Predicting Performance on Dichotomous Questions: Comparing Models for Large-Scale Adaptive Testing

Jill-Jênn Vie, Fabrice Popineau, Yolaine Bourda
LRI – Bât. 650 Ada Lovelace
Université Paris-Sud
91405 Orsay, France
{jjv,popineau,bourda}@lri.fr

Jean-Bastien Grill
Inria Lille - Nord Europe
40 avenue Halley
59650 Villeneuve-d'Ascq, France
grill@clipper.ens.fr

Éric Bruillard
ENS Cachan – Bât. Cournot
61 av. du Président Wilson
94235 Cachan, France
eric.bruillard@ens-cachan.fr

ABSTRACT

Computerized adaptive testing (CAT) is a mode of testing which has gained increasing popularity over the past years. It selects the next question to ask to the examinee in order to evaluate her level efficiently, by using her answers to the previous questions. Traditionally, CAT systems have been relying on item response theory (IRT) in order to provide an effective measure of latent abilities in possibly large-scale assessments. More recently, from the perspective of providing useful feedback to examinees, other models have been studied for cognitive diagnosis. One of them is the q-matrix model, which draws a link between questions and examinee knowledge components. In this paper, we define a protocol based on performance prediction to evaluate adaptive testing algorithms. We use it to evaluate q-matrices in the context of assessments and compare their behavior to item response theory. Results computed on three real datasets of growing size and of various nature suggest that tests of different type need different models.

Keywords

Adaptive assessment, computerized adaptive testing, cognitive diagnosis, item response theory, q-matrices

1. INTRODUCTION

Automated assessment of student answers has lately gained popularity in the context of online initiatives such as massive online open courses (MOOCs). Such systems must be able to rank thousands of students for evaluation or recruiting purposes and to provide personal feedback automatically for formative purposes.

For computerized adaptive tests (CAT), item response theory (IRT) provides the most common models [3]. IRT provides a framework to evaluate the performance of individual questions, called *items*, on assessments [6]. When the intention is more formative, examinees can receive a detailed feedback, specifying which knowledge components (KCs) are mastered and which ones are not [1]. Most of these models rely on a q-matrix specifying for each question the different KCs required to solve it.

We propose a protocol to evaluate adaptive testing algorithms and use it to compare the performances of the simplest IRT model, the 1-parameter logistic one, commonly known as Rasch model, with the simplest Q-matrix model. We expect to answer the following question: given a budget of questions of a certain dataset asked according to a certain adaptive selection rule, which model performs the best at predicting the answers of the examinee over the remaining questions? We managed to get satisfactory results, enabling us to state that no model dominates in all cases: according to the type of test, either the Rasch model or the q-matrix performs the best.

2. BACKGROUND AND RELATED WORK

2.1 Item Response Theory: Rasch Model

The Rasch model estimates the latent ability of a student by a unique real number θ modeled by a random variable and characterizes each question by one real number: its difficulty d, corresponding to the ability needed to answer the question correctly. Knowing those parameters, the probability of the event "the student of ability θ answers the question of difficulty d correctly", denoted by *success*, is modeled by:

$$\Pr\{success|\theta\} = \frac{1}{1 + e^{-(\theta - d)}}.$$

The aim is first to optimize the parameters d_j for each question j and θ_i for each student i in order to fit a given train dataset. Then, throughout the test, a probability distribution over θ_i is updated after each question answered, using the Bayes' rule.

2.2 Cognitive Diagnosis Model: Q-matrix

We now present a model that tries to be more informative about the student's knowledge components. Every student is modeled by a vector of binary values $(a_1, \ldots, a_K)$, called *knowledge vector*, representing her mastery of K distinct KCs. A q-matrix Q [7] represents the different KCs involved in answering every question. In the NIDA model considered here [3], Q_{ij} is equal to 1 if the KC j is required to succeed at question i, 0 otherwise. More precisely, we denote by s_i (g_i) the *slip* (*guess*) parameter of item i. The probability of a correct response at item i is $1 - s_i$ if all KCs involved are mastered, g_i if any required KC is not mastered.

The KCs are considered independent, thus the student's knowledge vector is implemented as a vector of size K indicating for each KC the probability of the student to master it. Throughout the test, this vector is updated using Bayes' rule. From this probability distribution and with the help of our q-matrix, we can derive the probability for a given student to answer correctly any question of the test.

3. ADAPTIVE TESTING FRAMEWORK

Our student data is a dichotomous matrix of size $N_S \times N_Q$ where N_S and N_Q denote respectively the number of students and the number of questions, and c_{ij} equals 1 if student i answered the question j correctly, 0 otherwise.

We detail our random subsampling validation method. Once the model has been trained, for each student of the *test* dataset, a CAT session is simulated. In order to reduce uncertainty at most, at each step we pick the question that maximizes the Fisher information and ask it to the student. The student parameters are updated according to her answer and a performance indicator at the current step is computed. To compare it to the ground truth, we choose the negative log-likelihood [5], that we will denote by "mean error".

4. EVALUATION

We compared an R implementation of the Rasch model (IRT) and our implementation of the NIDA q-matrix model (Q) for different values of the parameter K, the number of columns of the q-matrix. Our algorithms were tested over three real datasets:

SAT dataset [4]. Results from 296 students on 40 questions from the 4 following topics of a SAT test: Mathematics, Biology, World History and French.

