THE IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS (JAPANESE EDITION)

電子情報通信学会論文誌 D

情報・システム

VOL. J105-D NO. 11
NOVEMBER 2022

情報・システムソサイエティ

一般社団法人 電子情報通信学会

THE INFORMATION AND SYSTEMS SOCIETY
THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS

論　文

分類影響パラメータ数を最小化するベイジアンネットワーク分類器学習

菅原　聖太[†a)]　　植野　真臣[†b)]

Learning Bayesian Network Classifiers to Minimize the Class Variable Parameters

Shouta SUGAHARA[†a)] and Maomi UENO[†b)]

あらまし　本論では，分類に影響する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ベイジアンネットワーク分類器の構造学習手法を提案する．提案手法は以下の二つの学習ステップから構成される．第一ステップでは，目的変数から始まる変数順序それぞれに対して，周辺ゆう度を最大化する構造を探索する．第二ステップでは，第一ステップで得られた構造の中で，分類に影響するパラメータ数を最小にする構造を探索する．サンプルサイズが十分に大きいとき，提案手法は真の分類確率の推定を保証する．更に，提案手法は周辺ゆう度を最大化する一般的なベイジアンネットワーク構造学習法よりも計算時間が短い．リポジトリデータセットを用いた実験により，提案手法の優位性を示す．

キーワード　ベイジアンネットワーク，分類器，機械学習，確率的グラフィカルモデル

1. まえがき

ベイジアンネットワーク (Bayesian network: BN) 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) で表す確率的グラフィカルモデルである．BN 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 [1].

一般に BN の DAG 構造はデータから推定する必要があり，この問題を BN の構造学習と呼ぶ．構造学習では，候補構造から最適な学習スコアをもつ構造を探索するスコアベースアプローチが従来から行われてきた．本論では候補構造に制約を課さずに学習した BNC を General Bayesian Network (GBN) と呼ぶ．一般にスコアベースアプローチでは，パラメータ数最小 I-map への漸近一致性を有する，構造の周辺ゆう度 (Marginal Likelihood: ML) を学習スコアとして用いる．

ML を用いると，全変数の同時確率分布をモデル化する生成モデルとして BNC を学習できる．しかし，Friedman ら [1] は，BNC の構造学習スコアとして，生成モデルではなく，説明変数を所与とした目的変数の条件付き確率分布をモデル化する識別モデルのためのスコアを用いるべきだと主張した．そのような学習スコアとして，説明変数を所与とした目的変数の条件付き対数ゆう度 (Conditional Log Likelihood: CLL) が提案された．しかし，CLL を最大にするパラメータ推定式は閉形式で表せないため，構造の探索に効率的なアルゴリズムを用いることができず，学習時間が膨大になってしまう．これを解決するため，Carvalho ら [2], [3] は構造探索も効率にできるよう CLL を線形近似した approximated CLL (aCLL) を提案した．Grossman ら [4] は CLL をスコアとして，貪欲法の Hill-Climbing アルゴリズム [5] を用いて構造を探索する手法を提案した．Mihaljevic ら [6] は分類確率が等価な構造を探索空間から削減することで，CLL を用いた貪欲法の計算時間を短縮した．これらの近似手法で学習した BNC の方が，ML で学習した BNC よりも分類精度が高いことが報告されている．

しかし，ML 最大化より CLL 最大化の方がなぜ良い

† 電気通信大学大学院情報理工学研究科，調布市
Graduate School of Informatics and Engineering, The University of Electro-Communications, 1–5–1 Chofugaoka, Chofu-shi, 182–8585 Japan
a) E-mail: sugahara@ai.lab.uec.ac.jp
b) E-mail: ueno@ai.lab.uec.ac.jp
DOI:10.14923/transinfj.2022JDP7006

かという理由については未だ明らかにされていない．ML は推定構造に対してパラメータ数最小 I-map への漸近一致性が保証されており，サンプルサイズが大きいときに ML の分類精度が CLL に劣るのは奇異である．また，BNC の ML は閉形式で表せるため CLL より計算効率が良く，ML を大域的に最大化する構造を探索する厳密学習を効率的に行える．先行研究の比較実験では，ML を局所的に最大化する構造を探索する近似学習を行っているため，探索精度の悪さが影響したのかもしれない．近年の研究では ML を用いた厳密学習の効率的なアルゴリズムが多く提案されている [7]〜[15].

菅原ら [16]〜[18] は ML による厳密学習と CLL による近似学習によって得られた BNC の分類精度を比較した．結果として，サンプルサイズが大きいときは，ML による厳密学習は CLL による近似学習より高い分類精度を示すことが報告されている．しかし，サンプルサイズが小さくなると ML による厳密学習の分類精度が低くなり，最も単純な構造をもつ Naive Bayes よりも低い分類精度を示す場合もあった．特に，目的変数の親変数が多く子変数が少ないような構造を学習する場合に分類精度が低くなっていることが報告されている．その理由は，目的変数の親変数が多いと，パラメータ数が指数的に増えるため，一つのパラメータ学習のためのサンプルサイズが小さくなり，推定精度が悪くなってしまうからである．

この問題を緩和するため，菅原ら [16]〜[18] は，目的変数が親変数をもたず，説明変数が必ず目的変数の子となる Augmented Naive Bayes (ANB) 構造を制約とした BNC の厳密学習を提案した．彼らの手法で学習した構造は，真の従属性を不足なく表現する構造 (independence map: I-map) のうち，全パラメータ数が最小の ANB に漸近的に一致する．更に，彼らの ANB は ML で厳密学習した GBN と漸近的に等しい分類確率を表現することが示されている．意思決定に基づく AI システムの開発では，期待効用と期待損失を計算するために高精度な確率推論が要求されるため，上記の性質は有用である [19]〜[21]．しかし，真のモデルが BN に従っていない場合，ML 最大化は分類に影響するパラメータ数を最小にする保証がない．

この問題を解決するために，本論では，真のモデルが BN に従っていない場合でも，分類に影響するパラメータ数を最小とする I-map に学習構造が漸近的に一致する手法を提案する．提案手法では目的変数が親変数をもたないような構造集合 (NPCDAG: no parents class DAG) を探索空間とする．BN が表現可能な全ての分類確率は，NPCDAG のみで表現できることが証明されている [6]．また，NPCDAG のように目的変数に親変数がない構造は，GBN よりも高い分類精度をもつ傾向がある [16]〜[18].

本論では，ある変数順序のもとで ML を最大化する構造が，その順序に従う構造の中で分類に影響するパラメータ数最小の I-map に漸近的に一致することを証明する．この定理に基づき，提案手法は以下の二つのステップから構成される．第一ステップでは，目的変数から始まる全ての順序について，ML を最大化する構造をそれぞれ求める．第二ステップでは，第一ステップで得られた構造のうち分類に影響するパラメータ数を最小にする構造を探索する．結果として得られた構造は，真のモデルが BN に従うか否かにかかわらず，分類に影響する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 NPCDAG となる．また，提案手法は目的変数の親変数集合を探索しないため，GBN の厳密学習よりも計算時間が短い．更に，ベンチマークデータによる比較実験で，提案手法の分類精度が従来手法よりも有意に高いことを示す．

2. ベイジアンネットワーク

ベイジアンネットワーク (Bayesian network: BN) は，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である．今，離散確率変数集合 $\mathbf{V} = \{X_0, X_1, \cdots, X_i, \cdots, X_n\}$ において，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し，各変数 X_i が値 k をとるとき，$X_i = k$ と書く．また，構造 G における変数 X の親変数集合を $\mathbf{Pa}_X^G$ と表す．G を構成する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} で表すとすると，$\forall i, \mathbf{Pa}_{X_{\pi^i}}^G \subseteq \bigcup_{j=1}^{i-1}\{X_{\pi^j}\}$ が成り立つとき，π を G の変数順序という．更に，θ_{ijk} を $\mathbf{Pa}_{X_i}^G$ が j 番目のパターンをとったとき（$\mathbf{Pa}_{X_i}^G = j$ と書く）に $X_i = k$ となる条件付き確率 $P(X_i = k \mid \mathbf{Pa}_{X_i}^G = j)$ を示すパラメータとし，$\Theta_{ij} = \bigcup_{k=1}^{r_i}\{\theta_{ijk}\}$，$\Theta_i = \bigcup_{j=1}^{q_i}\{\Theta_{ij}\}$，$\Theta = \bigcup_{i=0}^{n}\{\Theta_i\}$ とする．ここで，$q_i = \prod_{l: X_l \in \mathbf{Pa}_{X_i}^G} r_l$ である．BN では，次式のように同時確率分布 $P(X_0, X_1, \cdots, X_n \mid G, \Theta)$ を各変数の条件付き確率パラメータの積に分解して表

