International Journal of Learning Technologies and Learning Environments
International Institute of Applied Informatics
2019, Vol. 2, No. 1, 19 – 34

Prediction of Success or Failure for Examination using Nearest Neighbor Method to the Trend of Weekly Online Testing

Hideo Hirose * †

Abstract

Using trends of estimated abilities in terms of item response theory for online testing, we can predict success/failure for term-end examinations for each student at early stages in courses. We applied the newly developed nearest neighbor method for determining the similarity of learning skills in the trends of estimated abilities, resulting in better prediction accuracy for success or failure. This paper shows that the use of the learning analytics incorporating trends for abilities is effective. ROC curve and recall precision curve are also utilized in the proposed method.

Keywords: success/failure prediction, item response theory, nearest neighbor, similarity, online testing, learning analytics.

1 Introduction

Since students of widely varying abilities are now enrolled in universities, it is crucial to identify students at risk of failing courses and/or dropping out as early as possible (see [27, 31]). However, the greater the varying abilities of students, the more we need methodologies for assisting students because conventional methods may not work when the numbers of staffs and classes are small. New assisting systems are required to solve such a difficulty.

To overcome this difficulty, we established online testing systems aimed at helping students who want to improve their mathematical skills. In such systems, we included learning check testing (LCT) for every class to measure student comprehension of lectures. The system has been successfully operating (see [16], [17]), and some computational results have been reported [19]. In addition, other relevant aspects have been investigated (see [18],[20], [21], [22], [26], [30]).

As indicated in [5], [6], and [27], the current focus on learning analytics is necessary in order to make a sustainable impact on the research and practice of learning and teaching. Using outputs obtained from the online testing, it is not so difficult to collect a

* Hiroshima Institute of Technology, Hiroshima, Japan
† This work was supported by JSPS KAKENHI Grant Number 17H01842.

large amount of learning data. We may be able to actively utilize the collected data to find optimal strate-gies for improving learning methods. It is also important to analyze the data theoretically (see [32]).

This paper is aimed at developing effective learning strategies of students at risk of failing courses and/or dropping out, using the large-scale learning data collected from online testings. In this paper, unlike conventional methods using correct answer rate (CAR) to identify proficiency of a student (e.g., see [19]), we use the *ability* obtained from item response theory (IRT, e.g., see [1], [7], [23]), and we introduce a new method to identify students at risk as early as possible using the IRT results.

This kind of research is part of the field of *educational data mining*, where learning analytics are used to find better learning methodologies. Referring to [2], these methodolo-gies fall into the following general categories: prediction, clustering, relationship mining, discovery with models, and distillation of data for human judgment. References [3], [24], and [29] are among them. Since this paper aims at the prediction of students at risk of failing courses and/or dropping out, a classification method is developed. In classification, decision trees, logistic regression (for binary predictions), and support vector machines are often applied (e.g., [28]). However, we introduce a newly developed method that uses the nearest neighbor method to compute the success/failure probability.

2 Weekly Online Testing

Analysis basic (i.e., calculus) and linear algebra are two fundamental subjects that mathematics teachers are involved in weekly online testings. Testing time duration is ten minutes, and m questions using multiple choice are provided to each testing; $m = 5$ is used in the first semester in 2017. The testings to check comprehension of each unit are incorporated into regular classes; for example, in the case of analysis basic, *differentiation* unit has a set of question items for testing, and calculus online testings consist of 14 different such sets. Each subject (analysis basic or linear algebra) consists of 16 units including midterm and end-term examinations; except for two examination classes, students have 14 lectures incorporating LCT; if we denote K as the number of opportunities that students take LCT, $K = 14$ in the first semester in 2017. In addition, we define the number of freshman students to be enrolled as N; in the first semester in 2017, N is approximately 1,100. Thus, we have user-item response matrices sized of $N \times mK$ to each subject at the end of the semester.

Figure 1 shows a part of such a response matrix; row and column correspond to student id and item (question) id, respectively; a red color element indicates that a student solved a problem item successfully, and a green color element means to be a failed response.

3 Ability Evaluation Using the IRT

In many cases, evaluation for learning skill is assessed by using correct answer rate (CAR) to questions; CAR values are obtained by the ratio of the number of correct answers to the number of given questions. Although this criterion is easily understood, it does not include effects from other students' scores.

Item response theory (IRT) provides us difficulties of the test items (problems) and the examinees' abilities together, resulting in evaluating examinees' abilities accurately

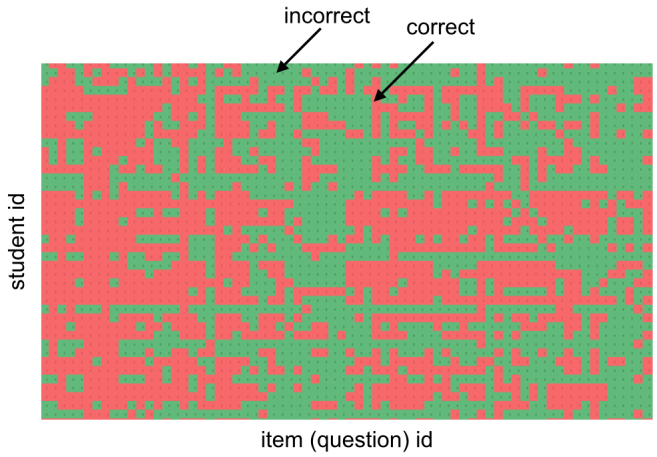

Figure 1: A part of a item response matrix (analysis basic in the first semester in 2017).

