Computational Causal Discovery and its Applications

Sisi Ma

University of Minnesota

sisima@umn.edu

April 6, 2019

Computational Causal Analytics at U of Minnesota

Constantin Aliferis, Professor, Director of IHI at UMN

Gyorgy Simon, Associate Professor

Computational Causal Analytics at U of Minnesota

Erich Kummerfeld, Assisstant Professor

Sisi Ma, Assisstant Professor

Where to find material related to this talk?

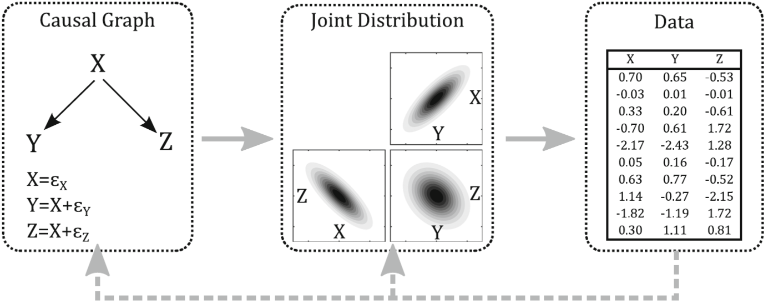
github.com/SisiMa1729/CausalAnalytics_Intro
github.com/SisiMa1729/Causal_Feature_Selection

Overview

What is Computational Causal Discovery?

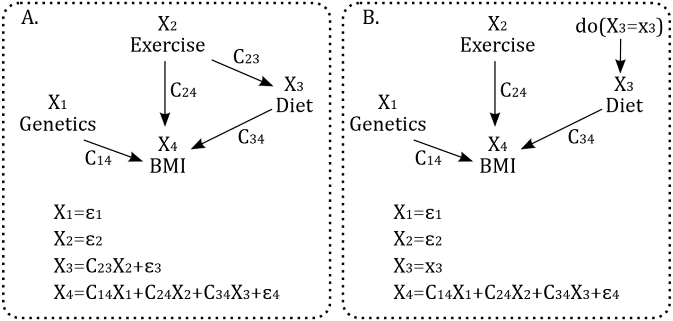
- Discovery of Causal Structure
- Estimation of Causal Effect
- Using a combination of observational data, experimental data, and prior knowledge.

Discovery of Causal Structure



Estimation of Causal Effect

$E(Y|do(X=x)) - E(Y|do(X=x'))$ is the average effect on Y when changing X from x to x'

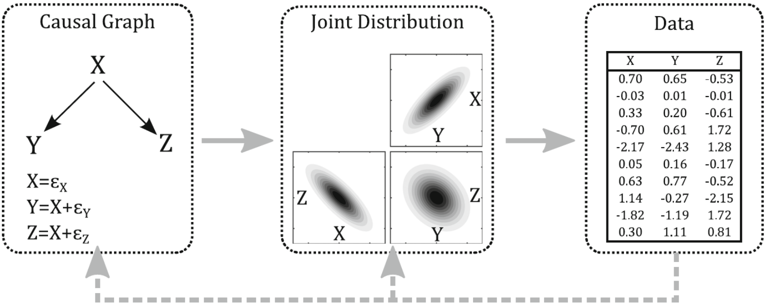


* comment: SEM?

Discovery of Causal Structure

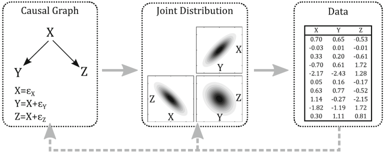
Correct effect estimation depends on identifying the correct causal structure.

Discovery of Causal Structure



X	Y	Z
0.70	0.65	-0.53
-0.03	0.01	-0.01
0.33	0.20	-0.61
-0.70	0.61	1.72
-2.17	-2.43	1.28
0.05	0.16	-0.17
0.63	0.77	-0.52
1.14	-0.27	-2.15
-1.82	-1.19	1.72
0.30	1.11	0.81

