12. ベイジアンネットワーク分類器

電気通信大学
情報理工学研究科
植野真臣

今後のスケジュール

4月7日　授業の概要とガイダンス
4月14日　ベイズの定理
4月21日　ベイズはどのように誕生したか？
4月28日　　ベイズはコンピュータ、人工知能の父である！！
5月12日　　アランチューリングとベイズ
5月19日　　ビリーフとベイズ
5月26日　　尤度と最尤推定(1)
6月2日　　尤度と最尤推定(2)
6月9日　　ベイズ推定と事前分布(1)
6月16日　ベイズ推定と事前分布（２）
6月 23日　　階層ベイズ
6月30日　　データサイエンス：ルービン因果推論
7月7日　ベイジアンネットワークと因果推論
7月14日　ベイジアンネットワーク分類器
7月28日　　国際会議で休講
8月 4日　　テストと総括

本日の目標

１．ベイジアンネットワーク分類器
２．Augmented Naive Bayes Classifiers (ANB)
３．大規模ANBの学習
４．分類影響パラメータ数最小化による
　　ベイジアンネットワーク分類器学習

１．ベイジアンネットワーク分類器

1.1 ベイジアンネットワーク分類器（Friedman et al.,1988）

離散確率変数集合$\mathbf{V} = \{X_0, X_1, \cdots, X_n\}$をもつベイジアンネットワークについて，$X_0$を目的変数，$\mathbf{F} = \{X_1, \cdots, X_n\}$を説明変数集合とする．いま$\mathbf{F}$の値$\mathbf{x} = \{x_1, \cdots, x_n\}$を得たとき，以下のように$X_0$の推定値$\hat{c}$を得る．

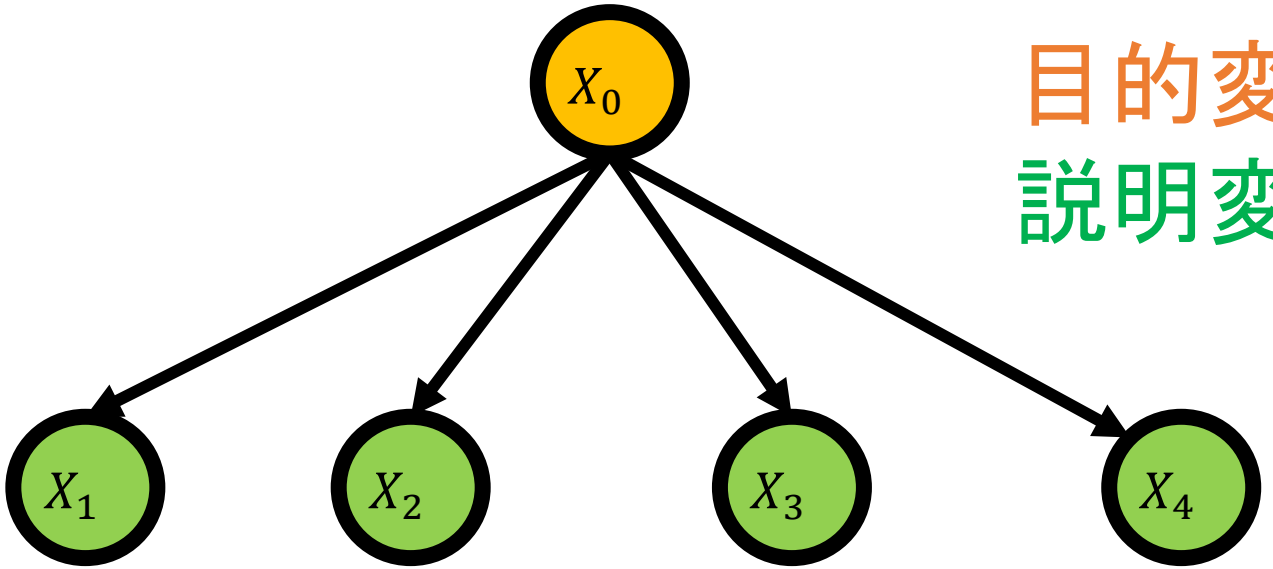
$$\begin{aligned}\hat{c} &= \underset{c=\{1,\cdots,r_0\}}{\operatorname{argmax}} P(X_0 = c \mid \mathbf{F} = \mathbf{x}) \\ &= \underset{c=\{1,\cdots,r_0\}}{\operatorname{argmax}} \frac{P(X_0 = c, \mathbf{F} = \mathbf{x})}{P(\mathbf{F} = \mathbf{x})} \\ &= \underset{c=\{1,\cdots,r_0\}}{\operatorname{argmax}} P(X_0 = c, \mathbf{F} = \mathbf{x}) .\end{aligned}$$

N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian Network Classifiers,” Machine Learning, vol.29, no.2, pp.131–163, 1997.

1.2 Friedman et al.(1997)による批判

周辺尤度で学習したベイジアンネットワークの分類精度が,
単純な構造をとるNaive Bayes(Minsky,1961)より劣ることが
多々ある.

Naïve Bayesの例



目的変数：X_0
説明変数：$X_1, \cdots, X_4$

N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian Network Classifiers,” Machine Learning, vol.29, no.2, pp.131–163, 1997.
Marvin Minsky. Steps toward Artificial Intelligence. In Proceedings of the IRE, volume 49, pp. 8–30, 1961.

1.2 Friedman et al.(1997)による批判

周辺尤度，MDL：同時確率分布$P(X_0, X_1, \cdots, X_n \mid G)$を表現する

生成モデルを学習

N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian Network Classifiers,” Machine Learning, vol.29, no.2, pp.131–163, 1997.

1.2 Friedman et al.(1997)による批判

周辺尤度，MDL：同時確率分布$P(X_0, X_1, \cdots, X_n \mid G)$を表現する

生成モデルを学習

分類が目的ならば，予測確率$P(X_0 \mid X_1, \cdots, X_n, G)$を表現する識別モデルを学習するべき

N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian Network Classifiers,” Machine Learning, vol.29, no.2, pp.131–163, 1997.

1.2 Friedman et al.(1997)による批判

$$\text{対数尤度} = \sum_{d=1}^{N} \log P\left(x_0^d, x_1^d, \ldots, x_n^d \mid G, \Theta\right)$$

$$= \underbrace{\sum_{d=1}^{N} \log P\left(x_0^d \mid x_1^d, \ldots, x_n^d, G, \Theta\right)}_{\text{分類に関与する } CLL(G, \Theta \mid D)} + \underbrace{\sum_{d=1}^{N} \log P\left(x_1^d, \ldots, x_n^d \mid G, \Theta\right)}_{\text{分類に関与しない}}$$

分類精度を高めるためには対数尤度$LL(D|G,\Theta)$ではなく，
条件付き対数尤度(Conditional Log Likelihood: CLL)のみを用いるべき．

N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian Network Classifiers,” Machine Learning, vol.29, no.2, pp.131–163, 1997.

定義1.3 Conditional MDL (Grossman and Domingos, 2004)

Conditional MDL(CMDL)スコアは以下で定義される.

$$CMDL(G, \Theta \mid D) = \frac{\log N}{2} \sum_{i=0}^{n} q_i(r_i - 1) - CLL(G, \Theta \mid D).$$

D. Grossman and P. Domingos, “Learning Bayesian Network classifiers by maximizing conditional likeli- hood,” Proceedings, Twenty-First International Con- ference on Machine Learning, ICML 2004, pp.361– 368, 2004.

1.4 CLLの問題点

CLLスコアは分解可能ではないため，構造探索に効率的なアルゴリズムが適用できず，厳密学習に膨大な時間がかかってしまう．

1.5 CLLの近似手法

- 構造探索に対して山登り法を適用した近似学習(Grossman et al.,2004)
- CLLが分解可能となるように近似したapproximated CLL(aCLL)スコア(Carvalho et al.,2013)
- CLLスコアが等価な構造を重複して探索しないような貪欲学習アルゴリズム(Mihaljević et al.,2018)

分類精度：
CLLを用いた近似手法　＞　周辺尤度を用いた近似手法

D. Grossman and P. Domingos. Learning Bayesian Network classifiers by maximizing conditional likelihood. In Proceedings of the International Conference on Machine Learning, pages 361–368, 2004.

A. M. Carvalho, P. Adão, and P. Mateus. Efficient Approximation of the Conditional Relative Entropy with Applications to Discriminative Learning of Bayesian Network Classifiers. Entropy, 15:2716–2735, 2013.

B. Mihaljević, C. Bielza, and P. Larrañaga. Learning Bayesian network classifiers with completed partially directed acyclic graphs. In Proceedings of the Interna tional Conference on Probabilistic Graphical Models, pages 272–283, 2018.

1.6 本当にCLLは良いのか？

- 周辺尤度最大化よりCLL最大化の方がなぜ良いのかという理由については未だ明らかになっていない.

- 周辺尤度を最大化する構造を厳密に学習できるにもかかわらず，先行研究では近似学習を行っている．このため，探索精度の悪さが結果に影響したのかもしれない.

1.7 周辺尤度厳密学習とCLL近似学習の分類精度比較

従来のCLLを用いた近似学習手法と，周辺尤度を用いた厳密学習手法の分類精度を比較する実験を行った.

1.8 周辺尤度厳密学習とCLL近似学習の分類精度比較

比較手法

- GBN(厳密)：周辺尤度を用いて厳密学習したBN
- Naive Bayes (Minsky,1961)
- BN-CMDL (Grossman and Domingos, 2004)：CMDLを用いて近似学習したBN
- BNC2P (Grossman and Domingos, 2004)：各変数が最大 2 つまでしか親を持たない構造を候補として，CLLを用いて近似学習したBN
- TAN-aCLL (Carvalho et al., 2013)：aCLLを用いて厳密学習したTAN
- BN(山登)：周辺尤度を用いて山登り法で近似学習したBN
- MC-DAGGES：CLLスコアが等価な構造を重複して探索しないような貪欲学習アルゴリズム(Mihaljević et al.,2018)

Marvin Minsky. Steps toward Artificial Intelligence. In Proceedings of the IRE, volume 49, pages 8–30, 1961.

D. Grossman and P. Domingos, “Learning Bayesian Network classifiers by maximizing conditional likeli- hood,” Proceedings, Twenty-First International Con- ference on Machine Learning, ICML 2004, pp.361– 368, 2004.

A. M. Carvalho, P. Ado, and P. Mateus. Efficient Approximation of the Conditional Relative Entropy with Applications to Discriminative Learning of Bayesian Network Classifiers. *Entropy*, 15(7): 2716–2735, 2013.

1.8 周辺尤度厳密学習とCLL近似学習の分類精度比較

No.	Dataset	Variables	Classes	Sample size	Naive-Bayes	GBN-CMDL	BNC2P	TAN-aCLL	GBN (山登)	MC-DAG GES	GBN (厳密)
1	Balance Scale	5	3	625	**0.9152**	0.3333	0.8560	0.8656	**0.9152**	0.7432	**0.9152**
2	banknote authentication	5	2	1372	0.8433	**0.8819**	0.8797	0.8761	**0.8819**	0.8768	0.8812
3	Hayes–Roth	5	3	132	0.8182	0.6136	0.6894	0.6742	0.7525	0.6970	0.6136
4	iris	5	3	150	0.7133	0.7800	0.8200	0.8200	0.8133	0.7800	**0.8267**
5	lenses	5	3	24	0.7500	0.8333	0.6667	0.7083	0.8333	0.8333	0.8333
6	Car Evaluation	7	4	1728	0.8571	**0.9497**	0.9416	0.9433	0.9416	0.9126	0.9416
7	liver	7	2	345	0.6319	0.6145	0.6290	**0.6609**	0.6029	0.6435	0.6087
8	MONK's Problems	7	2	432	0.7500	**1.0000**	**1.0000**	**1.0000**	0.8449	**1.0000**	**1.0000**
9	mux6	7	2	64	0.5469	0.3750	0.5625	0.4688	0.4063	**0.7656**	0.4531
10	LED7	8	10	3200	0.7294	0.7366	**0.7375**	0.7350	0.7297	0.7331	0.7294
11	HTRU2	9	2	17898	0.7031	0.7096	0.7070	0.7018	0.7188	0.7214	**0.7305**
12	Nursery	9	5	12960	0.6782	**0.7126**	0.6092	0.5862	**0.7126**	0.6322	**0.7126**
13	pima	9	2	768	0.8966	0.9086	0.9118	0.9130	0.9092	0.9093	0.9112
14	post	9	3	87	0.9033	0.5823	**0.9442**	0.9177	0.9291	0.9046	0.9340
15	Breast Cancer	10	2	277	**0.9751**	0.8917	0.9473	0.9488	0.7058	0.6354	**0.9751**
16	Breast Cancer Wisconsin	10	2	683	0.7401	0.6209	0.6823	0.7184	0.7094	**0.9780**	0.7184
17	Contraceptive Method Choice	10	3	1473	0.4671	0.4501	**0.4745**	0.4705	0.4440	0.4576	0.4542
18	glass	10	6	214	0.5561	0.5654	0.5794	0.6308	0.4626	0.5888	0.5701
19	shuttle-small	10	6	5800	0.9384	0.9660	0.9703	0.9583	0.9683	0.9586	0.9693
20	threeOf9	10	2	512	0.8164	**0.9434**	0.8691	0.8828	0.8652	0.8750	0.8887
21	Tic-Tac-Toe	10	2	958	0.6921	**0.8841**	0.7338	0.7203	0.6754	0.7557	0.8340
22	MAGIC Gamma Telescope	11	2	19020	0.7482	0.7849	0.7806	0.7631	0.7844	0.7781	0.7873
23	Solar Flare	11	9	1389	0.7811	0.8265	0.8315	0.8229	**0.8431**	0.8013	**0.8431**
24	heart	14	2	270	0.8259	0.8185	0.8037	0.8148	0.8222	**0.8333**	0.8259
25	wine	14	3	178	0.9270	**0.9438**	0.9157	0.9326	0.9045	**0.9438**	0.9270
26	cleve	14	2	296	0.8412	0.8209	0.8007	**0.8378**	0.7973	0.8041	0.7973
27	Australian	15	2	690	0.8290	0.8312	0.8348	0.8464	0.8420	0.8406	**0.8536**
28	crx	15	2	653	0.8377	0.8346	0.8208	0.8560	**0.8622**	0.8576	0.8591
29	EEG	15	2	14980	0.5778	0.6787	0.6374	0.6125	0.6732	0.6182	0.6814
30	Congressional Voting Records	17	2	232	0.9095	0.9698	0.9612	0.9181	**0.9741**	0.9009	0.9655
31	zoo	17	5	101	0.9802	0.9109	0.9505	1.0000	0.9505	0.9802	0.9307
32	pendigits	17	10	10992	0.8032	0.9062	0.8719	0.8700	0.9253	0.8359	**0.9290**
33	letter	17	26	20000	0.4466	0.5796	0.5132	0.5093	0.5761	0.4664	0.5761
34	ClimateModel	19	2	540	0.9222	**0.9407**	0.9241	0.9333	0.9370	0.9296	0.9000
35	Image Segmentation	19	7	2310	0.7290	0.7918	0.7991	0.7407	0.8026	0.7476	0.8156
36	lymphography	19	4	148	**0.8446**	0.7939	0.7973	0.8311	0.7905	0.8041	0.7500
37	vehicle	19	4	846	0.4350	0.5910	0.5910	0.5816	0.5461	0.5414	0.5768
38	hepatitis	20	2	80	0.8500	0.7375	**0.8875**	0.8750	0.8500	**0.8875**	0.5875
39	German	21	2	1000	0.7430	0.6110	0.7340	**0.7470**	0.7140	0.7180	0.7210
40	bank	21	2	30488	0.8544	0.8618	0.8928	0.8618	0.8952	0.8708	**0.8956**
41	waveform-21	22	3	5000	0.7886	0.7862	0.7754	0.7896	0.7698	0.7926	0.7846
42	Mushroom	22	2	5644	0.9957	**1.0000**	**1.0000**	0.9995	**1.0000**	0.9986	0.9949
43	spect	23	2	263	0.7940	0.7940	0.7903	0.8090	0.7603	0.8052	0.7378
	average				0.7764	0.7721	0.7936	0.7943	0.7867	0.7944	0.7963

1.9 高い分類精度を示す周辺尤度厳密学習

データセット	Variables	Classes	サンプルサイズ	Naive Bayes	GBN-CMDL	BNC2P	TAN-aCLL	GBN(貪欲)	MC-DAGGES	GBN(厳密)
HTRU2	9	2	17898	0.7031	0.7096	0.707	0.7018	0.7188	0.7214	0.7305
Nursery	9	5	12960	0.6782	0.7126	0.6092	0.5862	0.7126	0.6322	0.7126
MAGIC	11	2	19020	0.7482	0.7849	0.7806	0.7631	0.7844	0.7781	0.7873
EEG	15	2	14980	0.5778	0.6787	0.6374	0.6125	0.6732	0.6182	0.6814

サンプルサイズが大きいとき：

周辺尤度による厳密学習　＞　CLLによる近似学習

1.10 周辺尤度厳密学習の著しい分類精度低下

サンプルサイズの小さいデータセットでは
GBN(厳密)の分類精度が
他手法よりも著しく悪い.

No.	Dataset	Variables	Classes	Sample size	Naive-Bayes	GBN-CMDL	BNC2P	TAN-aCLL	GBN (山登)	MC-DAG GES	GBN (厳密)
1	Balance Scale	5	3	625	**0.9152**	0.3333	0.8560	0.8656	**0.9152**	0.7432	**0.9152**
2	banknote authentication	5	2	1372	0.8433	**0.8819**	0.8797	0.8761	**0.8819**	0.8768	0.8812
3	Hayes–Roth	5	3	132	0.8182	0.6136	0.6894	0.6742	0.7525	0.6970	0.6136
4	iris	5	3	150	0.7133	0.7800	0.8200	0.8200	0.8133	0.7800	**0.8267**
5	lenses	5	3	24	0.7500	0.8333	0.6667	0.7083	0.8333	0.8333	0.8333
6	Car Evaluation	7	4	1728	0.8571	**0.9497**	0.9416	0.9433	0.9416	0.9126	0.9416
7	liver	7	2	345	0.6319	0.6145	0.6290	**0.6609**	0.6029	0.6435	0.6087
8	MONK's Problems	7	2	432	0.7500	**1.0000**	**1.0000**	**1.0000**	0.8449	**1.0000**	**1.0000**
9	mux6	7	2	64	0.5469	0.3750	0.5625	0.4688	0.4063	**0.7656**	0.4531
10	LED7	8	10	3200	0.7294	0.7366	**0.7375**	0.7350	0.7297	0.7331	0.7294
11	HTRU2	9	2	17898	0.7031	0.7096	0.7070	0.7018	0.7188	0.7214	**0.7305**
12	Nursery	9	5	12960	0.6782	**0.7126**	0.6092	0.5862	**0.7126**	0.6322	**0.7126**
13	pima	9	2	768	0.8966	0.9086	0.9118	0.9130	0.9092	0.9093	0.9112
14	post	9	3	87	0.9033	0.5823	**0.9442**	0.9177	0.9291	0.9046	0.9340
15	Breast Cancer	10	2	277	**0.9751**	0.8917	0.9473	0.9488	0.7058	0.6354	**0.9751**
16	Breast Cancer Wisconsin	10	2	683	0.7401	0.6209	0.6823	0.7184	0.7094	**0.9780**	0.7184
17	Contraceptive Method Choice	10	3	1473	0.4671	0.4501	**0.4745**	0.4705	0.4440	0.4576	0.4542
18	glass	10	6	214	0.5561	0.5654	0.5794	0.6308	0.4626	0.5888	0.5701
19	shuttle-small	10	6	5800	0.9384	0.9660	0.9703	0.9583	0.9683	0.9586	0.9693
20	threeOf9	10	2	512	0.8164	**0.9434**	0.8691	0.8828	0.8652	0.8750	0.8887
21	Tic-Tac-Toe	10	2	958	0.6921	**0.8841**	0.7338	0.7203	0.6754	0.7557	0.8340
22	MAGIC Gamma Telescope	11	2	19020	0.7482	0.7849	0.7806	0.7631	0.7844	0.7781	0.7873
23	Solar Flare	11	9	1389	0.7811	0.8265	0.8315	0.8229	**0.8431**	0.8013	**0.8431**
24	heart	14	2	270	0.8259	0.8185	0.8037	0.8148	0.8222	**0.8333**	0.8259
25	wine	14	3	178	0.9270	**0.9438**	0.9157	0.9326	0.9045	**0.9438**	0.9270
26	cleve	14	2	296	0.8412	0.8209	0.8007	**0.8378**	0.7973	0.8041	0.7973
27	Australian	15	2	690	0.8290	0.8312	0.8348	0.8464	0.8420	0.8406	**0.8536**
28	crx	15	2	653	0.8377	0.8346	0.8208	0.8560	**0.8622**	0.8576	0.8591
29	EEG	15	2	14980	0.5778	0.6787	0.6374	0.6125	0.6732	0.6182	0.6814
30	Congressional Voting Records	17	2	232	0.9095	0.9698	0.9612	0.9181	**0.9741**	0.9009	0.9655
31	zoo	17	5	101	0.9802	0.9109	0.9505	1.0000	0.9505	0.9802	0.9307
32	pendigits	17	10	10992	0.8032	0.9062	0.8719	0.8700	0.9253	0.8359	**0.9290**
33	letter	17	26	20000	0.4466	0.5796	0.5132	0.5093	0.5761	0.4664	0.5761
34	ClimateModel	19	2	540	0.9222	**0.9407**	0.9241	0.9333	0.9370	0.9296	0.9000
35	Image Segmentation	19	7	2310	0.7290	0.7918	0.7991	0.7407	0.8026	0.7476	0.8156
36	lymphography	19	4	148	**0.8446**	0.7939	0.7973	0.8311	0.7905	0.8041	0.7500
37	vehicle	19	4	846	0.4350	0.5910	0.5910	0.5816	0.5461	0.5414	0.5768
38	hepatitis	20	2	80	0.8500	0.7375	**0.8875**	0.8750	0.8500	**0.8875**	0.5875
39	German	21	2	1000	0.7430	0.6110	0.7340	**0.7470**	0.7140	0.7180	0.7210
40	bank	21	2	30488	0.8544	0.8618	0.8928	0.8618	0.8952	0.8708	**0.8956**
41	waveform-21	22	3	5000	0.7886	0.7862	0.7754	0.7896	0.7698	0.7926	0.7846
42	Mushroom	22	2	5644	0.9957	**1.0000**	**1.0000**	0.9995	**1.0000**	0.9986	0.9949
43	spect	23	2	263	0.7940	0.7940	0.7903	0.8090	0.7603	0.8052	0.7378
	average				0.7764	0.7721	0.7936	0.7943	0.7867	0.7944	0.7963

1.10 周辺尤度厳密学習の著しい分類精度低下

BN(厳密)の分類精度が他手法より著しく悪い時のデータセット

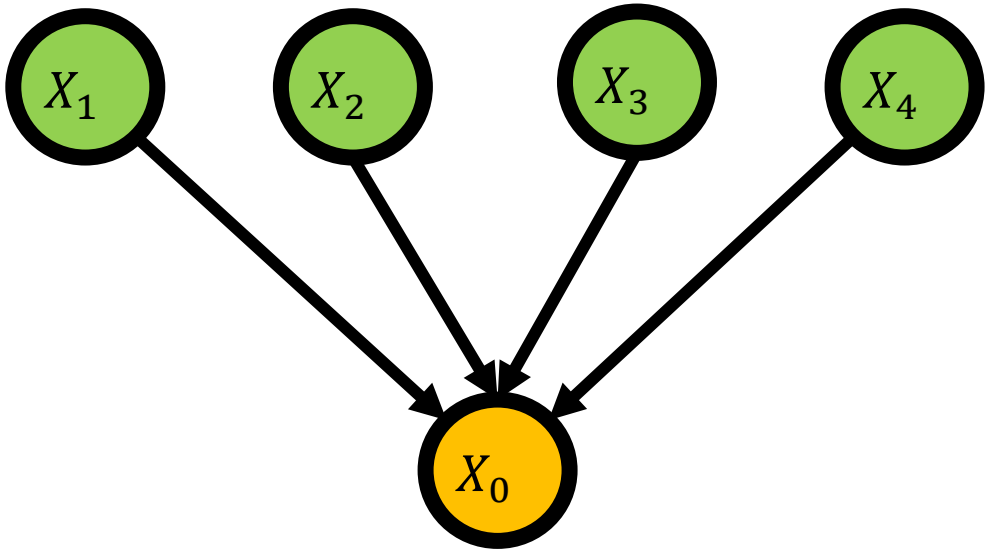
変数	サンプルサイズ	NB	BN-CMDL	BNC2P	TAN-aCLL	GBN(山登)	GBN(厳密)	GBN(厳密)の目的変数の親変数数
5	132	0.8182	0.8333	0.6364	0.6742	0.7879	0.6136	3.0
7	64	0.5469	0.3906	0.5625	0.4688	0.3750	0.4531	5.8
17	101	0.9802	0.8416	0.9505	1.0000	0.9406	0.9307	4.3

この時，周辺尤度による厳密学習では，
目的変数の親変数が多い構造を
学習していることがわかった．

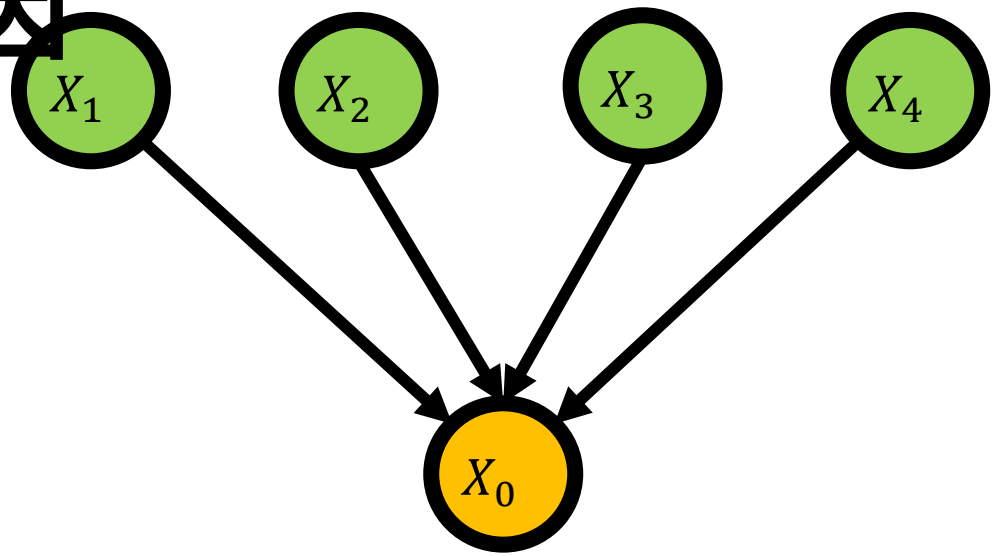
1.11 周辺尤度厳密学習の分類精度低下の原因

例. GBN(厳密)の分類精度が著しく悪い時の構造

親変数が多い



1.11 周辺尤度厳密学習の分類精度低下の原因



X_0の条件付き確率パラメータ
θ_{0jk}のEAP推定量

$$\frac{N'_{0jk} + N_{0jk}}{N'_{0j} + N_{0j}}$$

目的変数の親変数が増加

→ 目的変数の親変数集合の状態jのパターン数が指数的に増加

→ データが欠測するパターン数が増加

→ 目的変数のパラメータの推定精度が低下する.

