Student Assessment by Optimal Questionnaire Design

Melisa Aruci
Departmenti i Informatikës
Fakulteti i Shkencave të Natyrës, Tiranë, Albania
{melisa.aruci@gmail.com}

Giuseppina Lotito
"Gorjux-Tridente-Vivante"
70125 Bari, Italy
{giuseppina.lotito@istruzione.it}

Giuseppe Pirlo
Dipartimento di Informatica
Univ. di Bari, 70125 Bari, Italy
{giuseppe.pirlo@uniba.it}

Abstract

In this paper a new technique is presented for automatic design of optimal questionnaires. The technique, that is based on the Item Response Theory, performs multiple-choice item selection by a Genetic Algorithm. The experimental results demonstrate the validity of the proposed approach to adjust the characteristics of the questionnaire to the abilities of the student class.

1. Introduction

Computer-based student assessment is now considered a fundamental service of Learning Management Systems [Amelung2011; Dimauro2003; Romero2008; Greco2006b]. Although several types of computer-based systems for student's assessment have been proposed so far Multiple-choice Item on-line Questionnaires (MIQs) is the most diffuse approach [Lan2011; Romero2010] since they can be easily integrated into computer-based assessment systems [Kuechler2003; Romero2009]. When a MIQ is considered, students are asked to select the best possible answer from the choices provided on a list [Kuechler2003]. Data from MIQs can be used for providing personalized learning suggestions [Chu2006], for the analysis of individual targets [Yamanishi2001], for discovering the individual needs of the students [Pechenizkiy2008], for discovering rule patterns [Chen2009]. Unfortunately, the design of a questionnaire is a complex task that requires the selection of the set of items most advantageous for assessing the skill level of a student [Lan2011].

In this paper a new approach for optimal questionnaire design is proposed, based on the Item Response Theory (IRT). A questionnaire is considered as an entity that must be tailored according to the specific characteristics of the group of students to be assessed. The proposed approach uses a two-steps strategy. In the first step the system estimates item difficulty for a given student class with specific abilities. In the second step a Genetic Algorithm (GA) is used to determine the best set of items to be included in the questionnaire.

The organization of the paper is the following. Section 2 presents the problem of item evaluation by IRT. The problem of optimal questionnaire design is formally described in Section 3. Section 4 presents the genetic algorithm used for automatic questionnaire design. Section 5 presents the experimental results. Section 6 reports the conclusion.

2. Item Evaluation by IRT

IRT states that responses to a set of items can be explained by the existence of one or more latent traits, named abilities [Van der Linen1997; Fraley2000]. A main objective in item response modelling is to characterize the relation between a latent trait, θ, and the probability of item endorsement. This relation is typically referred to as the Item Characteristic Curve (ICC) and can be defined as the (nonlinear) regression line that represents the probability of endorsing an item (or an item response category) as a function of the underlying trait [Fraley2000]. For the purpose of this work, the Two-Parameter Logistic Model (2PLM) [Birnbaum1968] is considered. In this case, given the set of items $T=\{t_1, t_2,\ldots, t_j\ldots, t_M\}$, the probability that an individual with trait level θ_i will endorse item t_j is defined as a function [Birnbaum1968]:

$$P_j(\theta_i) = \frac{1}{1+e^{-\alpha_j(\theta_i-\beta_j)}} \quad (1)$$

where α_j and β_j are the item discrimination parameter and the item difficulty parameter, respectively. The difficulty parameter β_j represents the level of the latent trait necessary for an individual to have a 50% probability of endorsing the item; the item discrimination parameter α_j represents an item's ability to differentiate between people with contiguous trait levels. Of course, items are not equally informative across the entire range of the trait θ. In fact, an item yields the most information when θ_i equals β_j.In the IRT, an item is considered difficult if a high level of ability or knowledge is required to answer it correctly. Therefore, the difference $P_j(\theta_{max})$-$P_j(\theta_{min})$ can be used

to estimate the extent to which item t_j is effective to assess students in the range $[\theta_{min}, \theta_{max}]$: the greater the difference $P_j(\theta_{max})$-$P_j(\theta_{min})$ the better the item t_j.

