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

Good Students Look Back Previous Pages

Sachio HIROKAWA
Kyushu University, Japan
hirokawa@cc.kyushu-u.ac.jp

Abstract: Educational institutions have many expectations for the use of E-book. The top expectation is to evaluate and to improve the education system based on the accumulated learning activity log data. This paper applied machine learning to predict the learner's final score from e-Book browsing logs. The present paper evaluated the prediction performance of the good students with the final grade of 80 or more from their learning access logs. An experimental evaluation revealed that the prediction performance (accuracy) was only 64% if we use only the accessed page information. However, the accuracy was improved to 89% when consecutive browsing page transition information was used. Furthermore, it was confirmed that returning to the previous page as a feature of the highest grades student.s.

Keywords: e-Book Access Log, Student Score Prediction, SVM, Feature Selection

1. Introduction

According to (Rockinson-Szapkiw et.al 2013), e-Book is defined as "Texts that are digital and accessed via electronic screens". Even in early days of 2008, it is pointed out that e-book is effective for learning compared to conventional textbook (Shepperd, Grace & Koch 2008). Japanese government scheduled to use e-books for elementary , middle and high schools by 2020. Much attention is been paid to make good use of data kept as student learning logs. In fact, the research on learning analitics (LA) and on educational data mining (EDM) attract many researchers, for example in conferences of Learning Analytics (https://solaresearch.org/events/lak/) and Educational Data Mining (http://educationaldatamining.org/). Another reason of this current status is the progress of learning analytics platform. Actually, the data the present paper analyse is provided by the BookRoll system (Flanagan & Ogata 2017, Ogata et al. 2015, Ogata et.al.2017). The present author proposed a visualization of page view transision in (Hirokawa et al. 2015). However, the quantative evaluation of the visualization was not given then. The present paper applies the machine learning method SVM and feature selection to predict the high peformance students who achived more than 80 points in their evaluation. As the result, we observed that the transition information gives better performance compared to simple page view information. The accuracy of the former method is 0.64 and that of the later method is 0.89. Morever, as the result of optimal feature selection, it is turned out that the characteristice feature of high performance student is that they look back previous pages.

1. e-Book Access Log

1.1 Log Data

Final scores and learning log data of a total of 108 students (65,757 cases) of two classes of class 1 (53 people) and class 2 (55 people) are provided. The score of the student ranges from 0 to 100 points. The learning log data is composed of 13 attribute as shown below in one line. For other than 4 items, userid, pageno, contensid, eventtime, it was biased towards a small part. Therefore, in this paper we analyzed using classid and those 4 items. Items marked with * in Table 1 are items used in this paper. There are 6 differenct contents as shown in Table 2.

Table 1 Attributes in Student Learning Log

	attribute	example
0	*userid	ds126
1	action	https://w3id.org/xapi/adb/verbs/read
2	operationname	NEXT
3	makercolor	NULL
4	processcode	3
5	devicecode	tablet
6	makerposition	NULL
7	description	NULL
8	makertext	NULL
9	*pageno	1
10	*contents id	9f61408e3afb633e50cdf1b20de6f466
11	memotext	NULL
12	*eventtime	2017/11/22 5:55

Table 2 Access Count of Learning Content

class 1		class 2	
#access	content id	#access	content id
14691	No.1	27261	No.16
7959	No.11	5110	No.18
6159	No.12	4500	No.17

1.2 Page Access Transition

As a way of using page information in analysis, we analyzed by two methods using a page number regardless of the content id and a method using content id and page number as a set. Imagine, for example, that a student ds126 accessed the fourth page of the content No.11. The data of the student ds126 contains the page number "4" as well as the pair "11:4" of the conntentid 11 and the page number 4. Also, eventtime was used to extract page information that the same student accessed consecutively.

Table 3 shows a part of access logs in which we can see the pages the ds126 and ds112 viewed. We used the sequence of pages that a student viewed consecutively. The student has the paths 1-2 and 2-3. The student ds112 has the paths 10-9, 9-8 and 8-7. We used those paths in vectorizing the students. As the number of page transition steps, two kinds of analyzes were carried out, one with only two steps and one with five steps. In the example below, ds 126 has 3 steps with a maximum of 1-2-3, and ds 112 has 4 steps with a maximum length of 10-9-8-7.

