The Bayesian Approach and Suggested Stopping Criterion in Computerized Adaptive Testing

Loc Nguyen

***Abstract*—** The computer-based tests have more advantages than the traditional paper-based tests when there is the boom of internet and computer. Computer-based testing allows students to perform the tests at any time and any place and the testing environment becomes more realistic. Moreover, it is very easy to assess students' ability by using the computerized adaptive testing (CAT). The CAT is considered as the branch of computer-based testing but it improves the accuracy of test core when CAT systems try to choose items which are suitable to students' abilities; such items are called adaptive items. The important problem in CAT is how to estimate students' abilities so as to select the best items for students. There are some methods to solve this problem such as maximization likelihood estimation but we apply the Bayesian approach into computing ability estimates. In this paper, we suggest the stopping criterion for CAT algorithm: the process of testing ends only when student's knowledge becomes saturated (she/he can't do test better or worse) and such knowledge is her/his actual knowledge.

***Keywords*—**Bayesian inference, computerized adaptive test.

I. Introduction

A. Item Response Theory

Item Response Theory (IRT) [1] is defined as a statistical model in which examinees can be described by a set of one or more ability scores that are predictive, through mathematical models, linking actual performance on test items, item statistics, and examinee abilities. Note that the term "item" indicates test or exam. Given examinee *j* and item *i*, IRT is modeled as a function of a true ability of examinee *j* (denoted θ_j) and three parameters of item *i* (denoted a_i, b_i, c_i). This function so-called Item Response Function (IRF) or Item Characteristic Curve (ICC) function computes the probability of a correct response of examinee *j* to item *i*.

$$IRF(\theta_j) = \Pr(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{a_i(\theta_j - b_i)}}$$

ICC, a variant of logistic function, is plotted in following figure with $a_i=2.0$, $b_i=0$, $c_i=0.25$.

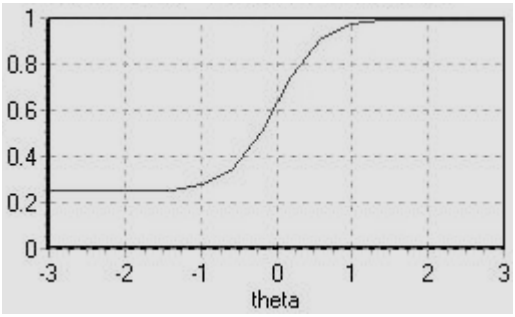

Fig. 1 Item Characteristic Curve

The horizontal axis θ is the scale of examinee's ability, ranging from *-3.0* to *+3.0*. The vertical axis is the probability of correct response to this item specified by three parameters: $a_i=2.0$, $b_i=0$, $c_i=0.25$. The left-hand of curve shows an easy item when the probability of correct response is high for low-ability. The center of curve shows a medium-difficulty item when the probability of correct response is around the middle of the ability scale. The right-hand of curve shows a difficult item when the probability of its correct response is low for most of ability scale. The lower asymptote at $c_i=0.25$ indicates the probability of correct response for examinee with lowest ability and otherwise for the upper asymptote at *1.0*.

ICC measures examinee's proficiency based on her/his ability and some properties of item. Every item *i* has three parameters a_i, b_i, c_i which are specified by experts or statistical data.

- The a_i parameter so-called *discriminatory parameter* tells how well the item discriminates between examinees whose abilities aren't different much. It defines the slope of the curve at the inflection point. The higher is the value of a_i, the steeper is the curve. In case of steep curve, there is a large difference between the probabilities of a correct response for examinees whose ability is slightly below of the inflection point and examinees whose ability is slightly above the inflection point.
- The b_i parameter so-called *difficult parameter* indicates how difficult the item is. It specifies the location of inflection point of the curve along the θ axis (examinee's ability). Higher value of b_i shifts the curve to the right and implicates that the item is more difficult.
- The c_i parameter so-called *guessing parameter* indicates that the probability of a correct response to item of low-

Loc Nguyen is with the University of Science, Ho Chi Minh city, Vietnam (corresponding author to provide phone: 84975250362; e-mail: ng_phloc@yahoo.com).

ability examinees is very close to c_i. It determines the lower asymptote of the curve.

