発表番号:66

データ拡張による因果グラフ的事前知識の予測モデリングへの活用

手嶋 毅志[1 2], 杉山 将[2 1]
1 東京大学　2 理化学研究所
2021年11月12日
第24回情報論的学習理論ワークショップ(IBIS2021)

概要

因果グラフとは	データ生成過程の定性的知識の簡潔な表現法
本研究の問い	因果グラフの知識を予測に活用できるか？
アイデア	データ拡張により統計的機械学習に取り入れる

本発表は “Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation.” (UAI 2021) に基づいています

本研究の全体像

因果グラフとは

- データ生成過程における変数間の依存関係を親子関係に反映した有向グラフ 

- 確率分布の条件付き独立性等を読み取れる

本研究の問い

因果グラフが所与のとき、その知識を
どのように教師付き学習に取り込めるか？

アイデアと結果

- データ拡張による方法を提案
- 理論、実験の両面から有効性を検証

メッセージ

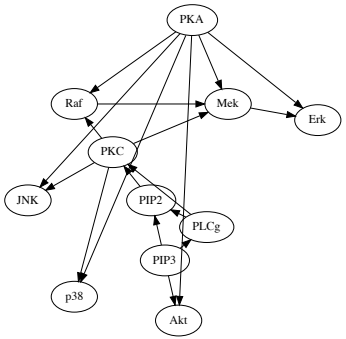
因果グラフで捉えた因果的メカニズムの知識は、
予測を目的とする統計的機械学習に役立つ

因果グラフとは

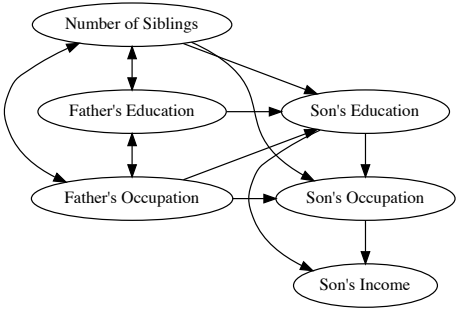
因果グラフ（CG）(e.g., Pearl, 2009)

データの生成過程の知識の簡潔な表現法 

- 頂点：データの各次元に対応する確率変数
- 有向辺：XがYの実現値を決める際に直接用いられているとき、XからYへの有向辺（矢線）を描く



Biology (Sachs et al., 2005)



Sociology (Duncan et al., 1972)

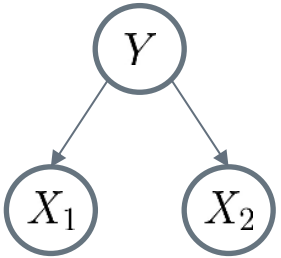
本発表では非巡回なもののみ考える．双方向辺は後のスライドで説明．

因果グラフの知識

因果グラフが示唆する知識

因果グラフが(ドメイン知識などから)既知のとき、
データ分布の条件付き独立性 (CI)を読み取れる(Pearl, 2009) (Richardson, 2003)

例:



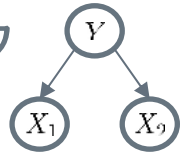
$$\Rightarrow \quad X_1 \perp\!\!\!\perp X_2 \mid Y$$

