

Pot-luck causality challenge: FACT SHEET (for a task solved)

Title: Brown_Tsamardinos

Participant name, address, email and website:

Laura Brown, Eskind Biomedical Library 4th floor, 2209 Garland Ave.

Nashville, TN 37232 USA laura.e.brown@vanderbilt.edu <http://www.dsl-lab.org>

Ioannis Tsamardinos, FORTH-ICS, N. Plastira 100, Vassilika Vouton GR-700 13

Heraklion Crete, GREECE tsamard@ics.forth.gr

Task(s) solved: LOCANET

Reference:

Bach's – Bach, F.R. and Jordan, M.I. NIPS, 2002

Causality Challenge Submission – Brown, L.E. and Tsamardinos, I. “A Strategy for Making Predictions Under Manipulation” 2008

MMHC – Tsamardinos, I. et al. Machine Learning, 2006

MMPC, MMB – Tsamardinos, I. et al. SIGKDD, 2003

Method:

Our submission for the LOCANET challenge relied on the results and procedures of the first causality challenge, from which the local networks were pruned. Details to the approach used for the first causality challenge are available in the paper for that challenge (available at the DSL website) but a general overview of the method and how the results are used for this task are presented.

Preprocessing: The preprocessing was tailored to each data set. For the REGED data set each variable was normalized so its mean was zero and standard deviation was one. For the SIDO data set, the variables were binary and no preprocessing was performed. For the CINA data set, variables that were not binary were treated as continuous and normalized; binary variables were all set to values of zero and one. For the MARTI data set, the preprocessed data by Dr. Guyon available on the challenge website was used.

Causal discovery: Once the initial data sets have been pre-processed, the next step of our procedure was to identify the skeleton structure of the Bayesian Network around the target variable recursively using the MMPC algorithm, up to three edges away from the target. This region of interest makes it practical to apply causal algorithms that cannot scale up to the sizes of all the networks in the challenge. The selection of a depth 3 in this case was for the previous challenge where we focused on identifying the Markov Blanket. For future work on the LOCANET challenge which focuses on learning a region out to depth 3, the MMPC recursion should run out to depth 4 and then be pruned back as a final step. In the next step of our analysis we tried to orient the edges of the region. For the case of continuous or mixed data, an adaptation of Bach's algorithm was used. For the case of binary data, MMHC was used to find the top scoring network. From the learned network, the region of depth 3 was extracted and submitted for analysis.

Feature selection: The recursive selection of variables to include in the region can be thought of as performing several iterations of feature selection.

Classification: None

Model selection/hyperparameter selection: Currently, a default set of parameters are used in the edge orientation procedure (parameters for score calculation either via BDeu score

in MMHC or parameters for the kernel in Bach’s algorithm). Future work for this challenge could also involve using many different parameters and perform model averaging over the results.

Results:

Table 1: Result table. The score of our method along with the top and lowest score for each data set are given. Three reference scores are also presented where applicable for comparison. The scores for REGED and MARTI are the second best submitted.

	REGED	SIDO	CINA	MARTI
Brown/Tsamardinos	0.27	3.46	2.23	0.36
Best Overall	0.22	3.31	1.70	0.21
Worst Overall	0.52	3.48	3.31	0.93
Reference A	0.01	0.64	0.64	0.02
Reference B	0.16	1.92	1.89	0.16
Reference C	3.08		1.67	3.01

Reference A: Truth graph with 20% of the edges flipped at random.

Reference B: Truth graph with connections symmetrized.

Reference C: Variables in the truth graph, fully connected.

Advantages:

The method gains in efficiency by rather than learning the entire network and pruning out the region it uses the recursive application of a local neighborhood identification method (MMPC) in a breadth-first search then orients the graph.

Limitations:

The results on CINA may be low because of the inappropriateness of the statistical tests used in MMPC for the mixed data. The MMPC algorithms have statistical tests provided for when the data is entirely binary or continuous (with a binary target); the mixed data set did not therefore match well to these methods. Also, as stated above the performance of the method may be improved by allowing the recursive procedure to run to a depth of 4 in order to better facilitate identification of all edges in the region of depth 3.

Implementation:

The methods are implemented in Matlab. The MMPC and MMHC algorithms are available from the Causal Explorer library, www.dsl-lab.org (please note, we were in part the developers of these methods and may have slightly extended or modified the code from the precise implementation available in Causal Explorer). Our method combined many algorithms and used the results from the previous challenge which are not available as a push-button application although the code and executables are available at the above website.

Keywords:

- Preprocessing or feature construction: normalization.
- Causal discovery: Bayesian Network,
- Feature selection: filter
- Classifier:
- Hyper-parameter selection:
- Other: