3M1-2

漸近一致性を有する大規模ベイジアンネットワーク学習

Consistent learning Bayesian networks with thousands variables

名取和樹 *1　宇都雅輝 *1　植野真臣 *1
Kazuki Natori　Masaki Uto　Maomi Ueno

*1電気通信大学大学院情報理工学研究科
Graduate school of Informatics and Engineering, The University of Electro-Communications

This study proposes a new constraint-based learning Bayesian network using Bayes factor. Specifically, our proposed method is a learning algorithm applying conditional independence test using Bayes factor to the recursive autonomy identification algorithm that is the state of art algorithm in constraint-based learning. The proposed method is expected to learn larger network structures than the traditional methods do because it greatly improves computational efficiency. This paper presents some experiments related to learning large network structures. Results show that the proposed method can learn surprisingly huge networks with thousands of variables.

1. はじめに

ベイジアンネットワークは，循環有向グラフ（Directed Acyclic Graph: DAG）と条件付き確率で定義される確率的グラフィカルモデルである．一般に，ベイジアンネットワークのグラフ構造は未知であり，データから推定する必要がある．この問題をベイジアンネットワークの構造学習と呼ぶ．

ベイジアンネットワークの構造学習は，一般に漸近一致性を有する学習スコアを用いてすべて構造の候補からスコアが高い構造を探索する．このアプローチはスコアベースアプローチと呼ぶ．しかし，構造の候補を全探索するため，ノード数の増加に伴い，探索数が指数的に増加する問題がある．この問題を緩和するために，動的計画法 [Silander+06]，A*探索 [Yuan+11]，幅優先分岐限定法 [Malone+11]，整数計画法 [Cussens11] と言った従来の探索アプローチを用いた構造学習法が提案されてきた．しかし，最先端手法でさえ，最大 60 ノード程度の学習が限界である．

一方で，因果モデルの研究分野では，大幅に計算量が削減できる制約ベースアプローチと呼ばれる構造学習法が提案されてきた．このアプローチでは，まずすべての 2 ノードの組合せについて条件付き独立性（Conditional Independence: CI）テストを行い無向グラフを推定する．その後，オリエンテーションルール [Verma+92] による方向付けにより DAG を得る．制約ベースアプローチの代表的な学習アルゴリズムとして，PC（Peter and Clark）アルゴリズム [Spirtes+00]，MMHC（Max-Min Hill Climb）アルゴリズム [Tsamardinos+06]，RAI（Recursive Autonomy Identification）アルゴリズム [Yahezkel+09] が知られている．しかし従来の CI テストは，漸近一致性を持たないため，これを用いた学習アルゴリズムは，真の構造を学習できる保証がない．

そこで本研究では，漸近一致性を有する CI テストを制約ベースアプローチに適用した新たな構造学習アプローチを提案する．具体的には，制約ベースアプローチの最先端手法である RAI アルゴリズムに Bayes factor を用いた CI テストを適用する．提案手法では，画期的に計算量を削減できるだけでなく，漸近一致性を有する大規模構造学習を実現できる．

連絡先: 電気通信大学大学院情報理工学研究科，
東京都調布市調布ヶ丘 1-5-1，042-484-8585，
{natori,uto,ueno}@ai.lab.uec.ac.jp

本論文では，数百，数千のノードを有するベンチマークネットワークを用いて，提案手法の有効性を示す．

2. ベイジアンネットワーク

ベイジアンネットワークは，確率構造に非循環有向グラフ（Directed Acyclic Graph: DAG）を仮定することにより同時確率分布を条件付き確率の積に分解できる．

今，$\{x_1, \cdots, x_N\}$ を N 個の離散確率変数集合 $\mathbf{X}$ とし，各変数 x_i は r_i 個の状態集合 $\{1, \cdots, r_i\}$ から一つの値をとるとする．ここで，変数 x_i が値 k をとるとき，$x_i = k$ と書く．このとき，ベイジアンネットワークの構造 g において，各変数 x_i の親変数集合 Π_{i} とした時の同時確率分布 $p(x_1, \cdots, x_N \mid g)$ は以下の通りに表現できる．

$$p(x_1, \cdots, x_N \mid g) = \prod_{i=1}^{N} p(x_i \mid \Pi_{\mathrm{i}}, g) \tag{1}$$

