Educational Question Mining At Scale: Prediction, Analysis and Personalization

Zichao Wang,[1][*] **Sebastian Tschiatschek,**[2][†] **Simon Woodhead,**[3] **José Miguel Hernández-Lobato,**[5][†]
Simon Peyton Jones,[4] **Richard G. Baraniuk,**[1] **Cheng Zhang**[4]

[1]Rice University   [2]University of Vienna   [3]Eedi   [4]Microsoft Research Cambridge   [5]University of Cambridge

{jzwang, richb}@rice.edu   sebastian.tschiatschek@univie.ac.at   jmh233@cam.ac.uk
simon.woodhead@eedi.co.uk   {simonpj, Cheng.Zhang}@microsoft.com

Abstract

Online education platforms enable teachers to share a large number of educational resources such as questions to form exercises and quizzes for students. With large volumes of available questions, it is important to have an automated way to quantify their properties and intelligently select them for students, enabling effective and personalized learning experiences. In this work, we propose a framework for mining insights from educational questions at scale. We utilize the state-of-the-art Bayesian deep learning method, in particular partial variational auto-encoders (p-VAE), to analyze real students' answers to a large collection of questions. Based on p-VAE, we propose two novel metrics that quantify question quality and difficulty, respectively, and a personalized strategy to adaptively select questions for students. We apply our proposed framework to a real-world dataset with tens of thousands of questions and tens of millions of answers from an online education platform. Our framework not only demonstrates promising results in terms of statistical metrics but also obtains highly consistent results with domain experts' evaluation.

Introduction

Online education platforms are transforming education by democratizing access to high-quality educational resources and personalizing learning experiences. A central instrument in today's online education scenarios is assessment questions (referred to as "questions" henceforth), which help teachers evaluate the students' abilities and help students reinforce the knowledge they are learning (Karpicke 2012; Karpicke and Roediger 2008; Connor-Greene 2000; Kovacs 2016; Koedinger et al. 2015). Questions are particularly important in *online* education, because teachers have more limited interactions with students; students answer records to questions serve as one of the few ways for teachers to interact with and understand their students (Twigg 2003). With the world reeling from the impact of the Covid-19 pandemic, there has been a rapid and massive increase in the number of students learning online (Bao 2020; Sandars et al. 2020; Ve-

rawardina et al. 2020). Questions as assessment and learning instruments have thus become even more prominent.

A key challenge to best utilize these questions is how to choose the most suitable ones for students. Only high quality and suitably difficult questions are beneficial for learning (Blanchette 2001); as a result, understanding these two properties of questions help guide the choice of questions. While in traditional classroom settings, manually examining and selecting questions to attend to the learning status of each individual student remains the best practice, this labor- and time-intensive procedure clearly does not scale to large-scale online education scenarios in which the number of questions and students can be massive. We thus desire an efficient tool for question analysis at scale and capable of computing question analytics including quality and difficulty and automatically selecting questions for students. The analytics will serve as side information that helps teachers and students select appropriate questions for their pedagogical and educational needs. The automatic question selection process enables personalized learning and adaptive testing experience when the number of students greatly exceeds teachers' capacities (Chang, Ying et al. 2009; Yan, Von Davier, and Lewis 2016). Both analytics and personalization are important to realize the values of massive questions in online education scenarios.

To address the above challenges, we aim to develop an AI solution for large-scale online educational question mining, providing both insights for question quality and difficulty and strategies for choosing questions for each student. This task involves 3 challenges. First, the number of questions, students, and answer records is extremely large and processing such data may be computationally expensive. Second, online educational data is highly incomplete because each student can only answer a small fraction of all available questions. Third, we need to design quantitative metrics and strategies to accurately identify question quality and to adaptively choose questions for each student. Overall, we need a solution that is efficient, handles highly sparse data, and automatically acquires educationally meaningful and actionable insights about questions.

Contributions

In this work, we collect a large real-world online educational dataset in the form of students' answers to multiple-

---

choice questions and develop a machine learning framework to quantify question quality and difficulty and provide personalized question selection strategy. We briefly summarize our framework below.

- We leverage the partial variational auto-encoder (p-VAE) (Ma et al. 2018, 2019) to efficiently handle the highly sparse educational data at a large scale. p-VAE models existing students' answers and predicts the potential answer to unseen questions in a probabilistic manner.

- We propose a novel information-theoretic metric to quantify the quality of each question based on p-VAE. We also define a metric to quantify question difficulty.

- We propose a novel information-theoretic strategy to sequentially select questions for each student. Our strategy chooses a sequence of questions which best identifies the student's learning status.

- We experimentally validate our framework on a new, large educational dataset and demonstrate state-of-the-art performances of our framework over various baselines.

Dataset

We analyze data from Eedi,[1] a renowned real-world online education platform used by over 100,000 thousand students and 25,000 teachers in over 16,000 schools. Eedi offers crowd-sourced, multiple-choice diagnostic questions to students from primary to high school (roughly between 7 and 18 years old). Each question has 4 answer choices and only one of them is correct. Currently, the platform focuses mainly on math questions. Figure 1 shows an example question from the platform. We use data collected from the 2018 – 2019 school year.

We organize the data in a matrix form where each row represents a student and each column represents a question. Each entry contains a number that represents whether the student has answered a question correctly (i.e., 0 represents the wrong answer and 1 represents the correct answer). Each student has only answered a tiny fraction of all questions and hence the matrix is extremely sparse. We thus removed questions that contain less than 50 answers and students who have answered less than 50 questions. Besides, when a student has multiple answer records to the same question, we keep the latest answer record.