1.11 周辺尤度厳密学習の分類精度低下の原因

実際に，目的変数の親変数が過多な構造では，
データが欠測するパターン数が増加していた.

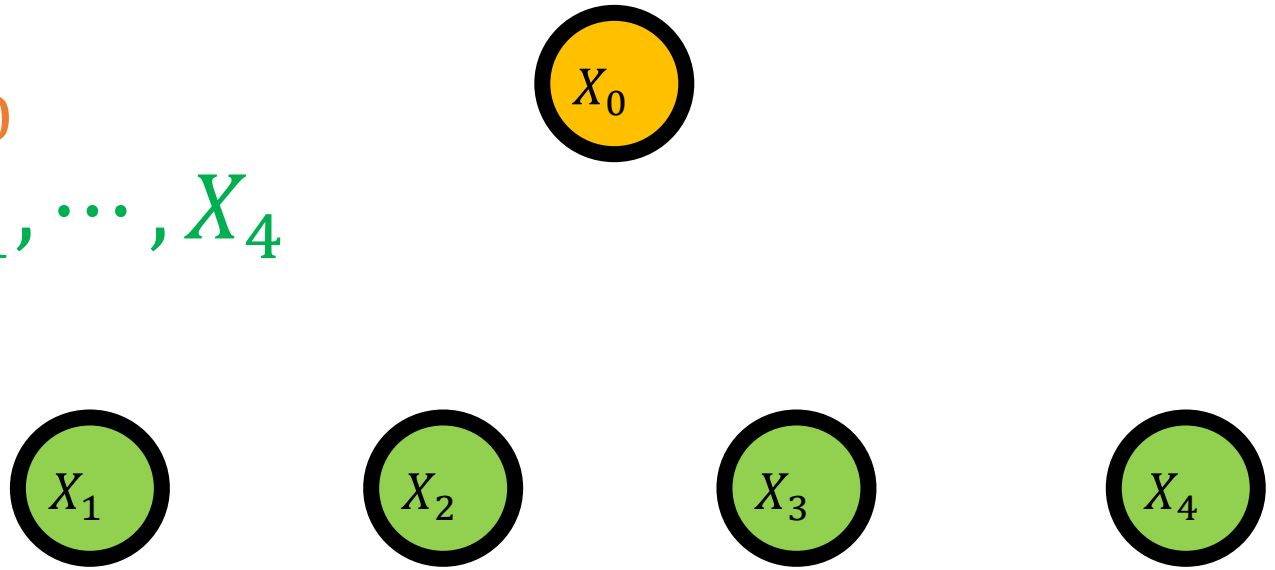
データセット	変数数	サムプルサイズ	目的変数の親変数数	データが欠測するパターン数
Hayes-Roth	5	132	3.0	17.2
mux6	7	64	5.8	5.2
zoo	17	101	4.3	20.3

2. Augmented Naive Bayes Classifiers (ANB)

2.1 Augmented Naïve Bayes Classifiers (ANB)

目的変数：X_0

説明変数：$X_1, \cdots, X_4$

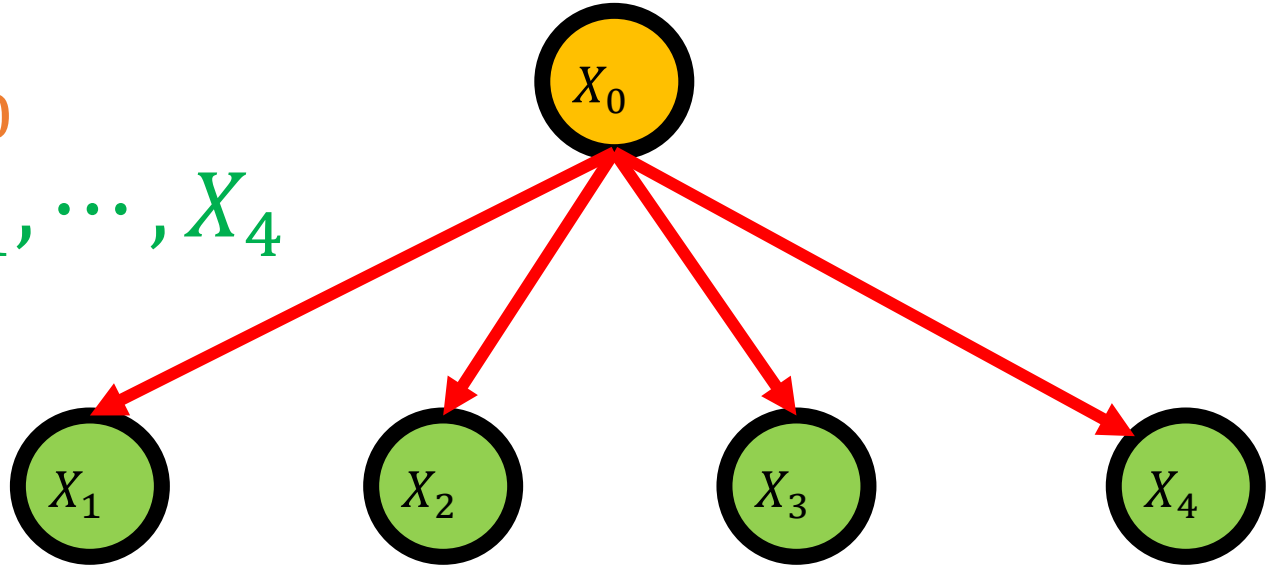


2.1 Augmented Naïve Bayes Classifiers (ANB)

まず強制的に目的変数から全説明変数へエッジを引く.

目的変数：X_0

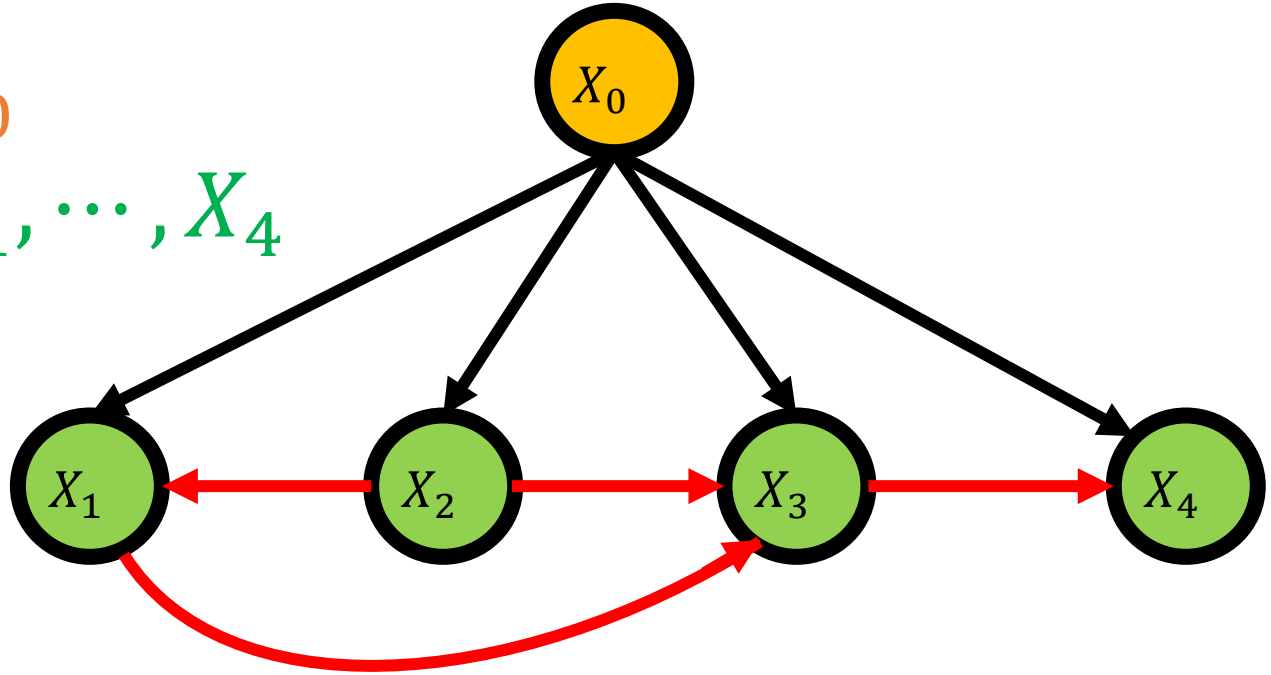
説明変数：$X_1, \cdots, X_4$



2.1 Augmented Naïve Bayes Classifiers (ANB)

まず強制的に目的変数から全説明変数へエッジを引く.

目的変数：X_0
説明変数：$X_1, \cdots, X_4$



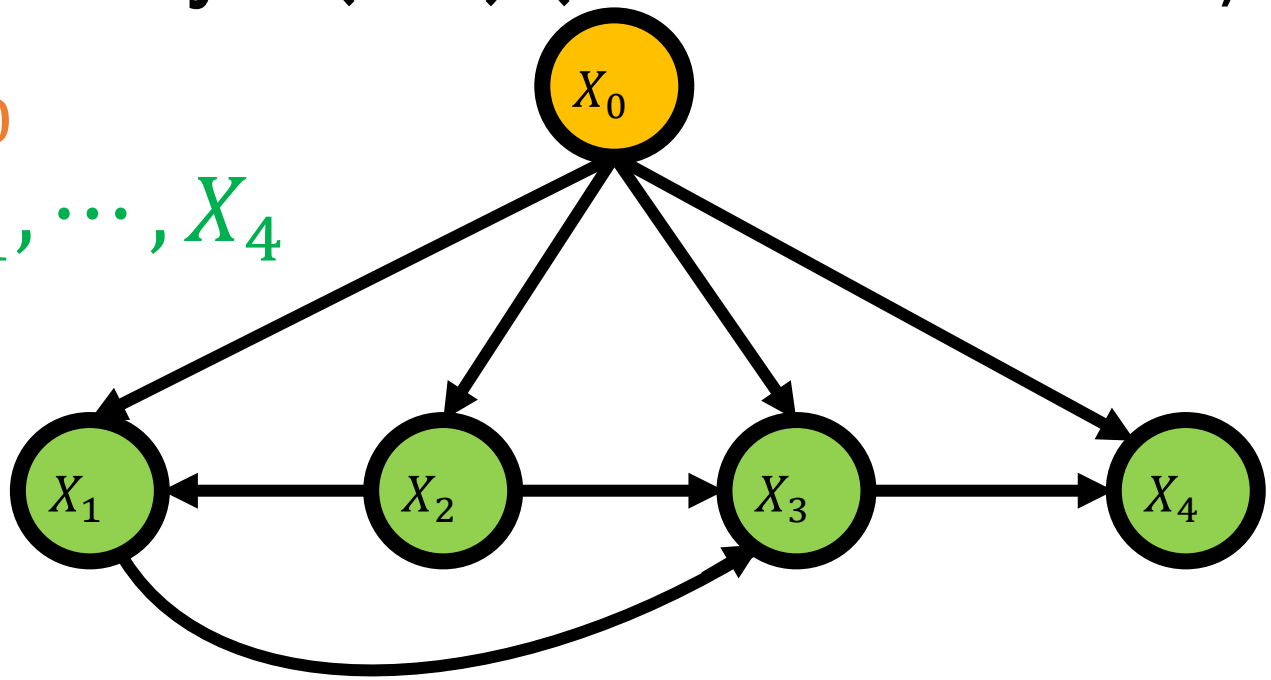
目的変数からのエッジを所与として
説明変数間の構造を厳密学習.

2.1 Augmented Naïve Bayes Classifiers (ANB)

目的変数から全説明変数にエッジが引かれている構造を，
Augmented Naïve Bayes(ANB)(Friedman et al.,1997)と呼ぶ.

目的変数：X_0

説明変数：$X_1, \cdots, X_4$



目的変数の親変数が0で，全ての説明変数を子変数とするため，先述した分類精度低下の問題が生じない.

N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian Network Classifiers,” Machine Learning, vol.29, no.2, pp.131–163, 1997.

2.2 ANBの厳密学習アルゴリズム

全てのANB構造の集合から，周辺尤度を最大にする構造を探索.

ANBでは目的変数は親を持たないため，以下のスコアを最大にすればよい.

$$BDeu_{ANB}(G,D) = BDeu(G,D) - LocalBDeu_0(\emptyset,D).$$

Silander and Myllymaki (2006)が提案した動的計画法による通常のベイジアンネットワークのための厳密学習アルゴリズムを，ANB学習に修正.

T. Silander and P. Myllyma¨ki. A Simple Approach for finding the Globally Optimal Bayesian Network Structure. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 445–452, 2006.

2.3 Notation

定義3.3 シンク (Pearl,1988)
子変数を持たない変数をシンクと呼ぶ.

定義3.4 最適親変数集合 (Silander and Myllymaki,2006)
変数集合$\mathbf{Z}, (X_0 \in \mathbf{Z})$のべき集合の中で$X_0$を含むものの集合を$\Pi(\mathbf{Z})$とすると，$X_i$の$\mathbf{Z}$に関する最適親変数集合は以下で定義される.

$$g_i^*(\mathbf{Z}) = \underset{\mathbf{W}\in\Pi(\mathbf{Z})}{\operatorname{argmax}} LocalBDeu_i(\mathbf{W}, D).$$

- 変数集合$\mathbf{Z}, (X_0 \in \mathbf{Z})$で構成されるANB構造の中で周辺尤度最大の構造を$G^*(\mathbf{Z})$で表す.
- $G^*(\mathbf{Z})$におけるあるシンクを$X_s^*(\mathbf{Z})$で表す.

T. Silander and P. Myllyma¨ki. A Simple Approach for finding the Globally Optimal Bayesian Network Structure. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 445–452, 2006.

2.4 ANBの厳密学習のステップ

１．説明変数$X_i \in \mathbf{F}$と変数集合$\mathbf{Z} \subseteq \mathbf{V} \setminus \{X_i\}, (X_0 \in \mathbf{Z})$の考えられる全ての組み合わせについて，ローカルスコア$LocalBDeu_i(\mathbf{Z}, D)$を計算する.

２．説明変数$X_i \in \mathbf{F}$と変数集合$\mathbf{Z} \subseteq \mathbf{V} \setminus \{X_i\}, (X_0 \in \mathbf{Z})$の考えられる全ての組み合わせについて，最適親変数集合$g^*(\mathbf{Z})$を計算する.

３．すべての変数集合$\mathbf{Z} \subseteq \mathbf{V}, (X_0 \in \mathbf{Z})$について，シンク$X_s^*(\mathbf{Z})$を計算する.

４．ステップ２と３を用いて$G^*(\mathbf{V})$を計算する.

2.5 ステップ3の計算方法

$G^*(\mathbf{Z})$はシンク$X_s^*(\mathbf{Z})$をもつ.

$G^*(\mathbf{Z})$において$X_s^*(\mathbf{Z})$は$g^*(\mathbf{Z} \setminus \{X_s^*(\mathbf{Z})\})$を親変数集合としてもっている.

$G^*(\mathbf{Z})$における$X_s^*(\mathbf{Z})$以外の変数は$G^*(\mathbf{Z} \setminus \{X_s^*(\mathbf{Z})\})$を構成する.

したがって,

$$X_s^*(\mathbf{Z}) = \underset{X_i \in \mathbf{Z} \setminus \{X_0\}}{\operatorname{argmax}} \{LocalBDeu_i(g_i^*(\mathbf{Z} \setminus \{X_i\}), D) + BDeu_{ANB}(G^*(\mathbf{Z} \setminus \{X_i\}), D)\}.$$

2.6 ステップ4の計算方法

$X_s^*(\mathbf{Z}) = \underset{X_i \in \mathbf{Z} \setminus \{X_0\}}{\operatorname{argmax}} \{LocalBDeu_i(g_i^*(\mathbf{Z} \setminus \{X_i\}), D) + BDeu_{ANB}(G^*(\mathbf{Z} \setminus \{X_i\}), D)\}$の式より，　$G^*(\mathbf{Z})$は$G^*(\mathbf{Z} \setminus \{X_s^*(\mathbf{Z})\})$と，$g^*(\mathbf{Z} \setminus \{X_s^*(\mathbf{Z})\})$から$X_s^*(\mathbf{Z})$に向かうエッジ集合と$X_s^*(\mathbf{Z})$に分解できる.

$G^*(\mathbf{V})$から再帰的に分解を行うことで，最終的にシンクとその最適親変数集合のペアn組に分解できる.

要素数の小さい$\mathbf{Z}$から順に$G^*(\mathbf{Z})$と$X_s^*(\mathbf{Z})$を計算していくことで，最後に$G^*(\mathbf{V})$が得られる.

2.7 ANB厳密学習アルゴリズムの効率性

このアルゴリズムで計算されるローカルスコア，最適親変数集合，シンクの数はそれぞれ$(n-1)2^{n-2}$，$(n-1)2^{n-2}$，2^{n-1}である.

ANB制約を課さないアルゴリズム(Silander and Myllymaki, 2006)におけるそれぞれの計算個数は$n2^{n-1}$，$n2^{n-1}$，2^nである.

ANB厳密学習アルゴリズムはANB制約を課さないアルゴリズムより約2倍速いと考えられる.

T. Silander and P. Myllyma¨ki. A Simple Approach for finding the Globally Optimal Bayesian Network Structure. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 445–452, 2006.

2.8 ANB厳密学習の漸近的性質①

定理2.8

$N \to \infty$のとき，厳密学習したANBはパラメータ数最小のI-map ANBに概収束する.

2.9 ANB厳密学習の漸近的性質②

定理2.9

以下の仮定１～３のもとで，厳密学習したANBは，真の構造G^*について，以下の関係を満たすような構造$\hat{G}$に概収束する.

任意の有限データセットD'について，

$$P\left(X_0 \mid \mathbf{F}, \hat{G}, D'\right) = P(X_0 \mid \mathbf{F}, G^*, D').$$

この関係を$\hat{G}$とG^*が分類等価であると呼ぶ.

仮定１　パーフェクトマップが存在する.

仮定２　すべての説明変数が，分類に影響を及ぼす変数集合（真の構造における目的変数のマルコフブランケット）に含まれる.

仮定３　真の構造において目的変数のマルコフブランケットに含まれる変数は目的変数と隣接する.

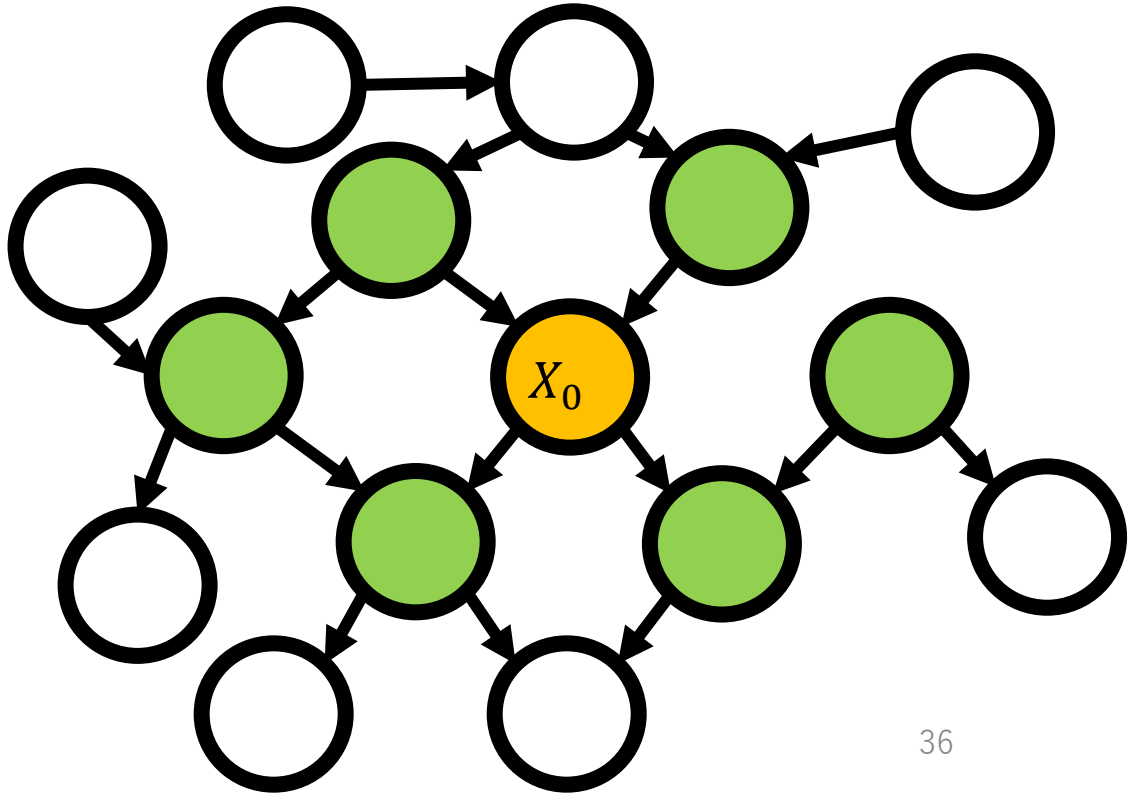
定義2.10 マルコフブランケット（Pearl, 1988）

変数集合**V**におけるX_0のマルコフブランケットとは，
以下を満たすような変数集合**M**である．

$$\forall X \notin \mathbf{M}, I(X, X_0 \mid \mathbf{M}).$$

特に，真の構造G^*における目的変数の子変数，親変数，共有する子変数をもつ変数の集合はマルコフブランケットである．

例．右の構造の場合，緑色で示された変数の集合がX_0のマルコフブランケットである．



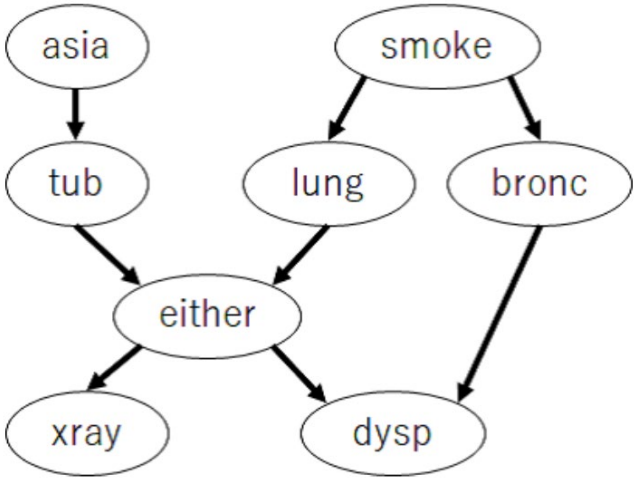
Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kanfmann, San Mateo, CA.

2.11 分類等価による利点

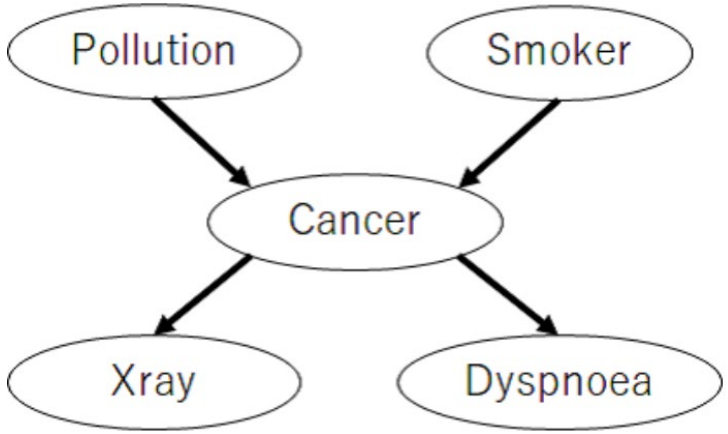
ANBの制約は一般にパラメータ数を増加させてしまうデメリットがあるが，仮定１～３のもとで，$N \to \infty$のとき，厳密学習されたANBは真の構造と全く同じ分類性能を持つ構造に概収束する．

2.12 定理の実証実験

仮定２と３を共に満たさないネットワークASIAと，共に満たすネットワークCancerを用いる.
真の構造とANB厳密学習それぞれの分類確率を計算し，Kullback-Leibler divergence (KLD)を測定した.
パラメータ数最小のANBと，ANB厳密学習の推定構造のstructural Hamming distance (SHD)を測定した.
SHDは構造間の距離のようなものを表す.
各ネットワークから100，500，1,000，10,000，50,000，100,000サンプル発生させ，それぞれについてKLDとSHDを測定した.



仮定２と３を共に満たさないネットワークASIA



仮定２と３を共に満たすネットワークCancer

2.13 パラメータ数最小のI-map ANBを学習できることの実証

Network	Variables	Sample size	SHD-(Proposal, I-map ANB)	KLD-(Proposal, True structure)
仮定２と３を満たさないネットワーク ASIA	8	100	3	2.31×10^{-2}
		500	2	1.24×10^{-1}
		1000	2	7.63×10^{-2}
		5000	1	3.67×10^{-3}
		10000	0	9.26×10^{-4}
		50000	0	6.28×10^{-4}
		100000	0	3.59×10^{-5}
仮定２と３を満たすネットワーク CANCER	5	100	1	8.79×10^{-2}
		500	1	2.43×10^{-3}
		1000	0	0.00
		5000	0	0.00
		10000	0	0.00
		50000	0	0.00
		100000	0	0.00

2.14 真の構造と分類等価な構造を学習できることの実証

Network	Variables	Sample size	SHD-(Proposal, I-map ANB)	KLD-(Proposal, True structure)
仮定２と３を満たさないネットワーク ASIA	8	100	3	2.31×10^{-2}
		500	2	1.24×10^{-1}
		1000	2	7.63×10^{-2}
		5000	1	3.67×10^{-3}
		10000	0	9.26×10^{-4}
		50000	0	6.28×10^{-4}
		100000	0	3.59×10^{-5}
仮定２と３を満たすネットワーク CANCER	5	100	1	8.79×10^{-2}
		500	1	2.43×10^{-3}
		1000	0	0.00
		5000	0	0.00
		10000	0	0.00
		50000	0	0.00
		100000	0	0.00

2.14 真の構造と分類等価な構造を学習できることの実証

Network	Variables	Sample size	SHD-(Proposal, I-map ANB)	KLD-(Proposal, True structure)
仮定２と３を満たさないネットワーク ASIA	8	100	3	2.31×10^{-2}
		500	2	1.24×10^{-1}
		1000	2	7.63×10^{-2}
		5000	1	3.67×10^{-3}
		10000	0	9.26×10^{-4}
		50000	0	6.28×10^{-4}
		100000	0	3.59×10^{-5}
仮定２と３を満たすネットワーク CANCER	5	100	1	8.79×10^{-2}
		500	1	2.43×10^{-3}
		1000	0	0.00
		5000	0	0.00
		10000	0	0.00
		50000	0	0.00
		100000	0	0.00

2.15 ANB厳密学習が分類精度を改善

	Naive-Bayes	GBN-CMDL	BNC2P	TAN-aCLL	BN(山登)	MC-DAG GES	BN(厳密)	ANB(厳密)
average	0.7764	0.7721	0.7936	0.7943	0.7867	0.7944	0.7963	0.8061
p-value (*ANB-BDeu* vs. the other methods)	0.00308	0.04136	0.00672	0.05614	0.06876	0.06010	0.22628	-

BN(厳密)で目的変数が親変数を多くもっていた時のデータセットの分類精度

変数数	サンプルサイズ	BN(厳密)	ANB(厳密)
5	132	0.6136	0.8333
7	64	0.4531	0.5469
17	101	0.9307	0.9505

2.16 変数選択

分類等価の定理では，すべての説明変数が真の構造における目的変数のマルコフブランケットに含まれることを仮定しているが，それは一般には成り立たない.