2.1 ベイジアンネットワークの構造学習

ベイジアンネットワークの構造学習では一般に，漸近一致性を有する学習スコアが最も高い構造を探索するスコアベースアプローチが利用されていきた．学習スコアには周辺尤度スコアが一般に用いられる．

今，変数 $x_i (i = 1, \cdots, N)$ の親ノード集合 Π_i が $j (j = 1, \cdots, q_i)$ 番目のパターンを取り，変数 x_i が状態値 $k (k = 1, \cdots, r_i)$ を取るときの出現頻度 n_{ijk} としたときの構造 g における周辺尤度 $p(\mathbf{D} \mid g)$ は以下となる．

$$p(\mathbf{D} \mid g) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \tag{2}$$

ここで，α_{ijk} はユーザが設定するハイパーパラメータを表す．また $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$，$n_{ij} = \sum_{k=1}^{r_i} n_{ijk}$ である．

[Heckerman+95] では，二つの構造が等価であるなら，それらのパラメータの同時確率密度は同一でなければならないという尤度等価を導入した．そして，尤度等価に矛盾しないディレクレ分布の条件として，以下のハイパーパラメータを提案した．

$$\alpha_{ijk} = \alpha p(x_i = k, \Pi_i = j \mid g^h) \tag{3}$$

ここで，α は Equivalent Sample Size (ESS) と呼ばれ事前知識の重みを示す擬似サンプルである．g^h はユーザの仮説であり，この構造を所与として ESS を α_{ijk} に分配する．この指標は，Bayesian Dirichlet equivalent (BDe) 呼ばれる．さらに，ESS をパラメータ数で除し，$\alpha_{ijk} = \alpha/r_i q_i$ としたスコアを提案している．このスコアは BDe の特殊形とみなすことができ，Bayesian Dirichlet equivalent uniform（BDeu）と呼ばれる．[Heckerman+95] や [Ueno 10][Ueno 11] の研究では，シミュレーション実験により無情報事前分布を用いた BDeu が最も有用であることを報告している．

この学習スコアが最も良い構造を得る厳密解探索手法として，動的計画法 [Silander+06]，A*探索 [Yuan+11]，幅優先分岐限定法 [Malone+11]，整数計画法 [Cussens11] を用いた構造学習法が提案されてきた．しかしこれらのアルゴリズムを用いても，最大 60 ノード程度の構造学習を限界とし，大規模構造学習の実現は難しい．

2.2 制約ベースアプローチによる構造学習

因果モデルの研究分野では，大幅に計算量を削減できる制約ベースアプローチと呼ばれる構造学習法が提案されてきた．このアプローチの基本的なアルゴリズムは以下の通りである．

1. データから推定される完全無向グラフを生成する．
2. CI テストにより辺を削除する．
3. (2) で得られた無向グラフに対してオリエンテーションルール [Verma+92] を用いて方向付けを行う．

2.3 CI テスト

制約ベースアプローチで用いられる CI テストは，一般に条件付き相互情報量（Conditional Mutual Information: CMI）を用いて 2 ノード間の独立性を検出する．

今，$\mathbf{Z}$ を所与として XY 間の $\mathrm{CMI}(X, Y \mid \mathbf{Z})$ は以下のように定義される．

$$\mathrm{CMI}(X, Y \mid \mathbf{Z}) = \sum_{x \in X, y \in Y, \mathbf{z} \in \mathbf{Z}} p(x, y, \mathbf{z}) \log \frac{p(x, y \mid \mathbf{z})}{p(x \mid \mathbf{z}) p(y \mid \mathbf{z})} \tag{4}$$

この値が，ユーザの設定した閾値以上の場合，従属であると判断する．

制約ベースアプローチにおける構造学習は，CI テストの精度に依存する．この CI テストは漸近一致性を持たないため，厳密な構造学習が保証されていない．

3. 漸近一致性を有する CI テスト

漸近一致性を有する CI テストとして，[Steck+02] は，Bayes factor を用いた CI テストを提案している．Bayes factor は，２つのモデルの周辺尤度の比で表される．