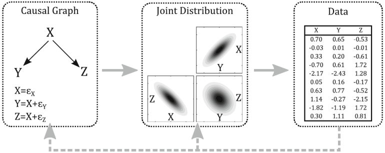
Discovery of Causal Structure



X	Y	Z
0.70	0.65	-0.53
-0.03	0.01	-0.01
0.33	0.20	-0.61
-0.70	0.61	1.72
-2.17	-2.43	1.28
0.05	0.16	-0.17
0.63	0.77	-0.52
1.14	-0.27	-2.15
-1.82	-1.19	1.72
0.30	1.11	0.81

- score based method (maximize likelihood of data given causal structure)
- constraint-based (joint distribution of data is constrained by causal structure)
- hybrid of the above two

Discovery of Causal Structure



- score based method (maximize likelihood of data given causal structure)
- constraint-based (joint distribution of data is constrained by causal structure)
- hybrid of the above two

Discovery of Causal Structures: Assumptions

Causal Markovian Condition: Every Markovian causal model M induces a distribution $P(X_1, ...X_i, ..., X_n)$ that satisfies the parental Markov condition relative to to causal diagram G associated with M; that is, each variable X_i is independent of all its nondescendants, given its parents in G. i.e.
$X_i \perp V - DE_i | PA_i$

This results in the important decomposition of the joint distribution $P(X_1, ..., X_i, ..., X_n) = \prod_i^n P(X_i|PA_i)$

Discovery of Causal Structures: Assumptions

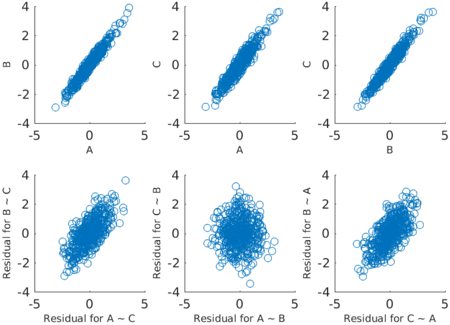
Faithfulness Condition: Let G be a causal graph and P a probability distribution generated by G. G and P satisfies the faithfulness condition if and only if every conditional independence true in P is entailed by the Causal Markovian condition applied to G.

Discovery of Causal Structures: Assumptions

Intuitively, causal Markov condition and faithfullness condition establish a one to one relationship between the causal graph (more precisely, Markov equivalent class) and the joint probablity distribution. This correspondance makes causal discovery possible.

Causal Graph and Conditional Independency

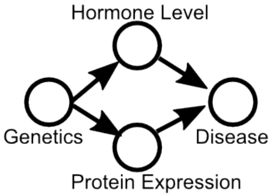
Conditional independence, What does it look like:
$A- > B- > C$; $A < -B < -C$; or $A < -B- > C$



$$A \not\perp B, A \not\perp C, B \not\perp C$$
$$A \not\perp B|C, A \perp C|B, B \not\perp C|A$$

Causal Graph and Conditional Independency

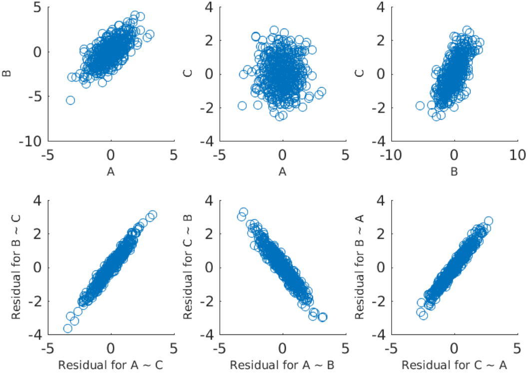
The concept of conditional independence:



Causal Graph and Conditional Independency

Conditional independence, What does it look like:

$A- > B < -C$



$A \not\perp B, A \perp C, B \not\perp C$

$A \not\perp B|C, A \not\perp C|B, B \not\perp C|A$

Causal Graph and Conditional Independency

Conditional Independence Tests:

- fisher's test
- χ^2, g^2
- conditional mutual information, distance correlation
- comparison of nested models

Causal Mechanism from Observational Data

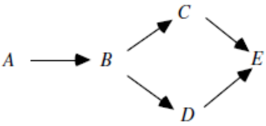
The demonstrated relationships among three variables can be extended to discover causal graphs among any number of variables.