研究上の問い

因果グラフの形で得られた事前知識は、どう予測に活用できるか？

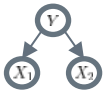
アイデア：データ拡張で独立性を取り込む

例（3変数のケース）

因果グラフ 

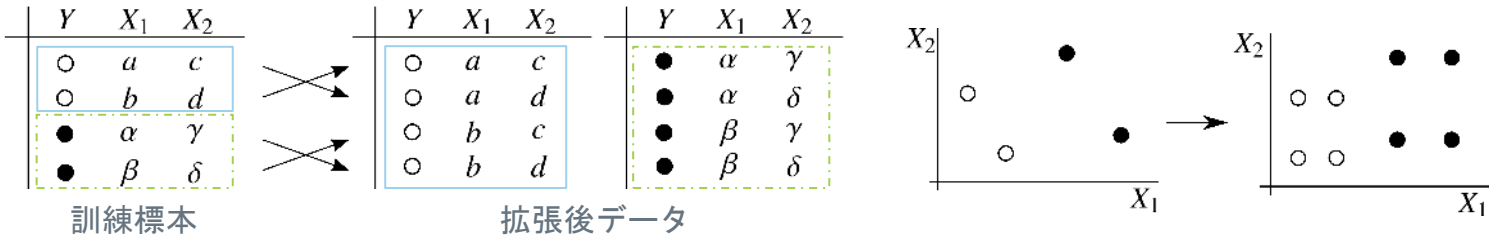
が与えられたもとで Y を (X_1, X_2) から予測する

アイデア：データ拡張

因果グラフ 

からは $X_1 \perp\!\!\!\perp X_2 \mid Y$ が従う

⇨学習データを Y でグループ化し、X_1 と X_2 を交換する



Q. 一般のグラフの場合はどうするか？

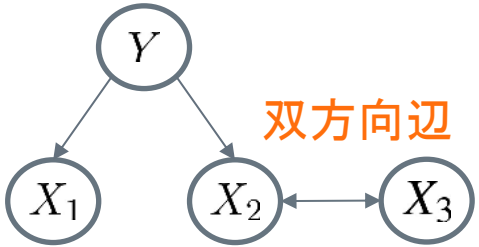
一般の因果グラフ

非巡回有向混合グラフ

(ADMGs; (Richardson, 2003) (Richardson et al., 2017))

非巡回有向グラフに、双方向辺も許したグラフ $\mathcal{G} = ([D], \mathcal{E}, \mathcal{B})$

潜在共通要因があるとき変数間に双方向辺を描く(準マルコフ因果モデル; Tian and Pearl, 2002).



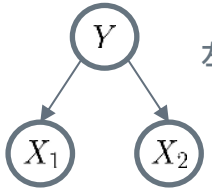
位相的ADMG分解

(Tian and Pearl, 2002) (Bhattacharya et al., 2020)

ADMG $\mathcal{G}$ を因果グラフに持つ確率変数 $\mathbf{Z} = (Z^1, \ldots, Z^D)$ は次を満たす:

$$p(\mathbf{Z}) = \prod_{j=1}^{D} p_{j|\mathrm{mp}(j)}(Z^j | \mathbf{Z}^{\mathrm{mp}(j)})$$

($\mathrm{mp}(j)$: 変数 Z^j のMarkov pillow
(「親」概念のADMGへの一般化))



左図のグラフ（双方向辺無し）の場合、

$$p(x_1, x_2, y) = p(x_1|y)p(x_2|y)p(y)$$

($\Leftrightarrow X_1 \perp\!\!\!\perp X_2 \mid Y$ に相当)

問題設定と目標

(X, Y)は区別せず $\mathbf{Z} = (Z^1, \ldots, Z^D)$ と表記

主仮定

$p(\mathbf{Z})$ は $\mathcal{G}$ に関して、位相的ADMG分解性を満たす

問題設定（教師付き学習）

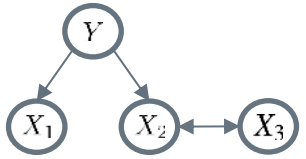
ラベル付きデータ $\mathcal{D} = \{\mathbf{Z}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p$ が所与

真のADMG $\mathcal{G}$ の推定値 $\hat{\mathcal{G}}$ (ADMG)が所与

目標：$R(f) = \mathbb{E}[\ell(f, \mathbf{Z})]$ が小さい予測器 $f : X \mapsto Y$ の学習

提案手法の導出

1. 位相的ADMG分解： $p(\mathbf{Z}) = \prod_{j=1}^{D} p_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)})$ ⟸



2. カーネル関数に基づく条件付き密度の推定量で置き換える

$$\hat{p}_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)}) := \frac{\sum_{i=1}^{n} \delta_{Z_i^j}(Z^j) K^j(\mathbf{Z}^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)})}{\sum_{k=1}^{n} K^j(\mathbf{Z}^{\mathrm{mp}(j)} - \mathbf{Z}_k^{\mathrm{mp}(j)})}$$

$K^j : \mathcal{Z}^{\mathrm{mp}(j)} \to \mathbb{R}_{\geq 0}$

経験条件付き密度(Stute, 1986)

3. リスクの定義式に代入すると、重み付き和によるリスク推定量になる

$$\hat{R}_{\mathrm{aug}}(f) = \int_{\mathcal{Z}} \ell(f, \mathbf{Z}) \prod_{j=1}^{D} \hat{p}_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)}) \mathrm{d}\mathbf{Z} = \sum_{\boldsymbol{i} \in [n]^D} \hat{w}_{\boldsymbol{i}} \cdot \ell(f, \mathbf{Z}_{\boldsymbol{i}})$$

