# 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

菅原聖太

2020 年 2 月 14 日

# 目次

# 表目次



# 図目次

# 1　まえがき

ベイジアンネットワークは，離散確率変数をノードとし，ノード間の条件付き独立性を非循環有向グラフ（Directed Acyclic Graph: DAG）で表し，同時確率分布を各ノードの親ノード集合を所与とした条件付き確率パラメータの積に分解する，確率的グラフィカルモデルである．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 [1].

一般にベイジアンネットワークの DAG 構造はデータから推定する必要があり，この問題をベイジアンネットワークの構造学習と呼ぶ．構造学習では，候補構造から最適な学習スコアを持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．一般にスコアベースアプローチでは，漸近一致性を有する，構造の周辺尤度（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尤度（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 approximate 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．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漸近一致性が保証されており，サンプルサイズが大きい時に 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 より計算効率がよく，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なっているため，探索精度の悪さが影響したのかもしれない．

そこで，本研究ではまず 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 の分類精度を比較する．結果として，ML 最大化による BNC は厳密学習することで，大きく精度が向上することがわかった．特にサンプルサイズが大きいときに，最も分類精度が高いことが示された．しかし，厳密学習ではサンプルサイズが小さくなると ML を最大化する BNC 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わかった．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，本論では，目的変数が親変数を持たず，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 の厳密学習を提案する．さらに，真の構造において説明変数が目的変数に隣接している場合は，提案手法は漸近的に最適な分類を保証すること示す．

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した．

# 2　ベイジアンネットワーク

## 2.1　定義

ベイジアンネットワークは，確率変数をノードとし，ノード間の条件付き独立性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，変数 $X_i$ の親変数集合を $G_i$ とし，ベイジアンネットワークの構造を $G = (G_0, G_1, \cdots, G_n)$ と定義する．さらに，$\theta_{ijk}$ を $G_i$ が $j$ 番目のパターンをとったとき（$G_i = j$ と書く）に $X_i = k$ となる条件付き確率 $P(X_i = k \mid G_i = j)$ を示すパラメータとし，$\Theta_{ij} = \bigcup_{k=1}^{r_i}\{\theta_{ijk}\}$，$\Theta = \bigcup_{i=0}^{n}\bigcup_{j=1}^{q_i}\{\Theta_{ij}\}$ とする．ここで，$q_i = \prod_{l:X_l \in G_i} r_l$ である．ベイジアンネットワークでは，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．

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

構造 $G$ に対して，確率分布 $P(\cdot)$ を厳密に表現するようなパラメータ $\Theta$ の値が存在するとき，$G$ は $P$ を含むと言う．

## 2.2　条件付き独立性

変数 $X, Y, Z$ について，構造 $G$ は $X, Y$ 間と $Y, Z$ 間にエッジを持ち，かつ $X, Y$ 間にエッジを持たないとする．$G$ において $X \to Z$，$Z \leftarrow Y$ となるようにエッジが引かれているとき，$X, Y, Z$ の結合関係を $Z$ に関する合流結合と呼び，$G$ は合流結合 $X \to Z \leftarrow Y$ を持つという．$G$ が合流結合を持たないとき，$X, Y, Z$ の結合関係を $Z$ に関する非合流結

合と呼び，$G$ は非合流結合 $X - Z - Y$ を持つという．ベイジアンネットワークは変数間の条件付き独立性を以下の有向分離によって表現する．

**定義 2.1** $G$ において $X, Y \in \mathbf{V}$ を結ぶ全ての有向経路について，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. $X$ と $Y$ を結ぶ経路上に変数 $Z \in \mathbf{Z}$ に関する非合流結合が存在する．
2. $X$ と $Y$ を結ぶ経路上に変数 $Z \notin \mathbf{Z}$ に関する合流結合が存在し，かつ $Z$ の子孫が $\mathbf{Z}$ に属さない．

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

$Dsep_G(X, Y \mid \mathbf{Z})$ のとき，$G$ は $X$ と $Y$ が $\mathbf{Z}$ を所与として条件付き独立であることを表す．また，合流結合と非合流結合は以下の性質を満たす．

**定理 2.1** （Koller and Friedman[6]）任意の三変数を $X, Y, Z \in \mathbf{V}$ とする．構造 $G$ が合流結合 $X \to Z \leftarrow Y$ を持つとき，

- $\forall \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y, Z\}, \neg Dsep_G(X, Y \mid \mathbf{Z}, Z)$,
- $\exists \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y, Z\}, Dsep_G(X, Y \mid \mathbf{Z})$,

がそれぞれ成り立ち，$G$ が非合流結合 $X - Z - Y$ を持つとき上記の二命題の否定がそれぞれ成り立つ．

証明は Koller and Friedman[6] を参照してほしい．

同時確率分布 $P$ 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 $I_P(X, Y \mid \mathbf{Z})$ で表す．真の同時確率分布を $P^{True}(X_0, X_1, \cdots, X_n)$ とし，$P^{True}$ の条件付き独立性と条件付き従属性を過不足なく表現できる構造として，以下を定義する．

**定義 2.2**　以下を満たす構造 $G^{True}$ を真の構造と呼ぶ.

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

しかし，任意の分布に対して真の構造がただ一つのみ存在するわけではない．そのため，DAG が表現できる条件付き独立性のみを扱う I-map を以下で導入する.

**定義 2.3**　同時確率分布 $P$ について，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 $P$ についてのインディペンデントマップ (independent map: I-map) という.

$$\forall X, Y \in \mathbf{V}, \forall \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}, Dsep_{G}(X, Y \mid \mathbf{Z}) \Rightarrow I_{P}(X, Y \mid \mathbf{Z}).$$

構造 $G$ が同時確率分布 $P$ について I-map であることは，$G$ が $P$ を含むことと同値である [6]．したがって，I-map となる構造のパラメータを一致推定量で推定すると，そのベイジアンネットワークが表現する確率分布は漸近的に真の分布に一致する.

## 2.3　ベイジアンネットワークのパラメータ推定

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 $\mathbf{d}^t = (x_0^t, x_1^t, \cdots, x_n^t)$ と表し，学習データを $D = (\mathbf{d}^1, \cdots, \mathbf{d}^t, \cdots, \mathbf{d}^N)$ と表す．$D$ が得られた時のベイジアンネットワーク $(G, \Theta)$ の尤度は以下で表される.

$$\begin{aligned} P(D \mid G, \Theta) &= \prod_{t=1}^{N} P(x_0^t, x_1^t, \cdots, x_n^t \mid G, \Theta) \\ &= \prod_{i=0}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \end{aligned} \tag{1}$$

ここで，$P(x_0^t, x_1^t, \cdots, x_n^t \mid G, \Theta)$ は $P(X_0 = x_0^t, X_1 = x_1^t, \cdots, X_n = x_n^t \mid G, \Theta)$ を表し，$N_{ijk}$ は $X_i = k$ かつ $G_i = j$ となる頻度を表す．式（1）の尤度を最大にする $\theta_{ijk}$ の最尤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，ベイジアンネットワークのパラメータ推定量として，$\theta_{ijk}$ の期待値である Expected A Posteriori (EAP) が用いられる．ベイジアンネットワークの構造 $G$ 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．

