Statistical Science
2019, Vol. 34, No. 1, 77–81
https://doi.org/10.1214/18-STS690
Main article: https://doi.org/10.1214/18-STS667

Comment: Strengthening Empirical Evaluation of Causal Inference Methods

David Jensen

Abstract. This is a contribution to the discussion of the paper by Dorie et al. (*Statist. Sci.* **34** (2019) 43–68), which reports the lessons learned from 2016 Atlantic Causal Inference Conference Competition. My comments strongly support the authors' focus on empirical evaluation, using examples and experience from machine learning research, particularly focusing on the problem of algorithmic complexity. I argue that even broader and deeper empirical evaluation should be undertaken by the researchers who study causal inference. Finally, I highlight a few key conclusions that suggest where future research should focus.

Key words and phrases: Causal inference, empirical evaluation, machine learning, algorithmic complexity, constructed observational studies, alignment.

1. INTRODUCTION

It is a great pleasure to read and comment on the paper by Dorie et al. (2019) summarizing lessons learned from the 2016 ACIC Competition. Competitions such as this are extremely difficult to plan and conduct. They require a vast amount of work, much of which can go unrecognized and unrewarded. The organizers deserve an enormous amount of credit for putting together such a well-run and comprehensive competition, as well as completing an extensive post-competition analysis and write up, which can be challenging to complete in the aftermath of the competition itself.

The authors designed an evaluation approach with several key characteristics. First, it evaluates methods based on their empirical performance on realistic data rather than their theoretical properties or their performance on entirely simulated data. Clearly, some aspects of the data are simulated, but these are kept to a minimum. Second, the evaluation employs a broad array of possible treatment assignment mechanisms and response surfaces, along with variation in other key aspects (e.g., functional forms and alignment). Third, the evaluation uses either black-box methods in which entrants have no role in applying the method or blind do-it-yourself (DIY) methods in which entrants do not have access to correct answers. A final element of the evaluation—structuring it as an explicit competition—motivated participation and ensured that even the DIY entries were blind.

My comments largely attempt to position the competition within a larger context, both in terms of how the research community should direct future evaluation efforts and in terms of the conclusions that readers should draw from the competition results. My comments also reflect my primary research communities—computer science in general and machine learning in particular—and attempt to convey a few hard-won lessons from these communities.

Specifically, my comments focus on three points. First, they support the authors' focus on empirical evaluation, using examples and experience from machine learning research. Second, they argue that even broader and deeper empirical evaluation should be undertaken by the entire research community. Finally, they highlight a few key conclusions that suggest where future research should focus.

David Jensen is Professor of Computer Science and Director of the Computational Social Science Institute, University of Massachusetts Amherst, Amherst, Massachusetts 01003, USA (e-mail: jensen@cs.umass.edu; URL: https://people.cs.umass.edu/~jensen/).

2. THE VALUE OF EMPIRICAL EVALUATION

It is challenging to design and conduct broad and empirical evaluations of methods for causal inference, and few examples of such evaluations exist. Given this, how much should we value them? The authors provide several reasons that we should value them highly, noting that individual researchers typically perform evaluations that: (1) compare relatively few methods and in potentially unfair ways; (2) employ evaluations that are not calibrated to realistic conditions; and (3) fall prey to the "file drawer effect" (in which a study's results influence whether it is published). Thus, the authors note, "[s]trong performance of a method in a paper written by its inventor is encouraging but should be interpreted cautiously..." I strongly concur.

However, another class of advantages is not made explicit in the paper. Causal inference has grown increasingly algorithmic (e.g., using complex methods for estimating non-parametric models), and the field now faces challenges similar to those long faced by researchers in many areas of computer science. One such challenge is algorithmic complexity. As the complexity of a given algorithm grows, its expected performance can become extremely difficult to estimate in theory, and thus empirical evaluation becomes correspondingly more useful. A related challenge is implementation detail. While the formal *specification* of a given algorithm may have desirable theoretical properties, liberties are often taken during implementation that can dramatically affect performance. Potential users care more about the performance of the *implementation*, rather than the specification, further increasing the utility of empirical evaluations. A final challenge, well-known to both statisticians and computer scientists, is how this complexity interacts with real data. While theoretical analyses of algorithm performance may be possible with idealized input data, many such estimates do not survive first contact with real data. This, too, increases the utility of broad and empirical evaluations.