Some popular algorithms:

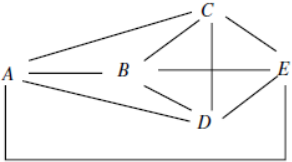
- constraint based: SGS, PC, HITON-PC, FCI
- score based: GES, FGES
- hybrid: MMHC,GFCI

Causal Mechanism from Observational Data

Example: PC algorithm: Skeleton Phase:



True Graph



Complete Undirected Graph

---

n = 0 No zero order independencies

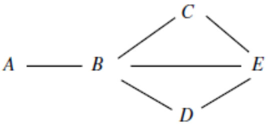
---

n = 1 First order independencies

$A \perp\!\!\!\perp C \mid B$ $A \perp\!\!\!\perp D \mid B$

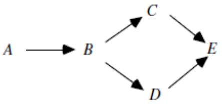
$A \perp\!\!\!\perp E \mid B$ $C \perp\!\!\!\perp D \mid B$

Resulting Adjacencies



Causal Mechanism from Observational Data

Example: PC algorithm:



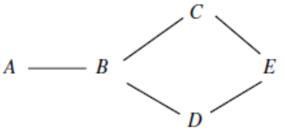
True Graph

Skeleton Phase (continued):

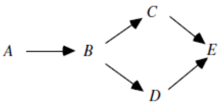
n = 2: Second order independencies

$B \perp\!\!\!\perp E \mid \{C,D\}$

Resulting Adjacencies

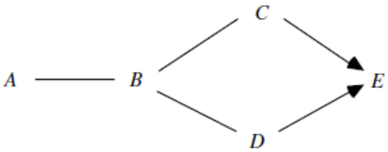


Example: PC algorithm:



True Graph

Orientation Phase:



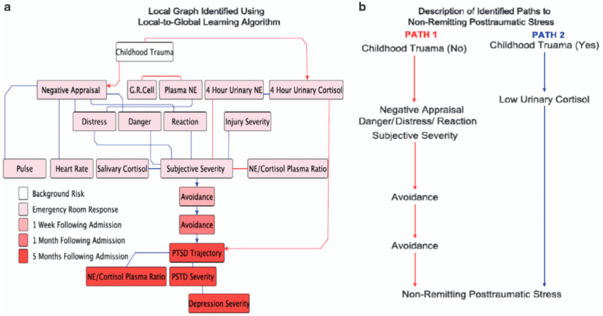
Note: Markov equivalent class

Applications of Causal Discovery Methods

- Causal Mechanism from Observational Data
- Causal Graph Guided Experimentation
- Causal Feature Selection

Causal Mechanism from Observational Data

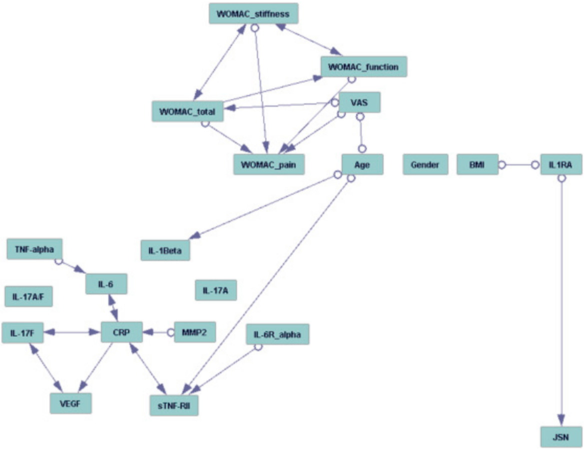
Biological Mechanisms of PTSD:



[Galatzer-Levy et al., 2017]

Causal Mechanism from Observational Data

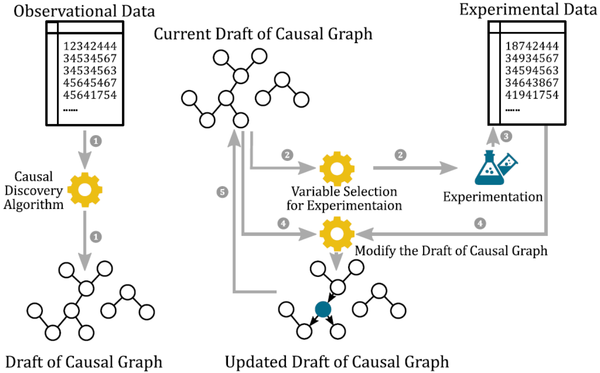
Biological Mechanisms of Osteoarthritis:



