Efficient and Robust Model Benchmarks with Item Response Theory and Adaptive Testing

Hao Song, Peter Flach*

University of Bristol (United Kingdom)

Received 21 January 2021 | Accepted 10 February 2021 | Published 23 February 2021

Abstract

Progress in predictive machine learning is typically measured on the basis of performance comparisons on benchmark datasets. Traditionally these kinds of empirical evaluation are carried out on large numbers of datasets, but this is becoming increasingly hard due to computational requirements and the often large number of alternative methods to compare against. In this paper we investigate adaptive approaches to achieve better efficiency on model benchmarking. For a large collection of datasets, rather than training and testing a given approach on every individual dataset, we seek methods that allow us to pick only a few representative datasets to quantify the model's goodness, from which to extrapolate to performance on other datasets. To this end, we adapt existing approaches from psychometrics: specifically, Item Response Theory and Adaptive Testing. Both are well-founded frameworks designed for educational tests. We propose certain modifications following the requirements of machine learning experiments, and present experimental results to validate the approach.

Keywords

Item Response Theory, Adaptive Testing, Model Evaluation, Benchmarks.

DOI: 10.9781/ijimai.2021.02.009

I. Introduction

Thanks to the recent popularity of machine learning and artificial intelligence techniques, researchers and practitioners now have a very considerable choice of models and learning algorithms when facing a given task. However, as choices come with deliberations, selecting an appropriate model is also becoming more challenging. Traditionally, model selection involves two steps:

1. Gather related work and explore existing comparisons.
2. Prepare a shortlist and run the models within the target task for more detailed and local comparisons.

However, given the number of research areas and datasets available now, few research papers provide a fully comprehensive benchmark on all related datasets. There is also a considerable risk of confirmation bias. People tend to focus on datasets where the proposed approach leads to improvements, making it even harder to obtain a fair and comprehensive view of different methods [1]. Regarding the second step above, given the rapid rise in computational demands among recent approaches, it is often impractical to simultaneously cover a broad set of experiments.

Despite the emergence of platforms such as OpenML [2] that aim to collect experimental results via standard configurations, it still requires relatively large numbers of new experiments once a novel task/method is introduced. These additional experiments could take a non-trivial time to run given OpenML's crowdsourcing nature. Although certain research areas and methods can come with formal guarantees, these only cover limited scenarios and most practices in the field still rely on experiments and empirical evaluations. Therefore, in this paper, we consider the problem of efficiently obtaining fair and reliable benchmarks on a set of models and datasets.

To get started, in this paper we focus on the typical setting of predictive machine learning. We assume some labelled datasets and several classifiers can be trained and tested on any possible combination. An experiment includes a set of evaluation measures, and we read the measurements to reflect the performance on any given model-dataset pair. We want to investigate approaches that can accurately quantify performance on a large variety of models and datasets while limiting the overall computational costs. For this purpose, we refer to the fields of psychometrics and testing in education and borrow the frameworks of Item Response Theory [3], [4] and Computerised Adaptive Testing [5], [6]. Both frameworks assume the same scenario, where a participant is assigned several items to answer (response). A typical example would be educational tests, where each student is a participant, and each test question is an item.

Item Response Theory (IRT) is a collection of probabilistic models built on the participants' responses to the items. In IRT, a representative setting assumes each participant has an ability parameter, and each item has a difficulty parameter. Both sets of parameters can affect the collected responses, but are not directly observable. IRT aims to learn these parameters from the collected responses, after which we can quantitatively interpret each participant's level and item with the parameter magnitudes. We can further use these parameters to perform statistical transformations, such as to rank students on their estimated abilities (rather than ranking them on the observed responses).

Computerised Adaptive Testing (CAT) is a framework further built on top of IRT. IRT expects the availability of many responses from different participant-item combinations. Sometimes a specific combination might not be necessary. For instance, it is less informative to give a more challenging question to a student who just failed to

* Corresponding author.
E-mail address: peter.flach@bristol.ac.uk

answer a much simpler one. The purpose of the CAT is to adaptively select the items according to previous responses so that the total number of items used in the test is kept at a relatively low level. Our work's central hypothesis is that IRT and CAT can be used – with some essential modifications – for benchmarking machine learning models.

This paper focuses on predictive machine learning tasks, where every dataset is an item, and each model class is a participant. We aim to investigate the possibilities of using the IRT and CAT frameworks to obtain accurate benchmarks on each model-dataset combination while limiting the total number of experiments. The main contributions of the paper include: (1) We adapt and modify the IRT and CAT frameworks to incorporate the need for model benchmarks as in machine learning. (2) We establish a set of experiments to investigate and compare a set of IRT and CAT approaches in a machine learning context, and demonstrate the benefits of having adaptive testing in typical predictive tasks. The outline of the paper is as follows. We first give a brief introduction of the existing approaches from both IRT and CAT in section II, following proposed modifications on them for our benchmarking requirements in section III. Experiments on some standard models and datasets will be presented in section IV, and finally, additional discussions and insights are provided in section V.

II. Background

This section gives an overview of basic concepts and methods in IRT and CAT and introduces necessary notation. We also discuss some existing work on applying IRT in machine learning.

A. Item Response Theory

Item Response Theory refers to a collection of methods that measure individual abilities, item (question) difficulties, and other potential attributes by checking individual responses to a set of items. IRT models are probabilistic models with latent variables, where the responses are the observations, and abilities, difficulties and other related parameters are the latent variables to be estimated. IRT models are of particular use when the responses distribute differently according to different items, and only averaging the responses does not adequately represent a participant's ability. IRT is therefore particularly suitable for analysing the results of educational exams and many physiological tests. When it comes to machine learning experiments, where different datasets typically come with varying baseline performance, IRT provides an opportunity to treat the performance gains among these datasets fairly.

In the following, we introduce two conventional IRT models and discuss their parameter settings and applications. We use θ to denote the parameter of a particular candidate, and δ and a for item parameters (some IRT models have more than two item parameters). The notation R denotes the random variable of the responses.