The above preprocessing steps lead to a final data matrix that consists of more than 17 million students' answer records with 123,889 students (rows) and 27,613 questions (columns), making it one of the largest educational datasets to date compared to a number of existing ones (Heffernan and Heffernan 2014; Stamper et al. 2010b,a). Additionally, each question is linked to one or more topics that describe the skills (VanLehn 1988; Chrysafiadi and Virvou 2013) that the question intends to assess. We have open-sourced this dataset (Wang et al. 2020) and it is available at https://eedi.com/projects/neurips-education-challenge.
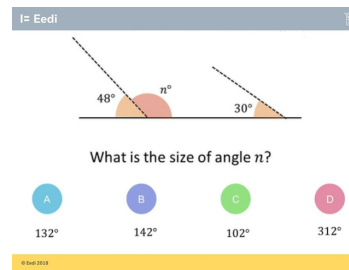
Figure 1: An example question from our new dataset.

Method

The extreme sparsity and massive quantity of our dataset brings challenges to analyze both the question quality because each student only answers a small fraction of potentially non-overlapping questions. To gain insights into such real-world educational data, we first need a model that predicts the missing data with uncertainty estimation. The missing data in our case is students' answers to unseen questions. With such a model, we can design different metrics to quantify question quality and difficulty.

The first step is formulated as the following probabilistic missing data imputation (matrix completion) problem. We have a data matrix X of size N by M, where N is the total number of students and M is the total number of questions. Each entry x_{ij} is binary which indicates whether student i has answered question j correctly.[2] The data matrix is only partially observed; we denote the observed part of the data matrix as X_O. Then, we would then like to accurately predict the missing entries in a *probabilistic* manner which enables the design of various metrics for question quality, difficulty, and selection strategy. Thus, we use the partial variational auto-encoder (p-VAE) (Ma et al. 2018), which is the state-of-the-art method for the above imputation task.

The second step is to quantify question quality and difficulty and adaptively select questions for students. Specifically, we define a difficulty measure using the full data matrix completed by p-VAE. We quantify question quality by measuring the value of information that each question carries using an information theoretical metric. Using Similar ideas, we sequentially select questions for each student based on a notion of information gain. Such "information" computation is made possible and efficient by our utilization of the p-VAE in the first step above. We present these two steps in detail in the remainder of this section.

Partial Variational Auto-encoder (p-VAE) for Student Answer Prediction

p-VAE (Ma et al. 2019) is a deep latent variable model that extends traditional VAEs (Kingma and Welling 2013; Rezende, Mohamed, and Wierstra 2014; Zhang et al. 2019) to handle missing data as in such education applications. VAEs assume that the responses x_i of student i is generated
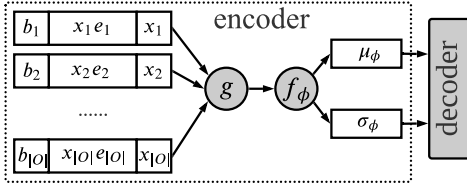
Figure 2: Illustration of the p-VAE model architecture.

from a latent variable $\mathbf{z}_i$:

$$p_\theta(\boldsymbol{X}) = \prod_{i=1}^N p_\theta(\boldsymbol{x}_i) = \prod_{i=1}^N \int p_\theta(\boldsymbol{x}_i|\mathbf{z}_i)\, p_\theta(\mathbf{z}_i)\, d\mathbf{z}_i$$

$$= \prod_{i=1}^N \int \prod_{j=1}^M p_\theta(x_{ij}|\mathbf{z}_i)\, p_\theta(\mathbf{z}_i)\, d\mathbf{z}_i\,,$$

where x_{ij} is the i-th student's answer to the j-th question. We use a deep neural network for the generative model $p_\theta(x_{ij}|\mathbf{z}_i)$ because of its expressive power. Of course, $\boldsymbol{x}_i$ contains missing entries because each student i only answers a small fraction of all questions. Unfortunately, VAEs can only model fully observed data. To model partially observed data, p-VAE extends traditional VAEs by exploiting the fact that, given $\boldsymbol{z}_i$, $\boldsymbol{x}_i$ is fully factorized. Thus, the unobserved data entries can be predicted given the inferred $\mathbf{z}_i$'s. Concretely, p-VAE optimizes the following *partial* evidence lower bound (ELBO):

$$\log p(\boldsymbol{X}_O) \geq \log p(\boldsymbol{X}_O) - D_{\mathrm{KL}}(q(\mathbf{z}|\boldsymbol{X}_O)\|p(\mathbf{z}|\boldsymbol{X}_O))$$

$$= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\boldsymbol{X}_O)}[\log p(\boldsymbol{X}_O|\mathbf{z}) + \log p(\mathbf{z}) - \log q(\mathbf{z}|\boldsymbol{X}_O)]$$

$$\equiv \mathcal{L}_{\mathrm{partial}}\,,$$

which is in the same form as the ELBO for VAE but only over the observed part of the data. Because $\boldsymbol{X}$ is binary, we use Bernoulli distribution as the likelihood function. However, we can also choose to use students' actual answers (A, B, C, or D) as $\boldsymbol{X}$ where the likelihood function becomes categorical. Investigation of categorical data format is left for future work.