[Attur et al., 2015]

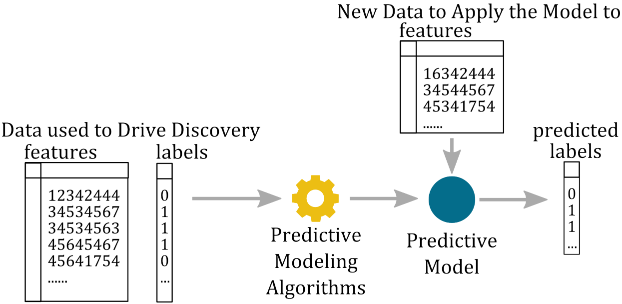
Causal Graph Guided Experimentation

Reduce number of Experiments needed to fully resolve the causal relations



[Statnikov et al., 2015]

Causal Feature Selection for Predictive Models



Causal Feature Selection for Predictive Models

Goals of Feature Selection:

- Improve model predictivity
- Enhance model interpretability
- Increase cost-efficiency for model deployment

Causal feature selection is well-suited for these goals.

Optimal Feature Set

Optimal feature set: Given a data set $\mathbb{D}$ (a sample from distribution $\mathbb{P}$) for variables $\mathbf{V}$, a learning algorithm $\mathbb{L}$, and a performance metric $\mathbb{M}$, a feature set $\mathbf{X} \subseteq \mathbf{V}$ is an optimal feature set of T if $\mathbf{X}$ maximizes the performance metric $\mathbb{M}$ for predicting T using learner $\mathbb{L}$ in the dataset $\mathbb{D}$.
[Statnikov et al., 2013, Kohavi and John, 1997]
Note: Optimal here is defined with respect to predictivity. There could be redundent features in a optimal feature set according to this definition.

Information Content of a Feature

The contribution of a given feature to the predictive performance of a model depends on what other features are present in the model.

Information Content of a Feature

Let $\mathbf{S_i} = \mathbf{V} - V_i$, the set of all features except V_i
Strong relevance: A feature V_i is strongly relevant to T iff there exists some v_i, t, and s_i for which $p(V_i = v_i, \mathbf{S_i} = s_i) > 0$, such that $p(T = t | V_i = v_i, \mathbf{S_i} = s_i) \neq p(T = t | \mathbf{S_i} = s_i)$.
[Kohavi and John, 1997]

Information Content of a Feature

Strongly relevant features are conditionally dependent of T given all other measured variables. In other words, knowing the value of features V_i gives additional information regarding the target of interest, even when we know the value of all other features.
[Kohavi and John, 1997]

Information Content of a Feature

Let $\mathbf{S_i} = \mathbf{V} - V_i$, the set of all features except V_i
Weak relevance: A feature V_i is weakly relevant to T iff it is not strongly relevant, and $\exists S_i' \subset S_i$ and some v_i, t, and s_i' for which $p(V_i = v_i, \mathbf{S_i'} = s_i') > 0$, such that
$p(T = t | X_i = v_i, S_i' = s_i') \neq p(T = t | S_i' = s_i')$.
[Kohavi and John, 1997]

Selection based on combination of features

Knowing the value of weakly relevant variable V_i do not give additional information regarding the target of interest, when we know the value of all other features; however if we only know the value of a subset of all other measured variables, knowing the value of V_i may provide additional information.
[Kohavi and John, 1997]

Selection based on combination of features

Irrelevance:A feature V_i is relevant if it is either weakly relevant or strongly relevant; otherwise, it is irrelevant.

Irrelevant features neither provide information regarding the target by itself nor in combination with any other features.

[Kohavi and John, 1997]

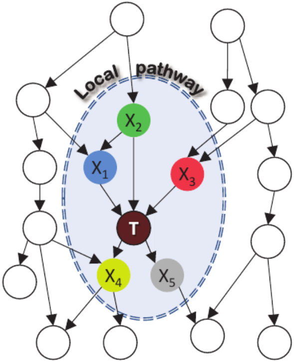
Selection based on combination of features

Given the definition of strong, weak and irrelevant variables, what is our stratergy?

