Pot-luck challenge: FACT SHEET

Repository URLs:
LOCANET `http://www.causality.inf.ethz.ch/data/LOCANET.html`
REGED `http://www.causality.inf.ethz.ch/repository.php?id=7`
SIDO `http://www.causality.inf.ethz.ch/repository.php?id=1`
CINA `http://www.causality.inf.ethz.ch/repository.php?id=6`
MARTI `http://www.causality.inf.ethz.ch/repository.php?id=8`

Task name: LOCANET
Title: LOcal CAusal NETwork
Authors: Isabelle Guyon, Alexander Statnikov, Constantin Aliferis
Contact: Isabelle Guyon, isabelle@clopinet.com, `http://clopinet.com/isabelle`

Key facts:
Data sets for discovering the local structure around a target variable. Time independent tasks. Learning causal structure from observational data. Four semi-artificial datasets (two using re-simulated data and two using real data augmented with artificial probe variables):

Dataset	Domain	Type	Features	Feat. #	Train #	Test #
REGED	Genomics	Re-simulated	Numeric	999	500	20000
SIDO	Pharmacology	Real + probes	Binary	4932	12678	10000
CINA	Econometrics	Real + probes	Mixed	132	16033	10000
MARTI	Genomics	Re-simulated	Numeric	999	500	20000

Abstract:

We designed four datasets for the purpose of benchmarking local causal discovery algorithms. These include two "re-simulated" datasets obtained from artificially generated data from models trained with real data and two datasets including real variables intermixed with artificial variables (called probes). There is no time dependency in the samples. We chose applications in marketing, pharmacology and bio-medicine spanning a high diversity of types of distributions. The datasets were used in two challenges in 2008 organized for the WCCI and NIPS conferences. A detailed technical report on the dataset design is available (Guyon et al., 2009). The website of the challenges remains open for post-challenge submissions (http://clopinet.com/causality).

Design:

We focused on some specific aspects of causal discovery:

Causality between random variables. We address causal relationships between random variables, as opposed to causal relationships between events, or objects.

No time dependency. Our everyday-life concept of causality is very much linked to time dependencies (the 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]

Learning from observational data. Only training data from a "natural" pre-manipulation distribution (observational data) is available for training. In other settings, experimental data may be available as well. Relatively small training sets are provided, making it difficult to infer conditional independencies and learning distributions.

Discovering local causal relationships. We focus on one particular variable of interest called "target" and design tasks requiring to uncover the variables, which are most closely related (*e.g.,* direct causes and consequences, Markov blanket, depth 3 network). The problem of local causal relationships is closely related to that of variable selection: (1) variables closely related to the target in a causal graph may be highly predictive; (2) the knowledge of causal relationships is useful to select the variables, which will remain predictive in post-manipulation distributions.

Predicting the consequences of manipulations. There is no predictive task in the pot-luck challenge LOCANET tasks, but our datasets were previously used for prediction tasks in the WCCI 2008 "causation and prediction challenge" (Guyon et al., 2008). They include test samples drawn from a "natural" pre-manipulation distribution and test samples drawn from various post-manipulation distributions, which can be used to assess predictive performances of the target variable. Post-challenge submissions can be made online at `http://www.causality.inf.ethz.ch/challenge.php`.

The type of causal relationships under consideration have often been modeled as Bayesian causal networks or structural equation models (SEM) (Pearl, 2000; Spirtes et al., 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, plus some noise. 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 SEM 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.

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.

We have adopted two strategies to design datasets suitable for benchmarks:

- **Re-simulated data:** We train a causal model (a causal Bayesian network or a structural equation model) with real data. The model is then used to generate artificial training and test data for the challenge. Truth values of causal relationships are known for the data generating model and used for scoring causal discovery results.
- **Real data with probe variables:** We use a dataset of real samples. Some of the variables may be causally related to the target and some may be predictive but non-causal. The nature of the causal relationships of the variables to the target is unknown (although domain knowledge may allow us to validate the discoveries to some extent). We add to the set of real variables a number of distractor variables called "probes", which are generated by an artificial stochastic process, including explicit functions of some of the real variables, other artificial variables, and/or the target. All probes are non-causes of the target, some are completely unrelated to the target. The identity of the probes in concealed.