例として，X_1 と X_2 間について，各ノードの共通の親ノード集合 $\mathbf{C}$ としたときの，従属モデルを g_1，独立モデルを g_2 とし，それぞれ図 1，図 2 に示す．

観測されたデータ集合 $\mathbf{D} = \{D_1, \cdots, D_n\}$ としたときの Bayes factor は $p(\mathbf{D} \mid g_1)/p(\mathbf{D} \mid g_2)$ で表される．Steck らでは，対数 Bayes factor を求め，0 以上か否かで CI テストを行った．

しかしこの手法は，ベイジアンネットワークの構造学習に応用されていない．

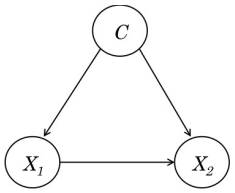

図 1: 従属モデル g_1

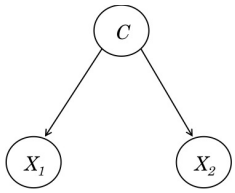

図 2: 独立モデル g_2

4. 提案手法

Bayes factor による CI テストは，漸近一致性を有するが，ベイジアンネットワークの構造学習に用いられていなかった．そこで本論文では，Bayes factor を用いた CI テストを制約ベースアプローチに適用した新たなアプローチを提案する．

制約ベースアプローチによる構造学習は，CI テストにより不要な辺を削除することで無向グラフを学習する．しかし [Steck+02] では，X_1，X_2 間を有向辺として CI テストを行っている．このとき，２ノードの各親変数パラメータ数が異なるため，CI テストの精度が不安定になる．

そこで本論文では，X_1，X_2 間を無向辺とした周辺尤度，式 (5)，(6) を導出し，それに基づく Bayes factor を用いた CI テストを提案する．

$$p(\mathbf{D} \mid g_1) = \prod_{j=1}^{q} \frac{\Gamma(r_1 r_2 \alpha_{g_1})}{\Gamma(r_1 r_2 \alpha_{g_1} + n_j)} \prod_{k_1=1}^{r_1} \prod_{k_2=1}^{r_2} \frac{\Gamma(\alpha_{g_1} + n_{jk_1k_2})}{\Gamma(\alpha_{g_1})} \tag{5}$$

$$p(\mathbf{D} \mid g_2) = \prod_{i=1}^{2} \prod_{j=1}^{q} \frac{\Gamma(r_i \alpha_{g_i})}{\Gamma(r_i \alpha_{g_i} + n_{ij})} \prod_{k_i=1}^{r_i} \frac{\Gamma(\alpha_{g_i} + n_{jk_i})}{\Gamma(\alpha_{g_i})} \tag{6}$$

式 (5) において，$n_{jk_1k_2}$ は変数 x_1，x_2 の共通の親変数集合 Π_i が $j (j = 1, \cdots, q)$ 番目のパターンを取るとき，変数 x_1，x_2 の状態値 $x_1 = k_1 (k_1 = 1, \cdots, r_1)$，$x_2 = k_2 (k_2 = 1, \cdots, r_2)$ となるデータの出現頻度を表す．式 (6) において，n_{jk_i} は共通の親変数集合 Π_i が j 番目のパターンを取るとき，変数 x_i の状態値 $x_i = k_i (k_i = 1, \cdots, r_i)$ となるデータの出現頻度を表す．また，α_{g_1}，α_{g_i} はハイパーパラメータを表す．

ここで，ハイパーパラメータ α_{g_1}，α_{g_i} の値を設定しなければならない．[Clarke+94] は，ジェフリーズの事前分布が事前分布のエントロピーのリスクを最小とすることを示し，ハイパーパラメータの漸近的な最適値として 1/2 を提案している．そこで本論文では，ハイパーパラメータの値として 1/2 を用いる．

本論文では，この漸近一致性を有する Bayes factor を用いた CI テストを制約ベースアプローチの RAI（Recursive Autonomy Identification）アルゴリズムに適用する．

5. RAI アルゴリズム

RAI(Recursive Autonomy Identification) アルゴリズム [Yahezkel+09] は，制約ベースアプローチにおいて最先端の学習アルゴリズムとして知られている．RAI アルゴリズムは，これまでの制約ベースアプローチでの精度悪化の原因であった高次の CI テストを抑える事を目的として開発された学習アルゴリズムである．RAI アルゴリズムの詳細をアルゴリズム 1 に示す．