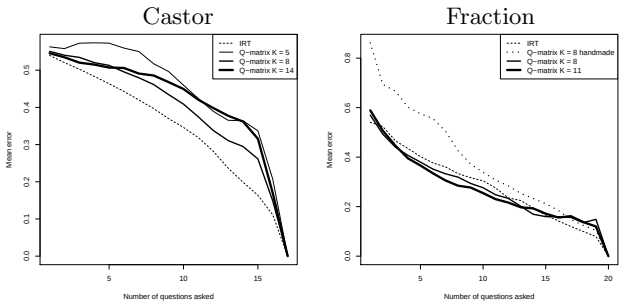
Fraction dataset [2]. Responses of 536 students to 20 questions about fraction subtraction.

Castor dataset. Answers of 6^{th} and 7^{th} graders competing in a K-12 Computer Science contest which was composed of 17 tasks. It is a 58939×17 matrix, where the (i, j) entry is 1 if contestant i got full score on task j, 0 otherwise.

Results are presented in Table 1 where the best performances are shown in bold. As a reference, 1.0 is the error obtained by the trivial algorithm affecting 1/2 to every probability. On the Castor dataset, IRT performs better than Q for any value of K throughout the whole test. On the Fraction dataset, the handmade q-matrix achieves the highest error. In the early questions of the test, Q algorithms for $K = 8$ and 11 perform slightly better than IRT. The Fraction dataset is a calculus test: it requires tangible, easy-to-define knowledge components. Therefore, after a few carefully chosen questions Q can estimate reasonably the performance of an examinee over the remaining ones. On the SAT dataset, IRT achieves the lowest error among all tested algorithms. We also observe that the variance increases throughout the test, probably because the behavior of the algorithm may vary substantially if the remaining questions are from a different topic than the beginning of the test.

5. DISCUSSION AND FUTURE WORK

Our comparison of the cognitive diagnosis model with IRT seems to indicate that q-matrices perform better on a certain type of tests; in the Fraction test, there are redundancies from one question to another in order to check that a notion is known and mastered. Conversely, IRT performs better on both the SAT test and Castor contest, which is remarkable given its simplicity. The fact that the SAT test is multidisciplinary explains the difficulty of all considered algorithms in predicting the answers, and the nature of Castor as a contest may require a notion of level instead of knowledge mastery. Therefore, in those cases, we will prefer to use the Rasch model. In order to confirm this behavior, we plan to test our implementation on many other datasets.



Castor	After 4 q.	After 10 q.	After 16 q.
Q $K = 2$	0.555 ± 0.004	0.456 ± 0.005	0.167 ± 0.012
Q $K = 5$	0.574 ± 0.004	0.460 ± 0.006	0.206 ± 0.016
Q $K = 8$	0.520 ± 0.004	0.409 ± 0.006	0.148 ± 0.013
Q $K = 11$	0.519 ± 0.004	0.462 ± 0.007	0.218 ± 0.014
Q $K = 14$	0.515 ± 0.003	0.449 ± 0.006	0.169 ± 0.014
IRT	**0.484 ± 0.003**	**0.346 ± 0.005**	**0.111 ± 0.010**
Fraction			
Q $K = 2$	0.464 ± 0.012	0.326 ± 0.013	0.196 ± 0.017
Q $K = 5$	0.440 ± 0.011	0.289 ± 0.014	**0.146 ± 0.013**
Q $K = 8$	0.407 ± 0.011	0.276 ± 0.015	0.159 ± 0.015
Q $K = 11$	**0.395 ± 0.009**	**0.255 ± 0.013**	0.156 ± 0.015
Q $K = 14$	0.422 ± 0.009	0.274 ± 0.014	0.180 ± 0.018
IRT	0.435 ± 0.012	0.304 ± 0.013	**0.142 ± 0.012**
Q* $K = 8$	0.596 ± 0.008	0.346 ± 0.007	0.182 ± 0.007
SAT			
Q $K = 2$	0.522 ± 0.007	0.417 ± 0.010	0.315 ± 0.018
Q $K = 5$	0.469 ± 0.007	0.365 ± 0.012	0.306 ± 0.019
Q $K = 8$	0.463 ± 0.007	0.367 ± 0.013	**0.242 ± 0.018**
Q $K = 11$	0.456 ± 0.008	0.364 ± 0.013	0.331 ± 0.023
Q $K = 14$	0.441 ± 0.007	0.350 ± 0.012	0.296 ± 0.021
IRT	**0.409 ± 0.008**	**0.285 ± 0.012**	**0.248 ± 0.022**

Table 1: Mean error of the different algorithms over the remaining questions of the Castor and Fraction datasets, after a certain number of questions have been asked. The dashed curve denotes the Rasch model (IRT), while the curves of growing thickness denote q-matrices (Q) of growing number of columns. The dotted curve in Fraction denotes the handmade q-matrix (Q*) [2].

6. ACKNOWLEDGEMENTS

We thank Chia-Tche Chang, Le Thanh Dung Nguyen and especially Antoine Amarilli for their valuable comments. We also thank Mathias Hiron for providing the Castor dataset. This work is supported by the Paris-Saclay Institut de la Société Numérique funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

7. REFERENCES

[1] Y. Cheng. When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74(4):619–632, 2009.

[2] L. T. DeCarlo. On the analysis of fraction subtraction data: The dina model, classification, latent class sizes, and the q-matrix. *Applied Psychological Measurement*, 2010.

[3] M. C. Desmarais and R. S. Baker. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1-2):9–38, 2012.

[4] M. C. Desmarais et al. Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization. In *4th International Conference on Educational Data Mining, EDM*, pages 41–50, 2011.

[5] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[6] R. K. Hambleton, H. Swaminathan, and H. J. Rogers. *Fundamentals of item response theory*. Sage, 1991.

[7] K. K. Tatsuoka. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4):345–354, 1983.