APPRAISAL OF QUESTION CATEGORIZATION FOR AN ADAPTIVE ASSESSMENT IN E-LEARNING

R. KAVITHA

Assistant Professor, Department of MCA,
Aloysius Institute of Management and Information Technology (AIMIT),
Mangalore, Karnataka, India
kavitha.rajamanii@gmail.com

Abstract

Due to the advancements in technology, e-learning plays a critical role in the field of advanced learning technologies. Each student is having their own way of learning and hence they cannot be assessed in unique way. In Intelligent Tutoring Systems, intellectual questions have to be given to the students. The objective here is to determine the item difficulties of questions to be posed in a test that are going to be used in e-learning. Computer Based Testing is used to collect user responses to sample items. According to these user responses, item difficulties have been found using different approaches. Consequently, best approach to find item difficulty has been determined by a simple classification tool. Since using this classification tool, the best method to find item difficulties is determined, items have been classified using RRT algorithm. This classification ended up with two different results that define the future work of this study. The critical objective is to dispense intellectual questions based on classification in a precise manner to the learners up to their ability without any loss of motivation and hence there is a convinced chance of performing well later there by intent of the Intelligent Tutoring System can be achieved.

Keywords: ***Intellectual question classification, Assessment in E-learning.***

1. Introduction

As this is a technological era, computer assisted assessments are becoming supplementary in on-line learning. The computers have to create a better learning and guiding environment than the traditional educational environment. In e-learning, computers can be used to not only to deliver the course, but also to assess the learner performance in that course. The critical benefit of having computer is automatic assessment and immediate feedback on the performance. Computer-based testing has been developing rapidly as substitute models of measurement, improvements in test organization, instant feedback to learners [1][7]. Thus, as an alternative of convince educators to use conventional techniques, e-learning software that assist them in teaching and measuring the success level of a course.

Computer based tests are viewed as not well enough in terms of effectiveness. The basis is that the questions posed during a session are not adapted for the specific skill of an individual learner. The same predefined set of questions is presented to all students participating in the assessment session, regardless of their ability [6]. The questions within this preset are classically chosen in such a way from low to advance. In this state of affairs, it is accepted that high-performance learners are posed with some questions that are below their level of ability. Likewise, low-performance learners are posed with questions that are above their level of ability. Normally, a CAT commences with a random question with an average difficulty. If the student answers the question correctly, a more difficult question follows. Conversely, if the response is incorrect, an easier question that is with lower estimate is presented next [4][5]. So, the questions have to be classified based on the item difficulties.

Section 2 deals with computer based testing techniques and their comparison. Section 3 concentrates on classification of questions based on Norm-referenced item analysis, Item Response Theory on item difficulty followed by the study that discloses the best algorithm that determines item difficulties. The last section deals

with the further suggestions that can be made.

2. Evolution and Approaches for Adaptive Testing

There are numerous works done on computer based testing applications. Traditionally, the testing focused mainly on paper and pen. In late 1980s, it extended to computer based tests which is based on demand. Instead of giving each examinee the same fixed test, CAT item selection adapts to the ability level of individual examinees. After each response the learner's ability estimate is updated and the successive item is selected to have best property at the new estimation. In fact, the idea of adaptive testing is as old as the practice of oral examinations. Good oral examiners have always known to tailor their questions to their impression of the learner's knowledge level.

The item response theory (IRT) has provided a concrete traction for CAT. The key in attribute of IRT is its modeling of response behavior with distinct parameters for the examinee's ability and the characteristics of the items. This growth led to extensive research by psychometricians to ensure that learners received scores that were fair and equitable. Computerization enables a much advances than just adaptive administration of traditional multiple-choice items.

Items are drawn from the CAT item pool by the item selection algorithm so that items are of appropriate difficulty for each individual examinee. Item pool contains a large number of calibrated items. The items have varying difficulties. Then a statistical analysis is performed to determine characteristics of the item such as its difficulty. Then items are placed in the CAT item pool; this includes the item's text, the classification of the item and the estimated item parameters which includes item difficulty.

The item selection algorithm incorporates sophisticated methods derived from item response theory (IRT) to choose items to administer to a particular learner so that (1) the test is of suitable difficulty (the item selection algorithm chooses more difficult items if an examinee answers items correctly and easier items if an examinee answers incorrectly); (2) the test provides precise measurement; and (3) all learners are administered items that span the same test content. Because the test for each individual is created during the process of test administration, there is not a test form. As a effect, the necessitate arose for effective methods to control for item-exposure as well as to detect items that have been compromised.

