CausalEdu: a real-world education dataset for temporal causal discovery and inference

Wenbo Gong WENBOGONG@MICROSOFT.COM
Microsoft Research

Digory Smith DIGORY.SMITH@EEDI.CO.UK
Eedi

Zichao Wang ZW16@RICE.EDU
Rice University

Craig Barton CRAIG.BARTON@EEDI.CO.UK
Eedi

Simon Woodhead SIMON.WOODHEAD@EEDI.CO.UK
Eedi

Nick Pawlowski NICK.PAWLOWSKI@MICROSOFT.COM
Microsoft Research

Joel Jennings JOELJENNINGS@MICROSOFT.COM
Microsoft Research

Cheng Zhang CHENG.ZHANG@MICROSOFT.COM
Microsoft Research

Abstract

Causal machine learning is an emerging field that aims to leverage machine learning techniques to tackle causal discovery and inference problems. Despite significant progress in recent years, one of the main limitations is the reliance on (semi-)synthetic datasets for evaluation, due to the difficulty in obtaining real-world ground truth data. This lack of real-world datasets and potential discrepancies between models and real-world scenarios pose significant challenges for real-time deployment. To address this issue, we propose a temporal causal dataset, *CausalEdu*, which contains student performance data on multiple-choice questions in math collected from Eedi's online learning platform. With the same platform, we conducted A/B tests, together with the expert opinions, to identify causal effects between different math concepts, providing the ground truth data. Our dataset supports both (1) causal discovery between various math concepts, and (2) the estimation of conditional average treatment effects (i.e. CATE) of learning one concept on the question accuracy of others. We believe that *CausalEdu* offers a unique opportunity for researchers to test their causal methods in real-world settings, uncover underlying challenges, and ultimately contribute to the development of innovative causal techniques.

Keywords: Causal discovery; Causal inference; Real-world dataset; Randomized control trials.

1. Introduction

Causal machine learning has emerged as a field dedicated to employing machine learning methods to address causal problems, including causal discovery and inference (Peters et al., 2017). Despite recent advancements in this domain, numerous unresolved challenges persist, such as missing data, selection bias, unobserved confounders, and other real-world complexities. A significant factor limiting progress in causal machine learning is the scarcity of real-world datasets with ground truth information. Evaluating causal machine learning techniques typically relies on synthetic or semi-synthetic datasets, the generative mechanisms of which may not accurately represent real-world scenarios. The primary challenge in utilizing real-world datasets is determining the ground truth causal relationships between variables and relevant causal quantities, such as (conditional) average treatment effect ((C)ATE). Acquiring this information often necessitates randomized control trials or expert opinions, which are frequently unavailable due to cost or ethical constraints.

To address this limitation, we introduce a novel timeseries causal dataset, *CausalEdu*, derived from the online education platform created by Eedi, a company focused on online education. This dataset captures the performance of school students aged 11 to 16 in multiple-choice questions related to mathematical concepts, referred to as *constructs*. To overcome the absence of real-world ground truth, we conducted A/B tests on the learning platform, together with the expert opinions regarding the causal connections between constructs, to obtain ground-truth causal relationships and the CATE. This unique feature makes *CausalEdu* stand out among other benchmark datasets for timeseries causal evaluation (e.g. DREAM3 (Madar et al., 2010) and Netsim (Smith et al., 2011)).

The impact and significance of our proposed dataset lie in its potential to serve as a valuable resource for causal researchers, enabling them to test and refine their causal methods using real-world data, and uncover previously unexplored problems. We anticipate that *CausalEdu* will facilitate the development of novel causal approaches, fostering advancements in the field and closing the gap between causal theory and real-world applications.

2. Background

One of the most significant challenges in education lies in determining the optimal sequence of subjects to facilitate personalized learning experiences. Topics can be broken down into elements, known as constructs, which represent the most fundamental units of learning. For instance, "Converting between cm and m" is a construct within the broader topic of "Units of length". Uncovering the causal relationships between these constructs can inform the development of more effective and tailored curricula.

Since direct measurement of a student's knowledge of a construct is not feasible, we rely on their responses to questions as a proxy. A particularly useful and simple question type is the *diagnostic question*. This multiple-choice question format, consisting of four possible answers, is designed to evaluate the knowledge of a single construct. Analyzing the correctness and accuracy of students' responses can reveal hidden causal relationships between constructs (e.g., some constructs may serve as prerequisites for others). Further, this information can then be leveraged for causal inference as well, which quantifies the impact of mastering one construct on others.