- Strongly relevant features should always be selected. Omitting Strongly relevant features will compromise predictive performance.
- Weakly relevant features do not improve predictive performance if all strongly relevant features are selected.
- Irrelevant features do not improve predictive performance.

[Kohavi and John, 1997]

Relationship between Causality and Feature Selection



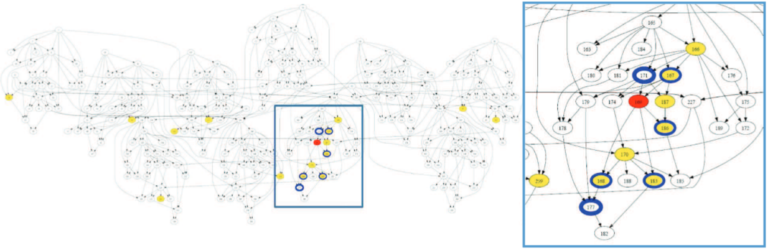
The Markov Boundary (direct causes + direct effects + direct causes of direct effects) is the minimal feature set that contain all information regarding the target of interest (i.e. strongly relevant) under faithfulness.

Benefit of Causal Feature Selection

- selected features have causal interpretations
- selected feature set is generally smaller compared to non-causal methods
- model generalize better under certain type of distribution shift

Benefit of Causal Feature Selection

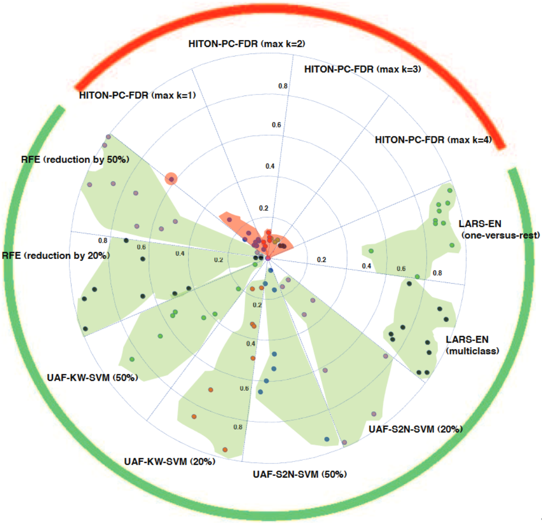
Example: selected features have causal interpretations.



[Aliferis et al., 2010a]

Benefit of Causal Feature Selection

Example: selected features have causal interpretations.



Benefit of Causal Feature Selection

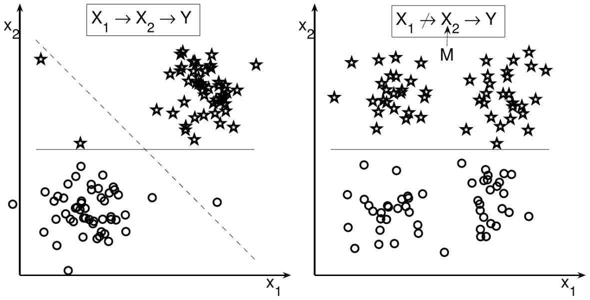
Example: selected feature set is generally smaller compared to non-causal methods, without compromising predictive performance.

	Predicitivity		***Reduction***	
Feature selection method	P-value	Nominal winner	P-value	Nominal winner
No feature selection	0.1890	Other	**<0.0001**	HITON-PC
RFE: 4 variants	0.9754	Other	**0.0046**	HITON-PC
	0.8030	Other	**0.0042**	HITON-PC
	0.1312	HITON-PC	0.3634	HITON-PC
	0.1008	HITON-PC	0.6816	Other
UAF-KruskalWallis-SVM: 4 variants	0.2248	Other	**0.0028**	HITON-PC
	0.0098	Other	**0.0004**	HITON-PC
	1.0000	HITON-PC	0.1414	HITON-PC
	0.3232	HITON-PC	0.3998	HITON-PC
UAF-Signal2Noise-SVM: 4 variants	0.0710	Other	**0.0018**	HITON-PC
	0.0752	Other	**0.0030**	HITON-PC
	0.4420	HITON-PC	0.7850	HITON-PC
	0.2820	HITON-PC	0.6604	HITON-PC
UAF-Neal-SVM: 4 variants	0.5046	Other	**<0.0001**	HITON-PC
	0.9782	HITON-PC	**<0.0001**	HITON-PC
	0.6980	HITON-PC	**0.0044**	HITON-PC
	0.3806	HITON-PC	**0.0186**	HITON-PC
Random Forest Variable Selection: 2 variants	0.6064	HITON-PC	0.3252	HITON-PC
	0.5050	HITON-PC	0.1338	Other
LARS-Elastic Net: 2 variants	1.0000	Other	0.1112	HITON-PC
	0.0832	HITON-PC	0.5216	Other

