Computerized Adaptive Testing Method using Integer Programming to Minimize Item Exposure

Yoshimitsu MIYAZAWA[*1] Maomi UENO[*2]

[*1]The National Center for University Entrance Examinations, Tokyo, Japan,
[*2]The University of Electro-Communications, Tokyo, Japan

Computerized adaptive testing (CAT) estimates an examinee's ability sequentially and selects test items that have the highest accuracy for estimating the ability. However, conventional CAT selects the same items for examinees who have equivalent ability. Therefore, tests cannot be used practically under circumstances in which the same examinee can take a test multiple times. As described herein, we propose CAT that minimizes item exposure and which adaptively selects different items for examinees of equal ability, while retaining accuracy. This paper presents the method's effectiveness through a simulation experiment and with item pools used by actual test providers. Results confirmed that 1) the proposed method yielded the shortest test length. 2) The proposed method controls exposure and selects different items for different examinees. The non-uniformity of estimation was low. 3) The average item exposure of the proposed method was the lowest.

1. Introduction

Computerized adaptive testing (CAT) estimates an examinee's ability after every answer and selects an item with the highest accuracy of estimating the ability[Linden 10][Ueno 10][Ueno 13].

This item selection presents an important benefit: the number of items to be selected and the testing time can be decreased for an examinee without reducing the accuracy. Furthermore, the abilities of all examinees are measured using the same degree of accuracy. Nevertheless, conventional CAT has the following two shortcomings.

1. Conventional CAT is inapplicable for conditions in which the same examinee takes a test multiple times because the same group of items tends to be selected when the same examinee takes a test multiple times.

2. Conventional CAT provides the same items to examinees of equivalent ability. Therefore, not all items in an item pool can be used effectively. Excessive item exposure engenders disclosure of the item contents to examinees, and might therefore degrade reliability[Way 98].

Items with estimated difficulty and discrimination parameters must be prepared in advance to conduct CAT. Huge amounts of time and other costs thereby arise for test item preparation, especially for high-stakes tests. Accordingly, it is desirable to use all items in an item pool for the implementation of a test. We propose a framework of new CAT that resolves these difficulties in this study.

Constrained CAT (CCAT) is proposed to resolve difficulties posed by excessive exposure [Linden 10, Linden 98]. The method reported by van der Linden et al.[Linden 98, Linden 10], a constrained CAT method, constitutes an item set in which the number of exposed items and necessary answer times, and selects an item with the most information from the item pool at every item selection. This method can control exposure and can select a different item for each examinee. However, deviation of accuracy in the selected items produces great differences among examinees in terms of the test length and accuracy.

We propose a CAT to resolve this difficulty. It can select a different item while maintaining equal accuracy, even for an examinee with the same ability. Specifically, we propose a CAT method that can select items 1) with uniform test length, and 2) with uniform accuracy among tests, but 3) the items are not identical.

This study proposes CAT item selection using integer programming. The approach of the proposed method is described as presented below. 1) A group that satisfies test constraints as Fisher information is assembled using integer programming for minimizing item exposure. 2) An item that has the highest information is selected from the group.

The proposed method uses integer programming for minimizing item exposure and satisfying upper and lower bounds of information. Accordingly, a different item set can be selected, even for an examinee with equal ability. Moreover, the improved diversity of item selection is expected to encourage thorough use of items in an item pool and to mitigate deviation in exposure.

This paper demonstrates the effectiveness of the proposed method through simulation experiments and experiments using actual data.

2. Item Response Theory

In CAT, ability is estimated based on Item Response Theory (IRT)[Lord 80, Lord 68] to select items with the highest estimation accuracy[Baker 04, Linden 16a, Linden 16b]. Item response theory is a recent test theory based on mathematical models for which practical use is in progress lately

Contact: Yoshimitsu MIYAZAWA, The National Center for University Entrance Examinations, 2-19-23, Komaba, Meguro-ku, Tokyo, Japan, 03-5478-1275, miyazawa@rd.dnc.ac.jp

in widely diverse areas related to computer testing.

The two-parameter logistic model (2PLM) has been used for many years as an item response model that is broadly applicable to such binary data. This study also adopts 2PLM, for which the probability of a correct answer given to item i by an examinee j with ability $\theta \in (-\infty, \infty)$ is assumed as

$$p(u_i = 1|\theta) = \frac{1}{1 + \exp[-1.7a_i(\theta - b_i)]}. \tag{1}$$

The standard error of ability estimation based on the item response theory is known to approach the reciprocal of Fisher information asymptotically [Lord 80]. Accordingly, item response theory usually employs Fisher information as an index representing the accuracy.

