Towards Efficient Item Calibration in Adaptive Testing

Eduardo Guzmán and Ricardo Conejo

Departamento de Lenguajes y Ciencias de la Computación,
E.T.S.I. Informática, Universidad de Málaga, Apdo. 4114, Málaga 29080, Spain
{guzman, conejo}@lcc.uma.es

Abstract. Reliable student models are vital for the correct functioning of Intelligent Tutoring Systems. This means that diagnosis tools used to update the student models must be also reliable. Through adaptive testing, student knowledge can be inferred. The tests are based on a psychometric theory, the Item Response Theory. In this theory, each question has a function assigned that is essential for determining student knowledge. These functions must be previously inferred by means of calibration techniques that use non-adaptive student test sessions. The problem is that, in general, calibration algorithms require huge sets of sessions. In this paper, we present an efficient calibration technique that just requires a reduced set of prior sessions.

1 Introduction

The construction of Intelligent Tutoring Systems (ITSs) requires the development of reliable mechanisms to supervise interaction with the students. One of the most common solutions to this end is testing. Generally, test-based diagnosis systems use heuristic solutions to infer student knowledge, but these solutions are in conflict with the aim of obtaining a reliable diagnosis. In contrast, adaptive testing theory guarantees this reliability, since it is based on a well-founded theoretical background.

The advantages of adaptive tests are that they require a smaller number of questions (called in this context *items*) than conventional tests. Each student usually takes different sequences of items, or even different items. Factors such as the items that must be posed to the student and when the test must finished are dynamically determined in relation to a previously established estimation of the student's knowledge.

However, one of the most important shortcomings of adaptive testing is that, in order to be used, items included in this type of tests require a preliminary calibration process. Through calibration, item characteristic functions are determined. These functions are vital to the proper functioning of an adaptive test. Thus, this disadvantage can be considered the most important, since it is essential to get valid and reliable adaptive testing based diagnosis. Calibration requires having available huge sets of test sessions previously done by students. These students were administered non-adaptive tests.

L. Ardissono, P. Brna, and A. Mitrovic (Eds.): UM 2005, LNAI 3538, pp. 402 – 406, 2005.

In previous papers [4], we presented an adaptive testing-based cognitive assessment model. This paper introduces the item calibration technique that has been developed. This technique is more efficient than conventional approaches, and the general requirements have been considerably relaxed. In particular, it reduces the number of prior test sessions needed.

This paper is structured as follows: The next section is dedicated to adaptive testing and Item Response Theory. In section 3 a brief description of the cognitive assessment model is outlined. In section 4, the mechanism used for item calibration is studied. Finally, Section 5 discusses the contributions of this paper and future tasks that we plan to accomplish.

2 Theoretical Background

Generally, in adaptive testing (a.k.a. *Computerized Adaptive Testing*) [10], items are posed one at a time. The final goal of an adaptive test is to estimate quantitatively the level of student knowledge as expressed by means of a numerical value (usually in the real number domain). The response model is the central element of the adaptive testing theory. This model supplies the underlying theoretical background. It is usually based on the *Item Response Theory* (IRT) [5]. IRT is a probabilistic theory that determines: how the student knowledge is inferred, how to calculate the most suitable item that must be posed to each student during the test, and when it must finish. It is based on two principles: a) Student performance in a test can be explained by means of his/her knowledge level. b) The performance of a student with a certain knowledge level answering an item can be probabilistically predicted and modeled by means of functions called c*haracteristic curves*.

There are hundreds of IRT-based models and different classification criteria of them. One of these criteria deals with how the models update the estimated student knowledge in terms of his/her response. Thereby, IRT-based models can be: (1) *Dichotomous models*: Only two possible scores are considered: correct or incorrect. A characteristic curve is enough to model each item, the *Item Characteristic Curve* (ICC). It expresses the probability that a student with a certain knowledge level has to answer the item correctly. (2) *Polytomous models*: The former family of models does not make any distinction in terms of the answer selected by the student. No partial credit is given. This means information loss. To overcome this problem, in this family of models each possible answer has a characteristic curve called *Trace Line* (TC). It expresses the probability that a student with a certain knowledge level will more than likely select this answer.