and fairly. In addition, adaptive testing using IRT selects the most appropriate items to examinees automatically, resulting in more accurate ability estimation and more efficient test procedures (see [11], [12], [13], [14], [15], [25]). Thus, we incorporated IRT evaluation method into the online testing systems. In this paper, we deal with the cases of the standard IRT evaluation using the two-parameter logistic function $P(\theta_i; a_j, b_j)$ shown below.

$$\begin{aligned} P(\theta_i; a_j, b_j) &= \frac{1}{1+\exp\{-1.7a_j(\theta_i - b_j)\}}, \\ &= 1 - Q(\theta_i; a_j, b_j), \end{aligned} \tag{1}$$

where θ_i expresses ability for student i, and a_j, b_j are constants in the logistic function for item j called the discrimination parameter and the difficulty parameter, respectively. The constant number 1.7 is used to fit the logistic distribution model to a standard normal distri-bution model. The corresponding likelihood function for all the examinees, $i = 1, 2, \ldots, N$, and all the items, $j = 1, 2, \ldots, n$, will become

$$L = \prod_{i=1}^{N} \prod_{j=1}^{n} \left(P(\theta_i; a_j, b_j)^{\delta_{i,j}} \times Q(\theta_i; a_j, b_j)^{1-\delta_{i,j}} \right), \tag{2}$$

where $\delta_{i,j}$ denotes the indicator function such that $\delta = 1$ for success and $\delta = 0$ for failure in answering a question. We adopt the IRT evaluation for students' abilities unlike the case in [19].

4 Trend of Estimated Students' Abilities Using Each Unit Response Matrix in the IRT

First, we show some trends for estimated abilities to each unit. This means that we use the response matrices $M_k(N, m)$, $k = 1, \ldots, K$. Then, we define $\theta_0(i, k)$ as student i's ability using the kth LCT response results, where each response matrix is a $N \times m$ size matrix.

Figure 2 shows a part of such a case for analysis basic; in this demonstration, N is about 100, and this corresponds to students for some department. The figure indicates that it seems difficult to discriminate students into certain categories. We see many up-and-down ability estimates in the ability trends from the 1st LCT to 14th LCT. The small

number of question items may make variance of the estimates large. That is, the ability estimates using each LCT response matrix are unreliable.

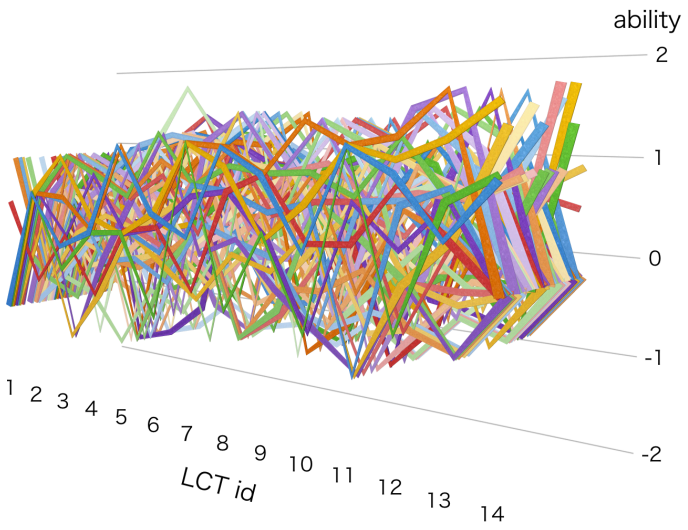

Figure 2: Some trends for the estimated abilities $\theta_0(i,k)$ to each unit (analysis basic in the first semester in 2017).

We can also see that mean trends of abilities to each student show a slight ascending tendency. However, this is actually resulting from ascending difficulties as lectures go forward, i.e., the more lectures students take, the more difficult the lecture level becomes. Thus, this tendency could be ignored.

5 Identifying Successful/Failed Students Using the Full Response Matrix in the IRT

To identify students at risk, the use of known two categorized groups could be helpful: one is successful students for the end-term examination, and the other is failed students. Figure 3 shows a histogram of estimated abilities of LCT to successful students overlaid a histogram of estimated abilities of LCT to failed students in the case of analysis basic in the first semester in 2017. The numbers of successful students and failed students are 921 and 206, respectively; the ratio of failed students to all the students is 0.18. Here, we have used full response matrices $M(N, mK)$ in estimation to obtain the most reliable estimates for abilities.

Except for very low values of ability estimates, the histograms indicate the normal distributions with different mean values (around 0.22 for successful students and -0.57 for failed students); the lowest estimates around -3.0 in both groups were resulting from the absence for testings. However, it seems very difficult to discriminate students into two groups by using certain ability threshold value. When we adopt the decision tree method, the most appropriate ability threshold value becomes to be -0.047.