解析的に導出できる

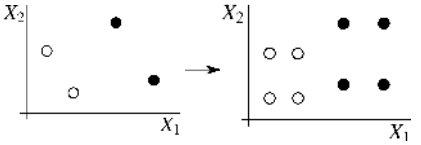
提案手法の導出

このリスク推定量の和の計算はデータ拡張として解釈可能

$$\hat{R}_{\mathrm{aug}}(f) = \sum_{\boldsymbol{i}\in[n]^D} \hat{w}_{\boldsymbol{i}} \cdot \ell(f, \mathbf{Z}_{\boldsymbol{i}})$$

$$\hat{w}_{\boldsymbol{i}} = \prod_{j=1}^{D} \frac{K^j(\mathbf{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}} - \mathbf{Z}^{\mathrm{mp}(j)}_{i_j})}{\sum_{k=1}^{n} K^j(\mathbf{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}} - \mathbf{Z}^{\mathrm{mp}(j)}_{k})}$$

$$\mathbf{Z}_{(1,1,2)} = \begin{bmatrix} A_1 \\ B_1 \\ C_2 \end{bmatrix}$$



$$\begin{bmatrix} A_1 \\ B_1 \\ C_1 \end{bmatrix} \begin{bmatrix} A_2 \\ B_2 \\ C_2 \end{bmatrix} \begin{bmatrix} A_3 \\ B_3 \\ C_3 \end{bmatrix}$$

$\mathbf{Z}_1$ $\mathbf{Z}_2$ $\mathbf{Z}_3$

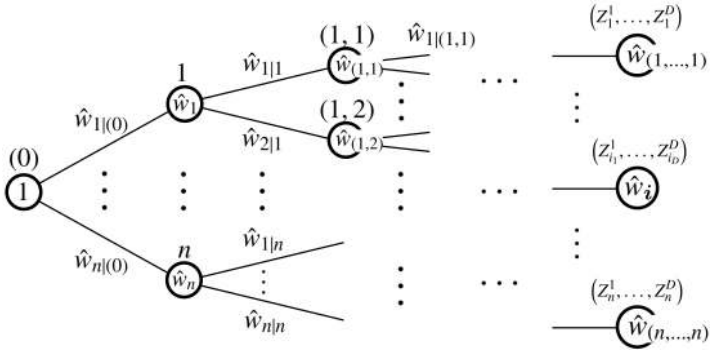
訓練標本

最初の3変数の例では、直観的に導出したデータ拡張法に一致する（重みは均等になる）

提案手法

- 重みはデータ拡張の過程で樹形図を構成することで再帰的に計算

$$\hat{R}_{\mathrm{aug}}(f) = \sum_{\boldsymbol{i} \in [n]^D} \hat{w}_{\boldsymbol{i}} \cdot \ell(f, \mathbf{Z}_{\boldsymbol{i}})$$



- 因果グラフの推定誤差を考慮し、加重平均を用いて学習

$$(1-\lambda)\hat{R}_{\mathrm{emp}}(f) + \lambda \hat{R}_{\mathrm{aug}}(f)$$

$\hat{R}_{\mathrm{emp}}$：経験リスク

$\lambda \in [0, 1]$

実験では全データセットで $\lambda = .5$ を使用

理論解析

Q. 提案手法は統計的にはどのように役立つのか？

設定および鍵となる仮定

- 真の因果グラフが存在し，これが完全に推定できている：$\hat{\mathcal{G}} = \mathcal{G}$
- 真の密度関数とカーネル関数が，滑らかさと有界性の仮定を満たす

定理（余剰リスク上界; インフォーマル）

$\hat{f} \in \arg\min_{f \in \mathcal{F}} \{\hat{R}_{\text{aug}}(f)\},\ f^* \in \arg\min_{f \in \mathcal{F}} \{R(f)\}$

$$R(\hat{f}) - R(f^*) \leq \underbrace{C_1 R_{\mathbf{H}} + C_p}_{\text{Kernel Bias}} + \underbrace{C_2 R_K + C_3 R_{\mathcal{F}.K}}_{\text{Complexity terms}} + \underbrace{C_4 \sqrt{\frac{\log(4D/\delta)}{2n}}}_{\text{Uncertainty}}$$

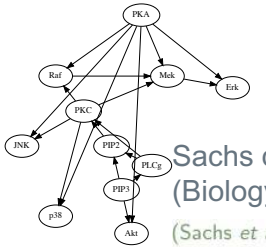
w/ high probability.