この問題を解決するには，事前に目的変数のマルコフブランケットのみを変数選択する必要がある.

仮定３が成り立つとき，目的変数のマルコフブランケットは目的変数の親変数と子変数（Parents and Children : PC）の集合に一致するため，これを探索する.

PC集合の厳密学習法として，SSL(Niinimäki and Parviainen, 2012)やS^2TMB(Gao and Ji,2017)があるが，変数数が増加すると計算時間が指数的に増加するため，30変数程度が学習の限界.

Teppo Niinimäki and Pekka Parviainen. Local Structure Discovery in Bayesian Networks. In Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI'12, pages 634–643. AUAI Press, 2012. ISBN 9780974903989.

Tian Gao and Qiang Ji. Efficient score-based Markov Blanket discovery. International Journal of Approximate Reasoning, 80:277–293, 2017. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2016.09.009.

2.17 PC探索手法

PCの厳密学習手法より効率的な手法として条件付き独立性検定(CIテスト)を用いた以下のような変数選択手法が知られている.

- MMPC (Tsamardinos et al. 2006)
- HITON-PC (Aliferis et al. 2003)
- PCMB (Pena 2007)

これらの手法では, 目的変数と説明変数の間でCIテストを行い,
独立性が検出された説明変数はPC集合から取り除く.
しかし, 上記の手法はすべてCIテストとして条件付き相互情報量や統計的検定を用いており, 漸近的に真の独立性を検出する保証がない.

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

C.F. Aliferis, I. Tsamardinos, and A. Statnikov. HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection. AMIA Annual Symposium proceedings, pages 21–25, 2003.

J.M. Pena, R. Nilsson, J. Björkegren, J. Tegnér. Towards scalable and data efficient learning of Markov boundaries International Journal of Approximate Reasoning, 45 (2) (2007), pp. 211-232

2.18 Bayes factor

漸近的に真の独立性を検出するCIテストとして，Steck and Jaakkola (2002)が周辺尤度を用いたBayes factorを提案している．

定義3.12 周辺尤度によるBayes factor（Steck and Jaakkola (2002)）

二変数X, Yと変数集合$\mathbf{Z}$について，周辺尤度を用いた対数Bayes factor $\log BF_D(X, Y \mid \mathbf{Z})$は以下で定義される．

$$\log BF_D(X, Y \mid \mathbf{Z}) = LocalBDeu_X(\mathbf{Z}, D) - LocalBDeu_Y(\mathbf{Z} \cup \{Y\}, D).$$

$BF_D(X, Y \mid \mathbf{Z}) \geq 1$のとき$I(X, Y \mid \mathbf{Z})$と判定し，
$BF_D(X, Y \mid \mathbf{Z}) < 1$のとき$\neg I(X, Y \mid \mathbf{Z})$と判定する．

Harald Steck and Tommi S. Jaakkola. On the Dirichlet Prior and Bayesian Regularization. In Proceedings of the 15th International Conference on Neural Information Processing Systems, NIPS'02, pages 713–720. MIT Press, 2002b.

2.19 Bayes factorを用いた変数選択を適用した ANB厳密学習は最高の分類精度を示す

	Naive-Bayes	GBN-CMDL	BNC2P	TAN-aCLL	GBN (山登)	MC-DAG GES	GBN (厳密)	ANB (厳密)	ANB (厳密，変数選択適用)
average	0.7764	0.7721	0.7936	0.7943	0.7867	0.7944	0.7963	0.8061	**0.8184**
p-value (*fsANB-BDeu* vs. the other methods)	0.00001	0.00014	0.00013	0.00280	0.00015	0.00212	0.00064	0.01101	-

分類精度：
ANB(厳密，変数選択適用) ＞ 比較手法 （有意水準 0.05）

３．大規模ANBの学習

3.1 スコアベースアプローチの問題点

スコアベースアプローチは，構造の探索数がノード数に対し指数的に増加してしまう．

3.2 厳密学習手法（学習可能な変数数）

- 動的計画法：29変数
  - T. Silander and P. Myllymaki, “A simple approach for finding the globally optimal Bayesian network structure,” in Uncertainty in Artificial Intelligence (UAI), 445–452, AUAI Press, 2006
- A* 探索：29変数
  - Yuan, C., and Malone, B., Learning optimal Bayesian networks: A shortest path perspective. Journal of Artificial Intelligence Research, 48, 23–65, 2013.
- 幅優先分枝限定法：33変数
  - Malone, B., Yuan, C., Hansen, E., and Bridges, Improving the scalability of optimal Bayesian network learning with external-memory frontier breadth-first branch and bound search. in Uncertainty in Artificial Intelligence, 479–488, 2011
- 整数計画法：60変数
  - J. Cussens, “Bayesian network learning with cutting planes,” in Uncertainty in Artificial Intelligence (UAI), 153–160, AUAI Press, 2011.

3.2 厳密学習手法（学習可能な変数数）

- 動的計画法：29変数
  - T. Silander and P. Myllymaki, “A simple approach for finding the globally optimal Bayesian network structure,” in Uncertainty in Artificial Intelligence (UAI), 445–452, AUAI Press, 2006.

厳密学習手法の問題点

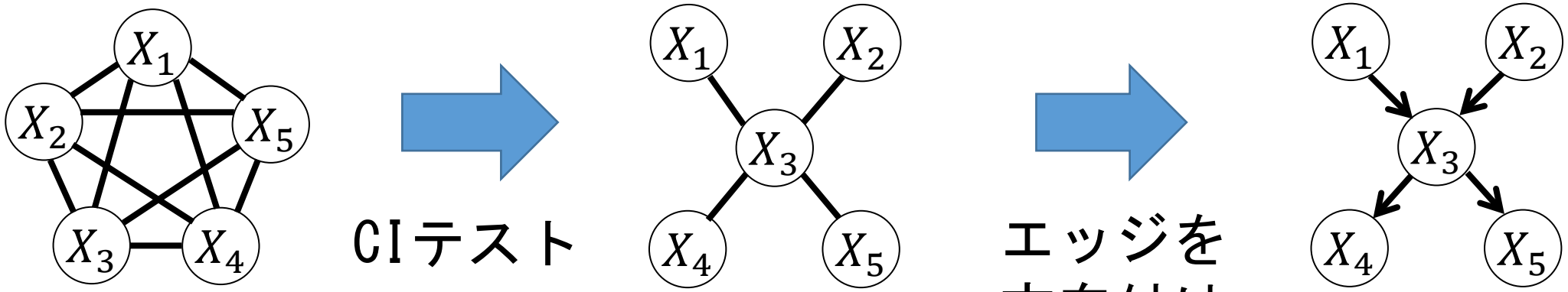
最先端手法を用いても60変数程度が限界

first branch and bound search. in Uncertainty in Artificial Intelligence, 479–488, 2011

- 整数計画法：60変数
  - J. Cussens, “Bayesian network learning with cutting planes,” in Uncertainty in Artificial Intelligence (UAI), 153–160, AUAI Press, 2011.

3.3 制約ベースアプローチ

- 漸近的に真の構造を学習する保証は持たないが効率的な学習法
- 条件付き独立性検定（CI テスト）によるエッジの削除とエッジの方向付けによる構造学習手法



完全無向グラフ　無向グラフ　推定構造

3.4 制約ベースアプローチの問題点

制約ベースアプローチの従来手法

- PCアルゴリズム（Spirtesら，2000）
- MMHCアルゴリズム（Tsamardinos，2006）
- RAIアルゴリズム（Yehezkel and Lerner，2009）

従来手法が用いるCIテストは漸近的に真の独立性を検出する保証がない.

P. Spirtes, C. Glymour, and R. Scheines, Causation, Prediction, and Search, MIT press, 2000.
I. Tsamardinos, L.E. Brown, and C.F. Aliferis, “The max-min hill-climbing Bayesian network structure learning algorithm,” Machine Learning, vol.65, no.1, pp.31–78, 2006.
R. Yehezkel and B. Lerner, “Bayesian network structure learning by recursive autonomy identification,” Journal of Machine Learning Research, vol.10, pp.1527–1570, 2009.

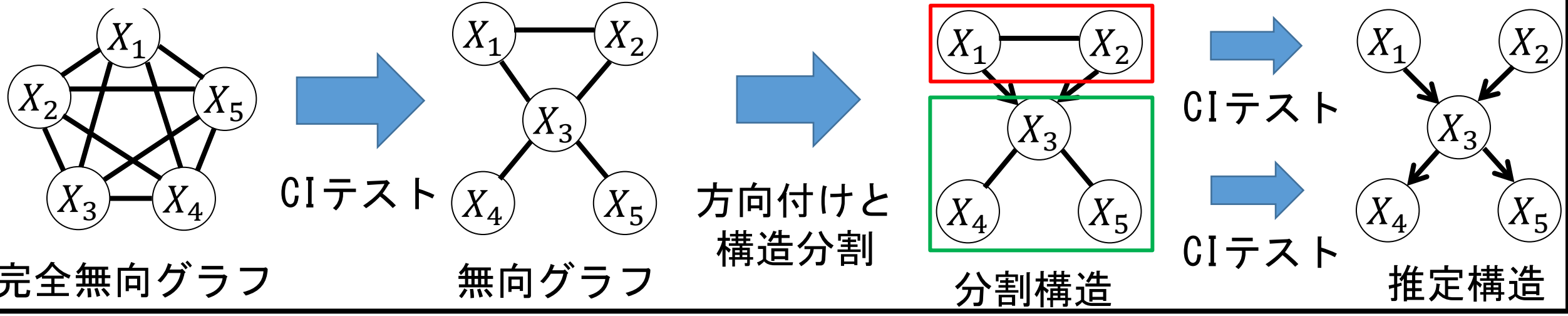
3.5 Bayes factorを用いたRAIアルゴリズム(Natoriら, 2017)

漸近一致性を有しつつ最も大規模な構造を学習できる手法

Bayes factor

$$CI(X,Y \mid \boldsymbol{Z}) = \begin{cases} \text{従属} & \left(\log BF_D(X,Y \mid \boldsymbol{Z}) \geq 0 \right) \\ \text{独立} & \left(\log BF_D(X,Y \mid \boldsymbol{Z}) < 0 \right) \end{cases}$$

RAIアルゴリズム



K. Natori, and M. Ueno, "Consistent learning Bayesian networks with thousands of variables," Advanced Methodologies for Bayesian Networks (Proceedings of Machine Learning Research), vol.73, pp.57–68, 2017.

3.6 Bayes factorを用いたRAIアルゴリズム

学習手順

入力：データ，完全無向グラフ，CIテストの次数$n_z = 0$

出力：推定されたグラフ

終了条件：各ノードが$n_z + 1$個未満の親ノードを持つ

１．各エッジに対してBayes factorを用いたCIテストとエッジの削除

２．各エッジを方向付け

３．グラフを部分グラフに分割

４．$n_z \leftarrow n_z + 1$として終了条件を満たすまで再帰的に動作

3.7 RAIANBアルゴリズムによる学習

完全無向グラフに制約を加え，CIテストの実行範囲を制限することでANB構造を学習する.

学習手順

入力：データ，説明変数からなる完全無向グラフ，
CIテストの次数$n_Z = 1$(条件部がX_0を必ず含む)，出力：推定されたグラフ

1．説明変数間の各エッジに対してBayes factorを用いたCIテストとエッジの削除

2．各エッジを方向付け

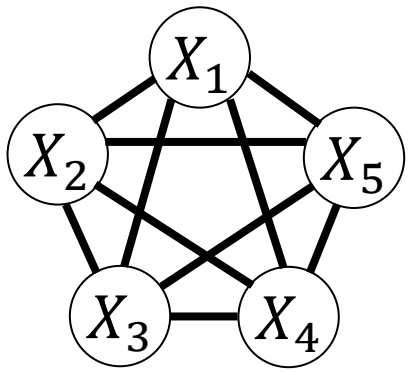
3．全体グラフを部分グラフに分割

4．$n_Z + 1$個未満の親を持つノードが存在する場合，$n_Z \leftarrow n_Z + 1$として再帰的に動作

5．目的変数から全ての説明変数へ向けてエッジを引く.

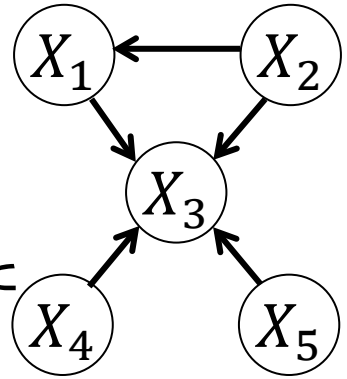
3.8 RAIANBの動作例

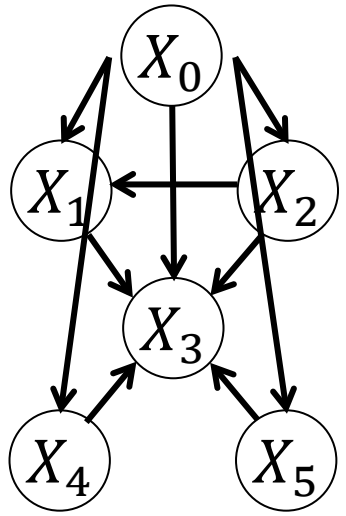
- 目的変数を含まない2変数に対してCIテストを行いエッジを削除
- 目的変数から全ての説明変数へ向けてエッジを方向付ける



説明変数からなる完全グラフ

X_0を所与とした
CIテスト





推定されたANB

56

3.9 RAIANBアルゴリズムの効率性

RAIANBアルゴリズムは以下の理由により通常のRAIアルゴリズムよりも高速に学習できる.

・RAIANBアルゴリズムでは目的変数と説明変数間のCIテストを行う必要がない.

・真の構造がANB構造である場合，最初から目的変数を所与としたCIテストを行うRAIANBの方が通常のRAIより早く独立性を検出するため，早期にエッジを除去することができる．これにより，アルゴリズム中の構造の分解が加速する.

3.10 RAIANBアルゴリズムの漸近的性質

定理4.10

$N \to \infty$のとき，RAIANBアルゴリズムで学習した構造は，パラメータ数最小のI-map ANBに概収束する.

3.11 RAIANBと従来手法の分類精度比較

- Naïve Bayes
  - 全ての説明変数が目的変数のみを親に持つ
- Tree Augmented Naïve Bayes (TAN)
  - 全ての説明変数が目的変数を親に持ち、説明変数間で木構造をとる
- 厳密学習手法 (数十変数が限界)
  - GBN(厳密)
    - 動的計画法で厳密学習したGBN
  - ANB(厳密)
    - 動的計画法で厳密学習したANB
- 制約ベース手法
  - RAI-GBN
    - CIテストにBayes factorを用いて構造学習したGBN
  - RAI-ANB
    - RAIANBアルゴリズムで学習したANB

3.12 小規模データにおける精度

	dataset	variable	number of data	classes	Naive Bayes	TAN	GBN-CMDL	BNC2P	TAN-aCLL	GBN (厳密)	ANB (厳密)	RAI-GBN	RAI-ANB
1	magic	11	19020	2	0.7447	0.7767	0.7849	0.7806	0.7631	**0.7865**	0.7863	0.7793	0.7790
2	Flare	11	1389	9	0.7804	0.7976	0.8265	0.8315	0.8229	**0.8430**	0.8265	0.8423	0.8178
3	heart	14	270	2	0.8296	0.8407	0.8185	0.8037	0.8148	**0.8444**	0.8148	0.7666	0.8333
4	wine	14	178	3	0.9205	0.9212	0.9438	0.9157	0.9326	0.9424	**0.9490**	0.9212	0.9150
5	Cleve	14	296	2	0.8309	0.8175	0.8209	0.8007	**0.8378**	0.8144	0.8309	0.7771	0.8212
6	Australian	15	690	2	0.8362	0.8304	0.8312	0.8348	0.8464	**0.8492**	0.8449	0.8405	0.8463
7	crx	15	653	2	0.8391	0.8483	0.8346	0.8208	**0.8560**	0.8481	0.8482	0.8544	0.8436
8	EEG	15	14980	2	0.5774	0.6298	0.6787	0.6374	0.6125	0.6843	**0.6937**	0.6421	0.6709
9	Congressional	17	232	2	0.9137	0.9398	0.9698	0.9612	0.9181	**0.9699**	**0.9699**	0.9655	0.9438
10	zoo	17	101	5	0.9709	0.9427	0.9109	0.9505	**1.0000**	0.9900	0.9700	0.8809	0.9418
11	pendigits	17	10992	10	0.7998	0.8477	0.9062	0.8719	0.8700	**0.9329**	0.9326	0.8757	0.9254
12	letter	17	20000	26	0.4456	0.4866	0.5796	0.5132	0.5093	0.5777	0.5950	0.5560	**0.6145**
13	ClimateModel	19	540	2	0.9203	0.9314	**0.9407**	0.9241	0.9333	0.9259	0.9055	0.9074	0.9203
14	ImageSegmentation	19	2310	7	0.7324	0.7510	0.7918	0.7991	0.7407	0.8233	**0.8290**	0.7839	0.8121
15	lymphography	19	148	4	0.8523	0.8109	0.7939	0.7973	0.8311	**0.8647**	0.7909	0.6842	0.8514
16	vehicle	19	846	4	0.4266	0.5472	0.5910	0.5910	0.5816	0.5910	**0.6417**	0.4893	0.6028
17	hepatitis	20	80	2	0.8750	0.8750	0.7375	0.8875	0.8750	**0.9250**	0.9000	0.8125	0.8875
18	German	21	5000	2	0.7440	0.7340	0.6110	0.7340	0.7470	0.7320	0.7420	0.7000	**0.7540**
19	bank	21	30488	2	0.8542	0.8774	0.8618	0.8928	0.8618	0.8954	0.8956	**0.8959**	0.8926
20	waveform-21	22	5000	3	0.7894	0.7914	0.7862	0.7754	0.7896	0.7938	**0.8048**	0.7328	0.7870
21	Mushroom	22	5644	2	0.9962	**1.0000**	**1.0000**	**1.0000**	0.9995	0.9946	**1.0000**	**1.0000**	**1.0000**
22	spect	23	263	2	0.7868	0.8101	0.7940	0.7903	0.8090	0.7759	**0.8207**	0.7937	0.8096
	Classification accuracy	Arithmetic average			0.7939	0.8094	0.8097	0.8143	0.8160	**0.8366**	0.8360	0.7955	0.8304
		p-value			0.0024	0.0117	0.0324	0.0099	0.0574	> 0.1	> 0.1	0.0013	-
	Runtime (s)	Arithmetic average			0.00	2.58	30.53	21.11	10.05	1790.93	500.76	26.06	3.14
		Geometric average			0.00	0.798	9.50	6.87	3.27	201.76	110.69	7.90	1.26

3.13 大規模データにおける精度

	dataset	variables	num of data	classes	Naive Bayes	TAN	RAI-GBN	RAI-ANB
1	kr-vs-kp	37	3196	2	0.8773	0.9239	0.9405	**0.9518**
2	Connect-4	43	67557	3	0.7212	0.7643	0.7467	**0.7973**
3	Flowmeters D	44	180	4	**0.8388**	**0.8388**	0.8055	0.8277
4	movement libras	91	360	15	0.5027	0.5388	0.1611	**0.5666**
5	dota2	117	102944	2	**0.5980**	0.5810	0.5435	0.5957
6	Musk1	167	478	2	0.6538	0.7565	0.6658	**0.8219**
7	Musk2	167	6598	2	0.7443	0.8408	0.8808	**0.9639**
8	Epileptic Seizure	179	11500	5	0.2344	0.3650	0.1886	**0.3820**
9	mfeat-fac	219	2000	10	0.3520	0.4590	0.3030	**0.4730**
10	semeion	257	1600	10	0.8556	0.8719	0.4106	**0.8794**
11	madelon	501	2000	2	0.5905	0.5270	**0.6280**	0.5830
12	pd speech features	755	756	2	0.7182	0.7897	0.7657	**0.8228**
13	pure-spectra-matrix	1301	571	20	0.9088	0.8984	0.4833	**0.9159**
	Classification accuracy	Arithmetic average			0.6612	0.7042	0.5787	**0.7370**
		p-value			0.0044	0.0012	0.0015	-
	Runtime (s)	Arithmetic average			0.0	545.7	2002.1	1665.9
		Geometric average			0.0	52.6	375.3	227.4

3.13 大規模データにおける精度

	dataset	variables	num of data	classes	Naive Bayes	TAN	RAI-GBN	RAI-ANB
1	kr-vs-kp	37	3196	2	0.8773	0.9239	0.9405	**0.9518**
2	Connect-4	43	67557	3	0.7212	0.7643	0.7467	**0.7973**
3	Flowmeters D	44	180	4	**0.8388**	**0.8388**	0.8055	0.8277
4	movement libras	91	360	15	0.5027	0.5388	0.1611	**0.5666**
5	dota2	117	102944	2	**0.5980**	0.5810	0.5435	0.5957
6	Musk1	167	478	2	0.6538	0.7565	0.6658	**0.8219**
7	Musk2	167	6598	2	0.7443	0.8408	0.8808	**0.9639**
8	Epileptic Seizure	179	11500	5	0.2344	0.3650	0.1886	**0.3820**
9	mfeat-fac	219	2000	10	0.3520	0.4590	0.3030	**0.4730**
10	semeion	257	1600	10	0.8556	0.8719	0.4106	**0.8794**
11	madelon	501	2000	2	0.5905	0.5270	**0.6280**	0.5830
12	pd speech features	755	756	2	0.7182	0.7897	0.7657	**0.8228**
13	pure-spectra-matrix	1301	571	20	0.9088	0.8984	0.4833	**0.9159**
	Classification accuracy	Arithmetic average			0.6612	0.7042	0.5787	**0.7370**
		p-value			0.0044	0.0012	0.0015	-
	Runtime (s)	Arithmetic average			0.0	545.7	2002.1	1665.9
		Geometric average			0.0	52.6	375.3	227.4

４. 分類影響パラメータ数最小化による ベイジアンネットワーク分類器学習

Shouta Sugahara, Koya Kato and Maomi Ueno: Learning Bayesian Network Classifiers to Minimize the Class Variable Parameters. Proceedings of the AAAI Conference on Artificial Intelligence, 38(18), 20540-20549. (2024) https://doi.org/10.1609/aaai.v38i18.30039.

