Development Projects for the Causality Workbench

Isabelle Guyon
Clopinet
955 Creston Road
Berkeley, CA 94708, USA

Jean-Philippe Pellet
IBM Zurich Research Lab
Sumerstrae 4, Postfach
8803 Rschlikon, Switzerland

Alexander Statnikov
New York University
545 First Avenue
New York, NY 10016, USA

Abstract

The Causality Workbench project provides an environment to test causal discovery algorithms. Via a web portal (`http://clopinet.com/causality`), we provide a number of resources, including a repository of datasets, models, and software packages, and a virtual laboratory allowing users to benchmark causal discovery algorithms by performing virtual experiments to study artificial causal systems. We regularly organize competitions. In this paper, we explore the opportunities offered by development applications.

Introduction

Uncovering cause-effect relationships is central in many aspects of everyday life in both highly industrialized and development countries: what affects our health, the economy, climate changes, world conflicts, and which actions have beneficial effects? Establishing causality is critical to guiding policy decisions in areas including medicine and pharmacology, epidemiology, climatology, agriculture, economy, sociology, law enforcement, and manufacturing. Urgent concerns include food supply/famine, and the spread of crop diseases.

One important goal of causal modeling is to predict the consequences of given *actions*, also called *interventions*, *manipulations* or *experiments*. This is fundamentally different from the classical machine learning, statistics, or data mining setting, which focuses on making predictions from observations. Observations imply no manipulation on the system under study whereas actions introduce a disruption in the natural functioning of the system. In the medical domain, this is the distinction made between "diagnosis" (prediction from observations) and "treatment" (intervention). For instance, smoking and coughing might be both predictive of respiratory disease and helpful for diagnosis purpose. However, if smoking is a cause and coughing a consequence, acting on the cause (smoking) can change your health status, but not acting on the symptom or consequence (coughing). Thus it is extremely important to distinguish between causes and consequences to predict the result of actions like predicting the effect of forbidding smoking in public places.

The need for assisting policy making while reducing the cost of experimentation and the availability of massive amounts of "observational" data prompted the proliferation of proposed causal discovery techniques (Glymour and Cooper 1999; Pearl 2000; Spirtes, Glymour, and Scheines 2000; Neapolitan 2003; Koller and Friedman 2009), but it is fair to say that to this day, they have not been widely adopted by scientists and engineers. Part of the problem is the lack of appropriate evaluation and the demonstration of the usefulness of the methods on a range of pilot applications. To fill this need, we started a project called the "Causality Workbench", which offers the possibility of exposing the research community to challenging causal problems and disseminating newly developed causal discovery technology. In this paper, we outline our setup and methods and the possibilities offered by the Causality Workbench to solve development problems.

What are "causal problems"?

Causal discovery is a multi-faceted problem. The definition of causality itself has eluded philosophers of science for centuries, even though the notion of causality is at the core of the scientific endeavor and also a universally accepted and intuitive notion of everyday life. But, the lack of broadly acceptable definitions of causality has not prevented the development of successful and mature mathematical and algorithmic frameworks for inducing causal relationships.

The type of causal relationships under consideration have often been modeled as causal Bayesian networks or structural equation models (SEM) (Pearl 2000; Spirtes, Glymour, and Scheines 2000; Neapolitan 2003). In the graphical representation of such models, an arrow between two variables $A \rightarrow B$ indicates the direction of a causal relationship: A causes B. A node in of the graph, labeled with a particular variable X, represents a mechanism to evaluate the value of X given the parent node variable values. For Bayesian networks, such evaluation is carried out by a conditional probability distribution $P(X|Parents(X))$ while for structural equation models it is carried out by a function of the parent variables and a noise model. Learning a causal graph can be thought of as a model selection problem: Alternative graph architectures are considered and a selection is performed, either by ranking the architectures with a global score (*e.g.,* a marginal likelihood, or a penalty-based cost function), or by retaining only graphs, which fulfill a number of constraints such as dependencies or independencies between subsets of variables. Bayesian networks and SEMs provide a convenient language to talk about the type of problem we are interested in, but we made an effort to design tasks, which do not preclude of any particular model. Our objective is not to reduce causality to a simple or convenient definition or to a family of models, which may induce simplifying assumptions that are either restrictive or unnecessary (*e.g.,* discrete variables, Gaussian distributions, linear effects, or no unobserved common causes), but rather to define tasks with clear objectives and give ourselves means of assessing how well these objectives are reached.

In designing our first benchmark tasks we have focused on some specific aspects of causal discovery:

Causality between random variables. We have so far addressed mostly causal relationships between random variables, as opposed to causal relationships between events, or objects.