The confusion matrix using this threshold is illustrated in table 1. The misclassification rate for this confusion matrix is 0.28. Limited to failed students, the decision tree predicted

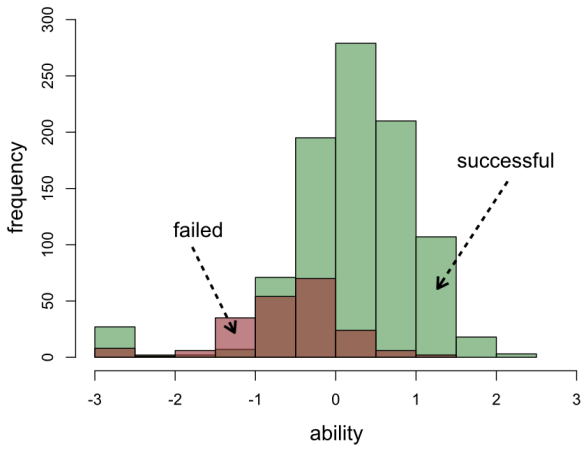

Figure 3: Histograms of estimated abilities for successful/failed two groups (analysis basic in the first semester in 2017).

that 446 students may fail, and eventually 169 students actually failed; the hitting ratio is 38%, and the result seems not to be useful.

Table 1: Confusion matrix determined by decision tree using full response matrix.

		predicted		
		successful	failed	total
observed	successful	644	277	921
	failed	37	169	206
	total	681	446	1127

threshold $= -0.047$

In addition to the LCT results, we have incorporated placement test (PT) results taken at the very beginning of the first semester. We have two kinds of PTs: one is a rather fundamental test and the other is an advanced test in high school level. Using the fundamental PT and the LCT results, we plotted correlations for these two tests in three groups in Figure 4 in the case of analysis basic in the first semester in 2017: first group is the successful in the end-term examination (score range is 60-100 expressed by green dots in the figure), second group is the badly failed group (score range is 0-39 expressed by red dots), and the rest is also the failed group (score range is 40-59 expressed by yellow dots). The horizontal axis means the ability values standardized to a standard normal distribution, and the vertical axis means the fundamental PT score. Although information using the computational results via IRT is added, it is still hard to find the boundaries to classify students into three groups or two successful/failed groups. In order to discriminate successful students from failed stu-dents much more clearly, it would be recommended to include other kind of information.

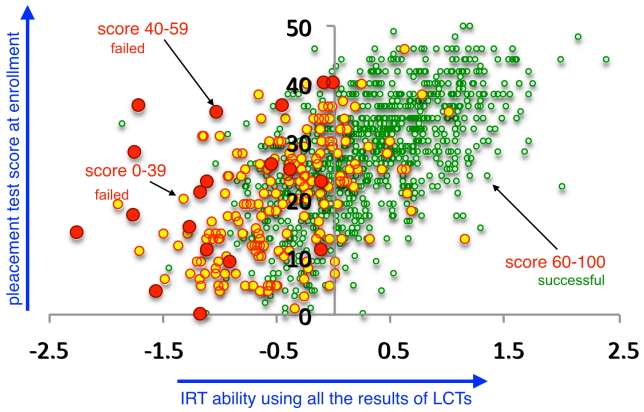

Figure 4: Correlations for the LCT results and the placement test results in three successful/failed groups (analysis basic in the first semester in 2017).

6 Trend of Estimated Students' Abilities Using Full Units Response Matrix in the IRT

We define $\theta_1(i,k)$ as student i's ability using response results from the 1st LCT to kth LCT, that is, the response matrix becomes a $N \times km$ size matrix. Figures 5 and 6 show trends of estimated abilities $\theta_1(i,k)$ for successful and failed groups. Looking at Figure 5, we can see that the estimated ability to each student seems to converge to a certain value as lectures go forward, and this means that the estimates become accurate. Figure 6 tells us that the estimated abilities show rather small variations around 0 value initially, but later they become lower as lectures go forward. Comparing to Figure 2, Figures 5 and 6 seem to characterize trends of estimated abilities for two groups with higher reliability than Figure 2 seems to. However, how can we use such a vague trend tendency to categorize the student groups into successful/failed students? We have to develop some tools to measure the similarity of the trend numerically.

7 Similarity Identification by Nearest Neighbor

In order to identify successful/failed students with much higher reliability in prediction, we here define the similarity via the nearest neighbor using the estimated ability trends as lectures goes forward. To do this, we use $\theta_1(i,k)$ defined in the previous section by incorporating the tentative response matrices $M_{m,k}(N,mk),\ k = 1,\ldots,K$ using LCTs from no.1 to no.k.

As an example to explain the similarity, we have provided Figure 7, where we can see three students' ability trends using estimated abilities from LCTs from no.1 to no.7. As lectures go forward, the estimated abilities seem to tend to certain values although the values are unreliable at early stages of the trends. We may assume that two final destination of success/failure may be the same if the estimated trends for abilities are close to each other. Although the use of only the full response matrices $M_{m,K}(N,mK)$ did not bear a reli-able results, we may expect that the trends of ability estimates by using response matrices $M_{m,k}(N,mk),\ k = 1,\ldots,K$ will give us much more information.