The LOCANET datasets include two re-simulated datasets and two real datasets with probes. They nicely complement each other: Re-simulated data provide us with full control over the data generative process and the truth values of all causal relationships, while real data with probes provide us with actual data distributions. The fact that truth values of causal relationships are known only for the probes affects the evaluation of causal discovery, which is less reliable than for artificial data.

Dataset description:

We formatted four datasets, including two re-simulated datasets (REGED and MARTI) and two real datasets with probes (CINA and SIDO). All datasets are thoroughly documented (including origin of the raw data, data preparation, past usage, and baseline results) in a Technical Report (Guyon et al., 2009). We briefly describe them:

REGED (REsimulated Gene Expression Dataset): The problem is to find genes, which could be responsible of lung cancer. The data are generated by a model derived from real human lung-cancer microarray gene expression data. From the causal discovery point of view, it is important to separate genes whose activity causes lung cancer from those whose activity is a consequence of the disease. The data include no hidden variable or missing data. The target variable is binary: it separates malignant samples (adenocarcinoma) from control samples (squamous).

SIDO (SImple Drug Operation mechanisms) contains descriptors of molecules which have been tested against the AIDS HIV virus. The target values indicate the molecular activity (+1 active, −1 inactive). The causal discovery task is to uncover causes of molecular activity among the molecule descriptors. This would help chemists in the design of new compounds, retaining activity, but having perhaps other desirable properties (less toxic, easier to administer). The molecular descriptors were generated programmatically from the three dimensional description of the molecule, with several programs used by pharmaceutical companies for QSAR studies (Quantitative Structure-Activity Relationship). For example, a descriptor may be the number of carbon molecules, the presence of an aliphatic cycle, the length of the longest saturated chain, etc.

CINA (Census Is Not Adult) is derived from census data (the UCI machine-learning repository Adult database). The data consists of census records for a number of individuals. The causal discovery task is to uncover the socio-economic factors affecting high income (the target value indicates whether the income exceeds 50K). The 14 original attributes (features) including age, workclass, education, marital status, occupation, native country, etc. are continuous, binary, or categorical. Categorical variables were converted to multiple binary variables (as we shall see, this preprocessing, which facilitates the tasks of some classifiers, complicates causal discovery).

MARTI (Measurement ARTIfact) is obtained from the same data generative process as REGED, a source of simulated genomic data. Similarly to REGED the data do not have hidden variables or missing data, but a noise model was added to simulate the imperfections of the measurement device. The goal is still to find genes, which could be responsible of lung cancer. The target variable is binary; it indicates malignant samples (adenocarcinoma) *vs.* control samples (squamous). The feature values representing measurements of gene expression levels are assumed to have been recorded from a two-dimensional microarray 32x32. The training set was perturbed by a zero-mean correlated noise model (**?**).

For the "causation and prediction challenge" (Guyon et al., 2008), the participants had to return predictions for the binary target variable on test data for three test set versions (version 0 from the unmanipulated distribution and versions 1, and 2 from the manipulated distribution). For the "pot-luck challenge", the participants needed only the training data (the same in all three versions) to produce the local causal structure.

Task of the LOCANET challenge:

The participants were asked to provide a depth 3 causal network (oriented graph structure) around the target, using only training data only for causal discovery. The submission format is via a text file containing the list of parents of the features of interest. The target is numbered 0. All other features are numbered with their column number in the data tables. Provide a file named: `<yourlastname>_<dataname>_feat.localgraph`. Example `Guyon_LUCAS_feat.localgraph`:

0: 1 5
1: 3 4
2: 1
6: 5
8: 6 9
9: 0 11
11: 0 10

Evaluation:

The participants of LOCANET were ranked on the basis of an average edit distance to the true causal relationship between the target and variables in the depth three network. Specifically, we considered only local directed acyclic graphs and encoded the relationship of a variable to the target variable as a string of up (u) and down (d) arrows, from the target:

Depth 1 relatives: parents (u) and children (d).