Algorithm 1 The RAI algorithm

Require: $\mathbf{V} = \{X_1, \ldots, X_N\}, \mathbf{D} = \{D_1, \ldots, D_n\}$
Ensure:
$G_{out} = RAI(N_s, G_{input}(\mathbf{V}_{input}, \mathbf{E}_{input}), G_{ex}(\mathbf{V}_{ex}, \mathbf{E}_{ex}), G_{all})$
1: **if** *all nodes in* G_{start} *have fewer than* $N_s + 1$ *potential parents* **then**
2: **return** $G_{out} = G_{all}$
3: **else**
4: **for** $Y \in G_{start}, X \in G_{ex}$ **do**
5: **if** $X \perp Y|\mathbf{S}, \exists \mathbf{S} \subset Pa(Y, G_{start}) \cup Pa_p(Y, G_{ex}) \setminus X$ *and* $|\mathbf{S}| = N_s$ **then**
6: *remove the edge between* X *and* Y *from* G_{all}
7: **end if**
8: **end for**
9: *Direct the edges in* G_{start} *using orientation rules*
10: **for** $Y \in G_{start}, X \in G_{start}$ **do**
11: **if** $X \perp Y|\mathbf{S}, \exists \mathbf{S} \subset Pa(Y, G_{ex}) \cup Pa_p(Y, G_{start}) \setminus X$ *and* $|\mathbf{S}| = N_s$ **then**
12: *remove the edge between* X *and* Y *from* G_{all} *and* G_{start}
13: **end if**
14: *Direct the edges in* G_{start} *using orientation rules*
15: *Group the nodes having lowest topological order into a descendant sub − structure* G_D
16: *Remove* G_D *from* G_{start} *temporarily and define the resulting unconnectedstructures as ancestor sub − structures* $G_{A_1}, \ldots, G_{A_k}$
17: **end for**
18: **for** $i = 1$ *to* k **do**
19: *Call* $RAI(N_s + 1, G_D, G_{ex_D}, G_{all})$
20: **end for**
21: *Define* $G_{ex_D} = \{G_{A_1}, \ldots, G_{A_k}, G_{ex}\}$ *as the exogenous set to* G_D
22: *Call* $RAI(N_s + 1, G_D, G_{ex_D}, G_{all})$
23: $G_{out} = G_{all}$
24: **end if**

まず 1 行目において，CI テストの次数 N_s とすると，各ノードが $N_s + 1$ より少ない潜在親変数を持つとき，RAI アルゴリズムの終了条件として定める．そして 2 行目から 14 行目は制約ベースアプローチの基本的なアルゴリズム (1) から (3) を表している．RAI アルゴリズムでは，15 行目で辺の方向付けを基に全体構造を親と子の部分構造に分割する．そして 18 行目から 22 行目で分割した子構造・親構造の部分構造毎で再帰的に RAI アルゴリズムを実行し構造学習を行う．これにより，これまでの制約ベースアプローチのアルゴリズムにおいて最も精度の良い構造学習を実現した．

6. 実験

本論文での提案手法の学習精度を評価するために，ベンチマークデータによる構造学習を行った．具体的には，提案手法と Steck らの手法，閾値に 0.05 と設定した CMI を用いた CI テストをそれぞれ RAI アルゴリズムに適用し，学習の精度評価を行った．また，A^* 探索を用いた構造学習法も比較手法に加え，学習時間が 12 時間で打ち切った．本論文では，真の構造と学習構造の距離を表す SHD（Structural Hamming Distance）[Tsamardinos+06] を評価指標として用いた．また本実験では，ベイジアンネットワークリポジトリ *bnlearn*[Scurari 10] に登録されている sachs（変数：11，辺数：17），win95pts（変数：76，辺数：112），andes（変数：223，辺数：338），munin（変数：1041，辺数：1397）を用いて，それぞれデータ数を 10000，20000，50000，100000，200000 と増やし，各データ数で 5 回学習を行った．

