Computers & Education 137 (2019) 91–103

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems

Konstantinos Pliakos[a,c], Seang-Hwane Joo[b,c], Jung Yeon Park[b,c,*], Frederik Cornillie[b,c], Celine Vens[a,c], Wim Van den Noortgate[b,c]

[a] KU Leuven, Department of Public Health and Primary Care, Faculty of Medicine, Campus Kulak, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium
[b] KU Leuven, Faculty of Psychology and Educational Sciences, Campus Kulak, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium
[c] imec - ITEC - KU Leuven, Belgium

ARTICLE INFO

Keywords:
Item response theory
Decision tree learning
Machine learning
Adaptive learning system
Cold-start problem

ABSTRACT

Adaptive learning systems aim to provide learning items tailored to the behavior and needs of individual learners. However, one of the outstanding challenges in adaptive item selection is that often the corresponding systems do not have information on initial ability levels of new learners entering a learning environment. Thus, the proficiency of those new learners is very difficult to be predicted. This heavily impairs the quality of personalized items' recommendation during the initial phase of the learning environment. In order to handle this issue, known as the cold-start problem, we propose a system that combines item response theory (IRT) with machine learning. Specifically, we perform ability estimation and item response prediction for new learners by integrating IRT with classification and regression trees built on learners' side information. The goal of this work is to build a learning system that incorporates IRT and machine learning into a unified framework. We compare the proposed hybrid model to alternative approaches by conducting experiments on two educational data sets. The obtained results affirmed the potential of the proposed method. In particular, the obtained results indicate that IRT combined with Random Forests provides the lowest error for the ability estimation and the highest accuracy in terms of response prediction. This way, we deduce that the employment of machine learning in combination with IRT could indeed alleviate the effect of the cold start problem in an adaptive learning environment.

1. Introduction

Over the last decade, online learning environments have received a rapidly growing attention. Technology-enhanced environments are deemed to have a greater potential than traditional classroom learning as they are capable of personalizing students' learning opportunities based on adaptive learning technologies (Albatayneh, Ghauth, & Chua, 2018; Truong, 2016; Ortigosa, Martín, & Carro, 2014; Kalyuga & Sweller, 2005; Shute & Towle, 2003; Brusilovsky, 1999). The goal of an adaptive learning system is to modify instructions using a set of predefined rules (Burgos, Tattersall, & Koper, 2007; Marcos-García, Martínez-Monés, & Dimitriadis, 2015) and to provide learning materials (or items) tailored to the behavior and needs of individual learners (Wauters, Desmet, & Van

* Corresponding author. KU Leuven, Faculty of Psychology and Educational Sciences, Campus Kulak, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium.
E-mail addresses: konstantinos.pliakos@kuleuven.be (K. Pliakos), seanghwane.joo@kuleuven.be (S.-H. Joo), ellie.park@kuleuven.be (J.Y. Park), frederik.cornillie@kuleuven.be (F. Cornillie), celine.vens@kuleuven.be (C. Vens), wim.vandennoortgate@kuleuven.be (W. Van den Noortgate).

https://doi.org/10.1016/j.compedu.2019.04.009
Received 6 June 2018; Received in revised form 15 April 2019; Accepted 17 April 2019
Available online 20 April 2019
0360-1315/

den Noortgate, 2010). An example is a system that selects items of an appropriate difficulty level. For such a system, it is crucial that the system builds up enough information about the learners' ability levels and predicts their responses to the learning items in a timely and accurate manner. However, one of the challenges of the process is that the system may have very limited (or no) information on initial ability levels of new learners when they enter a learning environment. In this case, it takes a long time until estimates of an acceptable accuracy are obtained of the learners' learning proficiency. Therefore, the system is likely to fail to recommend tailored items during the initial phase of the learning environment. This issue is referred to as the cold start problem (Schein, Popescul, Ungar, & Pennock, 2002). Studies showed that the cold start problem often makes new learners to abandon the system due to inappropriate first recommendations, which are experienced as frustrating (Bobadilla, Ortega, Hernando, & Bernal, 2012; Mackness, Mak, & Williams, 2010). Also lack of motivation, anxiety and boredom may be associated with the failure of adaptive item selection (Wauters et al., 2010; Klinkenberg, Straatemeier, & van der Maas, 2011; Ostrow, 2015; Jagust, Boticki, & So, 2018). It is therefore crucial to further develop methodologies and models that tackle this cold start problem.

In fact, this problem has received a considerable amount of attention in another context, the context of recommender systems that seek to predict the rating that a user would give in e-commerce or online streaming websites to an item based on his or her interest (e.g., books, movies, songs). Many studies (e.g., Barjasteh, Forsati, Ross, Esfahanian, & Radha, 2016; Contratres, Alves-Souza, Filgueiras, & DeSouza, 2018; Fernández-Tobías et al., 2016; Forsati, Mahdavi, Shamsfard, & Sarwat, 2014; Lika, Kolomvatsos, & Hadjiefthymiades, 2014; Ling, Lyu, & King, 2014, pp. 105–112; Menon, Chitrapura, Garg, Agarwal, & Kota, 2011; Pereira & Hruschka, 2015; Tang and McCalla, 2004a,b) proposed data mining and machine learning techniques (specifically, collaborative filtering algorithms) to address the cold-start problem using the side information about existing users (i.e., users' attributes) to make recommendations for new users with similar profiles. However, most of their approaches focus heavily on the prediction of the new user's rates on a given set of items, lacking the psychometric component i.e., assessment of the users' latent traits. In adaptive learning systems, however, getting insight in the latent ability level of persons is of crucial importance because of its role in evaluating how effectively the learning process is working and how the learner performed on those learning programs. Only a very limited number of studies (Tang and McCalla, 2004a,b, August; Sun, Cui, Xu, Shen, & Chen, 2018) have paid attention to the cold-start problem in the context of online-learning. Therefore, this study aims to answer the research question: *how can the effect of the new learner's side information (e.g., age, relevant courses taken, IQ, pre-test scores) be exploited in order to estimate the learner's initial ability and the corresponding performance on items with a variety of difficulty levels*?

With respect to the ability assessment, the use of item response theory (IRT; Van der Linden & Hambleton, 1997) is considered as one of the most recognized psychometric methods. The basic IRT model, the Rasch model (Rasch, 1960), is based on the idea that the probability of correctly solving an item is a logistic function of the difference between a person parameter and an item parameter, that are often interpreted as the person's ability parameter and the item's difficulty parameter. Fitting the model to responses of learners on a set of items allows to estimate the learners' ability levels and the item difficulties, which can be used afterwards to provide learners with the most informative item. The larger (smaller) the person's ability is compared to the item difficulty, the larger (smaller) the probability on a correct response. IRT models have a strong tradition in testing situations, because of several advantages (Hambleton & Jones, 1993; Van der Linden & Hambleton, 1997), which will be discussed below. Once a set of calibrated items in the bank is available (a measurement scale of items is constructed), new learners can be placed on the scale by assessing how successful the person responds to some of these calibrated items. IRT is often used in computerized adaptive testing (CAT; Van der Linden, 2009), in which after each response the ability estimate of a test taker is updated and an item is given with a difficulty that matches closely the ability estimate. In this way, shorter tests are sufficient for obtaining an accurate view on the learner's ability. The idea of selecting items whose difficulty matches the ability of the learner is also applicable to the online learning environments. Yet, while the goal of the adaptive testing is to gain efficiency in assessing test takers' ability level and examine their relative standing in the population, the goal of adaptive learning is to enhance the learning progress by providing more personalized learning items (Zhang & Chang, 2016).

