IRT-Item Response Theory Assessment for an Adaptive Teaching Assessment System

DR. MANSOOR AL-A'ALI

Department of Computer Science
College of Information Technology
University of Bahrain
KINGDOM OF BAHRAIN

Abstract:- Item Response Theory, IRT, is the study of scoring tests and questions based on assumptions concerning the mathematical relationship between examinee's ability (or other hypothesized traits) and questions' responses. Adaptive student tests, which are based on IRT, have many advantages over the conventional tests. The conventional method is used to give the traditional test scoring and estimating IRT questions' parameters. Our estimation of IRT questions' parameters is based on the Least Square method, a well-known statistical method. Our major goal is to minimize the number of questions in the adaptive test to reach the final level of the students by modifying the equation of estimation of student ability level. New factors are considered, namely initial student ability, subject's difficulty, number of exercises covered by the teacher and number of lessons covered by the teacher. Moreover, students' attitudes toward adaptive tests and their results were measured by a questionnaire. Our conventional exam results were compared with adaptive results and used in determining IRT parameters. We had modified the formula of estimating student ability level and had positive results in minimizing the number of questions in adaptive tests.

Key words:- IRT Item Response Theory, testing methods, adaptive testing, student assessment.

1 Introduction

Item Response Theory, IRT, is the study of scoring tests and questions based on assumptions concerning the mathematical relationship between examinee's ability (or other hypothesized traits) and questions' responses. Adaptive student tests, which are based on IRT, have many advantages over the conventional tests. The first advantage is that IRT adaptive testing contributes to the reduction of the length of the test because the adaptive test gives the most informative questions when the student shows a mastery level on a certain field. Secondly, the test may be better tailored for individual students.

Adaptive assessment is clearly a good way forward in teaching especially with availability of computers in all schools. As we know evaluation and assessment are an integral part of learning. A good objective test at the end of each learning objective can reveal a great deal about the level of understanding of the learner [Proc, 2003].

Conventional exams suffer from certain problems that must be considered carefully. First, assessments and assignments are normally given to test the different capabilities of students who are in the range of poor to excellent and thus all students have to meet or resolve the same standard questions, which mostly do not match their own capabilities, either lower or higher. This means that students with high capabilities may waste their time in solving average assignments without facing the excitement or challenge which would make the assignment or assessment interesting. Another aspect that we should keep in mind considerably is that a percent-correct score of 75 on a test containing easy items has a different meaning than a score of 75 on a test containing difficult items [Daniel, 2000]. Second, testing for knowledge and understanding in the context of a specific course typically involves administering the same set of test items or questions to all the students enrolled in the course, usually at the same sitting for the examination [Classical Test, 2000].

When we consider Classical Test Theory (CTT) we realize that CTT has a number of deficiencies. One of the problems of CTT is that it is test-oriented rather than item-oriented, i.e., in the classical true score model there is no regard for an examinee's response to a given item. As a result, CTT does not allow predictions to be made about how an individual or group of examinees will perform on a given item [Stephanie, 1997].
To formulate a complete understanding of the assessment process in schools, we look at some drawbacks of pencil and paper tests. Some drawbacks of conventional pencil and paper tests are scoring and feedback. Instructors need a lot of time before correcting tests' papers. This means that examinee cannot be informed about the result of the test just after its completion. We know that immediate feedback is also important to the students for psychological reasons. It is motivational; it helps them focus and informs them if they have to work harder.

One of the solutions to these problems is to use CATs Computer Adaptive Tests. Advantages of CAT can include shorter quicker tests, flexible testing schedules, increased test security, better control of item exposure, better balancing of test content areas for all ability levels, quicker test item updating, quicker reporting, and a better test-taking experience for the test-taker [John, 2000]. CATs are widely spread these days and they give good results in many educational degrees. CATs are used in many professional certification programs. Novell successfully introduced CATs into its certification program in 1991. The Educational Testing Service, the world’s largest testing organization, published the Graduate Record Exam (GRE) as an adaptive test in 1993. TOEFL is also using CAT. The Nursing Boards converted completely from paper-based testing to a computerized adaptive test in 1994.

Assessment can guide improvement-presuming they are valid and reliable-if they motivate adjustments to the educational system [Valerie & Brendon, 2003]. "The question is no longer whether assessment must incorporate technology. It is how to do it responsibly, not only to preserve the validity, fairness, utility, and credibility of the measurement enterprise but, even more so, to enhance it" [Bennett, 2002].

Intelligent tutoring systems permit the modeling of an individual learner and with that modeling comes the knowledge of how to perform an individualized assessment. Examinations can once again be tailored to meet the individual needs of a particular learner [Janson, 1996]. In contrast with paper-and-pencil multiple-choice tests, new assessments for complex cognitive skills involve embedding assessments directly within interactive, problem-solving, or open-ended tasks [Bennett & Persky, 2002].

One of our main objectives in this research was to study IRT in order to reduce the number of questions before reaching stability. It had been shown [Evangelia, 2001] and [Eggen, 2000] that after 13-15 questions, the level of students becomes stable. Therefore, the time of adaptive test is less than the pencil and paper test. Our hypothesis is that the starting level of the adaptive test is arbitrary. Our starting point in the adaptive test will start from the conventional level that we get from stage 1 of the method previously mentioned. Our target was get a stable level after 7 questions only and hence reducing the time of the test and the length of the test. In order to test our hypothesis we had to build a system and evaluate and enhance IRT. We measured the effectiveness of IRT through comparing the students' level in new system with teachers' student level.

Section 2 of the paper discusses the different types of tests. Section 3 discusses the details of IRT. Section 3 explains the concept testing methods and adaptive testing using IRT. Section 4 presents our results and findings through experimenting with IRT and enhancing it.