The challenge is to approximate the posterior of $\mathbf{z}_i$'s using a partial observed data vector. p-VAE uses a set-based inference network $q_\phi(\mathbf{z}_i|\boldsymbol{x}_{O_i})$, where $\boldsymbol{x}_{O_i}$ is the observed subset of answers for student i (Zaheer et al. 2017; Qi et al. 2017). $q_\phi(\mathbf{z}_i|\boldsymbol{x}_{Oi})$ is assumed to be Gaussian; Concretely, the mean and variance of the posterior of the latent variable is inferenced as

$$[\mu_\phi(\boldsymbol{x}_O), \sigma_\phi(\boldsymbol{x}_O)] = f_\phi(g(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_{ij},\ldots,\boldsymbol{s}_{|O|})),\quad (1)$$

where we have dropped the student index i for notation simplicity. $\boldsymbol{s}_{ij}$ is the observed answer value augmented by its location embedding, which is learned; $g(\cdot)$ is a permutation invariant transformation such as summation which outputs a fix sized vector; and $f_\phi : \mathbb{R}^M \to \mathbb{R}^K$ is a regular feedforward neural network. In this paper, we set $\boldsymbol{s}_{ij} = [x_{ij}, x_{ij}\boldsymbol{e}_j, b_j]$ where Figure 2 illustrates the network architecture of p-VAE.

Note that, in p-VAE, some parameters have natural interpretations. For example, the per-question parameters $[\mathbf{e}_j, b_j]$

can be collectively interpreted as a *question embedding* for each question j. The per-student latent parameter $\mathbf{z}_i$ can be interpreted as a *student embedding* for each student i. Furthermore, p-VAE quantifies the uncertainty of the student embedding $\mathbf{z}_i$ in the form of an (approximate) posterior distribution. This enables us to define information-theoretic metrics and strategies for question difficulty, question quality, and question selection, which we describe next.

Question Difficulty Quantification

For a group of questions answered by the same group of students, the difficulty level of the questions can be quantified by the incorrect rate of all students' answers. However, for real-world online education data, every question is answered only by a small fraction of students and by different subsets of students with different educational backgrounds. Thus, directly comparing the difficulty levels of the question from observational data is not accurate because an easy question, which may be answered by only less-skilled students, may be shown to be difficult if only observational data is used. Thanks to p-VAE, we can predict whether a student can correctly answer an unseen question. We achieves this by first predicting students' responses to all unanswered questions and then defining the difficulty level of question j as $\sum_i^N \frac{p(x_{ij}=0)}{N} \approx \frac{1}{N}\sum_i^N \hat{x}_{ij}$ where $\hat{x}_{ij}$ denotes if the student has answered the question correctly. Higher value implies that more students are predicted to answer this question correctly and that this is an easier question.

Question Quality Quantification

Working closely with education experts, we found that high-quality questions are considered to be those that best differentiate student abilities. When a question is simple, almost all students will answer it correctly. When a question is badly formulated, all students will provide incorrect answers or random guesses. In any of these cases, the question neither helps the teacher gain insights about the students' abilities nor helps students learn well. Thus, high-quality questions are the ones that can differentiate the students' abilities.

We thus formulate the following information theoretic objective to quantify the quality of question j:

$$R(j) = \mathbb{E}_{x_{ij}\sim p_\theta(x_{ij})}\Big[D_{\mathrm{KL}}[p_\phi(\mathbf{z}_i|x_{ij})\,\|\,p(\mathbf{z})]\Big],\quad (2)$$

$$\approx \frac{1}{S}\sum_{i=1}^S D_{\mathrm{KL}}\left[q_\phi(\mathbf{z}|x_{ij})|p(\mathbf{z})\right]\,,\quad (3)$$

where we have used Monte Carlo integration and replaced $p_\phi(\mathbf{z}_i|x_{ij})$ with $q_\phi(\mathbf{z}_i|x_{ij})$ for practical and efficient computation (Ma et al. 2019; Gong et al. 2019). j is the question index, x_{ij} is the i-th student's answer to the j-th question, which can be either binary indicating whether the student has answered it correctly or categorical which is the student's answer choice for this question. $\mathbf{z}$ is the latent embedding of students. The i-th student ability can be determined by the student's possible performance on all questions, which can be inferred from $\mathbf{z}_i$.

This objective measures the information gain of estimating the student ability by conditioning on the answer to question j. When R is large, the question is more informative on differentiating the student ability reflected in the student embedding $\mathbf{z}$, and thus it is considered as high-quality.

Personalized Question Selection

In practical online education scenarios, it is of great interest to adaptively select a small sequence of questions for students. Appropriately choosing these questions allows effective and efficient evaluation of students' abilities at scale.

We formulate the problem of personalized question selection as a Bayesian experiment design problem in an information theoretic manner. Specifically, we are interested in selecting a sequence of questions that is most informative in revealing the student's current state of learning. Inspired by (Ma et al. 2019; Gong et al. 2019), we formulate the following selection strategy which is similar to Eq. 3:

$$R_i(j, \boldsymbol{x}_O) = \mathbb{E}_{x_j \sim p(x_j|\boldsymbol{x}_O)} D_{\mathrm{KL}}[p(\mathbf{z}_j|\boldsymbol{x}_O, x_j)||p(\mathbf{z}_j|\boldsymbol{x}_O)]. \tag{4}$$

Intuitively, this selection strategy selects a question j that provides the most information, defined in KL divergence, on the student knowledge state summarized in $\mathbf{z}_i$. This objective is different from the one used in (Ma et al. 2019) as we do not have a particular target variable. Note that we have omitted the student index subscript i in Eq. 4 for succinctness. Similar to the previous quality metric computation, we use Monte Carlo integration to approximate Eq. 4.