1. Two-parameter Logistic Model

The two-parameter (per item) logistic model is defined as follows:

$$R \mid \Theta = \theta, \Delta = \delta, A = a \sim \textbf{Bernoulli}(\mu_{(\theta,\delta,a)}) \tag{1}$$

$$\mu_{(\theta,\delta,a)} = \frac{1}{1 + \exp\big(-a \cdot (\theta - \delta)\big)} \tag{2}$$

from which expectation and variance of R are obtained as follows:

$$\mathrm{E}\Big[R \mid \Theta = \theta, \Delta = \delta, A = a\Big] = \mu_{(\theta,\delta,a)} \tag{3}$$

$$\mathrm{Var}\Big[R \mid \Theta = \theta, \Delta = \delta, A = a\Big] = \mu_{(\theta,\delta,a)} \cdot \Big(1 - \mu_{(\theta,\delta,a)}\Big) \tag{4}$$

Here $R \in \{0, 1\}$ is a binary response variable indicating whether a particular individual answered a particular item correctly, $\theta \in \mathbb{R}$ is the individual's ability parameter, and $\delta \in \mathbb{R}$ is the item's difficulty parameter. The two-parameter logistic model additionally has a discrimination parameter a on the items, which controls how rapidly the response distribution changes when candidate ability varies. Therefore, assume we have two participants with different abilities, an item with high discrimination tends to have higher differences between the responses from the two participants, respectively. Positive discrimination indicates that higher ability leads to higher expectation on the responses, and vice versa. Besides the two-parameter setting, there also exists a few variants on Logistic IRT. The three-parameter setting further adds a guessing parameter which lower-bounds the response expectation. A multinomial setting can also be adapted to support categorical responses beyond the binary setting.

2. Three-parameter Beta Model

While the logistic model supports binary responses, a recently proposed IRT model extends the support to continuous response [7]:

$$R \mid \Theta = \theta, \Delta = \delta, A = a \sim \textbf{Beta}(\alpha_{(\theta,\delta,a)}, \beta_{(\theta,\delta,a)}) \tag{5}$$

$$\alpha_{(\theta,\delta,a)} = \left(\frac{\theta}{\delta}\right)^a \tag{6}$$

$$\beta_{(\theta,\delta,a)} = \left(\frac{1-\theta}{1-\delta}\right)^a \tag{7}$$

It can then be shown that:

$$\mathrm{E}\Big[R \mid \Theta = \theta, \Delta = \delta, A = a\Big] = \frac{\alpha_{(\theta,\delta,a)}}{\alpha_{(\theta,\delta,a)} + \beta_{(\theta,\delta,a)}} \tag{8}$$

$$\mathrm{Var}\Big[R \mid \Theta = \theta, \Delta = \delta, A = a\Big] = \frac{\alpha_{(\theta,\delta,a)} \cdot \beta_{(\theta,\delta,a)}}{(\alpha_{(\theta,\delta,a)} + \beta_{(\theta,\delta,a)})^2 \cdot (\alpha_{(\theta,\delta,a)} + \beta_{(\theta,\delta,a)} + 1)} \tag{9}$$

Here $R \in [0, 1]$ is a bounded continuous response, $\theta \in [0, 1]$, $\delta \in [0, 1]$ and $a \in \mathbb{R}$. Similar to the logistic case, a is a discrimination parameter that controls the change rate of responses according to the ratio between ability and discrimination. In addition to supporting continuous responses, one advantage is that the item characteristic curve of the three-parameter Beta model can have a variety of shapes beyond the usual sigmoid shape (for $a > 1$), including inverse-sigmoid ($0 < a < 1$), parabolic ($a = 1$) and even identity ($a = 1, \delta = 1/2$). For the cases with $a < 0$, the Beta model can give a symmetry shape to the cases with $a > 0$ with respect to the vertical line of $r = 0.5$.

3. Estimation of IRT Parameters

The estimation of IRT parameters proceeds as follows. We assume to have a bag of L items, denoted as $\mathbb{D} = \{1, ..., L\}$, and a bag of M participants, denoted as $\mathbb{F} = \{1, ..., M\}$. With a given experiment protocol, we can collect a set of N item-participant-response tuples, denoted as $\{(d_1, f_1, r_1), ..., (d_N, f_N, r_N)\}$. Here $d_i \in \mathbb{D}$, $f_i \in \mathbb{F}$ represents a particular item / participant respectively, and r_i is the corresponding response. Denote $\boldsymbol{\theta} = \{\theta_1, ..., \theta_M\}$ as the parameter vector of abilities of all participants, $\boldsymbol{\omega} = \{\omega_1, ..., \omega_L\}$ as the vector of item parameters, and $g(r; \theta, \omega)$ as the likelihood function of a selected IRT model. The maximum likelihood estimation can then be stated as:

$$(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\theta}}) = \arg\max_{(\boldsymbol{\omega},\boldsymbol{\theta})} \sum_{i=1}^{N} \ln\, g(r_i; \omega_{d_i}, \theta_{f_i}) \tag{10}$$

Among specific applications, we can also see a Bayesian treatment [7], [8], where the aim is to calculate the full posterior of the parameters, hence capture the corresponding uncertainties. In this work, we primarily use maximum likelihood fitting in order to keep the computational cost manageable.

B. Computerised Adaptive Testing

The fundamental idea of CAT is that, rather than testing a participant with all the questions or a random sequence of questions, the participant is given questions with practical difficulty selected in real-time based on the current estimate of ability. We can then update the ability estimate with the response to the selected question and select the next question. Therefore, it is quite common to apply CAT based on a pre-trained IRT model, where we have estimated the difficulties (and other parameters) and abilities on a pool of items/participants.

As a result, most CAT approaches include three main components: an IRT model, an item selection method, and an item exposure method. As the name suggests, an item selection method determines, given the current ability estimate, how we select an item with appropriate difficulty to be the next question to estimate the ability better. Intuitively, we do not want the item to be too complicated or too simple for the actual ability, as in both cases the responses do not give much additional information about the ability. We introduce two common item selection methods in the following sections.

On the other hand, the item exposure method controls the marginal probability that a particular item is selected for the participant. The motivation is that we do not want a small number of questions to be exposed to the participants often. Such high exposure can potentially leak these questions to further participants hence affect later responses. In this work, we focus on the item selection criterion and discuss item exposure methods at the end of the paper.