B. Computerized Adaptive Testing

Computerized Adaptive Testing (CAT) [3] is the iterative algorithm which begins providing examinee an (test) item so as to be best to her/his initial ability; after that the ability is estimated again and the process of item suggestion is continued until a stopping criterion is met. This algorithm aims to make a series of (test) items which are evaluated to become chosen items that suitable to examinee's ability. The set of items from which system picks ones up is called as (item) pool. The items having chosen and given to examinee compose the adaptive test. CAT includes following steps:

1. The initial ability of examinee must be defined and the items (in the pool) that have not yet been chosen are evaluated to choose the best one which is the most suitable to examinee's current ability estimate.
2. The best next item is chosen to give to examinee and the examinee responds. Such item is changed from pool to adaptive test.
3. A new ability estimate is computed based on the responses to all of the chosen items.
4. Steps 1 through 3 are repeated until a stopping criterion is met.

Note that the chosen item is also called the administered item and the process of choosing best item is also called the administration process. The ability estimate is the value of θ which is fits best to the model and reflects current proficiency of examinee in item but it is not imperative to define precisely the initial ability because the final ability estimate may not be closed to initial ability. The stopping criterion could be time, number of administered items, change in ability estimate, maximum-information of ability estimate, etc.

In step 1, there is the question: "how to evaluate the items to choose the best one". So each item i is qualified by the amount of information or entropy at given ability θ; such entropy is denoted $I_i(\theta)$. The best next item is the one that provides the most information or has highest value of $I_i(\theta)$.

$$I_i(\theta) = \frac{\Pr_i'(\theta)^2}{\Pr_i(\theta)(1-\Pr_i(\theta))}$$

Where $Pr_i(\theta)$ is the ICC function and $Pr_i'(\theta)$ is the first-order derivative of $Pr_i(\theta)$..

The entropy $I_i(\theta)$ reflects how much the item i matches examinee's ability. The item should be too easy or too difficult. Given ability θ, the sum of entropies over items in test which tells the qualification of such pool at ability θ is denoted $I(\theta)$.

$$I(\theta) = \sum_i I_i(\theta)$$

In step 3, there are some methods to compute the ability estimate such as maximum-likelihood, weighted likelihood [4], etc. But I aim to apply Bayesian approach into specifying ability estimate according to [2].

II. Bayesian Approach For CAT

In our method, CAT also includes four steps as above but we should redefine some concepts in order to take advantage of Bayesian rule. Suppose the indexes of items in pool are denoted as $i = \overline{1,I}$ and the indexes of these items in the adaptive test is denoted as $k = \overline{1,K}$ and so is the index of item i in the pool administered as the item k in the test is denoted i_k. Suppose $S_k = \{i_1, i_2, i_{k-1}\}$ is the set of previous *k-1* items that are administered and in a test now; of course they correspond with a set of *k-1* responses denoted as $U_k=\{u_{i_1}, u_{i_2},\ldots, u_{i_{k-1}}\}$. The set of items in the pool remaining after chosen items is denoted as $R_k=\{1,\ldots, I\} \setminus U_k$. The ICC function is written in general:

$$\Pr_i(\theta) = c_i + (1-c_i)\frac{e^{a_i(\theta-b_i)}}{1+e^{a_i(\theta-b_i)}}$$

Where θ is the examinee's ability and a_i, b_i and c_i are discriminatory parameter, difficult parameter and guessing parameter, respectively.

The likelihood function associated with the responses on the first *k–1* items is denoted as below:

$$L(\theta \mid u_{i_1},\ldots,u_{i_{k-1}}) = \prod_{j=1}^{k-1}\frac{(e^{a_{i_j}(\theta-b_{i_j})})^{u_{i_j}}}{1+e^{a_{i_j}(\theta-b_{i_j})}}$$

The second-order derivative of the likelihood reflects the curvature of the observed likelihood function at θ. The observed information measure is defined the negative of such second-order derivative:

$$J_{u_{i_1},\ldots,u_{i_{k-1}}}(\theta) = -\frac{\partial}{\partial\theta^2}\ln L(\theta \mid u_{i_1},\ldots,u_{i_{k-1}})$$

The expectation of observed information measure is expected information measure.

$$I_{u_{i_1},\ldots,u_{i_{k-1}}}(\theta) = E(J_{u_{i_1},\ldots,u_{i_{k-1}}}(\theta))$$

Suppose the prior for the unknown value of examinee's ability is assumed as $g(\theta)$. The function $g(\theta)$ is also called as the prior distribution of θ. According to Bayesian rule, the posterior distribution of ability estimate is computed as below:

$$g(\theta \mid u_{i_1},\ldots,u_{i_{k-1}}) = \frac{L(\theta \mid u_{i_1},\ldots,u_{i_{k-1}})g(\theta)}{\int L(\theta \mid u_{i_1},\ldots,u_{i_{k-1}})g(\theta)d\theta}$$