The above selection strategy enables selecting a sequence of personalized questions for each student in the following manner (which is different from Eq. 3 which can only select a single question). First, we initialize $\boldsymbol{x}_O = \emptyset$. Then, we compute the information reward according to Eq. 4 for each question and select question j outside the prediction targets (i.e., $x_j \notin \boldsymbol{x}_\psi$) with the maximum $\widehat{R}$ as the next one for student i. Finally, we set $\boldsymbol{x}_O \leftarrow \boldsymbol{x}_O \bigcup x_j$ and repeat the previous steps until we reach the desired number of selected questions.

Experiments

In this section, we demonstrate the applicability of our framework on the real-world educational dataset that we have introduced in the dataset section for student answer prediction, question difficulty, and quality quantification, and personalized question selection.

Student Answer Prediction

Setup We split the students (rows of the data matrix) into the train, validation, and test sets with an 80:10:10 ratio. Therefore, students that are in the test set are never seen in the training set. We train the model on the train set for 50 epochs using Adam optimizer (Kingma and Ba 2014) with a learning rate of 0.001. We train p-VAE on binary students' answer records (correct or incorrect answers). To evaluate imputation performance, we supply the trained p-VAE model a subset of the test set as input and compute the model's prediction accuracy and mean absolute error (MAE) on the rest of the test set.

Method	Accuracy ↑	Mean Abs. Err. ↓
Random	0.534	0.471
IRT	0.735	0.359
SVD	0.734	0.358
SVD++	0.737	0.352
Co-clustering	0.731	0.351
NMF	0.737	0.382
p-VAE (ours)	**0.739**	**0.343**

Table 1: Imputation performances of various methods. p-VAE remains a very strong competitor and slightly outperforms all baselines that we consider.

Results Table 1 shows the accuracy of p-VAE trained on binary answer records comparing to various baselines, including random imputation, the Rasch model (Rasch 1960), SVD, SVD++ (Koren 2008), Co-clustering (George and Merugu 2005) and Negative Matrix Factorization (NMF) (Luo et al. 2014). Note that SVD can also be interpreted as a multivariate Rasch model (Chalmers et al. 2012) which is another classic method in educational data mining. We did not compare with vanilla VAEs (Kingma and Welling 2013) because, as mentioned previously, they cannot handle an incomplete input data matrix.

We can observe that p-VAE achieves state-of-the-art performance on the dataset, slightly outperforming all baselines. In addition, p-VAE is the only method that not only accurately predicts students' responses but also computes a posterior distribution over student embeddings which enables the computation of question quality and personalized question selection. The remaining baselines either do not perform as well as p-VAE or do not compute any uncertainty information. There exist Bayesian versions of some baselines, such as Bayesian IRT (Swaminathan and Gifford 1982) or Bayesian sparse factor analysis (Lan, Studer, and Baraniuk 2014) models. Unfortunately, due to the high computational cost, these traditional Bayesian models do not scale to datasets as big as the one that we consider in this work. Thanks to the amortized inference, p-VAE is capable of efficient Bayesian inference at scale.

Implications Accurately predicting students' answers and estimate uncertainty lies at the core of educational data mining and adaptive testing. p-VAE achieves both with state-of-the-art performances which is a necessary prerequisite to the computation involved in the remaining experiments.

Question Difficulty Quantification

Setup With the complete data matrix imputed by p-VAE, we compute question difficulty by taking the average of all students' answers including observed and predicted answers. We resort to human evaluation to compare our framework's rankings. We ask the evaluator to provide a full difficulty ranking for all topics. We then compute the difficulty ranking using our framework and compute the Spearman correlation coefficient as a measure of the level of agreement between the model's and the expert's difficulty rankings.

Method	Spearman Correlation
random ordering	0.060
majority imputation	0.076
using observation	0.395
p-VAE imputation	**0.738**

Table 2: Spearman Correlation coefficients for question topic difficulty rankings between human expert and model prediction.

Method	T1	T2	T3	T4	T5
Random	0.4	0.56	0.44	0.44	0.44
Entropy	0.68	0.64	0.64	0.62	0.64
Ours	**0.72**	**0.64**	**0.68**	**0.62**	**0.72**

Table 3: Question quality ranking agreement between various methods and each evaluator (T1 through T5). Our metric achieves the best agreement with every evaluator.

Results Table 2 shows the Spearman correlation coefficients comparing expert's topic and scheme rankings, respectively, to our framework's and two other baselines' rankings. The baselines include random ordering, using majority imputation to fill the data matrix, and using the observed data alone. We see that our framework's ranking closely matches the human expert's ranking while baselines do not produce rankings that are any close to the expert's ranking.

Question Quality Quantification

Setup We compute question quality according to Eq. 3 and compare the pairwise rankings of question quality to those provided by human evaluators. Our 5 evaluators are all highly respected math teachers who have no prior information about this work. We resort to pairwise comparison, i.e., we give all 5 evaluators a pair of questions and ask them to give a preference on which question is of higher quality. We then compute the number of times our model agrees with each evaluator's choice of quality. Because this evaluation requires significant domain knowledge and because there are more than ten thousand questions and thus many more possible pairs, we only sample 80 pairs from a subset of our full dataset. Although limited, the evaluation of these samples provides preliminary evidence on whether the proposed quality metric agrees with domain experts. Figure 3 illustrates the interface that the evaluator sees when they provide the question quality labels. Each evaluator performed this task in isolation without knowledge of the other evaluators' labels.

We also consider two baseline metrics for question quality. The first metric randomly assigns a ranking to each question. The second metric computes the entropy of a question where the probability is the portion of correct or incorrect answers among all answer records for this question.