せる．

$$P(X_0, X_1, \cdots, X_n \mid G, \Theta) = \prod_{i=0}^{n} P(X_i \mid \mathbf{Pa}^G_{X_i}, \Theta).$$

BN の構造は確率分布の条件付き独立性を有向分離によって表す．有向分離の定義のため，まず合流点を以下で定義する．

［定義 2.1］ 構造 $G = (\mathbf{V}, \mathbf{E})$ に対し，道 ρ 上の三変数 $X, Y, Z \in \mathbf{V}$ について，X と Y が隣接せず，X と Y から Z にエッジが引かれているとき，かつそのときに限り Z を ρ における合流点と呼ぶ．

有向分離は以下のように定義される．

［定義 2.2］ G において $X, Y \in \mathbf{V}$ を結ぶ任意の道 ρ について，変数集合 $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$ が次のいずれかの条件を満たすとき，X と Y は $\mathbf{Z}$ によって有向分離されるという．

（1） ρ における合流点ではない変数 $Z \in \mathbf{Z}$ が ρ 上に存在する．

（2） ρ における合流点 Z が ρ 上に存在し，Z とその子孫は $\mathbf{Z}$ に属さない．

この関係を $Dsep_G(X, Y \mid \mathbf{Z})$ で表す．

真の同時確率分布において X と Y が $\mathbf{Z}$ を所与として条件付き独立であることを $I(X, Y \mid \mathbf{Z})$ で表す．また，I-map を以下で定義する．

［定義 2.3］ ベイジアンネットワーク構造 G が以下を満たすとき，G をインディペンデントマップ (independent map: I-map) という．

$$\forall X, Y \in \mathbf{V}, \forall \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\},$$
$$Dsep_G(X, Y \mid \mathbf{Z}) \Rightarrow I(X, Y \mid \mathbf{Z}).$$

I-map が表現する同時確率分布は漸近的に真の同時確率分布に収束する．

今，サンプルが N 個あり，各サンプルは独立で同一な分布に従うとする．t 番目のサンプルを $\mathbf{d}^t = \langle x^t_0, x^t_1, \cdots, x^t_n \rangle$ と表し，学習データを $D = \langle \mathbf{d}^1, \cdots, \mathbf{d}^t, \cdots, \mathbf{d}^N \rangle$ と表す．D が得られたときのベイジアンネットワーク $\langle G, \Theta \rangle$ のゆう度は以下で表される．

$$\begin{aligned} P(D \mid G, \Theta) &= \prod_{t=1}^{N} P(x^t_0, x^t_1, \cdots, x^t_n \mid G, \Theta) \\ &= \prod_{i=0}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \end{aligned} \tag{1}$$

ここで，$P(x^t_0, x^t_1, \cdots, x^t_n \mid G, \Theta)$ は $P(X_0 = x^t_0, X_1 = x^t_1, \cdots, X_n = x^t_n \mid G, \Theta)$ を表し，N_{ijk} は $X_i = k$ かつ $\mathbf{Pa}^G_{X_i} = j$ となる頻度を表す．式 (1) のゆう度を最大にする θ_{ijk} の最ゆう推定量は以下で表される．

$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{N_{ij}} \tag{2}$$

ここで，$N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ である．一般には，BN のパラメータ推定量として，θ_{ijk} の期待値である Expected A Posteriori (EAP) が用いられる．BN の構造 G に対し，式 (3) のようにパラメータの事前分布にディリクレ分布を仮定すると，式 (4) の事後分布 $p(\Theta_{ij} \mid D, G)$ が得られ，その事後分布から式 (5) のように EAP を求めることができる．

$$p(\Theta_{ij} \mid G) = \frac{\Gamma(\sum_{k=1}^{r_i} N'_{ijk})}{\prod_{k=1}^{r_i} \Gamma(N'_{ijk})} \prod_{k=1}^{r_i} \theta_{ijk}^{N'_{ijk}-1} \tag{3}$$

$$\begin{aligned} &p(\Theta_{ij} \mid D, G) \\ &\quad = \frac{\Gamma(\sum_{k=1}^{r_i}(N'_{ijk} + N_{ijk}))}{\prod_{k=1}^{r_i} \Gamma(N'_{ijk} + N_{ijk})} \prod_{k=1}^{r_i} \theta_{ijk}^{N'_{ijk}+N_{ijk}-1} \end{aligned} \tag{4}$$

$$\begin{aligned} \hat{\theta}_{ijk} &= \int \theta_{ijk} \cdot p(\Theta_{ij} \mid D, G) d\Theta_{ij} \\ &= \frac{N'_{ijk} + N_{ijk}}{N'_{ij} + N_{ij}} \end{aligned} \tag{5}$$

ここで，N'_{ijk} はディリクレ事前分布のハイパーパラメータを表し，$N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$ である．

BN のパラメータを推定するためには，最適な構造をデータから推定する必要がある．この問題を BN の構造学習と呼ぶ．構造学習では，候補構造から最適な学習スコアをもつ構造を探索するスコアベースアプローチが従来から行われてきた．一般に学習スコアとして周辺ゆう度 $P(D \mid G)$ (Marginal Likelihood: ML) が用いられ，ML を最大にする構造を最適解とする．ML を最大にする構造は漸近的に全パラメータ数が最小の I-map に一致する．この性質をパラメータ数最小 I-map への漸近一致性と呼ぶ．パラメータの事前分布がディリクレ分布と仮定すると，ML は次のように閉形式で表される．

$$P(D \mid G) = \prod_{i=0}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \tag{6}$$

式 (6) の ML は Bayesian Dirichlet (BD) と呼ばれる．Heckerman ら [5] は，同一の同時確率分布を表す構造は，それらの ML の値も同一でなければならないというゆう度等価を導入した．そして，ゆう度等価に矛盾しないディリ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ている．

$$N'_{ijk} = N'P(X_i = k, \mathbf{Pa}^G_{X_i} = j \mid G^h)$$

ここで，N' は Equivalent Sample Size (ESS) と呼ばれる事前知識の重みを示す擬似サンプルである．G^h はユーザの仮説構造であり，この構造を所与として N' を N'_{ijk} に分配する．このスコアは，Bayesian Dirichlet equivalent (BDe) と呼ばれる．更に，N' をパラメータ数で除し，$N'_{ijk} = N'/(r_i \cdot q_i)$ 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 ら [5] や Ueno [22]〜[24] は，無情報事前分布を用いた BDeu が最も有用であると報告している．

一方，$-\log$ML の近似である最小記述長 (Minimum Description Length: MDL) [25] は，BN と学習データ D の同時記述長を表す．

$$MDL(D \mid G, \Theta) = \frac{\log N}{2}\sum_{i=0}^{n} q_i(r_i - 1) - \log P(D \mid G, \Theta) \quad (7)$$

MDL を用いた学習では，式 (7) を最小にする構造を最適解とする．式 (7) の第一項は構造の複雑さに対するペナルティ項である．式 (7) の第二項は構造のデータへの当てはまりを反映するフィッティング項を表す対数ゆう度である．

logBDeu と MDL は，スコアとして次の性質を満たす．

$$Score(G) = \sum_{i=0}^{n} Score_i(\mathbf{Pa}^G_{X_i}). \quad (8)$$

ここで，$Score_i(\mathbf{Pa}^G_{X_i})$ は変数 X_i とその親変数集合 $\mathbf{Pa}^G_{X_i}$ のみに依存する関数であり，ローカルスコアと呼ぶ．例えば logBDeu の変数 X_i と親変数集合 $\mathbf{Pa}^G_{X_i}$ についてのローカルスコア $Score_i(\mathbf{Pa}^G_{X_i})$ は以下のように表せる．

$$Score_i(\mathbf{Pa}^G_{X_i}) = \sum_{j=1}^{q_i}\left(\log\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} + \sum_{k=1}^{r_i}\log\frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})}\right) \quad (9)$$

また，式 (8) を満たすスコアを分解可能であると言い，分解可能なスコアを用いると効率的に構造を探索できる [9], [14].