Algorithmic complexity also increases the scope for potential human biases. Inventors of new methods face a vast array of design choices in any sufficiently complex method, presenting them with what has been called a "garden of forking paths" (Gelman and Loken, 2013), only one of which is typically reported in a final paper even though large numbers of the paths may have been implicitly or explicitly tested using some limited collection of data sets. The authors allude to this problem, and it is worth reinforcing two of its implications. First, it increases the utility of evaluations such as the 2016 ACIC Competition because they amount to a type of pre-registration (van't Veer and Giner-Sorolla, 2016), and thus we should value their conclusions over alternatives. Second, given the multiplicity of methods tested, we should focus on *general* conclusions about the types of methods that perform well, rather than treating the Competition as a bake-off and focusing on the individual methods that perform best (as the authors themselves note).

For all of these reasons, broad and empirical evaluation strategies have long been emphasized in the field of machine learning (ML). Indeed, they are virtually required by most ML reviewers when confronted with a new modeling method. For certain classes of tasks, the ML research community has responded to this need with large suites of testing resources, including large repositories of data sets (e.g., Dheeru and Karra Taniskidou, 2017) and standard evaluation protocols (e.g., Provost et al., 1998, Bradley, 1997). Some of the best papers in the field routinely evaluate performance across a dozen or more data sets (a classic example is Domingos and Pazzani, 1997). These types of evaluations regularly yield surprises in machine learning research. Two longstanding examples that are frequently cited are the effectiveness of naive Bayesian classifiers (Domingos and Pazzani, 1997) and the analogous effectiveness of relatively simple bag-of-words models in information retrieval tasks (Lewis, 1998).

Of course, evaluating methods for causal inference (as opposed to, say, classification) poses special challenges, and the development of the resources necessary for broad and empirical evaluation is only just beginning. Perhaps the greatest challenge is that, in contrast to classification or regression tasks, the "ground truth" for causal inference requires knowledge of a different joint distribution than the one that generated the (observational) training data. Fortunately, the 2016 ACIC Competition joins several other recent efforts to develop such evaluation resources. These include two subsequent ACIC Competitions (Hahn, Dorie and Murray, 2018), at least three competitions organized for various machine learning conferences (Guyon et al., 2008, Guyon, Janzing and Schölkopf, 2010, Guyon, Statnikov and Batu, 2019), the DREAM *in silico* data sets that attempt to emulate single gene knockout experiments (e.g., Schaffter, Marbach and Floreano, 2011), observational data sets drawn from several exhaustive experiments on large-scale computational systems (Garant and Jensen, 2016), a recent

DARPA research program focused on evaluating methods for causal inference against data drawn from complex social simulations (DARPA, 2017), and the IBM Causal Inference Benchmarking Framework (Shimoni et al., 2018).

3. EXPANDING THE RANGE OF EMPIRICAL EVALUATION

The authors have provided an unusually strong addition to the set of resources for empirical evaluation of methods for causal inference, and the community needs many more such additions. None of the individual efforts mentioned above is comprehensive, yet the growing set of community resources could produce a major shift in how researchers guide their future research efforts (Cohen and Howe, 1988). This set of resources should continue to grow in both depth and breadth, particularly the latter. All evaluation approaches have blind spots, and researchers need a broad base of evidence from which to draw conclusions.

The authors themselves cite a number of the key challenges of causal inference that were not addressed in their competition, including non-binary treatment, non-continuous response, non-i.i.d. data, variation in sample size and number of covariates, measurement error, and weakening the assumptions of ignorability and overlap. This is an excellent list. I would add two additional items: (1) alternative causal inference tasks; and (2) real-world response surfaces. Each are described in more detail below.