To derive *CausalEdu*, we used Eedi's online learning platform for teachers to crowdsource responses to these questions. Students can engage with quizzes and lessons on the platform, and the resulting real-world data will be made available to form the *CausalEdu* dataset. To obtain the ground truth for causal discovery and inference, we also designed the randomized control experiments for a subset of constructs. Additionally for causal discovery, we also consult with experts in Eedi to gather their opinions regarding which constructs are the prerequisites for the others to maximize the ground-truth coverage of constructs.

3. Dataset details

3.1. Overview

The *CausalEdu* consists of four parts: (1) training files; (2) AB test lists; (3) ground truth files; and (4) meta information files regarding the constructs and users. The data is available at https://github.com/Eedi/CausalEdu.

3.2. Training data

Filename: `checkins_lessons_checkouts_training.csv`

This is the raw training data obtained from a real-world online education platform.

Quizzes consist of five "checkin" diagnostic questions. Students may attempt a lesson after each checkin question, although it is optional if they answered correctly. After completing the lesson, they must answer a "checkout" diagnostic question related to the same construct. If a student answers a checkin or checkout question incorrectly, they must retry the question before proceeding. Both attempts are recorded.

The dataset exclusively features data from student responses to mathematics content collected between February 1st, 2022, and August 3rd, 2022.

Table 1 is an illustration of the data records.

Table 1: Primary training data.

QuizSessionId	AnswerId	UserId	QuizId	QuestionId	IsCorrect	AnswerValue
8	57	5	232950	131432	0	2
8	58	5	232950	131432	0	3
8	None	5	232950	131432	None	None
8	59	5	232950	133665	1	4
8	60	5	232950	131433	1	1

CorrectAnswer	QuestionSequence	ConstructId	Type	Timestamp
4	2	433	Checkin	...06:15:01
4	2	433	CheckinRetry	...06:16:18
None	2	433	Lesson	...06:26:19
4	2	433	Checkout	...06:27:03
1	3	427	Checkin	...06:30:41

Each row in the table represents a quiz event with an associated *Timestamp*, either for a question attempt or a lesson. The *Type* for a question attempt can be ”Checkin”, ”CheckinRetry”, ”Checkout”, or ”CheckoutRetry”, while the *Type* for a lesson is ”Lesson”.

The *QuizSessionId* identifies a user’s quiz attempt, with the *QuizId* and *UserId* specifying the quiz and user, respectively. Each question in the quiz, identified by *QuestionId*, is linked to a single construct (*ConstructId*). The user’s answer is represented by *AnswerId*.

IsCorrect denotes whether a question was answered correctly (1) or incorrectly (0). Each question has four possible answers, with the student’s choice recorded as *AnswerValue* (1 to 4) and the correct option as *CorrectAnswer* (1 to 4).

QuestionSequence (1 to 5) records the question’s position in the quiz, linking checkin questions, lessons, and checkout questions.

The dataset includes both incomplete and complete quiz attempts, totaling over 65,000 quiz attempts from 6,400 students. This encompasses more than 470,000 diagnostic question answers and 37,000 lesson attempts.

3.3. AB test list

Filename: `constructs_input_test.csv`

The file contains a construct list that is used in A/B test. Due to the resources constraints, only some pairs have been tested in the A/B test (refer to `checkin_to_checkout.csv`). For the ones that are not in A/B test, one can find their causal relations in `construct_prerequisites_test.csv`, which stores the expert opinions.

Filename: `construct_experiments_input_test.csv`

The questionnaire (Table 2) for causal inference consists of rows containing queries to compute CATE. Each row describes a unique A/B experiment. The target construct is *QuestionConstructId*, while *TreatmentConstructId* and *ControlConstructId* represent the constructs taught in the treatment and control group, respectively. The *Year* indicates the year group of participating students.

Note that this file does not contain the full list of construct pairs we use in A/B test, we only includes some of them where the number of participants are sufficient. For the complete list, the user can find them in `construct_experiments_ates_test.csv`.

Table 2: Test questionnaire

QuestionConstructId	TreatmentConstructId	ControlConstructId	Year
471	469	2930	7
2034	2028	628	10

3.4. Supported causal quantities

Our dataset supports both causal discovery and inference tasks. For causal discovery, we obtain ground truth connections for a subset of constructs through A/B tests using both A/B test and expert opinions. Users can choose any structure learning metrics that support the evaluation on a subset of nodes, such as the F1 score.