1. Fisher Item Information

We start with the most commonly adopted approach for item selection, which uses Fisher information [9], [10]. Given the current candidate ability θ, a fitted IRT model with the likelihood function $g(r; \omega, \theta)$, and a set of L items with parameters $\{\omega_1, ..., \omega_L\}$, the Fisher item information (FII) on the j^{th} item is then calculated as:

$$\mathrm{FII}(\theta; g, \omega_j) = \mathrm{E}_{R\sim\mathbb{P}(\omega_j,\theta)}\left[\left(\frac{\partial \ln\, g(R;\omega_j,\theta)}{\partial\theta}\right)^2\right] \tag{11}$$

$$= \int_r \left(\frac{\partial \ln\, g(r;\omega_j,\theta)}{\partial\theta}\right)^2 g(r;\omega_j,\theta)\mathrm{d}\, r \tag{12}$$

Here $\mathbb{P}(\omega_j, \theta)$ refers to the corresponding probability measure of the IRT model. The Fisher item information calculates the variance of the likelihood gradient, so that we can find the item(s) that can potentially change the likelihood function to a more considerable extent.

2. Kullback-Leibler Item Information

As illustrated above, FII only depends on the current estimate of the ability parameter θ according to the local gradient. Alternatively, one can consider calculating the information based on both the current estimate θ and a potential estimate θ_*. By considering different potential θ_*, we might obtain more global information for the item selection process. This idea motivates the KL information (KLI) [10], [11], which is constructed based on the Kullback-Leibler divergence between the IRT likelihood g with current ability θ and the one with an updated ability θ_*. The divergence on the j^{th} item with parameter ω_j is defined as:

$$\mathrm{KL}_{\omega_j}(\theta_* \,||\, \theta) = \mathrm{E}_{R\sim\mathbb{P}(\omega_j,\theta)}\left[\ln \frac{g(R;\omega_j,\theta)}{g(R;\omega_j,\theta_*)}\right] \tag{13}$$

However, during application time we do not have access to the updated parameter θ_*, and hence cannot calculate the KL-divergence directly. As a solution, we consider the potential information from the j^{th} item to be the integrated divergence around the current ability θ, given the fact that the KL divergence is non-negative:

$$\mathrm{KLI}(\theta, g, \omega_j) = \int_{\theta_*=\theta-\epsilon}^{\theta+\epsilon} \mathrm{KL}_{\omega_j}(\theta_* || \theta)\mathrm{d}\theta_* \tag{14}$$

Hence, this KL item information is an aggregated gain around the current ability estimate, hence can be used to select the item with maximal information.

As mentioned, the main difference between FII and KLI is that the former only uses the local parameter estimates while the later obtains the information globally across different parameters [11]. The main benefit of the KLI approach is that it captures the changes in the ability parameter in both directions with a targeted range. Thus, it provides a way to merge the contributions from nearby regions on the item characteristic curve. On the other hand, FII is always based on the local gradient, requiring no extra configuration, which is more suitable when the ability estimate is closer to the actual value. KLI and FII can also prefer the same selection, particularly when the IRT model quantifies the responses well and has optimised likelihood on them. Later in the experiments, we adopt both of these two approaches to investigate their effectiveness for adaptive testing in machine learning empirically.

C. Applications in Machine Learning

There has been some recent work adopting the IRT framework for machine learning model analysis [7], [12], [13]. All three apply IRT on a model-instance level, seeing a model as a participant and treating a data instance within a given dataset as an item. In [12], [13] the authors use the Logistic model and discuss the interpretation of the learnt IRT parameters, including models like the always-correct model (e.g. predicts the ground truth). The response reflects whether a model correctly predicts the target class. In [7], the authors propose the three-parameter Beta model and learn its parameter in a Bayesian setting (e.g. posterior of the parameters). As the Beta IRT model supports bounded continuous response, in [7], the authors selected the correct class's predicted probability as the response.

III. Proposed Methods

The benchmarking methods we propose in this paper require some modifications on top of existing IRT and CAT methods to apply them to the problem of model-dataset evaluation. In general, we consider the following two requirements for the IRT and CAT methods. (1) They should support modelling continuous gain/loss measures standard in machine learning. (2) The corresponding item information should be obtainable analytically or through efficient approximations. Furthermore, we discuss the preference for non-negative discrimination in the scenario of a model-dataset benchmarking.

A. Modified Logistic IRT

The first modification is on the Logistic IRT family. Due to its original application scenario, the Logistic IRT family was used to model binary responses. As introduced above, to support CAT with a continuous response, the IRT needs to model a continuous response and provide the corresponding likelihood. The original Logistic IRT works on a Bernoulli assumption and the model estimates a mean parameter in the closed interval $[0, 1]$. While in Bernoulli distribution, the mean parameter is sufficient to calculate the likelihood, we need to consider another parameterisation for the continuous case. Although the Beta-3 IRT model uses the Beta likelihood and supports continuous response by default, it would also be valuable to keep an IRT model with sigmoid shape for better comparison. To achieve this, we replace the Bernoulli assumption with a logit-normal assumption in the IRT model. We use the original logistic function to calculate the mean of the response, and add a extra parameter s as the standard deviation:

$$R \mid \Theta = \theta, \Delta = \delta, A = a, S = s \sim \textbf{Logit-normal}(\mu_{(\theta,\delta,a)}, \sigma_{(s)}) \tag{15}$$

$$\mu_{(\theta,\delta,a)} = -a \cdot (\theta - \delta) \tag{16}$$

$$\sigma_s = s \tag{17}$$

The likelihood is then given as:

$$p(r \mid \theta, \delta, a, s) = \frac{1}{\sqrt{2\pi s^2}} \frac{1}{r \cdot (1-r)} \exp\left(- \frac{\left(\ln(\frac{1-r}{r}) + a \cdot (\theta - \delta)\right)^2}{2s^2} \right) \tag{18}$$

However, as the logit transform is not linear, the expectation (mean) and variance don't have closed forms:

$$\mathrm{E}[R \mid \theta, \delta, a, s] = \int_r p(r \mid \theta, \delta, a, s) \frac{1}{1 + \exp(r_i)} \mathrm{d}\, r \tag{19}$$

$$\mathrm{Var}[R \mid \theta, \delta, a, s] = \int_r p(r \mid \theta, \delta, a, s) \Big(\frac{1}{1 + \exp(r)} - \mathrm{E}[R \mid \theta, \delta, a, s] \Big) \mathrm{d}\, r \tag{20}$$

