Workshop on deep learning with educational data

Ryan Baker
University of Pennsylvania
Philadelphia, PA 19104
ryanshaunbaker@gmail.com

Joseph E. Beck
Worcester Polytechnic Institute
Worcester, MA, 01609
josephbeck@wpi.edu

Min Chi
North Carolina State University
Raleigh, NC 27695
mchi@ncsu.edu

Neil T. Heffernan
Worcester Polytechnic Institute
Worcester, MA, 01609
nth@wpi.edu

Mike Mozer
University of Colorado Boulder
Boulder, CO 80309
mozer@colorado.edu

1. WORKSHOP TOPIC

This workshop focuses on applications of deep learning for educational data. Deep learning is a machine learning approach using neural networks with multiple levels of representational transformation (i.e., hidden layers). Deep learning has been used in a variety of domains over the past five years with impressive results. Recently, it has been used for educational data sets with mixed results when compared to traditional modeling methodologies.

We are interested in work on a variety of topics with deep learning: new prediction and modeling problems, best practices for featurizing data, network architectures, approaches to pre-training and whether it is necessary, interpreting the learned models, end-to-end deep learning approaches with low-level non-symbolic data, toolkits people have developed, empirical results on known problems to help the field develop best practices. The workshop is also interested in negative results such as analyses of data sets and domains where deep learning fails to achieve state of the art performance.

2. GOALS OF WORKSHOP

The primary goal of this workshop is to provide a venue for researchers to present emerging work. There is not much prior art on applying deep learning to educational data, and it is unclear even what the scope of possible applications are: although most work has focused on student modeling, some work has focused on using deep learning to assist in scoring essays. Having a discussion about possible application areas will be productive.

In addition, this workshop will focus on recent big topics in deep learning for educational data. A paper published in 2016 "How deep is knowledge tracing" questions the need for deep models, and will be discussed at the workshop.

Finally, this workshop will provide researchers on deep learning for EDM a chance to get focused feedback on their work. Ensuring that the research is critiqued by a roomful of people interested in the topic is more useful to the presenters (and the community) than counting on haphazard interactions at the conference.

Sharing and Reusing Data and Analytic Methods with LearnSphere

Ran Liu
ranliu@cmu.edu

Kenneth Koedinger
koedinger@cmu.edu

John Stamper
jstamper@cs.cmu.edu

Philip Pavlik
ppavlik@memphis.edu

ABSTRACT

This workshop will explore LearnSphere, an NSF-funded, community-based repository that facilitates sharing of educational data and analytic methods. The workshop organizers will discuss the unique research benefits that LearnSphere affords. In particular, we will focus on Tigris, a workflow tool within LearnSphere that helps researchers share analytic methods and computational models. Authors of accepted workshop papers will integrate their analytic methods or models into LearnSphere's Tigris in advance of the workshop, and these methods will be made accessible to all workshop attendees. We will learn about these different analytic methods during the workshop and spend hands-on time applying them to a variety of educational datasets available in LearnSphere's DataShop. Finally, we will discuss the bottlenecks that remain, and brainstorm potential solutions, in openly sharing analytic methods through a central infrastructure like LearnSphere. Our ultimate goal is to create the building blocks to allow groups of researchers to integrate their data with other researchers in order to advance the learning sciences as harnessing and sharing big data has done for other fields.

Keywords

Learning metrics; data storage and sharing; data-informed learning theories; modeling; data-informed efforts; scalability.

1. INTRODUCTION

Due to a confluence of a boom of interest both in educational technology and in the use of data to improve student learning, student learning activities and progress are increasingly being tracked and stored. There is a large variety in the kinds, density, and volume of such data and to the analytic and adaptive learning methods that take advantage of it. Data can range from simple (e.g., clicks on menu items or structured symbolic expressions) to complex and harder-to-interpret (e.g., free-form essays, discussion board dialogues, or affect sensor information). Another dimension of variation is the time scale in which observations of student behavior occur: click actions are observed within seconds in fluency-oriented math games or in vocabulary practice, problem-solving steps are observed every 20 seconds or so in modeling tool interfaces (e.g., spreadsheets, graphers, computer algebra) in intelligent tutoring systems for math and science, answers to comprehension-monitoring questions are given and learning resource choices are made every 15 minutes or so in massive open online courses (MOOCs), lesson completion is observed across days in learning management systems, chapter/unit test results are collected after weeks, end-of-course completion and exam scores are collected after many months, degree completion occurs across years, and long-term human goals like landing a job and achieving a good income occur across lifetimes. Different paradigms of data-driven education research differ both in the types of data they tend to use and in the time scale in which that data is collected. In fact, relative isolation within disciplinary silos is arguably fostered and fed by differences in the types and time scale of data used [4, 5].

Thus, there is a broad need for an overarching data infrastructure to not only support sharing and use within the student data (e.g., clickstream, MOOC, discourse, affect) but to also support investigations that bridge across them. This will enable the research community to understand how and when long-term learning outcomes emerge as a causal consequence of real-time student interactions within the complex set of instructional options available [2]. Such an infrastructure will support novel, transformative, and multidisciplinary approaches to the use of data to create actionable knowledge to improve learning environments for STEM and other areas in the medium term and will revolutionize learning in the longer term.

LearnSphere transforms scientific discovery and innovation in education through a scalable data infrastructure designed to enable educators, learning scientists, and researchers to easily collaborate over shared data using the latest tools and technologies. LearnSphere.org provides a hub that integrates across existing data silos implemented at different universities, including educational technology "click stream" data in CMU's DataShop, massive online course data in Stanford's DataStage and analytics in MIT's MOOCdb, and educational language and discourse data in CMU's new DiscourseDB. LearnSphere integrates these DIBBs in two key ways: 1) with a web-based portal that points to these and other learning analytic resources and 2) with a web-based workflow authoring and sharing tool called Tigris. A major goal is to make it easier for researchers, course developers, and instructors to engage in learning analytics and educational data mining without programming skills.

2. SPECIFIC WORKSHOP OBJECTIVES

Broadly, this workshop offers those in the EDM community an exposure to LearnSphere as a community-based infrastructure for educational data and analysis tools. In opening lectures, the organizers will discuss the way LearnSphere connects data silos across universities and its unique capabilities for sharing data, models, analysis workflows, and visualizations while maintaining confidentiality.

More specifically, we propose to focus on attracting, integrating, and discussing researcher contributions to Tigris, the web-based workflow authoring and sharing tool. The goal of Tigris is to support any custom analysis method that can be applied to the datasets and to produce outputs in a standardized way that facilitates both quantitative and qualitative model comparisons. This workflow feature allows researchers to apply their own analysis methods to the vast array of datasets available in the educational data repository. It affords researchers the advantages of (1) using the built-in learning curve visualizations on the outputs of their own analysis workflows, (2) easily comparing their results both quantitatively and graphically to the outputs of