Table 3 Learning Log Sample

userid	pageno	eventtime
ds126	1	2017-11-22 05:55:24.0000000
ds126	2	2017-11-22 05:55:24.0000000
ds118	1	2017-11-22 05:55:24.0000000
ds146	1	2017-11-22 05:55:25.0000000
ds126	3	2017-11-22 05:55:25.0000000
ds112	10	2017-11-22 05:55:25.0000000
ds112	9	2017-11-22 05:55:25.0000000

ds112	8	2017-11-22 05:55:25.0000000
ds112	7	2017-11-22 05:55:25.0000000

2. Prediction of High Performance Students

In this paper, 108 students are vectorized using each student's access log information. Using a linear kernel of machine learning SVM as a positive example for students with a score of 80 or higher, the evaluation performance was evaluated. We used the set of student with a score of 80 or higher as the positive data and applied a linear kernel of machine learning SVM. We applied the feature selection (Sakai & Hirokawa 2012) based on the score of words obtained as the SVM model. We chose the top N positive words and the top N negative words in vectorization. We evaluated the prediction performance by taking the average of 3-fold cross-validation that have been run 10 times with partitions selected randomly for each run.

We conducted evaluation experiments with on three kinds of vectorizations, i.e., the vectorization based on page numbers and the vectorization (path2 and path5) based on the pages accessed consecutively. The path2 used two pages accessed consecutively as well as pages. The path5 used the pages accessed conscutively of the length less than or equal to five. An attribute "10-9" in path2 means that the student accessed the page 10 after the page 9. An attribute "10-9-8-7-6" means that the student accessed those page in that order. The values in Table 4 and Figure 1 means the prediction performance with respect to the vectorization methods, page-50, path2-300 and path5-900, where "50", "300" and "900" indicate the optimal feature selection N. For example, the accuracy for the method with page is 0.64. We can see that the accuracy is improved quite well with 0.83 for path2 and 0.89 for path5. In other words, we can conclude that there is a difference between high performance students and other students as to how the transition of the browsing page is more than just which page the student viewed.

Table 4 Optimal Prediction Performance of High Score Students

Attribute	page-50	path2-300	path5-900
Precision	0.6563	0.8291	0.8808
Recall	0.9011	0.9143	0.9581
F-measure	0.7558	0.8660	0.9164
accuracy	0.6447	0.8251	0.8915

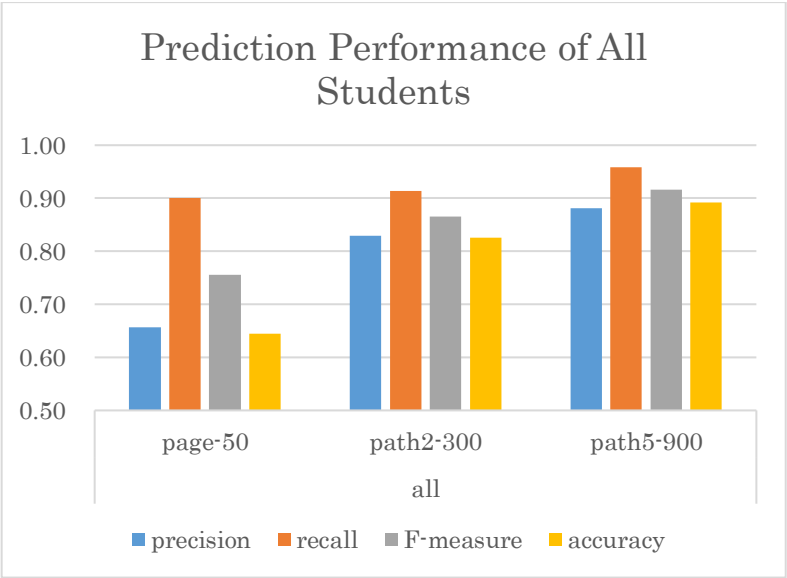


Figure 1 Otimal Prediction Performance of High Score Students

Figure 2 shows the accuracy when the number N of attribute selection is changed. In vectorization with only pages, changes in discrimination performance are hardly seen. On the other hand, in

path2 and path5, the accuracies are the highest at N=300 and N=900, indicating that the high peformance students can be characterized by the selected attribute set.