Multivariate problems. Many early efforts in causal studies have concentrated on the study of cause-effect relationships between a few variables. The availability of large observational datasets with thousands of recorded variables (in genomic studies with microarray data, in pharmacology with high throughput screening, in marketing with logs of internet customers, etc.) has drawn our attention to multivariate problems in which an array of eventually weak causes might influence an outcome of interest, called "target". Conversely, the study of causality between too few variables (*e.g.,* just two) is also a challenge since tests of conditional independence using covariates are used by many causal discovery algorithms.

Time dependency. Our everyday-life concept of causality is very much linked to time dependencies (causes precede their effects). However, many machine learning problem are concerned with stationary systems or "cross-sectional studies", which are studies where many samples are drawn at a given point in time. Thus, sometimes the reference to time is replaced by the notion of "causal ordering". Causal ordering can be understood as fixing a particular time scale and considering only causes happening at time t and effects happening at time $t + \delta t$, where δt can be made as small as we want. In practice, this means that the samples in our various training and test sets are drawn independently, according to a given distribution, which changes only between training and test set versions.[1] We are offering tasks with or without time dependencies.

Learning from observational or experimental data. We call *observational data*, data collected from a system let to evolve according to its own dynamics. In contract, *experimental data* is obtain as a result of interventions on the system of interest by an external agent who disrupts the system by imposing values to certain variables. Generally, experimenting is the only way to ascertain causal relationships. However, in many domains, experimenting is difficult and costly compared to collecting observational data. Hence, we have investigated settings in which only observational data are available for training. The tasks we collected also include settings in which both observational and experimental data are available.

We have so far mostly addressed two tasks of interest:

1. **Predicting the consequences of manipulations.** In one challenge we organized, our data included training samples drawn from a "natural" pre-manipulation distribution and test data drawn from various post-manipulation distributions (in which the values of a subset of variables has been set to given values by an external agent, bypassing the natural functioning of the system). The objective was to predict withheld values of a target variable, given a the values of a set of observed or manipulated variables.
2. **Discovering causal structures.** Causal graphs (*e.g.,* Bayesian networks or Structural Equation Models) are powerful to represent mechanisms at a level sufficient to reason and plan for future actions. A common exercise is to investigate whether the structure of such models can be reconstructed from artificial data generated by the models, in an effort to reassure ourselves that structures generated from real data may be meaningful.

The first task has a clear objective and it does not preclude of any particular modeling technique. In particular, it is not required to produce a causal graph. Operational definitions of causality (Glymour and Cooper 1999) use the notion of manipulation to evidence cause-effect relationships. Hence, predicting the consequences of manipulations is a "causal question" that can serve to evaluate causal models against non-causal models. The second task is more explicitly "causal", but evaluating its solutions on real data requires knowledge on the data generating systems, which we usually do not have in practice. There are many other causal questions, which we will progressively address (Guyon et al. 2010).

The Causality Workbench project

Our effort has been gaining momentum with the organization of two challenges, which each attracted over 50 participants. The first causality challenge we have organized (Causation and Prediction challenge, December 15 2007 - April 30 2008) allowed researchers both from the causal discovery community and the machine learning community to try their algorithms on sizable tasks of real practical interest in medicine, pharmacology, and sociology (see `http://www.causality.inf.ethz.ch/challenge.php`). The goal was to train models exclusively on observational data, then make predictions of a target variable on data collected after intervention on the system under study were performed. This first challenge reached a number of goals that we had set to ourselves: familiarizing many new researchers and practitioners with causal discovery problems and existing tools to address them, pointing out the limitations of current methods on some particular difficulties, and fostering the development

[1] When manipulations are performed, we must specify whether we sample from the distribution before or after the effects of the manipulation have propagated. Here we assume that we sample after the effects have propagated.

of new algorithms. The results indicated that causal discovery from observational data is not an impossible task, but a very hard one and pointed to the need for further research and benchmarks (Guyon et al. 2008). The Causal Explorer package (Aliferis et al. 2003), which we had made available to the participants and is downloadable as freeware, proved to be competitive and is a good starting point for researchers new to the field. It is a Matlab (R) toolkit supporting "local" causal discovery algorithms, efficient to discover the causal structure around a target variable, even for a large number of variables. The algorithms are based on structure learning from tests of conditional independence, as all the top ranking methods in this first challenge.