Results Table 3 shows the agreement between the models' question quality rankings and those provided by each of the
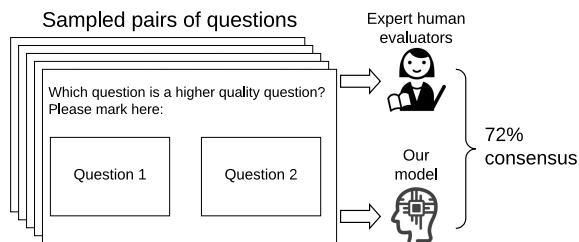

Figure 3: Illustration of question quality evaluation interface for the human evaluator. Our quality metric achieves a maximum of 72% agreement with human evaluators.

5 evaluators. We can see that our proposed quality metric achieves the highest agreement with human judgment. Also note that the entropy baseline, which is also inspired by information theory, although outperformed by our metric, is significantly better than random rankings. This preliminary result implies that further development of better question quality metrics via similar information-theoretic approaches has the potential to accurately capture human judgment of question quality and is a promising direction to pursue.

To visualize the high and low-quality questions that our framework selects, we show in Figure 4 two examples of high-quality questions (top row) and two examples of low-quality questions (bottom row) determined by our framework. For each pair, the left image shows the actual question, and the right image shows the stacked portion plot. The stacked portion plot shows the percentage of students in different ability ranges who have answered the question correctly. (the correct answer choice is always at the top, i.e., the red color part of the plot, and the remaining colors are the remaining three incorrect answer choices). The stacked portion plot is produced using the observed students' answer choices to the questions (i.e., A, B, C, or D choices).

In addition to the question content itself, we can gain some insights by examining and comparing the stacked portion plots. For example, We can see that high-quality questions better test the variability in students' abilities because fewer students with a lower ability score can answer them correctly, whereas more students with a higher ability score can answer them correctly. This phenomenon is not present in lower quality questions, where most of the students, regardless of their ability score, tend to answer them either correctly or incorrectly.

Implications Our difficulty and quality metrics efficiently and accurately provide data-driven analytics on questions which may help teachers in designing quizzes, homework sets, and exams that best suit their classes.

Personalized Question Selection

Setup To quantitatively evaluate the performance of our selection strategy, we proceed as follows. We first sample a subset of students from the dataset and sample 10% of each student's answer records as the prediction targets. For each student, we then sequentially choose 10 questions as prescribed in Eq. 4, reveal their answers, and predict the stu-
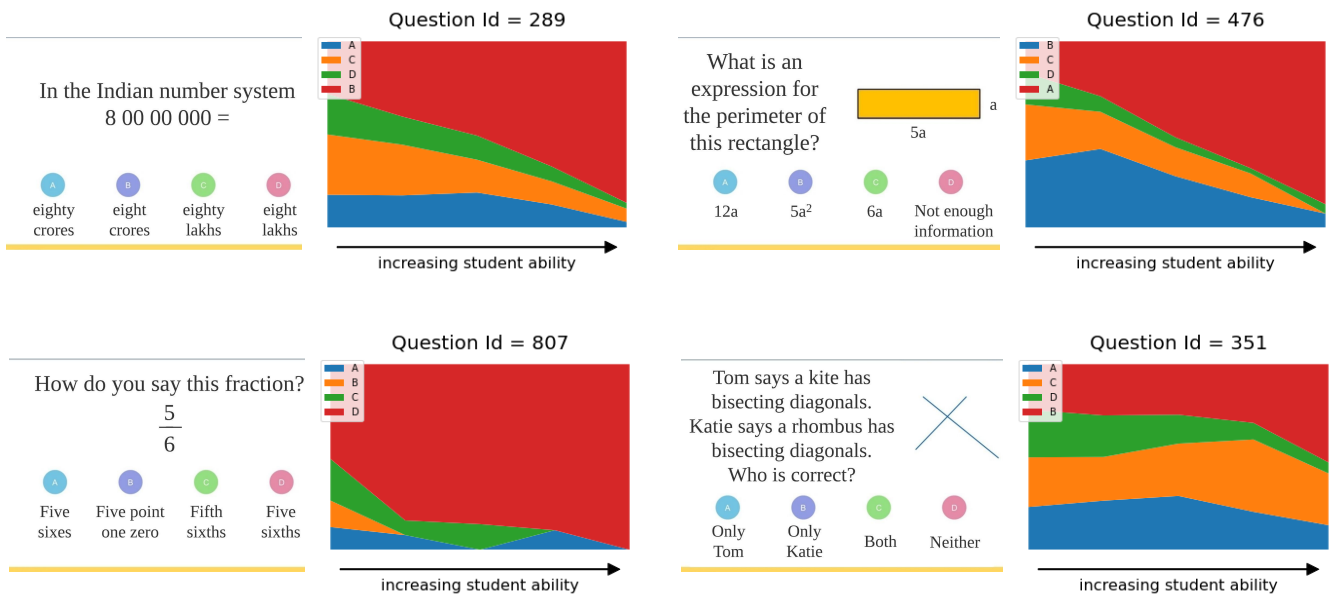
Figure 4: Two examples of high-quality questions (top row) and two examples of low-quality questions (bottom row) determined by our framework. For each pair, the left image shows the actual question, and the right image shows the stacked portion plot indicating the percentage of students who answered A, B, C, or D, where the top portion (red) always indicates the percentage of students who answered correctly. **Top row**: high-quality questions differentiate students' abilities. **Bottom left**: this low-quality question is too easy; students tend to answer it correctly despite their ability. **Bottom right**: these low-quality questions are either too easy or too difficult; the majority of the students tend to answer it either correctly or incorrectly despite their ability.
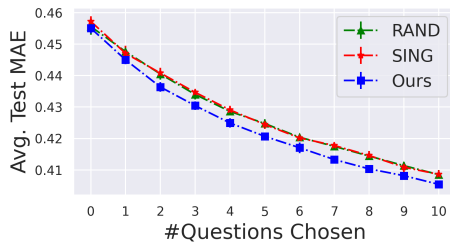

Figure 5: Our personalized question selection strategy outperforms baseline strategies for each question selected.

dent's answers using these 10 revealed answers through p-VAE. We report the average mean absolute error over all the sampled answer records at each of the 10 steps in the question selection process. We consider two baselines, a random selection strategy (RAND) and a single global optimal strategy (SING) which is similar to our proposed strategy but the reward is averaged over all students. Therefore, this strategy only selects one single sequence of questions for all students and cannot achieve personalization.