3. ベイジアンネットワーク分類器

3.1 ベイジアンネットワーク分類器による分類

BN 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 [1]．今，$X_1, \cdots, X_n$ を説明変数とし，X_0 を目的変数とした BNC を考える．説明変数のデータ $\langle x_1, \cdots, x_n\rangle$ が与えられたとき，目的変数の推定値 $\hat{c}$ は以下のように得られる．

$$\begin{aligned}
\hat{c} &= \underset{c\in\{1,\cdots,r_0\}}{\arg\max}\ P(c \mid x_1, \cdots, x_n, G, \Theta) \quad (10)\\
&= \underset{c\in\{1,\cdots,r_0\}}{\arg\max}\ \frac{P(c, x_1, \cdots, x_n \mid G, \Theta)}{P(x_1, \cdots, x_n \mid G, \Theta)}\\
&= \underset{c\in\{1,\cdots,r_0\}}{\arg\max}\ P(c, x_1, \cdots, x_n \mid G, \Theta)\\
&= \underset{c\in\{1,\cdots,r_0\}}{\arg\max}\ \prod_{i=0}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\left(\theta_{ijk}\right)^{1_{ijk}}\\
&= \underset{c\in\{1,\cdots,r_0\}}{\arg\max}\ \left[\prod_{j=1}^{q_0}\prod_{k=1}^{r_0}\left(\theta_{0jk}\right)^{1_{0jk}} \times \prod_{i:X_i\in\mathbf{Ch}}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\left(\theta_{ijk}\right)^{1_{ijk}}\right]
\end{aligned}$$

ここで，1_{ijk} は変数列 $\langle c, x_1, \cdots, x_n\rangle$ において $X_i = k$ かつ $\mathbf{Pa}^G_{X_i} = j$ のときに 1 をとり，それ以外のときは 0 をとる変数である．また，**Ch** は目的変数の子変数の集合である．式 (10) の最右辺から分かるように目的変数の分類に影響を及ぼす説明変数は，目的変数の親変数と子変数，及び目的変数と子を共有する変数のみである．これらの変数集合をマルコフブランケットと呼ぶ．

3.2 ベイジアンネットワーク分類器への制約

一般に，BN の構造学習で探索する候補構造はとり得る全ての構造であり，そのような候補構造に対して BDeu や MDL などを最適化して学習される BNC は General Bayesian Network (GBN) と呼ばれる（図 1 の

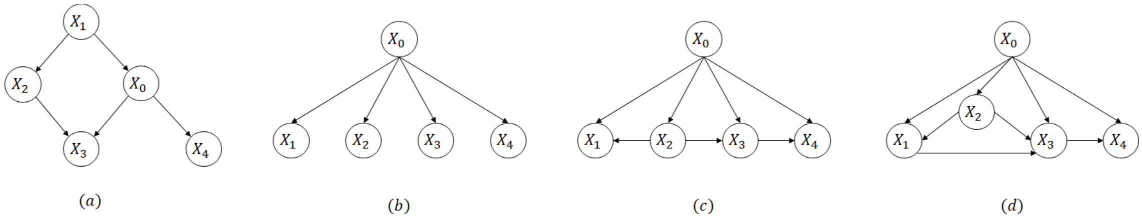

図 1　(a) GBN の例; (b) Naive Bayes; (c) TAN の例; (d) ANB の例
Fig. 1　(a) Example of GBN; (b) Example of Naive Bayes; (c) Example of TAN; (d) Example of ANB.

(a)）．つまり，分類器として用いられる，制約のない一般的な BN を GBN と呼ぶ．大きいネットワークでは GBN の学習に膨大な時間がかかってしまうため，候補構造に制約を入れて学習することが多い．例えば，GBN の下位構造として，全説明変数が目的変数のみを親にもつと仮定する Naive Bayes [26]（図 1 の (b)）や，全説明変数が目的変数を親にもち，説明変数間で木構造をとると仮定した Tree-Augmented Naive Bayes (TAN) [1]（図 1 の (c)）などが知られている．Naive Bayes の構造は一意に定まるため，構造学習の必要はない．ゆう度を学習スコアとした TAN の学習は多項式時間で学習でき，MDL によって近似的に学習した GBN と同等の分類精度をもつことが数値実験により示されている [1], [27]．また，Naive Bayes や TAN を一般化した，より表現力の高い制約として，全説明変数が目的変数の子変数とする制約のみで説明変数間の関係には制約をおかない Augmented Naive Bayes (ANB)（図 1 の (d)）[1] が知られている．

3.3　ベイジアンネットワーク分類器の学習

BDeu や MDL で学習した BNC は，全変数の同時確率分布をモデル化する生成モデルである．しかし，Friedman ら [1] は，BNC の構造学習には，説明変数を所与とした目的変数の条件付き確率分布をモデル化する識別モデルのためのスコアを用いるべきだと主張した．そのようなスコアとして，以下の，説明変数を所与とした目的変数の条件付き対数ゆう度 (Conditional Log Likelihood: CLL) が提案された．

$$\sum_{t=1}^{N} \log P(x_0^t \mid x_1^t, \cdots, x_n^t, G, \Theta)$$

しかし，CLL は分解可能ではないため，効率的な構造探索アルゴリズムを用いることができない．そこで，Grossman ら [4] は近似的な構造探索法として，構造に対しエッジを一つ追加，消去，反転のいずれかの操作を行ったときに最もスコアが良くなるようなエッジを選びその操作を行うというプロセスを繰り返して構造を更新する Hill-Climbing アルゴリズム [5] を用いた．Hill-Climbing アルゴリズムでは，任意のエッジの追加，消去，反転のどの操作を行ってもスコアが改善されないときに更新を終了し，そのときの構造を解とする．Mihaljeciv ら [6] は BN が表現できる分類確率を全て表現可能な構造の最小の集合として，minimal class-focused DAGs (MC-DAG) を提案した．MC-DAG は全てのエッジの終点が目的変数となっているような構造の集合である．更に，彼らは MC-DAG を探索空間として CLL をスコアを用いた貪欲探索法 MCDAGGES を提案した．

しかし，ML 最大化より CLL 最大化の方がなぜ良いかという理由についてはこれまで明らかにされていない．ML は推定構造に対してパラメータ数最小 I-map への漸近一致性が保証されており，サンプルサイズが大きいときに ML の分類精度が CLL に劣るのは奇異である．また，BNC の ML は分解可能であるため，ML による厳密学習は CLL による厳密学習とは異なり，現実的な時間で学習できる．先行研究の比較実験では，ML による近似学習を行っているため，探索精度の悪さが影響したのかもしれない．

菅原ら [16]〜[18] は BDeu による厳密学習と CLL による近似学習によって得られた BNC の分類精度を比較した．結果として，サンプルサイズが大きいときは，BDeu による厳密学習は CLL による近似学習より高い分類精度を示すことが報告されている．しかし，サンプルサイズが小さくなると BDeu による厳密学習の分類精度が低くなり，最も単純な構造をもつ Naive Bayes よりも低い分類精度を示す場合もあった．特に，目的変数の親変数が多く子変数が少ないような構造を学習する場合に分類精度が低くなっていることが報告されている．その理由は，目的変数の親変数が多いと，パラメータ数が指数的に増えるため，一つのパラメー

タ学習のためのサンプルサイズが小さくなり，推定精度が悪くなってしまうからである．

この問題を緩和するため，菅原ら [16]〜[18] は，目的変数が親変数をもたず，説明変数が必ず目的変数の子となる Augmented Naive Bayes (ANB) 構造を制約とした BNC の厳密学習を提案した．彼らの手法は全パラメータ数が最小の I-map ANB を漸近的に学習できる．更に，菅原ら [18] は以下の仮定 3.1 と 3.2 の下で，漸近的に ANB の厳密学習が真の構造と分類等価であることを示した．

［定義 3.1］ [28]
二つの構造 G, G' が，全説明変数に対する任意のインスタンス $\mathbf{d} = (x_1, \cdots, x_n)$ について $P(X_0 \mid \mathbf{d}, G) = P(X_0 \mid \mathbf{d}, G')$ となるとき，G と G' は分類等価という．

［仮定 3.1］ 以下を満たす構造 G^* が存在する．

$$\forall X, Y \in \mathbf{V}, \forall \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\},$$
$$Dsep_{G^*}(X, Y \mid \mathbf{Z}) \Leftrightarrow I(X, Y \mid \mathbf{Z}).$$

［仮定 3.2］ $\forall X \in \mathbf{V}$ について，X と X_0 は G^* において隣接する．

［定理 3.1］ 仮定 3.1 と 3.2 の下で，BDeu を用いて厳密学習された ANB は G^* と漸近的に分類等価である．

4. 分類に影響するパラメータ数最小化による BNC 学習

Sugahara ら [16]〜[18] の提案手法によって学習される I-map は，仮定 3.1 が成り立たない場合，すなわち真のモデルが BN に従っていない場合，以下で定義される分類に影響するパラメータの数 (the number of the class variable parameters: NCP) を最小化する保証がない．

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

ここで，$i = 0 \vee X_0 \in \mathbf{Pa}^G_{X_i}$ のとき $NCP_i(\mathbf{Pa}^G_{X_i}) = (r_i - 1)q_i$ であり，それ以外のとき $NCP_i(\mathbf{Pa}^G_{X_i}) = 0$ である．