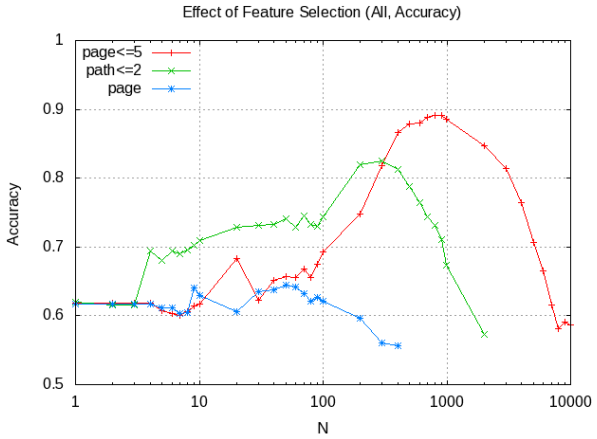


Figure 2 Effect of Feature Selection

3. Characteristic Feature of High Performance Students and Look Back Behavior

3.1 Characteristice Pages Accessed by High Performance Students

In this chapter, we will clarify the characteristics of the high performance students by looking at the attributes on the top of the score in detail. Table 5 shows positive characteristic attributes and negative attributes up to the top ten scores when the pages are attributes, that is, a characteristic pages accessed by those students. The column "df" shows the number of students of the attribute. Note that the different pages in th different content ids may have the same page number. Comparison and detailed analysis of different classes and contents are shown in the next section. In this chapter, common features independent of classes are extracted.

Table 5 Positive & Negative Attributes (pages)

positive attributes				negative attributes			
rank	score	df	page No	rank	score	df	page No
1	0.9619	104	9	1	−1.8282	1	250
2	0.9619	104	10	2	−1.0978	100	22
3	0.9605	81	68	3	−0.9976	1	182
4	0.9172	82	70	4	−0.9976	1	162
5	0.6428	90	40	5	−0.9109	83	67
6	0.6428	89	39	6	−0.8676	82	69
7	0.5587	98	26	7	−0.7298	101	21
8	0.5587	98	29	8	−0.6759	104	11
9	0.5587	98	27	9	−0.6759	104	13
10	0.3666	62	100	10	−0.6759	104	12

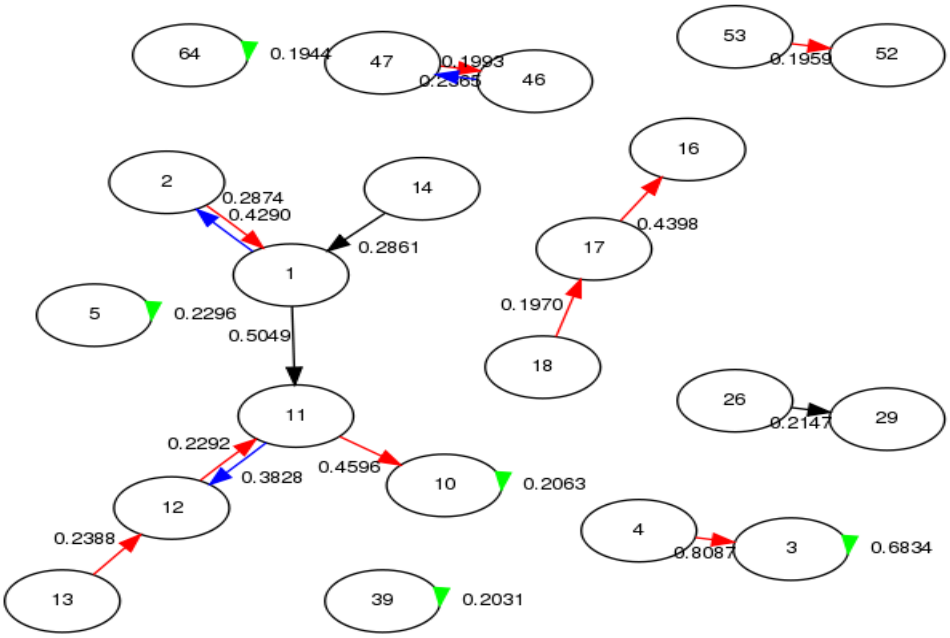
3.2 Characteristic Paths

Table 6 shows the positive attribute and the negative attribute up to the ten highest scores in the vectorization including the consecutive two pages of browsing transition information. Since the basic interface of the e-Book reading system is with PREV and NEXT, the movements to the adjacent page is the most frequent ones.