2 Adaptive Assessment Algorithm

The IRT algorithm aims to provide information about the functional relation between the estimate of the learner's proficiency on a concept and the likelihood that the learner will give the correct answer to a specific question [Evangelia, 2001]. Figure 1 is a schematic representation of an adaptive test.

![Figure 1: a schematic representation of an adaptive test](image-url)
In a conventional test, two matters are considered; the time and the length of the test. In adaptive test, possible termination criteria are:

1. The number of questions posed exceeds the maximum number of questions allowed, i.e., the number of questions is fixed.
2. The accuracy of the estimation of the learner's proficiency reaches some desired value.
3. Time limitations: most popular adaptive tests have a time limit. Although time limitation is not necessary in adaptive testing, it prevents the student to take a test for a very long time, gets tired and his tiredness affects negatively his score.
4. No more relevant items in the bank: when, in cases, the item bank is small or questions with difficulty level suitable for a student does not exist, the test must terminate at this point.

3 Item Characteristic Curve in IRT

The item characteristic curve (ICC) is the basic building block of item response theory. There are two technical properties of an item characteristic curve. The first one is the difficulty of the item. Under item response theory, the difficulty of an item describes where the item functions along the ability scale. The second technical property is discrimination, which describes how well an item can differentiate between examinees having abilities below the item location and those having abilities above the item location [Baker, 2001].

At each ability level, there will be a certain probability that an examinee with that ability will give a correct answer to the item. This probability will be denoted by \( P(\theta) \). Its formula is given by

\[
P(\theta) = \frac{1}{1 + e^{-L}} = \frac{1}{1 + e^{-a(\theta - b)}}
\]

Where:
- \( b \) is the difficulty parameter
- \( a \) is the discrimination parameter
- \( L = a(\theta - b) \) is the logistic deviate (logit)
- \( \theta \) is an ability level.

The importance of Item discrimination comes from that it relates the strength of the relationship between a test item and the underlying (and unobservable) attribute being measured, e.g., knowledge or learning [Classical Test, 2000].

In the case of a typical test item, this \( P(\theta) \) will be small for examinees of low ability and large for examinees of high ability [Baker, 2001]. Adding one more factor to the pervious factor the formula will be [Evangelia, 2001]:

\[
P(\theta) = c + \frac{1-c}{1 + e^{-2(\theta - b)}}
\]

where \( c \) factor is the guessing factor. Notice that \( a = 2 \) is given for the purpose of simplification.

Item information function (IIF) is considered a very important value in IRT. IFF is used in estimating the value of the ability parameter for an examinee. Moreover, it is related to the standard deviation of the ability estimation. If the amount of information is large, it means that an examinee whose true ability is at that level can be estimated with precision; i.e., all the estimates will be reasonably close to the true value. If the amount of information is small, it means that the ability cannot be estimated with precision and the estimates will be widely scattered about the true ability [Baker, 2001].

In statistical meaning of information, Sir R.A. Fisher defined information as the reciprocal of the precision with which a parameter could be estimated [Baker, 2001]. Statistically, the precision with which a parameter is estimated is measured by the variability of the estimates around the value of the parameter. The amount of information is given by the formula:

\[
I = \frac{1}{\sigma^2}
\]

where \( \sigma^2 \) is a measure of precision of the variance of the estimators.

The estimate \( \theta^1 \) is used based on maximizing the Likelihood function [Daniel, 2000].

The estimate \( \theta^1 \) is normally distributed with mean \( \theta \) and variance (\( \sigma^2 \)). The variance is:

\[
Var \ (\hat{\theta} \mid \theta) = \left[ -E \left( \frac{\partial^2}{\partial \theta \partial \theta} l(\theta) \right) \right]^{-1}
\]

\[
= \frac{1}{\sum_{i=1} I(\theta)}
\]

The estimate \( \theta^1 \) is used based on maximizing the Likelihood function [Daniel, 2000].
\[ I_i(\theta) = \frac{[P_i(Q_i(\theta))]^2}{P_i(\theta)(Q_i(\theta))} \]

4 Result and findings

Before estimating question parameters and starting the adaptive test, students were given conventional tests. Students answered about 100 questions in three stages. In each stage, they solved an exam of 34 questions. 45 students of first intermediate class were examined in these conventional tests. These questions were of first intermediate level. The subject of the questions was in 3-math topics of first intermediate level.

Least square method was used to estimate the difficulty level and discrimination parameters of each question. We used this method because it decreases the load of using computer processor and there is no use of statistical tables such as chi-square.

It is agreed that the difficulty level range from +3, which is very difficult level; to -3 which is very easy level [Baker, 2001].

After the conventional stage, we tested the adaptive stage on 24 students. Two of the students did not complete the exam. According to the hypothesis, we should modify the formula to shorten the number of questions to reach the final level of the students by modifying the equation of estimation of student ability level. We incorporated new factors into IRT, namely initial student ability, subject’s difficulty, number of exercises covered by the teacher and number of lessons covered by the teacher in order to make IRT more realistic and applicable. We had modified the formula of estimating student ability level and had positive results in minimizing the number of questions in adaptive tests.

5 Conclusion

This paper presented a description of adaptive testing based on IRT and experimented with IRT in order to evaluate its applicability, benefits and recommended enhancements for it. Our estimation of IRT questions’ parameters was based on the Least Square method, a well know statistical method. We demonstrated that it is possible to reduce the number of questions in the adaptive test to reach the final level of the students by modifying the equation of estimation of student ability level. We had modified the formula of estimating student ability level and had positive results in minimizing the number of questions in adaptive tests.

References:-