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

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

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

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．

## 2.4 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．この問題をベイジアンネットワークの構造学習と呼ぶ．構造学習では，候補構造から最適な学習スコアを持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．学習スコアとして事後分布 $P(G \mid D) = P(D \mid G)P(G)/P(D)$ が用いられる．構造学習を行うときは構造に関する情報を持っていない場合が普通であるため，$P(G)$ は一様分布と仮定する．また，$P(D)$ は構造 $G$ に影響されない定数であるため，$P(D \mid G)$ を最大化する構造が事後分布を最大化する構造である．したがって，学習スコアとして $P(D \mid G)$ が一般によく用いられ，このスコアを周辺尤度（Marginal

Likelihood: ML）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現する構造は，それらの ML の値も同一でなければならないという尤度等価を導入した．そして，尤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．

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

ここで，$N'$ は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．さらに，$N'$ 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[7][8][9] の研究では，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，$-$ logML の近似である最小記述長（Minimum Description Length: MDL）[10] は，ベイジアンネットワークと学習データ $D$ の同時記述長を表す．

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

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尤度である．

BDeu と負の MDL は以下で定義される漸近一致性を持つ [11].

**定義 2.4**　真の同時確率分布を $P^{True}(X_0, X_1, \cdots, X_n)$ とする．サンプルサイズ $N \to \infty$ のとき，学習スコア $Score(\cdot)$ が次の二条件を満たすならば，そのスコアは漸近一致性を持つと言う．

1. 構造 $G, G'$ について，$G$ が $P^{True}$ を含み $G'$ が $P^{True}$ を含まないとき，$Score(G) > Score(G')$.
2. 構造 $G, G'$ について，$G$ と $G'$ が $P^{True}$ を含み，$G$ のパラメータ数が $G'$ のパラメータ数より少ないとき，$Score(G) > Score(G')$.

すなわち，サンプルサイズが十分にあるとき，BDeu や MDL を用いれば，真の分布を含む構造の中でパラメータ数が最小の構造を学習できる．

さらに，logBDeu と MDL は，スコアとして次の性質を満たす．

$$Score(G) = \sum_{i=0}^{n} Score_i(G_i) \tag{8}$$

ここで，$Score_i(G_i)$ は変数 $X_i$ とその親変数集合 $G_i$ 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 $G_i$ についてのローカルスコア $Score_i(G_i)$ は以下のように表せる．

$$\begin{aligned} Score_i(G_i) = & \sum_{j=1}^{q_i} \left( \log \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \right. \\ & \left. + \sum_{k=1}^{r_i} \log \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \right) \end{aligned} \tag{9}$$

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 [12][13].

# 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

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 [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．説明変数のインスタンス $(x_1, \cdots, x_n)$ が与えられた時，目的変数の推定値 $\hat{c}$ は以下のように得られる．

$$
\begin{aligned}
\hat{c} &= \underset{c \in \{1,\cdots,r_0\}}{\arg\max}\ P(c \mid x_1, \cdots, x_n, G, \Theta) \qquad (10)\\
&= \underset{c \in \{1,\cdots,r_0\}}{\arg\max}\ \frac{P(c, x_1, \cdots, x_n \mid G, \Theta)}{P(x_1, \cdots, x_n \mid G, \Theta)}\\
&= \underset{c \in \{1,\cdots,r_0\}}{\arg\max}\ P(c, x_1, \cdots, x_n \mid G, \Theta)\\
&= \underset{c \in \{1,\cdots,r_0\}}{\arg\max} \prod_{i=0}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} (\theta_{ijk})^{1_{ijk}}\\
&= \underset{c \in \{1,\cdots,r_0\}}{\arg\max} \left[ \prod_{j=1}^{q_0} \prod_{k=1}^{r_0} (\theta_{0jk})^{1_{0jk}} \right.\\
&\qquad \left. \times \prod_{i:X_i \in \mathbf{Ch}_{X_0}} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} (\theta_{ijk})^{1_{ijk}} \right]
\end{aligned}
$$

ここで，$1_{ijk}$ はインスタンス $(c, x_1, \cdots, x_n)$ において $X_i = k$ かつ $G_i = j$ の時に 1 をとり，それ以外の時は 0 をとる変数である．また，$\mathbf{Ch}_{X_0}$ は目的変数の子変数の集合である．式（10）の最右辺からわ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，およ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．すなわち，構造 $G$ における $X_0$ と $X_0$ のマルコフブランケットからなる部分構造と，$G$ そのものは，以下で定義される分類等価の関係が成り立つ．

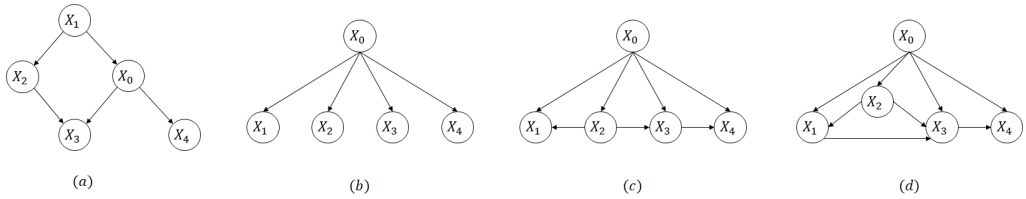


図 1　**（a）GBN の例;（b）Naive Bayes;（c）TAN の例;（d）ANB の例**

**定義 3.1**　二つの構造 $G, G'$ が $X_0$ について同じマルコフブランケット $M = \{X_1, \cdots, X_m\}$ を持つとする．$M$ に対する任意のインスタンス $\mathbf{d} = (x_1, \cdots, x_m)$ 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．

## 3.2　ベイジアンネットワーク分類器

一般に，ベイジアンネットワークの構造学習で探索する候補構造はとりう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（a））．つまり，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．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持つと仮定する Naive Bayes[14]（図 1 の（b））や，全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．尤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持つことが数値実験により示されている [1][15]．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なり説明変数間は DAG の仮定以外の制約を持たない 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，以下で表される，説明変数を所与とした目的変数の条件付き対数尤度（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時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時に更新を終了し，その時の構造を解とする．

一方，Carvalho ら [3] は，候補構造集合に ANB を仮定し，分解可能となるよう CLL を近似した aCLL（approximate Conditional Log Likelihood）を提案した．ここで，$t \in \{1, \cdots, N\}$，$c \in \{1, \cdots, r_0\}$ に対して，$J_{t,c} = P(c, x_1^t, \cdots, x_n^t \mid G, \Theta)$ とすると，CLL は次のように表せる．

$$\sum_{t=1}^{N} \log P(x_0^t \mid x_1^t, \cdots, x_n^t, G, \Theta) = \sum_{t=1}^{N} f(J_{t,1}, \cdots, J_{t,r_0})$$

ただし，

$$f(J_{t,1}, \cdots, J_{t,r_0}) = \log J_{t,x_0^t} - \log \left( \sum_{c=1}^{r_0} J_{t,c} \right)$$