As a result, the attribute of the page alone is not included in the upper level, and it can be seen that the page transition is effective for higher student identification. Further, looking at the difference between the positive page transition and the negative page transition in the positive page transition (that is, the upper student), as in the case of 4-3, 11-10, 17-16, 13-12, we can observe they are browsing back to. On the other hand, negative page transitions (i.e. lower students) have no such feature. Figures 3 and 4 visualize positive and negative page transitions as directed graphs. Red eges indicate reverse browsing, blue branches indicate normal order browsing.

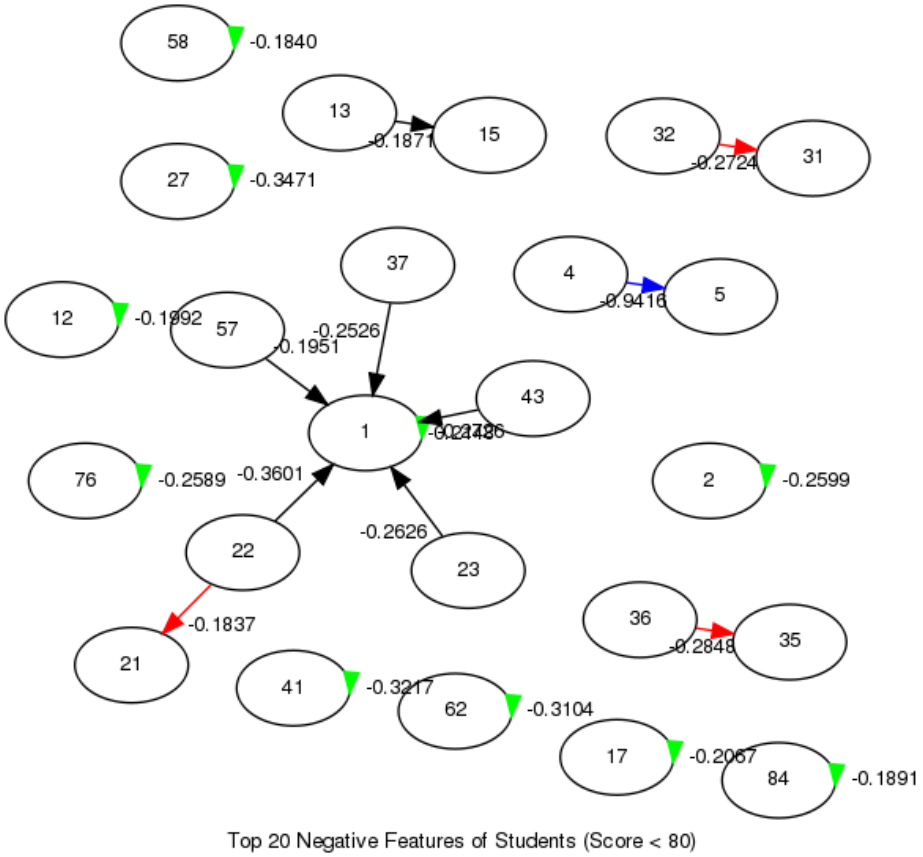
Table 6 Positive & Negative Attributes (pages & length 2 paths)

	positive attributes				negative attributes		
rank	score	df	attribute	rank	score	df	attribute
1	0.8087	59	4-3	1	-0.9416	103	4-5
2	0.6834	18	3-3	2	-0.3601	11	22-1
3	0.5049	12	1-11	3	-0.3471	14	27-27
4	0.4596	77	11-10	4	-0.3217	14	41-41
5	0.4398	73	17-16	5	-0.3104	10	62-62
6	0.4290	105	1-2	6	-0.2848	59	36-35
7	0.3828	103	11-12	7	-0.2726	108	1-1
8	0.2874	57	2-1	8	-0.2724	61	32-31
9	0.2861	9	14-1	9	-0.2626	12	23-1
10	0.2388	75	13-12	10	-0.2599	15	2-2