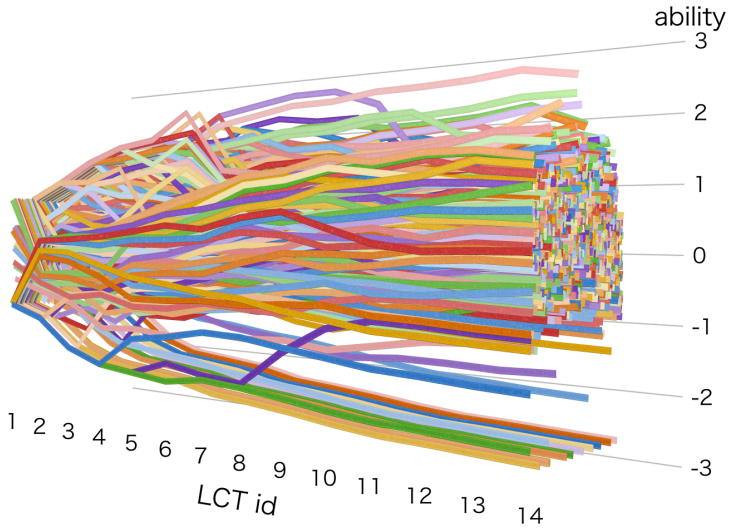

Figure 5: Trends of estimated abilities $\theta_1(i,k)$ for successful group (analysis basic in the first semester in 2017).

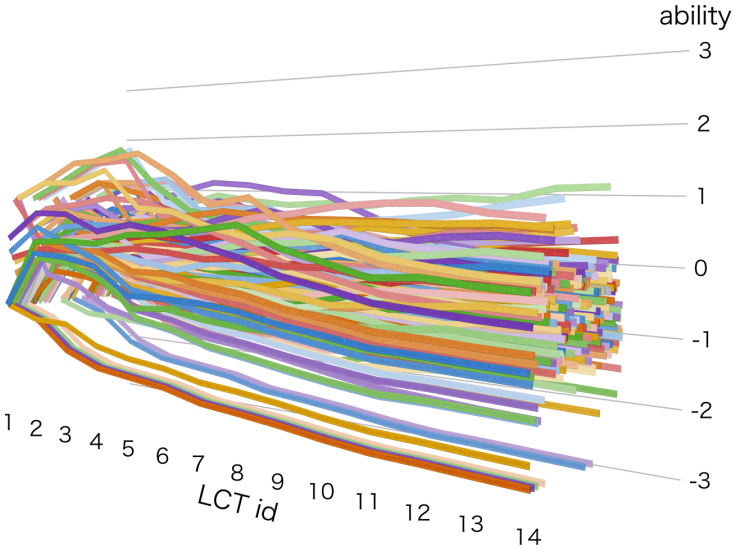

Figure 6: Trends of estimated abilities $\theta_1(i,k)$ for failed group (analysis basic in the first semester in 2017).

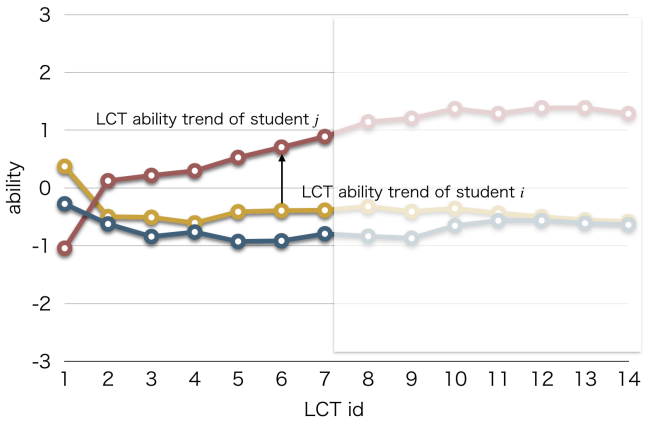

Figure 7: An example to explain the similarity via the nearest neighbor using the estimated ability trends using no.1 to no.7 LCTs.

We define the similarity of the two ability trends (i and j) by the following formula $S^k_{i,j}$ such that

$$S^k_{i,j} = \sqrt{\frac{1}{k}\sum_{l=1}^{k}(\theta_1(j,l)-\theta_1(i,l))^2},\ (i \neq j), \tag{3}$$

then, we can consider that $S^k_{i,j}$ expresses the mean distance between the trends of abilities for students i and j from the 1st LCT to kth LCT.

Sorting $S^k_{i,j}$ in ascending order in terms of j such as $S^k_{i,(1)} \leq \ldots, \leq S^k_{i,(N-1)}$, $S^k_{i,(j)}$ expresses the ordered statistics of $\{S^k_{i,j}\}$. We select the 10 least $S^k_{i,(j)}$ (i.e., $S^k_{i,(1)}, \ldots, S^k_{i,(10)}$), and obtain the mean value $\mu(i,k)$ of these end-term examination's success/failure indicator functions $\delta^k_{i,(j)}$, i.e., 1 for success and 0 for failure from jth final success/failure results. Then, $\mu(i,k) = 0, 0.1, \ldots, 0.9, 1$, and we can consider that $\mu(i,k)$ expresses the predicted value for success in the end-term examination. The computing method of values of $\mu(i,k)$ is related to the method of NNRMLR ([8], [9], [10]), where the nearest neighbor method is used to the regression, not to the classification. Even if the number of selected least $S^k_{i,(j)}$ is different from ten, the expected value for $\mu(i,k)$ would not be affected so much, although the standard deviation would be. We next show investigation results on the prediction accuracy using this similarity definition.