本論では，真のモデルが BN に従っていない場合でも，学習構造が NCP 最小の I-map に漸近的に一致する手法を提案する．提案手法では目的変数が親変数をもたないような構造集合 (NPCDAG: no parents class DAG) を探索空間とする．BN が表現可能な全ての分類確率は，NPCDAG のみで表現できることが証明されている [6]．また，NPCDAG のように目的変数に親変数がない構造は，GBN よりも高い分類精度をもつ傾向がある [16], [18]．

提案手法を導出するため，本論は以下で NCP 最小化に関する定理を証明する．今，変数集合 $\mathbf{V}$ からなる全ての変数順序集合を $\boldsymbol{\pi}(\mathbf{V})$ とすると，次の定理が成り立つ．

［定理 4.1］ $\forall \pi \in \boldsymbol{\pi}(\mathbf{V})$ について，π を所与として BDeu を最大化する構造は，π に従う I-map の中で NCP が最小の構造に漸近的に一致する．

証明は付録に記した．この定理に基づき，提案手法は以下の二つのステップから構成される．第一ステップでは，目的変数から始まる全ての変数順序について，BDeu を最大化する構造をそれぞれ求める．ここで，各変数順序を所与として得られた構造の NCP は異なっており，この中でその値を最小とする構造が求める NPCDAG となる．したがって，第二ステップでは，第一ステップで得られた構造のうち NCP 最小の構造を探索する．

提案手法の学習アルゴリズムの説明のため，用語の定義や表記を以下で定める．変数順序 π において変数 X に先行する変数の集合を $\mathbf{Pre}^{\pi}_X$ で表す．例えば，$\mathbf{Pre}^{\pi}_{X_{\pi^i}} = \bigcup_{j=1}^{i-1}\{X_{\pi^j}\}$ である．変数順序 π に従う構造の中で BDeu を最大にする構造を G^*_π で表す．変数 X_i と変数集合 $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X_i\}$ について，X_i の最適親変数集合を以下で定義する．

$$g^*_i(\mathbf{Z}) = \mathop{\arg\max}_{\mathbf{W} \subseteq \mathbf{Z}} Score_i(\mathbf{W}).$$

また，この章では構造 G を各変数の親変数集合からなるベクトル $(\mathbf{Pa}^G_{X_0}, \mathbf{Pa}^G_{X_1}, \cdots, \mathbf{Pa}^G_{X_n})$ で表す．変数集合 $\mathbf{Z}$ に対する変数順序のうち $X_{\pi^1} = X_0$ となるものの集合を $\boldsymbol{\pi}_0(\mathbf{Z})$ で表す．変数集合 $\mathbf{Z}$ で構成され，$\boldsymbol{\pi}_0(\mathbf{Z})$ に属する変数順序に従う全ての構造の中で BDeu を最大化する構造を $G^*(\mathbf{Z})$ で表す．ある変数が子変数をもたないとき，その変数をシンクと呼ぶ．

提案手法の第一ステップは，$\forall \pi_0 \in \boldsymbol{\pi}_0(\mathbf{V})$ について $G^*_{\pi_0}$ を求めることである．$G^*_{\pi_0} = (\emptyset, g^*_1(\mathbf{Pre}^{\pi_0}_{X_1}), \cdots, g^*_n(\mathbf{Pre}^{\pi_0}_{X_n}))$ と表せるため，$G^*_{\pi_0}$ を得るには $\forall \pi_0 \in \boldsymbol{\pi}_0(\mathbf{V}), \forall i \in \{1, \cdots, n\}$ について $g^*_i(\mathbf{Pre}^{\pi_0}_{X_i})$ を求めればよい．しかし，異なる二つの変数順序 π と π' について，$\mathbf{Pre}^{\pi}_X = \mathbf{Pre}^{\pi'}_X$ のとき，π の下での X の最適親変数集合は π' の下での X の最適親変数集合と等しい．この重複探索を避けるには，

$\forall i \in \{1, \cdots, n\}, \forall \mathbf{Z} \subseteq \mathbf{V} \setminus \{X_i\}$ に対する $g_i^*(\mathbf{Z})$ の計算のみ行えばよい．提案手法では，これらの最適親変数集合の探索アルゴリズムとして Silander ら [9] によって提案された動的計画法を用いる．

第二ステップでは，第一ステップで得られた構造の中で NCP を最小にする構造，すなわち $G^*(\mathbf{V})$ を探索する．この探索も Silander ら [9] の動的計画法を用いる．$G^*(\mathbf{Z})$ において，シンクが X_i のときその親変数集合は $g_i^*(\mathbf{Z} \setminus \{X_i\})$ である．$g_i^*(\mathbf{Z} \setminus \{X_i\})$ から X_i に引かれるエッジと変数 X_i を構造 $G^*(\mathbf{Z})$ から取り除いたものは $G^*(\mathbf{Z} \setminus \{X_i\})$ となる．したがって，$G^*(\mathbf{Z})$ におけるシンクを $X_s^*(\mathbf{Z})$ で表すと，以下が成り立つ．

$$X_s^*(\mathbf{Z}) = \arg\min_{X_i \in \mathbf{Z}} \{NCP_i(g_i^*(\mathbf{Z} \setminus \{X_i\})) + NCP(G^*(\mathbf{Z} \setminus \{X_i\}))\}. \quad (11)$$

$G^*(\mathbf{V})$ に対してこの分解を再帰的に行うと，最終的にシンクとその親変数集合を対とする $n+1$ 組 $(X_0, \emptyset), (X_1, g_1^*), \cdots, (X_n, g_n^*)$ が求まる．したがって，$G^*(\mathbf{V}) = (\emptyset, g_1^*, \cdots, g_n^*)$ が得られる．提案手法では目的変数の親変数集合を探索しないため，提案手法の計算時間は GBN の厳密学習の計算時間よりも短い．

本論の主定理は以下である．

［定理 4.2］ 提案手法の学習構造は全ての NPCDAG の中で NCP 最小の I-map に漸近的に一致する．

証明は付録に記した．BDeu 最大化とは異なり，提案手法は真のモデルが BN に従っているか否かによらず NCP を最小化することを保証する．したがって，提案手法は BDeu を用いた厳密学習よりも高精度な分類確率の推定が期待できる．

5. 評価実験

この章では提案手法の利点を示すための実験を行う．

5.1 実データを用いた分類精度比較

まず，実データを用いて以下の 10 種類の手法と提案手法の分類精度を比較する．

- *GBN-BDeu*: BDeu を用いて厳密学習した GBN
- *Naive Bayes*
- *GBN-CMDL*（Grossman ら [4]）: MDL のフィッティング項を CLL に置き換えた Conditional MDL (CMDL) を用いて近似学習した GBN
- *BNC2P*（Grossman ら [4]）: 各変数が最大二つまでしか親をもたない構造を候補として，CLL を用いて近似学習した BNC
- *TAN-aCLL*（Carvalho ら [3]）: aCLL を用いて厳密学習した TAN
- *gGBN-BDeu*: BDeu を用いて近似学習した GBN
- *MC-DAGGES*（Mihaljevic ら [6]）: MC-DAG を探索空間とした CLL スコアを用いた貪欲法
- *ANB-BDeu*（Sugahara ら [16]〜[18]）: BDeu を用いて厳密学習した ANB
- *fsANB-BDeu*（Sugahara ら [18]）: ベイズファクターによる変数選択後に BDeu を用いて厳密学習した ANB
- *NPCDAG-MNCP*: 提案手法の第二ステップにおいて NCP を最大化する構造を探索するように変更した手法

これら全ての手法は java で実装し，2.2 GHz の 10 コアプロセッサ (XEON) と 128 GB のメモリを搭載した PC で実験を行った．UCI レポジトリデータベース [29] の 24 個のベンチマークデータセットを用いた．菅原ら [16]〜[18] と同様に，各データセットに含まれる連続量はいずれも中央値を境に 2 値に離散化し，欠損値を含むサンプルはデータセットから除去した．いずれの手法においても，構造学習後の BNC のパラメータは全て EAP で推定した．*GBN-BDeu*，*ANB-BDeu*，*fsANB-BDeu* と提案手法の ESS については，10 分割交差検証を用いて $\{1, 10, 100, 1000\}$ から定めた．

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，分類精度として表 1 に示した．表 1 のデータセットは，全変数がとり得る値のパターン数でサンプルサイズを割ったもの (sample per pattern: SPP) で昇順に上から並んでいる．提案手法とその他の手法との優位性を示すため，分類精度の多重検定手法として標準的に用いられる Hommel の多重検定 [30], [31] を行った．検定の p 値を表 1 の最下部に示した．表 2 は *GBN-BDeu*，*ANB-BDeu*，*fsANB-BDeu* と提案手法で学習した構造の平均 NCP をそれぞれ示している．更に，表 3 は *GBN-BDeu*，*fsANB-BDeu* と提案手法の平均実行時間を示している．

表 1 より，提案手法は *fsANB-BDeu* を除く全ての手法に対して，有意水準 10% のもとで有意に分類精度が高かった．SPP の大きい 19 番や 20 番のデータセットでは *Naive Bayes*，*TAN-aCLL*，*BNC2P* は提案手法よりも分類精度が低い．これらの手法は親変数数に制限を設けており，親変数数の上限が小さいと表現力が低下してしまう [32]．SPP の大きい 20 番や 23 番のデー

表 1　各手法の分類精度
Table 1　Classification accuracies of the respective methods.