Not only IRT, also machine learning techniques can be valuable for adaptive item selection. Response predictions for new learners can be made by addressing the ability estimation as a regression task based on machine learning. The system can predict the new learner's responses by using first a machine learning model to estimate the ability parameter of this new learner and then use this estimated ability parameter to predict the responses with IRT. Alternatively, the response prediction can be addressed as a multi-target prediction task (Kocev, Vens, Struyf, & Džeroski, 2013). In this case, the system can, for example, employ a decision tree-based learning model in a multi-output setup (Kocev et al., 2013). The machine learning model is learned on a training set of learners containing their descriptive features (background information) and their responses.

This study proposes a hybrid approach by combining the strength of IRT models with machine learning. Specifically, the approach integrates the Rasch model with the use of classification and regression trees (Breiman et al., 1984) trained on side information (i.e., learner attributes/features). In this way, the model potentially can surpass the cold start problem and make reasonable predictions for new learners. We suggest the use of decision tree-based learning methods (i.e., single decision trees or ensembles, such as Random Forests; Breiman, 2001) among various machine learning methods because of their interpretability and visualization properties. In addition, when they are extended to ensembles, their predictive performance is greatly improved (Fernández-Delgado, Cernadas, & Amorim, 2014). In our case, this means that we can get more accurate predictions of a student's latent ability and responses. In order to validate the effectiveness of the proposed hybrid approach, we conducted experiments using educational data sets including background information on learners and items.

The novelty of our approach lies in its capability of incorporating learner features (a) to estimate the new learner's initial ability level when getting engaged in an e-learning environment; and (b) to predict the corresponding responses to a given item bank. In particular, when the estimation of the ability of the learners is concerned, the hybrid system employs:

1. IRT to estimate the ability of the learners for which we already have data on their responses to items.
2. A regression tree-based method trained using the features that characterize the learners and the IRT generated abilities.

In summary, the overarching goal is to develop a method that integrates decision tree-based techniques and IRT for predicting the response pattern and estimating the latent ability of new learners. As the current study aims to address the cold start problem in adaptive learning environments, our focus is to investigate the performance of the hybrid method against one of the most common approaches used in computerized adaptive systems i.e., assuming at the start of the algorithm that the new learner has an average ability (e.g., Van der Linden & Veldkamp, 2004). When it comes to adaptive learning systems, usually there is an initial phase where new learners are given some items and their ability is estimated based on the responses to those first items. However, in case these recommendations are incompatible (e.g., too easy, too difficult) then the learner gets frustrated or even abandons the effort (Mackness et al., 2010). We investigate the possibility to build a system that improves the performance of the initial phase of an adaptive learning system, without making use of prior ability tests.

The rest of the paper is organized as follows. In the next section, we start by presenting previous studies that are relevant to this work. In section 2, we describe general frameworks of IRT and decision tree-based methods. We then propose a hybrid approach combining the two methods in order to address the cold-start problem. Next, in section 3, we introduce two educational data sets (one for an educational testing and the other for an online learning environment) for the evaluation of our approach and present the experimental strategy. The results are demonstrated in section 4. We discuss our findings and provide concluding remarks in section 5.

1.1. Related work

Although several recent studies proposed methods to tackle the cold-start problem in recommender systems, their applications in an educational domain (i.e., the adaptive learning systems) are an underexplored topic.

Like in a typical educational testing environment, the majority of the online learning platforms currently do not use any prior information on new learners for personalized learning (Thai-Nghe, Horvath, & Schmidt-Thieme, 2011). After an item bank is calibrated by IRT modelling, which means that item difficulties have become available, the naive method renders initial items without knowing anything about the new learners. The item can be selected randomly or by using the average ability estimate of the pre-existing learners as the starting value for the new learners. Thus, the prediction performance for new learners is not very efficient in the sense that it may take longer for the learning system to estimate the learner's ability level with sufficient accuracy.

A number of studies tackled the cold-start problem in relation to recommender systems in general. The recommender systems refer to information filtering and decision support systems that seek to predict the ratings or preferences a user would give to an item (e.g, movies, music, books, and products) in e-commerce or online streaming sites. The systems typically utilize a variety of collaborative filtering (CF) algorithms that generate automatic predictions about the user by collecting information from other users who shared similar ratings or preferences. To handle the cold-start problem, several studies used data mining techniques that incorporate user features (age, gender, and social contact) in the CF. Said and Bellogín (2014), Guo, Zhang, and Yorke-Smith (2013), and Vozalis and Margaritis (2004) proposed a modified version of k-nearest neighbors (k-NN) by adding a user demographic vector to the user profile and embedding it in the CF. Similarly, Son, Minh, Cuong, and Canh (2013) proposed using a fuzzy clustering method that incorporates the demographic features in the filtering system. Fernández-Tobías et al. (2016) proposed adding the users' personality information to a matrix factorization (MF) model that incorporates user features to improve the recommendation where there are no ratings for the new users. Contratres et al. (2018) showed that the user cold-start issue can be alleviated by using sentiment analysis based on support vector machine (SVM) (Burges, 1998) in the recommender systems.

Likewise, despite the recent popularity and prolificacy of the cold-start problem in the recommender systems, not much research has been done in the domain of adaptive learning systems. However, there have been various studies applying machine learning in educational systems in general. The majority of them harness the prediction accuracy of machine learning to develop predictive models for students. These models are often trained over students' demographic characteristics or other kinds of student related attributes/features (e.g., school progress, number of books at home, dyslexia, dyscalculia, etc.), targeting at performing grade or drop-out predictions. More specifically, Kotsiantis (2012) built a decision support system to predict students' performance. The system was trained on students' demographic features and marks in written assignments addressing student grade prediction as a regression problem. Kai, Almeda, Baker, Heffernan, and Heffernan (2018) used a decision tree to classify students into two groups – productive persistence or wheel-spinning. Rovira, Puertas, and Igual (2017) employed machine learning for students' grades and dropout intention prediction. The authors also proposed a personalized course recommendation model. Course preferences as well as course completeness ratios were studied using decision tree learning in (Hsia, Shie, & Chen, 2008). Lykourentzou, Giannoukos, Nikolopoulos, Mpardis, and Loumos (2009) proposed a dropout prediction method for e-learning courses using a combination of machine learning techniques. Vie, Popineau, Bruillard, and Bourda (2018) proposed a determinantal point process to select adaptive items for new learners, using ability and difficulty estimates calibrated by a cognitive diagnosis model. Park, Joo, Cornillie, and Van der Maas (2018) proposed a psychometric method to reduce the new learner cold-start problem, zooming in on the adaptive learning systems. Based on an explanatory IRT model trained by learner-item interaction data and learner features (e.g., age, gender, learning disability), their method first provides initial ability estimates for the new learners based on his or her profiles, then allows to make recommendations for the most informative items. Based on the previous studies, it is clear that background information of learner could contribute significantly to obtain models that precisely predict the learner's performance. In addition, more learner's information can provide more precise and accurate prediction in the context of machine learning in educational assessment.

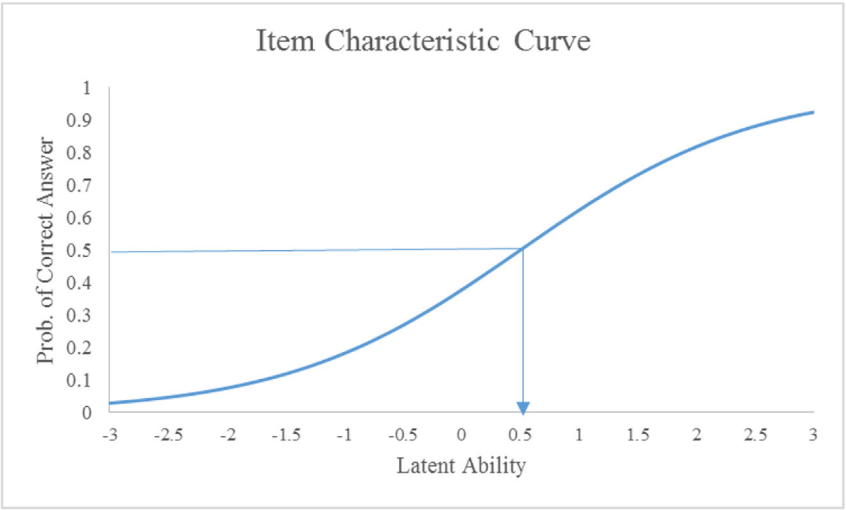

Fig. 1. Item characteristic curve (ICC) of the Rasch model for an item (difficulty $\beta = 0.5$).