4.1 NCP最小のI-map

BN分類器では，分類確率の推定に全ての変数のパラメータを用いるわけではなく，分類確率の推定に必要な変数のパラメータのみを用いる.
従来手法はパラメータ数最小のI-map ANBが得られることは保証するが，分類確率の推定に全ての変数のパラメータ数（Number of Class variable Parameters: NCP）を最小化した方が高精度な分類器が構成できると考えられる.

$$NCP(G) = \sum_{i=0}^{n} NCP_i(\mathbf{Pa}_{X_i}^G), \quad NCP_i\left(\mathbf{Pa}_{X_i}^G\right) = \begin{cases} (r_i - 1)q_i & (i = 0 \ \vee X_0 \in \mathbf{Pa}_{X_i}^G) \\ 0 & (\text{その他の場合}) \end{cases}$$

4.2 変数順序

変数順序： 構造Gの各変数を要素とするベクトルπに対し，πのi番目の要素をX_{π_i}で表すと，$\forall i, \mathbf{Pa}_{X_{\pi_i}}^{G} \subseteq \bigcup_{j=1}^{i-1}\{X_{\pi_j}\}$ が成り立つ時，πを変数順序という．

ここで，$\mathbf{Pa}_{X_i}^{G}$は変数X_iの親変数集合．

例） 変数順序(X_0, X_1, X_2, X_3)に従う構造

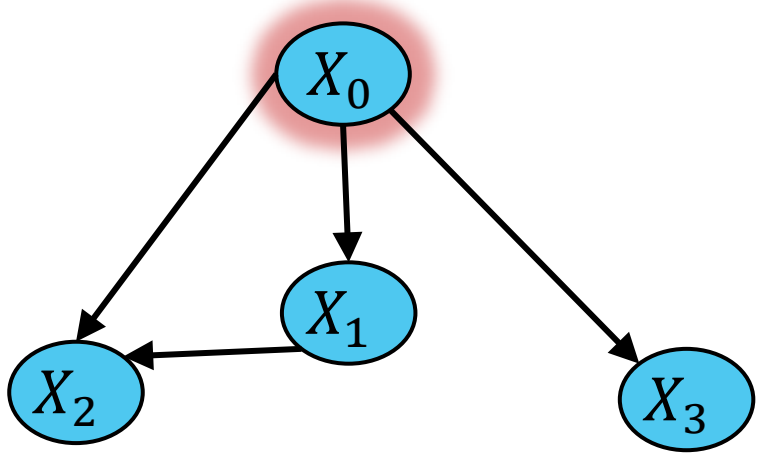
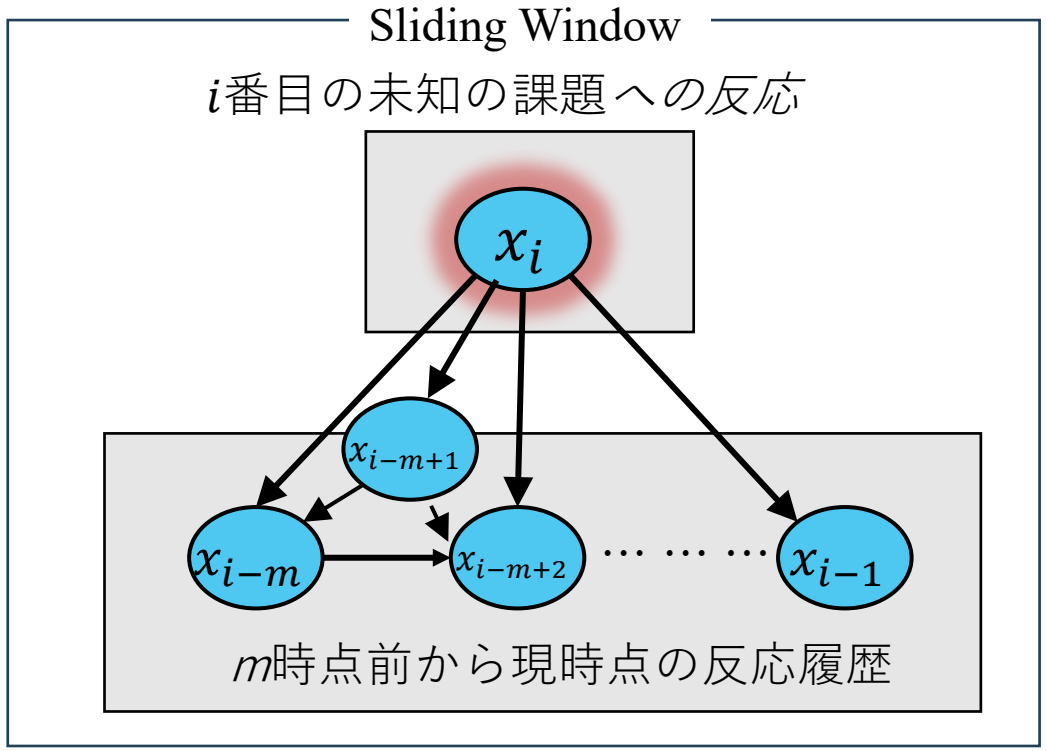


図 変数順序(X_0, X_1, X_2, X_3)に従う構造の例

4.2 変数順序


Sliding Window
i番目の未知の課題への反応
x_i
x_{i-m+1}
x_{i-m}
x_{i-m+2}
x_{i-1}
m時点前から現時点の反応履歴

定理 4.3

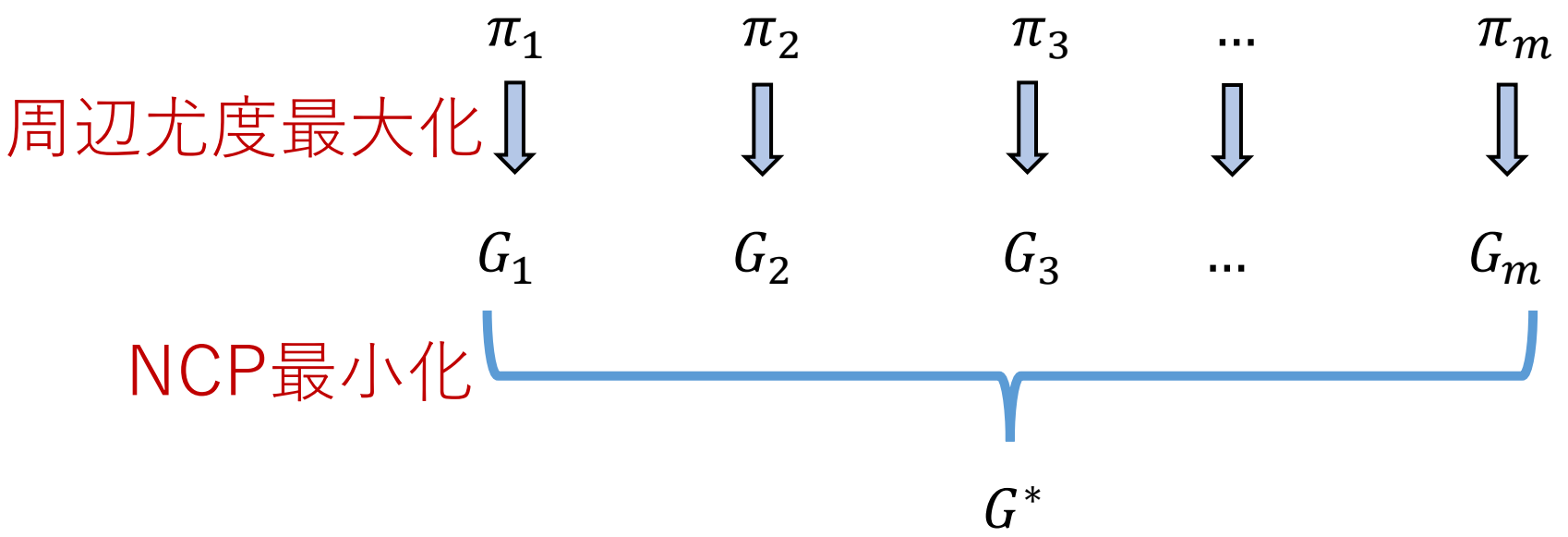
変数順序を所与として周辺尤度を最大にする構造は，その変数順序に従う真の分類確率に漸近収束する構造の中で目的変数パラメータ数が最小の構造に一致する．

4.4　学習手順

以下の2つのステップから構成される

1. 目的変数から始まる全ての変数順序について，周辺尤度を最大化する構造をそれぞれ求める
2. 第1ステップで得られた構造のうち目的変数パラメータ数
   最小の構造を探索する

4.5　学習手順の図



π_i: 目的変数を
先頭に持つ変数順序

G^*：目的変数パラメータ数を最小にして
真の分類確率に漸近収束する構造

4.6　最短パス探索問題への定式化

目的変数パラメータ数をコストとした重み付きグラフ(NPCリバースオーダグラフ)の最短パス探索問題として定式化.

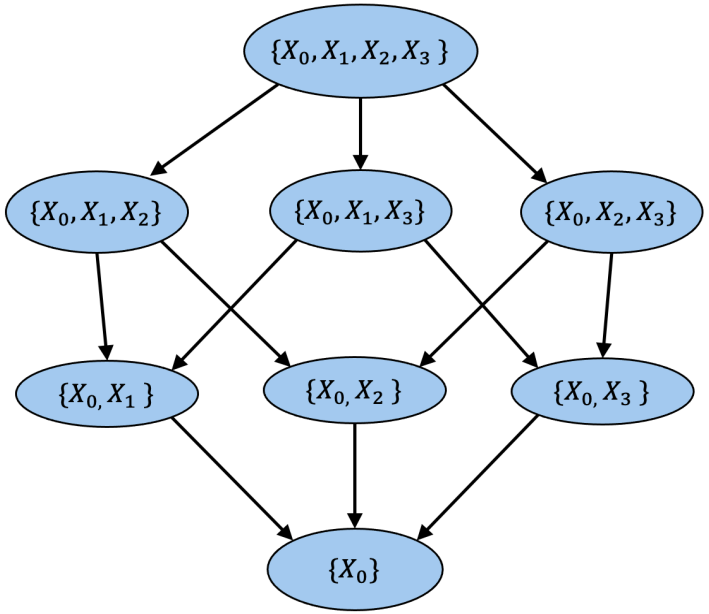
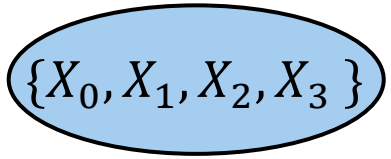
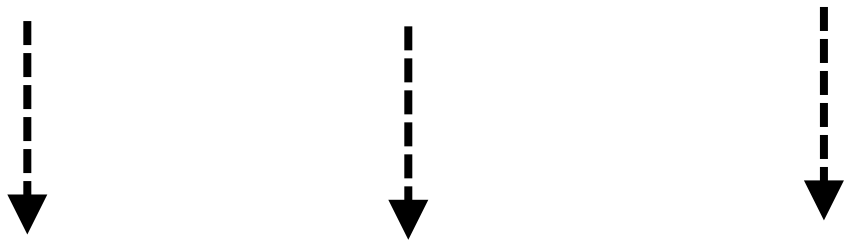


図　　四変数に対するNPCリバースオーダグラフ

4.7　幅優先探索

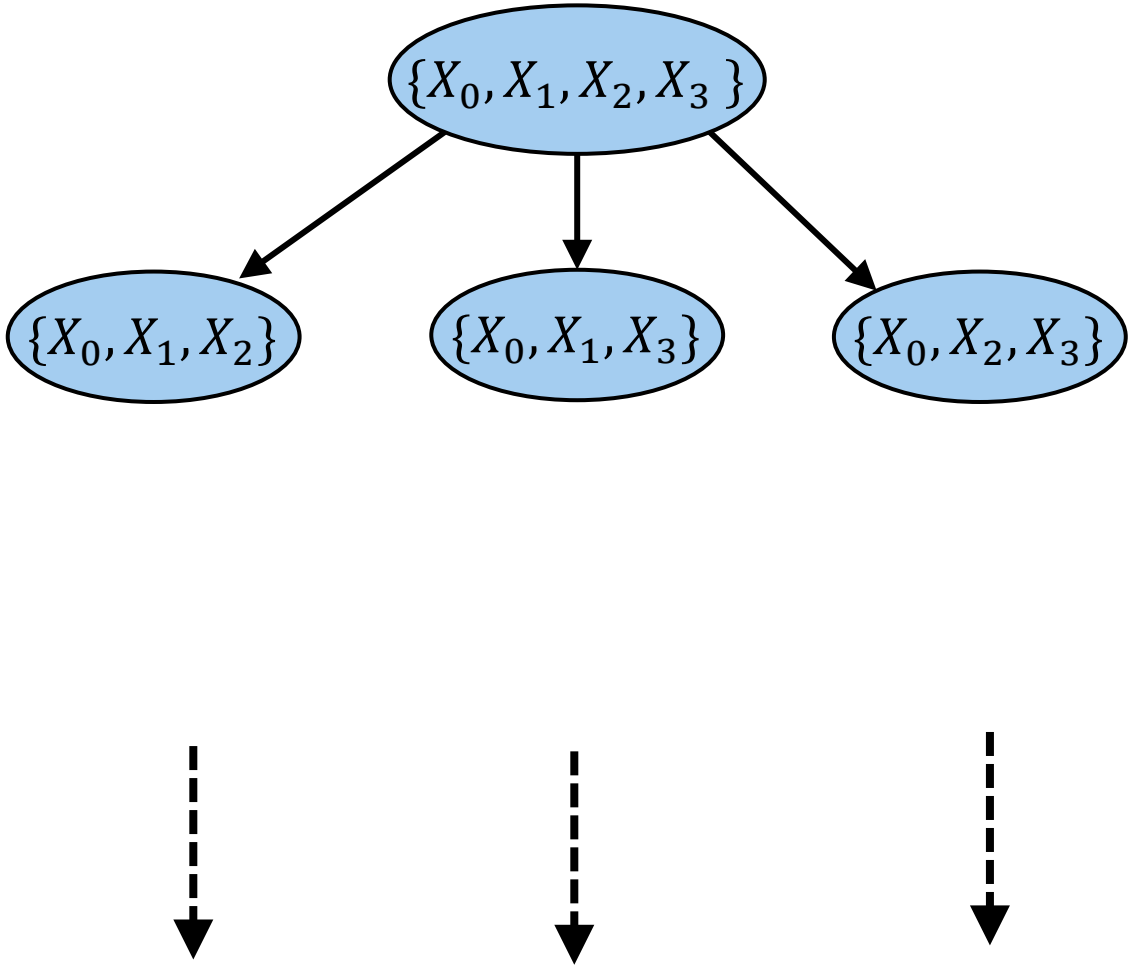
NPCリバースオーダグラフを幅優先探索する





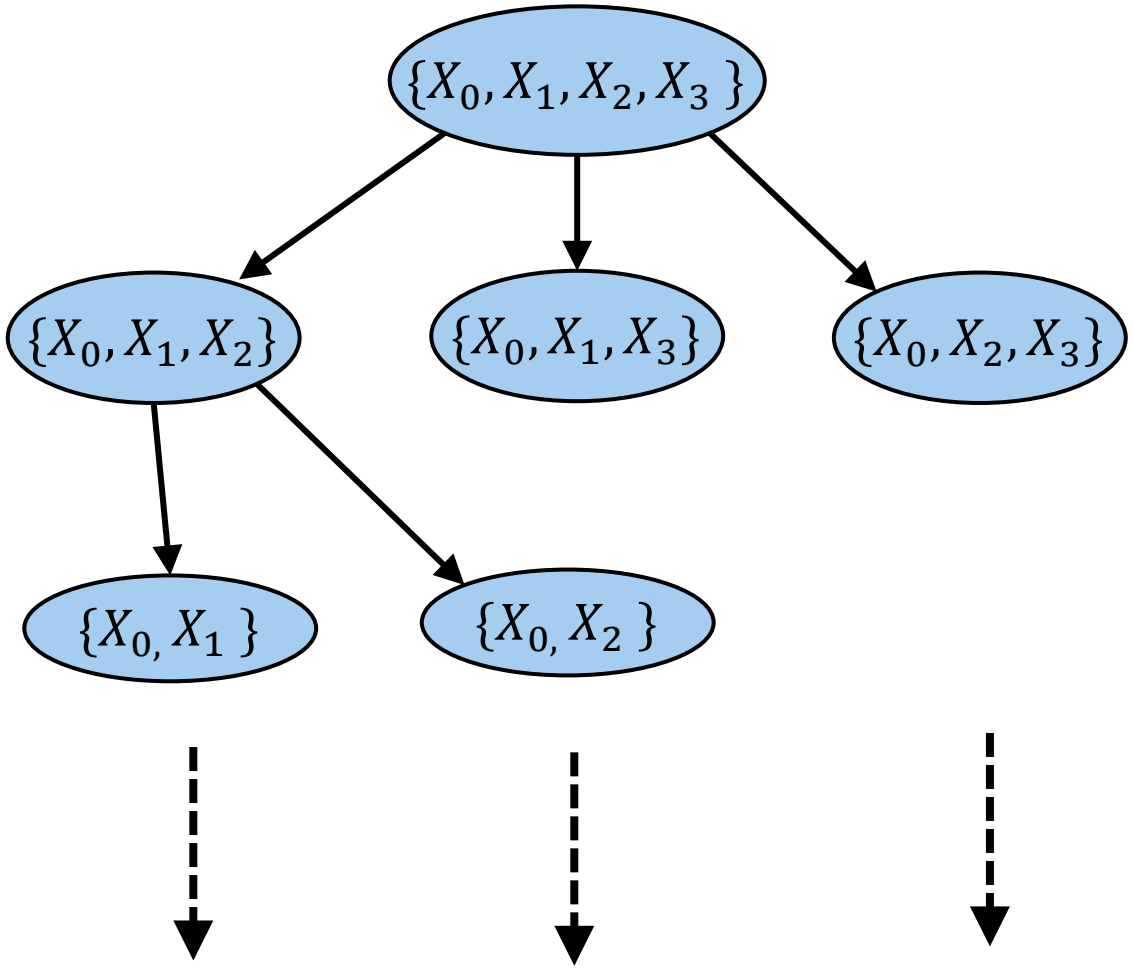
4.7　幅優先探索

NPCリバースオーダグラフを幅優先探索する



4.7　幅優先探索

NPCリバースオーダグラフを幅優先探索する



4.8　幅優先探索の問題

- 変数数の増加に伴い指数的に計算時間が増加する
  20変数程度の構造学習が限界

- 探索の終了まで構造を得ることができない

4.9　深さ優先分枝限定法

幅優先探索の問題を解決する効率的な学習法：深さ優先分枝限定法

定理5.7

本手法はNCPを最小にして真の分類

確率に漸近収束する構造を学習する

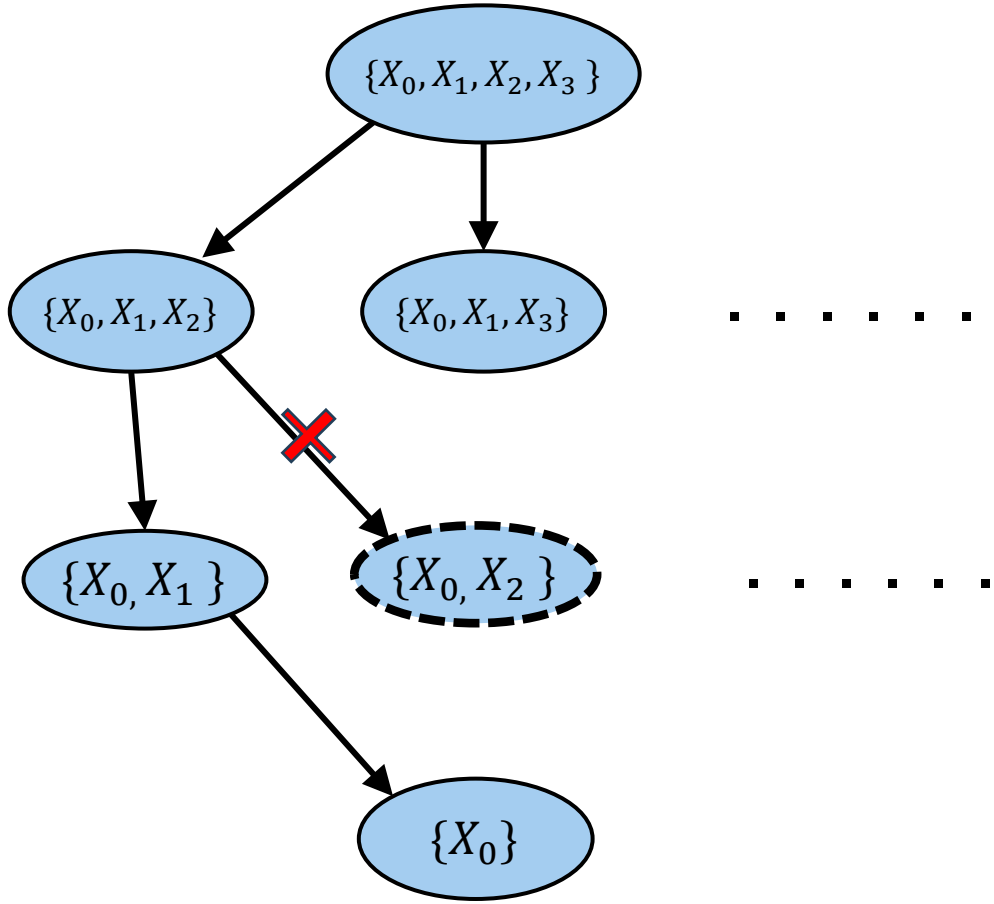
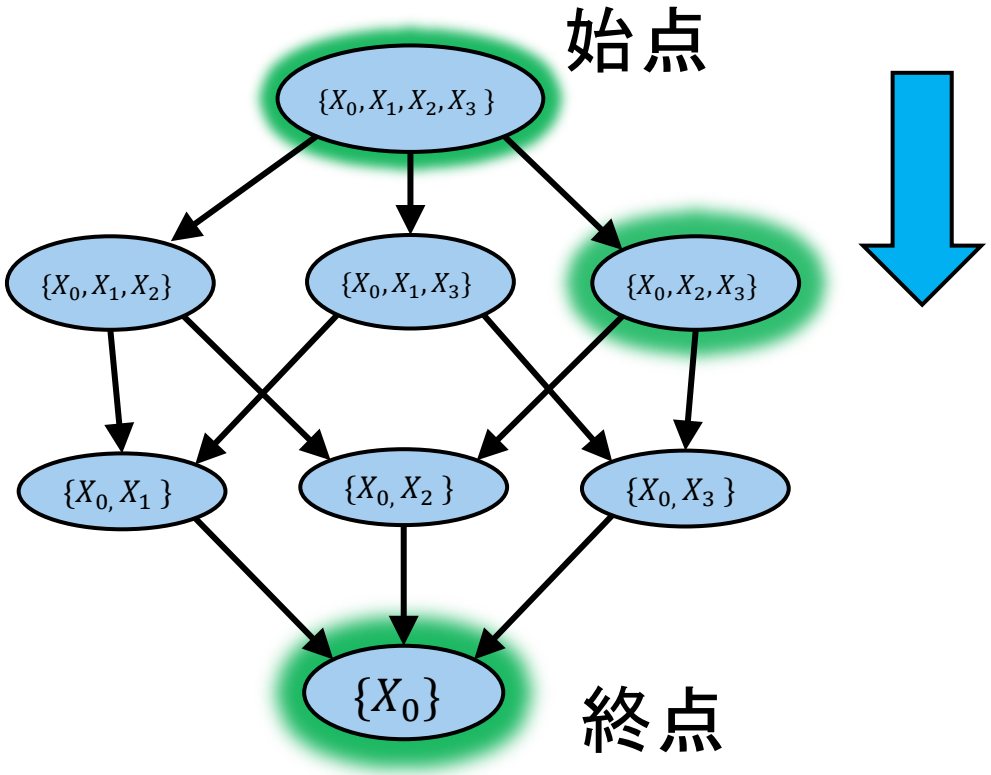