Polytomous models usually require a smaller number of items per test than the dichotomous ones. Nonetheless, dichotomous models are most commonly used in adaptive testing environments. The main reason is that the calibration process is harder in polytomous models. Instead of calibrating one curve per item, a set of TCs must be determined per item. This means that the prior set of non-adaptive test sessions is greater. While a test of dichotomous items requires several hundreds of prior test sessions, a test of polytomous items requires several thousands [4].

3 The Cognitive Assessment Model

This model assumes that the declarative knowledge in a certain subject (or course) can be represented by means of a hierarchy of topics (or concepts), forming the curriculum. All these topics are related by means of aggregation relations. Accordingly, this curriculum can be seen as a granularity hierarchy [6]. These topics symbolize knowledge pieces, where leaf nodes represent a unique concept or a set of concepts inseparable from the assessment point of view.

In order to assess the student knowledge state in part of (or in the whole) curriculum, items must be created and linked to the topics they assess. Thus, items are student knowledge evidence providers. The relationship between an item and a topic expresses that the item is used to assess the topic. Thanks to the aggregation relation between topics, if an item provides evidence about the student knowledge in a topic *T*, it will provide evidence of the knowledge in all preceding topics of *T* in the curriculum hierarchy. This relation is supported by means of characteristic curves as will be explained in a posterior subsection.

For this cognitive model, an IRT-based model has been developed. It uses a discrete scale to measure the knowledge level, where the number of knowledge levels in which the students can be classified is a configurable parameter. Let *K* be the number of knowledge levels, student knowledge can be found between *0* (absence of knowledge) and *K-1* (full knowledge). Accordingly, characteristics curves turn into vectors, i.e. a probability value per knowledge level. This model is also polytomous. Therefore, for each pair item answer-topic assessed, there will be a different TC. Consequently, the number of item TCs is equal to the topics it assesses, multiplied by the number of possible answers. A restriction must be imposed to ensure the maintenance of all probabilistic properties: for each pair item-topic the sum of all the TCs must be equal to one in each knowledge level.

This response model uses a non-parametric approach. This means that, characteristic curves are not constrained by any model. [9] indicates that parametric models are commonly used without checking if they actually are appropriate for calibration input data, and this is unacceptable from a statistical perspective. The goal of calibration is to infer the TCs that represent the real student behavior while taking a test, not to force the TC shape to fit certain model far away from this behavior. In addition, the use of a non-parametric approach facilitates the calibration process, as will be shown in the next section.

4 Item Calibration

Kernel smoothing [7] is a statistical technique very popular thanks to its simplicity. It has been traditionally used to determine non-parametric regression curves. It is based on the principle that given a set of observations *X* and a function *m*, the set of observations next to *x*, should contain information about the value of *m* in *x*. Accordingly, to estimate the value of *m(x)* it is possible to use some kind of local average of the data closest to *x* [8].

Some psychometricians have previously used kernel smoothing in adaptive testing [7]. In our cognitive assessment model, kernel smoothing is used to calibrate

the TCs of our polytomous response model. Accordingly, using kernel smoothing, the TCs will be determined for each pair item-topic. The procedure for calibrating the set of TCs of all items that assess certain topic *C* has the following steps:

1) *Prior student session compilation*: From all test sessions available, all of them that involved the topic *C* are collected. The information of these sessions required for calibration is the answer that each student selected per item. Information on any other item not involving topic *C* is purged.
2) *Score computation*: For each student, his/her score is computed. This is done heuristically, since it is useful just for ordering the students' performance in the test. For instance, one of the ways to do this is by calculating the percentage of items successfully answered.
3) *Score transformation*: The percentage obtained in the former phase is transformed into a temporary knowledge level. It is made by calculating the corresponding quartile in a standard normal distribution. After that, this value is mapped to the discrete scale used to represent the knowledge level.
4) *Session sort*: Student test sessions are ordered in terms of their temporary knowledge level.
5) *Smoothing*: For each item, their TCs are computed using Equation 1. $p(u_i=r_j|\theta_k)$ is the probability value of the TC vector of the answer j of the item i for the knowledge level k.