For causal inference, we focus on the estimation of the Conditional Average Treatment Effect (CATE). Specifically, we define CATE as:

$$\mathrm{CATE}(c_I, c_R, Y_{c_t}, T) = \mathbb{E}_{p(Y_{c_t}\,|\,\mathrm{do}(B=c_I),T)}[Y_{c_t}] - \mathbb{E}_{p(Y_{c_t}\,|\,\mathrm{do}(B=c_R),T)}[Y_{c_t}], \tag{1}$$

where T represents the year group information, Y_{c_t} is the average question accuracy for the target construct, c_I, c_R denotes the treatment and control lesson construct, respectively, and B stands for the lesson assignment variable.

Using this quantity, one can evaluate model inference performance with the Root Mean Square Error (RMSE) as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(\mathrm{CATE}'_k - \mathrm{CATE}^*_k)^2}, \tag{2}$$

where N is the number of queries in the questionnaire, CATE'_k represents the predicted CATE for query k, and CATE^*_k denotes the ground truth CATE.

3.5. Ground truth data

A/B testing procedure: To obtain real-world ground truth, we collected additional intervention data by conducting several A/B tests to establish the relationships between completing a lesson on a given construct and performance on questions related to other constructs. Students were randomly divided into treatment and control groups, with both groups taking an initial quiz consisting of five check-in questions from selected constructs.

Following the check-in questions, the treatment group received a lesson on a related construct, while the control group received a lesson on an unrelated construct, as determined by domain experts. Subsequently, both groups took a quiz containing five check-out questions based on the same constructs as the check-in questions to assess whether the lessons improved their understanding of the constructs. All student performance data were recorded by the online platform.

Filename: `checkin_to_checkout.csv`

This ground truth file is for discovering causal relationships between different constructs. Table 3 illustrates one row of this file.

Table 3: Ground truth causal relations

LessonConstructId	QuestionConstructId	n00	n01	n10	n11	Count	p010_m	k010_m
70	1270	8	38	3	30	79	0.83	0.78

Since this file is used to identify underlying causal relations, we do not have explicit treatment or control groups. Instead, we aim to compare question accuracy before and after a specific lesson (i.e., check-in and check-out question accuracy) to determine whether the construct associated with the lesson is a prerequisite or not.

In particular, *LessonConstructId* refers to the ID of the construct associated with the lesson taught to the students. *QuestionConstructId* is the construct ID associated with the check-in and check-out questions.

The next four columns indicate the overall performance of the users. 0 denotes an incorrect answer, and 1 denotes a correct answer. The former digit represents the check-in question, and the latter one indicates the check-out question. Thus, *n01* represents the total number of users who answered the check-in question incorrectly but managed to answer the check-out question correctly after learning the lesson. *UserCount* is the total number of users, which is the sum of the previous four columns.

The next six columns are the statistics we computed to determine whether causal relations exist between constructs. In particular, we made two hypotheses, and each leads to different statistics (i.e., starting with letter *p* or *k*). It depends on the users to decide which one is more appropriate.

Hypothesis 1 Our hypothesis is that learning a lesson cannot be harmful to a student's knowledge. Therefore, the students in *n10* can be treated as outliers or noise. Thus, we will remove them when computing the effect of learning a construct. In particular, we have

$$p010\_m = \frac{n01}{n01 + n00} \tag{3}$$

which represents the probability of correctly answering the check-out questions among the students who did not know this construct before the lesson. Since there are four choices for each question, we determine that the lesson construct is a prerequisite to the question construct if $p010_m > 0.25$ to rule out random guessing.

Hypothesis 2 The main difference compared to Hypothesis 1 is that we do not remove *n10* but assume the reason they answered the check-in question correctly is due to random guessing. Namely, *n10* should be treated as they still do not understand the question construct even after learning the lesson. Therefore, we have

$$k010\_m = \frac{n01}{n00 + n01 + n10} \tag{4}$$

which represents the same probability as *p010_m*. Similarly, we determine their causal relations by observing if $k010\_m > 0.25$.

Filename: `construct_prerequisites_test.csv`

This file contains the expert opinions regarding which constructs are the prerequisites for the others. Table 4 shows an example row.

Table 4: Expert opinions

ConstructId	**SubjectId**	**PrerequisiteConstructIds**
856	33	{483, 76}

The *ConstructId* represents a unique identifier associated with each construct, while the *SubjectId* refers to the ID of the subject to which the construct belongs. The *PrerequisiteConstructIds* column contains a tuple of construct IDs that are considered prerequisites for the given construct, based on expert opinions.

This file serves to complement the limited number of construct pairs obtained from the A/B test due to resource constraints. It is important to note that, in cases where contradictions arise between expert opinions and A/B test results, the A/B test results should take precedence.