8 Identifying Successful/Failed Students Using the Similarity of the Trends of Estimated Students' Abilities in the IRT

We consider typical three cases in using LCT response results: 1) LCTs from no.1 to no.4, 2) LCTs from no.1 to no.7, 3) LCTs from no.1 to no.11.

Figure 8 shows a bar chart for the predicted numbers of students to be failed in the end-term examination in the case of analysis basic in the first semester in 2017. Upper green parts express the observed successful numbers of students; lower orange parts express the observed failed numbers of students. In the figure, we see a notation of $p \geq 0.3$, e.g., which is the same as $\mu(i,4) \geq 0.3$ when using LCTs from no.1 to no.4, and other

notations are expressed in a similar manner. For example, in the case of 2) LCTs from no.1 to no.7, and $p \geq 0.4$, we predicted that 173 students are to be failed in which 69 students are actually failed and 104 students are actually successful. These numbers are also seen in table 2.

Although the observed failed number of students, 206, is larger than the predicted value, the hitting ratio, 0.40, shows larger value to some extent than that shown in section 4 where the size of the response matrix is the maximum. Looking at all the bars in the figure, it should be noted that all the three cases using LCTs from no.1 to no.4, LCTs from no.1 to no.7, and LCTs from no.1 to no.11 reveal that the hitting ratios are larger than that shown in section 5 as long as $p \geq 0.4$.

From the confusion matrix in table 2, we can easily obtain the misclassification rates as shown in table 3. For example, to the cases $p \geq 0.3$, $p \geq 0.4$, and $p \geq 0.5$ using LCTs from no.1 to no.7 which used almost half of LCT, the misclassification rates are 0.28, 0.22, 0.18. All the misclassification rates in table 3 are smaller than or equal to that computed in section 5 which used all the LCT results in computing the IRT abilities.

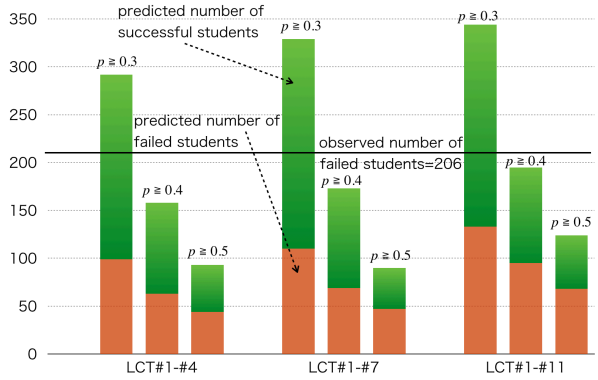

Figure 8: Numbers of successful/failed students using the similarity of the trends of estimated students' abilities (analysis basic in the first semester).

Table 4 shows the hitting ratios of the number of actually failed students to the number of predicted failed students corresponding to table 2. Since the hitting ratio using all the LCT results was 0.38 as mentioned in section 5, the hitting ratios using the nearest neighbor similarity are superior to that using the IRT abilities from all the LCT results.

9 Discussions

Comparing to the misclassification rate in the condition that only the numbers of suc-cess/failures are known, the predicted misclassification rates seem not to be informative so much. That is, in the analysis basic case the success rate is 0.82 and the failure rate is 0.18, then misclassification rate will be 0.18 if we assume that all the students are successful in the end-term examination. The estimated misclassification rates in table 3 are comparative at most or worse than that in the case mentioned above. However, it is totally absurd that we admit all the students are successful; we cannot find any students at risk. The hitting ratio is 0.

We actually want to know the students at risk, and an important point is that we canfind such students with high probability. From this viewpoint, the high hitting ratios are

Table 2: Confusion matrix determined by the nearest neighbor (analysis basic)

LCT #1-#4	$p \geq 0.3$		predicted	
		successful	failed	total
	successful	728	193	921
observed	failed	107	99	206
	total	835	292	1127
LCT #1-#4	$p \geq 0.4$		predicted	
		successful	failed	total
	successful	826	95	921
observed	failed	143	63	206
	total	969	158	1127
LCT #1-#4	$p \geq 0.5$		predicted	
		successful	failed	total
	successful	872	49	921
observed	failed	162	44	206
	total	1034	93	1127
LCT #1-#7	$p \geq 0.3$		predicted	
		successful	failed	total
	successful	702	219	921
observed	failed	96	110	206
	total	798	329	1127
LCT #1-#7	$p \geq 0.4$		predicted	
		successful	failed	total
	successful	817	104	921
observed	failed	137	69	206
	total	954	173	1127
LCT #1-#7	$p \geq 0.5$		predicted	
		successful	failed	total
	successful	878	43	921
observed	failed	159	47	206
	total	1037	90	1127
LCT #1-#11	$p \geq 0.3$		predicted	
		successful	failed	total
	successful	710	211	921
observed	failed	73	133	206
	total	783	344	1127
LCT #1-#11	$p \geq 0.4$		predicted	
		successful	failed	total
	successful	821	100	921
observed	failed	111	95	206
	total	932	195	1127
LCT #1-#11	$p \geq 0.5$		predicted	
		successful	failed	total
	successful	865	56	921
observed	failed	138	68	206
	total	1003	124	1127