Results   Figure 5 reports the average MAE comparing our proposed strategy with RAND and SING after 10 runs. Our strategy achieves lower MAE at every step of the question selection process, demonstrating its effectiveness in choosing a sequence of questions that minimizes prediction error.

To demonstrate that our strategy is personalized for each student, we show in figure 6 the sequence of questions for 10 students. The x-axis is the step (number) of the question selected and the y-axis is the question ID. Each color represents the selected question sequence for one student. We can see that our strategy picks a very different set of questions for these 10 students. We also show the topics of the selected questions for two students in Table 4 to gain some understanding of how our strategy chooses questions. We can see that the questions chosen for these two students are very different: the questions for the first students emphasize a few recurring topics while the questions for the second students cover a diverse set of topics.

Implications   Our question selection strategy provides an efficient and automatic method to sequentially select questions for students. This can be used in large-scale learning and adaptive testing scenarios in which teachers cannot attend to each individual students but still would like to have a quick way to assess each student's skills.

Related Work

AI in Education   There is a vast and rapidly expanding literature on using AI for advancing education. A sample of examples includes knowledge tracing (Lan et al. 2014; Rasch 1960; Vie and Kashima 2019; Lan, Studer, and Baraniuk 2014; Naito et al. 2018; Piech et al. 2015); various automation including grading (Waters, Tinapple, and Baraniuk 2015; Lan et al. 2015), feedback generation (Wu et al. 2019) and quiz question generation (Wang et al. 2018); and under-

Questions	Student #1	Student #2
Q1	Squares, Cubes, etc.	Perimeter and Area
Q2	Simplifying Expressions by Collecting Like Terms	Line Symmetry
Q3	Squares, Cubes, etc	Missing Lengths
Q4	Mental Multiplication and Division	Measuring Angles
Q5	Mental Multiplication and Division	Line Symmetry
Q6	Factors and Highest Common Factor	Basic Angel Facts
Q7	Factors and Highest Common Factor	Angles
Q8	Factors and Highest Common Factor	Measuring Angles
Q9	Simplifying Expressions by Collecting Like Terms	Perimeter and Area
Q10	Simplifying Expressions by Collecting Like Terms	Transformations

Table 4: Example of the topics of the 10 questions that our framework chooses for 2 students.

Figure 6: Illustration of the questions that our framework selects for 10 students. Each student gets a personalized sequence of questions.

standing students' online and offline behaviors such as collaboration (Waters, Studer, and Baraniuk 2014) and cheating (Levitt and Lin 2015).

Most related to our work is prior literature on acquiring educational insights by analyzing students' answer records to questions. Notable examples include the Rasch model (Rasch 1960) that outputs a scalar question difficulty and a scalar student ability; the sparse factor analysis (SPARFA) model (Lan et al. 2014) which extends the Rasch model to output multi-dimensional question and student analytics; the time-varying SPARFA model (Lan, Studer, and Baraniuk 2014) capable of processing time-varying students' answer records which is a more realistic scenario. These models, however, are limited to their high computational complexity and thus are difficult to apply to large scale educational data analysis. Our work complements prior work in that we develop a framework capable of efficiently analyzing large educational data sets while at the same time extracting educationally meaningful insights.

Missing Data Imputation The missing Data Imputation method is used in many real-world applications, such as recommender systems. Many existing approaches for recommender systems rely on linear methods because of their efficiency and scalability (Salakhutdinov and Mnih 2007; Parhi, Pal, and Aggarwal 2017; Stern, Herbrich, and Graepel 2009). More recent literature introduces nonlinear models using deep learning for improved model capacity, notably with variational autoencoders (Chen and de Rijke 2018; Karamanolakis et al. 2018; Li and She 2017; Nazabal et al. 2018; Wang, Shi, and Yeung 2015). However, many of these models do not handle missing data without some ad-hoc modifications to the data, i.e., zero imputation as in (Kuchaiev and Ginsburg 2017; Liang et al. 2018; Sedhain et al. 2015). Such an ad-hoc way of replacing missing values in the data matrix significantly changes the original data distribution and can negatively impact the imputation results. In contrast to the above methods, p-VAE leverages amortized Bayesian inference with a special model architecture to efficiently handle missing data and quantify uncertainty, which is ideal for our application.

Our work builds on (Ma et al. 2018, 2019; Gong et al. 2019; Mattei and Frellsen 2018), all of which take advantage of the desirable properties of p-VAE for recommender systems and information acquisition framework, respectively. Our work further advances such prior work in that we extend, to the first of our knowledge, p-VAE to the application of education data mining with a novel information criterion to extract educationally meaningful insights.

Conclusions

In this paper, we develop a framework to analyze questions in online education platforms on a large scale. Our framework combines the recently proposed partial variational auto-encoder (p-VAE) for efficiently processing large scale, partially observed educational datasets, and novel metrics and strategy for automatically producing a suite of meaningful and actionable insights about quiz questions. We demonstrate the applicability of our framework on a real-world educational dataset, showcasing the rich and interpretable information including question difficulty, question quality, and personalization that our framework obtains from millions of students' answer records to multiple-choice questions.