Alternative causal inference tasks—The causal inference task addressed in the Challenge is simple: estimate the effect of a single binary treatment on a single continuous outcome variable at a specific time for all treated individuals. However, a wide range of alternative tasks are entirely plausible. For example, many real-world scenarios require: (1) estimating the effects of repeated interventions over time, (2) intervening on multiple dependent data instances, (3) manipulating multiple treatment variables simultaneously, (4) estimating the temporal trajectory of an outcome variable; (5) considering multiple outcome variables (perhaps with constraints or an overall cost function on their joint values); or (6) some combination of these elements. In addition, some realistic tasks could allow a method to abstain from providing any causal estimate for specific instances. Providing the means to assess performance on such tasks (particularly in the form of a competition) can be an effective way to drive research interest in such tasks and to broaden what the community considers to be reasonable research topics.

Real-world response surfaces—The response surfaces considered in the challenge were impressively diverse, but it is unclear to what extent they correspond to the response surfaces likely to be encountered in practice. While the covariates used in the challenge were drawn from real-world distributions, the treatment assignment and response surfaces were not. This is particularly concerning because one of the primary conclusions of the paper is that methods that flexibly model the response surface perform best. How much does this conclusion depend on the particular distribution of response surfaces in the competition?

As the authors note, one option for generating realistic response surfaces is to evaluate using "constructed observational studies." These studies use data from a randomized experiment to estimate treatment effect, then construct an observational data set, make an observational estimate using the constructed observational data, and compare the two estimates. In the cases cited by the authors, the observational data sets are constructed by either: (1) combining data from treated individuals in the experimental data with data from untreated individuals drawn from an alternative data source (LaLonde and Maynard, 1987, Hill, Reiter and Zanutto, 2004); or (2) allowing an alternative pool of subjects to self-select (Shadish, Clark and Steiner, 2008). Constructed observational studies have the advantage of producing real-world treatment assignment and real-world response surfaces. However, as the authors note in Section 2.1, such studies have a set of related problems: (1) they represent only a single data generating process; (2) we cannot know whether ignorability is satisfied; and (3) additional uncertainty is introduced by the comparison of two estimates (as opposed to an estimate and a known parameter value).

Intriguingly, however, there exists an alternative approach to constructing observational studies that mitigates at least one of these problems. This approach (Garant and Jensen, 2016) uses exhaustive experimental data in which all potential outcomes can be assessed (in the three examples we provide, the exhaustive experimental data is generated by manipulating large-scale computational systems). The approach then samples non-randomly from such data to create constructed observational data. As with the constructed observational studies critiqued by the authors, the estimated causal effect from the observational analysis is then compared to the experimental estimate. However, this approach has the advantage of having a re-

alistic response surface. In contrast to the data generated for the competition, only the treatment assignment is synthetic. In addition, because the treatment assignment can be determined entirely from known covariates, ignorability is satisfied. This approach is hardly a panacea—each data set still represents only a single data-generating process and it still requires comparing two estimates. However, it is one more option to consider in producing a range of empirical evaluation resources.

4. IMPLICATIONS

Given the strong reasons to trust the validity of the competition results, the conclusions of the authors are of great interest. I applaud the care and thoughtfulness of the authors in drawing conclusions from the large array of quantitative data generated by the competition. I will focus on two conclusions that should be of particular interest to researchers.

First, the authors note that methods that focus on flexible modeling of the response surface performed best, even when the method did not also model the assignment mechanism. This will surprise a large group of researchers who focus primarily on modeling treatment assignment, particularly since modeling treatment assignment appears so much simpler (at least in the case of binary treatment). This provocative finding emphasizes the need to better understand the range of likely real-world response surfaces and the extent to which the competition results would hold over that range.

Second, the authors call out "lack of alignment" as one of the most challenging aspects of the data. Alignment refers to the correspondence between the manner in which specific covariates affect both the assignment mechanism and the response surface. The authors note that: "Lack of alignment creates difficulty because if there are many covariates available to a researcher and only a subset of these are true confounders (and indeed perhaps only certain transformations of these act as true confounders) then methods that are not able to accurately privilege true confounders are potentially at a disadvantage. *Of course most of the submissions did not explicitly do this.*" (emphasis added).