No.	Dataset	Variables	Sample size	SPP	Naive-Bayes	GBN-CMDL	BNC2P	TAN-aCLL	gGBN-BDeu	MC-DAG GES	GBN-BDeu	ANB-BDeu	fsANB-BDeu	NPCDAG -MNCP	Proposed method
1	Lymphography	19	148	1.63×10^{-7}	0.8446	0.7939	0.7973	0.8311	0.7905	0.8041	0.7500	0.8108	0.7905	0.6014	0.8108
2	Breast Cancer Wisconsin	10	683	3.42×10^{-7}	0.9751	0.8917	0.9473	0.9488	0.7094	0.9780	0.9751	0.9751	0.9751	0.8389	0.9751
3	Hepatitis	20	80	7.63×10^{-5}	0.8500	0.7375	0.8875	0.8750	0.8500	0.8875	0.6125	0.5750	0.8500	0.7875	0.7875
4	Zoo	17	101	1.03×10^{-4}	0.9802	0.9109	0.9505	1.0000	0.9505	0.9802	0.9228	0.9406	0.9406	0.9802	0.9307
5	Australian	15	690	2.97×10^{-4}	0.8290	0.8312	0.8348	0.8464	0.8420	0.8406	0.8507	0.8203	0.8594	0.8594	0.8493
6	Vehicle	19	846	8.07×10^{-4}	0.4350	0.5910	0.5910	0.5816	0.5461	0.5414	0.5898	0.6217	0.6135	0.6123	0.6241
7	Breast Cancer	10	277	8.33×10^{-4}	0.7401	0.6209	0.6823	0.7184	0.7058	0.6354	0.7256	0.6968	0.7437	0.6390	0.7473
8	Image Segmentation	19	2310	1.26×10^{-3}	0.7290	0.7918	0.7991	0.7407	0.8026	0.7476	0.8255	0.8273	0.8290	0.8338	0.8294
9	Congressional Voting Records	17	232	1.77×10^{-3}	0.9095	0.9698	0.9612	0.9181	0.9741	0.9009	0.9655	0.9483	0.9353	0.8491	0.9655
10	Heart	14	270	2.44×10^{-3}	0.8259	0.8185	0.8037	0.8148	0.8222	0.8333	0.8370	0.8037	0.8185	0.7222	0.8333
11	Solar Flare	11	1389	3.72×10^{-3}	0.7811	0.8265	0.8315	0.8229	0.8431	0.8013	0.8431	0.8215	0.8387	0.8195	0.8431
12	Wine	14	178	7.24×10^{-3}	0.9270	0.9438	0.9157	0.9326	0.9045	0.9438	0.9270	0.9270	0.9101	0.9326	0.9438
13	Letter	17	20000	1.17×10^{-2}	0.4466	0.5796	0.5132	0.5093	0.5761	0.4664	0.6434	0.6434	0.6434	0.6418	0.6434
14	Pendigits	17	10992	1.68×10^{-2}	0.8032	0.9062	0.8719	0.8700	0.9253	0.8359	0.9342	0.9332	0.9317	0.9326	0.9343
15	Contraceptive Method Choice	10	1473	5.99×10^{-2}	0.4671	0.4501	0.4745	0.4705	0.4440	0.4576	0.4792	0.4481	0.4610	0.4498	0.4413
16	Glass	10	214	6.97×10^{-2}	0.5561	0.5654	0.5794	0.6308	0.4626	0.5888	0.5888	0.6355	0.5911	0.6493	0.5888
17	Hayes–Roth	5	132	2.29×10^{-1}	0.8182	0.6136	0.6894	0.6742	0.7525	0.6970	0.6212	0.7879	0.7652	0.8333	0.8333
18	Balance Scale	5	625	3.33×10^{-1}	0.9152	0.3333	0.8560	0.8656	0.9152	0.7432	0.9152	0.9152	0.9152	0.9152	0.9152
19	Lenses	5	24	3.33×10^{-1}	0.7500	0.8333	0.6667	0.7083	0.8333	0.8333	0.8333	0.7500	0.8750	0.8750	0.8750
20	EEG	15	14980	4.57×10^{-1}	0.5778	0.6787	0.6374	0.6125	0.6732	0.6182	0.7246	0.7212	0.7212	0.7178	0.7165
21	LED7	8	3200	2.50×10^{0}	0.7294	0.7366	0.7375	0.7350	0.7297	0.7331	0.7303	0.7303	0.7288	0.7294	0.7316
22	Iris	5	150	3.13×10^{0}	0.7133	0.7800	0.8200	0.8200	0.8133	0.7800	0.8267	0.8156	0.8156	0.8200	0.8267
23	HTRU2	9	17898	3.50×10^{1}	0.8966	0.9086	0.9118	0.9130	0.9092	0.9093	0.9141	0.9141	0.9141	0.9133	0.9140
24	Banknote authentication	5	1372	4.29×10^{1}	0.8433	0.8819	0.8797	0.8761	0.8819	0.8768	0.8812	0.8812	0.8812	0.8812	0.8819
	average				0.7643	0.7498	0.7766	0.7798	0.7774	0.7681	0.7882	0.7893	0.8062	0.7848	0.8101
	p-value				0.0080	0.0001	0.0047	0.0300	0.0037	0.0071	0.0466	0.0188	0.1802	0.0800	-

表 2　*GBN-BDeu*，*ANB-BDeu*，*fsANB-BDeu*，*NPCDAG-MNCP* と提案手法によって学習された構造の NCP
Table 2　The number of the class variable parameters (NCP) of the learned structures by *GBN-BDeu*, *ANB-BDeu*, *fsANB-BDeu*, *NPCDAG-MNCP* and the proposed method.

				NCP				
No.	Variables	Sample size	SPP	ANB-BDeu	GBN-BDeu	fsANB-BDeu	NPCDAG -MNCP	Proposed method
1	19	148	1.63×10^{-7}	219126	216535	44388	800975	319
2	10	683	3.42×10^{-7}	179	150	179	12799	159
3	20	80	7.63×10^{-5}	5880	2011	469	21953	5
4	17	101	1.03×10^{-4}	816	4455	449	28479	1385
5	15	690	2.97×10^{-4}	447	65	95	4551	133
6	19	846	8.07×10^{-4}	556	987	700	13931	1404
7	10	277	8.33×10^{-4}	105	20	41	2517	31
8	19	2310	1.26×10^{-3}	2464	3840	2464	20159	4156
9	17	232	1.77×10^{-3}	121	24	107	6391	9
10	14	270	2.44×10^{-3}	58	32	71	3783	25
11	11	1389	3.72×10^{-3}	1570	19	305	9719	8
12	14	178	7.24×10^{-3}	59	36	59	4031	27
13	17	20000	1.17×10^{-2}	18386	18360	18386	30783	12287
14	17	10992	1.68×10^{-2}	11215	11042	11215	21759	8487
15	10	1473	5.99×10^{-2}	196	40	147	1247	33
16	10	214	6.97×10^{-2}	444	354	480	1535	610
17	5	132	2.29×10^{-1}	40	176	45	191	29
18	5	625	3.33×10^{-1}	50	48	50	50	50
19	5	24	3.33×10^{-1}	18	8	9	11	9
20	15	14980	4.57×10^{-1}	2703	2850	2703	3839	1856
21	8	3200	2.5×10^{0}	187	186	85	959	167
22	5	150	3.13×10^{0}	28	18	28	242	19
23	9	17898	3.5×10^{1}	77	69	103	511	201
24	5	1372	4.29×10^{1}	31	25	29	31	15
	Average			11032	10890	3442	41269	1309

表 3　*GBN-BDeu*，*fsANB-BDeu* と提案手法の平均計算時間（秒）
Table 3　Average runtime (s) of *GBN-BDeu*, *fsANB-BDeu*, and the proposed method.