$$P(\mathrm{u_i} = \mathrm{r_j} \mid \theta_K) = \sum_{s=1}^{N} w_{sjk} u_{sji} \tag{1}$$

where N is the number of the prior student sessions. u_{sji} is equal to 1 if the student s selected the answer j of the item i. Its value is zero otherwise. w_{sjk} is a weight computed as follows:

$$w_{sjk} = \frac{F((\theta_k - \theta_s)/h)}{\sum_{a=1}^{N} F((\theta_k - \theta_a)/h)} \tag{2}$$

where F is the so-called *kernel function*.

6) *Iterative refinement*: This step is optional. Using the calibrated TCs obtained in the previous step, the student real knowledge levels in topic *C* are computed. These new values can be used as a feedback to recalibrate the TCs. This process should continue until the values of the student knowledge levels and the TC values remain unchanged.

This calibration procedure must be repeated for all the topics of the curriculum. Once all the TCs have been calibrated, any time they will be used (now in adaptive tests), they could be updated with these new test session results. Accordingly, this process could be repeated, automatically or on demand, getting more accurate estimations of the characteristic curves.

Conventional calibration techniques are iterative procedures that require too much time [9]. In contrast, through kernel smoothing, calibration is a non-iterative procedure (even when the refinement step is carried out, it just requires a few

iterations). Using this calibration technique, the number of prior student sessions can be reduced, yet reasonable estimations are still obtained[1].

5 Conclusions and Future Work

The main contribution of this paper is a calibration technique that makes feasible the use of adaptive testing with a polytomous response model. This method is based on kernel smoothing. It requires a reduced number of prior student sessions in comparison to the conventional calibration algorithms. This calibration technique has been included in a polytomous response model.

This algorithm just represents the starting point of this research. Exhaustive experiments must be carried out in order to study its behavior and to determine the prerequisites for the minimum requirements of the prior student sessions necessary to obtain reasonable calibration results.

A prototype of the cognitive model and the calibration technique is currently implemented in the SIETTE system (http://www.lcc.uma.es/siette) [1]. It is a web-based system that can be used as a diagnosis tool inside web-based ITSs, or as an independent testing application. It allows teachers to include new items and tests through an elicitation tool.

References

1. Conejo, R.; Guzmán, E.; Millán, E.; Pérez-de-la-Cruz, J. L., Trella, M. and Ríos, A. SIETTE: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, 14 (2004). 29-61.
2. Eubank, R. Spline smoothing and nonparametric regression. Deker, New York (1988).
3. Guzmán, E. and Conejo, R. A library of templates for exercise construction in an adaptive assessment system. *Technology, Instruction, Cognition and Learning (TICL)*, 2(1-2). (2004). 21-43.
4. Guzmán, E. and Conejo, R. A Model for Student Knowledge Diagnosis Through Adaptive Testing. *LNCS, 2363. ITS 2004*. Springer Verlag; 2002: 12-21.
5. Lord, F. M. *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1980.
6. McCalla, G. I. and Greer, J. E. Granularity-Based Reasoning and Belief Revision in Student Models. In: Greer, J. E. and McCalla, G., eds. *Student Modeling: The Key to Individualized Knowledge-Based Instruction*. Springer Verlag; 1994; 125 39-62.
7. Ramsay, J.O. Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika* **56**, 611-630 (1991).
8. Simonoff, J.S. Smoothing Methods in Statistics. Springer-Verlag, New York (1996).
9. Stout, W. Nonparametric Item Response Theory: A Maturity and Applicable Measurement Modeling Approach. *Applied Psychological Measurement* **25**, 300-306 (2001).
10. van der Linden, W. J. and Glas, C. A. W. *Computerized Adaptive Testing: Theory and Practice*. Netherlands: Kluwer Academic Publishers; 2000.

[1] Experimental results have not been included due to lack of space.