3. Question Classification for Adaptive Testing In E-Learning

3.1. Data gathering and preprocessing

The dataset introduced here consists of a test taken by 103 students with 10 multiple choice questions with 4 options.
In CAT, items are drawn from the item pool by a simple item selection algorithm so for each individual examinee, appropriate items are delivered. All items in the CAT item pool range in difficulty values. If a learner gets an item right, an item having a greater difficulty is selected from the item pool and delivered to the examinee as the next question of the test. If he/she gets it wrong, then an item having a smaller difficulty is selected from the item pool and delivered to the learner as the next question of the test.

Table .1 Nominal question levels

Classification of Question	Numerical Depiction
Very Easy	-1
Easy	-0.5
Middle	0
Hard	0.5
Very Hard	1

Primarily, the initial item difficulties of the questions using the appropriate algorithm to be found, so that a CAT can be developed using these initial item difficulties. In a classification problem, in addition to calculating the item difficulties, output classes have to be determined to create a model. Questions are classified in 5 different categories as very easy, easy, middle, hard and very hard as listed in above Table1. The problem here is to determine the item difficulties and place them into right nominal question levels.

3.2. Approach to find item difficulty

3.2.1 Norm-Referenced Item Analysis

A norm-referenced test (NRT) is a type of assessment, or evaluation in which the tested individual is compared to a sample of his or her peers. The term "normative assessment" refers to the process of comparing one test-taker to his or her peers. The goal is to rank the entire set of individuals in order to make comparisons of their performances relative to one another. The strength of multiple-choice tests depends upon an organized selection of items with regard to both content and level of learning. Although most teachers try to select items that sample the range of content covered in class, they often fail to consider the level of unfairness and level of complexity of the items they use.

Item discrimination and item difficulty can be calculated by evaluating the test takers as in norm-referenced item analysis supposed by [2]. Item difficulty is a measure of overall difficulty (p) of the test item. The lower the p, the more difficult a particular item is. Whereas, item discrimination tells how good a question is for separating high and low performers. It is more important for an item to be discriminable than it is to be difficult. For norm-referenced item analysis, test takers should be sorted in descending order first. Then specify, number of people in high and low groups and number of people in high and low groups who get a particular answer right. Using these two groups, item discrimination index and item difficulty index can be calculated using the below formulas:

$$\text{Item Discrimination Index: } a = (Up / U) - (Lp / L) \quad (1)$$

$$\text{Item Difficulty Index: } p = (Up + Lp) / (U + L) \quad (2)$$

Where,

Up = Number of high performers with question right
Lp = Number of low performers with question right
U = Number of high performers
L = Number of Low performers

Table 2. Item Discrimination and Difficulty using NRT

Question Item ID	Up	Lp	Item Discrimination Index	Item Difficulty Index	Classification Level
1	26	11	0.4240	0.5441	0
2	25	10	0.3805	0.5147	0
3	30	13	0.5087	0.6324	-1
4	23	10	0.3894	0.4853	0.5
5	20	15	0.1903	0.5147	0
6	28	14	0.4104	0.6176	-0.5
7	20	8	0.3398	0.4118	0.5
8	27	12	0.4612	0.5735	-0.5
9	28	8	0.5583	0.5294	0.5
10	18	8	0.2995	0.3824	1

When the value of a is high, then better the item is capable of separating high and low performance. If $a = 1$, this means the entire high performance group and none in the lower performance group get a particular question right. But, this is not often seen. It is rare to have a=1. An item has an acceptable level of discrimination if $a >= 0.30$, p and a are not independent probabilities. Discrimination indexes less than 0.30 are sometimes acceptable if there is a very high p value.

3.2.2 Item Response Theory

Item difficulty can be determined by using another IRT approach which uses the formula below:

$$ID = MSCA/SCAE \quad (3)$$

Where,

ID = item difficulty
MSCA = Minimum Sum of Correct Answers
SCAE = Sum of Correct Answers of Each Question

Among all questions, the least answered one is the 10th question. So it has the greatest ID. By having these Item difficulties, the question levels can be classified.

Table 3. Item Difficulty using IRT

Question Item ID	MSCA	SCAE	Item Difficulty	Classification Level
1	38	54	0.7037	0
2	38	53	0.7170	0
3	38	65	0.5846	-0.5
4	38	51	0.7450	0
5	38	56	0.6786	-0.5
6	38	60	0.6333	-0.5
7	38	42	0.9048	0.5
8	38	57	0.6667	-0.5
9	38	54	0.7037	0
10	38	38	1	1

According to this algorithm, there is no item tagged as very easy. Forwhy an item can be very easy if and only if converges to zero.

3.2.3 Choosing the Best Algorithm for Determining Item Difficulties

Since item difficulties of both tests are calculated in two different ways, now a classification algorithm called RandomTree is used to determine which of the methods above the best for determining the item difficulty is.