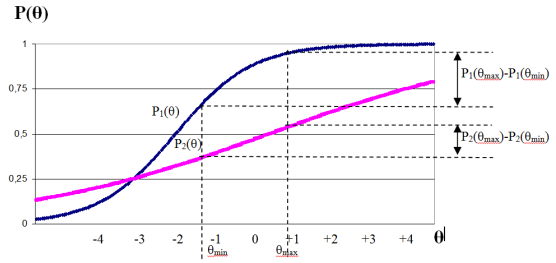

Figure 1. Item Effectiveness Estimation by ICCs

Figure 1 shows the ICCs of two items t_1 and t_2. In this case, the results indicated that t_1 is better than t_2 for assessing the students in the range $[\theta_{min}, \theta_{max}]$, since $P_1(\theta_{max})$-$P_1(\theta_{min}) > P_2(\theta_{max})$-$P_2(\theta_{min})$.

3. A Theoretical Approach to Optimal MIQ Design by GA

In this paper the problem of optimal MIQ design is considered as an optimization process in which - from the set of M items T - the subset of N items (N<M) more suitable for investigating the latent abilities of the set of students belonging to the skill range $[\theta_{min}, \theta_{max}]$ is selected.

Formally, let $T=\{t_1, t_2,\ldots, t_j\ldots, t_M\}$ be the set of M items available, and $S=\{s_1, s_2,\ldots, s_i\ldots, s_N\}$ the set of N students under consideration, being θ_i the trait ability level of the i-th student, i=1,2,…,N. The problem of optimal MIQ design concerns the selection from T of the subset $Q=\{t_{i_p} \mid p=1,2,\ldots,P$ with $(1\le i_p\le M$ and $i_p\ne i_q$ for $p\ne q)\}$, which maximize the fitness function:

$$F(Q) = P^Q(\theta_{\max}) - P^Q(\theta_{\min}) \qquad (2)$$

where:

$\theta_{max}=\max\{\theta_i \mid i=1,..,N\}$ and $\theta_{min}=\min\{\theta_i \mid i=1,..,N\}$.

4. Optimal MIQ Design by GA

A binary-coded genetic algorithm was considered is used to solve the optimization problem in eq. (2), since genetic algorithms have potential for solving non-linear optimization problems, in which the analytical expression of the object function is not known [Michalewicz1996; Goldberg1989]. The genetic approach is based on the following phases [Baeck1996]. The initial – population $Pop=\{\Phi_1, \Phi_2, \ldots,\Phi_\kappa,\ldots,\Phi_{Npop}\}$ of random individuals was created. In our tests N_{pop} has been set to 20. since some preliminary experiments have shown N_{pop}=20 is a good trade-off between convergence speed of the genetic algorithm and its capability to escape from local extrema. In our approach, each individual (that is a MIQ) is represented by a vector $\Phi_k=\langle h_1, h_2,\ldots, h_j,\ldots, h_M\rangle$, where each gene h_j was a Boolean value: h_j=0 means that j-th item of T (i.e. the item t_j) was not included in MIQ; h_j=1 means that j-th item of T (i.e. the item t_j) was included in Q. Of course, since P items must be included into the questionnaire Q, the following normalization procedure was performed for each individual Φ_k. In particular, let be $P'=h_1+h_2+\ldots+h_M$, if P'>P then select randomly (P'-P) genes equal to 1 and set them to 0; if P' < P then select randomly (P-P') genes equal to 0 and set them to 1. Successively, the fitness function was computed for each individual Φ_k of the population, according to eq. (2).

From the initial - population, the following four genetic operations were used to generate the new populations of individuals:

i) Individual Selection. In the selection procedure $N_{pop}/2$ random pairs of individuals were selected for crossover, according to a roulette-wheel strategy. This associates a selection probability to each individual. The higher the fitness function of the individual, the higher the selection probability [Baeck1996].

ii) Crossover. In our approach, a one-point crossover was used [Baeck1996]. In this case, for each pair of individuals selected for crossover, a random integer ν ($1< \nu \le M$) was chosen and the child individuals are defined according to the following rule:

- $h^a_s=h^a_s$ and $h^b_s=h^b_s$, if $s < \nu$;
- $h^a_s=h^b_s$ and $h^b_s=h^a_s$, if $s \ge \nu$.

iii) Mutation. In this approach a uniform mutation operator is considered. Let $\Phi_\kappa=\langle h_1, h_2,\ldots, h_M\rangle$ be an individual, the uniform mutation operator changed (inverted) each gene of the individual according to a mutation probability, Mut_prob (Mut_prob=0.02 in our tests). After mutation, the normalization procedure was also applied to all individuals Φ_k, k=1,2,…,Pop, in order

to ensure that each questionnaire has a number of items equal to P,

iv) Elitist Strategy. From the N_{pop} individuals generated by the above-described operations, one individual was randomly removed and the individual with the maximum fitness in the previous population was added to the current population [Baeck1996].