As both integrations involve the probability density function, the most straightforward solution here is to consider Monte-Carlo numeric integration (e.g., importance sampling):

$$\mathrm{E}[R \mid \theta, \delta, a, s] = \frac{1}{Q} \sum_{i=1}^{Q} \frac{1}{1 + \exp(r_i)} \tag{21}$$

$$\mathrm{Var}[R \mid \theta, \delta, a, s] = \frac{1}{Q} \sum_{i=1}^{Q} \Big(\frac{1}{1 + \exp(r)} - \mathrm{E}[R \mid \theta, \delta, a, s] \Big) \tag{22}$$

$$r_i \sim \textbf{Normal}(-a \cdot (\theta - \delta), s) \tag{23}$$

Here Q is the number of samples used in the calculation. In general, the approximation will be more accurate when using a larger Q. Although there is also a certain analytic approximation for the expected value (e.g. the probit approximation), we keep the sampling approach as it is also required to calculate item information as discussed later. With these modifications, the IRT model and corresponding CAT approaches can work with any bounded continuous response. While other possible extensions support continuous response [14], [15], we experimented particularly with the logistic and Beta-3 models given their close connection.

B. Approximate Item Information

The second modification also aims to incorporate continuous responses. While using binary responses, both Fisher item information and KL item information can be derived analytically [11]. Such closed forms generally are no longer possible when switching to IRT models with continuous response. However, as the integration in both Fisher item information and part of KL item information calculates an expectation over a density function, we can approximate them again with Monte-Carlo sampling. For FII, the approximation is given as:

$$\mathrm{FII}(\theta; g, \omega_j) = \frac{1}{Q} \sum_{i=1}^{Q} \Big(\frac{\partial \ln\, g(r_i; \omega_j, \theta)}{\partial \theta} \Big)^2$$

And similarly for KL divergence:

$$\mathrm{KL}_{\omega_j}(\theta_* \,||\, \theta) = \frac{1}{Q} \sum_{i=1}^{Q} \Big[\ln \frac{g(r_i; \omega_j, \theta)}{g(r_i; \omega_j, \theta_*)} \Big]$$

For both approximations we have $r_i \sim \mathbb{P}(\omega_j, \theta)$ to be random samples from the corresponding distribution.

While the calculation of FII is done with this single step, KLI still requires a further approximation to solve the second integration around θ, where we can consider a simple trapezoidal rule given ϵ is relatively small and $\mathrm{KL}_{\omega j}(\theta||\theta) = 0$:

$$\begin{aligned} \mathrm{KLI}(\theta, g, \omega_j) &= \frac{\epsilon}{2} \Big(\mathrm{KL}_{\omega_j}(\theta - \epsilon || \theta) + \mathrm{KL}_{\omega_j}(\theta || \theta) \Big) + \\ &\quad \frac{\epsilon}{2} \Big(\mathrm{KL}_{\omega_j}(\theta - \epsilon || \theta) + \mathrm{KL}_{\omega_j}(\theta || \theta) \Big) \\ &= \frac{\epsilon}{2} \Big(\mathrm{KL}_{\omega_j}(\theta - \epsilon || \theta) + \mathrm{KL}_{\omega_j}(\theta + \epsilon || \theta) \Big) \end{aligned}$$

With these approximation approaches, both item information quantities can be calculated efficiently, which is relevant as item information needs to be calculated for every candidate dataset at every step of the adaptive testing process.

C. The Constraint of Non-negative Discrimination

For typical IRT models, positive discrimination indicates the item has better average responses from candidates with higher ability estimates. In contrast, items with negative discrimination can be seen as tricky ones that cause stronger candidates to be more likely to give the wrong response than lower-ability candidates. In [13], the authors discuss the interpretation of negative discrimination in machine learning with each data instance being an item. One of their observations is that negative discrimination is often observed on instances within the regions where their opposite label dominates. A similar discussion can also be found in [7] with the Beta-3 IRT model. In this setting, the correct response from a model (candidate) when facing a data instance (item) is the instance's correct label. Assume we have a bag of instances with a Bayes optimal probability of 0.9 to be a positive class, and we can then conclude that models with a higher ability estimate should be more likely to give the correct response (positive). However, as there is still a probability of 0.1 for an instance to be negative, an optimistic prediction from a good model becomes the wrong response for these instances. It is clear that negative discrimination indeed describes the situation for these minority instances, and having negative discrimination parameters is essential for the IRT model to fit the responses correctly.

We now switch to the dataset configuration addressed in this paper, where each participant is still a model, but each item is changed to be a particular dataset. We consider a response to be a single performance measurement obtained via fitting the model on a random training fold and measuring the model with the remainder of the dataset. We assume all performance metrics to be calculated as gain measures so that a higher measurement indicates a better response for the IRT models. Therefore, if a model has a stronger ability, we expect it to have a higher averaged performance on most of the datasets, meaning it statistically fits well with a variety of training sets (i.e., it can capture a large function space) and also generalises to unseen test sets (i.e., no over-fitting). The question is then if we can design a dataset so that stronger models tend to have a lower (expected) performance, which is the requirement for negative discrimination to occur. The first possibility to have a averaged lower performance on a given dataset is that the dataset is hard to separate, that is, there is little pattern to be learnt from any part of the dataset. However, for such a dataset we expect most models to perform similarly as the labels are not dependent on the features, indicating a 0 discrimination is more suitable than negative values. The second possibility for a model to perform poorly on a dataset is that the model over-fits the training set. While this can happen with a particular combination of the training set and test set, it is less likely to occur when considering the averaged performance from a large number of random training sets and test sets. Furthermore, as discussed above, a model needs to be robust against over-fitting on most datasets to be estimated with higher ability. Therefore, it does not appear realistic to postulate that

a specific dataset can cause more robust models to be more vulnerable to over-fitting.