	GBN-BDeu	*fsANB-BDeu*	Proposed method
Average	1690	4375	1498

タセットでは *gGBN-BDeu* と *GBN-CMDL* は提案手法よりも分類精度が低い．厳密学習の方が近似学習よりも高い精度で構造を推定できることがこの理由として考えられる．

NPCDAG-MNCP は提案手法の第二ステップにおいて NCP を最大化する構造を探索するため，表 2 から分かるように，*NPCDAG-MNCP* は全てのデータセットにおいて提案手法より NCP が大きい．特に 1 番，2 番，9 番，10 番のデータセットにおいて *NPCDAG-MNCP* の NCP は提案手法の NCP を大きく上回っており，これらのデータセットでは *NPCDAG-MNCP* は提案手法より著しく低い分類精度を示している．この結果の理由として，*NPCDAG-MNCP* ももつ大きな NCP が過学習を引き起こし，分類確率の推定精度が低下したと考えられる．

SPP の小さな 3 番のデータセットでは，提案手法は *GBN-BDeu* と *ANB-BDeu* よりも著しく高い分類精度を示している．このデータセットでは，提案手法は *GBN-BDeu* と *ANB-BDeu* より NCP が小さいことが表 2 より分かる．このため，提案手法は SPP が小さいときに過学習を防ぎ，結果として 3 番のデータセットにおいて高い分類精度を示したと考えられる．

SPP の大きなデータセットでは，*GBN-BDeu*，*ANB-BDeu*，*fsANB-BDeu* は提案手法とほとんど同じ分類精度を示している．特に，提案手法と *fsANB-BDeu* の分類精度は有意差も認められていない．これらの理由として，表 1 の多くのデータセットで仮定 3.1 と 3.2 を満たしていることが考えられる．次節では，仮定 3.1 と 3.2 が成り立たないときに提案手法が *GBN-BDeu*，

表 4　*ANB-BDeu*，*GBN-BDeu*，*fsANB-BDeu* と提案手法によって学習された構造の NCP と，4 手法が推定した分類確率と真の分類確率間の平均 KLD

Table 4　The NCPs of structures learned by *ANB-BDeu*, *GBN-BDeu*, *fsANB-BDeu*, and the proposed method, and the average KLDs between the true posteriors of the class variable and those of the four methods.

		ANB-BDeu				GBN-BDeu				fsANB-BDeu				Proposed method			
Network	Sample size	aveKLD	NCP	Imin(NP)	MB	aveKLD	NCP	Imin(GBN)	MB	aveKLD	NCP	Imin(NP)	MB	aveKLD	NCP	Imin(NP)	MB
Cancer (Structure (1))	100	5.11×10^{-2}	9	×	4	2.70×10^{-2}	5	×	2	5.87×10^{-2}	7	×	3	2.70×10^{-2}	5	×	2
	1,000	4.11×10^{-2}	9	×	4	2.67×10^{-2}	5	×	2	3.97×10^{-2}	7	×	3	2.67×10^{-2}	5	×	2
	10,000	1.27×10^{-3}	11	○	4	1.27×10^{-3}	8	○	4	1.27×10^{-3}	11	○	4	1.27×10^{-3}	11	○	4
	100,000	8.46×10^{-5}	11	○	4	8.46×10^{-5}	8	○	4	8.46×10^{-5}	11	○	4	8.46×10^{-5}	11	○	4
Asia (Structure (2))	100	8.81×10^{-2}	21	×	7	5.21×10^{-2}	5	×	2	1.04×10^{-1}	13	×	5	4.40×10^{-2}	3	×	1
	1,000	3.70×10^{-2}	21	×	7	3.40×10^{-2}	9	×	3	4.89×10^{-2}	7	×	3	6.38×10^{-2}	7	×	2
	10,000	2.58×10^{-2}	21	×	7	3.60×10^{-3}	10	×	5	2.88×10^{-2}	15	×	5	2.46×10^{-2}	11	×	4
	100,000	1.94×10^{-3}	25	×	7	2.72×10^{-4}	10	○	5	1.32×10^{-2}	17	×	5	2.72×10^{-4}	13	○	5
Markov net (Structure (3))	100	2.20×10^{-1}	29	×	3	2.08×10^{-1}	3	×	1	2.20×10^{-1}	29	×	3	8.18×10^{-2}	5	×	1
	1,000	6.47×10^{-2}	29	×	3	1.38×10^{-2}	17	×	2	6.47×10^{-2}	29	×	3	6.63×10^{-2}	5	×	1
	10,000	2.96×10^{-3}	29	×	3	2.96×10^{-3}	27	×	3	2.96×10^{-3}	29	×	3	4.43×10^{-4}	17	○	2
	100,000	1.20×10^{-3}	29	×	3	1.20×10^{-3}	29	×	3	1.20×10^{-3}	29	×	3	7.94×10^{-5}	17	○	2

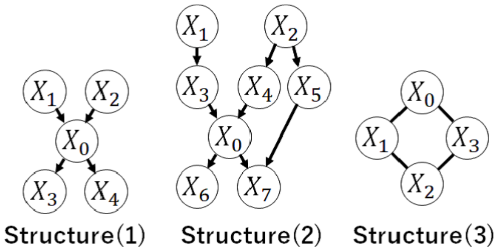

図 2　構造：(1) Cancer ネットワーク，(2) Asia ネットワーク，(3) 閉路をもつマルコフネットワーク

Fig. 2　Structures: (1) the Cancer network, (2) the Asia network, and (3) a Markov network with a cycle.

ANB-BDeu，*fsANB-BDeu* よりも高精度に分類確率を推定できることをシミュレーション実験で示す．

表 3 より，提案手法の計算時間は *GBN-BDeu* の計算時間より短い．これは，提案手法では目的変数の親変数集合を探索しないためである．また，提案手法の計算時間は *fsANB-BDeu* の計算時間より短い．*fsANB-BDeu* では変数選択手法におけるハイパーパラメータの最適化のために，厳密学習を複数回実行しなければならないため，その分だけ計算時間が長くなっていると考えられる．

5.2　シミュレーションデータを用いた分類確率推定精度の比較

この実験では，*GBN-BDeu*，*ANB-BDeu*，*fsANB-BDeu* と提案手法によって推定された分類確率と真の分類確率の Kullback–Leibler divergence (KLD) をそれぞれ比較する．仮定 3.1 と 3.2 を満たす Cancer ネットワーク [33]（図 2-(1)），仮定 3.1 を満たし仮定 3.2 を満たさない Asia ネットワーク [33]（図 2-(2)），仮定 3.1 と 3.2 を両方満たさないマルコフネットワーク（図 2-(3)）をそれぞれ真のモデルとした実験を行う．

図 2 の三つのネットワークからサンプルサイズが百，千，一万，十万のデータセットをそれぞれ発生させ，各データセットに対して，*GBN-BDeu*，*ANB-BDeu*，*fsANB-BDeu* と提案手法の四つの手法で構造を学習する．表 4 は学習した構造の NCP と，説明変数がとり得る全てのパターンについて，推定分類確率と真の分類確率間の KLD の平均をとったものを示している．表 4 における"Imin(GBN)"と"Imin(NP)"は，学習構造が GBN の中で NCP を最小にする I-map であるか否か，NPCDAG の中で NCP を最小にする I-map であるか否かをそれぞれ示している．また，表 4 における"MB"は各学習構造における目的変数のマルコフブランケットの大きさを示している．

表 4 より，Cancer ネットワークでは，$N \geq 10000$ のときに 4 手法の KLD はほとんど同じ値をとっている．Cancer ネットワークは仮定 3.1 と 3.2 を満たすため，4 手法とも NCP 最小の I-map を学習できている．

同様に，Asia ネットワークにおいて，$N \geq 10000$ のときに提案手法と *GBN-BDeu* は NCP 最小の I-map を学習できており，それぞれの KLD はほとんど同じ値をとっている．しかし，*ANB-BDeu* と *fsANB-BDeu* は，提案手法と *GBN-BDeu* よりも大きな NCP と大きな KLD を示している．Asia ネットワークのように仮定 3.2 を満たさないモデルでは，*ANB-BDeu* と *fsANB-BDeu* の学習構造は NCP 最小の I-map に漸近的に一致することを保証しない．

図 2 の (3) のマルコフネットワークでは，$N \geq 10000$ のときに提案手法の方が *GBN-BDeu* よりも NCP が小さい．このとき，提案手法は NCP 最小の I-map を学習

できているが，*GBN-BDeu* は学習できていない．このように，仮定 3.1 を満たさないモデルでは，*GBN-BDeu* が NCP を最小化する保証がない．一方で，提案手法は定理 4.1 で示したように，仮定 3.1 が成り立たない場合でも NCP の最小化を保証する．更に，表 4 の"MB"から分かるように，$N \geq 10000$ のときに提案手法は *GBN-BDeu* よりも目的変数のマルコフブランケットが小さい．

結果として，この章では以下の提案手法の利点を実験により示した．(1) *fsANB-BDeu* を除く全ての手法に対して，有意水準 10% のもとで有意に分類精度が高い．(2) 仮定 3.1 と 3.2 を満たさないようなモデルに対しても学習構造が NCP 最小の I-map NPCDAG に漸近的に一致することを保証する．(3) 小サンプルのとき，過学習を防ぐ．(4) *GBN-BDeu* と *fsANB-BDeu* よりも計算時間が短い．

6. むすび

本論文では，学習構造が NCP 最小の I-map NPCDAG に漸近的に一致する手法を提案した．真のモデルが BN に従わないとき，BDeu 最大化では NCP の最小化を保証できなかった．一方で，提案手法は真のモデルが BN に従うか否かによらず，NCP を最小化して真の分類確率に漸近収束する．このため，提案手法は BDeu を用いた厳密学習よりも高精度に分類確率を推定できる．更に，提案手法は目的変数の親変数集合を探索しないため，GBN の厳密学習より計算時間が短い．今後の課題として，提案手法にモデル平均手法 [34]～[42] を適用する．

文　献

[1] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” Mach. Learn., vol.29, no.2, pp.131–163, 1997.