Operations (i),(ii),(iii),(iv) were then repeated until N_{iter} successive populations of individuals were generated (N_{iter}=50 in our tests). When the process stopped, the optimal questionnaire was obtained by the best individual of the last-generated population.

5. Experimental Results

In order to evaluate the new technique for optimal questionnaire design, a well-defined simulated dataset was considered. First a set of MT*N random responses simulating the answers of N of students to a set of MT items was generated automatically.

The experiment included two steps: (1) the ability estimation step; (2) the optimal test design step.

1) In the ability estimation step the student models (i.e. the trait ability level of each student) were estimated. After data simulation, the ICC of each item was evaluated using the 2PLM model and the trait ability level of each student was computed. For the purpose, the Marginal maximum likelihood estimation was considered, where the hidden student variables are chosen to maximize the likelihood of the data, according to the approach proposed in the literature [Bock and Aitkin 1981]. Finally, the skill range of the set of students [θ_{min}, θ_{max}] was determined.
2) In the optimal test design step the optimal MIQ was designed for the specific set of students under consideration. In the test step, a new set of M items named Full Set (FSM) was generated and the optimal questionnaire T^*_P was defined by automatically picking out the optimal subset of P items from FSM, for the given set of simulated students with a range equal to [θ_{min}, θ_{max}].

Table I shows the experimental results obtained with the simulation procedure. In this case, we considered N=20 students and MT=100 items. Successively, the ability of each student was estimated according to the approach of Bock and Aitkin [Bock and Aitkin 1981] and the skill range [θ_{min}, θ_{max}]=[2.20, 3.31] of the student set was determined. The test step was carried out using the questionnaire FSM of M items (M=50 in our test) and other MIQ obtained by selecting the optimal subset T*P of P items out of M (P=10,15,20 in our test). In order to evaluate the effectiveness of the proposed approach, the ability estimated when using the optimal questionnaire T*P was compared with the average ability determined when using the random-generated MIQs of P items, where item selection was performed randomly. In particular, each value T^{rnd}_P is the average ability calculated when taking into account 10 MIQs, each one realized by selecting P random items from FSM. In order to estimate the effectiveness of the MIQs for student assessment we considered the following measures:

- A_FSQ(i) the ability of the i-th student estimated through the Full Set questionnaire FSM of M items;
- $A\_T^*_{P(i)}$ the ability of the i-th student estimated through the optimal questionnaire T*P of P items;
- $A\_T^{rnd}_{P(i)}$ the ability of the i-th student estimated by averaging the abilities determined through 10 random-generated P items questionnaires.

Hence the accuracy of $T^*_{P(i)}$ and $T^{rnd}_{P(i)}$ to assess student ability was estimated, respectively, by the standard deviations:

$$SD(FSQ\_T^*{}_p) = \sqrt{\sum_{i=1}^{N}\left[A\_FSQ(i) - A\_T^*{}_p(i)\right]^2}$$

and

$$SD(FSQ\_T^{rnd}{}_p) = \sqrt{\sum_{i=1}^{N}\left[A\_FSQ(i) - A\_T^{rnd}{}_p(i)\right]^2}$$

Of course, the comparison between SD(FSQ_T*$_P$) and SD(FSQ_T$^{rnd}_P$) reported in Table I provides a useful information about the capability of the proposed approach in selecting optimal subsets of items for questionnaire design, able to assess students more precisely than using randomly selected items.

Table I. Experimental Results

	P	SD(FSQ_T*$_P$)	SD(FSQ_T$^{rnd}_P$)
Simulated Data	*10*	*0.27*	*0.62*
	15	*0.21*	*0.47*
	20	*0.13*	*0.25*

6. Conclusion

This paper presents a new technique for optimal questionnaire design based on the IRT. The aim of this work is twofold. First, the problem of optimal questionnaire design is considered as an optimization problem. Second, a genetic algorithm is proposed for optimal questionnaire design and its effectiveness is

demonstrated. The algorithm automatically selected the best set of items for the specific range of ability of the students under consideration.