図　四変数に対するNPCリバースオーダグラフ

4.10　NPCリバースオーダーグラフと枝刈り

深さ優先分枝限定法の計算時間は枝刈りによって向上する

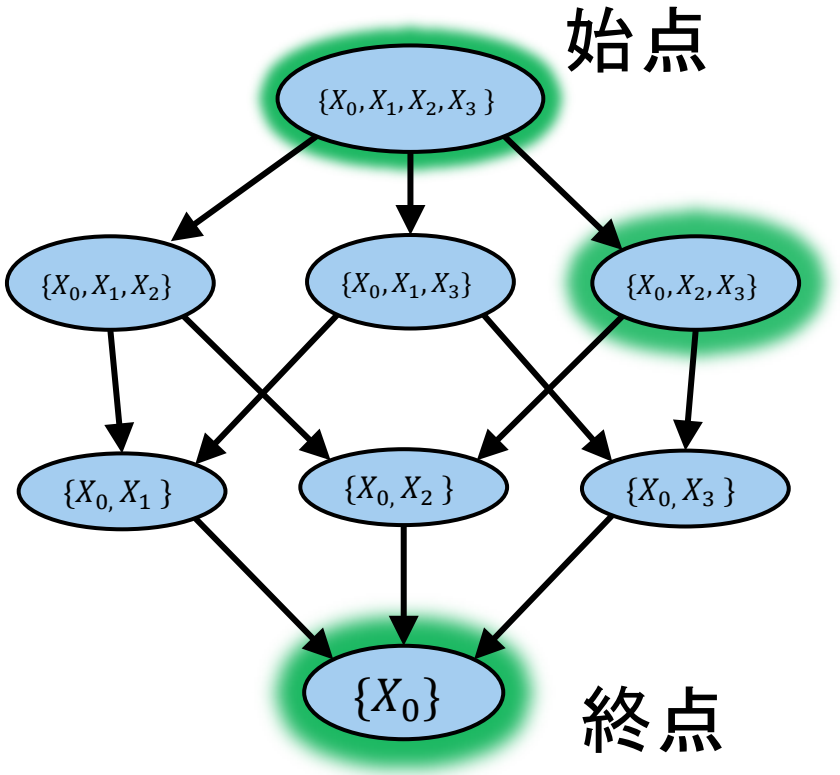


始点からあるノードUまでの
最短パスのコスト

図 NPCリバースオーダグラフ

4.10　NPCリバースオーダーグラフと枝刈り

深さ優先分枝限定法の計算時間は枝刈りによって向上する



始点からあるノードUまでの
最短パスのコスト

＋

あるノードUから終点までの
コストの下限値

図 NPCリバースオーダグラフ

4.10　NPCリバースオーダーグラフと枝刈り

深さ優先分枝限定法の計算時間は枝刈りによって向上する

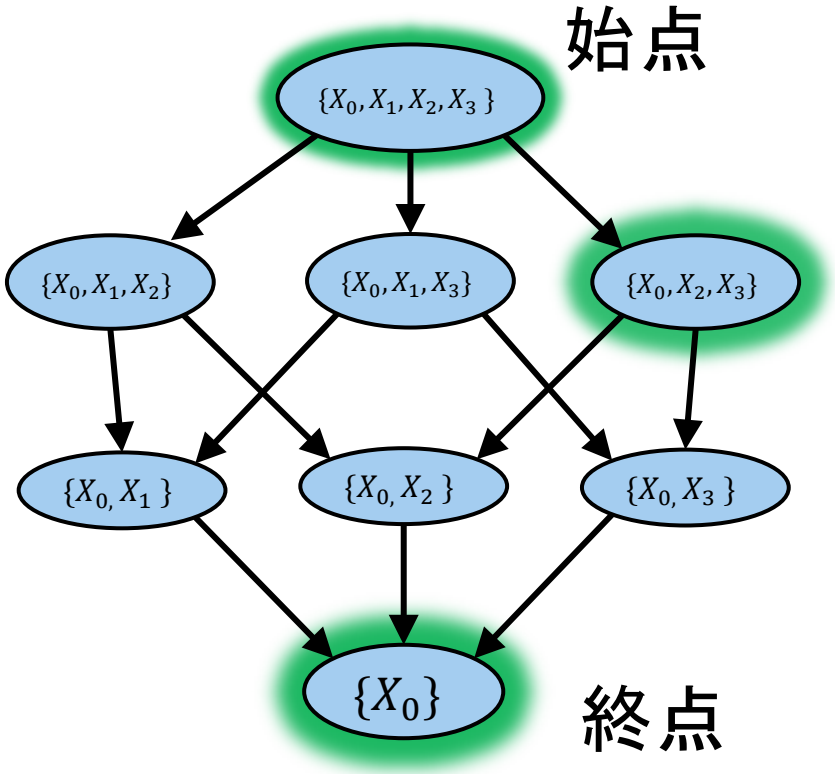


図 NPCリバースオーダグラフ

始点からあるノードUまでの
最短パスのコスト

＋

あるノードUから終点までの
コストの下限値

≧ これまでに得られている
最短パスのコスト

4.10　NPCリバースオーダーグラフと枝刈り

深さ優先分枝限定法の計算時間は枝刈りによって向上する

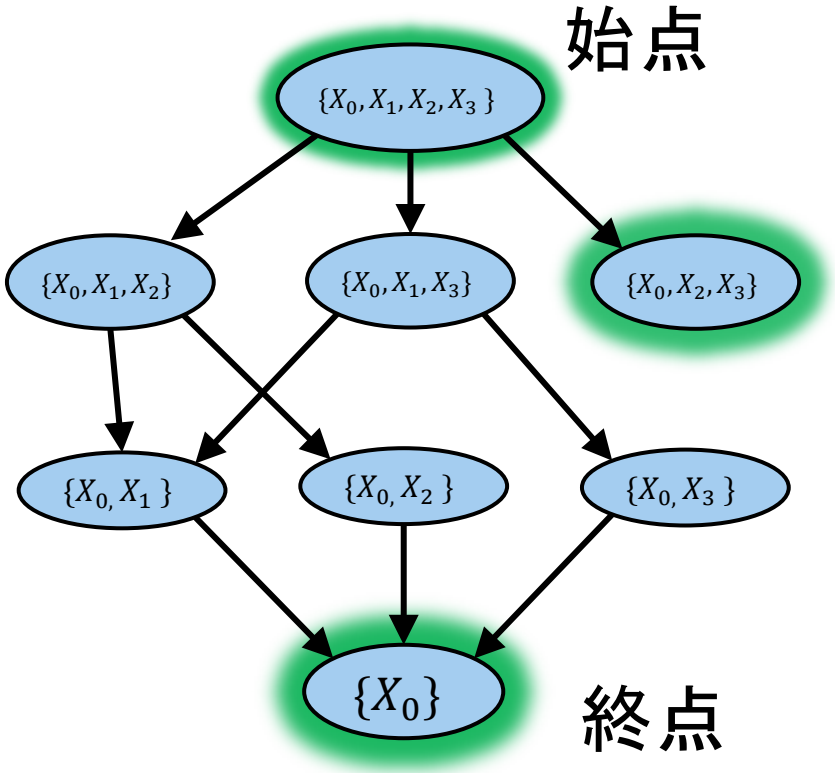


図 NPCリバースオーダグラフ

始点からあるノードUまでの
最短パスのコスト

＋

あるノードUから終点までの
コストの下限値

≧ これまでに得られている
最短パスのコスト

4.10　NPCリバースオーダーグラフと枝刈り

深さ優先分枝限定法の計算時間は枝刈りによって向上する

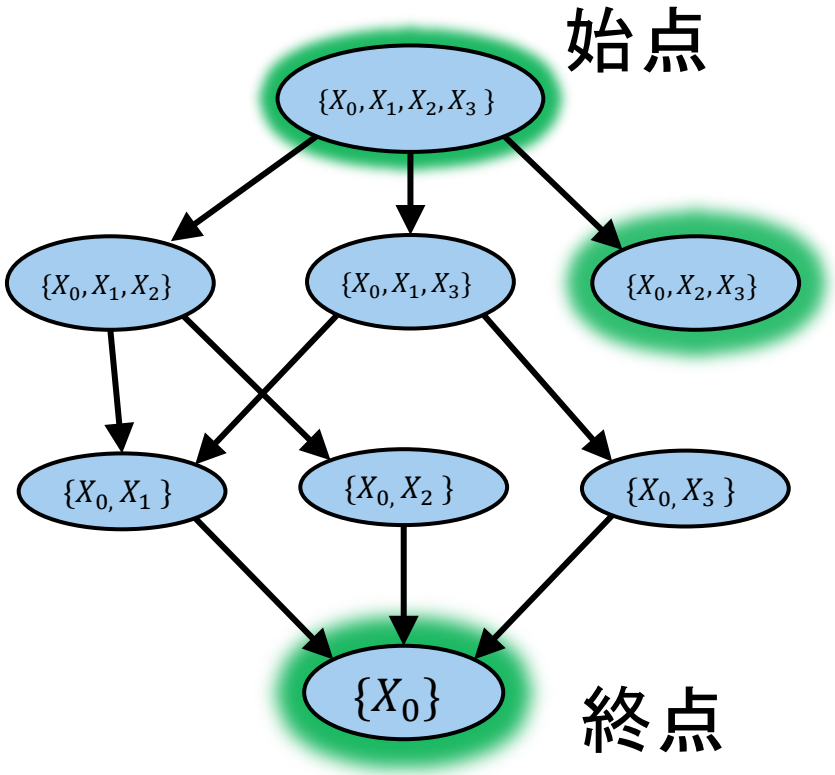


図 NPCリバースオーダグラフ

始点からあるノードUまでの
最短パスのコスト

＋

あるノードUから終点までの
コストの下限値

≧ これまでに得られている
最短パスのコスト

4.11　コストの下限値

定理4.11

Naive Bayesの目的変数パラメータ数
　　≤ 目的変数パラメータ数を最小にする構造の目的変数パラメータ数

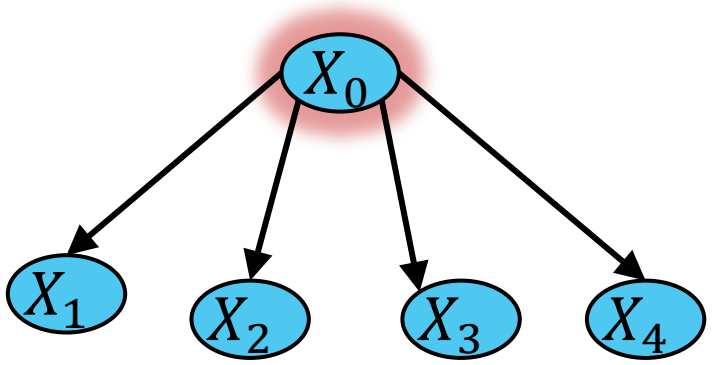


図 Naive Bayes

4.12　深さ優先分枝限定法の利点

1. 枝刈りにより幅優先探索に比べて計算時間を削減できる

2. 実行途中にメモリ等のリソースが不足してもそれまでの最適な構造を得ることが可能になる

4.13　評価実験

- 比較手法:
  1. Naive Bayes
  2. ANB厳密学習（変数選択適用）
  3. 幅優先探索によりNCP最小の構造を学習
  4. 深さ優先分枝限定法: Naive Bayesによる下限値を用いた深さ優先分枝限定法

  パラメータはすべて期待事後推定量(expected a posterior: EAP)で推定した.
- 実データ:
  - UCIレポジトリデータベースに登録されているベンチマークデータセット
- 実験手順:
  - 各手法，各データセットに対して，10分割交差検証によるテストデータの平均一致率を求め，分類精度とし，計算時間を測定した

4.14　分類精度

データセット	変数数	データ数	NaiveBayes	ANB	幅優先探索	深さ優先 分枝限定法
Image Seg	19	2310	0.9286	0.9468	0.9550	0.9558
Pendigits	17	10992	0.8805	0.9636	0.9601	0.9609
Letter	17	20000	0.7384	0.8454	0.8608	0.8616
Lympho	19	148	0.8446	0.7770	0.8041	0.7905
EEG	15	14980	0.6874	0.7644	0.7304	0.7285
WCW	10	683	0.9751	0.9751	0.9751	0.9751
Zoo	17	101	0.9406	0.9505	0.9505	0.9307
Hepatitis	20	80	0.8500	0.5750	0.7875	0.8000
Wine	14	178	0.9831	0.9663	0.9775	0.9775
Australian	15	690	0.8464	0.8420	0.8551	0.8507
Vehicle	19	846	0.4350	0.6253	0.6019	0.5827
BreastCancer	10	277	0.7401	0.7040	0.7401	0.7401
Heart	14	270	0.8444	0.8407	0.8074	0.8074
HTRU2	9	17898	0.9689	0.9779	0.9783	0.9784
CVR	17	232	0.9095	0.9483	0.9655	0.9698
Solar Flare	11	1389	0.7811	0.8229	0.8431	0.8431
Glass	10	214	0.5561	0.6449	0.6262	0.6075
CMC	10	1473	0.4671	0.4481	0.4623	0.4467
Hayes-Roth	5	132	0.8182	0.7879	0.8333	0.8333
Balance Scale	5	625	0.9152	0.9152	0.9152	0.9152
Lenses	5	24	0.7500	0.7500	0.8750	0.8750
Iris	5	150	0.9400	0.9400	0.9467	0.9467
LED7	8	3200	0.7294	0.7294	0.7316	0.7325
Banknote	5	1372	0.9249	0.9410	0.9410	0.9410
平均			0.8106	0.8201	0.8385	0.8354

4.15　計算時間と枝刈り回数

No.	Dataset	Variables	幅優先探索	深さ優先 分枝限定法	枝刈り回数
1	Lenses	5	0.0131	**0.0100**	3.9
2	Hayes-Roth	5	0.0170	**0.0149**	3.3
3	Iris	5	0.0181	**0.0134**	3.2
4	Balance Scale	5	**0.0134**	0.0170	4.0
5	Banknote	5	**0.0188**	0.0268	4.0
6	LED7	8	**0.1214**	0.1371	27.6
7	HTRU2	9	**0.2785**	0.3636	58.7
8	BC	10	0.3236	**0.1719**	176.1
9	BCW	10	0.3175	**0.1221**	98.0
10	Solar Flare	11	0.8851	**0.4564**	326.2
11	Wine	14	16.0481	**4.6310**	5945.4
12	Heart	14	9.9224	**3.8845**	6275.3
13	Australian	15	18.3376	**8.1574**	22030.4
14	EEG	15	166.2700	**24.3754**	42485.8
15	Zoo	17	459.3530	**21.2139**	29600.5
16	Congressional	17	427.0858	**34.8079**	42485.8
17	Pendigits	17	744.8170	**145.3891**	29600.5
18	Letter	17	530.7353	**99.6420**	22001.9
19	Lymphography	19	555.6909	**97.3051**	189638.4
20	Image Segmentation	19	5588.0012	**261.5876**	154339.2
21	Hepatitis	20	10044.8238	**250.7541**	386621.3
	average		883.9567	**45.3848**	44368.1

計算時間が短い方を赤字

4.16　大規模データセットを用いた評価実験

- 比較手法
  小規模での実験と同様の手法
- 実データ
  小規模での実験より大規模な31～116変数のベンチマークデータセット
- 実験手順
  - 各手法，各データセットに対して，10分割交差検証によるテストデータの平均一致率を求め，分類精度とし，計算時間を測定した
  - 構造学習は，6時間の制限時間を設け，超過する場合は打ち切った

4.17　分類精度

No.	Dataset	Variables	Naïve Bayes	ANB	幅優先探索	深さ優先 分枝限定法
1	wdbc	31	0.9139	TO	TO	0.9350
2	turkiye	33	0.3442	TO	TO	0.4897
3	ionosphere	35	0.7550	TO	TO	0.8832
4	kr-vs-kp	37	0.6640	TO	TO	0.9252
5	Flowmeters_D	44	0.8333	TO	TO	0.8833
6	Parkinson	48	0.7625	TO	TO	0.7708
7	PAMAP2	53	0.6864	TO	TO	0.8634
8	spam	58	0.8794	TO	TO	0.9331
9	molecular	61	0.9433	TO	TO	0.9464
10	Nuclear	75	0.9303	TO	TO	0.9914
11	MI	116	0.9154	TO	TO	0.9375
	average		0.7843	-	-	0.8690

TO: 6時間以内に構造を得ることができなかったことを表す

4.18　さらなる精度の向上

- 深さ優先分枝限定法で学習した構造についてCLLを最大化するようパラメータを推定することで，さらなる精度の向上が期待できる.
- しかし，CLLを最大化するパラメータは解析的に解けないため，勾配法で数値的に求める必要があるが，ほとんどの場合局所解に陥ってしまう問題がある.
- 構造がコーダルグラフである場合，CLLが単峰性を有するため，パラメータの初期値によらず勾配法で大域的最適解が得られることが知られている(Roos et.al，2005).
- 学習構造に対しコーダルグラフとなるようエッジを追加し，CLL最大推定量でパラメータ推定することで，局所解の問題を解決できる.

T. Roos, H. Wetting, P. Grunwald, P. Myllymaki, H. Tirri, “On Discriminative Bayesian Network Classifiers and Logistic Regression,” Machine Learning, vol.59, pp.267–296, 2005.

4.19　CLL最大化推定のアルゴリズム

1. 分類に影響する目的変数パラメータ数(NCP)を最小にして真の分類確率に漸近収束する構造を学習する.
2. 得られた構造をモラル化し，その後コーダル化する.
3. ２で得られたコーダルグラフGをマルコフネットワークの構造として扱い，マルコフネットワークのCLLを最大化するようパラメータを推定する.　訓練データを$D = \cup_{d=1}^{N}\left(x_0^d, x_1^d, \ldots, x_n^d\right)$とすると，パラメータ$\phi_C(\mathbf{x})$（$X_0 \in C$）についてのCLLの偏導関数は以下である.

$$\frac{\partial}{\partial \phi_C(\mathbf{x})} CLL(\boldsymbol{\Theta}) = N_{\mathbf{x}} - \sum_{d=1}^{N} P\left(\mathbf{x}^{X_0} \mid x_1^d, \ldots, x_n^d\right),$$

ただし，$N_{\mathbf{x}}$は訓練データ中の$\mathbf{x}$の頻度であり，$\mathbf{x}^{X_0}$は$\mathbf{x}$におけるX_0の値である.
この等式を用いた勾配法によってパラメータを推定する.

マルコフネットワークの構造がコーダルグラフの場合，CLLは単峰性を有する.
ステップ2を行うことで，勾配法によってCLL最大化推定量の大域解が得られる.

4.20　分類精度評価実験

- 比較手法:
  1. 深さ優先分枝限定法(EAP):深さ優先分枝限定法で学習した構造のパラメータをEAPで推定した分類器
  2. RF: Random forest
  3. DL: Deep learning
  4. 深さ優先分枝限定法(CLL): 深さ優先分枝限定法で学習した構造についてコーダルグラフとなるようエッジを追加した後，CLLを最大化するようパラメータを推定した分類器
- 実データ:
  - UCIレポジトリデータベースに登録されているベンチマークデータセット
- 実験手順:
  - 各手法，各データセットに対して，10分割交差検証によるテストデータの平均一致率を求め，分類精度とし，計算時間を測定した

4.21　分類精度

データセット	変数数	サンプルサイズ	深さ優先分枝限定法(EAP)	RF	Deep	深さ優先分枝限定法(CLL)
Hayes-Roth	5	132	0.8333	0.8088	0.7725	0.8032
Balance Scale	5	625	0.9152	0.8287	0.9840	0.9856
Banknote authentication	5	1372	0.9410	0.9432	0.9403	0.9428
Hepatitis	20	80	0.8000	0.8625	0.8750	0.8375
Zoo	17	101	0.9307	0.9500	0.9300	0.9604
Pendigits	17	10992	0.9609	0.9914	0.9899	0.9743
平均			0.8968	0.8974	0.9153	0.9173

まとめ

- 周辺尤度による厳密学習手法は，CLLによる近似学習手法と比べて，必ずしも劣るわけではない.
- サンプルサイズが小さいときにBNCでは目的変数の親変数数が増え，子変数が減ると分類精度が劣化するので，目的変数が全説明変数を子変数として持つ制約をおくANB構造の周辺尤度スコアを用いた厳密学習法がCLLを含む従来手法の分類精度を有意に改善することができる.
- RAIANBアルゴリズムは1000変数を超えるデータを学習でき，従来のRAIアルゴリズムよりも分類精度が高い.
- 深さ優先分枝限定法は真のモデルがBNに従っていない場合でも，NCP最小のI-mapを学習できる.
- 深さ優先分枝限定法の学習構造のパラメータをCLL最大化推定量で推定した場合，ランダムフォレストやディープラーニングと同等以上の分類精度を示す.

整数計画法による
ベイジアンネットワーク分類器の学習

○植野 真臣、稲村 健太郎、加藤 弘也、菅原 聖太
電気通信大学

1. 本研究

離散変数の確率的分類機として最高精度を持ち、解釈性も持つベイジアンネットワーク分類器（AAAI 2024)が注目されている.

しかし、計算量が大きく３０変数程度しか学習できない.

整数計画法を用いた効率的な学習アルゴリズムを提案し、ネットワークの大規模化を目指す.

2.分類問題での深層学習・ランダムフォレストとの比較

- 深層学習・ランダムフォレスト：　予測精度　○　解釈性　×
- ベイジアンネットワーク　　：　予測精度　×　解釈性　○

分類問題での深層学習・ランダムフォレストとの違い

- 深層学習・ランダムフォレストは説明変数を所与として目的変数を予測する識別モデル
- ベイジアンネットワークは同時確率分布を予測する生成モデル

データ数が十分に大きい場合、ベイジアンネットワークは同時確率分布に対して漸近一致制を持つが、変数数に対して予測しなければならないパターン数が指数的に増加して　識別モデルとしての予測精度は深層学習やランダムフォレストに勝てない。

3. 1. ベイジアンネットワーク分類器 [AAAI2024]

分類に影響する目的変数パラメータ数(NCP)の最小化

$$NCP(G) = \sum_{i=0}^{n} NCP_i(\mathbf{Pa}(X_i, G))$$

$$NCP_i(\mathbf{Pa}(X_i, G)) = \begin{cases} (r_i - 1)q_i & i = 0 \vee \\ & X_0 \in \mathbf{Pa}(X_i, G) \\ 0 & otherwise \end{cases}$$

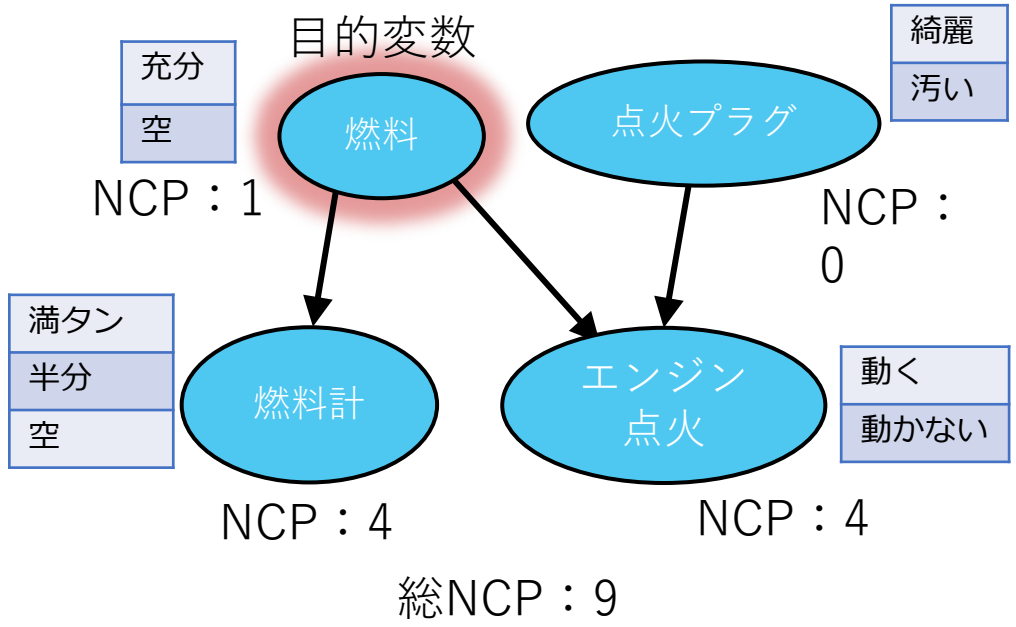


図3: NCPの計算例

G：離散確率変数$\mathbf{V} = \{X_0, X_1, \ldots, X_n\}$をノードとし，
ノード間の依存関係をエッジで表す非循環有向グラフ，
r_i：各変数X_iが取りうる状態数，
$\mathbf{Pa}(X_i, G)$：GにおけるX_iの親変数集合，q_i：$\mathbf{Pa}(X_i, G)$が取りうるパターン数

メリット：真のモデルがBNに従わない場合も真の分類確率に漸近的に一致.

Shouta Sugahara, Koya Kato and Maomi Ueno: Learning Bayesian Network Classifiers to Minimize Class Variable Parameters. In the 38th AAAI Conference on Artificial Intelligence (AAAI 2024)

3.2.Sugahara (AAAI2024) アルゴリズム

1. 目的変数X_0が親変数を持たず，各説明変数が親として持てる変数を表したすべての制約について，周辺尤度を最大化する構造を列挙
   i. 各変数$X_i \in \mathbf{V}$と親変数集合のみから成るネットワークの周辺尤度を計算.
   ii. すべての制約の中で，X_iのすべての親変数集合から，手順(i)の周辺尤度が最大になるものを決定.

2. 第一ステップで得られた構造の中でNCP最小の構造を探索
   - NCPをコストとした探索グラフの最短パス探索問題として定式化
   - 深さ優先探索で最短パスを求めている.
   - 逐次的に最適な構造を更新する深さ優先探索に枝刈りを適用

→NCPを最小にし，真の分類確率に漸近的に一致

- メリット：
  1. 枝刈りにより効率的な探索が可能
  2. 実行途中にメモリ等のリソースが不足しても，それまでの最適な構造を得られる.

3.3. Sugahara (AAAI2024) アルゴリズム問題点

- 第一ステップ手順(ii)，第二ステップの探索を別々に行っており，探索の効率が悪い．

- 第一ステップの探索において最大親変数数に制限を設けても，第二ステップの探索において必要な空間計算量が$O(2^n)$のまま
  ↓
  30変数程度でメモリオーバにより学習が打ち切られる．

4. 提案手法

整数計画法を用いたベイジアンネットワーク分類器の学習

メリット

1. 第一ステップ手順(ii)，第二ステップの探索を一つの目的関数で同時に実施→効率的な探索を実現
2. 最大親変数数をdに制限にしたとき，空間計算量を$O\left(n\sum_{j=0}^{d}\binom{n-1}{j}\right)$に削減可能.

4.1. 提案手法の目的関数

$$\text{maximize} \quad \sum_{X,\mathbf{W}} Score_X(\mathbf{W}) I(\mathbf{W} \to X) - \gamma \times \sum_{X,\mathbf{W}} NCP_X(\mathbf{W}) I(\mathbf{W} \to X)$$

周辺尤度の和

NCPの大きさ

γ：チューニングパラメータ
$Score_X(\mathbf{W})$：変数集合$\mathbf{W}$が変数Xの親であるときの周辺尤度
$I(\mathbf{W} \to X)$：BNCにおいて$\mathbf{W}$がXの親であるなら1, そうでなければ0

4.2. 提案手法の制約

目的変数が親変数を持たないBNCであるための制約

1. 各変数$X \in \mathbf{V}$の親変数集合$\mathbf{W}$がただ一つである

$$\forall X : \sum_{\mathbf{W}} I(\mathbf{W} \to X) = 1$$

2. 構造に循環を含まない

$$\forall C \subseteq \mathbf{V} : \sum_{X \in C} \sum_{\mathbf{W}:\mathbf{W} \cap C = \emptyset} I(\mathbf{W} \to X) \geq 1$$

3. 目的変数が親変数を持たない

$$I(\emptyset \to X_0) = 1$$

4.3. 全体の定式化

$$\text{maximize} \quad \sum_{X,\mathbf{W}} Score_X(\mathbf{W}) I(\mathbf{W} \to X) - \gamma \times \sum_{X,\mathbf{W}} NCP_X(\mathbf{W}) I(\mathbf{W} \to X)$$

$$\text{subject to} \quad \forall X : \sum_{\mathbf{W}} I(\mathbf{W} \to X) = 1$$

$$\forall C \subseteq \mathbf{V} : \sum_{X \in C} \sum_{\mathbf{W}: \mathbf{W} \cap C = \emptyset} I(\mathbf{W} \to X) \geq 1$$

$$I(\emptyset \to X_0) = 1$$

$$\forall X, \mathbf{W} : I(\mathbf{W} \to X) \in \{0, 1\}$$

4.4. 提案手法の時間計算量

$$O\left(2^n \Sigma_{j=0}^{d} \binom{n-1}{j}\right)$$

(例)最大親変数数を3に制限したときの時間計算量

AAAI 2024 → $O(n2^n)$

提案手法→ $O\left(2^{n^4}\right)$

時間計算量は増

4.5. 提案手法の空間計算量

$$O\left(n\sum_{j=0}^{d}\binom{n-1}{j}\right)$$

(例)最大親変数数を3に制限したときの空間計算量

AAAI2024 の手法 $\rightarrow O(2^n)$

提案手法$\rightarrow O(n^4)$

空間計算量は減

5. 1.小規模ネットワークの評価実験

- 比較手法:
  - Naive Bayes
  - gGBN：周辺尤度を用いて近似学習したGBN
  - GBN：周辺尤度を用いて厳密学習したGBN
  - ANB：ANB構造を制約とした学習手法
  - 幅優先探索によるNCP最小化(以下 AAAI2024 幅優先と表記)
  - 深さ優先分枝限定法によるNCP最小化(以下 AAAI2024 深さ優先と表記)
  - 提案手法(ハイパーパラメータは0.05、IBM CPLEX)
- 実データ:
  UCIレポジトリデータベースに登録されているベンチマークデータセット
- 実験手順:
  各手法，各データセットに対して，10分割交差検証によるテストデータの平均一致率を求めて分類精度とした.