In accordance with this discussion, in this paper we assume the discrimination parameter to be non-negative. In practice, we can achieve this either via constrained optimisation during the estimation of IRT parameters, or directly by estimating the logarithm of the discrimination parameters via unconstrained approaches. We adopt the latter in our implementation, within a stochastic gradient descent and automatic differentiation framework. Alternatively, one can also do it the Bayesian way, which assumes a prior distribution that makes positive discrimination more likely. However, as we only consider the maximum likelihood case in this paper, we leave this option as future work.

IV. Empirical Evaluation

This section experimentally investigates the performance of the IRT and CAT-based benchmarking methods introduced in this paper. We assess their performance with the following two experiments.

1. To compare different IRT models, we evaluate their performance to make inferences over unseen responses (several standard machine learning evaluation measures).
2. To assess the utility of the CAT-based method, we examine the efficiency of different item selection methods, in terms of the amount of computation costs it saves from testing the entire collection of datasets.

We first introduce the experimental setup. For the first IRT experiment, we compare the inference errors on responses using a standard train-test split. Regarding the CAT methods, we compare the final root mean squared error (RMSE) on the inferred response and the convergence speed, given the test sequences and the validation sets.

A. Setup

As response targets, we selected six evaluation measures commonly used in predictive machine learning: (1) multi-class accuracy, (2) Brier score, (3) log-loss, (4) weighted averaged binary accuracy, (5) weighted averaged binary AUC, and (6) weighted averaged binary F-measure. All these losses are bounded within $[0, 1]$ except the log-loss, which requires post-processing. We rescaled the averaged log-loss to the range of $[0, 1]$ with the exponential operation, which is an invertible calculation and ensures the final density function is valid on both scales. Furthermore, we use the negative value of Brier score and log-loss to fit the IRT models, so that they become gain measures (i.e., larger values indicate better results), in line with the other evaluation measures.

We select a set of datasets and model classes (described below) and run each model-dataset combination with an even random train-test split ten times. We use these results to train both Beta-3 and Logistic IRT models.

We use the 165 datasets provided by PMLB [16], which is a pre-processed collection of UCI datasets on various classification tasks. For computational efficiency, for all the datasets with more than $10,000$ instances, we sample it down to $10,000$ instances while approximately keeping the marginal distribution of the target variable.

We selected 9 model classes from the sklearn package: (1) multi-layer perceptron (MLP), (2) K nearest neighbours (KNN), (3) support vector machine (SVM), (4) pseudo Gaussian process (GP), (5) decision tree (TREE), (6) random forest (RF), (7) Ada boosting (ADA), (8) naive Bayes (NB), and (9) logistic regression (LR).

We selected eight different parameter settings for each model class to form different model instances, resulting in a total number of 72 models. For instance, for the MLP we choose a range of hidden units in a two-layer setting. Regarding the GP, here we call it pseudo models as the sklearn implementation does not support sparse covariance matrix hence can not scale to large datasets. We hence perform a simple random sampling on the training set. We first randomly select one data point for each class, then further sample random data points from the entire training set.

B. Evaluation of IRT Approaches

The first experiment we performed was to investigate whether the IRT models can accurately model and infer the performance measurements. As introduced in section II and III, the IRT models can estimate a distribution of the responses given each dataset and model combination. Therefore, we can evaluate each IRT model's goodness by evaluating the quality of these estimated distributions. Here we consider evaluating each distribution's mean, which is the estimated average performance measurement between the corresponding model and dataset. In general, we expect the estimated average response from a good IRT model to be close to some previously unseen measurements during future tests.

For this purpose, we perform ten times random split experiments on the collected responses from the 165 datasets and 72 models. We then divided the collected responses into a training set and a test set. We use the training set to estimate the IRT models' parameters, and the test set to verify the expected responses from each IRT model. Given the continuous responses, we use the root mean squared errors (RMSE) as the metric to evaluate the IRT models. Table I gives the results; notice here the RMSE is calculated after re-scaling all the evaluation metrics (e.g. the log-loss is re-scaled to $[0, 1]$). Additionally, the raw global mean and standard deviations of all the evaluation measures are also given. As the results show, both IRT models infer the evaluation measures well, with most RMSE values smaller than 0.05, which is considerably lower than the population standard deviation. For most evaluation measures, Logistic IRT and Beta-3 IRT perform similarly. Fig. 1 shows the item characteristic curves of both IRT approaches on the chess dataset and labour dataset with multi-class accuracy as an evaluation metric. Both IRT approaches tend to assign the same order to the ability parameters, as we can observe a similar pattern with the responses marked by the black points. Although in the bottom figures the two item characteristic curves are quite different from each other around the edges of the figures, it is noticeable that the curves behave similarly around the region with dense black points. This observation can help illustrate how the two different IRT approaches share similar RMSE values in the final results.

TABLE I. The Inference Errors (RMSE) of Both IRT Models on Different Evaluation Measures (top Two Rows), and the Global Mean and Standard Deviation of the Original Evaluation Measures (Bottom Two Rows)

	Acc	BS	LL	W-Acc	W-AUC	W-F1
Logistic	0.01349	0.00555	0.04379	0.00922	0.01763	0.05150
Beta-3	0.01569	0.00625	0.04151	0.02367	0.01498	0.05060
Global Mean	0.71878	0.79850	0.42725	0.79339	0.76487	0.58888
Global Std	0.21352	0.12954	0.29986	0.16509	0.18878	0.34585

