The University of Manchester

MANCHESTER 1824

Structure Learning of Bayesian Networks: Challenges and Opportunities

Zhigao Guo (zhigao.guo@manchester.ac.uk)

Research Group: Machine Learning and Robotics

Usual machine learning, usually based on strict assumptions, aims at finding a model that best fit the available (small amounts of) observational data but often fail to generalize to the future observations. One reason is it ignores the underlying causal mechanisms that holds in both past and future observations. Causal models provide solutions in representing the causal relationships between variables of the investigated system. Bayesian network (BN), as a type of causal models, has attracted attentions of machine learning community because of its successful applications in fields, such as fault diagnosis, automatic driving, and medical decision making. Structure learning of Bayesian networks, as part of causal discovery issue, is crucial. After decades of study, statisticians and computer scientists have contributed various learning approaches. Within the past three years, we conducted extensive literature review and empirical study and therefore saw challenges and rewards of BN structure learning. Through this poster, we present the work we have done and findings and insights that deserve attraction.

Motivation

- Numerous Bayesian Network structure learning (BNSL) algorithms have been proposed in the past 30 years. However, there is no agreement on which one is ”best”.
- Most algorithms are based on a set of assumptions, such as complete data and causal sufficiency and tend to be evaluated with data that conform to those assumptions. However, real-world data often does not obey those assumptions.

Work

- We tested on three well-established networks (Asia, Alarm, and Pathfinder) with up to 109 variables, and three real-world application networks (Sports, ForMed, Property) with up to 88 variables.
- We generated data of noise, such as missing values, incorrect values, latent variables, merged states and their combinations.
- We investigated the performance of 15 well-established BNSL algorithms. They are PC-stable, FGES, FCI, GFCI, RFCI-BSC, Inter-IAMB, MMHC, GS, HC, Tabu, HC, H2PC, SaiyanH, ILP, WINASOBS, NOTEARS.
- We considered implementations (with defaul parameters) of tested algorithms from software or packages, including bnlearn (R), r-causal (R), GOBNILP (C), BLIP (Java), Bayesys (Java), and NOTEARS (Python).
- The algorithms were evaluated in terms of metrics, such as F1, SHD, BSF and time complexity.
- This work involved learning over 7,000 graphs with a total learning runtime of seven months.