5.1.小規模ネットワークでの実験結果

表1: 小規模の分類精度

No.	Dataset	Variables	Sample size	SPP	Naive-Bayes	gGBN	GBN	ANB	AAAI2024 幅優先	AAAI2024 深さ優先	提案手法 $\gamma = 0.05$
1	Lymphography	19	148	1.63×10^{-7}	0.8378	0.7905	0.7500	0.8108	0.7635	0.7838	0.7905
2	Breast Cancer Wisconsin	10	683	3.42×10^{-7}	0.9751	0.7094	0.9751	0.9751	0.9737	0.9737	0.9737
3	Hepatitis	20	80	7.63×10^{-5}	0.8500	0.8500	0.6125	0.5750	0.8250	0.8000	0.8250
4	Zoo	17	101	1.03×10^{-4}	0.9802	0.9505	0.9228	0.9406	0.9505	0.9208	0.9703
5	Australian	15	690	2.97×10^{-4}	0.8290	0.8420	0.8507	0.8203	0.8493	0.8449	0.8580
6	Vehicle	19	846	8.07×10^{-4}	0.4314	0.5461	0.5898	0.6217	0.6050	0.5843	0.5946
7	Breast Cancer	10	277	8.33×10^{-4}	0.7364	0.7058	0.7256	0.6968	0.7076	0.7365	0.7184
8	Image Segmentation	19	2310	1.26×10^{-3}	0.7290	0.8026	0.8255	0.8273	0.8264	0.8320	0.8264
9	Congressional Voting Records	17	232	1.77×10^{-3}	0.9095	0.9741	0.9655	0.9483	0.9698	0.9655	0.9612
10	Heart	14	270	2.44×10^{-3}	0.8259	0.8222	0.8370	0.8037	0.8222	0.8333	0.8222
11	Solar Flare	11	1389	3.72×10^{-3}	0.7804	0.8431	0.8431	0.8215	0.8431	0.8409	0.8398
12	Wine	14	178	7.24×10^{-3}	0.9270	0.9045	0.9270	0.9270	0.9494	0.9494	0.9494
13	Letter	17	20000	1.17×10^{-2}	0.4466	0.5761	0.6434	0.6434	0.6290	0.6303	0.6237
14	Pendigits	17	10992	1.68×10^{-2}	0.8032	0.9253	0.9342	0.9332	0.9368	0.9373	0.9314
15	Contraceptive Method Choice	10	1473	5.99×10^{-2}	0.4671	0.4440	0.4792	0.4481	0.4616	0.4396	0.4742
16	Glass	10	214	6.97×10^{-2}	0.5514	0.4626	0.5888	0.6355	0.5794	0.6036	0.6122
17	Hayes-Roth	5	132	2.29×10^{-1}	0.8333	0.7525	0.6212	0.7879	0.8333	0.8333	0.8333
18	Balance Scale	5	625	3.33×10^{-1}	0.9152	0.9152	0.9152	0.9152	0.9152	0.9152	0.9152
19	Lenses	5	24	3.33×10^{-1}	0.7083	0.8333	0.8333	0.7500	0.8750	0.8750	0.8750
20	EEG	15	14980	4.57×10^{-1}	0.5778	0.6732	0.7246	0.7212	0.7155	0.7135	0.7115
21	LED7	8	3200	2.50×10^{0}	0.7294	0.7297	0.7303	0.7303	0.7316	0.7325	0.7275
22	Iris	5	150	3.13×10^{0}	0.7133	0.8133	0.8267	0.8156	0.8200	0.8200	0.8133
23	HTRU2	9	17898	3.50×10^{1}	0.8966	0.9092	0.9141	0.9141	0.9140	0.9140	0.9141
24	Banknote authentication	5	1372	4.29×10^{1}	0.8433	0.8819	0.8812	0.8812	0.8819	0.8819	0.8812
	average				0.7624	0.7774	0.7882	0.7893	0.8070	0.8067	0.8101

5.1.小規模ネットワークでの実験結果

表1: 小規模の分類精度

No.	Dataset	Variables	Sample size	SPP	Naive-Bayes	gGBN	GBN	ANB	AAAI2024 幅優先	AAAI2024 深さ優先	提案手法 $\gamma = 0.05$
1	Lymphography	19	148	1.63×10^{-7}	0.8378	0.7905	0.7500	0.8108	0.7635	0.7838	0.7905
2	Breast Cancer Wisconsin	10	683	3.42×10^{-7}	0.9751	0.7094	0.9751	0.9751	0.9737	0.9737	0.9737
3	Hepatitis	20	80	7.63×10^{-5}	0.8500	0.8500	0.6125	0.5750	0.8250	0.8000	0.8250
4	Zoo	17	101	1.03×10^{-4}	0.9802	0.9505	0.9228	0.9406	0.9505	0.9208	0.9703
5	Australian	15	690	2.97×10^{-4}	0.8290	0.8420	0.8507	0.8203	0.8493	0.8449	0.8580
6	Vehicle	19	846	8.07×10^{-4}	0.4314	0.5461	0.5898	0.6217	0.6050	0.5843	0.5946
7	Breast Cancer	10	277	8.33×10^{-4}	0.7364	0.7058	0.7256	0.6968	0.7076	0.7365	0.7184
8	Image Segmentation	19	2310	1.26×10^{-3}	0.7290	0.8026	0.8255	0.8273	0.8264	0.8320	0.8264
9	Congressional Voting Records	17	232	1.77×10^{-3}	0.9095	0.9741	0.9655	0.9483	0.9698	0.9655	0.9612
10	Heart	14	270	2.44×10^{-3}	0.8259	0.8222	0.8370	0.8037	0.8222	0.8333	0.8222
11	Solar Flare	11	1389	3.72×10^{-3}	0.7804	0.8431	0.8431	0.8215	0.8431	0.8409	0.8398
12	Wine	14	178	7.24×10^{-3}	0.9270	0.9045	0.9270	0.9270	0.9494	0.9494	0.9494
13	Letter	17	20000	1.17×10^{-2}	0.4466	0.5761	0.6434	0.6434	0.6290	0.6303	0.6237
14	Pendigits	17	10992	1.68×10^{-2}	0.8032	0.9253	0.9342	0.9332	0.9368	0.9373	0.9314
15	Contraceptive Method Choice	10	1473	5.99×10^{-2}	0.4671	0.4440	0.4792	0.4481	0.4616	0.4396	0.4742
16	Glass	10	214	6.97×10^{-2}	0.5514	0.4626	0.5888	0.6355	0.5794	0.6036	0.6122
17	Hayes-Roth	5	132	2.29×10^{-1}	0.8333	0.7525	0.6212	0.7879	0.8333	0.8333	0.8333
18	Balance Scale	5	625	3.33×10^{-1}	0.9152	0.9152	0.9152	0.9152	0.9152	0.9152	0.9152
19	Lenses	5	24	3.33×10^{-1}	0.7083	0.8333	0.8333	0.7500	0.8750	0.8750	0.8750
20	EEG	15	14980	4.57×10^{-1}	0.5778	0.6732	0.7246	0.7212	0.7155	0.7135	0.7115
21	LED7	8	3200	2.50×10^{0}	0.7294	0.7297	0.7303	0.7303	0.7316	0.7325	0.7275
22	Iris	5	150	3.13×10^{0}	0.7133	0.8133	0.8267	0.8156	0.8200	0.8200	0.8133
23	HTRU2	9	17898	3.50×10^{1}	0.8966	0.9092	0.9141	0.9141	0.9140	0.9140	0.9141
24	Banknote authentication	5	1372	4.29×10^{1}	0.8433	0.8819	0.8812	0.8812	0.8819	0.8819	0.8812
	average				0.7624	0.7774	0.7882	0.7893	0.8070	0.8067	0.8101

5.2.大規模ネットワークでの評価実験

- 比較手法:
  - Naive Bayes
  - 深さ優先分枝限定法によるNCP最小化(以下 AAAI2024 深さ優先と表記)
  - 提案手法(ハイパーパラメータは0.05, IBM CPLEX)
- 実データ:
  UCIレポジトリデータベースに登録されているベンチマークデータセット
- 実験手順:
  - 各手法，各データセットに対して，10分割交差検証によるテストデータの平均一致率を求めて分類精度とした.
  - AAAI2024の手法と提案手法は構造学習中のメモリ使用量を測定した.
  - 最大親変数数を3にした.

5.2.大規模ネットワークでの実験結果

表2: 大規模の分類精度

No.	Dataset	Variables	Sample size	Naive-Bayes	AAAI2024 深さ優先	提案手法 $\gamma = 0.05$
25	Phishing	31	11055	0.9276	0.9337*	0.9433
26	Postures	31	23906	0.8290	0.8727*	0.8760
27	connect-4	43	67557	0.7058	0.6751*	0.7138
28	PAMAP2	53	174915	0.6862	0.8266*	0.8430
29	spam	58	4601	0.8794	0.9063*	0.9107
	average			0.8056	0.8429	0.8574

※ "*"は, メモリオーバによって打ち切りが発生し, それまでに得た最もNCPの小さい構造を用いたときの分類精度に付与されている.

5.2.大規模ネットワークでの実験結果

表2: 大規模の分類精度

No.	Dataset	Variables	Sample size	Naive-Bayes	AAAI2024 深さ優先	提案手法 $\gamma = 0.05$
25	Phishing	31	11055	0.9276	0.9337*	0.9433
26	Postures	31	23906	0.8290	0.8727*	0.8760
27	connect-4	43	67557	0.7058	0.6751*	0.7138
28	PAMAP2	53	174915	0.6862	0.8266*	0.8430
29	spam	58	4601	0.8794	0.9063*	0.9107
	average			0.8056	0.8429	0.8574

※ "*"は, メモリオーバによって打ち切りが発生し, それまでに得た最もNCPの小さい構造を用いたときの分類精度に付与されている.

提案手法はAAAI2024手法と比較して分類精度が向上した.

5.3.大規模ネットワークでのメモリ使用量(58変数)

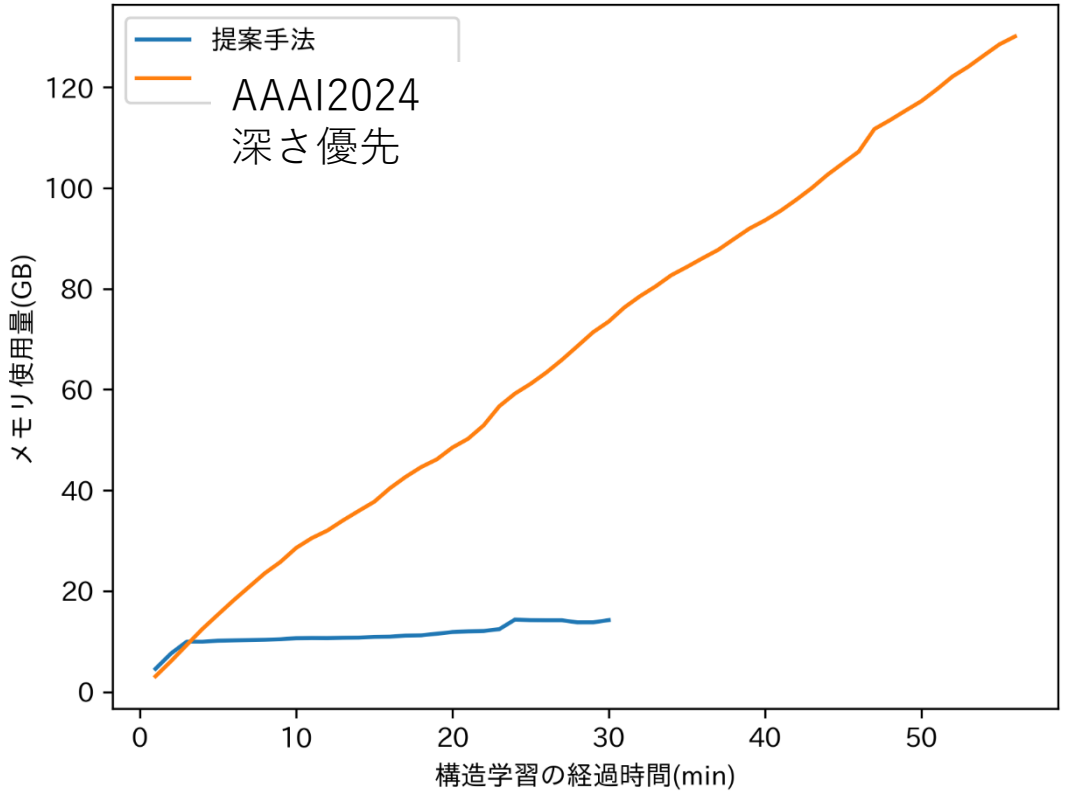
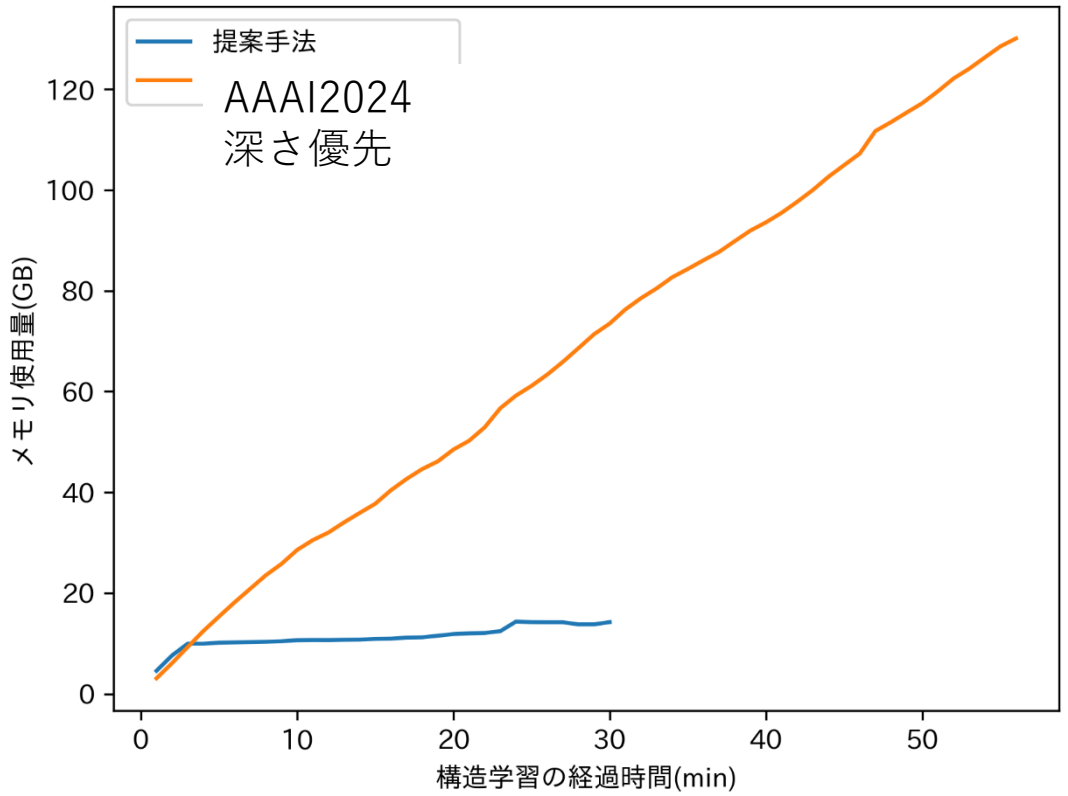


図4: 5番の構造学習中のメモリ使用量

5.3.大規模ネットワークでのメモリ使用量(58変数)



メモリオーバ

図4: 5番の構造学習中のメモリ使用量

提案手法はAAAI2024の手法と比較して構造学習中のメモリ使用量が少ない.

6. むすび

- 提案手法:
  整数計画法を用いた新たなベイジアンネットワーク分類器学習
  1. NCPを最小にして真の分類確率に漸近的に収束する構造を学習するための目的関数を導入
  2. 目的変数が親変数を持たないBNCを学習するための制約を導入
  3. 第一ステップ手順(ii)，第二ステップの探索を一つの目的関数で同時に実施
     →効率的な探索を実現

- 結果:
  - ■時間計算量は×　であるが　空間計算量は　〇
  - ■大規模ベイジアンネットワーク分類器では空間計算量の制限が大きいので実験では，学習できるネットワークの変数数を従来の30程度から58変数までアップデートできた.

7. 今後の大規模ネットワーク学習へのアイデア

Two-step Algorithm :

１．時間計算量の小さいが　空間計算量の大きい　深さ優先探索を用いて　メモリが許す限りに最適なベイジアンネットワーク分類器を学習する.

２．深さ優先探索で途中まで学習されたベイジアンネットワーク分類器の構造を初期値に入力し、時間計算量は大きいが空間計算量の小さい整数計画法を用いてさらに最適なベイジアンネットワーク分類器を学習する.

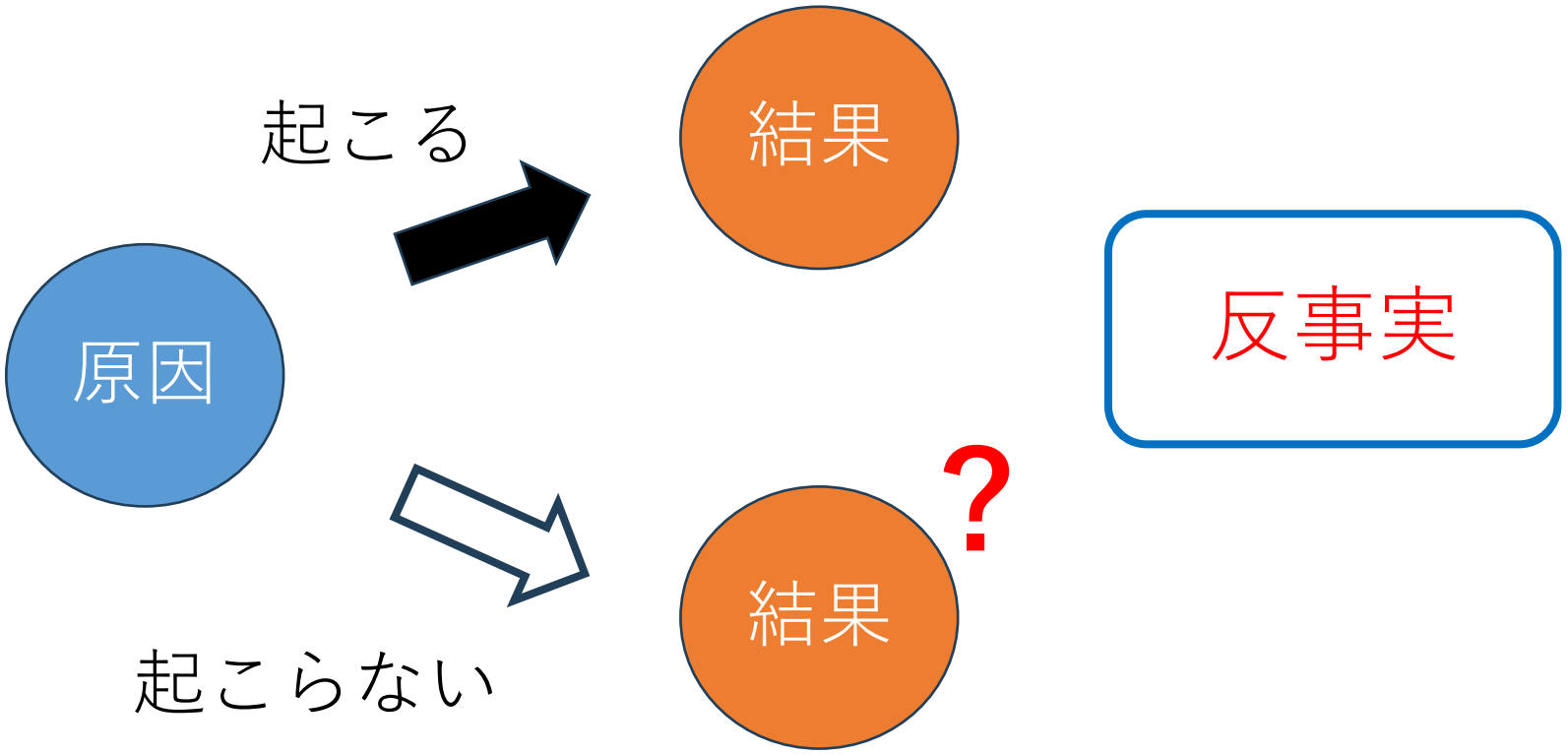
2025年5月28日

ベイジアンネットワーク分類器を用いた傾向スコア推定による因果推論

電気通信大学大学院情報理工学研究科
佐久間理奈　加藤弘也　植野真臣
2025年度人工知能学会全国大会　2S5-GS-2

因果推論

原因が結果にどのように影響を与えているのかを明らかにする



Rubin因果モデル

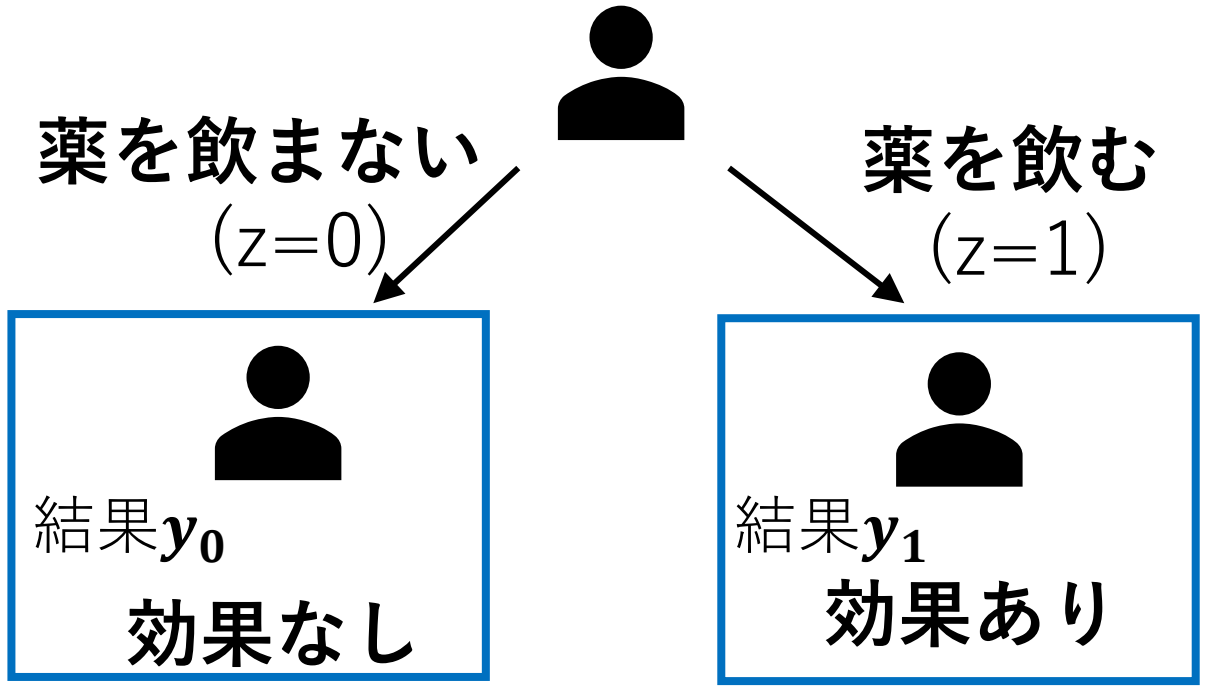
反事実の仮定をおいたRubin因果モデル[1]は, 介入の有無による結果を比較して因果推論を行う統計的手法

例) 頭痛薬の効果があるか確かめたい

介入z：頭痛薬を投与する

y_1：薬を飲んだ時の効き目

y_0：飲まなかったときの効き目



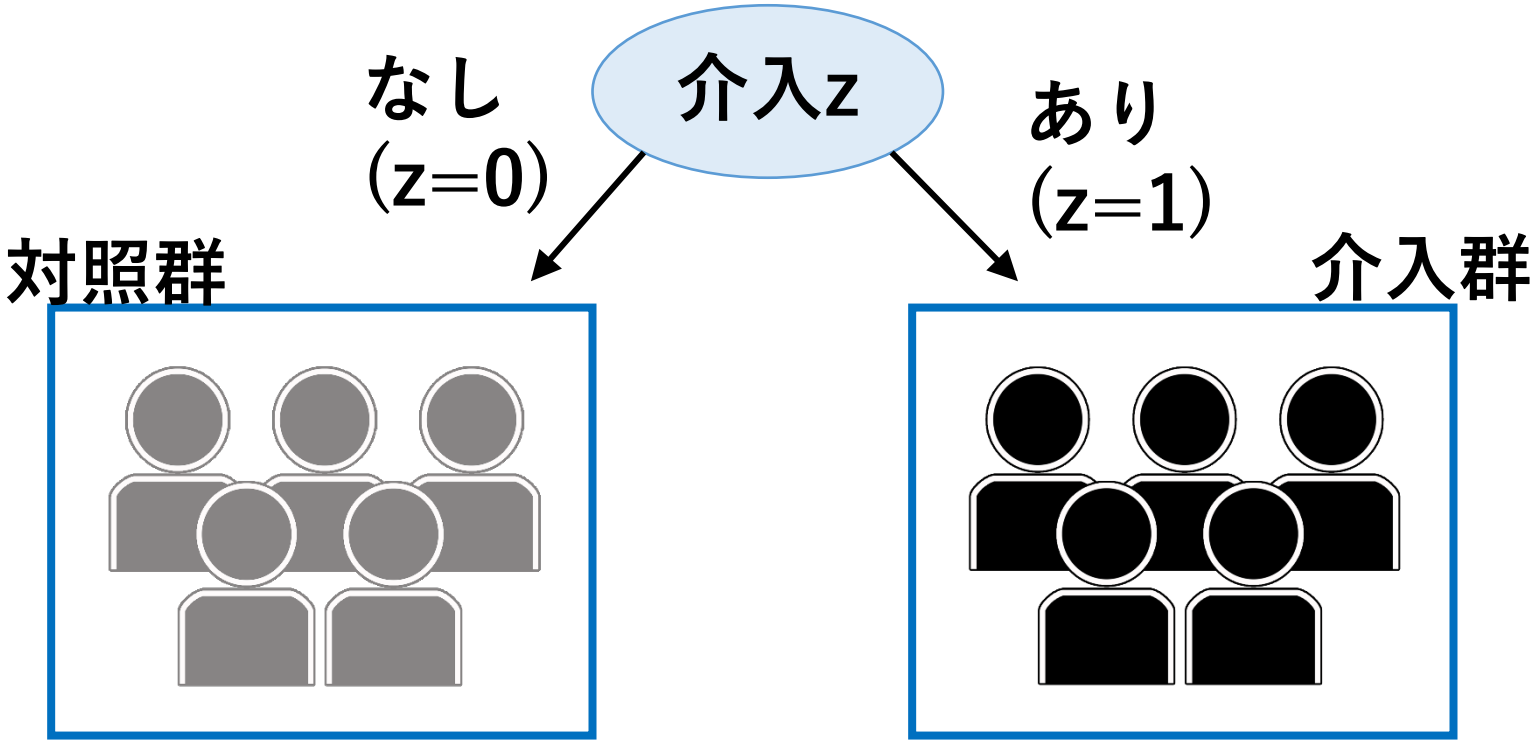
因果効果 $= y_1 - y_0$

しかし, 同一の対象では介入の有無を同時に観測できない

[1] Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688–701.

無作為化比較試験

対象をランダムに介入群と対照群に分けて介入の効果を評価する方法



無作為化比較試験では, $(y_1, y_0) \perp z$が成り立つため,

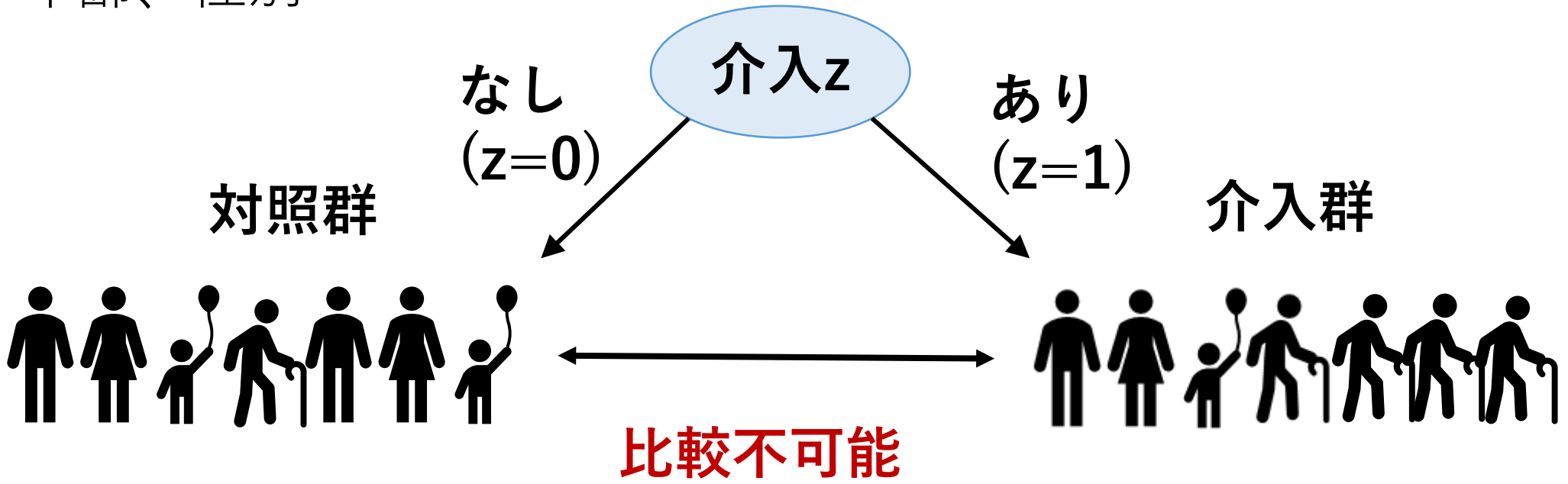
$$E(y_j) = E(y_j \mid z = j), (j = 1, 0)$$

となるので, ***因果効果または平均介入効果*(Average Treatment Effect : ATE)は**

$$\boldsymbol{ATE = E(y_1 - y_0) = E(y_1|z = 1) - E(y_0|z = 0)}$$

無作為化比較試験ができない場合

共変量x：対象の特徴
例）年齢、性別



共変量xに偏りがあり，平均介入効果を推定できない

$$ATE \neq E(y_1 \mid z = 1) - E(y_0 \mid z = 0)$$

傾向スコア

Rosenbaum & Rubin [2] は傾向スコアを用いた共変量調整法を提案した.