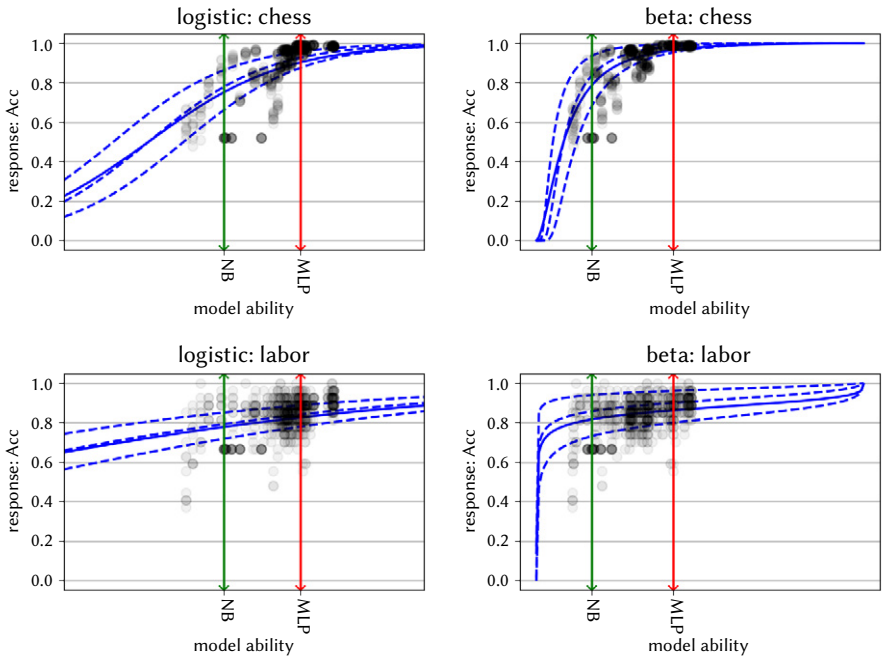


Fig. 1. The estimated item characteristic curves of the Logistic IRT and Beta-3 IRT models on two datasets evaluated with multi-class accuracy. The blue line indicates the mean value of responses, and the three dashed lines mark the 25%, 50% and 75% percentile of the responses. The grey dots mark the collected responses of the XXX models. The green line and the red line indicate the estimated ability of a naive Bayes classifiers and a multi-layer perceptron classifier respectively.

TABLE II. The Index of the Selected Dataset At Some Locations of the Adaptive Testing Sequence (Gradient Boosting Classifier and Multi-class Accuracy). For Example, the Logistic-FII Approach Select the 93 Dataset for the First Test, and Proceed With the 114 Dataset for the Second Test, This Selection Process Continues Till All the Datasets Have Been Tested

	Test 1	Test 2	Test 3	Test 10	Test 50	Test 100	Test 150
Logistic (Fisher)	93	114	147	135	79	140	42
Logistic (KL)	114	93	147	135	79	148	70
Logistic (Random)	20	80	89	39	27	164	61
Beta-3 (Fisher)	107	115	156	55	113	116	45
Beta-3 (KL)	107	1	43	124	113	23	45
Beta-3 (Random)	116	38	34	87	15	3	92

C. Evaluation of IRT and CAT Pairs

For the second experiment, we use different IRT and item selection approaches to test a set of different classifiers. We selected five classifiers from the sklearn package with their default settings as the candidate model: (1) gradient boosting classifier (GBC), (2) multi-layer perceptron (MLP), (3) support vector machine (SVM), (4) random forest (RF), and (5) logistic regression (LR). Here the GBC classifier is not used during the fitting of IRT models, while other classifiers have different parameter settings compared to those in the IRT model estimation process. While this group of classifiers doesn't cover all the model types as seen in the previous experiment, we select them due to their differences (e.g. linear v.s. nonlinear, ensemble v.s. standalone).

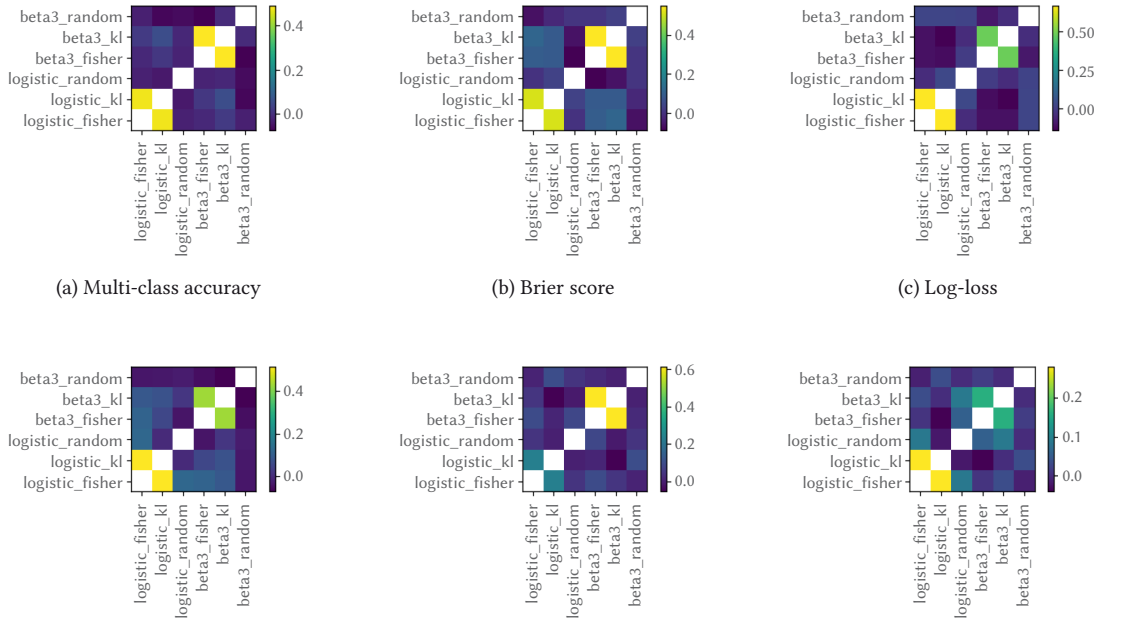
We run these models with all the datasets ten times using the same setting as in the previous experiment. The performance measurements are collected and used as a validation set. During adaptive testing, each time we update the model ability, we use the trained IRT to infer the expected value of responses (performance measures). We then calculate the corresponding RMSE the validation set to evaluate different IRT and CAT approaches. In principle, a better IRT-CAT combination should eventually have a lower RMSE and a faster convergence speed to the final RMSE.

We start by assuming the candidate model has average ability, then keep testing the model and updating its ability until we have tested all the datasets. We record the selected dataset at each test step, and the RMSE calculated using the validation set. Here we first analyse the results on the gradient boosting classifier (GBC) with multi-class accuracy as an example. Table II and Table III show the indices of the selected dataset and the RMSE on the averaged response on some locations of the testing sequences, respectively. It can be seen that both item information approaches pick similar datasets around the beginning of the sequence. This result can be observed with the logistic case, where test 1, 2, 3 , and 10 all select the same combination of datasets, and the order only differs between the first two tests. As discussed, FII and KLI can give similar selections when the IRT approach models the responses well. Hence our observation here agrees with the low RMSE as shown in the previous experiment. To further verify this observation, we calculate the pair-wise correlation with Kendall's Tau among the entire testing sequences for the GBC with all six performance metrics, and the results are shown in Fig. 2. The correlation between the two item information quantities with the same IRT approach can be clearly observed for the entire test sequence of 165 datasets across all metrics.