2. Methods

In the following, we propose a hybrid approach that combines Item Response Theory and Decision Tree-based learning. First, we describe in more detail both components.

2.1. Item response theory

IRT has been widely used in educational and psychological settings, especially in educational assessments, to assess persons' abilities or to develop learner's cognitive or non-cognitive measurements. For example, in large-scale assessments such as Trends in International Mathematics and Science Study (TIMSS; Mullis & Martin, 2013), tests are often constructed and evaluated using various IRT models by analyzing characteristics of items, such as item discrimination and item bias. For more detail information about test and measurement theory, see Allen and Yen (2001) and McDonald (2013).

In general, IRT models often assume that the probability of a correct response to an item follows a logistic or a normal ogive curve. That is, the probability of correct response to the item *j* increases as learner *i*'s ability increases monotonically with boundaries between 0 and 1. Equation (1) shows the conditional probability function of the correct response for learner *i* to item *j* using the Rasch model,

$$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \quad (1)$$

where X_{ij} is a dichotomously scored observed variable (with 1 referring to a correct response, 0 to an incorrect response), θ_i is a latent ability parameter for learner *i*, and β_j is item difficulty parameter for item *j*. Also, Fig. 1 shows the item characteristics curve (ICC) of the Rasch model for the item with difficulty parameter, β_j, of 0.5. As shown in Fig. 1, the probability of correctly responding to the item increased from 0 to 1 as the learner's ability increases. Also note that when the latent ability is equal to 0.5 (item difficulty), the probability of correctly responding to the item is 0.5. This implies that the probability of giving the correct answer is determined by item difficulty and latent ability; when a learner's ability is equal to the item difficulty, the probability of given the correct answer is 0.5.

In educational and psychological measurement, several alternative IRT models were developed. For instance, if one is not only interested in the item difficulties but also in how well the items discriminate between persons with higher and lower ability, the two-parameter logistic (2 PL) model can be utilized that includes a second item parameter reflecting the slope of the logistic curve. The three-parameter logistic (3 PL) model, adds to the item difficulty and discrimination parameters a guessing parameter, accounting for situations where the probability of a correct answer for a learner with low ability can be increased by guessing the item. As a result, the ICC of the 3 PL model has a lower asymptotic boundary across the latent continuum (Birnbaum, 1968).

One of the advantages of IRT models over traditional Classical Test Theory (CTT) is that the item parameters in principle do not depend on the characteristics of the samples that are used to calibrate the items. The calibration sample does not have to come from the same population as the persons that are tested afterwards with the calibrated items. The second advantage of the IRT modeling is that it enables researchers to conduct more sophisticated item and latent ability analysis using item information functions. Consequently, more advanced testing and learning environments can be further developed using the IRT modeling. For example, the IRT models have been implemented in adaptive learning system (e.g., Wauters et al., 2010). In their paper, the authors explored and illustrated the feasibility of applying the Rasch model for adaptive learning system.

2.2. Decision tree learning

One of the most exhaustively studied fields in machine learning and data mining is supervised learning (Jordan & Mitchell, 2015). The instances in supervised learning are represented by features and are associated with targets. The task is to predict a target value by building a prediction function over a training set of instances with known target (Witten, Frank, Hall, & Pal, 2016). Using this prediction function, we can perform predictions for new (unseen) data. Among the various prediction tasks, classification and regression are the most common. In classification, one is interested in predicting categorical values (i.e., class labels where each instance is assigned to). In regression, one is interested in predicting numerical values. Furthermore, both tasks have a multi-output extension. More specifically, the assumption in single target prediction tasks is that an instance corresponds to only one class out of two (binary classification) or more classes (multi-class classification). However, in many applications this assumption does not hold and instances may belong simultaneously to more than one class. For example in the field of text mining, a document can be associated with many topics at the same time. In social media, an uploaded image can be associated with many tags. This kind of data are called multi-label data and the machine learning methods that build prediction functions over these data are called multi-label methods (Tsoumakas, Katakis, & Vlahavas, 2009; Tsoumakas & Katakis, 2006).

Decision Tree learning (Breiman, Friedman, Olshen, & Stone, 1984) is among the most popular machine learning methods. It is used mainly for classification and regression but also for many other tasks. Decision trees consist of nodes and edges that connect those nodes. Every node usually has an ingoing edge connecting it with its parent node and outgoing edges connecting it with its children. The first node of the tree is called the root while the nodes without an output edge are called leaves. Decision tree learning knows many advantages, such as scalability, computational efficiency, and interpretability. Although there are predictors which are more powerful, the interpretability and visualization properties of decision trees are great advantages.

Following the Predictive Clustering Tree (PCT) framework (Blockeel & De Raedt, 1998), decision trees are constructed with a top-down induction method. Every node is considered to be a cluster of the data. The root node is a cluster that contains all the training instances. Starting from the root node, all the nodes are recursively split by applying a test to one of the features that describe the instances. The best split is found by evaluating a split quality criterion. This criterion is based on the machine learning task at hand. For example, information gain is usually used in classification while variance reduction is used in regression. The tree growing procedure stops when a stopping criterion is fulfilled or no further split can be performed. The final nodes are called leaves and the prediction of the target variables is based on a function called the prototype. For classification, this function is the majority class assigned to the instances in the leaf. For regression, this function is the average of the target values that correspond to the instances in a leaf. When a new instance arrives, it traverses the tree ending up in a leaf node. The target value that corresponds to that leaf is assigned to the new instance. The path that is followed by an instance when it traverses the tree consists of a set of tests and can be used thereby as a rule, providing this way interpretability of the performed prediction. An example of a decision tree is demonstrated in Fig. 2. The first split is based on the feature “Age”, instances (e.g., students) whose age is below (or equal to) 10 follow the left child node and the others the right one. The next two splits are based on the gender and on whether an instance has or not dyslexia.

PCTs are able to predict multiple targets at the same time by using an appropriate split criterion and prototype function. Such trees are called multi-output or multi-target decision trees (Kocev et al., 2013). The corresponding prediction task is called multi-label classification if the targets are binary values and multi-target regression if they are continuous values. The quality V of a split is obtained by computing the split quality criterion for every output and summing the corresponding values V_j, i.e., $V= \sum_{j=1}^{N} V_j$. The prototype function returns the average target *vector* of the training instances in a leaf. When it comes to multi-label classification, this corresponds to a vector composed of the estimated probabilities of the labels (Vens, Struyf, Schietgat, Džeroski, & Blockeel, 2008).

Random Forests (RF) (Breiman, 2001) are an ensemble learning method composed of a collection of multiple decision trees. An important characteristic of this method is the diversity that is enforced among the trees. This diversity is obtained by using bootstrap replicates of the training set and random selection of the features describing the samples. More specifically, each decision tree of the ensemble is constructed on a random subset of the training set. Every node of that tree is split by computing the best possible split among a random subset of selected feature candidates. The final prediction is yielded as the average of the predictions of individual trees. Random Forests surpass the lack of variance and the overfitting drawbacks of single decision trees.