The experimental results confirm the effectiveness of the new approach in adapting questionnaire to the abilities of a given set of students.

References

[Amelung2011] Amelung M., Krieger K., Rosner D. (2011) "E-Assessment as a Service", IEEE TLT, Vol. 4, No. 2, pp. 35-46.

[Baeck1996] Baeck T. (1996) Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolution Programming, Genetic Algorithms. New York: Oxford Univ. Press.

[Birnbaum1968] Birnbaum A. (1968) "Some latent trait models and their use in inferring an examinee's ability", Statistical theories of mental test scores, F. M. Lord and M. R. Novick (eds.), Addison-Wesley, pp. 397-472.

[Bock and Aitkin 1981] Bock,R. and Aitkin,M. (1981) "Marginal Maximum Likelihood Estimation of Item Parameters: Applications of an EM Algorithm", Psychometrika, Vol. 46, pp. 443-459.

[Chen2009] Chen Y., Wenig C. (2009)"Mining fuzzy association rules from questionnaire data", Knowledge-Based Systems Journal, Vol. 22, No. 1, pp. 46-56.

[Chu2006] Chu H.C., Hwang G. J., Tseng J.C.R., Hwang G. H. (2006) "A computerized approach to diagnosing student learning problems in health education", Asian Journal of Health and Information Sciences, Vol. No. 1, pp. 43-60.

[Dimauro2003] Dimauro G., Impedovo S., Pirlo G. (2003) "Traditional learning toward on-line learning", Proc.TEL'03, Italy, pp. 355-360.

[Fraley2000] Fraley R.C., Waller N. G., Brennan K. A. (2000) "An Item Response Theory Analysis of Self-Report Measures of Adult Attachment", Journal of Personality and Social Psychology, Vol. 78, No. 2, pp. 350-365.

[Goldberg1989] Goldberg D.E. (1989) Genetic Algorithms in Search, Optimization and Machine Learning, Addison–Wesley, New York.

[Greco2006b] Greco N., Impedovo D., Pirlo G. (2006b) "New Steps toward the Effective Evaluation of E-learning Activities: A Participant-Based Approach", WSEAS Trans. Adv. Eng. Educ, Issue 7, Vol. 3, pp. 662-666.

[Impedovo2006] Impedovo S., Lucchese M.G., Pirlo G. (2006) "e-Examinations: an Advanced Methodology for Student's tests on e-Learning University Courses" , WSEAS Trans. Adv. Eng. Educ., Issue 5, Vol. 3, pp. 361- 366.

[Kuechler2003] Kuechler W. L., Simkin M. G. (2003) "How well do multiple choice tests evaluate student understanding in computer programming classes?",J.Inf.Syst.Educ.,Vol.14,No.4,pp.389–399.

[Lan2011] Lan C. H., Graf S., Lai K. R., Kinshuk (2011) "Enrichment of Peer Assessment with Agent Negotiation", IEEE TLT, Vol. 4, No. 1, pp. 35-46.

[Michalewicz1996] Michalewicz Z. (1996) Genetic Algorithms + Data Structure=Evolution Programs, Springer Verlag, Berlin, Germany.

[Pechenizkiy2008] Pechenizkiy M., Calders T., Vasilyeva E., De Bra P. (2008) "Mining the student assessment data: lessons drawn from a small scale case study", International Conference on Educational Data Mining, Spain, pp. 187-191.

[Romero2010] Romero C., Ventura S. (2010) "Educational Data Mining: A Review of the State-of-the-Art", IEEE TSMC -Part C: Applications and Reviews, Vol. 40, No. 6, pp. 601–618.

[Romero2009] Romero C., Ventura S., De Bra P. (2009) "Using Mobile and Web-Based Computerized Tests to Evaluate University Students", Computer Applications in Engineering Education, Vol. 17, No. 4, 435-447.

[Romero2008] Romero C., Ventura S., Salcines P. E. (2008) "Data mining in course management systems: Moodle case study and tutorial", Computers & Education, Vol.51, No.1,pp.368-384.

[Van der Linen1997] Van der Linen W.J., Hambleton R.K. (eds.) (1997) Handbook of modern item response theory, Springer, New York.

[Yamanishi2001] Yamanishi K. and Li H. (2001) "Mining from open answers in questionnaire data", Proceedings of the Seventh ACM SIGKDD, pp. 443-449.