We can observe a similar pattern on the RMSE sequence decay on averaged responses. Both FII and KLI led to quite similar RMSE values around the beginning of the sequence with the first 3 tests. While the two item information approaches have lower RMSE values at the

TABLE III. Root Mean Squared Error of the Expected Response At Some Locations of the Adaptive Testing Sequence (Gradient Boosting Classifier and Multi-class Accuracy)

	Initial	Test 1	Test 2	Test 3	Test 10	Test 50	Test 100	Test 150
Logistic (Fisher)	0.12102	0.08409	0.08411	0.08354	0.08368	0.08466	0.08438	0.08375
Logistic (KL)	0.12103	0.12099	0.08525	0.08462	0.08393	0.08480	0.08452	0.08380
Logistic (Random)	0.12103	0.09959	0.09138	0.09038	0.08542	0.08380	0.08362	0.08362
Beta-3 (Fisher)	0.10100	0.07818	0.07828	0.07821	0.07817	0.07814	0.07809	0.07811
Beta-3 (KL)	0.10100	0.07818	0.07808	0.07803	0.07832	0.07855	0.07841	0.07843
Beta-3 (Random)	0.10100	0.08526	0.08236	0.08264	0.08444	0.07814	0.07834	0.07811



(a) Multi-class accuracy (b) Brier score (c) Log-loss

(d) Weighted averaged binary accuracy (e) Weighted averaged binary AUC (f) Weighted averaged binary F-measure

Fig. 2. Kendall's Tau between the adaptive testing sequences of the gradient boosting classifier on all six evaluation metrics, a brighter yellow colour indicates a stronger correlation and a deeper blue colour corresponds to a weaker correlation.

early stage, the random selection also performs remarkably well and gets to a relatively low RMSE value at test 10 in the logistic case. All RMSE values are very similar within each IRT approach from test 50 onwards.

To quantitatively adaptive testing sequence on the RMSE, we now examine the number of tests required before the inference error converges to a certain level according to the end of the test sequence. To calculate this number, given a test sequence of inference errors (RMSE), denoted as $(r_1, ..., r_T)$, we first select the final inference error r_T at the end of the entire sequence and construct a consequence region of $|r_T - \epsilon|$. In this experiments we set ϵ to be 0.05 of the minimum RMSE in the sequence, then we can obtain the convergence point c so that for $\forall i \geq c$ we have $|r_i - \epsilon| \leq \epsilon$.

Table IV lists the convergence points for the five different candidate classifiers and six evaluation measures. It is noteworthy that there are various cases where it took only 1 or 2 tests (out of 165 datasets) before the testing sequence reaches the convergence point. While both Fisher item information and KL item information require a smaller number of tests than random selection, we can still observe a few cases where random selection gives the fastest convergence. We hypothesise the randomness causes this within the model testing procedure. As the evaluation measurements can differ even on the same combination of dataset and model configuration, specific measurements cause a high bias on the item selection information, which leaves random selection a suitable backup choice. To obtain the best efficiency of adaptive testing, it is therefore suggested to calculate both item information and perform random selection while adaptive testing is required, so that the fastest convergence can always be achieved.

V. Conclusion

This paper introduced a novel framework to effectively benchmark a set of predictive models on an extensive collection of datasets. Instead of performing experiments on all possible datasets, we propose to model the similarity and dependency among different models and datasets to infer their experimental results without actually running all train-test cycles. Furthermore, we adopt the adaptive testing technique and uses the uncertainties on the unknown measurements to automatically decide a testing sequence for any unseen model based on the previous observations.

We performed a range of experiments, from which some general conclusions can be drawn. First of all, the choice of the IRT model plays an essential role in the benchmark. A suitable IRT model can indeed lead to better inference on the test results, without spending much effort on further testing. Which IRT model is most suited for which machine learning evaluation metric warrants further research.

TABLE IV. Number of Tests Before the Inferred Response Error (RMSE) Converges to the Overall Test Results

	GBC	MLP	SVM	RF	LR
Logistic (Fisher)	2	15	36	2	4
Logistic (KL)	3	17	36	3	4
Logistic (Random)	8	3	10	8	5
Beta-3 (Fisher)	2	2	5	2	15
Beta-3 (KL)	2	2	5	2	16
Beta-3 (Random)	13	16	27	5	8
Median	2.5	9	18.5	2.5	6.5

(a) Accuracy, Median: 5

	GBC	MLP	SVM	RF	LR
Logistic (Fisher)	2	1	10	88	18
Logistic (KL)	2	1	5	42	2
Logistic (Random)	7	1	13	4	3
Beta-3 (Fisher)	2	1	5	2	24
Beta-3 (KL)	2	1	2	2	19
Beta-3 (Random)	12	3	2	9	147
Median	2	1	5	6.5	18.5

(b) Brier Score, Median: 3

	GBC	MLP	SVM	RF	LR
Logistic (Fisher)	61	1	80	2	2
Logistic (KL)	21	1	54	3	30
Logistic (Random)	2	45	32	21	17
Beta-3 (Fisher)	13	97	31	147	9
Beta-3 (KL)	46	4	31	151	13
Beta-3 (Random)	43	125	2	97	6
Median	32	24.5	31.5	59	11

(c) Log loss, Median: 25.5

	GBC	MLP	SVM	RF	LR
Logistic (Fisher)	2	1	7	2	1
Logistic (KL)	2	1	4	2	1
Logistic (Random)	2	7	4	5	1
Beta-3 (Fisher)	2	1	1	2	7
Beta-3 (KL)	2	1	1	2	7
Beta-3 (Random)	2	12	18	2	14
Median	2	1	4	2	4