Simple variable selection is unlikely to be effective in cases that lack alignment, because the question is not whether a variable is necessary for accurately modeling *either* treatment assignment *or* outcome, but whether a given variable participates in confounding and whether conditioning on that variable will remove the confounding. As Pearl (2009) and others note, this is not a simple inference, but requires substantial knowledge about the causal dependence structure among all variables. Of course, this knowledge is exactly is what learned by methods that infer causal graphical models from observational data, and strong formal theory exists to use that structure to identify the correct adjustment set for any particular causal inference (Spirtes, Glymour and Scheines, 2000, Pearl, 2009). The challenges created by lack of alignment argue strongly for using methods based on causal graphical models, both because such methods produce the knowledge necessary to infer the correct set of variables on which to condition and because the joint causal structure efficiently factors the joint distribution into much simpler and more easily estimated conditional distributions.

A final note: As the authors point out, some of their most interesting findings are negative—very few of the features sometimes conjectured to be major differentiators proved to be important in predicting performance of specific methods across a wide range of data generating processes. This provides strong guidance to future research. Researchers should either: (1) proceed as if these factors are unimportant; (2) identify special cases in which they are; or (3) identify and correct errors in this particular approach to empirical evaluation.

5. CONCLUSIONS

Again, congratulations to the authors on conducting an extremely successful and informative competition that will provide lasting value to the research community.

REFERENCES

BRADLEY, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. **30** 1145–1159.

COHEN, P. R. and HOWE, A. E. (1988). How evaluation guides AI research: The message still counts more than the medium. *AI Mag*. **9** 35.

DARPA (2017). Ground Truth (GT). Broad agency announcement. Defense Sciences Office. Defense Advanced Research Projects Agency, U.S. Dept. Defense. HR001117S0031.

DHEERU, D. and KARRA TANISKIDOU, E. (2017). UCI machine learning repository.

DOMINGOS, P. and PAZZANI, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn*. **29** 103–130.

DORIE, V. and HILL, J. and SHALIT, U. and SCOTT, M. and CERVONE, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statist. Sci*. **34** 43–68.

GARANT, D. and JENSEN, D. (2016). Evaluating causal models by comparing interventional distributions. ArXiv Preprint arXiv:1608.04698.

GELMAN, A. and LOKEN, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Dept. Statistics, Columbia Univ., New York, NY.

GUYON, I., JANZING, D. and SCHÖLKOPF, B. (2010). Causality: Objectives and assessment. In *Causality*: *Objectives and Assessment* 1–42.

GUYON, I., STATNIKOV, A. and BATU, B. (2019). *Cause-Effect Pairs in Machine Learning*. *Springer Series on Challenges in Machine Learning*. Springer. To appear.

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 *Causation and Prediction Challenge* 1–33.

HAHN, P. R., DORIE, V. and MURRAY, J. S. (2018). Atlantic Causal Inference Conference (ACIC) data analysis challenge 2017.

HILL, J. L., REITER, J. P. and ZANUTTO, E. L. (2004). A comparison of experimental and observational data analyses. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. *Wiley Ser. Probab. Stat*. 49–60. Wiley, Chichester. MR2134801

LALONDE, R. and MAYNARD, R. (1987). How precise are evaluations of employment and training programs: Evidence from a field experiment. *Evaluation Review* **11** 428–451.

LEWIS, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *European Conference on Machine Learning* 4–15. Springer.

PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166

PROVOST, F. J., FAWCETT, T., KOHAVI, R. et al. (1998). The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the International Conference on Machine Learning* **98** 445–453.

SCHAFFTER, T., MARBACH, D. and FLOREANO, D. (2011). GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* **27** 2263–2270.

SHADISH, W. R., CLARK, M. H. and STEINER, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *J. Amer. Statist. Assoc*. **103** 1334–1343. MR2655714

SHIMONI, Y., YANOVER, C., KARAVANI, E. and GOLDSCHMNIDT, Y. (2018). Benchmarking framework for performance-evaluation of causal inference analysis. ArXiv Preprint arXiv:1802.05046.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR1815675

VAN'T VEER, A. E. and GINER-SOROLLA, R. (2016). Preregistration in social psychology—A discussion and suggested template. *J. Exp. Soc. Psychol*. **67** 2–12.