傾向スコア$e(x)$：共変量 x を所与としたときの対象が介入を受ける確率

$$e(x) = p(z = 1|y_1, y_0, x) = p(z = 1|x)$$

このとき $E(y_j \mid x) = E(y_j \mid e(x)), (j = 1,0)$

$$(y_1, y_0) \perp z \mid e(x), \quad 0 < p(z = 1 \mid e(x)) < 1$$

この条件のもと, $E(y_j \mid e(x)) = E(y_j \mid e(x), z = j), (j = 1,0)$

となり, $\mathrm{ATE} = E_x(E(y_1 \mid e(x), z = 1) - E(y_0 \mid e(x), z = 0))$

が成り立つため，逆確率重み付け法(Inverse probability weighting : IPW)でATEを求められる[3].

$$\mathrm{ATE} = \frac{1}{N}\sum_i^N \frac{y_{1i}}{e(x_i)} z_i - \frac{1}{N}\sum_i^N \frac{y_{0i}}{1 - e(x_i)}(1 - z_i)$$

Nは対象全体の数, y_{1i}は対象iが介入を受けた時の結果変数, y_{0i}は対象iが介入を受けなかった時の結果変数, x_iは対象iの共変量, z_iは対象iの介入の有無

[2] Rosenbaum, P. R., and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.
[3]K. Hirano, G.W. Imbens, and G. Ridder, “Efficient estimation of average treatment effects using the estimated propensity score,” Econometrica,

ロジスティック回帰を用いた傾向スコア推定

傾向スコアの推定にはロジスティック回帰[2]が最も用いられる

しかし

傾向スコアのlogitが共変量の線形関数で表現できない場合,
傾向スコアの推定に一致性がなくなる[3]

$$\mathrm{logit}\big(\hat{e}(x)\big) = \log\frac{\hat{e}(x)}{1-\hat{e}(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + x_n\beta_n$$

ここで$\hat{e}(x)$は介入を受ける確率(傾向スコア)の推定値, β_0は定数項, $\beta_1, \beta_2, \ldots, \beta_n$ は各説明変数の回帰係数, $x_1,\ x_2, \ldots, x_n$ は説明変数である.

因果効果推定の精度は傾向スコア推定の精度に大きく依存する[4,5] ため,
この手法は適さない

[3] Sant'Anna, P. H., Song, X., and Xu, Q. Covariate distribution balance via propensity scores. *Journal of Applied Econometrics*, 37(6):1093–1120, 2022.
[4] Hainmueller, J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20 (1):25–46, 2012.
[5] Imai, K. and Ratkovic, M. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,

機械学習を用いた傾向スコア推定

- 勾配ブースティング[6]
- ニューラルネットワーク[7]
- 決定木[7,8]

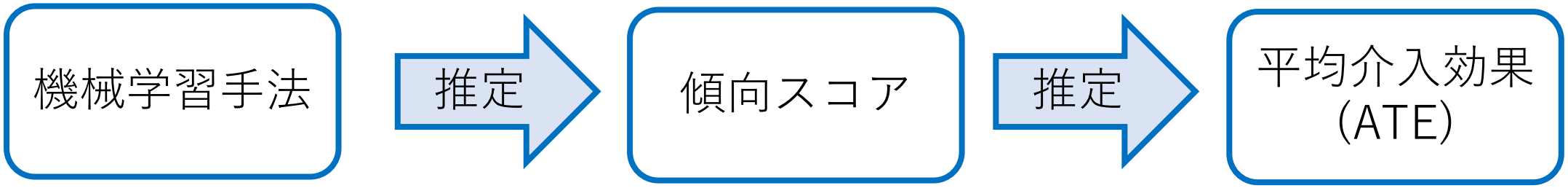
などの手法が知られているが, 漸近的に真の確率を推定する保証がない

[6] McCaffrey, D. F.; Ridgeway, G.; and Morral, A. R. 2004. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4): 403.
[7] Setoguchi, S.; Schneeweiss, S.; Brookhart, M. A.; Glynn, R. J.; and Cook, E. F. 2008. Evaluating uses of data mining technique in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6): 546– 555.
[8] Westreich, D.; Lessler, J.; and Funk, M. J. 2010. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of clinical epidemiology*, 63(8).

研究の目的

傾向スコアを利用して平均介入効果(ATE)を
高精度に推定したい

精度向上



提案手法

介入zを受けるかどうかを目的変数とし, 共変量xと結果変数yを
説明変数とした真の分類確率に漸近的に一致するベイジアン
ネットワーク分類器を利用し, 傾向スコアを推定する

ベイジアンネットワーク

ベイジアンネットワークとは, 離散確率変数をノードとし，ノード間の条件付き従属関係を非循環有向グラフ(Directed Acyclic Graph: DAG)で表し，各ノードの親ノード集合を所与とした条件付き確率で表現される確率的グラフィカルモデル

n個の変数集合 $\mathbf{V} = \{X_1, \cdots, X_n\}$を持つベイジアンネットワークは$(G, \Theta)$
で表現される.

Gは$\mathbf{V}$に対応するノード集合によって構成されるDAG

ΘはGの各エッジに対応する条件付き確率パラメータ集合$p(X_i | \mathbf{Pa}(X_i, G))\ (i = 1, \ldots n)$である.

($\mathbf{Pa}(X_i, G)$は変数X_iの親変数集合)

- DAGを仮定すると, 同時確率分布を条件付き確率の積に分解できる.

$$p(X_1, \ldots, X_N) = \prod_{i=1}^{n} p(X_i \,|\mathbf{Pa}(X_i, G))$$

ベイジアンネットワーク分類器

ベイジアンネットワークにおける一つのノードを目的変数とし，その他のノードを説明変数としたベイジアンネットワーク分類器（Bayesian Network Classifier: BNC）は，離散変数を扱う分類器として知られている[9]

分類器として用いられる，制約のない一般的なベイジアンネットワークをGBNと呼ぶ

[9] N. Friedman, D. Geiger, and M. Goldszmidt, Bayesian Network Classifiers, Machine Learning, vol. 29, pp. 131--163, 1997.

目的変数パラメータ数最小化によるベイジアンネットワーク分類器

分類に影響する目的変数パラメータ数
(the number of the class variable parameters: NCP)の最小化[10]

$$NCP(G) = \sum_{i=0}^{n} NCP_i(\mathbf{Pa}(X_i, G))$$

$$NCP_i(\mathbf{Pa}(X_i, G)) = \begin{cases} (r_i - 1)q_i & i = 0 \vee \\ & X_0 \in \mathbf{Pa}(X_i, G) \\ 0 & otherwise \end{cases}$$

G：離散確率変数$\mathbf{V} = \{X_0, X_1, \ldots, X_n\}$をノードとし，
ノード間の依存関係をエッジで表す非循環有向グラフ，
r_i：各変数X_iが取りうる状態数，
$\mathbf{Pa}(X_i, G)$：GにおけるX_iの親変数集合，q_i：$\mathbf{Pa}(X_i, G)$が取りうるパターン数

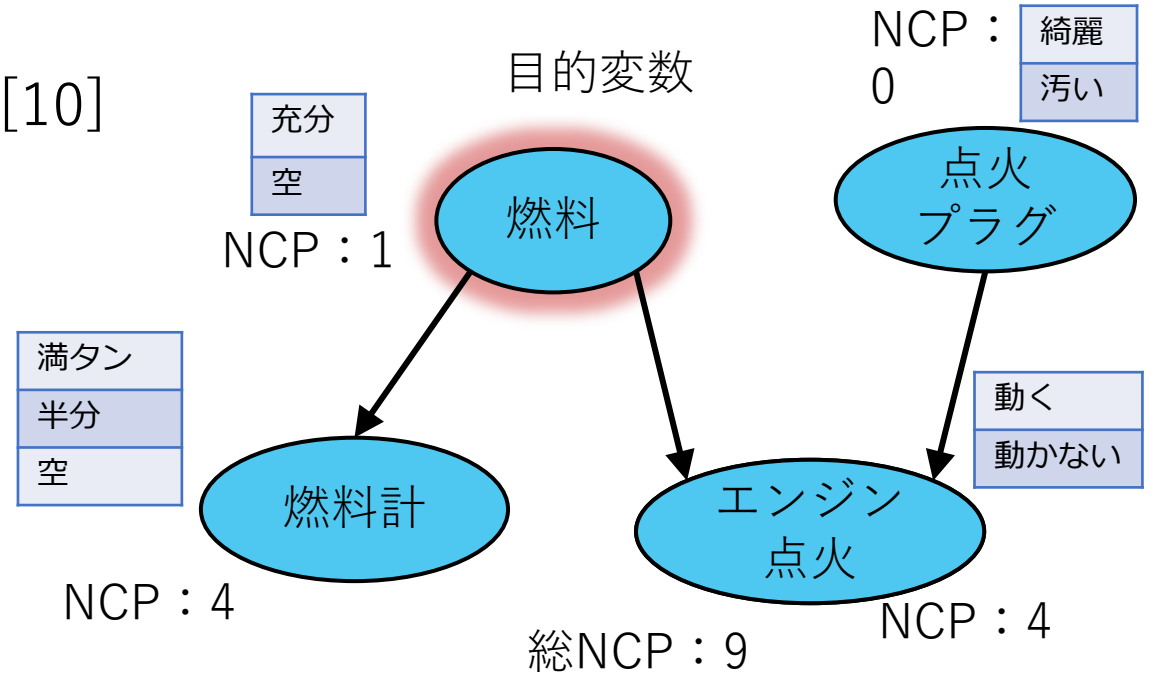


図: NCPの計算例

利点: 真のモデルがベイジアンネットワークに従わない場合でも真の分類確率に漸近的に一致する

[10] Sugahara, S., Kato, K., and Ueno, M. (2024). Learning Bayesian Network Classifiers to Minimize the Class Variable Parameters. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(18), 20540-20549.

ベイジアンネットワーク分類器を用いた傾向スコアの推定方法

学習したベイジアンネットワーク分類器の構造をG,
条件付き確率パラメータ集合をΘとすると, 以下の式で傾向スコアが推定できる

$$
\begin{aligned}
&e(x \mid G, \Theta) = p(z = 1 \mid x, G, \Theta) \\
&= p(z = 1 \mid x_1, x_2, \ldots, x_n, G, \Theta) \\
&= \frac{p(z = 1, x_1, x_2, \ldots, x_n \mid G, \Theta)}{p(x_1, x_2, \ldots, x_n \mid G, \Theta)}
\end{aligned}
$$

提案手法の利点

真のモデルがベイジアンネットワークに従わない場合
でも真の傾向スコアを漸近的に推定できる

解釈のしやすさ
因果関係を視覚的に表現することで結果を理解しやすくなる

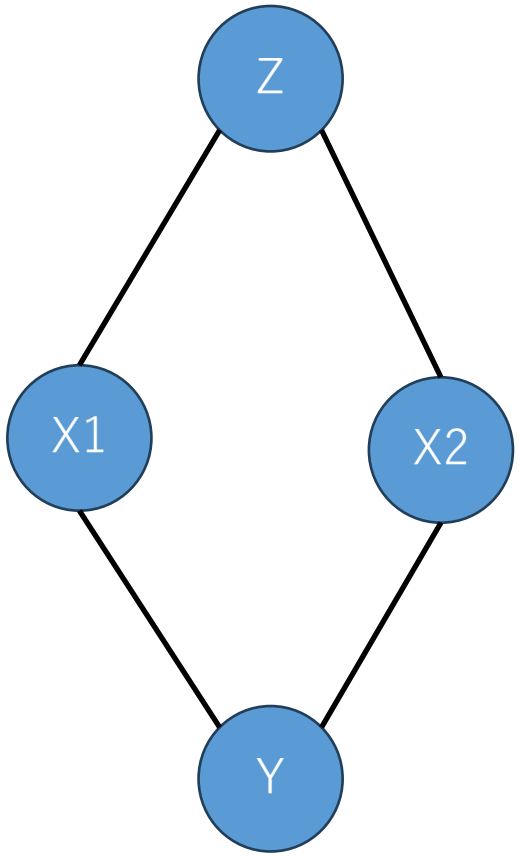
自動的な変数選択
データから構造を学び、状況に応じた最適なモデルを構築可能

実験

- マルコフネットワークのシミュレーションデータを用いた評価実験

マルコフネットワークはベイジアンネットワークでは表現できない条件付き独立性を表現

真のモデルをすべての手法と異なるものにできる



マルコフネットワークの真の構造

- 実データセットJobsを用いた評価実験

- 実データセットTwinsを用いた評価実験

比較手法

- ロジスティック回帰(LR)[1]
- Boosted CART(BOOST)[11]
  - 反復回数20 000回, 収縮パラメーター0 .0005
- ニューラルネットワーク(NN)[7]
  - 隠れ層1層10ノード[7],epochs100,batch32
- ベイジアンネットワーク分類器
  - GBNとNCPMIN(提案手法)

[11] Brian K. Lee, Justin Lessler, and Elizabeth A. Stuart. Improving propensity score weighting using machine learning. Statistic in Medicine. 2010 Feb 10; 29(3): 337–346.

シミュレーション実験手順

1. 対象の数 $n = 100, 1000, 10000, 100000$ の各場合で訓練，テスト用のシミュレーションデータを発生させる
2. 各手法を用いて傾向スコアの推定を行う訓練データで学習し、テストデータに適用する
3. 2で推定した傾向スコアを用いてATEの推定を行う
4. 1 ~ 3 を 10 回繰り返す

評価指標

- 推定したATEに対しBiasとRMSEとMAE
- 推定した傾向スコアに対しカルバックライブラー情報量

Biasの定義

$$Bias = E\left[\widehat{ATE} - ATE_{true}\right]$$

ここで $\widehat{ATE}$ はATEの推定値，ATE_{true} は真のATEとする

ATE推定の実験結果

- サンプルサイズが小さいと
  ニューラルネットワークや
  ロジスティック回帰で精度が高い

- サンプルサイズが大きいと
  提案手法(NCPMIN)の精度が高い

samplesize	LR	NN	BOOST	GBN	NCPMIN
			Bias		
100	0.3297	0.1844	3.6938	-0.4504	-0.5406
1000	0.0652	0.0352	-0.0403	0.0266	-0.0039
10000	0.0494	0.0288	0.0216	0.0385	0.0211
100000	0.0342	0.0194	MO	0.0032	0.0008
			RMSE		
100	0.5378	0.3955	11.1671	1.7477	1.7484
1000	0.1350	0.1430	0.2831	0.3318	0.3183
10000	0.0588	0.0722	0.0401	0.0669	0.0400
100000	0.0352	0.0374	MO	0.0092	0.0089
			MAE		
100	0.4285	0.3254	4.0066	0.8779	0.8805
1000	0.0963	0.1271	0.1738	0.2240	0.2112
10000	0.0494	0.0566	0.0337	0.0529	0.0336
100000	0.0342	0.0322	MO	0.0076	0.0068

傾向スコア推定の実験結果

- 傾向スコアの推定精度はBOOSTが高い
- サンプルサイズが大きくなるとBOOSTはメモリ不足により推定できず, 提案手法(NCPMIN)の推定精度が高くなる

samplesize	LR	NN	BOOST	GBN	NCPMIN
100	0.1101	0.1099	0.0561	0.1382	0.2107
1000	0.0924	0.0540	0.0064	0.0273	0.0308
10000	0.0890	0.0101	0.0006	0.0053	0.0006
100000	0.0885	0.0055	MO	0.0004	0.0001

BOOSTの傾向スコア推定精度が高いがATE推定精度が低い問題の考察

平均介入効果ATE(Average Treatment Effect)

$$\mathrm{ATE} = \frac{1}{N}\sum_i^N \frac{y_{1i}}{e(x_i)} z_i - \frac{1}{N}\sum_i^N \frac{y_{0i}}{1-e(x_i)}(1-z_i)$$

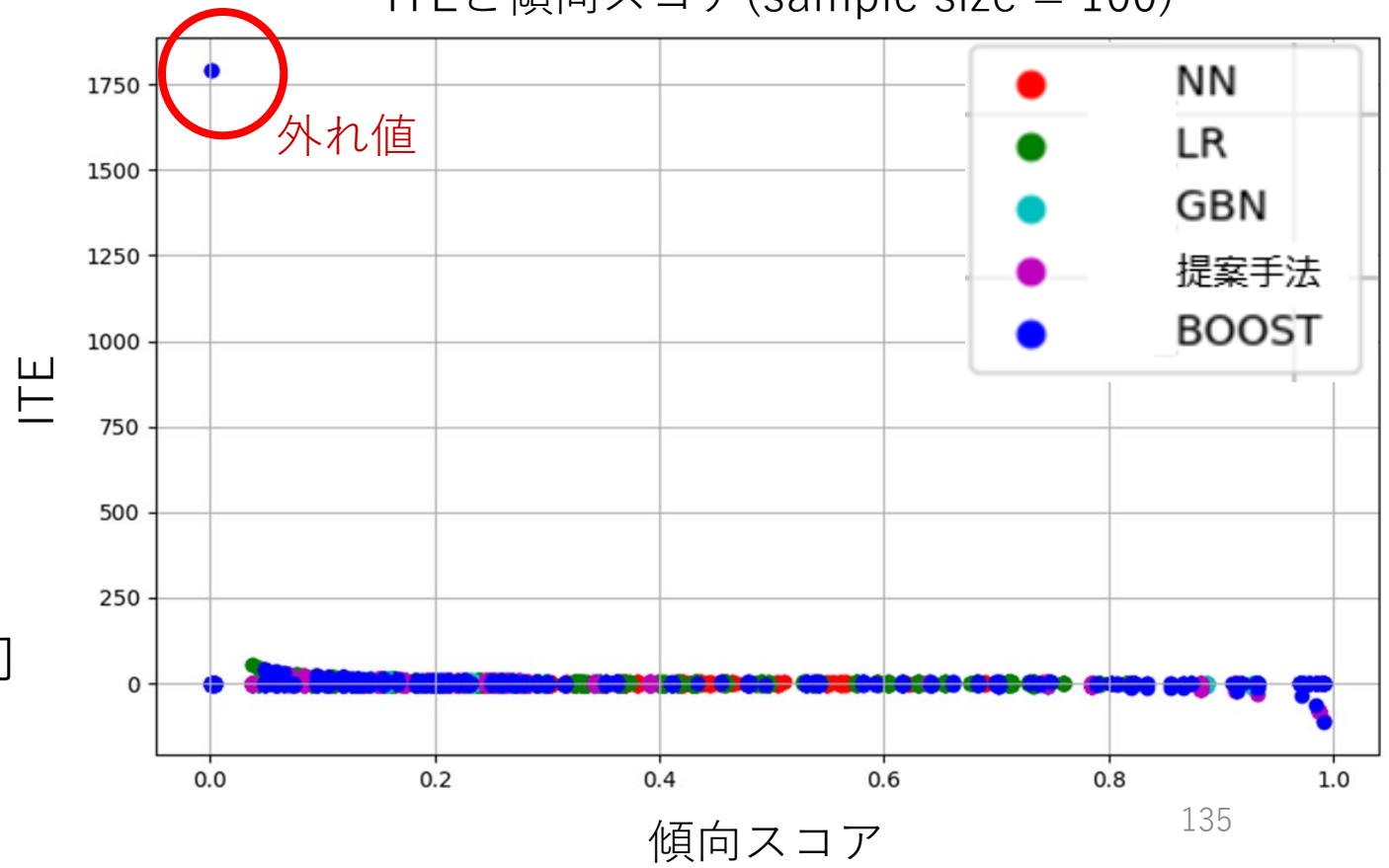
推定した傾向スコアが0や1に
近いときに絶対値が大きい値をとる

個別介入効果ITE(Individual Treatment Effect)
：各データの介入効果

$$\mathrm{ITE} = \frac{y_{1i}}{e(x_i)} z_i - \frac{y_{0i}}{1-e(x_i)}(1-z_i)$$

BOOSTは傾向スコアを0や1に
近い値に推定する場合がある

個別介入効果(ITE)が大きくなる傾向
があり, ATEの推定精度が低くなる



Jobsデータセット

概要

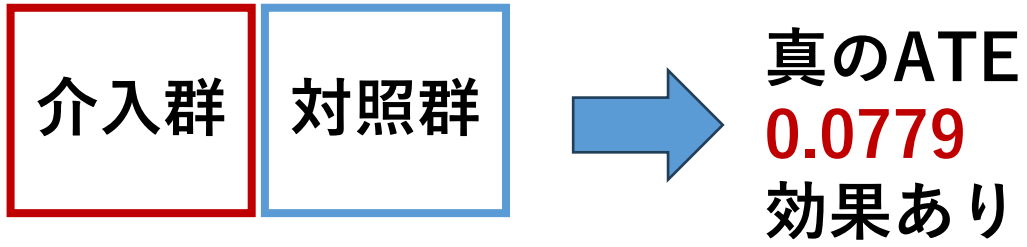
- Jobsデータセット[12]は職業訓練プログラムの効果を評価するための実データセット
- 参加者の雇用状況や収入に関する情報が含まれている

データの内容

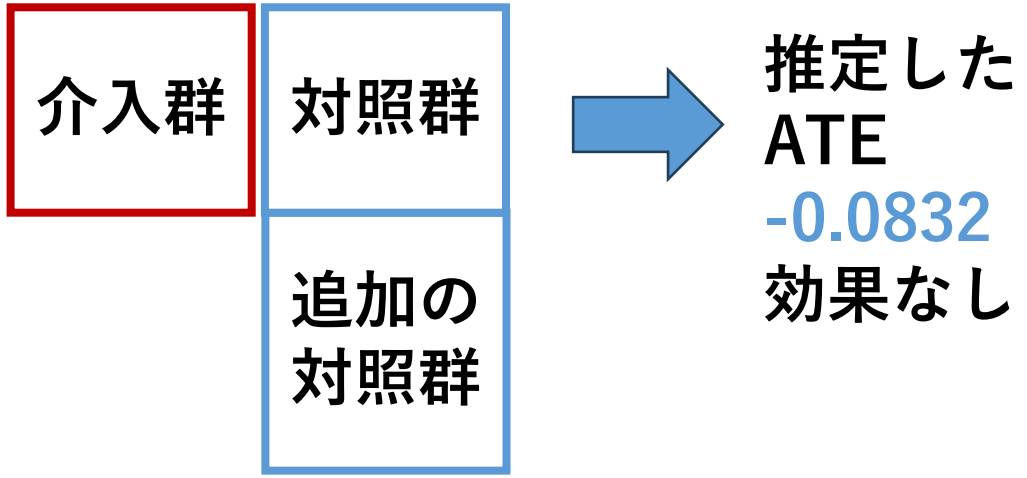
- 共変量：年齢, 最終学歴, 民族(black, hispanic), 婚姻状況, 大学の学位を持っているか, 1975年の収入の7つの変数
- 介入：職業訓練への参加の有無
- 効果：介入後に就職できたかどうか

データの特徴

無作為化比較試験



バイアスのあるデータセットを作成



[12]Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American statistical Association, 94(448):1053–1062,

実験手順

1. Jobsデータを訓練データ：テストデータ ＝ 8 : 2 で分ける
2. 各手法を用いて傾向スコアの推定を行う
   訓練データで学習し, テストデータに適用する
3. 2 で推定した傾向スコアを用いてATEの推定を行う
4. 1 ~ 3 を 5 回繰り返す

評価指標

推定したATEに対しBiasとRMSEとMAE

ATE推定の実験結果

	LR	NN	BOOST	GBN	提案手法
Bias	-0.3755	-0.5486	-0.2939	13.0158	0.0060
RMSE	0.3883	0.5530	0.3384	26.5959	0.1877
MAE	0.3755	0.5486	0.2939	13.4092	0.1515

提案手法の精度が
最も高くなる

Biasの定義

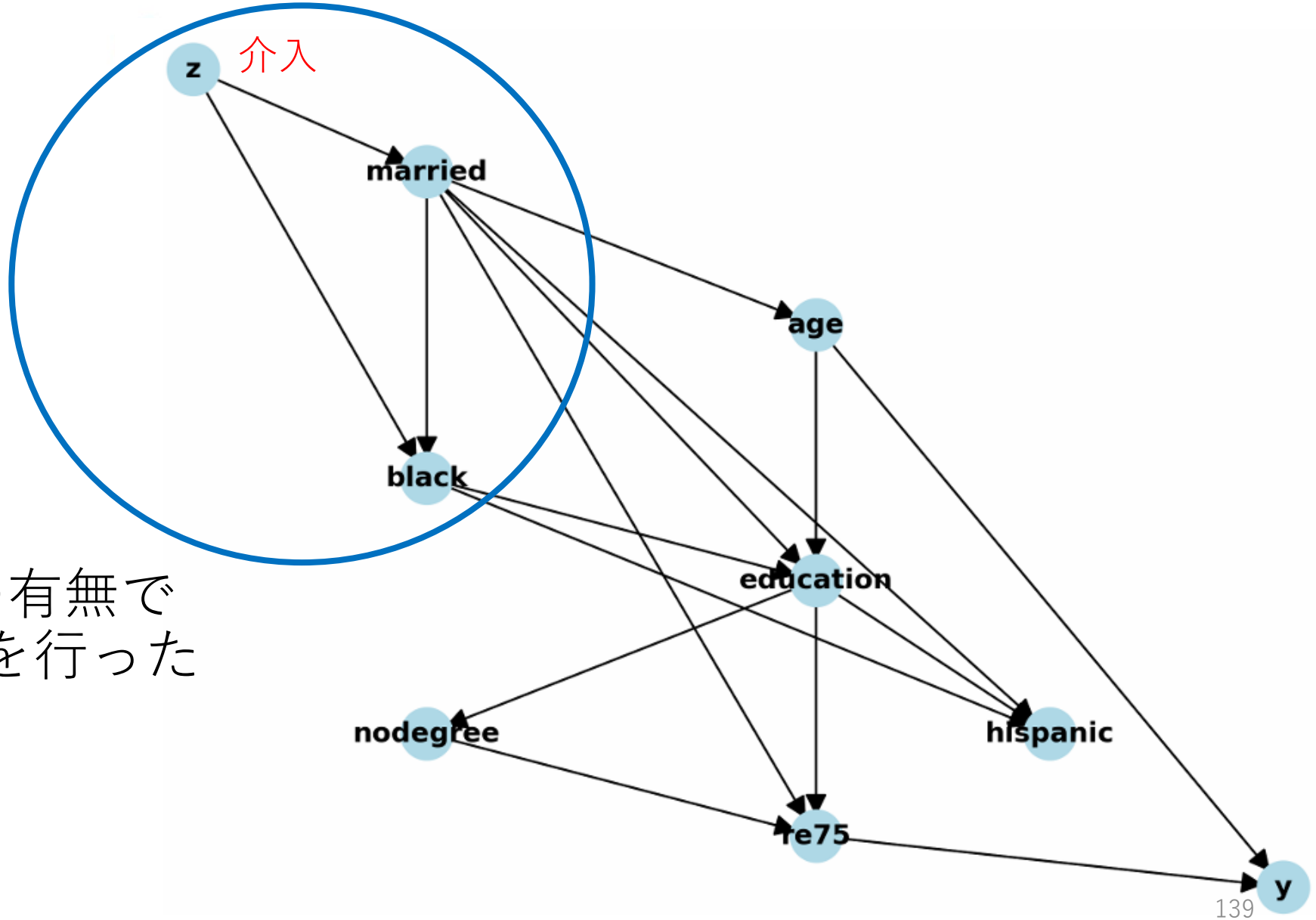
$$Bias = E[\widehat{ATE} - ATE_{true}]$$

より, 以下が成り立つ.

$$E\left[\widehat{ATE}\right] = Bias + E[ATE_{true}] = Bias + 0.0779$$

Bias の結果から, 職業訓練に正の効果があると推定できたのは
GBNと提案手法のみである

Jobsデータの確率的因果構造



提案手法は対象が
黒人の方か, 結婚の有無で
傾向スコアの推定を行った

Twinsデータセット

概要

1989年から1991年の米国での全出生記録から得られた実データセット[13]

乳幼児の出生体重による死亡率を評価するために使用される

データの特徴

体重が重い方　z = 1　　軽い方　z = 0

死亡率: 14.92%　　16.91%

真のATEは-0.0248

データの内容

出生体重が2kg未満の同性の双子を選択

共変量：年齢, 性別, 母親の健康状態, 妊娠, 両親の社会経済的背景などの46変数

介入：双子のうち体重が重い方

効果：１年後に死亡しているかどうか

介入の選択

共変量ベクトルをx, 共変量の個数をdとすると,

$$z|\mathbf{x} \sim Bern(Sigmoid(\mathbf{w}^{\mathrm{T}}\mathbf{x} + n))$$

ここで

$$\mathbf{w}^{\mathrm{T}} \sim \mathcal{U}\left((0.1, 0.1)^{d\times 1}\right), n \sim \mathcal{N}(0, 0.1)$$

[13] Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. The Quarterly Journal of Economics, 120(3):1031–1083, 2005.