である．今，$(J_{t,1}, \cdots, J_{t,r_0})$ が対称ディリクレ分布に従うことを仮定すると，$f$ を以下の $\hat{f}$ で近似できる．

$$\hat{f}(J_{t,1}, \cdots, J_{t,r_0}) = \log J_{t,x_0^t} + \sum_{c=1}^{r_0} \beta \log J_{t,c} + \gamma$$

ここで，$\beta$ と $\gamma$ は，2 点 $A = -\log(\sum_{c=1}^{r_0} J_{t,c})$ と $B = \sum_{c=1}^{r_0} \log J_{t,c}$ をランダムに多数発生させ，$A = \beta B + \gamma$ を満たすように推定するパラメータである．Carvalho らはこの近似を用いて以下の aCLL スコアを提案した．

$$\begin{aligned} aCLL(D \mid G) &= \sum_{t=1}^{N} \hat{f}(J_{t,1}, \cdots, J_{t,r_0}) \\ &= \sum_{t=1}^{N} \left( \log J_{t,x_0^t} + \sum_{c=1}^{r_0} \beta \log J_{t,c} + \gamma \right) \\ &\propto \sum_{i=1}^{n} \sum_{j=1}^{q_i^*} \sum_{k=1}^{r_i} \sum_{c=1}^{r_0} \left( N_{ijck} + \beta \sum_{c'=1}^{r_0} N_{ijc'k} \right) \log(\theta_{ijck}) \end{aligned}$$

ここで，$N_{ijck}$ は $X_i = k$ かつ $G_i \setminus \{X_0\} = j$ かつ $X_0 = c$ となる頻度を表し，$\theta_{ijck}$ は $G_i \setminus \{X_0\} = j$ かつ $X_0 = c$ の時に $X_i = k$ となる条件付き確率パラメータを表す．また，$q_i^* = \prod_{l:X_l \in G_i \setminus \{X_0\}} r_l$ である．ハイパーパラメータ $N'_{ijck} > 0$ に対して，aCLL を最大にするパラメータは次のように推定できる．

$$\hat{\theta}_{ijck} = \frac{N_{ij+ck}}{N_{ij+c}}$$

ここで，

$$N_{ij+ck} = \begin{cases} N_{ijck} + \beta \sum_{c'=1}^{r_0} N_{ijc'k} \\ (\text{if } N_{ijck} + \beta \sum_{c'=1}^{r_0} N_{ijc'k} \geq N'_{ijck}) \\ N'_{ijck} \\ (\text{if } N_{ijck} + \beta \sum_{c'=1}^{r_0} N_{ijc'k} < N'_{ijck}), \end{cases}$$

$$N_{ij+c} = \sum_{k=1}^{r_i} N_{ijc+k}$$

である．最適な $\beta$ と $\gamma$ に対して，aCLL は CLL の最小分散不偏推定量である．これらの

近似手法で学習した BNC の方が，ML（BDeu や MDL など）によって近似的に学習した BNC よりも分類精度が高いことが報告されてきた．

しかし，ML に対する CLL の優位性は理論的に示されていない．先行研究の比較実験では，ML による近似学習を行なっているため，探索精度の悪さが影響した可能性がある．そこで，次章では ML によって厳密に学習した BNC と CLL によって近似的に学習した BNC の分類精度を比較する．

# 4　GBN の厳密学習と識別モデルの比較実験

本章では，ML によって厳密に学習した BNC と CLL によって近似的に学習した BNC の分類精度を比較する．ML として用いるスコアは BDeu とし，擬似サンプル（ハイパーパラメータ）の値は Ueno[8, 9] の提案に従い 1.0 とした．この実験では次の 6 つの手法を比較する．

- 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 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 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

近似学習の構造探索法としては Hill-Climbing を用い，厳密学習の構造探索手法としては動的計画法 [12] を用いた．UCI レポジトリデータベース [16] の 43 個のベンチマークデータセットを用い，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．

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．ただし，表 1 に記載されている ANB-BDeu については 6 章で議論する．各データセットに対し，各手法の中で最も高い分類精度を太字で示している．また，分類精度と学習した GBN-BDeu の構造との関係をみるため

表 1　GBN-BDeu，ANB-BDeu と従来手法の分類精度（太字は最大の分類精度）

| No. | Dataset | Variables | Sample size | Classes | Naive-Bayes | GBN-CMDL | BNC2P | TAN-aCLL | gGBN-BDeu | GBN-BDeu | ANB-BDeu |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Balance Scale | 5 | 625 | 3 | **0.9152** | 0.3333 | 0.8560 | 0.8656 | **0.9152** | **0.9152** | **0.9152** |
| 2 | banknote authentication | 5 | 1372 | 2 | 0.8433 | **0.8819** | 0.8797 | 0.8761 | **0.8819** | 0.8812 | 0.8812 |
| 3 | Hayes-Roth | 5 | 132 | 3 | **0.8182** | 0.6136 | 0.6894 | 0.6742 | 0.7525 | 0.6136 | **0.8182** |
| 4 | iris | 5 | 150 | 3 | 0.7133 | 0.7800 | 0.8200 | 0.8200 | 0.8133 | **0.8267** | 0.8200 |
| 5 | lenses | 5 | 24 | 3 | 0.7500 | **0.8333** | 0.6667 | 0.7083 | **0.8333** | **0.8333** | 0.7500 |
| 6 | Car Evaluation | 7 | 1728 | 4 | 0.8571 | **0.9497** | 0.9416 | 0.9433 | 0.9416 | 0.9416 | 0.9427 |
| 7 | liver | 7 | 345 | 2 | 0.6319 | 0.6145 | 0.6290 | **0.6609** | 0.6029 | 0.6087 | 0.6348 |
| 8 | MONK' s Problems | 7 | 432 | 2 | 0.7500 | **1.0000** | **1.0000** | **1.0000** | 0.8449 | **1.0000** | **1.0000** |
| 9 | mux6 | 7 | 64 | 2 | 0.5469 | 0.3750 | **0.5625** | 0.4688 | 0.4063 | 0.4531 | 0.5469 |
| 10 | led7 | 8 | 3200 | 10 | 0.7294 | 0.7366 | **0.7375** | 0.7350 | 0.7297 | 0.7294 | 0.7294 |
| 11 | HTRU2 | 9 | 17898 | 2 | 0.7031 | 0.7096 | 0.7070 | 0.7018 | 0.7188 | **0.7305** | 0.7188 |
| 12 | Nursery | 9 | 12960 | 3 | 0.6782 | **0.7126** | 0.6092 | 0.5862 | **0.7126** | **0.7126** | 0.6782 |
| 13 | pima | 9 | 768 | 9 | 0.8966 | 0.9086 | 0.9118 | 0.9130 | 0.9092 | 0.9112 | **0.9141** |
| 14 | post | 9 | 87 | 5 | 0.9033 | 0.5823 | **0.9442** | 0.9177 | 0.9291 | 0.9340 | 0.9181 |
| 15 | Breast Cancer | 10 | 277 | 2 | **0.9751** | 0.8917 | 0.9473 | 0.9488 | 0.7058 | **0.9751** | **0.9751** |
| 16 | Breast Cancer Wisconsin | 10 | 683 | 2 | **0.7401** | 0.6209 | 0.6823 | 0.7184 | 0.7094 | 0.7184 | 0.7040 |
| 17 | Contraceptive Method Choice | 10 | 1473 | 3 | 0.4671 | 0.4501 | **0.4745** | 0.4705 | 0.4440 | 0.4542 | 0.4650 |
| 18 | glass | 10 | 214 | 6 | 0.5561 | 0.5654 | 0.5794 | 0.6308 | 0.4626 | 0.5701 | **0.6449** |
| 19 | shuttle-small | 10 | 5800 | 6 | 0.9384 | 0.9660 | 0.9703 | 0.9583 | 0.9683 | 0.9693 | **0.9716** |
| 20 | threeOf9 | 10 | 512 | 2 | 0.8164 | **0.9434** | 0.8691 | 0.8828 | 0.8652 | 0.8887 | 0.8730 |
| 21 | Tic-Tac-Toe | 10 | 958 | 2 | 0.6921 | **0.8841** | 0.7338 | 0.7203 | 0.6754 | 0.8340 | 0.8497 |
| 22 | MAGIC Gamma Telescope | 11 | 19020 | 2 | 0.7482 | 0.7849 | 0.7806 | 0.7631 | 0.7844 | 0.7873 | **0.7874** |
| 23 | Solar Flare | 11 | 1389 | 9 | 0.7811 | 0.8265 | 0.8315 | 0.8229 | **0.8431** | **0.8431** | 0.8229 |
| 24 | heart | 14 | 270 | 2 | **0.8259** | 0.8185 | 0.8037 | 0.8148 | 0.8222 | **0.8259** | 0.8185 |
| 25 | wine | 14 | 178 | 3 | 0.9270 | **0.9438** | 0.9157 | 0.9326 | 0.9045 | 0.9270 | 0.9270 |
| 26 | cleve | 14 | 296 | 2 | **0.8412** | 0.8209 | 0.8007 | 0.8378 | 0.7973 | 0.7973 | 0.8277 |
| 27 | australian | 15 | 690 | 2 | 0.8290 | 0.8312 | 0.8348 | 0.8464 | 0.8420 | **0.8536** | 0.8246 |
| 28 | crx | 15 | 653 | 2 | 0.8377 | 0.8346 | 0.8208 | 0.8560 | **0.8622** | 0.8591 | 0.8515 |
| 29 | EEG | 15 | 14980 | 2 | 0.5778 | 0.6787 | 0.6374 | 0.6125 | 0.6732 | 0.6814 | **0.6864** |
| 30 | Congressional Voting Records | 17 | 232 | 2 | 0.9095 | 0.9698 | 0.9612 | 0.9181 | **0.9741** | 0.9655 | 0.9483 |
| 31 | zoo | 17 | 101 | 5 | 0.9802 | 0.9109 | 0.9505 | **1.0000** | 0.9505 | 0.9307 | 0.9505 |
| 32 | pendigits | 17 | 10992 | 10 | 0.8032 | 0.9062 | 0.8719 | 0.8700 | 0.9253 | **0.9290** | 0.9279 |
| 33 | letter | 17 | 20000 | 26 | 0.4466 | 0.5796 | 0.5132 | 0.5093 | 0.5761 | 0.5761 | **0.5935** |
| 34 | ClimateModel | 19 | 540 | 2 | 0.9222 | **0.9407** | 0.9241 | 0.9333 | 0.9370 | 0.9000 | 0.8426 |
| 35 | Image Segmentation | 19 | 2310 | 7 | 0.7290 | 0.7918 | 0.7991 | 0.7407 | 0.8026 | 0.8156 | **0.8225** |
| 36 | lymphography | 19 | 148 | 4 | **0.8446** | 0.7939 | 0.7973 | 0.8311 | 0.7905 | 0.7500 | 0.7770 |
| 37 | vehicle | 19 | 846 | 4 | 0.4350 | 0.5910 | 0.5910 | 0.5816 | 0.5461 | 0.5768 | **0.6253** |
| 38 | hepatitis | 20 | 80 | 2 | 0.8500 | 0.7375 | **0.8875** | 0.8750 | 0.8500 | 0.5875 | 0.6250 |
| 39 | german | 21 | 1000 | 2 | 0.7430 | 0.6110 | 0.7340 | **0.7470** | 0.7140 | 0.7210 | 0.7380 |
| 40 | bank | 21 | 30488 | 2 | 0.8544 | 0.8618 | 0.8928 | 0.8618 | 0.8952 | **0.8956** | 0.8950 |
| 41 | waveform-21 | 22 | 5000 | 3 | 0.7886 | 0.7862 | 0.7754 | 0.7896 | 0.7698 | 0.7846 | **0.7966** |
| 42 | Mushroom | 22 | 5644 | 2 | 0.9957 | **1.0000** | **1.0000** | 0.9995 | **1.0000** | 0.9949 | **1.0000** |
| 43 | spect | 23 | 263 | 2 | 0.7940 | 0.7940 | 0.7903 | 0.8090 | 0.7603 | 0.7378 | **0.8240** |
| | average | | | | 0.7764 | 0.7721 | 0.7936 | 0.7943 | 0.7867 | 0.7963 | **0.8061** |
| | p-value | | | | 0.0031 | 0.0414 | 0.0067 | 0.0561 | 0.0629 | 0.2263 | - |

に，10 分割交差検証における GBN-BDeu の学習構造の目的変数の親変数数の平均を表 2 の"Parents"に示した．表 2 の"Children"は，10 分割交差検証における GBN-BDeu の学習構造の目的変数の子変数数の平均を示している．表 2 の"Sparse data"は $N_{0j} = 0$ となる空データをとる $j$ のパターン数の平均を示している．ただし，表 2 に記載されている MBsize，Missing variables，Extra variables，Max parents については 6 章で言及する．

表 2　GBN-BDeu と MANB-BDeu の学習構造の詳細

| No. | Dataset | Variables | Classes | Sample size | Parents | Children | Sparse data | MBsize | Missing variables | Extra variables | Max parents |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Balance Scale | 5 | 3 | 625 | 0.4 | 3.6 | 0.0 | 4.0 | 0.0 | 0.0 | 1.0 |
| 2 | banknote authentication | 5 | 2 | 1372 | 0.0 | 2.0 | 0.0 | 4.0 | 0.0 | 0.0 | 4.0 |
| 3 | Hayes–Roth | 5 | 3 | 132 | 3.0 | 0.0 | 17.2 | 3.0 | 0.0 | 0.0 | 1.0 |
| 4 | iris | 5 | 3 | 150 | 1.8 | 1.2 | 0.0 | 3.0 | 0.0 | 0.0 | 2.0 |
| 5 | lenses | 5 | 3 | 24 | 1.1 | 1.0 | 0.0 | 2.0 | 0.0 | 0.1 | 1.1 |
| 6 | Car Evaluation | 7 | 4 | 1728 | 2.0 | 3.0 | 0.0 | 5.0 | 0.0 | 0.0 | 2.0 |
| 7 | liver | 7 | 2 | 345 | 0.0 | 1.9 | 0.0 | 3.4 | 1.6 | 0.0 | 2.0 |
| 8 | MONK’s Problems | 7 | 2 | 432 | 3.0 | 0.0 | 0.0 | 3.0 | 0.0 | 0.0 | 3.0 |
| 9 | mux6 | 7 | 2 | 64 | 5.8 | 0.0 | 5.2 | 5.8 | 0.2 | 0.0 | 1.0 |
| 10 | led7 | 8 | 10 | 3200 | 0.9 | 6.1 | 0.0 | 7.0 | 0.0 | 0.0 | 1.0 |
| 11 | HTRU2 | 9 | 2 | 17898 | 1.8 | 4.2 | 0.0 | 7.0 | 0.0 | 0.0 | 3.0 |
| 12 | Nursery | 9 | 5 | 12960 | 4.0 | 3.0 | 0.0 | 7.0 | 0.0 | 0.0 | 3.0 |
| 13 | pima | 9 | 2 | 768 | 1.4 | 1.7 | 0.0 | 4.0 | 0.0 | 0.2 | 2.0 |
| 14 | post | 9 | 3 | 87 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 15 | Breast Cancer | 10 | 2 | 277 | 0.9 | 8.0 | 0.0 | 8.0 | 0.0 | 0.9 | 1.0 |
| 16 | Breast Cancer Wisconsin | 10 | 2 | 683 | 0.7 | 0.3 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 |
| 17 | Contraceptive Method Choice | 10 | 3 | 1473 | 0.7 | 0.8 | 0.0 | 1.5 | 0.5 | 0.2 | 1.2 |
| 18 | glass | 10 | 6 | 214 | 0.6 | 3.1 | 0.0 | 4.2 | 0.8 | 0.1 | 1.6 |
| 19 | shuttle-small | 10 | 6 | 5800 | 2.0 | 4.0 | 0.0 | 7.0 | 0.0 | 0.0 | 5.0 |
| 20 | threeOf9 | 10 | 2 | 512 | 5.0 | 2.1 | 0.0 | 7.6 | 1.4 | 0.0 | 2.4 |
| 21 | Tic-Tac-Toe | 10 | 2 | 958 | 1.2 | 2.2 | 0.0 | 5.1 | 0.9 | 0.2 | 3.0 |
| 22 | MAGIC Gamma Telescope | 11 | 2 | 19020 | 0.0 | 6.1 | 0.0 | 8.0 | 0.0 | 0.0 | 4.0 |
| 23 | Solar Flare | 11 | 9 | 1389 | 0.8 | 0.2 | 0.0 | 0.1 | 0.9 | 0.9 | 1.0 |
| 24 | heart | 14 | 2 | 270 | 1.8 | 4.2 | 0.0 | 6.0 | 1.0 | 0.3 | 1.5 |
| 25 | wine | 14 | 3 | 178 | 1.7 | 5.3 | 0.0 | 7.0 | 0.0 | 1.1 | 2.1 |
| 26 | cleve | 14 | 2 | 296 | 1.8 | 4.5 | 0.0 | 6.3 | 0.7 | 0.3 | 2.0 |
| 27 | Australian | 15 | 2 | 690 | 1.4 | 2.8 | 0.0 | 4.2 | 0.8 | 0.3 | 2.2 |
| 28 | crx | 15 | 2 | 653 | 1.3 | 2.8 | 0.0 | 2.9 | 1.1 | 1.3 | 2.0 |
| 29 | EEG | 15 | 2 | 14980 | 0.4 | 8.2 | 0.0 | 12.8 | 0.2 | 0.0 | 5.0 |
| 30 | Congressional Voting Records | 17 | 2 | 232 | 1.3 | 2.6 | 0.1 | 5.2 | 1.8 | 1.0 | 2.8 |
| 31 | zoo | 17 | 5 | 101 | 4.3 | 1.6 | 20.3 | 6.9 | 3.1 | 0.5 | 3.5 |
| 32 | pendigits | 17 | 10 | 10992 | 2.6 | 13.4 | 0.1 | 16.0 | 0.0 | 0.0 | 5.6 |
| 33 | letter | 17 | 26 | 20000 | 2.9 | 9.1 | 0.0 | 13.0 | 0.0 | 0.0 | 5.0 |
| 34 | ClimateModel | 19 | 2 | 540 | 1.8 | 4.4 | 0.0 | 15.9 | 1.1 | 0.7 | 14.0 |
| 35 | Image Segmentation | 19 | 7 | 2310 | 0.7 | 10.4 | 0.0 | 12.7 | 0.3 | 0.5 | 4.1 |
| 36 | lymphography | 19 | 4 | 148 | 1.6 | 5.9 | 0.2 | 9.0 | 1.0 | 4.1 | 8.0 |
| 37 | vehicle | 19 | 4 | 846 | 1.1 | 5.1 | 0.1 | 9.0 | 2.0 | 1.1 | 3.6 |
| 38 | hepatitis | 20 | 2 | 80 | 1.3 | 6.1 | 0.4 | 13.1 | 1.9 | 2.9 | 10.7 |
| 39 | German | 21 | 2 | 1000 | 1.1 | 2.8 | 0.0 | 3.9 | 2.1 | 0.2 | 1.2 |
| 40 | bank | 21 | 2 | 30488 | 4.1 | 2.0 | 32.5 | 9.9 | 0.1 | 0.0 | 5.0 |
| 41 | waveform-21 | 22 | 3 | 5000 | 3.8 | 10.1 | 0.0 | 13.9 | 0.1 | 0.6 | 4.0 |
| 42 | Mushroom | 22 | 2 | 5644 | 1.3 | 3.3 | 9.0 | 6.1 | 12.9 | 0.3 | 5.2 |
| 43 | spect | 23 | 2 | 263 | 2.0 | 3.4 | 0.0 | 6.4 | 2.6 | 1.3 | 2.5 |

表 1 を見ると，GBN-BDeu はサンプルサイズの大きいデータセット 11 番，12 番，22 番，32 番，40 番について，比較手法 6 つの中で分類精度が最も高い．この結果から，ML によって学習した BNC が必ずしも CLL によって学習した BNC より分類精度が低いとは限らないことがわかる．しかし，サンプルサイズの小さいデータセットである 3 番，9 番，31 番では GBN-BDeu の分類精度が著しく低い．これらのデータセットでは，目的変数の子変数が少なく，親変数が多いことが表 2 の"Children"，"Parents"からわかる．また，これらのデータセットでは"Sparse data"の値が大きいことがわかる．このように，

目的変数の親変数が多いと，親変数集合 $G_0$ の状態数 $q_0$ が指数的に増えるため，各状態 $G_0 = j$ に対して $N_{0j}$ が小さくなる．したがって，目的変数のパラメータ推定値（式（5））の精度が悪くなってしまう．さらに，式（10）より，目的変数の子変数が少ないほど目的変数のパラメータが分類に大きく影響を及ぼす．以上から，目的変数の親変数が多く子変数が少ないことが GBN-BDeu の分類精度の著しい低下の原因だと考えられる．次章でこの問題を解決する手法とそのアルゴリズムを提案する．

# 5 Augmented Naive Bayes 制約を持つ BNC の厳密学習

4 章では，サンプルサイズの小さいデータセットにおいて GBN-BDeu が目的変数の親変数が多く子変数が少ない構造を学習する場合があり，それが分類精度の著しい低下を招くことを示した．GBN-BDeu は目的変数のマルコフブランケットとそれ以外の変数の構造を同時に学習するが，分類には目的変数のマルコフブランケットのみが影響する．したがって，GBN-BDeu の目的変数のマルコフブランケットを説明変数集合として，上記の問題を緩和するような制約下で構造を学習すれば，分類精度を向上できる可能性がある．目的変数の親変数数と子変数数を制約する手法として，目的変数が親変数を持たず，全説明変数を子変数として持つ Augmented Naive Bayes（ANB）が知られている．本論では，GBN-BDeu の目的変数についてのマルコフブランケットのみを説明変数集合とし，ANB 構造を制約として BDeu により厳密学習する BNC（以降，このモデルを Markov blanket Augmented Naive Bayes （MANB）と呼ぶ）を提案する．ANB はこれまで識別モデルとして扱われてきたため，BDeu を最大化して学習することはなかった．本論の提案は，識別モデルの学習ではなく，生成モデルとしての GBN の学習に，目的変数の親変数が増えないように ANB 構造に制約することである．本論では，BDeu を用いて厳密学習した MANB を MANB-BDeu と呼ぶ．

## 5.1 MANB-BDeu の性質

本節では，サンプルサイズ $N \to \infty$ のときにおける MANB-BDeu の漸近的性質を示す．GBN-BDeu における $X_0$ のマルコフブランケットを $M = \{X_1, \cdots, X_m\}$ とすると，以下が成り立つ．

**補題 5.1** $N \to \infty$ のとき，MANB-BDeu の学習構造 $G^L$ は真の同時確率分布 $P^{True}(X_0, M)$ について I-map である．

**証明** $P^{True}(X_0, M)$ について I-map となる構造が MANB の構造空間に存在するとき，BDeu の漸近一致性により，$N \to \infty$ のもとで MANB-BDeu は $P^{True}(X_0, M)$ についてパラメータ数を最小とする I-map となる構造を学習する [11]．よって，$P^{True}(X_0, M)$ の分布形によらず $P^{True}(X_0, M)$ について I-map となる構造が MANB の構造空間に存在することを示せばよい．MANB の構造空間は $M \cup \{X_0\}$ からなる完全グラフを含む．完全グラフは $P^{True}(X_0, M)$ の分布形によらず $P^{True}(X_0, M)$ について I-map である．□

また，補題 5.1 より，以下が成り立つ．

**定理 5.1** $N \to \infty$ のとき，MANB-BDeu の学習構造 $G^L$ は真の条件付き確率分布 $P^{True}(X_0 \mid X_1, \cdots, X_n)$ を含む．

**証明** GBN-BDeu の学習構造を $G$ とする．$G$ における $X_0$ のマルコフブランケット $M$ の性質から，

$$P(X_0 \mid X_1, \cdots, X_n, G) = P(X_0 \mid M, G)$$

が成り立つ．BDeu が持つ漸近一致性の条件 1 より，$N \to \infty$ のとき GBN-BDeu は $P^{True}(X_0, X_1, \cdots, X_n)$ を含むため，

$$P^{True}(X_0 \mid X_1, \cdots, X_n) = P^{True}(X_0 \mid M)$$

が成り立つ．よって，$P^{True}(X_0 \mid M)$ を含む構造は $P^{True}(X_0 \mid X_1, \cdots, X_n)$ を含む．

ここで，$P^{True}(X_0 \mid M)$ は以下のように $P^{True}(X_0, M)$ で表すことができる．

$$P^{True}(X_0 \mid M) = \frac{P^{True}(X_0, M)}{\sum_M P^{True}(X_0, M)}$$

よって，$P^{True}(X_0, M)$ を含む構造は $P^{True}(X_0 \mid M)$ を含む．したがって，$P^{True}(X_0, M)$ を含む構造は $P^{True}(X_0 \mid X_1, \cdots, X_n)$ を含む．補題 5.1 より，$G^L$ は $P^{True}(X_0, M)$ を含むから，$G^L$ は $P^{True}(X_0 \mid X_1, \cdots, X_n)$ を含む．□

定理 5.1 より，真の構造が存在しなくても，MANB-BDeu は漸近的に真の条件付き確率分布 $P^{True}(X_0 \mid X_1, \cdots, X_n)$ を表現することを保証する．

一方で，MANB-BDeu では $X_0$ が親変数を持つことができず，有向分離の表現に制限があるため，以下で示すようにエッジ数の多い冗長な構造を学習してしまう場合がある．

**補題 5.2**　真の構造 $G^{True}$ が存在するとする．$\forall X, Y \in M$ について $G^{True}$ が合流結合 $X \rightarrow X_0 \leftarrow Y$ を持つとき，$P^{True}(X_0, M)$ について I-map となる MANB 構造 $G$ は $X, Y$ 間にエッジを持つ．

**証明**　$G^{True}$ は合流結合 $X \rightarrow X_0 \leftarrow Y$ を持つため，定理 2.1 と真の構造の定義 2.2 より，

$$\forall \mathbf{Z} \in M \setminus \{X, Y\}, \neg I_{P^{True}}(X, Y \mid X_0, \mathbf{Z}).$$

ここで，$G$ が $X, Y$ 間のエッジを持たないと仮定すると，$G$ は ANB 制約のため非合流結合 $X - X_0 - Y$ を持つ．よって定理 2.1 より，

$$\exists \mathbf{Z} \in M \setminus \{X, Y\}, Dsep_G(X, Y \mid X_0, \mathbf{Z}).$$

これは，$G$ が $P^{True}(X_0, M)$ について I-map であるという仮定に反する．よって，$G$ は $X, Y$ 間にエッジを持つ．□

補題 5.2 より，真の構造において $X_0$ が親変数を持つとき，$P^{True}(X_0, M)$ について I-map となる構造を学習する MANB-BDeu はエッジ数の多い冗長な構造を学習してしまう．エッジ数が多くなるとパラメータ数が増加するため，同時確率分布の収束が遅くなる．し

かし，以下で示すように，任意の構造 $G$ に対して，$X_0$ が親変数を持たないような構造 $G'$ が存在して，$G'$ は $G$ と分類等価である．

**補題 5.3** （Mihaljević ら [17]）任意の構造 $G$ における $X_0$ の親変数集合を $\mathbf{Pa}_{X_0}$ とする．$G$ における $X_0$ とそのマルコフブランケットから構成される部分構造を $G'$ とする．$G'$ に以下の操作を加えた構造は $G$ と分類等価である．

1. $\forall X \in \mathbf{Pa}_{X_0}$ について，$G'$ における $X$ から $X_0$ へのエッジを反転させる．
2. $\forall X, Y \in \mathbf{Pa}_{X_0}$ について，$G'$ に $XY$ 間のエッジを加える．

証明は Mihaljević ら [17] を参照してほしい．補題 5.3 より，真の構造と分類等価な MANB 構造が存在する場合がある．そのような MANB 構造を学習できれば，分類で用いる条件付き確率分布 $P(X_0 \mid X_1, \cdots, X_n)$ の推定値の収束速度が真の構造に一致し，最適であることがわかる．次の定理 5.2 では，以下の仮定 1 を満たすとき $N \to \infty$ のもとで MANB-BDeu の学習構造が真の構造 $G^{True}$ と分類等価であることを示す．

**仮定 1**　真の構造 $G^{True}$ が存在し，$G^{True}$ において $X_0$ のマルコフブランケットの全ての要素と $X_0$ が隣接している．

**定理 5.2**　仮定 1 が成り立つとき，$N \to \infty$ のもとで MANB-BDeu の学習構造は $G^{True}$ と分類等価である．

**証明**　補題 5.3 に示される操作によって得られる構造を $H$ とする．仮定 1 より，$H$ と $G^{True}$ において $M$ の各要素と $X_0$ は隣接しているため，

$$\forall X \in M, \forall \mathbf{Z} \subseteq M \setminus \{X\},\ \neg Dsep_H(X_0, X \mid \mathbf{Z}) \wedge \neg I_{P^{True}}(X_0, X \mid \mathbf{Z}). \tag{11}$$

$G^{True}$ における $X_0$ の親変数集合を $\mathbf{Pa}_{X_0}^{True}$ とする．$M$ に属する二変数について，片方または両方が $\mathbf{Pa}_{X_0}^{True}$ に属さないとき，$H$ と $G^{True}$ において二変数は $X_0$ に関して合流結合しない．したがって，二変数は $G^{True}$ と $H$ において $X_0$ が所与でないとき有向分離

されないため，

$$\forall X, Y \in M,\ s.t.\ \neg(X \in \mathbf{Pa}_{X_0}^{True} \wedge Y \in \mathbf{Pa}_{X_0}^{True}),$$
$$\forall \mathbf{Z} \subseteq M \setminus \{X, Y\},\ \neg Dsep_H(X, Y \mid \mathbf{Z}) \wedge \neg I_{P^{True}}(X, Y \mid \mathbf{Z}). \quad (12)$$

式 (11) と (12) より，

$$\forall X, Y \in M \cup \{X_0\},\ s.t.\ \neg(X \in \mathbf{Pa}_{X_0}^{True} \wedge Y \in \mathbf{Pa}_{X_0}^{True}),$$
$$\forall \mathbf{Z} \subseteq M \setminus \{X, Y\},\ \neg Dsep_H(X, Y \mid \mathbf{Z}) \wedge \neg I_{P^{True}}(X, Y \mid \mathbf{Z}). \quad (13)$$

ここで，命題 $p, q$ について $\neg p \wedge \neg q$ が真のとき $p \Leftrightarrow q$ も真であるため，式 (13) から次式が導ける．

$$\forall X, Y \in M \cup \{X_0\},\ s.t.\ \neg(X \in \mathbf{Pa}_{X_0}^{True} \wedge Y \in \mathbf{Pa}_{X_0}^{True}),$$
$$\forall \mathbf{Z} \subseteq M \setminus \{X, Y\},\ Dsep_H(X, Y \mid \mathbf{Z}) \Leftrightarrow I_{P^{True}}(X, Y \mid \mathbf{Z}). \quad (14)$$

また，$G^{True}$ において $X_0$ を所与とすると $\mathbf{Pa}_{X_0}^{True}$ の各要素は互いに有向分離されないため，このとき $G^{True}$ は $M \cup \{X_0\}$ 間について $H$ と同一の有向分離を表現する．したがって，

$$\forall X, Y \in M, \forall \mathbf{Z} \subseteq M \setminus \{X, Y\},$$
$$Dsep_H(X, Y \mid \mathbf{Z}, X_0) \Leftrightarrow I_{P^{True}}(X, Y \mid \mathbf{Z}, X_0). \quad (15)$$

式 (14) と (15) より，以下が成り立つ．

$$\forall X, Y \in M \cup \{X_0\},\ s.t.\ \neg(X \in \mathbf{Pa}_{X_0}^{True} \wedge Y \in \mathbf{Pa}_{X_0}^{True}),$$
$$\forall \mathbf{Z} \subseteq M \cup \{X_0\} \setminus \{X, Y\},\ Dsep_H(X, Y \mid \mathbf{Z}) \Leftrightarrow I_{P^{True}}(X, Y \mid \mathbf{Z}). \quad (16)$$

また，$H$ において $\mathbf{Pa}_{X_0}^{True}$ の各要素間が隣接することから，

$$\forall X, Y \in \mathbf{Pa}_{X_0}^{True}, \forall \mathbf{Z} \subseteq M \cup \{X_0\} \setminus \{X, Y\},\ \neg Dsep_H(X, Y \mid \mathbf{Z}). \quad (17)$$

式 (16) と (17) より

$$\forall X, Y \in M \cup \{X_0\}, \forall \mathbf{Z} \subseteq M \cup \{X_0\} \setminus \{X, Y\},$$
$$Dsep_H(X, Y \mid \mathbf{Z}) \Rightarrow I_{P^{True}}(X, Y \mid \mathbf{Z}). \quad (18)$$

式 (18) より $H$ は $P^{True}(X_0, M)$ について I-map である．補題 5.2 より，$P^{True}(X_0, M)$ について I-map となる MANB 構造は $\mathbf{Pa}_{X_0}^{True}$ に属する二変数間がエッジで結ばれなければならない．式 (16) より，それ以外の二変数間について $H$ が表現する条件付き独立性は真の条件付き独立性に一致している．すなわち，$H$ は $P^{True}(X_0, M)$ について I-map となる MANB 構造の中でパラメータ数が最小であるから，MANB-BDeu の学習構造は BDeu の漸近一致性により $N \to \infty$ のとき $H$ に一致する．したがって，$N \to \infty$ のとき MANB-BDeu の学習構造は $G^{True}$ と分類等価となる．□

定理 5.2 より，$N \to \infty$ のとき，MANB-BDeu の学習構造は $P^{True}(X_0 \mid M)$ を含むだけでなく，真の条件付き確率分布 $P^{True}(X_0 \mid X_1, \cdots, X_n)$ への収束速度が真の構造と等しくなる．以上から，MANB-BDeu は $N$ が小さいときは目的変数の親変数数が過多になりにくいため分類精度の著しい低下を抑えると期待でき，$N$ が大きいときでも分類性能が真の構造と厳密に等しい構造が得られる．

## 5.2 MANB-BDeu の厳密学習

MANB の厳密学習では，GBN の目的変数についてのマルコフブランケット $M = \{X_1, \cdots, X_m\}, (m = |M|)$ と目的変数 $X_0$ によって構成される全ての ANB 構造の中で，スコアを最大にする構造を探索する．ANB では目的変数が親変数を持たないため，全ての ANB 構造の目的変数についてのローカルスコアは等しい．このため，厳密解を得るには，構造全体のスコアから目的変数のローカルスコアを除いた $Score_{ANB}(G) = Score(G) - Score_0(\phi)$ を最大化する ANB 構造を探索すればよい．しかし，ANB の全構造数は説明変数数 $m$ の増加に対して超指数関数的に増加するため，構造を全探索すると計算ステップ数が $\mathcal{O}(m!2^{\binom{m}{2}})$ と膨大になってしまう．そこで本論では全構造を列挙せずに，変数間の半順序を推定することにより構造の計算ステップ数を大幅に削減するアルゴリズムを提案する．以下では提案アルゴリズムで用いる表記や用語を定義する．$M \cup \{X_0\}$ の