[2] A.M. Carvalho, T. Roos, A.L. Oliveira, and P. Myllymäki, “Discriminative learning of bayesian networks via factorized conditional log-likelihood,” J. Mach. Learn. Res., vol.12, pp.2181–2210, 2011.

[3] 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, vol.15, no.7, pp.2716–2735, 2013.

[4] D. Grossman and P. Domingos, “Learning Bayesian Network classifiers by maximizing conditional likelihood,” Proc. 21st Int. Conf. Mach. Learn., ICML 2004, pp.361–368, 2004.

[5] D. Heckerman, D. Geiger, and D.M. Chickering, “Learning Bayesian networks: The combination of knowledge and statistical data,” Mach. Learn., vol.20, no.3, pp.197–243, 1995.

[6] B. Mihaljević, C. Bielza, and P. Larrañaga, “Learning Bayesian network classifiers with completed partially directed acyclic graphs,” Proc. 9th Int. Conf. Probabilistic Graphical Models, vol.72, pp.272–283, Proc. Mach. Learn. Res., 2018.

[7] M. Koivisto and K. Sood, “Exact Bayesian structure discovery in Bayesian networks,” J. Mach. Learn. Res., vol.5, pp.549–573, 2004.

[8] A.P. Singh and A.W. Moore, “Finding optimal Bayesian networks by dynamic programming,” Technical Report, Carnegie Mellon University, 2005.

[9] T. Silander and P. Myllymäki, “A simple approach for finding the globally optimal Bayesian network structure,” Proc. Uncertainty in Artificial Intelligence, pp.445–452, 2006.

[10] C.P. deCampos and Q. Ji, “Efficient structure learning of Bayesian networks using constraints,” J. Mach. Learn. Res., vol.12, no.12, pp.663–689, 2011.

[11] B.M. Malone, C. Yuan, E.A. Hansen, and S. Bridges, “Improving the scalability of optimal Bayesian network learning with external-memory frontier breadth-first branch and bound search,” Proc. Uncertainty in Artificial Intelligence, p.479–488, 2011.

[12] C. Yuan and B. Malone, “Learning optimal bayesian networks: A shortest path perspective,” J. Artificial Intelligence Research, vol.48, no.1, pp.23–65, 2013.

[13] J. Cussens, “Bayesian network learning with cutting planes,” Proc. 27th Conf. Uncertainty in Artificial Intelligence, pp.153–160, 2012.

[14] M. Barlett and J. Cussens, “Advances in Bayesian Network Learning Using Integer Programming,” Proc. 29th Conf. Uncertainty in Artificial Intelligence, pp.182–191, 2013.

[15] J. Suzuki, “A theoretical analysis of the BDeu scores in Bayesian network structure learning,” Behaviormetrika, vol.44, no.1, pp.97–116, 2017.

[16] S. Sugahara, M. Uto, and M. Ueno, “Exact learning augmented naive Bayes classifier,” Proc. 9th Int. Conf. Probabilistic Graphical Models, vol.72, pp.439–450, Proc. Mach. Learn. Res., PMLR, 2018.

[17] 菅原聖太，植野真臣，“Augmented naive bayes 制約を持つベイジアンネットワーク分類器の厳密学習，” 信学論（D），vol.J103-D, no.4, pp.301–313, 2020.

[18] S. Sugahara and M. Ueno, “Exact learning augmented naive bayes classifier,” Entropy, vol.23, no.12, 2021.

[19] F.V. Jensen and T.D. Nielsen, Bayesian Networks and Decision Graphs, 2nd edition, Springer Publishing Company, Incorporated, 2007.

[20] A. Darwiche, Modeling and Reasoning with Bayesian Networks, Cambridge University Press, 2009.

[21] A. Darwiche, “Three Modern Roles for Logic in AI,” Proc. 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, pp.229–243, 2020.

[22] M. Ueno, “Learning likelihood-equivalence Bayesian networks using an empirical Bayesian approach,” Behaviormetrika, vol.35, no.2, pp.115–135, 2007.

[23] M. Ueno, “Learning networks determined by the ratio of prior and data,” Proc. Uncertainty in Artificial Intelligence, pp.598–605, 2010.

[24] M. Ueno, “Robust learning Bayesian networks for prior belief,”

Proc. Uncertainty in Artificial Intelligence, pp.689–707, 2011.
[25] J. Rissanen, Stochastic Complexity in Statistical Inquiry Theory, World Scientific Publishing, 1989.
[26] M. Minsky, "Steps toward Artificial Intelligence," Proc. IRE, vol.49, pp.8–30, 1961.
[27] M.G. Madden, "On the classification performance of TAN and general Bayesian networks," Knowledge-Based Systems, pp.489–495, 2009.
[28] S. Acid, L.M. deCampos, and J.G. Castellano, "Learning Bayesian network classifiers: Searching in a space of partially directed acyclic graphs," Mach. Learn., vol.59, no.3, pp.213–235, 2005.
[29] M. Lichman, "UCI machine learning repository," 2013. http://archive.ics.uci.edu/ml
[30] G. Hommel, "A stagewise rejective multiple test procedure based on a modified bonferroni test," Biometrika, pp.383–386, 1988.
[31] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," J. Mach. Learn. Res., vol.7, pp.1–30, 2006.
[32] C.X. Ling and H. Zhang, "The representational power of discrete Bayesian networks," J. Mach. Learn. Res., vol.3, pp.709–721, 2003.
[33] M. Scutari, "Learning Bayesian networks with the bnlearn R package," J. Statistical Software, Articles, vol.35, no.3, pp.1–22, 2010.
[34] J. Tian, R. He, and L. Ram, "Bayesian model averaging using the K-best Bayesian network structures," Proc. 26th Conf. Uncertainty in Artificial Intelligence, pp.589–597, UAI'10, 2010.
[35] Y. Chen and J. Tian, "Finding the k-best equivalence classes of bayesian network structures for model averaging," Proc. National Conf. Artificial Intelligence, vol.4, pp.2431–2438, Jan. 2014.
[36] R. He, J. Tian, and H. Wu, "Structure learning in Bayesian networks of a moderate size by efficient sampling," J. Mach. Learn. Res., vol.17, no.101, pp.1–54, 2016.
[37] E.Y.-J. Chen, A. Choi, and A. Darwiche, "Learning Bayesian networks with non-decomposable scores," Graph Structures for Knowledge Representation and Reasoning, pp.50–71, 2015.
[38] E.Y.-J. Chen, A.C. Choi, and A. Darwiche, "Enumerating equivalence classes of Bayesian networks using ec graphs," Proc. 19th Int. Conf. Artificial Intelligence and Statistics, vol.51, pp.591–599, 2016.
[39] E.Y.-J. Chen, A. Darwiche, and A. Choi, "On pruning with the MDL score," Int. J. Approximate Reasoning, vol.92, pp.363–375, 2018.
[40] Z. Liao, C. Sharma, J. Cussens, and P. van Beek, "Finding all Bayesian network structures within a factor of optimal," 33rd AAAI Conf. Artificial Intelligence, pp.7892–7899, 2018.
[41] S. Sugahara, I. Aomi, and M. Ueno, "Bayesian network model averaging classifiers by subbagging," Proc. 10th Int. Conf. Probabilistic Graphical Models, vol.138, pp.461–472, Proc. Mach. Learn. Res., PMLR, 2020.
[42] S. Sugahara, I. Aomi, and M. Ueno, "Bayesian network model averaging classifiers by subbagging," Entropy, vol.24, no.5, 2022.
[43] J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, 2000.
[44] D.M. Chickering, "Learning equivalence classes of Bayesian-network structures," J. Mach. Learn. Res., vol.2, pp.445–498, 2002.
[45] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers, San Francisco, CA, USA, 1988.

