Causal Discovery with Tetrad in LearnSphere's Tigris

Richard Scheines
Carnegie Mellon University
scheines@cmu.edu

Peter Schaldenbrand
Carnegie Mellon University
pschalde@cs.cmu.edu

Kenneth Koedinger
Carnegie Mellon University
koedinger@cmu.edu

ABSTRACT

This tutorial will explore causal discovery algorithms in Tetrad[1] implemented in LearnSphere's Tigris workflow tool.[2] The Tetrad software suite contains algorithms that search for causal models from observational and experimental datasets (Spirtes et al., 2000), and has been productively applied to many educational datasets.[3] LearnSphere's Tigris is an online workflow authoring tool and data mining infrastructure for custom analyses of new and existing data formats, including the educational data repository DataShop (Stamper et al., 2011). The tutorial will teach the fundamentals of causal discovery with hands on work in Tetrad, and teach integrated causal data-mining with hands on work in Tigris. Attendees will gain experience sharing their results and methods in Tigris with others as well as connecting their analyses to thousands of datasets available in DataShop.

Keywords

Causal discovery; data storage and sharing; analysis reproducibility; data-informed learning theories; modeling.

1. INTRODUCTION

As more educational technologies are created, more data is being logged and available to researchers. Mining this data for information concerning what pattern of student behaviors cause better learning outcomes is crucial if we are to intervene, either on the design of the online material, or on the student's behavior more directly. The gold standard for determining causal relationships is randomized controlled experiments, but these cost money and time, and are often unable to answer questions about the mechanisms by which an intervention causes learning. Tetrad contains algorithms that can mine both experimental and observational datasets for information about the causal relationships that might have produced the observed data. It has been used to guide follow-up experiments in educational research (Rau and Scheines 2012). With an abundance of datasets available and causal questions at the center of all kinds of science, educational data-mining for causation is an increasingly important task.

Tetrad[1] is a standalone Java program developed by the Philosophy department at Carnegie Mellon University and more recently in conjunction with the Department of Biomedical Informatics at the University of Pittsburgh.[4] The aim of the program is to provide sophisticated methods of creating, searching for, estimating, and testing causal and statistical models of both experimental and observational data. No prior programming experience is needed to use Tetrad.

Much of the functionality of Tetrad has been included in an online workflow tool called Tigris. The LearnSphere project created Tigris to connect multiple data sources to analytical tools that are open source and available for collaboration. 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 (Stamper et al., 2011), massive online course data in Stanford's DataStage and analytics in MIT's MOOCdb (Veeramachaneni, Halawa, et al., 2014), and educational language and discourse data in CMU's new DiscourseDB (Jo et al., 2016). Researchers can share their data mining code in the Tigris tool by adding it to the open source repository[5] such as the Performance Factors Analysis (Pavlik et al., 2009) program or the Feature Extractor (Veeramachaneni, O'Reilly, et al., 2014).

The advantages of the Tetrad implementation in Tigris are numerous. Since Tigris is a web based tool, workflows can be shared amongst anyone with a computer. There is no need to worry about having Java on your computer or downloading different versions of Tetrad. The owner of a workflow can share their Tetrad analysis on Tigris with whomever they wish. In addition to shareability, Tetrad on Tigris is connected to the educational data repository DataShop. This connection allows for easy access and permission control of data in a Tetrad analysis.

Tigris workflows are executed and saved remotely on secure servers. LearnSphere has a scalable infrastructure that allows for computation on large datasets, and so the limitations on computation are not influenced by the user's machine. The analyses are run remotely, so the user can close their computers and the code still runs. The analyses can be accessed from any computer and are safely saved.

In this tutorial, we will explore the causes of learning in online courses using Tetrad in Tigris. The organizers will explain how to interpret the output generated by algorithms in Tetrad and how to build a Tigris workflow that utilizes data in DataShop to determine the causes of learning with Tetrad. The attendees will gain access to a dataset that was collected from an online college course. The data contains variables such as pages viewed, activities performed, quiz scores, and final exam grades. Using

[1] http://www.phil.cmu.edu/tetrad/

[2] https://pslcdatashop.web.cmu.edu/LearnSphere

[3] Scheines, et al. 2005; Rau et al., 2013; Koedinger, Scheines, and Schaldenbrand, 2018.

[4] See https://www.ccd.pitt.edu/

[5] https://github.com/LearnSphere/WorkflowComponents

this data, the attendees will generate causal models in a similar fashion to (Koedinger et al., 2015) and (Koedinger et al., 2018).