ATE推定の実験結果

	LR	NN	BOOST	提案手法
Bias	0.0689	0.0498	0.0317	0.0051
RMSE	0.0784	0.0984	0.0439	0.0417
MAE	0.0689	0.0796	0.0393	0.0328

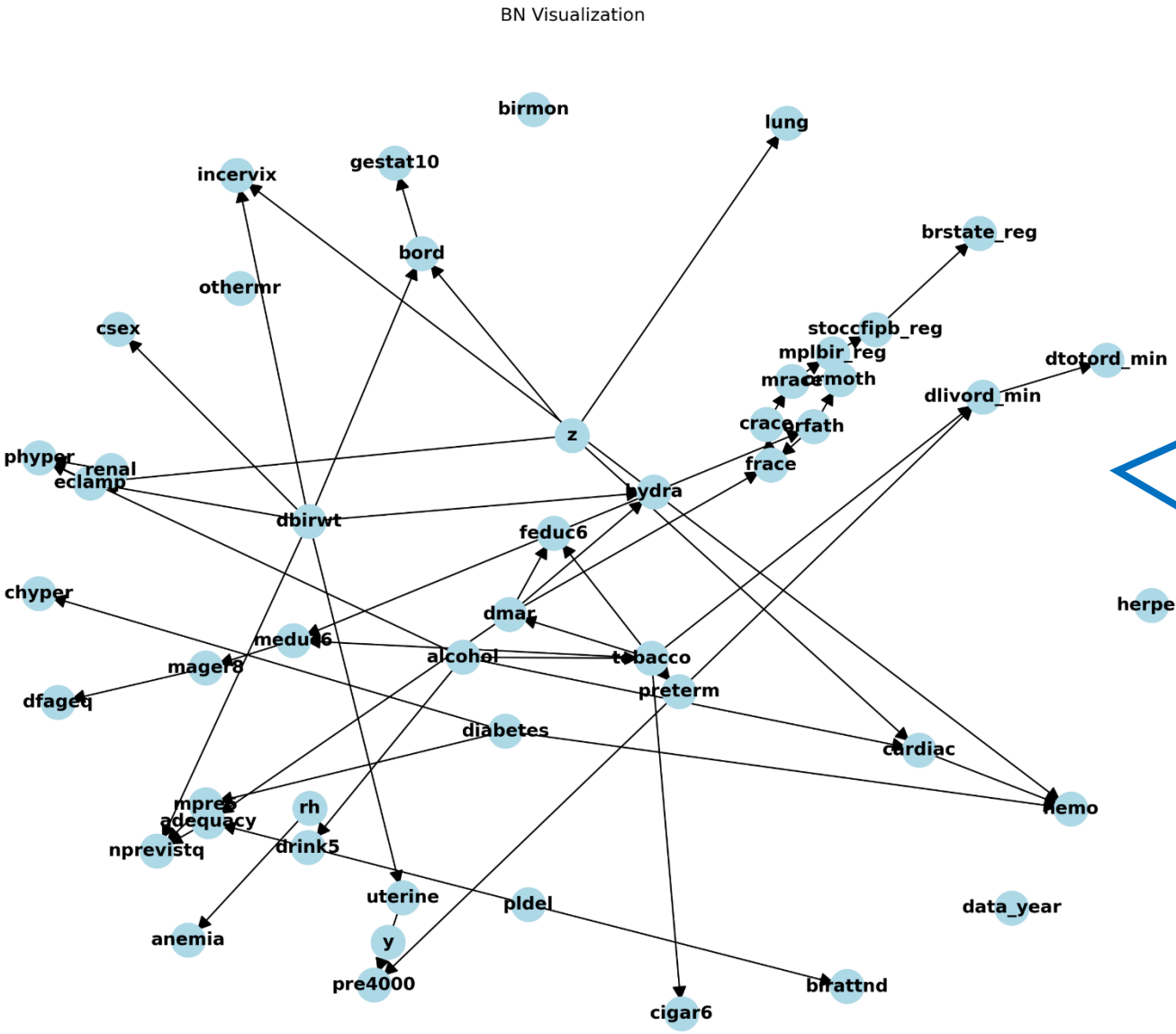
提案手法の精度が最も高くなる

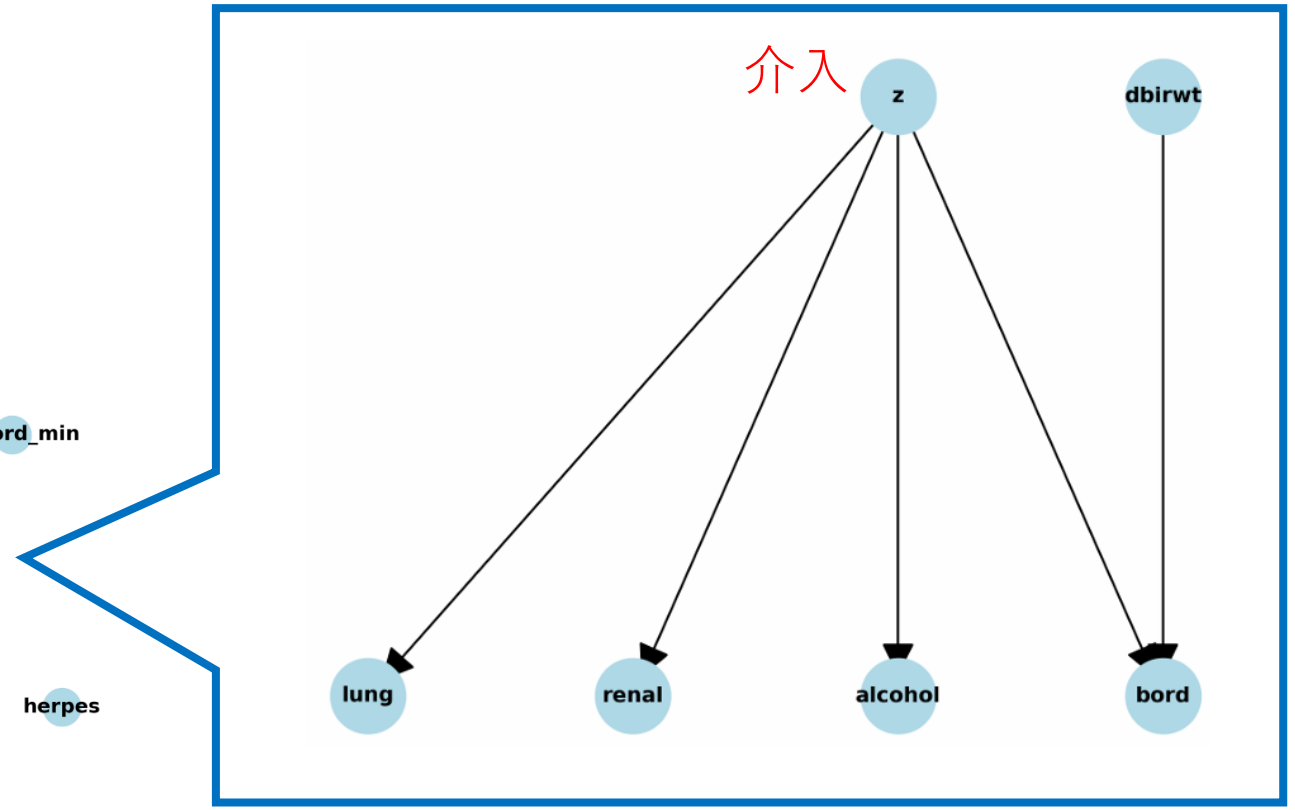
Biasの定義より, 以下が成り立つ.

$$E\left[\widehat{ATE}\right] = Bias + E[ATE_{true}] = Bias - 0.0248$$

Bias の結果から, 推定したATEが負の値をとる, つまり, 体重が重い方が死亡率が低いと推定できたのは提案手法のみである

Twinsデータの確率的因果構造





提案手法は急性及び慢性の肺疾患(lung), 腎臓病(renal), 飲酒 (alcohol), 生まれた順番(bord), 体重(dbirwt)で傾向スコアの推定を行った

むすび

提案手法
真の傾向スコアに漸近的に一致するベイジアンネットワーク分類器を用いた傾向スコア推定法を提案した

結果
シミュレーションデータを用いた評価実験と実データ(jobsとtwins)を用いた評価実験でATE推定は提案手法の精度が高い

今後の研究

- 提案手法は離散値しか扱えないため、連続値を扱えるようにしてより高精度な傾向スコア推定、ATE推定を行いたい

付録:GBNの分類精度が低い要因

- GBNの分類精度が低い要因:
  学習した構造の目的変数の親変数が増大し,
  一つのパラメータ学習のサンプルサイズが小さくなるため

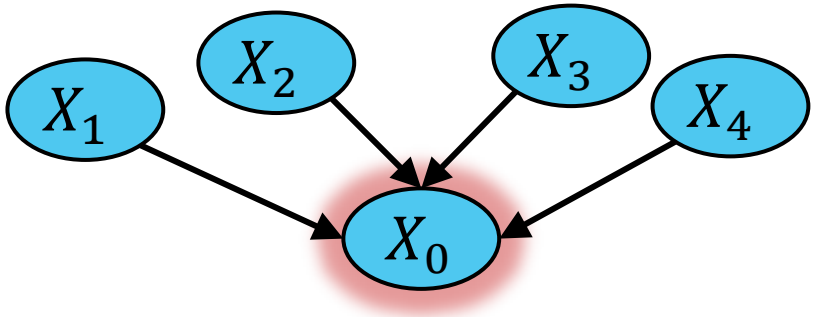


図2: 目的変数の親変数が多いBNの例

[2] S. Sugahara, M. Uto, and M. Ueno, “Exact learning augmented naive Bayes classifier,” Proceedings of the 9th International Conference on Probabilistic Graphical Models, vol.72, pp.439–450, Proceedings of Machine Learning Research, PMLR,2018.

[3] S. Sugahara and M. Ueno, “Exact learning augmented naive Bayes classifier,” Entropy, vol.23, no.12, pp. 1703, 2021.

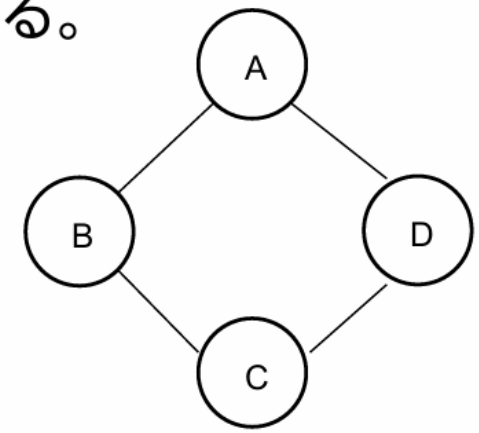
マルコフネットワークとベイジアンネットワーク

無向グラフでの表現

- $I(A, C|B, D)_G \; and \; I(B, D|A, C)_G$

もしくは$\mathrm{A} \perp C|B, D \; and \; \mathrm{B} \perp D|A, C$

は有向グラフでは表現できないが無向グラフでは表現できる。

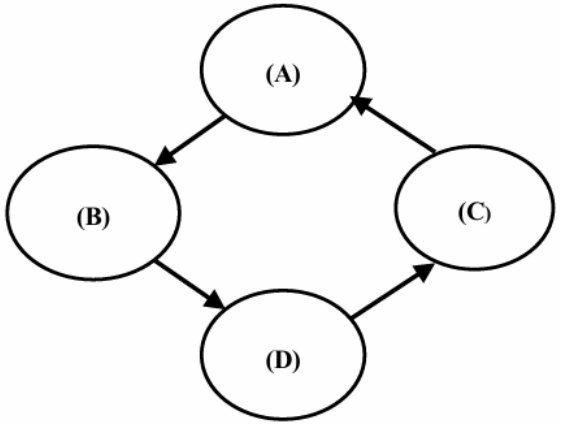


循環構造

- $I(A, C|B, D)_G \; and \; I(B, D|A, C)_G$

もしくは$\mathrm{A} \perp C|B, D \; and \; \mathrm{B} \perp D|A, C$

はベイジアンネットワークでは表現できない。



目的変数の推定値 c

説明変数のデータ$\mathbf{x}=\{x_1, \ldots, x_n\}$を得たとき

$$\hat{c} = arg \max_{c \in \{1,\ldots,r_0\}} p(X_0 = c \,|(\mathbf{x}, G, \boldsymbol{\theta}))$$

n個の変数集合 $\mathbf{V} = \{X_1, \cdots, X_n\}$を持つベイジアンネットワークは$(G, \Theta)$
で表現される.

Gは$\mathbf{V}$に対応するノード集合によって構成されるDAG

ΘはGの各エッジに対応する条件付き確率パラメータ集合$p(X_i|\,\mathbf{Pa}(X_i, G)\,)\ (i = 1, \ldots n)$である.

($\mathbf{Pa}(X_i, G)$は変数X_iの親変数集合)

条件付パラメータ集合Θ

- 条件付パラメータ集合Θ

$$\Theta = \bigcup_{i=0}^{n} \bigcup_{j=1}^{q_i} \bigcup_{k=1}^{r_i} \{\theta_{ijk}\}$$

Pa(Xi,G) がj 番目のパターンをとったとき(Pa(Xi,G) = j)に,
Xi = kとなる条件付き確率パラメータとする.
qi はPa(Xi,G)の取りうるパターン数

EAP

- θ ijk の推定法には, その期待値であるExpected A Posteriori (EAP) が最も良く用いられる.
- サンプルがN 個 あるデータが得られた時のEAPは条件付き確率パラメータの事前分布にディリクレ分布 を仮定すると以下で表される

$$\hat{\theta}_{ijk} = \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}$$

ここで, N_{ijk} は $X_i = k$ かつ $\mathbf{Pa}(X_i, G) = j$ となる頻度を表す. また, α_{ijk} はディリクレ事前分布のハイパーパラメータを表し, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$, $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ である.

傾向スコア

Rosenbaum & Rubin [2] は傾向スコアを用いた共変量調整法を提案した.

傾向スコアe(x)：共変量 x を所与としたときの対象が介入を受ける確率

$$e(x) = p(z = 1|y_1, y_0, x) = p(z = 1|x)$$

このとき

$$(y_1, y_0) \perp z \mid e(x), \quad 0 < p(z = 1 \mid e(x)) < 1$$

この条件のもと，$E(y_j \mid e(x)) = E(y_j \mid e(x), z = j), (j = 1,0)$

となり，$ATE = E_x(E(y_1 \mid e(x), z = 1) - E(y_0 \mid e(x), z = 0))$

が成り立つため，逆確率重み付け法(Inverse probability weighting : IPW)でATEを求められる[3].

$$\mathrm{ATE} = \frac{1}{N}\sum_i^N \frac{y_{1i}}{e(x_i)} z_i - \frac{1}{N}\sum_i^N \frac{y_{0i}}{1 - e(x_i)}(1 - z_i)$$

Nは対象全体の数, y_{1i}は対象iが介入を受けた時の結果変数, y_{0i}は対象iが介入を受けなかった時の結果変数, x_iは対象iの共変量, z_iは対象iの介入の有無

[2] Rosenbaum, P. R., and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.
[3]K. Hirano, G.W. Imbens, and G. Ridder, “Efficient estimation of average treatment effects using the estimated propensity score,” Econometrica,

研究概要

- 因果効果を推定するために傾向スコアが提案され，
  広く用いられている[1]

- 傾向スコア推定のために多くのアプローチが提案されているが，
  既存の手法では，確率値を正確に推定できない

- 確率値を正確に推定できるベイジアンネットワーク分類器を用いた
  傾向スコア推定法を提案する

[1] Rosenbaum, P. R., and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

平均介入効果

対象が介入を受けた場合の結果変数をy_1, 介入を受けなかった場合の結果変数をy_0とする. また, 変数zは対象が介入を受けたとき, $z = 1$, 介入を受けなかったとき, $z = 0$を取るとする.

$z = 1$となる対象の集合を介入群, $z = 0$となる対象の集合を対照群とすると,
介入群, 対照群の平均介入効果(ATE : Average Treatment Effect)は以下のように定義される[2]

$$ATE = E(y_1 - y_0) = E(y_1) - E(y_0)$$

$E(y_1)$: すべての対象が介入を受けたときの結果の期待値
$E(y_0)$: すべての対象が介入を受けなかったときの結果の期待値

しかし, 同一の対象では, y_1, y_0を同時に観測できない

[2] Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688–701.

無作為化比較試験

無作為化比較試験では, $(y_1, y_0) \perp z$が成り立つため,

$$E(y_j) = E(y_j \mid z = j), (j = 1, 0)$$

となるので,

$$ATE = E(y_1 - y_0) = E(y_1 \mid z = 1) - E(y_0 \mid z = 0)$$

となり, 介入群での結果の期待値 – 対照群での結果の期待値で
因果効果を求めることができる

しかし、多くの分野では無作為化比較実験を行うことが困難である.

そのような実験では介入群と対照群で共変量の分布が合っていないため

$$ATE \neq E(y_1 \mid z = 1) - E(y_0 \mid z = 0)$$

であり, ATEを計算することができない

傾向スコア

Rosenbaum & Rubin [2] は傾向スコアを 用いた共変量調整法を提案した.

共変量をxとすると, 傾向スコア$e(x)$とはある対象が介入を受ける確率を表したスコアである[2]. 変数zは対象が介入を受けたとき, $z=1$, 介入を受けなかったとき, $z=0$を取るとする.

$$e(x) = p(z = 1 \mid y_1, y_0, x) = p(z = 1 \mid x)$$

このとき

$$(y_1, y_0) \perp z \mid e(x), 0 < p(z = 1 \mid e(x)) < 1$$

が成り立つため, IPW(逆確率重み付け法)でATEを求められる[3].

$$\mathrm{ATE} = \frac{1}{N}\sum_i^N \frac{y_{1i}}{e(x_i)} z_i - \frac{1}{N}\sum_i^N \frac{y_{0i}}{1-e(x_i)}(1-z_i)$$

Nは対象全体の数, y_{1i}は対象iが介入を受けた時の結果変数, y_{0i}は対象iが介入を受けなかった時の結果変数, x_iは対象iの共変量, z_iは対象iの介入の有無

[2] Rosenbaum, P. R., and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

[3]K. Hirano, G.W. Imbens, and G. Ridder, “Efficient estimation of average treatment effects using the estimated propensity score,” Econometrica, vol.71, no.4, pp.1161–1189, 2003.

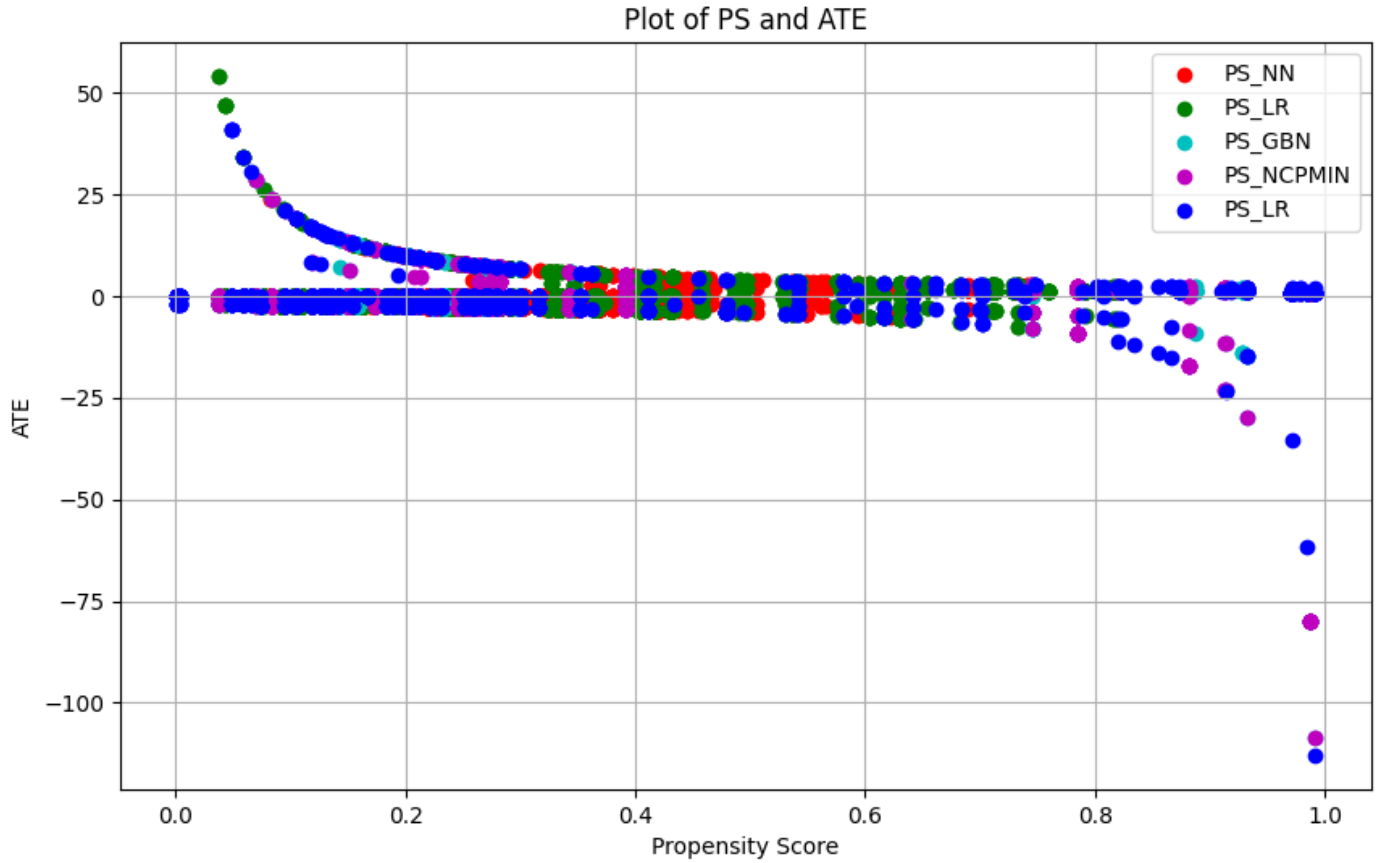
ベイジアンネットワーク分類器を用いた傾向スコアの推定方法

学習したベイジアンネットワーク分類器の構造をG,
条件付き確率パラメータ集合をΘとすると, 以下の式で傾向スコアが推定できる

$$
\begin{aligned}
&e(x \mid G, \Theta) = p(z = 1 \mid x, G, \Theta) \\
&= p(z = 1 \mid x_1, x_2, \ldots, x_n, G, \Theta) \\
&= \frac{p(z = 1, x_1, x_2, \ldots, x_n \mid G, \Theta)}{p(x_1, x_2, \ldots, x_n \mid G, \Theta)}
\end{aligned}
$$

Boostの傾向スコア推定とATE推定

- 外れ値をなくしたとき



Strong ignorabilityの仮定

Strong ignorability[1]: zや(y_1, y_0)に影響を及ぼす共変量xについて,

$$(y_1, y_0) \perp z \mid x, \qquad 0 < p(z = 1 \mid x) < 1$$

この条件のもとでは, $E(y_j \mid x) = E(y_j \mid x, z = j), (j = 1,0)$

となるため, $ATE = E_x(E(y_1 \mid x, z = 1) - E(y_0 \mid x, z = 0))$

で求めることが可能になる