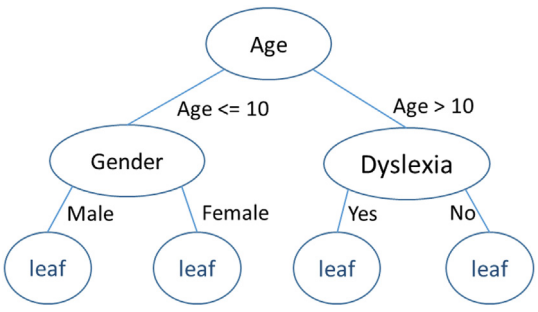

Fig. 2. An example of a decision tree.

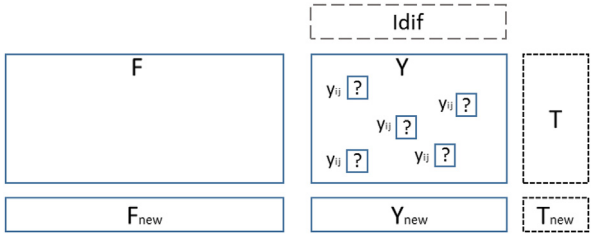

Fig. 3. Illustration of the hybrid prediction model.

2.3. A hybrid approach

We propose a hybrid system that combines IRT and decision tree-based learning. The inputs in the hybrid system are the features of learners (i.e. background variables, such as school, age, gender, etc.) and the responses of learners that were already involved in the learning environment. The values of interest are the new learners' abilities and responses. In more detail, the available data for the system are the features of a set of learners and their already known responses. The system is built on these data and the goal is to predict the abilities and responses of new learners when they enter the system. Let the existing learners' features (i.e., learners already included in the system) be denoted as F, their abilities as T, their responses as Y and their "unknown" or missing responses as Y_u.

The hybrid prediction model is illustrated in Fig. 3. First, an IRT model is applied to Y, estimating item difficulties (Idif) and learners' abilities (T), as displayed in Section 2.1. Note that any missing responses ($y_{ij} \varepsilon Y_u$) in the target space are also predicted by the IRT model (a 1 is predicted if the probability on a correct response that is expected using the estimated Idif and T is 0.5 or higher, 0 otherwise). Next, a regression tree-based model is trained on the existing learners' feature space (F) and their target space (T). When a new learner arrives, in this way, the trained regressor allows to predict the latent ability parameter of the new learner (T_{new}; an ability parameter estimate of the new learner) by using his or her background information (i.e., features). In summary, the machine learning model is trained using existing data, specifically the features and the ability parameter of the existing learners. Next, when the model has been trained, it can predict the ability parameter of the new learner (T_{new}).

Finally, in order to predict responses of the new learner (Y_{new}), we consider two approaches in our hybrid system. Fig. 4 illustrates how the system goes on to the two types of predictions. The first approach (at the bottom right) predicts responses of the new learner (Y_{new}) based on the estimated ability parameter of the new learner (T_{new}) and the estimated item difficulties (Idif). Given the two components are known, it is natural that IRT formula (see Equation (1)) allows to estimate the learner's responses to each individual item in Y_{new}. In the second prediction approach (at the bottom left), on the other hand, the response prediction can be viewed as a multi-target prediction problem. Therefore, the system predicts the new learners' responses (Y_{new}) by means of a multi-output learning model (i.e., a model that learns to predict multiple responses simultaneously). During the training process of the decision tree-based learning model, the splits are based on the features of the learners included in the training and the variance reduction is computed on their responses. To be clear, the response matrix of existing learners, Y, has been previously completed by using IRT to predict any missing values. When a new user arrives, the trained model can perform predictions based on background information (i.e., features) of the new learner.

Note that although here we choose tree-based learning methods (i.e., decision trees, random forests) for the machine learning

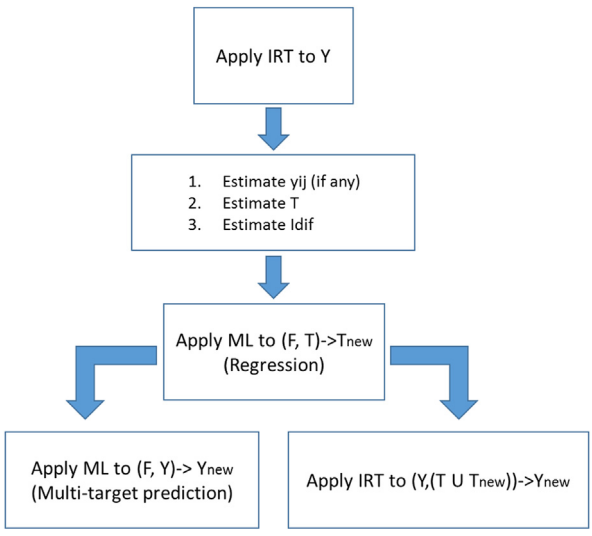

Fig. 4. Outline of the hybrid system.

counterpart of our approach, other methods could be used as well. For example, one could use Support Vector Machines (SVM), k Nearest Neighbors (k-NN), or linear regression. However, we promote the employment of tree-based machine learning methods due to their interpretability advantages. It is often desirable to have interpretable models and gain insights while performing predictions. Only through interpretability we can understand and trust the decisions made by machine learning (Doshi-Velez & Kim, 2017). Moreover, tree-based methods are scalable and generally very fast predictors. The methodology used in this hybrid approach is similar to methods handling sparse data in educational assessment. Because large scale assessments often make use of incomplete designs (each person only answers a subset of items), researchers often use the existing responses to calibrate item and latent ability parameters simultaneously (Gottschall, West, & Enders, 2012; Rose, von Davier, & Xu, 2010). They use background variables to fit the model and predict the nonresponses and the latent ability distribution. Here we investigate the employment of machine learning models and more specifically tree-based learning models to perform predictions. We combine IRT and machine learning to predict the new student's responses and latent ability in the learning system. The goal is then to use these predictions as starting values in an adaptive learning system.

3. Evaluation of the hybrid approach

To illustrate and compare the new hybrid system, we apply it on two real datasets, described below. For the implementation and experimental evaluation of our system we used the machine learning library Scikit-learn (Python) (Pedregosa et al., 2011). Scikit-learn contains all the machine learning algorithms used in this study as well as the relevant evaluation metrics that were employed. We also used the library NumPy for data handling purposes. When it comes to IRT, we used the IRT implementation in the programming language R and specifically the ltm package (Rizopoulos, 2006).

3.1. Assessment dataset

The first dataset consists of 2044 students and 20 items that were used for a statistics exam at our university (further information about university was erased due to double blind restrictions). In addition, the background information associated with the students who took the exam was collected (i.e., study program, gender, school progress, language they speak at home, number of books at home, Socioeconomic Status (SES), dyslexia, dyscalculia, ADHD, ASS, other learning problem, school type, subsidized education, rural urban, concentration, province, language of their friends, hours of math per week, academic self concept (math), academic self concept, attest 2nd grade, math ambitions, parents' attitude toward math). In total, 23 background variables were provided. The number of books at home, SES, concentration, hours of math per week, academic self concept (math), academic self concept, math ambitions, and parents' attitude toward math were recorded as continuous variables and the other variables were categorical variables. Among the categorical variables, some dichotomous variables related to a learning disability such as dyslexia, dyscalculia, ADHD, ASS and other learning problem were included. Those variables were recorded as 1 if a student has a learning disability, and 0 otherwise. We included all 23 background variables because they were closely relevant to the performance assessment and we wanted as many variables as possible in order to let the machine learning techniques learn from these data and “decide” which are important and which not. The responses from the students were recorded as a dichotomous variable (1 = correct, and 0 = incorrect). The proportion of correct responses is .55.