The first challenge explored an important problem in causal modeling, but is only one of many possible problem statements. The second challenge called "competition pot-luck" aimed at enlarging the scope of causal discovery algorithm evaluation by inviting members of the community to submit their own problems and/or solve problems proposed by others. The challenge started September 15, 2008 and ended November 20, 2008, see `http://www.causality.inf.ethz.ch/pot-luck.php`. One task proposed by a participant drew a lot of attention: the cause-effect pair task. The problem was to try to determine in pairs of variables (of known causal relationships), which one was the cause of the other. This problem is hard for a lot of algorithms, which rely on the result of conditional independence tests of three or more variables. Yet the winners of the challenge succeeded in unraveling 8/8 correct causal directions (Zhang and Hyvärinen 2009).

A number of lessons were drawn from these first evaluations:

- From an algorithmic perspective, we stubbled on the multivariate problem. Moving from a multivariate variable selection task in an *i.i.d.* setting (Guyon et al. 2006) to a similar task in a non *i.i.d.* setting (Guyon et al. 2008) magnified the problem of "overfitting". This problem is familiar to machine learning scientists: in the *i.i.d.* setting, multivariate algorithms struggle to outperform univariate algorithms (selecting variables for their individual predictive power). Additionally, in the *causation and prediction challenge*, "causal" variable selection algorithms struggled to outperform non-causal algorithms. Hence it is important to match well tasks to methods: find causal problems, which truly benefit from the causal discovery arsenal and admit that some causal problems may as well be solved by traditional statistical methods.
- From a methodology perspective, we realized that learning causal relationships reliably from observational data only may not be realistic. Experiments are needed to firm up hypotheses made by analyzing observational data. This is particularly critical in a multivariate setting where errors cumulate and propagate. Our Virtual Lab, briefly described in the last section, lets researchers conduct virtual experiments to benchmark hybrid techniques capitalizing both on observational data and designed experiments.
- From a practical point of view, we learnt that for many applications focusing on a particular target variable, it is more important to rank candidate causes of the target in order of potential impact on the target than to unravel the overall causal structure of the covariates. This may require developing entirely new approaches.
- Finally, from a challenge design perspective, we discovered that there is a lot of value in offering to the participants the possibility of contributing problems, which can evidence the power of causal discovery algorithms (Guyon, Janzing, and Schölkopf 2009). We intend to organize new "pot-luck challenges" in which the participants can contribute problems. This type of challenge may be a nice forum of development centric problems.

Part of our benchmarking effort is dedicated to collecting problems from diverse application domains. Via the organization of competitions, we have successfully channeled the effort or dozens of researchers to solve new problems of scientific and practical interest and identified effective methods. However, competition without collaboration is sterile. Recently, we have started introducing new dimensions to our effort of research coordination: stimulating creativity, collaborations, and data exchange. We are organizing regular teleconference seminars. We have created a data repository for the Causality Workbench already populated by 15 datasets. All the resources, which are the product of our effort, are freely available on the Internet at `http://clopinet.com/causality`.

Causal problems in the development world

Causal structure learning is highly relevant in development issues, as there are several domains where there is a lot of data, a limited understanding of the causal relationships involved, and a motivation to make predictions under interventions on some of the variables. We have identified a number of areas of interest:

- **Epidemiology – Preventing new pandemics**: The recent Mexican pandemic flu has reminded us that we are still vulnerable to the burst and spread of new diseases, which are difficult to keep under control. Epidemiology has long been one of the main areas of application of causal modeling (Rubin 1974; Herskovits and Dagher 1997; J.M. Robins 2000). Epidemiologists have also embraced the new tools of genomics and proteomics to investigate gene-environment interactions (Vinei and Kriebel 2006; Jenab et al. 2009).
- **Agriculture – Food supply and famine avoidance**: The World Health Organization estimates that one-third of the world is well-fed, one-third is under-fed one-third is starving. Of the many factor potentially affecting the availability and the price of food (including climatic variation, regional conflicts, reserves and supply from other parts of the world), which ones should be the focus of attention and what policy should be put in place are important causal questions? Crop yield optimization is one of the oldest areas of causal studies, which gave birth to the first formal mathematical methodology for designing experiments by Fisher, in his book "The Design of Experiments" (1935). The problem of supply an demand also

touches to econometrics, an active area of causal studies (late Prof. Clive Granger received the 2003 Nobel prize in economics for his causality-related work).