In 2PLM, the Fisher information is defined when item i provided to an examinee with ability θ using the following equations[Birnbaum 68].

$$I_i(\theta) = \frac{[p'(u_i = 1|\theta)]^2}{p(u_i = 1|\theta)[1 - p(u_i = 1|\theta)]} \tag{2}$$

where

$$p'(u_i = 1|\theta) = \frac{\partial}{\partial\theta} p(u_i = 1|\theta). \tag{3}$$

That result implies that the examinee ability can be ascertained near ability θ using an item with much Fisher information $I_i(\theta)$. Accordingly, it is expected that ability estimation can be implemented by selecting items with much Fisher information at a given ability for each examinee. Based on this concept, an item selection method of computerized based testing, CAT, presents items with much Fisher information. The total of Fisher information of an item set contained in a test presented to an examinee is called test information, which represents the test estimation accuracy.

3. Computerized Adaptive Testing

In CAT, items are selected from a given item set with known item parameters, using the following procedures.

1. The examinee ability is initialized.
2. An item that maximizes Fisher information for given ability is selected from the item pool and is presented to an examinee.
3. The estimated ability of the examinee is updated from the correct/wrong answer data to the item.
4. Procedures 2 and 3 are repeated until the update difference of the estimated ability of the examinee reaches a constant value ϵ or less.

Consequently, for a small number of items to be selected compared with a fixed test, repeating item selection based on maximizing information and estimation of an examinee ability engenders high ability estimation accuracy.

Unfortunately, in CAT, it is highly likely that the same set of items will be selected for examinees who have equal ability. Conventional CAT cannot be used practically under situations in which the same examinee can take a test multiple times. In addition, to follow the normal distribution for ability items with high information, the average value $\theta = 0$ is frequently selected. Therefore, some items in an item pool might not be used effectively. Excessive exposure of items leads to disclosure of the item contents to examinees, and might therefore degrade the test reliability[Way 98].

To resolve this difficulty, we propose a CAT method that minimizes item exposure. It can select a different item while maintaining the same accuracy, even for an examinee with equal ability.

4. Proposed Method

The concept of the proposed method is to assemble a group with test constraints as Fisher information, but with different items using integer programming, to select an item with higher information, and to minimize item exposure from the group. Details of the proposed method are described below.

1. The estimated ability of an examinee is initialized.
2. The group maximizing Fischer information is assembled using the integer programming presented below.

$$\text{Minimize } y = \sum_{i=1}^{I} e_i x_i \tag{4}$$

subject to

$$\sum_{i=1}^{I} I(\theta_k) x_i \geq r_k, k = 1, 2, \cdots, K \tag{5}$$

$$\sum_{i=1}^{I} I(\theta_k) x_i \leq s_k, k = 1, 2, \cdots, K \tag{6}$$

$$\sum_{i=1}^{I} x_i = n(\text{Test length}) \tag{7}$$

To expand the method proposed by Adema (1989) [Adema 92] in which test forms are assembled using integer programming, we propose an optimization problem in which we embed an objective function minimizing item exposure. The lower boundaries and the upper boundaries of test information function at a set of the examinee's ability, $\Theta = \{\theta_1, \ldots, \theta_K\}$, is r_k and s_k. Also, $I(\theta_k)$ denotes the test information function at the examinee's ability θ_k. The exposure count of item i is e_i. If item i is selected into the group, then $x_i = 1$; otherwise $x_i = 0$.

3. An item is selected from a group. Then response data are generated with the given true ability and item parameters.
4. Ability $\hat{\theta}$ is estimated by expected a posteriori (EAP) [Baker 04].

表 1: Results obtained using simulation data

Item pool size	methods	Avg. test length	non-uniformity of estimation	Max. No. exposure item	Avg. exposure item
	CAT	19.99 (2.26)	0.03	1000	39.98 (120.93)
500	CCAT	23.99 (6.2)	0.21	60	47.98 (23.15)
	Proposal	9.35 (2.15)	0.09	65	18.7 (17.1)
	CAT	21.33 (2.15)	0.03	1000	21.33 (82.22)
1000	CCAT	28.37 (8.98)	0.86	30	28.37 (6.75)
	Proposal	9.73 (2.46)	0.09	42	9.73 (8.74)
	CAT	22.48 (2.1)	0.02	1000	11.24 (57.47)
2000	CCAT	28.56 (11.35)	2.33	15	14.28 (3.19)
	Proposal	9.61 (2.66)	0.08	19	4.8 (4.02)

表 2: Results obtained using actual data

Item pool size	methods	Avg. test length	non-uniformity of estimation	Max. No. exposure item	Avg. exposure item
	CAT	14.82 (3.41)	0.05	1000	15.15 (78.5)
978	CCAT	26.56 (6.36)	0.19	30	27.15 (8.42)
	Proposal	11.99 (3.15)	0.07	75	12.26 (10.31)

5. Procedures 2–4 are repeated until the update difference of the estimated ability decreases to ϵ or less.

The proposed method selects an item from a different group for each examinee. Accordingly, a different item is expected to be selected, even to an examinee of equivalent ability. Furthermore, improved diversity of item selection is anticipated to encourage thorough use of items in an item pool and to mitigate deviation in exposure.

5. Simulation Experiment

The simulation experiment procedure is the following.