A Rapidly-exploring Random Tree (RRT) is a data structure and algorithm designed for efficiently searching high-dimensional search spaces. Simply put, the tree is constructed in such a way that any sample in the space is added by connecting it to the closest sample already in the tree. According to RRT, classification results for norm-referenced item analysis and IRT are shown in Table 4.

Table 4. Classification Results for NRT, IRT according to RRT

Approach	Correctly Classified Cases	Incorrectly Classified Cases	Total Number of Cases	Percentage of Correctly Classified Cases	Percentage of Incorrectly Classified Cases
Norm-Referenced Item Analysis	680	350	1030	66.02%	33.98%
IRT	612	418	1030	59.42%	40.58%

As seen from the results the questions are not classified perfectly, to correct this problem data is made nominal and RRT is applied to the data again. Using the nominal data, the results are fairer. According to RRT, classification results for norm-referenced item analysis and IRT on nominal data are shown below in Table 5.

Table 5. Classification Results for NRT, IRT on Nominal Data according to RRT

Approach	Correctly Classified Cases	Incorrectly Classified Cases	Total Number of Cases	Percentage of Correctly Classified Cases	Percentage of Incorrectly Classified Cases
Norm-Referenced Item Analysis	989	41	1030	96.02%	3.98%
IRT	920	110	1030	89.32%	10.68%

Best method to determine the item difficulties is obtained as Norm-Referenced Item Analysis as a consequence of taking both item discrimination and item difficulty into consideration. Another important thing is to make

data nominal before trying to run any classification method on it.

4. Conclusions and Further Enhancement

In the education area, there is an immense need to have tools to monitor test results as well as more precise tools to identify questions that are most likely to be benefited by learners according to the knowledge level.

The applications of item response theory modeling can help to create these tools. Identification of items that are informative helps educators to understand the domains they are measuring as well as the populations they measure. Besides the complexities of the numerous IRT models themselves as to what circumstances are appropriate to use IRT and which model to use. The numerous available IRT software in the market are not user-friendly and often yield different results (parameter and trait estimates) because of the different estimation processes used by the software. Research applying IRT models are appearing more. Together, a better understanding of the models and applications of IRT will emerge and IRT will be as commonly used. Work is still required in defining constructs and related domains of content, drafting items to measure the constructs, field testing, test norming, and conducting reliability and validity studies. However in the sample of this study, best method to determine the item difficulties is obtained as Norm-Referenced Item Analysis as a consequence of taking both item discrimination and item difficulty into consideration. Another important thing is to make data nominal before trying to run any classification algorithm on it.

The further work to done on this starts with questioning the size of item pool. These items will be classified and the importance of the size of item pool will be tried to be determined by comparing the classification of items in different pools. Another work to be done on this research is to determine whether other classification algorithms yield better results or not.

References

[1] Akdemir, O., Oguz, A. (2008). Computer – based testing: An alternative for the assessment of Turkish undergraduate students . Computers and Education, 51, 1198 – 1204.
[2] Brown, J. D. (1995). Developing norm-reference d language tests for program-level decision making . In J. D. Brown & S.O. Yamashita (Eds.). Language Testing in Japan (pp. 40-47). Tokyo: Japan Association for Language Teaching.
[3] Chen, C., Lee, H., Chen, Y. (2005). Personalized e-learning system using Item Response Theory . Computers & Education. 44 – 3, 237 – 255.
[4] Lilley, M., Barker, T. (2002). The development and e valuation of a computer-adaptive testing application for English language . In Proceedings of the 6th computer-assisted assessment conference. Loughborough University, United Kingdom.
[5] Lilley, M., Barker, T. (2003). An evaluation of a computer-adaptive test in a UK University context . In Proceedings of the 7th computer-assisted assessment conference. Loughborough University, United Kingdom.
[6] Lilley, M., Barker, T., Britton, C., (2004). The development and evaluation of a software prototype for computer-adaptive testing . Computers & Education, 43, 109 – 123.
[7] Mills, C. N. (Ed.). (2002). Computer-based testing: Building the foundation for future assessment . NJ: Lawrence Erlbaum.

Author Biography

R. Kavitha is from Tamilnadu, India. She did MCA (Bharathiar University, Tamilnadu), MPhil., in Computer Science (Bharathidasan University, Tamilnadu). Also, she got PGDHET, a degree in Higher Education Technology from Bharathidasan University which is meant exclusively for academicians working in colleges. She is a rank holder in under graduate.

She is having 9 years of teaching experience in colleges. She is currently working as Assistant Professor in Department of MCA, Aloysius Institute of Management and Information Technology (AIMIT), Mangalore, Karnataka, India. Her research area of interest is educational data mining. She presented 4 papers in National Conferences and 2 papers in International Conferences. Now, she is involved in preparing an e-learning content for the subject of Data Structures.