Table 3: Misclassification rates by using the decision tree (analysis basic)

	LCT #1-#4	LCT #1-#7	LCT #1-#11
$p \geq 0.3$	0.27	0.28	0.25
$p \geq 0.4$	0.21	0.22	0.19
$p \geq 0.5$	0.19	0.18	0.17

Table 4: Hitting ratios of the number of actually failed students to the number of predicted failed students (analysis basic)

	LCT #1-#4	LCT #1-#7	LCT #1-#11
$p \geq 0.3$	0.34	0.33	0.39
$p \geq 0.4$	0.40	0.40	0.49
$p \geq 0.5$	0.47	0.52	0.55

informative to tell such students that you may fail if you insist to continue the same behavior as the current behavior. In the analysis basic case, 18% students failed, and we could identify about half of such students.

Using the obtained value of p which means the estimated failure probability using the trends of accumulated IRT results, we will be able to make alert to students for possible failures in the coming end-term examination. One method is to use the estimated value directly such that "you will fail in the end-term examination with higher probability than p". However, it seems that two-value information of failure or success is much clearer to students such that "you will fail in the end-term examination as long as you leave your learning style unchanged".

In such a situation, the threshold values for p will be informative when we alert students to the signal for possible failures. ROC (receiver operating characteristic) [4] curve may help to find such a threshold value. Figure 9 shows ROC curves when we use LCTs from no.1 to no.4, LCTs from no.1 to no.7, and LCTs from no.1 to no.11, in the case of analysis basic in the first semester. When we abbreviate false positive rate and true positive rate to FPR = FP / (FP + TN) and TPR = TP / (TP + FN), respectively, ROC curve represents the relationship between FPR and TPR, where FP, TN, TP, and FN are false positive, true negative, true positive, and false negative, respectively. In the figure, false positive rate in the abscissa means the ratio of the number of actually successful students in the predicted failed students to the total number of actually successful students, and true positive rate in the ordinate means the ratio of the number of actually failed students in the predicted failed students to the total number of actually failed students. In Figure 9, we see that $0.2 \leq p \leq 0.4$ could be used for the threshold value. However, from the viewpoint of importance of true positive rather than false positive, we recommend using the case of $p = 0.4$ in this case. We paid attention much to the true positive, i.e., students at risk.

To understand the hitting ratio shown in table 4, the recall precision curve may be useful. Figure 10 shows the recall precision curves when we use LCTs from no.1 to no.4, LCTs from no.1 to no.7, and LCTs from no.1 to no.11, in the case of analysis basic in the first semester. Recall and precision mean TP / (TP + FN) and TP / (TP + FP), respectively,

and the recall precision curve represents these two relationship. In the figure, recall in the abscissa means the ratio of the number of actually failed students in the predicted failed students to the total number of actually failed students, and precision in the ordinate means the ratio of the number of actually failed students in the predicted failed students to the total number of predicted failed students. Precision is equivalent to the hitting ratio shown in table 4. In section 5, we mention that the hitting ratio is 0.38 when we used all the LCTs in the semester. However, we see that this value is lower than those of 0.40, 0.40, 0.49 for $p \geq 0.4$ in LCTs from no.1 to no.4, LCTs from no.1 to no.7, and LCTs from no.1 to no.11 cases seen in table 4.

Figure 11 shows the predicted numbers of successful students and failed students to each p using results from LCTs from no.1 to LCT no.11 in the case of analysis basic in the first semester. We can see that the number of predicted successful students are becoming smaller when p is larger than 0.4.

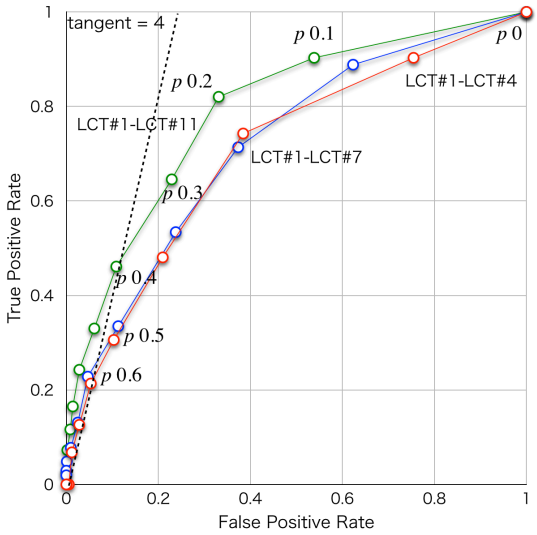

Figure 9: ROC curve (analysis basic in the first semester). False Positive Rate in the abscissa means the ratio of the number of actually failed students in the predicted failed students to the total number of actually successful students. True Positive Rate in the ordinate means the ratio of the number of actually failed students in the predicted failed students to the total number of actually failed students.

10 Concluding Remarks

Nowadays, it is crucial to identify students at risk of failing courses and/or dropping out as early as possible. By adopting online testing systems such as learning check testing (LCT) for every class to measure student comprehension of lectures, we can accumulate information for learning analytics. This paper is aimed at producing effective learning strategies for students at risk by utilizing the learning analytics obtained from the online testing.

