Automatic Identification of Quasi-Experimental Designs for Scientific Discovery

David Jensen, Andrew Fast, Brian Taylor, Marc Maier, and Matthew Rattigan

Knowledge Discovery Laboratory
Department of Computer Science
University of Massachusetts Amherst
{jensen,afast,btaylor,maier,rattigan}@cs.umass.edu

Abstract

We briefly describe recent research on the automatic identification of quasi-experimental designs, a family of methods used in the medical, social, and economic sciences to discover causal knowledge from observational data. These methods are widely used for manual discovery, but recent advances in knowledge representation and databases have made it possible to automate aspects of their use. We report on a prototype system for automatically identifying quasi-experimental designs and suggest future work.

Introduction

Quasi-experimental designs (QEDs) specify conditions under which causal knowledge can be inferred from observational data. By applying a QED, an investigator can identify and exploit fortuitous conditions in non-experimental data that emulate the conditions of intentional experiments. QEDs are widely used in medicine, social science, and economics, and they represent an important class of methods for scientific discovery.

However, the potential to automate this class of methods has never been systematically studied. Until very recently, identifying and exploiting QEDs has been an exclusively manual activity limited to a relatively small number of trained researchers.

In a recent paper, we have provided the first demonstration that QEDs can be identified algorithmically. Our system—Automated Identification of Quasi-experiments (AIQ)—applies a standard Prolog theorem-prover to a knowledge base expressed in first-order logic that represents the schema of a given domain, prior knowledge about causal dependencies in that domain, and the necessary and sufficient conditions for QEDs. AIQ is a proof-of-concept that larger and more capable systems could automatically identify QEDs in a wide range of observational data sets, identifying "natural experiments" that would otherwise go unrecognized and unexploited.

Copyright © 2008, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Why Discover Causal Knowledge?

Quasi-experimental designs support inferences about causal knowledge. By *causal knowledge*, we mean the assertion that manipulating one variable will make another vary. This dependence can be probabilistic and involve intervening variables, but it goes beyond mere association to represent the expected outcomes of specific actions.

Hypotheses about causality are one important class of scientific hypotheses. Other classes include existential hypotheses (e.g., there exists an element with atomic mass 68), compositional hypotheses (atoms consist of protons, neutrons, and electrons), and associational hypotheses (smoking and cancer are correlated). Causal hypotheses are typically valued more highly than associational hypotheses because causation implies association, whereas association does not necessarily imply causation.

Inferring that A causes B requires meeting three conditions: (1) establishing *statistical association* between the values of A and B; (2) establishing the *direction* of causality, if it exists, from A to B (e.g., based on temporal criteria); and (3) eliminating the effects of all potential *common causes* of A and B.

Eliminating common causes is a key focus of experimental and quasi-experimental designs. These designs typically employ control, randomization, or modeling to eliminate common causes. *Control* holds the values of potential common causes constant, so they affect neither potential causes (often called "treatments") nor potential effects ("outcomes"). *Randomization* assigns subjects to treatments randomly, so that potential common causes cannot systematically affect outcomes. *Modeling* attempts to statistically estimate the effects of common causes so they can be factored out of assessments of association and so that any remaining association between treatment and outcome indicates causal dependence.

Control and randomization require direct manipulation of the conditions under which data are generated and thus cannot be applied to observational data. This leaves only modeling, an approach that has been actively pursued in statistics (e.g., Holland & Rubin 1988), artificial intelli-

gence (Pearl 2000), and philosophy (Spirtes, Glymour, Scheines 2000). While the developments to date in modeling have been impressive, and new developments continue, the problem of causal discovery is far from solved. Problems of correct model specification, latent variables, computational tractability, and high sample complexity continue to constrain real-world applications of modeling.

Quasi-Experimental Designs

QEDs are a family of methods for exploiting fortuitous situations in observational data that emulate control and randomization (Campbell & Stanley 1963; Shadish, Cook, and Campbell 2002). QEDs are templates for causal inference that increase the statistical power of inferences by selecting subsets of data that reduce or eliminate some common causes. QEDs bring back a form of control and randomization to the analysis of non-experimental data, and thus expand the range of approaches that can be applied to such data.

There are many examples of QEDs, including: (1) *twin designs* which control the values of some potential common causes within specified pairs of data instances; (2) *non-equivalent control group designs* which compare temporal responses of treated instances to a control group of similar untreated instances; (3) *regression discontinuity designs* which use cases in which treatment is assigned entirely based on the value of a single known variable.

QEDs are particularly useful when experiments are deemed unethical (e.g., studies of smoking and cancer in humans), when experiments are impractical (studies of how increased state penalties affect drunk driving), or when data have already been collected for other purposes and researchers want an inexpensive precursor to potential future experiments.

Automatic Identification

Identifying opportunities to apply QEDs is currently a painstaking manual process. It requires highly specific knowledge of the domain, the available data, and QEDs themselves. Despite the wide use of QEDs, many opportunities to apply these methods for causal discovery still go unrecognized.

Fortunately, recent developments in the technologies and applications of databases and machine learning present new opportunities for automating the discovery of QEDs. First, the increasing complexity of the relational and temporal structure of databases provides the necessary scope for application of QEDs. Second, the increasing size of databases provides the ability to identify subsets of data with the necessary statistical power for valid causal inferences. Finally, recently devised relational and temporal knowledge representations provide the ability to explicitly represent the causal knowledge necessary to drive the identification of QEDs (Jensen 2008).

We have developed AIQ, a prototype system that automatically identifies QEDs (Jensen et al. 2008). [1] It takes input in the form of a standard entity-relationship diagram annotated with temporal extents and frequencies, as well as any existing domain knowledge about known causes. It produces output in the form of a specification of a QED, including a treatment variable, an outcome variable, and the set of records that form the units (data instances) necessary to conduct a hypothesis test.

AIQ searches a space of potential temporal streams and data instances constructed from the existing tables in a relational database. It matches potential treatments, outcomes, and units to the specifications of one common QED (the non-equivalent control group design) and outputs valid specifications. These instantiated designs can then be validated by human investigators and applied if valid.

Future Work

We are currently pursuing several lines of additional research. First, we are examining all known designs to identify their necessary and sufficient conditions and determine how to formally represent the knowledge necessary to automatically identify them. Second, we are characterizing the conditions under which QEDs provide advantages over existing modeling-based methods for causal discovery. Third, we are pursuing a range of applications of QEDs to provide case studies and examples.

References

Campbell, D.T, and Stanley, J.C. 1963. *Experimental and Quasi-Experimental Designs for Research*. Rand McNally.

Holland, P., and Rubin, D. 1988. Causal inference in retrospective studies. *Evaluation Review* 12: 203–231.

Jensen, D; Fast, A.; Taylor, B.; and Maier, M. 2008. Automatic identification of quasi-experimental designs for discovering causal knowledge. *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 372–380.

Jensen, D.D. 2008. Beyond prediction: Directions for probabilistic and relational learning. *17th International Conference on Inductive Logic Programming*. 4-21.

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

Shadish, W.R.; Cook, T.D.; and Campbell, D.T. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Spirtes, P.; Glymour, C.; Scheines, R. 2000. *Causation, Prediction, and Search, 2nd ed.* Cambridge, MA: MIT Press.

[1] Source code for AIQ is available at: http://kdl.cs.umass.edu/causality/index.html