- **Sociology – Violence, law enforcement, conflict management**: Violence is an important obstacle to development. Can we identify the causes of social conflicts and crises and prevent them? There is also a long history of use of causal models in social sciences, and particularly structural equation models (SEM) (Haavelmo 1943)[2].
- **Ecology – Water supply and environment preservation, desertification, deforestation**: According to (Geist 2005) desertification has three major types causes: meteorological (precipitation variations, atmospheric dust, air temperature, elevated atmospheric CO_2), ecological (nutrient cycling, plan growth/regeneration/mortality, microbial dynamics, plant cover, herbivory life cycles, evapotranspiration), and human or socio-economic (loss of habitat, overexploitation, spread of exotic pests and weeds, pollution, climate changes). Can we contribute to identifying causative mechanisms of desertification?
- **Humanities – Education, culture and language preservation**: The future of development countries rests to a large extent on education. Can we identify the education deficiencies and their causes as well as the causes of the disappearance of local know-how and cultural identity. For instance, according to UNESCO there are around 6000 languages spoken worldwide today, but half of the world's population communicates in only 8 languages and more than 3000 languages are now spoken by fewer than 10000 people.

The key to involving the AAAI community in solving these problems will be to identify useful databases and make them available in a friendly format. We count on the collaboration of other interested researchers to make this happen as there are many source of data publicly available, which we could potentially use.

The World Health Organization (WHO)[3], the United Nations (UN)[4] and many other institutions are making available a wealth of statistics, which lend themselves to statistical modeling. Our challenge will include integrating data from various sources. For example, for the famine problem, one may need to rely upon satellite data,[5], geographical conflict information,[6], and general statistics from WHO, UN, etc.

Validating methodologies and findings is challenging, but can be done both by studying retrospective data and comparing predictions against the results of simulations. Realistic simulators, such as the epidemiology simulator STEM[7] are valuable resources.

Finally, we also have information available, which can be used to identify the most effective actions[8]. These data can be used for learning from experimental data.

On-going and planned activities

Our current challenge on "Active Learning" (`http://clopinet.com/al`), incorporates several datasets on applications of interest to the development community, including:

- **Pharmacology:** Discovery of molecules and mechanisms of action against the HIV virus and tuberculosis.
- **Ecology:** Identification of vegetation cover from remote sensing.
- **Culture preservation:** Electronic annotation of ancien arabic manuscripts.

The website of the challenge will remain open after the competition ends and we will encourage researchers from the AAAI Spring Symposium on Artificial Intelligence for Development to use it to gain familiarity with causal tasks.

Methods for learning cause-effect relationships without experimentation (learning from observational data) are attractive because observational data is often available in abundance and experimentation may be costly, unethical, impractical, or even plain impossible. Still, many causal relationships cannot be ascertained without the recourse to experimentation and the use of a mix of observational and experimental data might be more cost effective. We implemented a *Virtual Lab* allowing researchers to perform experiments on artificial systems to infer their causal structure. The design of the platform is such that:

- Researchers can submit new artificial systems for others to experiment.
- Experimenters can place queries and get answers.
- The activity is logged.
- Registered users have their own virtual lab space.

We have released a first version `http://www.causality.inf.ethz.ch/workbench.php`.
We plan to attach to the virtual lab sizeable realistic simulators such as the Spatiotemporal Epidemiological Modeler (STEM), an epidemiology simulator developed at IBM, now publicly available: `http://www.eclipse.org/stem/`.

Our planned challenge ExpDeCo (Experimental Design in Causal Discovery) will benchmark methods of experimental design in application to causal modeling. The goal

[2]As an indication, the structural equation model journal was ranked 1^{st} in Social Sciences and Mathematical Methods in 2009 by Thomson Reuters Journal Citation (R).

[3]World Health Statistical Information System: `http://www.who.int/whosis/en/index.html`.

[4]United Nations food and agriculture statistics: `http://www.fao.org/economic/ess/en/`.

[5]NASA satellite data to predict famine: `http://www.scientificblogging.com/news/using_nasa_satellites_to_predict_famine`.

[6]Nobel prize foundation conflict maps: `http://nobelprize.org/educational_games/peace/conflictmap/`.

[7]Model of the Mexican Pandemic publicly available for the STEM simulator: `http://www.eclipse.org/stem/download_sample.php?file=UsaMexicoDemo.zip`.

[8]For instance the web sites rating charities like `http://www.charitynavigator.org/`.

will be to identify effective methods to unravel causal models, requiring a minimum of experimentation, using the Virtual Lab. A budget of virtual cash will be allocated to participants to "buy" the right to observe or manipulate certain variables, manipulations being more expensive that observations. The participants will have to spend their budget optimally to make the best possible predictions on test data. This setup lends itself to incorporating problems of relevance to development projects, in particular in medicine and epidemiology where experimentation is difficult while developing new methodology.