付　　録

1. 定理 4.1 の証明

定理 4.1 の証明では，以下の定理と定義を用いる．

［定理］（BN における局所独立性）[43]

BN 構造 $G = (\mathbf{V}, \mathbf{E})$ について，変数 X の子孫でない変数の集合を $\mathbf{ND}_G(X)$ で表すと，以下が成り立つ．

$$\forall X \in \mathbf{V}, Dsep_G(X, (\mathbf{ND}_G(X) \setminus \mathbf{Pa}_X^G) \mid \mathbf{Pa}_X^G).$$

［定義］（パラメータ数最小 I-map への漸近一致性）[44]

二つの構造 $G_1 = (\mathbf{V}, \mathbf{E}_1)$ と $G_2 = (\mathbf{V}, \mathbf{E}_2)$ について，あるスコア関数 $Score$ が漸近的に以下の二つを満たすとき，$Score$ は漸近一致性をもつという．

- G_1 が I-map で G_2 が I-map でないとき，$Score(G_1) > Score(G_2)$.
- G_1 と G_2 が I-map であり，G_1 のもつパラメータ数が G_2 のもつパラメータ数より少ないとき，$Score(G_1) > Score(G_2)$.

［定義］（スコアの局所一致性）[44]

変数 X, Y について，エッジ $Y \to X$ をもたない構造を $G_1 = (\mathbf{V}, \mathbf{E}_1)$ とし，$G_1 = (\mathbf{V}, \mathbf{E}_1)$ にエッジ $Y \to X$ を加えたものを G_2 とする．あるスコア関数 $Score$ が漸近的に以下の二つを満たすとき，$Score$ は局所一致性をもつという．

- $I(X, Y \mid \mathbf{Pa}_X^{G_1}) \Rightarrow Score(G_1) > Score(G_2)$.
- $\neg I(X, Y \mid \mathbf{Pa}_X^{G_1}) \Rightarrow Score(G_1) < Score(G_2)$.

BDeu スコアはパラメータ数最小 I-map への漸近一致性と局所一致性をもつことが知られている [44]．また，定理 4.1 の証明のため，以下の補題を証明する．

［補題］ $\mathbf{X}, \mathbf{Y}, \mathbf{A}, \mathbf{B}$ をそれぞれ互いに素な変数集合とすると，以下が成り立つ．

$$\neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A}) \Rightarrow \neg I(\mathbf{X}, \mathbf{B} \mid \mathbf{A} \cup \mathbf{Y}) \vee \neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A} \cup \mathbf{B}).$$

［証明］ 条件付き独立の分解性 [45] より，$I(\mathbf{X}, (\mathbf{Y} \cup \mathbf{B}) \mid \mathbf{A}) \Rightarrow I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A}) \wedge I(\mathbf{X}, \mathbf{B} \mid \mathbf{A})$ が成り立つ．対偶をとると，$\neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A}) \vee \neg I(\mathbf{X}, \mathbf{B} \mid \mathbf{A}) \Rightarrow \neg I(\mathbf{X}, (\mathbf{Y} \cup \mathbf{B}) \mid \mathbf{A})$. したがって，明らかに以下が成り立つ．

$$\neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A}) \Rightarrow \neg I(\mathbf{X}, (\mathbf{Y} \cup \mathbf{B}) \mid \mathbf{A}). \tag{A·1}$$

また，条件付き独立の交差性 [45] より，$I(\mathbf{X}, \mathbf{B} \mid$

$\mathbf{A} \cup \mathbf{Y}) \wedge I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A} \cup \mathbf{B}) \Rightarrow I(\mathbf{X}, (\mathbf{Y} \cup \mathbf{B}) \mid \mathbf{A})$ が成り立つ．対偶をとると，

$$\neg I(\mathbf{X}, (\mathbf{Y} \cup \mathbf{B}) \mid \mathbf{A}) \Rightarrow \neg I(\mathbf{X}, \mathbf{B} \mid \mathbf{A} \cup \mathbf{Y}) \vee \neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A} \cup \mathbf{B}). \qquad (\mathrm{A}\cdot 2)$$

式 (A·1)，(A·2) より，$\neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A}) \Rightarrow \neg I(\mathbf{X}, \mathbf{B} \mid \mathbf{A} \cup \mathbf{Y}) \vee \neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A} \cup \mathbf{B})$. □

上記の定理，定義，補題を用いて，以下のように定理 4.1 を証明する．

［定理 4.1］ $\forall \pi \in \boldsymbol{\pi}(\mathbf{V})$ について，π を所与として BDeu を最大化する構造は，π に従う I-map の中で NCP が最小の構造に漸近的に一致する．

［証明］ 変数順序 π に従う任意の I-map を $G_\pi = (\mathbf{V}, \mathbf{E}_\pi)$ で表し，変数順序 π に従う構造の中で BDeu を最大にするものを $G^*_\pi = (\mathbf{V}, \mathbf{E}^*_\pi)$ で表す．BDeu のパラメータ数最小 I-map への漸近一致性 [44] より，G^*_π は I-map である．定理 4.1 が成り立つための十分条件は $\mathbf{E}^*_\pi \subseteq \mathbf{E}_\pi$ であるため，これを背理法で示す．変数順序 π に従うある I-map $G'_\pi = (\mathbf{V}, \mathbf{E}'_\pi)$ が存在し，$\mathbf{E}^*_\pi \not\subseteq \mathbf{E}'_\pi$ となると仮定する．この仮定から，$\exists X, Y \in \mathbf{V}, (Y \to X) \in \mathbf{E}^*_\pi \wedge (Y \to X) \notin \mathbf{E}'_\pi$ が成り立つ．$\mathbf{A} = \mathbf{Pa}^{G^*_\pi}_X \setminus \{Y\}$ とすると，$(Y \to X) \in \mathbf{E}^*_\pi$ と BDeu の局所一致性 [44] より，$\neg I(X, Y \mid \mathbf{A})$ が成り立つ．また，$\mathbf{B} = \mathbf{Pre}^\pi_X \setminus \mathbf{Pa}^{G^*_\pi}_X$ とすると，$\neg I(X, Y \mid \mathbf{A})$ と補題より，$\neg I(X, \mathbf{B} \mid \mathbf{A} \cup \{Y\}) \vee \neg I(X, Y \mid \mathbf{A} \cup \mathbf{B})$ が成り立つ．すなわち，$I(X, \mathbf{B} \mid \mathbf{A} \cup \{Y\}) \Rightarrow \neg I(X, Y \mid \mathbf{A} \cup \mathbf{B})$ が成り立つ．G^*_π における局所独立性より $I(X, \mathbf{B} \mid \mathbf{A} \cup \{Y\})$ であるため，以下が成り立つ．

$$\neg I(X, Y \mid \mathbf{A} \cup \mathbf{B}). \qquad (\mathrm{A}\cdot 3)$$

一方で，G'_π において，X と Y は隣接しておらず，$\mathbf{A} \cup \mathbf{B}$ の要素で X と Y の共通の子孫は存在しないため，以下が成り立つ．

$$Dsep_{G'_\pi}(X, Y \mid \mathbf{A} \cup \mathbf{B}). \qquad (\mathrm{A}\cdot 4)$$

式 (A·3) と式 (A·4) が同時に成り立つことは G'_π が I-map であることに矛盾する．ゆえに，定理 4.1 が証明された．□

2. 定理 4.2 の証明

以下のように定理 4.2 を証明する．

［定理 4.2］ 提案手法の学習構造は全ての NPCDAG の中で NCP 最小の I-map に漸近的に一致する．

［証明］ 提案手法の第一ステップでは目的変数から始まる変数順序それぞれに対して，変数順序に従う構造の中で BDeu が最大のものを探索する．定理 4.1 より，サンプルサイズが十分に大きければ，第一ステップで得られた構造の中に，目的変数が親をもたない構造の中で NCP 最小の I-map が存在する．第二ステップでは第一ステップで得られた構造の中で NCP が最小のものを探索する．したがって，提案手法の学習構造は目的変数が親をもたない構造の中で NCP 最小の I-map（すなわち NCP 最小の I-map NPCDAG）に漸近的に一致する．□

（2022 年 1 月 26 日受付，5 月 30 日再受付，
7 月 21 日早期公開）

菅原　聖太

2020 電気通信大学大学院情報理工学研究科博士前期課程了．同年，電気通信大学大学院情報理工学研究科博士後期課程入学，現在に至る．

植野　真臣　（正員）

1992 神戸大学大学院教育学研究科了，1994 東京工業大学大学院総合理工学研究科了．博士（工学）．東京工業大学，千葉大学，長岡技術科学大学を経て 2006 より電気通信大学助教授，2013 より教授，現在に至る．