In step 3, after the responses to *k-1* items, it is necessary to determine the ability estimate denoted $\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}}$ or $\hat{\theta}$ in brief. According to [2], $\hat{\theta}$ is the expectation of the posterior distribution $g(\theta \mid u_{i_1},\ldots,u_{i_{k-1}})$.

$$\hat{\theta} = \hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}} = E(\theta \mid u_{i_1},\ldots,u_{i_{k-1}})$$
$$= \int \theta g(\theta \mid u_{i_1},\ldots,u_{i_{k-1}})d\theta$$

This technique based on the posterior distribution of θ has

an advantage when the ability estimate $\widehat{\theta}$ always exists and is easy to compute. Another method used to compute ability estimate is to determine the value $\widehat{\theta}$ that maximizes the likelihood function in over all possible values of θ. Such $\widehat{\theta}$ is considered as the ability estimate.

$$\widehat{\theta} = \widehat{\theta}_{u_{i_1},...,u_{i_{k-1}}} = \arg\max_{\theta}(L(\theta \mid u_{i_1},...,u_{i_{k-1}}))$$

This method isn't optimal because maybe the maximizer $\widehat{\theta}$ isn't found out and the essence of how to determine $\widehat{\theta}$ comes back the hazard of solving the expectation maximization (EM) problem. It isn't asserted that there is always solution to such problem. In general, ability estimation is the most important task including three stages:

- At the beginning of the item-selection procedure, we need to specify the prior distribution $g(\theta)$, for example, Gaussian density function.
- During the test, we determine the entropy of each item and compute the ability estimate by using the posterior distribution (Bayesian rule).
- At the end of the test, the final estimate is determined reflecting the examinee's mastery over the test.

III. Suggested Stopping Criterion

In normal the stopping criterion in CAT algorithm is often the number of (test) items, for example, if the test has ten items then the examinee's final estimate is specified at 10^{th} item and the test ends. This form is appropriate to examination in certain place and certain time and user is the examinee who passes or fails such examination.

Suppose in situation that user is the learner who wants to gains knowledge about some domain as much as possible and she/he doesn't care about passing or failing the examination. In other words, there is no test or examination and the learners prefer to study themselves by doing exercise. There is an exercise and the items are questions that belong to this exercise. It is possible to use another stopping criterion in which the exercise ends only when the learner can't do it better or worse. At that time her/his knowledge becomes saturated and such knowledge is her/his actual knowledge. The standard deviation of student's ability estimate is used to assess the saturation of her/his knowledge. In fact, the standard deviation denoted σ is the root of the variance of the posterior distribution of θ.

$$\mathrm{var}(\theta) = E(\theta - E(\theta \mid u_{i_1},...,u_{i_{k-1}}))^2$$

$$= \int (\theta - E(\theta \mid u_{i_1},...,u_{i_{k-1}}))^2 g(\theta \mid u_{i_1},...,u_{i_{k-1}}) d\theta$$

$$\sigma(\theta) = \sqrt{\mathrm{var}(\theta)}$$

Given threshold ξ, if the standard deviation σ is less than ξ then the CAT algorithm stops. There is no restriction for the number of (question) items in exercise.

IV. Conclusion

We recognized that CAT gives us the excellent tools for assessing student's ability. The CAT algorithm includes four steps in which step 3 is the most important when student's ability estimate is determined. There is an advantage of Bayesian approach when the ability estimate which is the expectation of posterior probability always exists and is easy to compute. However, the quality of posterior probability depends on the prior probability which may be pre-defined by experts. In the future trend, we intend to find out the technique for learning training data so as to specify precisely prior probabilities.

Moreover we propose the stopping criterion for CAT algorithm in which given threshold ξ, if the standard deviation of student's ability estimate is less than ξ then the CAT algorithm stops. The goal of this technique is that the exercise ends only when the student can't do it better or worse. It means that her/his knowledge becomes saturated and such knowledge is her/his actual knowledge. This method is only suitable to training exercises because there is no restriction for the number of (question) items in exercises. Conversely, in the formal test, the examinee must finish such test right before the decline time and the number of items in formal test is fixed.

References

[1] Frank B. Baker. The basics of item response theory. Published by the ERIC Clearinghouse on Assessment Evaluation 2001.

[2] Bock, R. D. and Mislevy, R. J. Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6, 431-444 (1988).

[3] Wim J. van der Linden and Gees A.W. Glas. Computerized Adaptive Testing: Theory and Practice. Kluwer Academic Publishers ©2002. ISBN: 0-7923-6425-2.

[4] Warm, T. A. Weighted likelihood estimation of ability in item response theory with tests of finite length. Psychometrika, 54, 427-450 (1989).