Top 20 Positive Features of Students (Score >= 80)

Figure 3 Top 20 Positive Paths



Top 20 Negative Features of Students (Score < 80)

Figure 4 Top 20 Negative Paths

4. Difference in Classes

4.1 Distribution of Final Scores and Distribution of Page Access Counts

There are two kinds of data provided by the organizer the workshop. Because there are differences in browsing pattern, this chapter conducted detailed analysis for each class. In this paper, we analyzed them as class 1 and class 2 respectively. Figure 5 shows the distribution of class 1 and class 2. In class 2, the scores are distributed around 80 points. On the other hand, class 1 has the largest number of 100 points. Figure 6 shows the distribution of the number of accesses for the contents with the top three accessed contents. Meanwhile, the number of accesses of contents is less than 3 for other contents. We observe that there is a big difference between the data of class 1 and class 2 both in grades and contents access distribution.

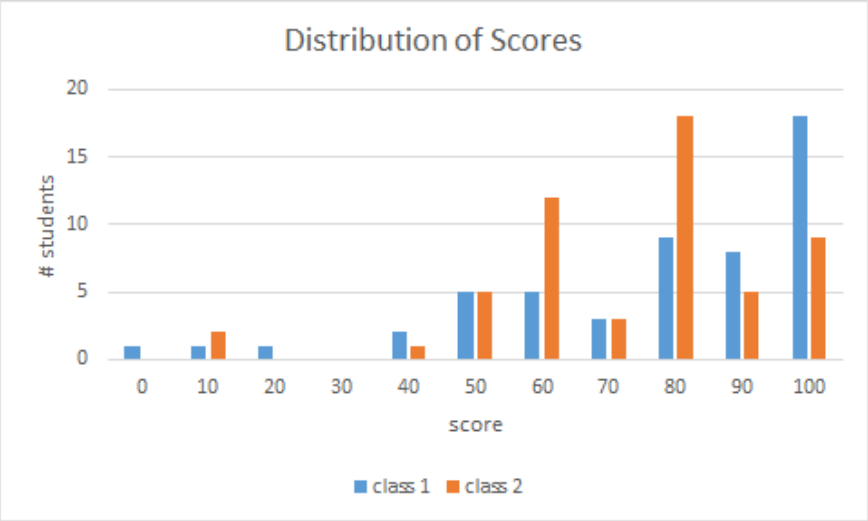


Figure 5 Distibution of Scores of Class 1 & Class 2

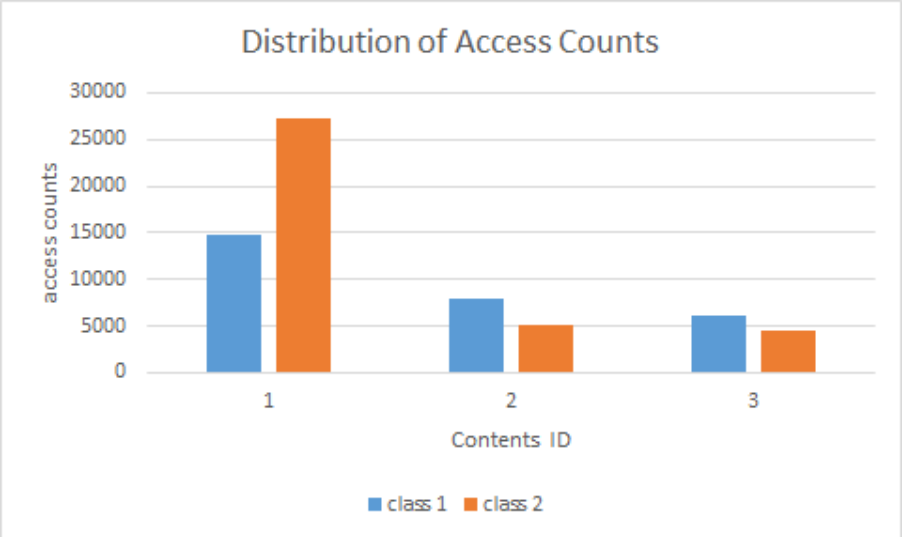


Figure 6 Distribution of Top Three Access Contens of Class 1 & Class 2

4.2 Prediction Performance of Each Class

In concrete experiments, to each element of the student information vector we added pairs of a path and the class as a new attribute. For example, when a student of class 1 browsed page number 11, it is expressed as "1:11" as an attribute. The transitioning from page 11 to page was expressed as an attribute "1:11-12".