Filename: `construct_experiments_ates_test.csv`

This file contains the ground truth data for causal inference. Table 5 demonstrates an example row of the file.

Table 5: Ground truth causal relations

TreatmentLessonConstructId	QuestionConstructId	Year	ControlLessonContructId
206	211	7	3119

Table 6: Ground truth causal relations

ControlUsersCount	TreatmentUsersCount	ate_p_1_	ate_k_1_
73	94	0.033	-0.019

To obtain the ground truth causal inference quantity (i.e., CATE), we divided the students into treatment and control groups. *TreatmentLessonConstructId* (c_I in Eq.1) specifies the construct ID associated with the lesson taught to the treatment group. *QuestionConstructId* is the construct ID (c_t in Eq.1) related to the check-out questions after learning the lesson. *Year* indicates the year group of the students (T in Eq.1). *ControlLessonConstructId* is the lesson construct Id (c_R in Eq.1) taught at the control group. *ControlUsersCount* and *TreatmentUsersCount* are the total number of students in control and treatment groups, respectively. *ate_p_1_* and *ate_k_1_* are the CATE values based on the previous two hypothesis and Eq.1.

ate_p_1_ This quantity is computed based on Hypothesis 1, where we disregard the students who correctly answer the check-in but fail on the check-out questions. We have:

$$p\_1\_ = \frac{n01 + n11}{n00 + n01 + n11} \tag{5}$$

$$ate\_p\_1\_ = p\_1\_(\text{treatment}) - p\_1\_(\text{control}) \tag{6}$$

In this case, $p\_1\_$ represents the probability of correctly answering the check-out questions.

ate_k_1_ This quantity is computed based on Hypothesis 2, where we assume the students in *n10* still do not understand the construct even after the lesson. Therefore:

$$k\_1\_ = \frac{n01 + n11}{n00 + n01 + n11 + n10} \tag{7}$$

$$ate\_k\_1\_ = k\_1\_(\text{treatment}) - k\_1\_(\text{control}) \tag{8}$$

3.5.1. META INFORMATION

Filename: `topic_pathway_metadata.csv`

The topic pathway of Eedi's dataset is a carefully curated sequence of topics, recommended by a team of expert mathematics teachers. Each row in the table represents a pair of questions (`CheckinQuestionId` and `CheckoutQuestionId`) within a topic quiz (`QuizId`). The

QuestionSequence column indicates the position of the questions within the quiz, while the QuizSequence determines the position of the quiz within the topic pathway. Quizzes are organized into Levels, which cater to different YearGroups. Each question pair addresses a specific ConstructId, which is associated with a single SubjectId. Questions can be directly linked to multiple subjects, and these connections are recorded as comma-separated values in the QuestionSubjectIds column.

Filename: subject_metadata.csv

In the dataset, each subject is identified by a unique SubjectId and has an associated Name. The table has a self-referencing structure, where the ParentId of one subject corresponds to the SubjectId of another subject. The Level column indicates the number of degrees of separation between a subject and the top-level subject, "Maths."

Filename: student_metadata.csv

Each student in the dataset is uniquely identified by a UserId. The Gender column includes one of the following values: "male," "female," "other," or "unspecified." Instead of recording a student's complete date of birth, the dataset contains the MonthOfBirth, represented as the first day of the month in which they were born. The YearGroup is self-reported by the user and does not rely on the MonthOfBirth. The dataset follows the UK year group numbering system, as documented on the UK government website: https://www.gov.uk/national-curriculum.

In the UK, additional funding is provided to schools for disadvantaged students through the "pupil premium" program. The IsPupilPremium column indicates whether a student is part of this program (1) or not (0).

If any demographic information is unknown or unprovided, the corresponding cell in the table is left blank.

4. Conclusion

In this paper, we introduced a temporal causal dataset called *CausalEdu* derived from a real-world online education platform. The main advantage of *CausalEdu* is that its ground truth are obtained through A/B tests compared to (semi-)synthetic dataset. It provides a unique opportunity for the researchers to evaluate their model performance in real-world causal discovery and inference tasks. We believe through our dataset, researchers can gain a deeper understanding of underlying problems of the existing causal model and ultimately, this knowledge can contribute to the development of novel causal methodologies for real-world applications.

References

Aviv Madar, Alex Greenfield, Eric Vanden-Eijnden, and Richard Bonneau. Dream3: network inference using dynamic context likelihood of relatedness and the inferelator. PloS one, 5(3): e9803, 2010.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference: foundations and learning algorithms. The MIT Press, 2017.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. Neuroimage, 54(2):875–891, 2011.