2. ORGANIZATIONAL DETAILS

2.1 Type of Event

Tutorial

2.2 Proposed Schedule

Table 1. Proposed Half-day Schedule

Time	Item
12:30	Introductions
1:00	Causal discovery using Tetrad
2:30	Demonstrate building a Tigris workflow that utilizes DataShop data and Tetrad
3:10	Coffee Break
3:20	Hands on: Import an educational dataset into Tigris and determine the causes of learning
4:20	Participants analyze their own data with Tetrad
5:00	Closing / High-level Discussion

2.3 Activities

Attendees will view presentations from the organizers, do hands on exercises in Tetrad, create accounts in LearnSphere, gain access to datasets in DataShop, and create Tigris workflows. The Tigris workflow will utilize Tetrad and give the attendees hands on experience with causal discovery algorithms. Discussions will be led by the organizers to determine which causal relationships need to be explored more in education and how Tetrad might be used to understand how students learn in future research.

2.4 Connection to Other Tutorials

The other proposed tutorial using LearnSphere's Tigris is a broad introduction to the capabilities of Tigris. This tutorial would focus on causal discovery and dive deeper both into Tetrad and into the Tetrad tools available in Tigris. Our workshop is thus proposed as a second half of the day tutorial. Attendees can learn about Tigris as a whole in the morning, then gain experience using the causal discovery algorithms in Tigris in the afternoon. Afternoon workshop attendees do not need to attend the morning workshop, however. We will make it self-contained.

2.5 Expected Numbers

We expect 15-20 participants based on previous tutorials.

2.6 Required Equipment

Projector and screen will be required by organizers. Attendees will need to bring laptops and will need adequate internet connectivity.

3. ORGANIZERS

3.1 Richard Scheines

RICHARD SCHEINES is the Dean of the Mariana Brown Dietrich College of Humanities and Social Sciences at Carnegie Mellon University and a Professor in the Department of Philosophy. His research is on causal discovery, and in particular, the problem of learning about causation from statistical evidence. The theoretical and computational dimensions of Dr. Scheines's work are implemented in the Tetrad causal discovery tool. Building efficient and practically useful algorithms for causal discovery is as much computer science as philosophy, and thus Dr. Scheines has a courtesy appointment in the Machine Learning Department at CMU.

In addition to Dr. Scheines's research into causal and statistical models, he has proven to be dedicated to building and researching the effectiveness of educational software, such as determining the effectiveness of online material in a causal and statistical reasoning course (Scheines et al., 2005). Because of this work, Dr. Scheines has a courtesy appointment in the Human-Computer Interaction Institute at CMU. Dr. Scheines has won the *Best Paper Award* at both the 1st (Shih, Koedinger, and Scheines 2008) and 6th (Rau, Scheines, Aleven, and Rummel 2013) International Conferences on Educational Data Mining. He hopes to instill the importance of causality in educational data mining during this tutorial and display the power of Tetrad within the Tigris workflow tool.

3.2 Peter Schaldenbrand

PETER SCHALDENBRAND is a developer on the LearnSphere project. He has authored many analysis components in the Tigris workflow tool and has implemented Tetrad's causal search algorithms into Tigris. Mr. Schaldenbrand's research focuses on exploring the causes of learning within educational technology and mining as much useful information from observed educational datasets as possible. He has a passion for increasing the power of pedagogical instruction, lowering the cost of education, and making effective learning tools available to all people. Mr. Schaldenbrand hopes to teach tutorial attendees about the exciting tools he has developed in Tigris and also listen to feedback from attendees to make Tigris an even more effective educational data mining tool.

3.3 Kenneth Koedinger

KENNETH R. KOEDINGER is a professor of Human Computer Interaction and Psychology at Carnegie Mellon University. Dr. Koedinger has an M.S. in Computer Science, a Ph.D. in Cognitive Psychology, and experience teaching in an urban high school. His multidisciplinary background supports his research goals of understanding human learning and creating educational technologies that increase student achievement. His research has contributed new principles and techniques for the design of educational software and has produced basic cognitive science research results on the nature of student thinking and learning.

Dr. Koedinger has a long history of research in the educational data mining domain as well as the causes of learning. Recently he has been researching the importance of resources within online courses. Dr. Koedinger has authored over 180 peer-reviewed publications and has been funded by over 30 grants. He has received many best paper awards including papers at the 1st (Shih, Koedinger, & Scheines 2008) and 5th (Koedinger, McLaughlin, Stamper 2012) International Conferences on Educational Data Mining. Dr. Koedinger is a co-founder of Carnegie Learning, Inc. and leads LearnSphere. DataShop and the Tigris workflow tool were developed under his direction.

4. OBJECTIVES AND OUTCOMES

One of the objectives of this tutorial is to introduce the importance of causal analysis to the educational data mining community. Many papers focus on finding interesting correlations in data such as (San Pedro et al., 2013) and (Slater et al., 2016). These correlations can be fascinating but only give small clues into how to make actionable interventions in curricula, intelligent tutoring systems, or classrooms to increase learning. This tutorial will demonstrate that important causal relationships can be mined from observational data and that these relationships can give insight into real interventions on student learning.

This tutorial will also display the importance of making analyses sharable and introduce easily reproducible workflow capabilities in Tigris. Software versioning, data permissions, and hidden calculations plague the reproducibility of research results. In Tigris, versioning is hidden from the user allowing anyone with a computer to use the tools. Data permissions are handled through the safe infrastructure of DataShop, and all calculations and parameters are visible in the Tigris workflow pipeline. This tutorial will demonstrate that doing research in Tigris is a great way to support reproducibility and shareability.

One outcome of this tutorial is that attendees will be familiar with using data from DataShop in a Tigris workflow. They will request access to a dataset with which they will import into a Tigris workflow under the guidance of the organizers. They will then gain familiarity with the Tetrad algorithms and analyze the results with the organizers.

Once the attendees are familiar with creating causal models in a workflow that the organizers have supervised, they can move onto working on their own research questions. Tutorial attendees will discuss what causal relationships they would like to investigate and try to expand on the use cases of Tetrad. They may present a workflow that utilizes Tetrad and their own datasets to the other attendees. This tutorial will promote the importance of causal reasoning in education with the intention of making productive interventions on student learning.

5. REFERENCES

[1] Jo, Y., Tomar, G., Ferschke, O., Rosé, C. P., & Gašević, D. (2016, April). Pipeline for expediting learning analytics and student support from data in social learning. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 542-543). ACM.

[2] K. Koedinger, J. Kim, J. Jia, E. McLaughlin and N. Bier, Learning is not a spectator sport: Doing is better than watching for learning from a MOOC, in *ACM Conf. Learn at Scale*, 2015.

[3] Koedinger, K.R., McLaughlin, E.A., & Stamper, J.C. (2012). Automated student model improvement. In Yacef, K., Zaïane, O., Hershkovitz, H., Yudelson, M., & Stamper, J. (Eds.), *Proceedings of the 5th International Conference on Educational Data Mining*, pp. 17-24. Chania, Greece.

[4] Koedinger, K. R., Scheines, R., & Schaldenbrand, P. (2018). Is the Doer Effect Robust Across Multiple Data Sets? In *Proc. of the 11th Intl. Conference on Educational Data Mining*. Buffalo, NY.

[5] Pavlik Jr., P. I., Cen, H., & Koedinger, K. R. (2009). Performance Factors Analysis - A New Alternative to Knowledge Tracing. *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, (AIED09) pp. 531-538. Amsterdam, IOS Press.

[6] Rau, M., and Scheines, R. (2012). Searching for Variables and Models to Investigate Mediators of Learning from Multiple Representations, in *Proceedings of the 5th International Conference on Educational Data Mining*

[7] Rau, M., Scheines, R., Aleven, V., and Rummel, N. (2013). Does Representational Understanding Enhance Fluency – or Vice Versa? Searching for Mediation Models. *Proceedings of the 6th International Conference on Educational Data Mining*.

[8] M. San Pedro, R. Baker, A. Bowers and N. Heffernan, Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School, in *Proc. of the 6th International Conf. on Educational Data Mining*, 2013.

[9] R. Scheines, G. Leinhardt, J. Smith and K. Cho, Replacing Lecture with Web-Based Course Materials, *Journal of Educational Computing Research*, vol. 32, no. 1, pp. 1-26, 2005.

[10] Shih, B., Koedinger, K.R., and Scheines, R. (2008). A response time model for bottom-out hints as worked examples. In *Proceedings of the First International Conference on Educational Data Mining*. 117-126

[11] S. Slater, R. Baker, J. Ocumpaugh, P. Inventado, P. Scupelli and N. Heffernan, Semantic Features of Math Problems: Relationships to Student Learning and Engagement, in *Proc. of the 9th International Conf. on Educational Data Mining*, 2016.

[12] P. Spirtes, C. N. Glymour, R. Scheines, Causation, Prediction, and Search, 2nd edition, MIT Press 2000.

[13] Stamper, J., Koedinger, K.R., Baker, R., Skogsholm, A., Leber, B., Demi, S., Yu, S., Spencer, D. (2011) Managing the Educational Dataset Lifecycle with DataShop. In Kay, J., Bull, S. and Biswas, G. (eds). *Proceeding of the 15th International Conference on Artificial Intelligence in Education* (AIED2011)

[14] Veeramachaneni, K., Halawa, S., Dernoncourt, F., O'Reilly, U. M., Taylor, C., & Do, C. (2014). Moocdb: Developing standards and systems to support MOOC data science. *arXiv preprint*. arXiv:1406.2015.

[15] Veeramachaneni, K., O'Reilly, U. M., & Taylor, C. (2014). Towards Feature Engineering at Scale for Data from Massive Open Online Courses. *arXiv preprint*. arXiv:1407.5238.