Our framework is highly flexible, which enables further improvements and extensions to obtain richer educational insights. For example, one extension is to customize the information-theoretic metrics for extracting various other information of interest. Another extension is to adapt the p-VAE model for time-series data, where we can work with a more realistic yet challenging scenario that students' states of knowledge change over time. To facilitate and encourage future research, we have open sourced our dataset in the form of a competition for AI in education; see https://eedi.com/projects/neurips-education-challenge for more details.

Acknowledgements

References

Bao, W. 2020. COVID-19 and online teaching in higher education: A case study of Peking University. *Human Behavior and Emerging Technologies* 2(2): 113–115.

Blanchette, J. 2001. Questions in the online learning environment. *Journal of distance education* 16(2): 37–57.

Chalmers, R. P.; et al. 2012. mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software* 48(6): 1–29.

Chang, H.-H.; Ying, Z.; et al. 2009. Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics* 37(3): 1466–1488.

Chen, Y.; and de Rijke, M. 2018. A Collective Variational Autoencoder for Top-N Recommendation with Side Information. In *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*, 3–9.

Chrysafiadi, K.; and Virvou, M. 2013. Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications* 40(11): 4715–4729.

Connor-Greene, P. A. 2000. Assessing and Promoting Student Learning: Blurring the Line between Teaching and Testing. *Teaching of Psychology* 27(2): 84–88.

George, T.; and Merugu, S. 2005. A Scalable Collaborative Filtering Framework Based on Co-Clustering. In *Proc. IEEE International Conference on Data Mining*, 625–628. IEEE Computer Society.

Gong, W.; Tschiatschek, S.; Turner, R.; Nowozin, S.; Hernández-Lobato, J. M.; and Zhang, C. 2019. Icebreaker: Element-wise Active Information Acquisition with Bayesian Deep Latent Gaussian Model. In *Proc. Conference on Neural Information Processing Systems*.

Heffernan, N. T.; and Heffernan, C. 2014. The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. *Journal of Educatinal Data Mining* 24: 470–497.

Karamanolakis, G.; Cherian, K. R.; Narayan, A. R.; Yuan, J.; Tang, D.; and Jebara, T. 2018. Item recommendation with variational autoencoders and heterogeneous priors. In *Workshop on Deep Learning for Recommender Systems*, 10–14.

Karpicke, J. 2012. Retrieval-Based Learning: Active Retrieval Promotes Meaningful Learning. *Current Directions in Psychological Science* 21(3): 157–163.

Karpicke, J. D.; and Roediger, H. L. 2008. The Critical Importance of Retrieval for Learning. *Science* 319(5865): 966–968.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Koedinger, K. R.; Kim, J.; Jia, J. Z.; McLaughlin, E. A.; and Bier, N. L. 2015. Learning is Not a Spectator Sport: Doing is Better Than Watching for Learning from a MOOC. In *Proc. ACM Conference on Learning at Scale*, 111–120.

Koren, Y. 2008. Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model. In *Proc. SIGKDD Conference on Knowledge Discovery and Data Mining*, 426–434.

Kovacs, G. 2016. Effects of In-Video Quizzes on MOOC Lecture Viewing. In *Proc. ACM Conference on Learning at Scale*, 31–40.

Kuchaiev, O.; and Ginsburg, B. 2017. Training Deep AutoEncoders for Collaborative Filtering. *arXiv preprint arXiv:1708.01715* .

Lan, A. S.; Studer, C.; and Baraniuk, R. G. 2014. Time-varying Learning and Content Analytics via Sparse Factor Analysis. In *Proc. SIGKDD Conference on Knowledge Discovery and Data Mining*, 452–461.

Lan, A. S.; Vats, D.; Waters, A. E.; and Baraniuk, R. G. 2015. Mathematical Language Processing: Automatic Grading and Feedback for Open Response Mathematical Questions. In *Proc. ACM Conference on Learning at Scale*, 167–176.

Lan, A. S.; Waters, A. E.; Studer, C.; and Baraniuk, R. G. 2014. Sparse Factor Analysis for Learning and Content Analytics. *Journal of Machine Learning Research* 15: 1959–2008.

Levitt, S. D.; and Lin, M.-J. 2015. Catching cheating students. Technical report, National Bureau of Economic Research.

Li, X.; and She, J. 2017. Collaborative Variational Autoencoder for Recommender Systems. In *Proc. SIGKDD Conference on Knowledge Discovery and Data Mining*, 305–314.

Liang, D.; Krishnan, R. G.; Hoffman, M. D.; and Jebara, T. 2018. Variational Autoencoders for Collaborative Filtering. *arXiv preprint arXiv:1802.05814* .

Luo, X.; Zhou, M.; Xia, Y.; and Zhu, Q. 2014. An Efficient Non-Negative Matrix-Factorization-Based Approach to Collaborative Filtering for Recommender Systems. *IEEE Transactions on Industrial Informatics* 10(2): 1273–1284.

Ma, C.; Gong, W.; Hernández-Lobato, J. M.; Koenigstein, N.; Nowozin, S.; and Zhang, C. 2018. Partial VAE for hybrid recommender system. In *Proc. Conference on Neural Information Processing Systems workshop on Bayesian Deep Learning*.

Ma, C.; Tschiatschek, S.; Palla, K.; Hernandez-Lobato, J. M.; Nowozin, S.; and Zhang, C. 2019. EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE. In *Proc. International Conference on Machine Learning*, volume 97, 4234–4243.