3.2. Learning dataset

For evaluation purposes, we also used a dataset collected using the Statistics-Online learning environment (Kadengye, Ceulemans, & Van den Noortgate, 2015). The Statistics-Online was designed as an item-based e-learning environment for students in the Educational Sciences, Speech Therapy and Audiology Sciences program at our university. The Statistics-Online environment supplements students' learning by providing exercises and feedback on students' responses to the questions. The example dataset was obtained from the one of modules (regression analysis) in Statistics-Online environment and the dataset consists of 145 multiple choice items. The item responses were obtained as follows: 1) students were allowed to log into the environment at any choice of times, 2) randomly ordered items were given to students, 3) students' correctness (1 or 0) were recorded and 4) the corresponding feedback on why the selected answer is correct or wrong was given in order to enhance learning. Note that the total number of items per student varied as the median number of items was 13 and the mean of items per student was 20 with a minimum of 10 and maximum of 124. The proportion of correct responses in this dataset is 0.59. The total number of students engaged was 229 and consequently there was a number of missing responses in the dataset. The dataset recorded the background information of students as well: gender, campus, year, program, previous statistics courses, total hours of study sessions. The background variable ‘total hours of study sessions’ was recorded as a continuous variable and the other variables as categorical variables. Similar to the previous example, we included all these variables because they could be useful information to explain the students' mathematical learning process and the previous study also provided an evidence (Kadengye et al., 2015) for that. The 2-dimensional visualization of the used datasets is illustrated in Fig. 5. The projections were made using principal component analysis (PCA).

3.3. Evaluation strategy

We evaluated our system using the two datasets in a 10 fold cross validation (10-CV) setting (Kohavi, 1995). In *k*-fold cross validation the benchmark dataset is randomly divided in k subsets of equal size. Next, a subset is selected as the “test” subset for the

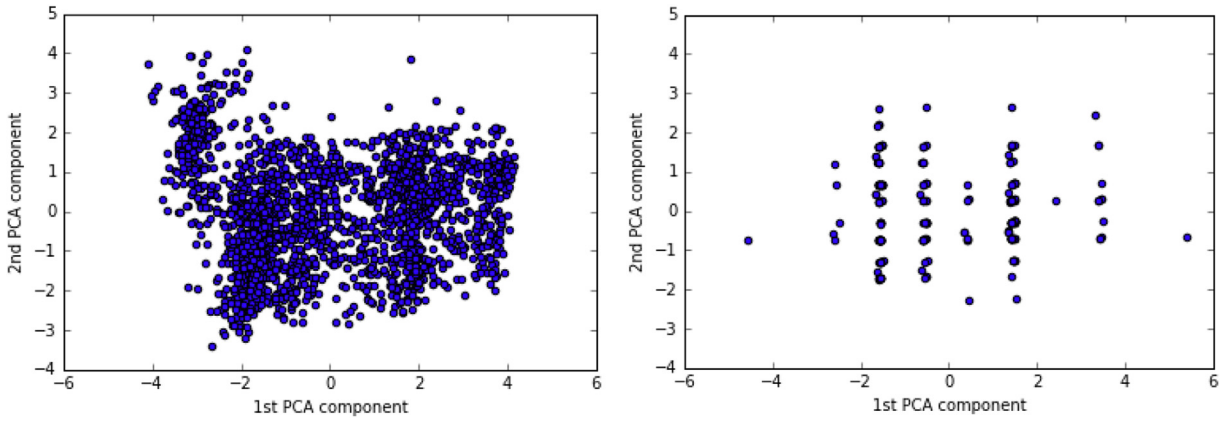

Fig. 5. 2-D projection of the two employed datasets (on the left, the assessment dataset and on the right, the learning dataset).

evaluation of the predictions. This "test" subset represents the new learners in our system. The remaining *k*-1 subsets are used for "training" the model. The process is repeated k times, each time using a different subset as the "test" subset. In the end, the final evaluation result is the average of the *k* obtained results from the *k* folds.

The predictions of the learners' response patterns that were made by our model were compared to the observed (true) values. Note that response patterns of the test datasets are directly observable, thus, it is possible to compare the predicted responses to the true responses. However, the latent abilities are not observable before the employment of IRT. To compute the accuracy of the latent trait estimate prediction in the cross validation setting, we considered the IRT-estimated latent ability parameters from each training subset as "true" values and then estimated the latent ability parameters for the test subset.

For comparison purposes, we used Support Vector Machine Regression denoted as SVR (Burges, 1998), Linear Regression, and k Nearest Neighbors (k-NN). These machine learning models are often used in educational studies. The k-NN or alternative versions of it was used in Said and Bellogín (2014), Guo et al. (2013), and Vozalis and Margaritis (2004). SVM was employed by Contratres et al. (2018). SVM was also employed in a different setting (drop-out or grades prediction) in Lykourentzou et al. (2009) and Rovira et al. (2017). Finally, we compare the approach with imputing the ability values for new users by randomly drawing from the ability estimates from existing users, or by taking the overall mean ability estimate. The approach of using the average values to handle the cold-start problem was followed by Thai-Nghe et al. (2011).

For Random Forests, 100 trees and a stopping criterion of 3 samples per leaf were used. For single trees, the stopping criterion was 30 samples per leaf. For SVM, the radial basis function (RBF) kernel (Burges, 1998) was used. The γ parameter was set equal to the inverse of the number of features and the C parameter was selected from a range of (0.01, 0.1, 1, 10, 100). In particular, it was optimized using the training set in an internal 3-fold CV. For the inner tuning, 3-fold CV was used instead of 5 or 10 for computational reasons. For k-NN, the number of nearest neighbors (k) was selected from a range of (3, 5, 7, 9, 11). It was optimized in an internal 3-fold CV.

To evaluate the accuracy of the predictions made by our system, we employed several evaluation criteria. In particular, the mean squared error (MSE) was used for the evaluation of the estimated (predicted) latent abilities for new learners (i.e., learners from the test set). The MSE is often used as a measure of the "precision" of the estimated (predicted) parameters and is defined as:

$$MSE(\theta_j, \hat{\theta}_j) = \frac{1}{\mathrm{n}} \sum_{j=1}^{n} (\theta_j - \hat{\theta}_j)^2 \tag{2}$$

where θ_j is the true latent ability parameter, $(\hat{\theta}_j)$ is the predicted latent ability parameter, and n is the total number of learners in the test set. Of course, because we work with real datasets, rather than with simulated datasets, we do not really know the abilities of the learners from the test set. Therefore, we consider the ability estimates obtained when using IRT on the whole dataset as the true values.

When it comes to the predicted response patterns for the new learners, the evaluation measures that were employed are the Area Under Receiver Operating Characteristic curve (AUROC) and Area Under Precision Recall curve (AUPR). Note that a ROC curve represents the relation between true positive rates TP/(TP + FN) and false positive rates FP/(FP + TN) at various probability thresholds. A Precision Recall Curve is defined as the Precision: TP/(TP + FP) against the Recall: TP/(TP + FN) at various probability thresholds. The true positive rate is the same as recall, and is also denoted as sensitivity, while the false positive rate is also denoted as (1-specificity). Both AUROC and AUPR were used in a micro-average setup. In case of totally random predictions, the AUROC is approximately equal to 0.5 and AUPR is equal to the proportion of the positive class. AUPR is known to provide a better image than AUROC in case of heavily imbalanced or skewed data (i.e., the frequency of one class is substantially higher than the other class) (Boyd, Eng, & Page, 2013; Davis & Goadrich, 2006).

Table 1
MSE Results of the Hybrid Approaches. Value in bold indicates the smallest MSE. Value in parenthesis indicates the standard deviation in the MSEs across folds.

