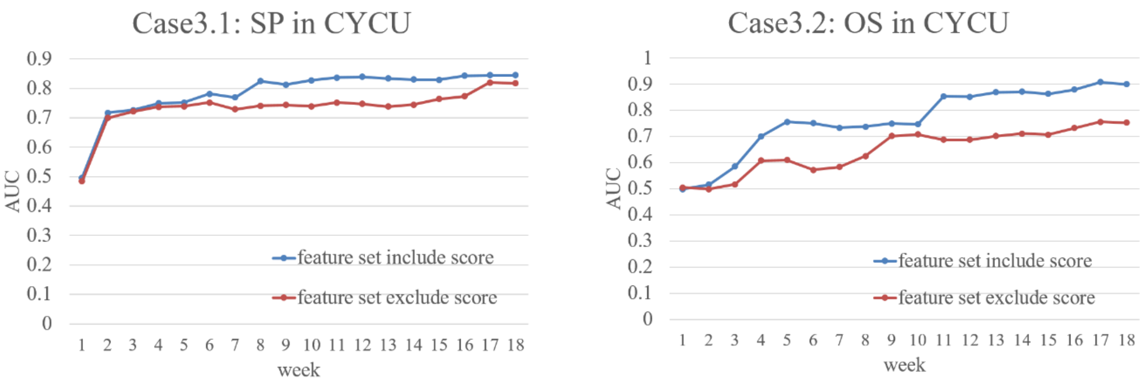
The situation of three influence factors for each dataset

Factor	**KU**		**NCU**		**CYCU**	
	Case1.1	**Case1.2**	**Case2.1**	**Case2.2**	**Case3.1**	**Case3.2**
Small samples	Y(53)	Y(55)	Y(59)	N(128)	N(135)	N(125)
Limited feature categories	Y(1)	Y(1)	N(7)	N(7)	N(6)	N(6)
Small sig. rate	Y (.06)	Y (.46)	Y (.2)	N(1.0)	N(1.0)	N(1.0)

Note: Whether dataset have the problem caused from the factor (the value of the factor)

To further study how students' scores influence on the prediction performance, the case 3.1 and 3.2 of CYCU datasets were not only divided into 18 weekly datasets, but also computed the AUC of classifications by using feature set include and exclude score information. Figure 1 shows the AUC of two classifications by using feature set include and exclude score information over weeks. In figure 1, (a) and (b) were showed the AUC for case 3.1 and 3.2 over weeks, respectively. For the CYCU datasets, there are 4 score related features which consist of online

online_hw, online_quiz, pgm and quiz features listed in Table 5. Therefore, the feature sets include and exclude score information were consist of 16 features and 12 features, respectively. From Table 5, online_hw and online_quiz indicate the sum of online homework and exam, respectively; pgm and quiz indicate the sum of projects' score and exam score.



(a) The AUC over weeks for case 3.1 (b) The AUC over weeks for case 3.2

Figure 1. The AUC of two classifications by using feature set include and exclude score information over weeks for CYCU datasets

For the case 3.1, the values of feature of online_hw and online_quiz can be obtained at the weeks of 4, 6, 9, 10, 14, 15, 16, 17, 18; the value of pgm feature can be obtained at weeks of 7, 14 and 17; the value of quiz feature can be obtained at weeks of 8 and 17. From figure 1, the distances of the two classified AUCs by using the feature set including and excluding the score information can be separated at sixth week, and then pulled to a larger distance at eighth week. This may be because that the online_hw and online_quiz can be obtained two scores at sixth week, and the scores of pgm and quiz can be firstly obtained at week 7 and 8, respectively. In other words, the values of online_hw, online_quiz, pgm and quiz features can be obtained at eighth week resulting the larger distance of two AUCs of classifications.