[Aliferis et al., 2010a]

Benefit of Causal Feature Selection

Example: Model built with causal feature selection methods generalize better under certain distribution shifts.



(a) No manipulation. (b) Manipulation of X_2.

[Guyon et al., 2007]

Markov Boundary

There are a number of algorithms that discovery the local causal neibourhood of the data.

- PC-simple[Bühlmann et al., 2010]
- HITON-PC[Aliferis et al., 2003, Aliferis et al., 2010a, Aliferis et al., 2010b]
- HITON-MB[Aliferis et al., 2010a, Aliferis et al., 2010b]
- MMPC [Tsamardinos et al., 2006]
- IAMB[Tsamardinos et al., 2003]
- MBFS[Ramsey, 2006]

Take Home Message

- Causal structural discovery methods can identify causal relationships from observational and experimental data with and without domain knowledge.
- Computational causal discovery methods have wide applications in knowledge discovery and predictive modeling.
- `github.com/SisiMa1729/CausalAnalytics_Intro`
- `github.com/SisiMa1729/Causal_Feature_Selection`

References I

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. (2010a).
Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation.
Journal of Machine Learning Research, 11(Jan):171–234.

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. (2010b).
Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: Analysis and extensions.
Journal of Machine Learning Research, 11(Jan):235–284.

References II

Aliferis, C. F., Tsamardinos, I., and Statnikov, A. (2003).
Hiton: a novel markov blanket algorithm for optimal variable selection.
In *AMIA Annual Symposium Proceedings*, volume 2003, page 21. American Medical Informatics Association.

Attur, M., Statnikov, A., Samuels, J., Li, Z., Alekseyenko, A., Greenberg, J., Krasnokutsky, S., Rybak, L., Lu, Q., Todd, J., et al. (2015).
Plasma levels of interleukin-1 receptor antagonist (il1ra) predict radiographic progression of symptomatic knee osteoarthritis.
Osteoarthritis and cartilage, 23(11):1915–1924.

References III

Bühlmann, P., Kalisch, M., and Maathuis, M. H. (2010).
Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm.
Biometrika, 97(2):261–278.

Galatzer-Levy, I., Ma, S., Statnikov, A., Yehuda, R., and Shalev, A. (2017).
Utilization of machine learning for prediction of post-traumatic stress: a re-examination of cortisol in the prediction and pathways to non-remitting ptsd.
Translational psychiatry, 7(3):e1070.

Guyon, I., Aliferis, C., and Elisseeff, A. (2007).
Causal feature selection.
Computational methods of feature selection, pages 63–82.

References IV

Kohavi, R. and John, G. H. (1997).
Wrappers for feature subset selection.
Artificial intelligence, 97(1-2):273–324.

Ramsey, J. (2006).
A pc-style markov blanket search for high dimensional datasets.
Technical Report.

Statnikov, A., Lytkin, N. I., Lemeire, J., and Aliferis, C. F. (2013).
Algorithms for discovery of multiple markov boundaries.
Journal of Machine Learning Research, 14(Feb):499–566.

References V

Statnikov, A., Ma, S., Henaff, M., Lytkin, N., Efstathiadis, E., Peskin, E. R., and Aliferis, C. F. (2015).
Ultra-scalable and efficient methods for hybrid observational and experimental local causal pathway discovery.
The Journal of Machine Learning Research, 16(1):3219–3267.

Tsamardinos, I., Aliferis, C. F., Statnikov, A. R., and Statnikov, E. (2003).
Algorithms for large scale markov blanket discovery.
In *FLAIRS conference*, volume 2, pages 376–380.

References VI

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78.