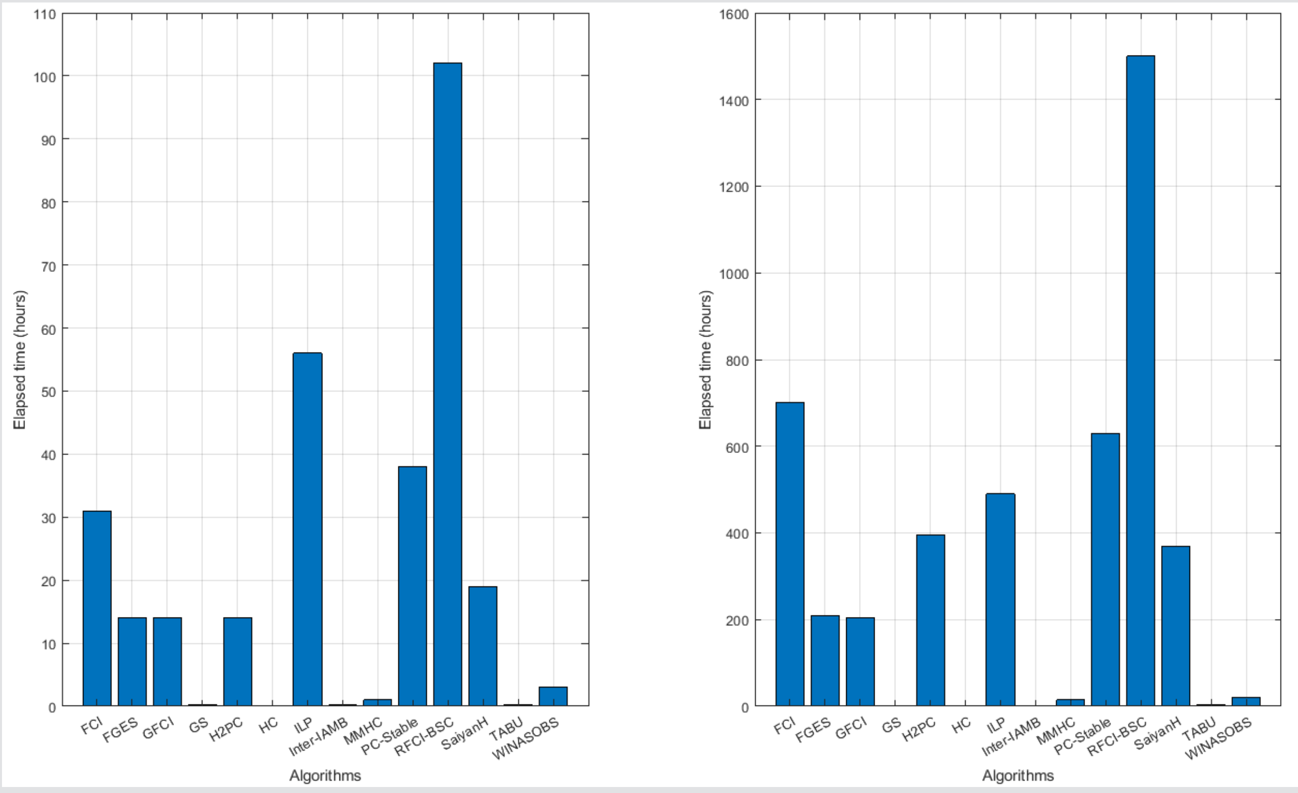
Results

Algorithms	F1		SHD		BSF	
	Average rank	Overall rank	Average rank	Overall rank	Average rank	Overall rank
FCI	7.7	9th	6.57	7th	7.67	9th
FGES	7.5	8th	7.83	10th	7.1	8th
GFCI	6.87	7th	6.87	9th	6.97	6th
GS	11.87	14th	10.43	13th	11.9	14th
H2PC	6.13	5th	5.1	3rd	6.97	6th
HC	3.63	2nd	4.77	2nd	3.17	2nd
ILP	4.8	3rd	6.43	5th	4.13	3rd
Inter-IAMB	10	12th	8.6	12th	10.43	12th
MMHC	7.77	10th	6.47	6th	8.6	11th
NOTEARS	12	15th	13	15th	12	15th
PC-Stable	8.1	11th	6.83	8th	8	10th
RFCI-BSC	11.5	13th	10.9	14th	11.47	13th
SaiyanH	5.33	5th	8	11th	4.77	5th
Tabu	3.27	1st	4.43	1st	3.1	1st
WINASOBS	6.3	6th	5.87	5th	6.17	5th

Table: Average and overall ranked performance of the algorithms over all case studies in noise-free experiment, as determined by each of the three metrics.

Algorithms	F1		SHD		BSF	
	Average rank	Overall rank	Average rank	Overall rank	Average rank	Overall rank
FCI	8.67	11th	8.67	12th	8.23	11th
FGES	7.15	8th	7.12	8th	7.37	8th
GFCI	7.26	9th	6.91	7th	7.60	10th
GS	11.74	15th	9.54	13th	11.68	14th
H2PC	5.66	5th	4.96	3rd	6.26	5th
HC	3.60	1st	4.92	2nd	3.03	1st
ILP	5.17	3rd	6.72	6th	4.35	3rd
Inter-IAMB	9.79	12th	7.82	9th	9.98	12th
MMHC	6.51	6th	4.66	1st	7.59	9th
NOTEARS	11.65	14th	12.83	15th	12.51	15th
PC-Stable	7.59	10th	7.87	10th	7.15	7th
RFCI-BSC	11.5	13th	11.05	14th	11.54	13th
SaiyanH	5.27	4th	7.87	11th	5.16	4th
Tabu	3.62	2nd	4.99	4th	3.13	2nd
WINASOBS	6.54	7th	5.49	5th	6.77	6th

Table: Average and overall ranked performance of the algorithms over all case studies in noise-based experiments determined by the three metrics.



The cumulative runtime of the algorithms over noise-free (left) and noise-based (right) experiments.

Findings

- Performance of algorithms tested on traditional synthetic data drops on the real-world data.
- A higher fitting score does not necessarily imply a more accurate causal graph.
- Score-based algorithms are generally superior to the constraint-based algorithms.

Challenges and Opportunities

- Causal discovery, especially from complicated real-world data (with hidden variables, selection bias, measurement errors), is far from trustworthy.
- Causality plays crucial roles in explainable AI and causal elements enhance the interpretability of AI models.

References

N. Kitson, A. Constantinou, Z. Guo, Y. Liu, K. Chobtham
A survey of Bayesian Network structure learning
https://arxiv.org/abs/2109.11415

A. Constantinou, Y. Liu, K. Chobtham, Z. Guo, N. Kitson
Large-scale empirical validation of Bayesian Network structure learning algorithms with noisy data
International Journal of Approximate Reasoning, 131: 151-188.