Variable Construction and Causal Modeling of Online Education Messaging Data: Initial Results

S. FANCSALI
Carnegie Mellon University, USA
Apollo Group, Inc., USA

Preliminary results are described of search for variable constructions from "raw" student messaging data in an online forum. Variable constructions are judged based upon their contribution to a novel measure of "causal predictability" for a target learning outcome using graphical search procedures for causal modeling.

Key Words and Phrases: variable construction; dimensionality reduction; Bayesian networks; causality; causal discovery; online education

1. INTRODUCTION

Research on causal discovery from observational data using graphical causal models (e.g., Spirtes, *et al.* 2000; Pearl 2000) focuses almost entirely on situations in which data are available at an appropriate unit/level of analysis (i.e., well-defined random variables that capture salient features of interest). However, in many social science settings, "raw" data manifests in log files and other sources whose variables are not suited to straight-forward causal interpretations. In online education settings, for example, we might have access to logs of student behaviors, online forum messages, and other aspects of student interactions with courseware. In such cases, the problem of causal discovery is two-fold: in addition to discovering relevant causal relations, we must also discover, from complex, high-dimensional, "raw" data, the very variables that figure in those causal relations.

While there is a great deal of work in statistics and machine learning on dimensionality reduction (feature extraction and selection), little work is aimed at the construction of variables, let alone for causal modeling. Arnold, *et al.* (2006) developed a method for feature discovery with education data, but did not use it for causal modeling. We extend some of those ideas to the problem of causal discovery by simultaneously searching for both suitable constructed variables and resulting graphical causal models. Specifically, we search over different deterministic functions of the underlying "raw" variables and evaluate those constructed variables by informativeness with regards to the underlying causal structure (including suitability for manipulation) in online course environments.

2. DATA AND METHOD

We consider data from 646 learners that used a message forum for an online, graduate-level economics course and consider only raw data from forum logs. With this data, we seek to develop models of the causes of student learning outcomes as measured by instructor-assigned course grades and independently assessed final exams scores. While a plethora of other data sources and (behavioral) causes of these learning outcomes might be considered, the focus on an "impoverished" data set provides an interesting and difficult test of our method.

We iteratively construct and evaluate sets of constructed variables—deterministic functions of base characteristics of student messaging behavior (e.g., length of messages, message counts, etc.). At the first stage, we consider simple functions (max, average, variance, etc.) of these base characteristics and continue at later stages to consider

Author's addresses: S. Fancsali, Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA, USA. Product Strategy and Development, Apollo Group, Inc., San Francisco, CA, USA. E-mail: sfancsal@andrew.cmu.edu

discretizations as well as interactions among them. At each stage, a variety of strategies can be used to prune the large set of possible constructions to a smaller number.

From any particular set of constructed variables, we search for causal structure using the FCI algorithm (Spirtes, *et al.* 2000). FCI can infer the presence of unmeasured (or "latent") confounders of observed variables, which are almost certainly present given our (deliberately) "impoverished" data. FCI outputs a partial ancestral graph (PAG), which encodes all of the causal models consistent with the observed data. We deploy a novel method to assess the value of this PAG, and thus the set of constructed variables. We first enumerate all of the directed acyclic graphs (DAGs) that are compatible with the causal explanation provided by the PAG. From each DAG, we build a regression model for a target variable (i.e., learning outcome) based upon the target's parents in that DAG. We then assess the set of constructed variables by the average R^2 value over all of these regression models (i.e., assuming every DAG represented by the PAG is equiprobable). We seek the set of constructed variables that maximizes "causal predictability" as assessed by this "average causal R^2" value.

3. RESULTS

Table I provides a comparison of average causal R^2 values for constructed variables found via the above search procedure compared to those achieved using ad hoc variables previously specified in a pilot study for this data (Fancsali, forthcoming). With this novel analysis, we find substantial improvement in the acquisition of causal knowledge for both of the measured learning outcomes. FCI search over ad hoc constructed variables and target learning outcomes indicates that any relationships between the two are confounded by unmeasured common causes, thus providing no "causally predictive" information (i.e., causal R^2 values of 0.0) with respect to these learning outcomes. Importantly, this new method unambiguously learns that appropriate constructed variables are direct causes of the course grade, recovering aspects of known "ground truth" for the instructed-assigned grade. Search for constructed variables is a ripe area for future research and will likely have many fruitful applications for both predictive and causal modeling. Future work will deploy similar search on data with richer semantic content (e.g., data from intelligent tutoring systems) to ideally lead to more causally interpretable variable constructions.

Table I. Average Causal R^2 Values for Ad Hoc Constructed Variables vs. Constructed Variables from Search

	course grade	**final exam**
ad hoc construction	0.0	0.0
variable construction search	0.37	0.13

REFERENCES

ARNOLD, A., BECK, J., AND SCHEINES, R. 2006. Feature Discovery in the Context of Educational Data Mining: An Inductive Approach. *Proceedings of the AAAI2006 Workshop on Educational Data Mining*, Boston, MA, 7-13.

FANCSALI, S. forthcoming. Variable Construction for Predictive and Causal Modeling of Online Education Data. *Proceedings of the First International Conference on Learning Analytics and Knowledge.* Banff, Alberta, Canada.

PEARL, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge UP.

SPIRTES, P., GLYMOUR, C., AND SCHEINES, R. 2000. *Causation, Prediction, and Search*. Second Edition. Cambridge, MA: MIT.