Mattei, P.-A.; and Frellsen, J. 2018. missiwae: Deep generative modelling and imputation of incomplete data. *arXiv preprint arXiv:1812.02633* .

Naito, J.; Baba, Y.; Kashima, H.; Takaki, T.; and Funo, T. 2018. Predictive Modeling of Learning Continuation in Preschool Education Using Temporal Patterns of Development Tests.

Nazabal, A.; Olmos, P. M.; Ghahramani, Z.; and Valera, I. 2018. Handling Incomplete Heterogeneous Data using VAEs. *arXiv preprint arXiv:1807.03653* .

Parhi, P.; Pal, A.; and Aggarwal, M. 2017. A survey of methods of collaborative filtering techniques. In *Proc. International Conference on Inventive Systems and Control*, 1–7.

Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L. J.; and Sohl-Dickstein, J. 2015. Deep Knowledge Tracing. In *Proc. Advances in Neural Information Processing Systems*, 505–513.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proc. Conference on Computer Vision and Pattern Recognition*, 652–660.

Rasch, G. 1960. *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Oxford, England: Nielsen & Lydiche.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proc. International Conference on Machine Learning*.

Salakhutdinov, R.; and Mnih, A. 2007. Probabilistic Matrix Factorization. In *Proc. Conference on Neural Information Processing Systems*, 1257–1264.

Sandars, J.; Correia, R.; Dankbaar, M.; de Jong, P.; Goh, P. S.; Hege, I.; Masters, K.; Oh, S.-Y.; Patel, R.; Premkumar, K.; et al. 2020. Twelve tips for rapidly migrating to online learning during the COVID-19 pandemic. *MedEdPublish* 9.

Sedhain, S.; Menon, A. K.; Sanner, S.; and Xie, L. 2015. Autorec: Autoencoders meet collaborative filtering. In *Proc. International World Wide Web Conference*, 111–112.

Stamper, J.; Niculescu-Mizil, A.; Ritter, S.; Gordon, G.; and Koedinger, K. 2010a. Algebra I 2008-2009. *Challenge data set from KDD Cup 2010 Educational Data Mining Challenge* .

Stamper, J.; Niculescu-Mizil, A.; Ritter, S.; Gordon, G.; and Koedinger, K. 2010b. Bridge to Algebra 2008-2009. *Challenge data set from KDD Cup 2010 Educational Data Mining Challenge* .

Stern, D. H.; Herbrich, R.; and Graepel, T. 2009. Matchbox: large scale online bayesian recommendations. In *Proc. International World Wide Web Conference*, 111–120. ACM.

Swaminathan, H.; and Gifford, J. A. 1982. Bayesian Estimation in the Rasch Model. *Journal of Educational Statistics* 7(3): 175–191.

Twigg, C. A. 2003. Improving Learning and Reducing Costs: New Models for Online Learning. *EDUCAUSE Review* 38(5): 28–38. ISSN 1527-6619.

VanLehn, K. 1988. Student modeling. *Foundations of Intelligent Tutoring Systems* 55: 78.

Verawardina, U.; Asnur, L.; Lubis, A. L.; Hendriyani, Y.; Ramadhani, D.; Dewi, I. P.; Darni, R.; Betri, T. J.; Susanti, W.; and Sriwahyuni, T. 2020. Reviewing Online Learning Facing the Covid-19 Outbreak. *Talent Development & Excellence* 12.

Vie, J.-J.; and Kashima, H. 2019. Knowledge Tracing Machines: Factorization Machines for Knowledge Tracing. In *Proc. AAAI Confeence on Artificial Intelligence*, volume 33, 750–757.

Wang, H.; Shi, X.; and Yeung, D.-Y. 2015. Relational Stacked Denoising Autoencoder for Tag Recommendation. In *Proc. AAAI Confeence on Artificial Intelligence*, 3052–3058.

Wang, Z.; Lamb, A.; Saveliev, E.; Cameron, P.; Zaykov, Y.; Hernández-Lobato, J. M.; Turner, R. E.; Baraniuk, R. G.; Barton, C.; Jones, S. P.; et al. 2020. Diagnostic questions: The neurips 2020 education challenge. *arXiv preprint arXiv:2007.12061* .

Wang, Z.; Lan, A. S.; Nie, W.; Waters, A. E.; Grimaldi, P. J.; and Baraniuk, R. G. 2018. QG-net: A Data-driven Question Generation Model for Educational Content. In *Proc. ACM Conference on Learning at Scale*, 7:1–7:10.

Waters, A. E.; Studer, C.; and Baraniuk, R. G. 2014. Collaboration-type identification in educational datasets. *Journal of Educatinal Data Mining* 6(1): 28–52.

Waters, A. E.; Tinapple, D.; and Baraniuk, R. G. 2015. BayesRank: A Bayesian Approach to Ranked Peer Grading. In *Proc. ACM Conference on Learning at Scale*, 177–183.

Wu, M.; Mosse, M.; Goodman, N.; and Piech, C. 2019. Zero Shot Learning for Code Education: Rubric Sampling with Deep Learning Inference. In *Proc. AAAI Confeence on Artificial Intelligence*, volume 33, 782–790.

Yan, D.; Von Davier, A. A.; and Lewis, C. 2016. *Computerized multistage testing: Theory and applications*. CRC Press.

Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R. R.; and Smola, A. J. 2017. Deep sets. In *Proc. Conference on Neural Information Processing Systems*, 3391–3401.

Zhang, C.; Bütepage, J.; Kjellström, H.; and Mandt, S. 2019. Advances in Variational Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(8): 2008–2026.