To find students at risk as early as possible, we have proposed a newly developed method to identify students likely fail by analizing the similarity of the trends of estimated

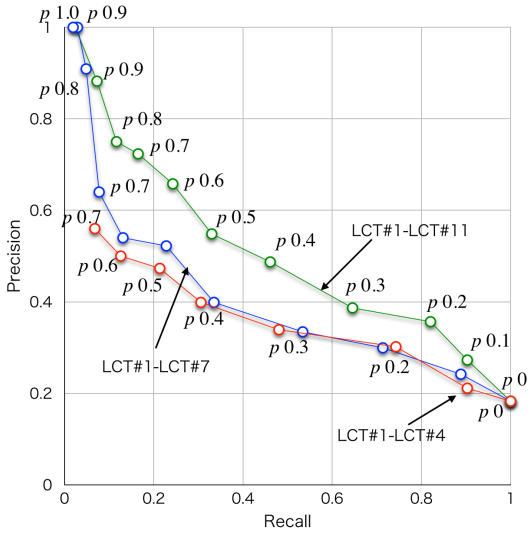

Figure 10: Recall Precision curve (analysis basic in the first semester). Recall in the abscissa means the ratio of the number of actually failed students in the predicted failed students to the total number of actually failed students. Precision in the ordinate means the ratio of the number of actually failed students in the predicted failed students to the total number of predicted failed students.

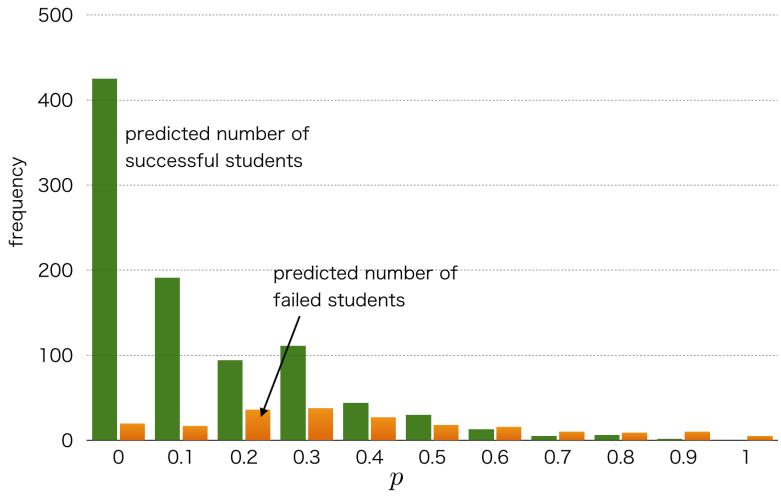

Figure 11: Bar charts for predicted numbers of successful students and failed students to each p using results from LCT no.1 to LCT no.11 (analysis basic in the first semester).

students' abilities in item response theory. The method uses the nearest neighbor methodology for determining the similarity of learning skill in the trends of estimated abilities. In the cases of analysis basic subject in the first semester in 2017, the proposed method can identify at an almost half of the students who subsequently failed the end-term examination from the early stages. This result is superior to the hitting ratio when we use the full data from the first to the last online testing results. We have applied ROC curve and recall precision curve to find the optimal threshold value for failure probability in precisely investigating the accuracy of the proposed method.

Therefore, pedagogical implications focusing on the findings of this research are the following: 1) it is important to accumulate learning data such as LCT week by week in order to support students at risk, 2) even if only the first half of the ability trends are available, we can predict the risk of failure at the end-term examination using the similarities of ability trends more accurately than using only the full of LCT abilities.

Acknowledgment

The author would like to thank mathematical staffs at Hiroshima Institute of Technology.

References

[1] R. de Ayala, The Theory and Practice of Item Response Theory. Guilford Press, 2009.

[2] R.S.J.D. Baker, Data mining for education. *In B. McGaw, P. Peterson, E. Baker, (Eds.) International Encyclopedia of Education (3rd edition)*, Elsevier, 2010.

[3] R.S.J.D. Baker, K. Yacef, The State of Educational Data Mining in 2009: A Review and Future Visions, Journal of Educational Data Mining, 1(1), 2009. pp.1-16.

[4] J.P. Egan, Signal detection theory and ROC analysis, Series in Cognition and Perception. Academic Press, New York, 1975.

[5] N. Elouazizi, Critical Factors in Data Governance for Learning Analytics, Journal of Learning Analytics, 1, 2014, pp. 211-222.

[6] D. Gasevic, S. Dawson, and G. Siemens, Let's not forget: Learning analytics are about learning, TechTrends, 59, 2015, pp. 64-71.

[7] R. Hambleton, H. Swaminathan, and H. J. Rogers, Fundamentals of Item Response Theory. Sage Publications, 1991.

[8] H. Hirose, Y. Soejima, K. Hirose, NNRMLR: A Combined Method of Nearest Neighbor Regression and Multiple Linear Regression, 6th International Workshop on e-Activity, pp.351-356, 2012.

[9] H. Hirose, NNRMLR2: An Improved Combined Method of Nearest Neighbor Regression and Multiple Linear Regression, 3rd IMS Asia Pacific Rim Meetings, Paper: 220534, 2014.