1. An item pool comprising 500, 1000, or 2000 items is generated. The true values of parameters of each item are set randomly from $a_i \sim U(0,1)$, $b_i \sim N(0,1)$.
2. The true ability of an examinee is sampled from $\theta \sim N(0,1)$.
3. The estimated ability of an examinee is initialized to $\hat{\theta} = 0$.
4. An item is selected from an item pool using each method. Then response data are generated with the given true ability and item parameters.
5. The ability $\hat{\theta}$ is estimated using EAP.
6. Procedures 4 and 5 are repeated until the update difference of the estimated ability decreases to ϵ or less. Also, ϵ is set to 0.05, which is used conventionally for actual CAT[Linden 10].
7. Procedures 2–6 are repeated 1000 times to obtain statistical values for the following indices using a delivery pattern and answer data obtained: a) the length of a test, b) the non-uniformity of ability estimation accuracy, and c) the exposure of each item.

Table 1 presents the results. Results confirmed that the proposed method yielded the shortest test length under all conditions. Selecting an item with less information for the initial value of θ rather than selecting an item with more information is known to achieve faster convergence of ability estimation when the initial value of θ is distant from the true ability of an examinee. The proposed method constrains the number of items with uniformity conditions. Therefore, it has a property of not selecting items with an extremely large amount of information only to a certain estimated value. This property shortens the test length, thereby reducing the exposure of items. The proposed method also indicates the smallest standard deviation of test length under all conditions. The proposed method selects items from an item set of uniform information. Therefore, it can render the number of items to take for convergence of the estimated ability uniform.

The non-uniformity of estimation using conventional CAT was lowest. It repeatedly selects some item sets with much information. However, the proposed method controls exposure and selects different items for each examinee. The non-uniformity of estimation was low because of the deviation of measurement accuracy in the selected items.

The maximum exposure was the least when CCAT was used. Results demonstrate that CCAT can constrain maximum exposure directly in the same manner as the fraction of different items. This result is interpreted as attributable to the exposure setting to use as many items in an item pool as possible in this experiment. The averages of exposure obtained using the proposed method were the lowest. Actually, CCAT constrained only the maximum of exposure directly, so that the deviation of exposure could not be controlled.

6. Simulation Using Real Data

This chapter explains evaluation of the effectiveness of the proposed method using real data. An experiment was conducted using an item pool of real data. The item pool contained 978 items. Table 2 presents the experimentally obtained results. Table 2 suggests the following characteristics: 1) The test length produced using the proposed method is the shortest. The standard deviation of test length produced using the proposed method is the smallest, which suggests that the test length selected to examinees has little dispersion. 2) Non-uniformity of estimation by the proposed method is the second highest to CAT, so the same accuracy was maintained using the proposed method. 3) The maximum exposure was smallest by CCAT. The average tended to be small when obtained using the proposed methods.

7. Conclusions

This study has examined a proposed CAT implementation in which different items can be selected, even by an examinee with equivalent ability, while maintaining equal accuracy. Specifically, we proposed a method by which a group is assembled using integer programming: an item is selected from the group. Using simulation experiments and experiments using real data, some points have been verified as benefits of the proposed method.

参考文献

[Adema 92] Adema, J. J.: Implementations of the branch-and-bound method for test construction problems, *Methodika*, Vol. 6, No. 2, pp. 99–117 (1992)

[Baker 04] Baker, F. B. and Kim, S.-H. eds.: *Item Response Theory: Parameter Estimation Techniques*, CRC Press (2004)

[Birnbaum 68] Birnbaum, A.: Some latent trait models and their use in inferring an examinee's ability, in Lord, F. M. and Novick, M. R. eds., *Statistical theories of mental test scores*, pp. 397–479, Addison-Wesley (1968)

[Linden 98] Linden, van der W. J. and Reese, L. M.: A model for optimal constrained adaptive testing, *Applied Psychological Measurement*, Vol. 22, No. 3, pp. 259–270 (1998)

[Linden 10] Linden, van der W. J. and Glas, C. A. W. eds.: *Elements of Adaptive Testing*, Springer (2010)

[Linden 16a] Linden, van der W. J. ed.: *Handbook of Item Response Theory, Volume One: Models*, Chapman and Hall/CRC (2016)

[Linden 16b] Linden, van der W. J. ed.: *Handbook of Item Response Theory, Volume Two: Statistical Tools*, Chapman and Hall/CRC (2016)

[Lord 68] Lord, F. and Novick, M. R.: *Statistical Theories of Mental Test Scores*, Addison-Wesley (1968)

[Lord 80] Lord, F. M.: *Applications of Item Response Theory To Practical Testing Problems*, Lawrence Erlbaum Associates (1980)

[Ueno 10] Ueno, M. and Songmuang, P.: Computerized adaptive testing based on decision tree, in *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on*, pp. 191–193 (2010)

[Ueno 13] Ueno, M.: Adaptive testing based on bayesian decision theory, in *International Conference on Artificial Intelligence in Education*, pp. 712–716 (2013)

[Way 98] Way, W. D.: Protecting the integrity of computerized testing item pools, *Educational Measurement: Issues and Practice*, Vol. 17, pp. 17–27 (1998)