Depth 2 relatives: spouses (du), grand-children (dd), siblings (ud), grand-parents (uu). Depth 3 relatives: great-grand-parents (uuu), uncles/aunts (uud), nices/nephews (udd), parents of siblings (udu), spouses of children (ddu), parents in law (duu), children of spouses (dud), great-grand-children (ddd).

A confusion matrix C_{ij} was computed, recording the number of relatives confused for another type of relative, among the 14 types of relatives in depth 3 networks. A cost matrix A_{ij}, was applied to account for the distance between relatives (computed with an edit distance as the number of substititions, insertion, or deletion to go from one string to the other, using the string description described above). The score of the solution was computed as:

$$S = \sum_{ij} A_{ij} C_{ij}$$

There are additional details on how to handle ties. We provide the Matlab code to compute this score (Guyon, 2009). For artificially generated data (REGED and MARTI), the ground truth for the target local neighborhood was determined by the generative model. For real data with artificial "probe" variables (SIDO and CINA), we do not have ground truth for the relationships of the real variables to the target. The score was therefore computed on the basis of the artificial variables only.

After the challenge, we also computed other metrics of evaluation. For particular features subsets (parents, children, parents and children, Markov blanket[2], all relatives up to depth 2, all relatives up to depth 3), we computed precision and recall (/em aka sensitivity or true positive rare), defined as follow:
Precision: NumberGoodFound / NumberFound
Recall: NumberGoodFound / NumberGood.

We also evaluated the predictive power of the Markov blanket by training a reference classifier (linear ridge regression) and testing on unmanipulated test data.

Results and conclusions:

Ten participants entered the challenge. All the details of the analysis and fact sheets for some of the entries are available on-line at: `http://www.causality.inf.ethz.ch/data/LOCANET.html`.
The methods included: Structure learning using independence tests (Brown & Tsamardinos and Zhou, Wang, Yin & Geng), combinations of score-based and structure learning methods (de-Prado-Cumplido & Antonio Artes-Rodrigues and Tillman & Ramsey), combinations of feature selection and structure methods (Olsen, Meyer & Bontempi), and ensemble methods (Mwebaze & Quinn).

The edit distance scores of the participants were fairly poor. On REGED and MARTI, the best ranking entries were empty graphs. On CINA, the best ranking entry had results worse than the fully connected graph (with symmetric connections). On SIDO, the best result was barely better than that of the empty graph. From the point of view of the precision and recall metrics, structure learning methods gave the most promising results

2. We call Markov blanket the set of parents, children, and spouses of the target variable.

(highest precision), but all methods gave a poor recall, particularly for SIDO. We performed additional qualitative analyses in CINA using the semantics of the identifiers of the true variables to see whether the uncovered relationships made sense. It is unclear whether using the tools of causal discovery brought us a lot more information that simple correlation would have:
- most features cited as cause or effect of the target rank among the most correlated features,
- there is usually no consensus on the causal direction among the participants,
- when there is a large consensus on the causal direction, the result is sometimes suspicious given the semantics of the feature,
- a simple ranking in order of correlation yields nested feature subsets always more predictive than the Markov blanket.

Overall these results point to the need to improve the reliability of causal discovery from observational data.

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. Additional support was provided by the U.S. National Institute of Health under grant 2R56LM007948-04A1.

References

I. Guyon. Scoring code for the locanet taks. `http://www.causality.inf.ethz.ch/data/LocanetScoreCode.zip`, October 2009.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Design and analysis of the causation and prediction challenge. In *JMLR W&CP*, volume 3, pages 1–33, WCCI2008 workshop on causality, Hong Kong, June 3-4 2008.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Datasets of the causation and prediction challenge. Technical Report, `http://eprints.pascal-network.org/archive/00004566/`, 2009.

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

Judea Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, March 2000.

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