Table 7 and Figure 7 show the optimal discrimination performance of the high performance students for class 1 and class 2 F-measure and accuracy are both around 80% in class

1, but 95% or more in class 2. Figures 8 and 9 show the change in accuracy when the number of attributes is changed. In both classes, it turns out that they are stably optimized around a specific N.

Table 7 Optimal Prediction Performance for Class 1 & Class 2

target	class 1			class 2		
attribute	page−1	path2−40	path5−100	page−30	path2−200	path5−400
precision	0.6639	0.8305	0.8400	0.6624	0.9469	0.9432
recall	1.0000	0.8524	0.8872	0.9096	0.9774	1.0000
F−measure	0.7921	0.8338	0.8552	0.7586	0.9599	0.9695
accuracy	0.6639	0.7799	0.8047	0.6702	0.9553	0.9652

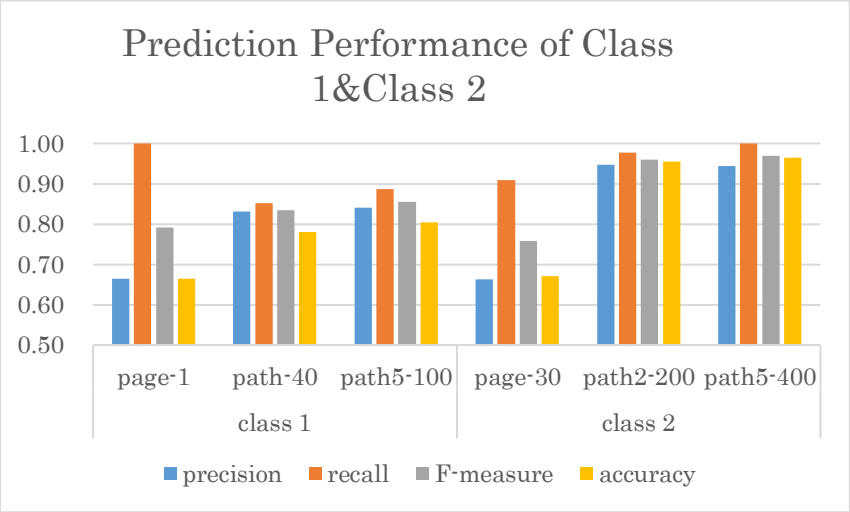


Figure 7 Optimal Prediction Perforance of Class 1 & Class 2

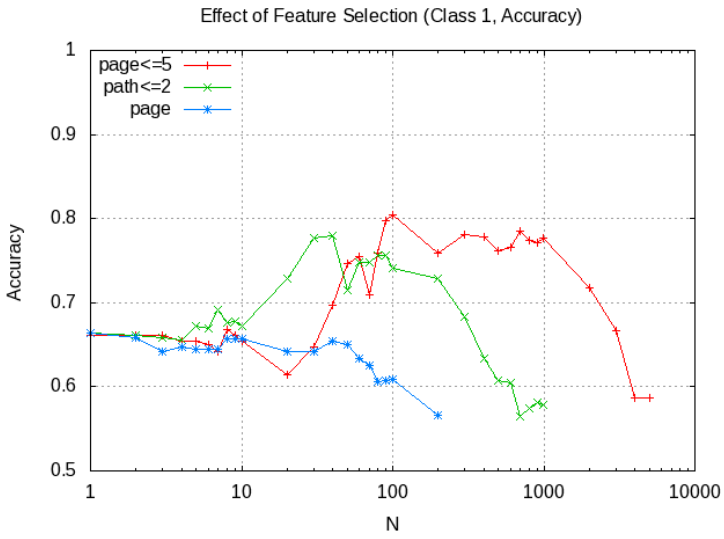


Figure 8 Effect of Feature Selection (Class 1)