Methods	Assessment Data	Learning Data
Regression Tree	0.5208 (0.0525)	0.2826 (0.0949)
Random Forest Regressor	**0.4496 (0.0348)**	**0.2426 (0.0907)**
SVR	0.5000 (0.0484)	0.2595 (0.1122)
Linear Regression	0.5098 (0.0511)	0.2429 (0.0717)
kNN	0.516 (0.0534)	0.2639 (0.0790)
Mean	0.7671 (0.0523)	0.3349 (0.1443)
Random	6.1768 (0.3072)	5.9915 (1.3251)

4. Results

First, the performance of the system in predicting the ability parameter of the new learners (T_{new}) is described. In Table 1, the obtained regression results are presented in terms of MSE. A comparison with the approach of imputing random values drawn from the estimated abilities from the existing users or the mean value, shows that the performance of using machine learning is relatively effective for predicting the abilities of new learners: for both datasets used in our study, the MSE results are substantially better. As it is reflected in Table 1, Random Forests slightly outperform all the other methods.

The results regarding the prediction of the responses of new learners are presented in Table 2. As described in Section 2.3, predictions of the responses of a new learner can be achieved by using IRT based on the known item parameters and the latent ability parameter predicted by the regression step (machine learning) that precedes. For comparison purposes, we present experimental results using IRT based on learner ability estimation with Regression Tree (RT_IRT), Random Forest (RF_IRT), SVR (SVR_IRT), k-NN (kNN_IRT), and Linear Regression (LR_IRT). We also included IRT results based on the mean value of the abilities of the learners used in the training set (MV_IRT). The obtained results are demonstrated in Table 2 in terms of AUROC and AUPR. The ROC curves of the two datasets are also displayed in Fig. 6. As it is shown, the model succeeds in predicting responses of new learners, as both AUROC and AUPR are much higher than the expected values using random predictions (i.e., 0.5 for AUROC and the positive class frequency for AUPR). Moreover, all the machine learning-based approaches outperform the naive approach MV_IRT. Although the differences are small, the best results were obtained by IRT based on Random Forest ability estimates. This is in line with the finding described above that Random Forest achieves the best performance in terms of MSE. It is interesting to notice that basic Linear Regression (LR)

Table 2
AUROC and AUPR Results for the Hybrid Approaches. Note: RT_IRT = Regression Tree with IRT approach, RF_IRT = Random Forest with IRT approach, LR_IRT = Linear Regression with IRT approach, SVR_IRT = Support Vector Machine Regression with IRT approach, MV_IRT = Mean Value with IRT approach.

	Assessment Data		Learning Data	
Combined Methods	AUROC	AUPR	AUROC	AUPR
RT_IRT	0.7071 (0.0079)	0.7366 (0.0073)	0.7405 (0.0278)	0.7890 (0.0521)
RF_IRT	**0.7164 (0.0054)**	**0.7454 (0.0044)**	**0.7464 (0.0223)**	**0.7952 (0.0480)**
LR_IRT	0.7071 (0.0094)	0.7356 (0.0145)	0.7438 (0.0195)	0.7902 (0.0440)
SVR_IRT	0.7107 (0.0105)	0.7386 (0.0160)	0.7364 (0.0273)	0.7774 (0.0525)
kNN_IRT	0.7077 (0.0075)	0.7355 (0.0111)	0.7406 (0.0235)	0.7891 (0.0377)
MV_IRT	0.6505 (0.0077)	0.6745 (0.0174)	0.7182 (0.0402)	0.7599 (0.0607)
Multi-target prediction				
Decision Tree	0.7106 (0.0086)	0.7414 (0.0107)	0.7441 (0.0261)	0.7934 (0.0479)
Random Forest	**0.7248 (0.0061)**	**0.7535 (0.0049)**	**0.7449 (0.0223)**	**0.7975 (0.0469)**

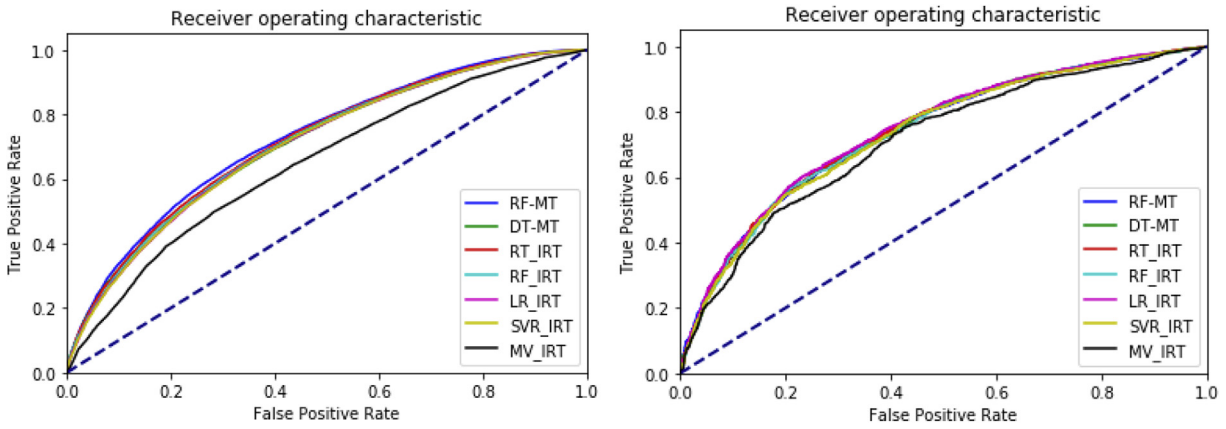

Fig. 6. ROC curves for the 1st (left) and the 2nd (right) dataset.

is generally comparable to the other methods. In particular, it is the third best method for the first dataset and the second best for the second dataset. Alternatively, one could predict directly the responses of a new learner by applying a Decision Tree or a Random Forest model in a multi-output setup, addressing this way the problem as a multi-target prediction task. As it is shown in Table 2, the multi-output Random Forest approach outperforms the multi-output decision trees, as well as all the other approaches, including RF_IRT. At this point, it has to be mentioned that even in that case the integration of machine learning with IRT is still needed as IRT is initially used to predict the missing data (i.e., y_{ij}, described in Section 2.3). However, this multi-output approach does not make use of learner's abilities or item difficulties in order to predict the responses for new learners. Note that this imputation of missing values takes place only in the training phase. Although predictions are made for all the responses in the test set, predictions that correspond to missing responses (i.e, imputed responses) are excluded from the evaluation process (i.e., computation of AUPR, AUROC).

5. Discussion and conclusion

In the current study, we presented an approach that combines psychometric modelling with machine learning techniques. We proposed that a hybrid model can be used as an alternative approach to address the cold start problem by predicting learner's ability in the initial stage of adaptive learning in online learning systems. More specifically, the proposed approach starts with estimating existing learner's abilities based on IRT analysis. Then, a tree-based method is used by regressing the estimated abilities on a set of features that characterize the learners. Of many machine learning techniques available, we selected decision-tree based methods because of their interpretability advantage. Based on the results, new learners' abilities and their responses on the items can be predicted. To empirically investigate our approach, we employed two educational datasets: an assessment and a learning dataset. For validation purposes, we compared the proposed method to alternative approaches including machine learning methods that are not based on decision trees, as well as random prediction.

Results from both empirical datasets suggest that our approach generally leads to relatively precise predictions of the new learner's ability and item responses. More specifically, IRT combined with random forests demonstrated the best performance, outperforming other machine learning approaches as well as naive approaches based on computing the average of the target values or performing random predictions. The hybrid approach outperformed the alternative ones in terms of MSE for learner's ability estimation and in terms of AUROC and AUPR for learner's response prediction. In particular, it substantially outperforms naive approaches (such as approaches based on random predictions) that are currently used for the estimation of new learners' abilities and responses.