- 複雑度項は、通常の経験リスク最小化のRademacher複雑度よりサンプルサイズ依存性が改善：過学習の抑制効果を示唆（データ増加⇨過学習がより困難に）
- しかしカーネル関数による近似に由来するバイアス項（近似誤差）が出現（条件付き分布の条件部分の等号を、カーネル関数で近似しているため）

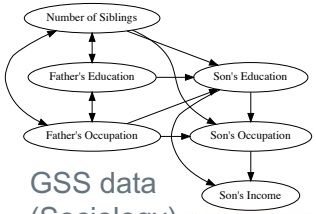
実験的評価

データと実験設定

- UCIレポジトリの6つのデータセット (Dua *et al.*, 2017)
- うち2つはドメイン知識に基づく因果グラフが存在



Sachs data (Biology) (Sachs *et al.*, 2005)



GSS data (Sociology) (Duncan *et al.*, 1972)

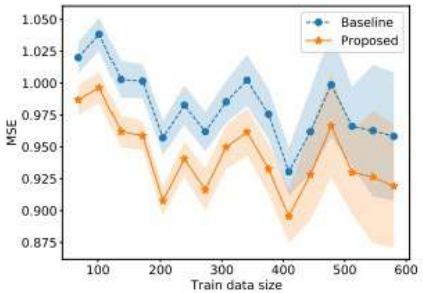
NAME	#VAR	#OBS
Sachs	11	853
GSS	6	1380
Boston Housing	14	506
Auto MPG	7	392
White Wine	12	4898
Red Wine	12	1599

- それ以外のデータセットにはDirectLiNGAMを適用し因果グラフを推定 (Shimizu *et al.*, 2011)

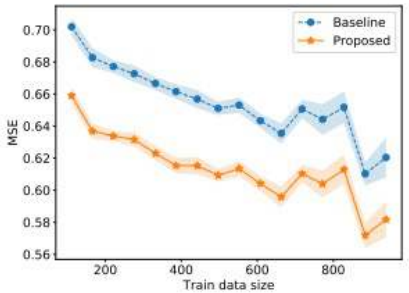
モデルと比較対象

- 予測器の仮説集合：勾配ブースティング回帰木 (Friedman, 2001) (Chen & Guestrin, 2016)
- ベースライン：工夫なしの教師付き学習 $\hat{f} \in \arg\min_{f \in \mathcal{F}} \{\hat{R}_{\mathrm{emp}}(f) + \Omega(f)\}$

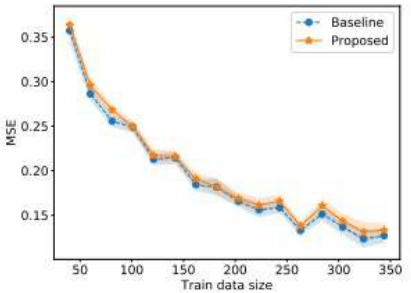
実験結果



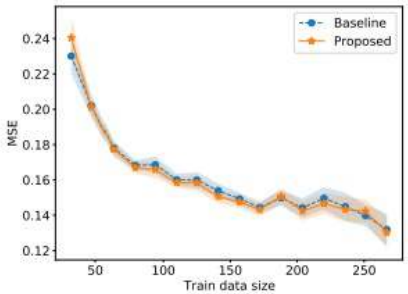
(a) Sachs data.



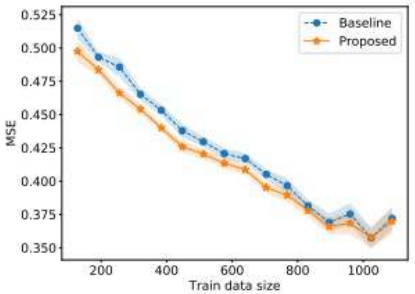
(b) GSS data.



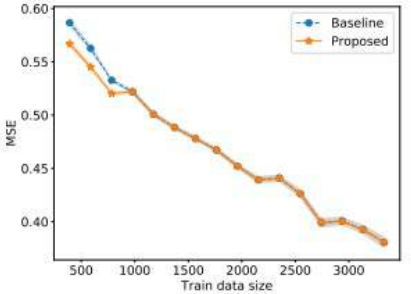
(c) Boston Housing data.



(d) Auto MPG data.



(e) Red Wine data.



(f) White Wine data.

- 特に小サンプル領域での予測性能が改善

まとめ

- 因果グラフで表現された事前知識を予測モデルの学習に活用するためのデータ拡張法を提案

- 提案法はデータ点を増やすことで過適合を抑制、一方でカーネル近似の近似誤差と複雑度が発生（条件付き密度の条件部をカーネル関数で表現）

- 実験的には追加の誤差と複雑度に見合う程度の性能改善が、特に小データ領域で、ドメイン知識による因果グラフが入手可能なときみられた