ベンチマークネットワークである sachs，win95pts，andes，munin についてそれぞれ構造学習を行った結果を図 3，4 に示す．

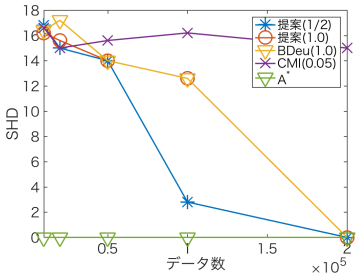

図 3: sachs の SHD

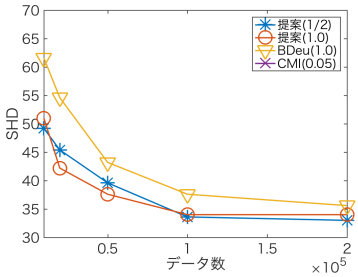

図 4: win95pts の SHD

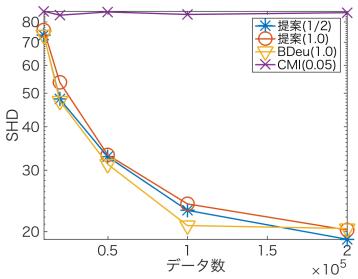

図 5: andes の SHD

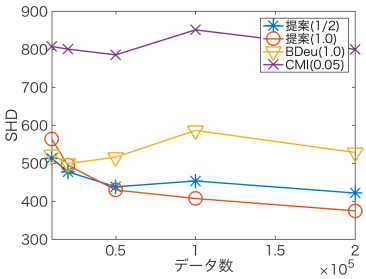

図 6: munin の SHD

横軸はデータ数，縦軸は 5 回の SHD の平均を示し，比較手法はそれぞれ，ハイパーパラメータに 1/2，1 を設定したものを “提案（1/2）”，“提案（1.0）”，Steck らの手法を “BDeu（1.0）”，CMI を “CMI(0.05)”，A^* 探索を用いた構造学習法を A^* と示している．

まず CMI は，2 つのベンチマークデータの学習結果は共に他の手法に比べて劣る精度となり，データ数を増加しても，精度は変化せず，漸近一致性がないことが確認された．また A^* は，sachs を用いた実験では，学習することができたが，その他のベンチマークネットワークでは，学習できなかった．

sachs を用いた実験結果では，すべてのデータ数において A^* は SHD が 0 に収束した．ハイパーパラメータに 1/2 を設定した提案手法は，すべてのデータ数において A^* を除いた他

手法と比べ，良い精度となりデータ数 200000 で SHD が 0 に収束した．ハイパーパラメータに 1.0 を設定した提案手法は，Steck らの手法とほぼ同等の精度を有したままデータ数 200000 で SHD が 0 に収束した．これにより，提案手法は漸近的に真の構造を学習できることが確認できた．

win95pts を用いた実験結果では，CMI はすべてのデータ数において非常に悪い精度となり，図**??**中に示すことができなかった．ハイパーパラメータに 1/2 と 1.0 を設定した提案手法は，すべてのデータ数においてほぼ同等の精度となり，Steck らの手法と比べ良い精度となった．

andes を用いた学習結果では，データ数が 10000 の時，提案手法と Steck らの手法はほぼ同じ精度を有しているが，100000 までは Steck らの手法が最も良い精度となっている．データ数が十分に多い 200000 では，ハイパーパラメータに 1/2 と設定した提案手法が最も良い精度となった．

munin を用いた学習結果では，データ数が 20000 まで提案手法と Steck らの手法は同等の精度となったが，データ数 50000 以上では，提案手法が Steck らの手法と比べ良い精度となった．またデータ数 50000 まではハイパーパラメータにそれぞれ 1/2，1.0 を設定した提案手法は同等の精度となった．しかし，データ数が 100000 以上では，ハイパーパラメータに 1.0 を設定した提案手法が最も良い精度となってしまった．変数数 1000 を超えるネットワークでは，データ数 200000 でもデータ数が少ないため，精度が不安定になったと考えられる．そのため，今後より多くのデータ数での実験によりハイパーパラメータに 1/2 を設定した提案手法の有効性を検証していく必要がある．