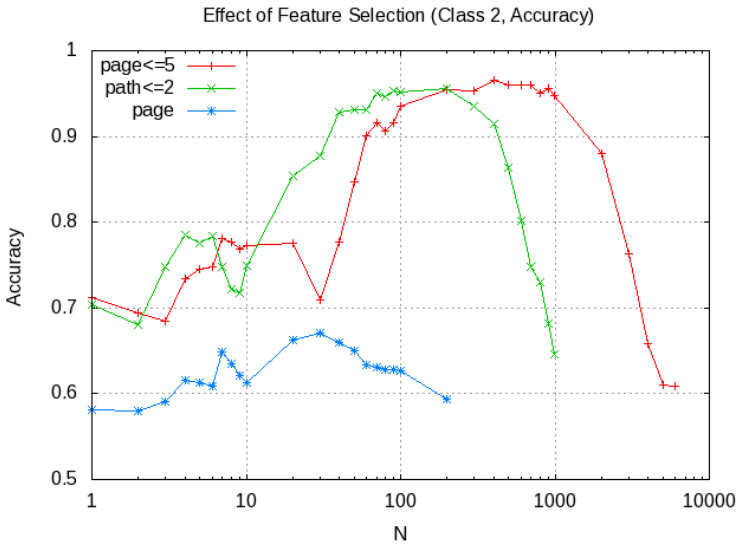


Figure 9 Effect of Feature Seletion (Class 2)

5. Conclusion and Further Work

Learning activity analysis based on real data has become possible by computerization of teaching materials and accumulation of learning behavior logs. In this paper, 108 e-book browsing logs of students were analyzed. By applying machine learning SVM and attribute selection to browsing information, we identified high performance students with more than 80 points. With the information on the browsing pages, the identification performance (accuracy) was 64%. But the accuracy reached 89% if we use the transition information of the browsing pages.

Looking at the to scored attributes in the discrimination model for the top grades students, the view pattern such as 4-3, 11-10, 17-16, were obtained as characteristic features that return to the preveious page. At this stage, we do not know the reason for that pattern. A hypothesis may be that good students return to the last when they concounter any difficulty at the time of learning. In this paper, learning behavior of students is vectorized with information of browsing page and viewing page transition, and classification of two classes, upper and lower students, was done. Because RMSE and AUC are recommended as evaluation indices of workshop, we conducted the regression. Specifically, the experiment was carried out by regression of SVM-light with a linear kernel. The results are as shown in Table 8 and Table 9, but they are very poor. However, in vectorization, each student is expressed as a Boolean vector of 1, 0. Regression analysis by normalizing frequency is a future task.

Table 8 RMSE with Different Vectorizations

	page	path2	path5
all	22.282	22.0643	22.0389
class 1	26.1996	26.0596	26.3405
class 2	18.6815	18.2745	18.4375

Table 9 AUC with Different Vectorizations

	page	path2	path5
all	0.5521	0.6075	0.6254
class 1	0.5588	0.5812	0.5956
class 2	0.5396	0.6322	0.6535

References

Flanagan, B., and Ogata H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy. *Proceedings of ICCE2017* (pp.333-338).

Hirokawa, S., Yin, C., Wang, J., Oi, M., and Ogata, H. (2015). Visualization of e-Book Learning Logs, *Proceedings of ICCE2015* (pp.659-664).

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K.,and Yamada, M., (2015). E-Book-based learning analytics in university education, *Proceedings of ICCE 2015* (pp.401-406).

Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., Wang, J.,and Hirokawa S. (2017), Learning Analytics for E-Book-Based Educational Big Data in Higher Education, In: Yasuura H., Kyung CM., Liu Y., Lin YL. (Eds) *Smart Sensors at the IoT Frontier* (pp.327-350), Springer, Cham.

Sakai, T., and Hirokawa S. (2012), Feature Words that Classify Problem Sentence in Scientific Article, *Proceedings of iiWAS2012* (pp.360-367).

Shepperd, J.A., Grace, J.L., and Koch E.J.(2008), Evaluating the electronic textbook: is it time to dispense with the paper text?. *Teaching Of Psychology*, 35(1), 2-5.

Rockinson-Szapkiw, A., Courduff, J., Carter, K., and Bennett D.(2013), Electronic versus traditional print textbooks: A comparison study on the influenc e of university students' learning, *Computers & Education*, 63, 259-266.