Our study has important implications for researchers and practitioners who use adaptive learning systems. Relatively simple approaches have been commonly used (e.g., taking a random or average value) in practical situations to address the cold-start problem. Based on our study results, using the proposed hybrid approach, one can estimate the initial ability of learners more precisely. The cold-start problem yields lack of information about learner's initial ability which sequentially leads to inaccurate predictions in the initial stage of adaptive learning system. Thus, our proposed method can be easily implemented prior to the adaptive learning system. Instead of using relatively simple approaches including random or averaged values, the proposed method can provide more accurate estimations of the learner's ability in the beginning of the system using learner's background information. Once the hybrid model predicts the learner's ability, the adaptive learning system can update the learner's ability more precisely based on the learner's performance.

With the more accurate ability parameter estimate, the adaptive learning system can choose items which difficulty levels more closely match with the learner's current need (even from the very beginning). For example, items can be chosen such that the expected probability of correct responses for the learner is 0.70 (i.e., items with a difficulty level a little bit lower than the learner's current ability level so as not to lose motivation). But depending on the learner's personal preference, the adaptive learning system can select items resulting in a probability of .50 for more challenging path or .90 for easier path, for instance.

The performance of the machine learning methods were similar, especially when it comes to AUROC and AUPR results. However,

besides effectiveness, in comparison to other approaches, the tree-based methods (i.e., decision trees, random forest) have particular advantages such as that they are scalable, and interpretable (i.e., they provide an explanation of the predictions). In addition, they are computationally efficient and can also handle categorical variables. Furthermore, it makes performing predictions for new learners feasible, surpassing a serious bottleneck of IRT. To the best of our knowledge, this is the first work that combines the psychometric model and machine learning to resolve the cold-start problem of new learners in the online learning environments. We recommend peer researchers in adaptive learning systems or in education in general to consider the proposed approach, not only as a tool to perform predictions but also as a means to analyze their samples. In addition, it has to be noted that the proposed approach is rather generic and therefore applicable to other domains.

We recognize limitations of the study. For example, although this study evaluated the accuracy of the method by predicting future student's responses and ability estimates, the practical applicability of the method has not been investigated. The method could be applied for adaptive testing in e-learning environments, and it should be examined to what extent the combined methods can improve the efficiency of adaptive testing. In future work, we plan to combine the proposed method for addressing the cold-start problem with adaptive learning algorithms such as Bayesian Knowledge Tracing (Corbett & Anderson, 1995), Performance Factor Analysis (Pavlik, Cen, & Koedinger, 2009), Deep Knowledge Tracing (Piech et al., 2015), and examine the improved efficiency. A second limitation is that the approach was only applied on two datasets, with specific characteristics. For instance, the sample size for the second dataset can be considered relatively small (N = 229), and comparison results might be different for larger datasets. Also, there is sparseness in the dataset because not all the items were administered to every student in the e-learning environment. In the future, more learning datasets with relatively large number of sample sizes should be used to evaluate the IRT combined with machine learning method.

Furthermore, it would be interesting to compare this approach with more advanced IRT models which include explanatory IRT (e.g., Wauters, Desmet, & Van den Noortgate, 2012). Given that the explanatory IRT model can implement background variables of learners for computing the probability of correct answers using the context of multilevel modelling, it would be worthwhile to compare the performances of machine learning approaches with the explanatory IRT model in the context of learning environment. Another extension of IRT that was proposed recently is IRTree (IRTree; De Boeck & Partchev, 2012). The IRTree model was developed based on a tree structure and allows for multiple sources of individual variations for a response scale. Later, an extension of the IRTree model was further formulated to incorporate multiple parametric forms, dimensionality and choice of covariates (Jeon & De Boeck, 2016). In regard of predicting multiple skills within a course, we acknowledge that the IRT model we used in this study assumes a unidimensional skill space; in the future work, it will be interesting to integrate current machine learning model into multi-dimensional IRT models (that allow that the multiple skills are correlated among each other) to predict the new learner's proficiency in a fine-grained way. In addition, when multiple courses are involved, the ability estimates obtained from one course may be valuable background information that can be used in the model for another course. Another interesting topic of future research would be to compare the prediction accuracy between the IRTree model and the hybrid approach (combined with uni- and multi-dimensional IRTs). A simulation study could be conducted under various conditions to empirically show the difference between those approaches.

Finally, in a broad sense, the current study can provide ideas about teaching and learning in any educational setting. For example, in a traditional classroom, it is also likely that the teacher has no information about new students in the beginning of a course. Based on the same principle as our approach, the teacher will handle the cold-start problem by predicting their ability level using the new learner's side information (e.g., relevant courses taken) and observations (such as attitude or participation). Besides that, the teachers will be able to exploit the interpretability advantage provided by our proposed approaches. We also hope our methodological approach can provide insights when electronic devices (e.g., adaptive e-textbook) become more common in the future classroom.

Acknowledgements

"This research includes a methodological approach from the LEarning analytics for AdaPtiveSupport (LEAPS) project, funded by imec (Kapeldreef 75, B-3001, Leuven, Belgium) and the Agentschap Innoveren & Ondernemen. The LEAPS project aimed to develop a self-learning analytical system to enable adaptive learning. This support system can be integrated into educational games and in software supporting professional communication and persons with dyslexia. The study also was partially carried out within imec's Smart Education research programme, with support from the Flemish government."

References

Albatayneh, N., Ghauth, K., & Chua, F. (2018). Utilizing Learners' Negative Ratings in Semantic Content-based Recommender System for e-Learning Forum. *Journal of Educational Technology & Society, 21*(1), 112–125.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory.* Waveland Press.

Barjasteh, I., Forsati, R., Ross, D., Esfahanian, A. H., & Radha, H. (2016). Coldstart recommendation with provable guarantees: A decoupled approach. *IEEE Transactions on Knowledge and Data Engineering, 28*, 1462–1474.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.). *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

Blockeel, H., & De Raedt, L. (1998). Top-down induction of first-order logical decision trees. *Artificial Intelligence, 101*, 285–297.

Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based System, 26*, 225–238.

Boyd, K., Eng, K. H., & Page, C. D. (2013). Area under the precision-recall curve: Point estimates and confidence intervals. *Joint European conference on machine learning and knowledge discovery in databases* (pp. 451–466). Berlin, Heidelberg: Springer.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees.* Wadsworth, New York: Chapman and Hall.