We are planning another challenge called CoMSICo for "Causal Models for System Identification and Control", which is more ambitious in nature because it will perform a continuous evaluation of causal models rather than separating training and test phase. In contrast with ExpDeCo in which the organizers will provide test data with prescribed manipulations to test the ability of the participants to make predictions of the consequences of actions, in CoMSICo, the participants will be in charge of making their own plan of action (policy) to optimize an overall objective (e.g., improve the life expectancy of a population, improve the GNP, etc.) and they will be judged directly with this objective, on an on-going basis, with no distinction between "training" and "test" data. This challenge will also be via the Virtual Lab. The participants will be given an initial amount of virtual cash, and, as previously, both actions and observations will have a price. New in CoMSICo, virtual cash rewards will be given for achieving good intermediate performance, which the participants will be allowed to re-invest to conduct additional experiments and improve their plan of action (policy). The winner will be the participant ending up with the largest amount of virtual cash.

While some of our core activities focus on benchmarking algorithms, we realize that discovery and problem-solving are critical aspects, which cannot always be quantitatively evaluated on the short run, yet need immediate attention from the research community. With the collaboration of other researchers we will enroll at the AAAI Spring Symposium on Artificial Intelligence for Development, we intend to populate our repository with development-related problems. We intend to involve interested participants in the organization of a new "pot-luck challenges", specifically on development problems.

Conclusion

Our program of data exchange and benchmark will make available to the research community problems that tie into development projects in various application domains. Causal discovery is a problem of fundamental and practical interest in many areas of science and technology and there is a need for assisting policy making in all these areas while reducing the costs of data collection and experimentation. Hence, the identification of efficient techniques to solve causal problems will have a widespread impact. Our activities, such as teleconference seminars, data and tool exchange, competitions and post-competition collaborative experiments, will cement collaborations between researchers and ensure a rapid and broad dissemination of the methods and results.

Acknowledgments

This project is an activity of the Causality Workbench supported by the Pascal network of excellence funded by the European Commission and by the U.S. National Science Foundation under Grant N0. ECCS-0725746. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We are very grateful to all the members of the causality workbench team for their contribution to organizing past challenges and in particular to our co-founders Constantin Aliferis, Greg Cooper, André Elisseeff, and Peter Spirtes.

References

Aliferis, C. F.; Tsamardinos, I.; Statnikov, A.; and Brown, L. 2003. Causal explorer: A probabilistic network learning toolkit for biomedical discovery. In *2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS)*. Las Vegas, Nevada, USA: CSREA Press.

Geist, H. 2005. *The causes and progression of desertification*. Ashgate Studies in Environmental Policy and Practice.

Glymour, C., and Cooper, G., eds. 1999. *Computation, Causation, and Discovery*. Menlo Park, California, Cambridge, Massachusetts, London, England: AAAI Press/The MIT Press.

Guyon, I.; Gunn, S.; Nikravesh, M.; and Editors, L. Z. 2006. *Feature Extraction, Foundations and Applications*. Studies in Fuzziness and Soft Computing. With data, results and sample code for the NIPS 2003 feature selection challenge. Physica-Verlag, Springer.

Guyon, I.; Aliferis, C.; Cooper, G.; Elisseeff, A.; Pellet, J.-P.; Spirtes, P.; and Statnikov, A. 2008. Design and analysis of the causation and prediction challenge. In *JMLR W&CP*, volume 3, 1–33.

Guyon, I.; Aliferis, C.; Cooper, G.; Elisseeff, A.; Pellet, J.-P.; Spirtes, P.; and Statnikov, A. 2010. Causality workbench. In Illaria, P. M.; Russo, F.; and Williamson, J., eds., *Causality in the Sciences*. Oxford University Press, (to appear).

Guyon, I.; Janzing, D.; and Schölkopf, B. 2009. Causality: Objectives and assessment. In *JMLR W&CP*, volume 6, 1–42.

Haavelmo, T. 1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11.

Herskovits, E. H., and Dagher, A. P. 1997. Application of Bayesian networks to health care.

Jenab, M.; Slimani, N.; Bictash, M.; Ferrari, P.; and Bingham, S. 2009. Biomarkers in nutritional epidemiology: applications, needs and new horizons. *Human Genetics*.

J.M. Robins, M.A. Hernan, B. B. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* 11(5):550–560.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Neapolitan, R. E. 2003. *Learning Bayesian Networks*. Prentice Hall series in Artificial Intelligence. Prentice Hall.

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

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688–701.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. Cambridge, Massachusetts, London, England: The MIT Press.

Vinei, P., and Kriebel, D. 2006. Causal models in epidemiology: past inheritance and genetic future. *Environ Health* 5(21).

Zhang, K., and Hyvärinen, A. 2009. Distinguishing causes from effects using nonlinear acyclic causal models. In *JMLR W&CP*, volume 6.