For the case 3.2, the values of feature of online_hw and online_quiz can be obtained at the weeks of 3, 4, 5, 7, 9, 14, 15, 16, 17, 18; the value of quiz feature can be obtained at weeks of 5 and 9; the value of pgm feature can be obtained at weeks of 11 and 17. From Figure 1, the distances of the two classified AUCs by using the feature set including and excluding the score information can be separated at third week, and then pulled to a larger distance at fifth and eleventh weeks. This may be because that the online_hw and online_quiz can be firstly obtained at third week, and the scores of pgm and quiz can be firstly obtained at week 11 and 5, respectively. We have get the values of online_hw, online_quiz and quiz feature at fifth week, and then the value of pgm was firstly obtained at eleventh week. This is the causing reason for the larger distance of two AUCs of classifications at fifth and eleventh weeks.

5. Conclusion

To investigate the prediction performances of various classifications in education field, this study aims to construct students' academic performance classification based on different categories of students' tracking logs. For the comparing six datasets, the values of prediction accuracy are range from 0.65 to 0.96 showed in Table 2. These results are similar and consistent with the previous studies (Hu et al., 2014; Villagrá-Arnedo et al., 2016). Based on the above results, we can predict students' academic performance based on different categories of students' tracking logs.

We have further explored the influence factors of prediction performance of classifications in various datasets. This study aims to investigate the influence of sample size, feature category, and the significant features on the performance of classifications. For the two cases of 1.1 and 1.2 in KU, the low prediction performance was caused by the small samples, only one type of feature category, and small number of significant features. Although there are 7 feature categories in NCU, the prediction performance still very low for case 2.1. This result is caused by small samples and the small number of significant features. The sample size, feature categories, and significant features are sufficient large for cases of 2.2, 3.1, and 3.3. Consequently, the three cases can have achieved good prediction performance by using DT, LR, and XGBoost. Finally, we have also investigated how the score related features influence on prediction performance over weeks. From figure 1, the AUC can be pulled larger when the values of score related features were obtained. In the other words, according to the AUC of case 3.1 and 3.2 in CYCU datasets showed in figure 1, the score related features have greatly effects on prediction performance.

Acknowledgements

This work was supported by Ministry of Science and Technology, Taiwan under grants MOST-105-2511-S-008-003-MY3, and MOST-106-2511-S-008 -004 -MY3.

References

Asif, R., Merceron, A., & Pathan, M. K. (2014). Predicting student academic performance at degree level: a case study. *International Journal of Intelligent systems and applications*, *7*(1), 49.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.

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

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 215-242.

Flanagan, B., Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy. In *International Conference on Computers in Education (ICCE2017)*, 333-338.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*(2), 215-223.

Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, *36*, 469-478.

John, G. H., & Langley, P. (1995, August). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (pp. 338-345). Morgan Kaufmann Publishers Inc.

Lu, O.H.T, Huang, A.Y.Q, Huang, J.C.H., Lin A.J.Q., Ogata, H., Yang, S.J.H.*, (2018). Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning, *Educational Technology & Society*, *21*(2), 220-232.

Lu, O.H.T, Huang, A.Y.Q, Huang, J.C.H., Yang, S.J.H.* (2017). Applying learning analytics for improving students engagement and learning outcomes in a MOOCs enabled collaborative programming course, *Interactive Learning Environments*, *25*(2), 220-234.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015, January). E-Book-based learning analytics in university education. In *International Conference on Computer in Education (ICCE 2015)*, 401-406.

Oladokun, V. O., Adebanjo, A. T., & Charles-Owaba, O. E. (2008). Predicting students' academic performance using artificial neural network: A case study of an engineering course. *The Pacific Journal of Science and Technology*, *9*(1), 72-79.

Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end games. In *Machine learning*(pp. 463-482). Springer, Berlin, Heidelberg.

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, *68*, 458-472.

Villagrá-Arnedo, C. J., Gallego-Durán, F. J., Compañ, P., Llorens Largo, F., & Molina-Carmona, R. (2016). Predicting academic performance from behavioural and learning data.

Yamada, M., Oi, M., & Konomi, S. I. (2017). Are Learning Logs Related to Procrastination? From the Viewpoint of Self-Regulated Learning. *International Association for Development of the Information Society*.

Yoo, J., & Kim, J. (2014). Can online discussion participation predict group project performance? Investigating the roles of linguistic features and participation patterns. *International Journal of Artificial Intelligence in Education*, *24*(1), 8-32.