7. まとめ

本論文では，漸近一致性を有する CI テストを制約ベースアプローチに適用した新たな構造学習アプローチを提案した．具体的には，Bayes factor を用いた CI テストを導出し，これを制約ベースアプローチの最先端手法である RAI アルゴリズムに適用した手法を提案した．ベンチマークネットワークによる学習評価において，小規模ネットワークでは，提案手法が A^* と同様に真の構造を学習できることが確認できた．さらに，これまでのスコアベースアプローチでは学習できない，数百，数千といったノードを有する構造学習が実現できることを示せた．また，従来から用いられている CMI を用いた CI テストを用いた場合と比べ，精度が向上しデータ数の増加により漸近的に真の構造に近づくことが確認できた．

しかしながら，1000 変数以上のネットワークを用いた評価では，200000 データまでとデータ数が少ないことから，提案手法が不安定な結果となった．そのため，より多くのデータ数を用いた実験により提案手法の有効性を検証する必要がある．

今後は，さらに多くのベンチマークネットワークとデータ数により，提案手法の有効性を検証する．

参考文献

[Clarke+94] Clarke, Bertrand S. and Barron, Andrew R. ”Jeffreys’ prior is asymptotically least favorable under entropy risk.”, *Journal of Statistical Planning and Inference*, (1994).

[Cussens11] J. Cussens ”Bayesian network learning with cutting planes.”, *Proc. of the 27th Int. Conf. Uncertainty in Artificial Intelligence*, (2011).

[Heckerman+95] D. Heckerman and D. Geiger and D.M. Chickering ”Learning Bayesian Networks: The Combination of Knowledge and Statistical Data”, *Machine Learning*, (1995).

[Malone+11] Malone, B. and Yuan, C. and Hansen, E. and Bridges, S ”Improving the Scalability of Optimal Bayesian Network Learning with External-Memory Frontier Breadth-First Branch and Bound Search.”, *Proc. of the 27th Int. Conf. Uncertainty in Artificial Intelligence*, (2011).

[Natori+15] Natori K., Uto, M., Nishiyama, Y., Kawano, S., and Ueno, M. ”Constraint-Based Learning Bayesian Networks Using Bayes Factor.” *Workshop on Advanced Methodologies for Bayesian Networks*, Springer, (2015).

[Scurari 10] Scutari, M. ”Learning Bayesian Networks with the bnlearn R Package”, *Journal of Statistical Software*, (2011).

[Silander+06] Silander, T. and Myllymaki, P. ”A simple approach for finding the globally optimal Bayesian network structure”, *Proc. of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, (2006).

[Spirtes+00] Spirtes, P. and Glymour, C. and Scheines, R: ”Causation, Prediction, and Search”, MIT press (2000).

[Steck+02] Steck, H., and Tommi S. Jaakkola. ”On the Dirichlet prior and Bayesian regularization.” *Int. Conf. on Neural Information Processing Systems*, (2002).

[Tsamardinos+06] Tsamardinos, Ioannis and Brown, Laura E. and Aliferis, Constantin F. ”The Max-min Hill-climbing Bayesian Network Structure Learning Algorithm”, *Machine Learning*, (2006).

[Yahezkel+09] Yehezkel, R. and Lerner, B. ”Bayesian Network Structure Learning by Recursive Autonomy Identification”, *Journal of Machine Learning Research (JMLR)*, (2009).

[Yuan+11] Yuan, C. and Malone, B. and Xiaojian, W. ”Learning Optimal Bayesian Networks Using A* Search” *Proc. of the 22nd Int. Joint Conf. on Artificial Intelligence*, (2011).

[Verma+92] Verma, T., and Pearl, J. ”An algorithm for Deciding if a Set of Observed Independencies Has a Causal Explanation”, *Proc. of Conference on Uncertanty in Artificial Intelligence*, (1992).

[Ueno 10] Ueno, M. ”Learning networks determined by the ratio of prior and data”, *Proc. of the 26th Int. Conf. on Uncertainty in Artificial Intelligence*, (2010).

[Ueno 11] Ueno, M. ”Robust learning Bayesian networks for prior belief”, *Proc. of the 27th Int. Conf. on Uncertainty in Artificial Intelligence*, (2011).