[10] H. Hirose, NNRMLR3: Further Improved Combination Method of Nearest Neighbor Regression and Multiple Linear Regression, 2nd International Symposium on Applied Engineering and Sciences, 2014.

[11] H. Hirose and T. Sakumura, Test evaluation system via the web using the item response theory, in Computer and Advanced Technology in Education, 2010, pp.152-158.

[12] H. Hirose, T. Sakumura, Item Response Prediction for Incomplete Response Matrix Using the EM-type Item Response Theory with Application to Adaptive Online Ability Evaluation System, IEEE International Conference on Teaching, Assessment, and Learning for Engineering, 2012, pp.8-12.

[13] H. Hirose, Yu Aizawa, Automatically Growing Dually Adaptive Online IRT Testing System, IEEE International Conference on Teaching, Assessment, and Learning for Engineering, 2014, pp.528-533.

[14] H. Hirose, Y. Tokusada, K. Noguchi, Dually Adaptive Online IRT Testing System with Application to High-School Mathematics Testing Case, IEEE International Conference on Teaching, Assessment, and Learning for Engineering, 2014, pp.447-452.

[15] H. Hirose, Y. Tokusada, A Simulation Study to the Dually Adaptive Online IRT Testing System, IEEE International Conference on Teaching, Assessment, and Learning for Engineering, 2014, pp.97-102.

[16] H. Hirose, Meticulous Learning Follow-up Systems for Undergraduate Students Using the Online Item Response Theory, 5th International Conference on Learning Technologies and Learning Environments, 2016, pp.427-432.

[17] H. Hirose, M. Takatou, Y. Yamauchi, T. Taniguchi, T. Honda, F. Kubo, M. Imaoka, T. Koyama, Questions and Answers Database Construction for Adaptive Online IRT Testing Systems: Analysis Course and Linear Algebra Course, 5th International Conference on Learning Technologies and Learning Environments, 2016, pp.433-438.

[18] H. Hirose, Learning Analytics to Adaptive Online IRT Testing Systems “Ai Arutte” Harmonized with University Textbooks, 5th International Conference on Learning Technologies and Learning Environments, 2016, pp.439-444.

[19] H. Hirose, M. Takatou, Y. Yamauchi, T. Taniguchi, F. Kubo, M. Imaoka, T. Koyama, Rediscovery of Initial Habituation Importance Learned from Analytics of Learning Check Testing in Mathematics for Undergraduate Students, 6th International Conference on Learning Technologies and Learning Environments, 2017, pp.482-486.

[20] H. Hirose, Dually Adaptive Online IRT Testing System, Bulletin of Informatics and Cybernetics Research Association of Statistical Sciences, 48, 2016, pp.1-17.

[21] H. Hirose, Difference Between Successful and Failed Students Learned from Analytics of Weekly Learning Check Testing, Information Engineering Express, Vol 4, No 1, 2018, pp.11-21.

[22] H. Hirose, A Large Scale Testing System for Learning Assistance and Its Learning Analytics, Proceedings of the Institute of Statistical Mathematics, Vol.66, No.1, 2018, pp.79-96.

[23] W.J.D. Linden and R.K. Hambleton, Handbook of Modern Item Response Theory. Springer, 1996.

[24] C. Romero, S. Ventura, P.G. Espejo, C. Hervas, Data Mining Algorithms to Classify Students. Proceedings of the First International Conference on Educational Data Mining, 2008, pp.8-17.

[25] T. Sakumura and H. Hirose, Making up the Complete Matrix from the Incomplete Matrix Using the EM-type IRT and Its Application, Transactions on Information Processing Society of Japan (TOM), 72, 2014, pp.17-26.

[26] T. Sakumura, H. Hirose, Bias Reduction of Abilities for Adaptive Online IRT Testing Systems, International Journal of Smart Computing and Artificial Intelligence (IJSCAI), 1, 2017, pp.57-70.

[27] G. Siemens and D. Gasevic, Guest Editorial - Learning and Knowledge Analytics, Educational Technology & Society, 15, 2012, pp.1-2.

[28] K, Spoon, J. Beemer, J.C. Whitmer, J.F. J.P. Frazee, J. Stronach, A.J. Bohonak, R.A. Levine, Random Forests for Evaluating Pedagogy and Informing Personalized Learning, Journal of Educational Data Mining, 8(2), 2016. pp.20-50.

[29] M. Sweeney, H. Rangwala, J. Lester, A. Johri, Next-Term Student Performance Prediction: A Recommender Systems Approach, Journal of Educational Data Mining, 8(1), 2016. pp.22-50.

[30] Y. Tokusada, H. Hirose, Evaluation of Abilities by Grouping for Small IRT Testing Systems, 5th International Conference on Learning Technologies and Learning Environments, 2016, pp.445-449.

[31] R. J. Waddington, S. Nam, S. Lonn, S.D. Teasley, , Improving Early Warning Systems with Categorized Course Resource Usage, Journal of Learning Analytics, 3, 2016, 263-290.

[32] A.F. Wise and D.W. Shaffer, Why Theory Matters More than Ever in the Age of Big Data, Journal of Learning Analytics, 2, pp. 5-13, 2015.