(d) Weighted averaged binary accuracy, Median: 2

	GBC	MLP	SVM	RF	LR
Logistic (Fisher)	2	2	3	6	2
Logistic (KL)	2	2	3	4	2
Logistic (Random)	2	7	17	2	3
Beta-3 (Fisher)	15	2	1	2	2
Beta-3 (KL)	15	2	4	10	2
Beta-3 (Random)	1	5	6	24	18
Median	2	2	3.5	5	2

(e) Weighted averaged binary AUC, Median: 2.5

	GBC	MLP	SVM	RF	LR
Logistic (Fisher)	3	1	1	24	4
Logistic (KL)	53	4	1	2	5
Logistic (Random)	11	3	7	2	3
Beta-3 (Fisher)	119	2	167	26	8
Beta-3 (KL)	2	167	88	8	116
Beta-3 (Random)	7	166	167	51	120
Median	9	3.5	47.5	16	6.5

(f) Weighted averaged binary F-measure, Median: 7.5

Secondly, we have demonstrated that adaptive testing can effectively reduce the total number of experiments. For most evaluation measures, we can observe a significant decay on the inference error with a small number of tests, leading to a significant reduction of model benchmarking costs.

One of the most promising directions for future research is to incorporate this adaptive testing framework into the development cycle of machine learning approaches. Modern data-driven approaches usually require many train-test runs to optimise their configuration and hyper-parameters. Although a range of approaches have been proposed in auto-ML and neural architecture search [17], most approaches still require to perform large-scale experiments on the given datasets to obtain the search points. With the assistance of adaptive testing, we can further attempt to reduce such search costs by selecting the most promising datasets. Another direction is to look beyond predictive machine learning tasks. Recent work has made significant progress on non-predictive tasks such as random data generation and neural-based density estimation. Both areas can potentially benefit from adaptive testing considering their significant computational demands during training. Item exposure control [18] is also worth further consideration in the benchmarking process, which allows us to further control the rate that a particular dataset is used.

References

[1] M. Hutson, "Artificial intelligence faces reproducibility crisis," 2018. [Online]. Available: https://science.sci-encemag.org/content/359/6377/725, doi: 10.1126/sci-ence.359.6377.725.

[2] J. Vanschoren, J. N. Van Rijn, B. Bischl, L. Torgo, "OpenML: networked science in machine learning," *ACM SIGKDD Explorations Newsletter*, vol. 15, no. 2, pp. 49–60, 2014.

[3] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, V. McCauley, "Testing the test: Item response curves and test quality," *American Journal of Physics*, vol. 74, no. 5, pp. 449–453, 2006.

[4] W. J. van der Linden, R. K. Hambleton, *Handbook of modern item response theory*. Springer Science & Business Media, 2013.

[5] B. F. Green, R. D. Bock, L. G. Humphreys, R. L. Linn, M. D. Reckase, "Technical guidelines for assessing computerized adaptive tests," *Journal of Educational Measurement*, vol. 21, no. 4, pp. 347–360, 1984.

[6] D. J. Weiss, G. G. Kingsbury, "Application of computerized adaptive testing to educational problems," *Journal of Educa-tional Measurement*, vol. 21, no. 4, pp. 361–375, 1984.

[7] Y. Chen, T. S. Filho, R. B. Prudencio, T. Diethe, P. Flach, "β3-IRT: A new item response model and its applications," in *AISTATS 2019*, vol. 89 of *Proceedings of Machine Learning Research*, 2019, pp. 1013–1021.

[8] B. P. Veldkamp, M. Matteucci, "Bayesian computerized adaptive testing," *Ensaio: Avaliação e Políticas Públicas em Educação*, vol. 21, no. 78, pp. 57–82, 2013.

[9] R. R. Meijer, M. L. Nering, "Computerized adaptive testing: Overview and introduction," 1999. [Online]. Available: https://doi.org/10.1177/01466219922031310, doi: 10.1177/01466219922031310.

[10] H.-H. Chang, "Psychometrics behind computerized adaptive testing," *Psychometrika*, vol. 80, no. 1, pp. 1–20, 2015.

[11] H.-H. Chang, Z. Ying, "A global information approach to computerized adaptive testing," *Applied Psychological Measurement*, vol. 20, no. 3, pp. 213–229, 1996.

[12] R. B. Prudêncio, J. Hernández-Orallo, A. Martínez-Usó, "Analysis of instance hardness in machine learning using item response theory," in *Second International Workshop on Learning over Multiple Contexts in ECML 2015. Porto, Portugal, 11 September 2015*, 2015.

[13] F. Martínez-Plumed, R. B. Prudêncio, A. Martínez-Usó, J. Hernández-Orallo, "Making sense of item response theory in machine learning," in *Proceedings of the Twentysecond European Conference on Artificial Intelligence*, 2016, pp. 1140–1148, IOS Press.

[14] K. SHOJIMA, "A noniterative item parameter solution in each em cycle of the continuous response model," *Educational technology research*, vol. 28, no. 1-2, pp. 11–22, 2005.

[15] F. Samejima, "Graded response model," in *Handbook of modern item response theory*, Springer, 1997, pp. 85–100.

[16] R. S. Olson, W. La Cava, P. Orzechowski, R. J. Urbanowicz, J. H. Moore, "Pmlb: a large benchmark suite for ma-chine learning evaluation and comparison," *BioData mining*, vol. 10, no. 1, p. 36, 2017.

[17] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, F. Hutter, "Efficient and robust automated machine learning," in *Advances in neural information processing systems*, 2015, pp. 2962–2970.

[18] J. Sympson, R. Hetter, "Controlling item-exposure rates in computerized adaptive testing," pp. 973–977, 1985.

Hao Song

HS obtained his PhD at the University of Bristol in 2017, and is currently a postdoctoral researcher at the University of Bristol. His research interests include quantifying different types of uncertainties within the machine learning pipeline, particularly for different probabilistic out-puts and corresponding evaluation metrics.

Peter Flach

PF is Professor of Artificial Intelligence at the University of Bristol and has over 30 years experience in machine learning and data mining, with particular expertise in mining highly-structured data and the evaluation and improvement of machine learning models. He was PC co-chair of KDD'09 and ECML-PKDD'12 and has edited and authored several books, including *Machine Learning: the Art and Science of Algorithms that Make Sense of Data*.