Brusilovsky, P. (1999). In C. Rollinger, & C. Peylo (Vol. Eds.), *Special issue on intelligent systems and teleteaching: Vol. 4*, (pp. 19–25). Künstliche Intelligenz.
Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*, 121–167.
Burgos, D., Tattersall, C., & Koper, R. (2007). Representing adaptive and adaptable units of learning. In B. Fernández-Manjón, J. M. Sánchez-Pérez, J. A. Gómez-Pulido, M. A. Vega-Rodríguez, & J. Bravo- Rodríguez (Eds.). *Computers and education*. Dordrecht: Springer.
Contratres, F. G., Alves-Souza, S. N., Filgueiras, L. V. L., & DeSouza, L. S. (2018). Sentiment analysis of social network data for cold-start relief in recommender systems. In Á. Rocha, H. Adeli, L. Reis, & S. Costanzo (Vol. Eds.), *Trends and advances in information systems and technologies. WorldCIST'18 2018. Advances in intelligent systems and computing: Vol. 746*. Cham: Springer.
Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278.
Davis, J., & Goadrich, M. (2006). June). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning* (pp. 233–240). ACM.
De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software, 48*, 1–28.
Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning.* eprint arXiv:1702.08608.
Fernández-Tobías, I., Braunhofer, M., Elahi, M., et al. (2016). Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Modeling and User-Adapted Interaction, 26*, 221.
Fernández-Delgado, M., Cernadas, Barro S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research, 15*, 3133–3181.
Forsati, R., Mahdavi, M., Shamsfard, M., & Sarwat, M. (2014). Matrix factorization with explicit trust and distrust side information for improved social recommendation. *ACM Transactions on Information Systems, 32*, 1–38.
Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research, 47*, 1–25.
Guo, G., Zhang, J., & Yorke-Smith, N. (2013, August). A novel bayesian similarity measure for recommender systems. *IJCAI '13 proceedings of the twenty-third international joint conference on artificial intelligence* (pp. 2619–2625). .
Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*, 38–47.
Hsia, T. C., Shie, A. J., & Chen, L. C. (2008). Course planning of extension education to meet market demand by using data mining techniques–an example of Chinkuo technology university in Taiwan. *Expert Systems with Applications, 34*, 596–602.
Jagust, T., Boticki, I., & So, H.-. J. (2018). Examining competitive, collaborative and adaptive gamification in young learners' math learning. *Computer & Education, 125*, 444–457.
Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods, 48*, 1070–1085.
Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*, 255–260.
Kadengye, D. T., Ceulemans, E., & Van den Noortgate, W. (2015). Modeling growth in electronic learning environments using a longitudinal random item response model. *Journal of Experimental Education, 83*, 175–202.
Kai, S., Almeda, M. V., Baker, R., Heffernan, C., & Heffernan, N. (2018). Decision tree modeling of wheel- spinning and productive persistence in skill builders. *Journal of Educational Data Mining, 10*(1), 36–71.
Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educational Technology Research and Development, 53*(3), 83–93.
Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice on Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education, 57*, 1813–1824.
Kocev, D., Vens, C., Struyf, J., & Džeroski, S. (2013). Tree ensembles for predicting structured outputs. *Pattern Recognition, 46*, 817–833.
Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the fourteenth international joint conference on artificial intelligence: Vol. 2*, (pp. 1137–1143). San Mateo, CA: Morgan Kaufmann 12.
Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review, 37*, 331–344.
Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications, 41*, 2065–2073.
Ling, G., Lyu, M. R., & King, I. (2014). Ratings meet reviews, a combined approach to recommend. *Proceedings of the 8th ACM conference on recommender systems*.
Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education, 53*, 950–965.
Mackness, J., Mak, S., & Williams, R. (2010). The ideas and reality of participating in a MOOC. *Networked learning conference* (pp. 266–275). United Kingdom: University of Lancaster.
Marcos-García, J. A., Martínez-Monés, A., & Dimitriadis, Y. (2015). DESPRO: A method based on roles to provide collaboration analysis support adapted to the participants in CSCL situations. *Computers & Education, 82*, 335–353.
McDonald, R. P. (2013). *Test theory: A unified treatment.* Psychology Press.
Menon, A. K., Chitrapura, K. P., Garg, S., Agarwal, D., & Kota, N. (2011). Response prediction using collaborative filtering with hierarchies and side-information. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 141–149). .
Mullis, I. V. S., & Martin, M. O. (Eds.). (2013). *TIMSS 2015 assessment frameworks*. Retrieved from Boston College, TIMSS & PIRLS.
Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior, 31*(1), 527–541.
Ostrow, K. (2015). Motivating learning in the age of the adaptive tutor. In C. Conati, N. Heffernan, A. Mitrovic, & M. Verdejo (Vol. Eds.), *Artificial intelligence in education. AIED 2015. Lecture notes in computer science: Vol. 9112*. Cham: Springer.
Park, J. Y., Joo, S. H., Cornillie, F., Van der Maas, H. L. J., & Van den Noortgate, W. (2018). An explanatory item response theory method for alleviating the cold-start problem in adaptive learning environments. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1166-9 [Epub ahead of print].
Pavlik, P. I., Jr., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis - a new alternative to knowledge tracing. In V. Dimitrova, & R. Mizoguchi (Eds.). *proceedings of the 14th international conference on artificial intelligence in education*. England: Brighton.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
Pereira, A. L. V., & Hruschka, E. R. (2015). Simultaneous co-clustering and learning to address the cold start problem in recommender systems. *Knowledge-Based Systems, 82*, 11–19.
Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., et al. (2015). Deep knowledge tracing. *Proceedings of the 29th conference on neural information processing systems, Montreal, Canada*.
Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*.
Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.
Rose, N., von Davier, M., & Xu, X. (2010). Modeling non-ignorable missing data with item response theory (IRT). *ETS Research Report Series, 2010*(1), i–53.
Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PLoS One, 12*, e0171207.
Said, A., & Bellogín, A. (2014). Comparative recommender system evaluation: Benchmarking recommendation frameworks. *Proceedings of the 8th ACM conference on recommender systems* (pp. 129–136). .
Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). Methods and metrics for cold start recommendations. *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 253–260). ACM.
Shute, V., & Towle, B. (2003). Adaptive e-learning. *Educational Psychologist, 38*(2), 105–114.
Son, L. H., Minh, N. T. H., Cuong, K. M., & Canh, N. V. (2013). An application of fuzzy geographically clustering for solving the cold-start problem in recommender

systems. *Proceeding of 5th IEEE international conference of soft computing and pattern recognition (SoCPaR 2013)* (pp. 44–49). .
Sun, G., Cui, T., Xu, D., Shen, J., & Chen, S. (2018). A heuristic approach for new-item cold start problem in recommendation of micro open education resources. *International conference on intelligent tutoring systems* (pp. 212–222). Cham: Springer.
Tang, T., & McCalla, G. (2004a). Evaluating a smart recommender for an evolving E-learning system: A simulation-based study. *Advances in artificial intelligence. Lecture notes in computer science: Vol. 3060*, (pp. 439–443). Berlin, Heidelberg: Springer.
Tang, T., & McCalla, G. (2004b). Utilizing artificial learners to help overcome the cold-start problem in a pedagogically-oriented paper recommendation system. *International conference on adaptive hypermedia and adaptive web-based systems*(pp. 245–254). Berlin, Heidelberg: Springer August.
Thai-Nghe, N., Horvath, T., & Schmidt-Thieme, L. (2011). Factorization models for forecasting student performance. *Conference: Proceedings of the 4th international conference on educational data mining, Eindhoven, The Netherlands, July 6-8*.
Truong, H. (2016). Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities. *Computers in Human Behavior, 55*(PB), 1185–1193.
Tsoumakas, G., & Katakis, I. (2006). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining, 3*, 309–310.
Tsoumakas, G., Katakis, I., & Vlahavas, I. (2009). Mining multi-label data. *Data mining and knowledge discovery handbook*. Springer US.
Van der Linden, W. J. (2009). Constrained adaptive testing with shadow tests. In W. Van der Linden, & C. Glas (Eds.). *Elements of adaptive testing. Statistics for social and behavioral Sciences*. New York, NY: Springer.
Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
Van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*, 273–291.
Vens, C., Struyf, J., Schietgat, L., Džeroski, S., & Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. *Machine Learning, 73*, 185–214.
Vie, J. J., Popineau, F., Bruillard, É., & Bourda, Y. (2018). *International Journal of Artificial Intelligent Education*. https://doi.org/10.1007/s40593-017-0163-y.
Vozalis, M., & Margaritis, K. G. (2004, August). Collaborative filtering enhanced by demographic correlation. *AIAI symposium on professional practice in AI, of the 18th world computer congress*.
Wauters, K., Desmet, P., & Van den Noortgate, W. (2010). Adaptive item based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning, 26*, 549–562.
Wauters, K., Desmet, P., & Van den Noortgate, W. (2012). Item difficulty estimation: an auspicious collaboration between data and judgment. *Computers & Education, 58*, 1183–1193.
Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques.* US: Morgan Kaufmann.
Zhang, S., & Chang, H. (2016). From smart testing to smart learning: How testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning, 1*, 67–92.