部分集合 $W$ で構成される ANB 構造の中で，$Score_{ANB}$ を最大にする構造を $G^*(W)$ とする．子変数を持たない変数をシンクと呼び，$G^*(W)$ におけるシンクを $X_s^*(W)$ と定義する．$W$ のべき集合に対して，$X_0$ を含む部分集合を $\Pi(W)$ と定義する．$X_i \in M$ の親変数集合 $G_i \subseteq M \cup \{X_0\} \setminus \{X_i\}, (X_0 \in G_i)$ について，$\Pi(G_i)$ を $X_i$ の親変数候補集合とよぶ．次式のように，$\Pi(G_i)$ の中でローカルスコアを最大にする親変数集合を最適親変数集合とよぶ．

$$g_i^*(\Pi(G_i)) = \arg\max_{g \in \Pi(G_i)} Score_i(g)$$

提案アルゴリズムは以下の 5 つのプロセスからなる．

1. GBN-BDeu を学習し，目的変数 $X_0$ のマルコフブランケット $M$ を取り出す．GBN-BDeu の学習法は Silander ら [12] が提案しており，それをそのまま用いる．以降の学習プロセスは，Silander らの GBN-BDeu の学習アルゴリズムをそのまま適用したものではなく，$M \cup \{X_0\}$ についての学習構造が ANB の制約を満たすように Silander らのアルゴリズムを改良したものである．
2. 各説明変数 $X_i \in M$ の親変数集合 $G_i \subseteq M \cup \{X_0\} \setminus \{X_i\}, (X_0 \in G_i)$ の各組合せについて logBDeu のローカルスコア（式（9））を計算する．全てのローカルスコアの計算ステップ数は $m2^{m-1}$ である．
3. 各説明変数 $X_i \in M$ の親変数集合 $G_i \subseteq M \cup \{X_0\} \setminus \{X_i\}, (X_0 \in G_i)$ の各組合せについて $g_i^*(\Pi(G_i))$ を計算する．全ての最適親変数集合の計算ステップ数は $m2^{m-1}$ である．
4. 全ての変数集合 $W \subseteq M \cup \{X_0\}, (X_0 \in W)$ に対して，$X_s^*(W)$ を計算する．全てのシンクの計算ステップ数は $2^m$ である．
5. プロセス（3）と（4）の結果を用いて $G^*(M \cup \{X_0\})$ を計算する．

プロセス（4）は以下の性質に基づいている．$G^*(W)$ はシンク $X_s^*(W)$ を必ず持っており，$G^*(W)$ において $X_s^*(W)$ は最適な親変数集合 $g_s^*(\Pi(W \setminus \{X_s^*(W)\}))$ を持っていな

ければならない．さらに，$G^*(W)$ において残りの $W \setminus \{X_s^*(W)\}$ で構成される構造も最適でなければならないため，次式が成り立つ．

$$X_s^*(W) = \underset{X_i \in W \setminus \{X_0\}}{\arg\max} \{Score_i(g_i^*(\Pi(W \setminus \{X_i\}))) + Score_{ANB}(G^*(W \setminus \{X_i\}))\} \tag{19}$$

式（19）によって $G^*(M \cup \{X_0\})$ をシンクとそれ以外の変数からなる構造に分解できる．この分解によりシンクを順に削除して新しいシンクを求めることができる．これを再帰的に適用して順に得られるシンク $X_s^*(W)$ とその最適親変数集合 $g_s^*(\Pi(W \setminus \{X_s^*(W)\}))$ の組をそれぞれ $(X_{s_1}, g_{s_1}^*), \cdots, (X_{s_i}, g_{s_i}^*) \cdots, (X_{s_m}, g_{s_m}^*)$ とすると，$G^*(M \cup \{X_0\})$ における変数 $X_{s_i}$ の親変数集合は $g_{s_i}^*$ である．以上により $G^*(M \cup \{X_0\})$ を求めることができる．構造探索を行うプロセス（3）と（4）における最適親変数集合とシンクの計算ステップ数は合わせて $\mathcal{O}(m2^m)$ であり，全ての構造パターンの計算ステップ数 $\mathcal{O}(m!2^{\binom{m}{2}})$ より小さい．このように提案アルゴリズムはシンクの順序付けによって，全ての構造パターンを計算する場合より，大幅に計算ステップ数を削減できる．

次に，最も計算時間のかかるプロセス（3）の最適親変数集合の計算と，プロセス（4）のシンクの計算の具体的なアルゴリズムを説明する．$g_i^*(\Pi(G_i))$ は $G_i$ そのものか，$G_i$ から一つだけ説明変数を取り除いた親変数候補集合 $\Pi(G_i \setminus \{X\}), (X \in G_i \setminus \{X_0\})$ についての最適親変数集合のうち，ローカルスコアが最大のものである [12]．したがって，次式が得られる．

$$Score_i(g_i^*(\Pi(G_i))) = \max(Score_i(G_i), Score1(G_i)), \tag{20}$$

ここで，

$$Score1(G_i) = \max_{X \in G_i \setminus \{X_0\}} Score_i(g_i^*(\Pi(G_i \setminus \{X\})))$$

である．式（20）によって $X_i \in M$ の全ての親変数候補集合に対して最適親変数集合を再帰的に求めることができる．しかし，再帰的な計算では，既に計算済みの最適親変

数集合を重複して計算してしまう．例えば，$g_1^*(\Pi(\{X_0, X_2, X_3, X_4\}))$ を計算する場合，$g_1^*(\Pi(\{X_0, X_2, X_3\}))$，$g_1^*(\Pi(\{X_0, X_2, X_4\}))$，$g_1^*(\Pi(\{X_0, X_3, X_4\}))$ を再帰的に計算する必要がある．しかし，$g_1^*(\Pi(\{X_0, X_2, X_3\}))$，$g_1^*(\Pi(\{X_0, X_2, X_4\}))$ の計算時に，それぞれ $g_1^*(\Pi(\{X_0, X_2\}))$ を重複して計算してしまう．したがって，式（20）を再帰的に計算する場合は，既に計算済みかどうかの確認が必要となり，計算時間が増加してしまう．そこで，再帰的な計算は行わず，Silander ら [12] と同様に，重複計算がないことを保証できる計算順序として $G_i$ のビット列表記に関する辞書式順序で式（20）を適用する．ここで，変数集合 $W \subseteq M \cup \{X_0\}$ のビット列表記は以下の長さ $m+1$ のビット列で定義される．

$$(b_m, \cdots, b_i, \cdots, b_1, b_0), \quad b_i = \begin{cases} 1 & (\text{if } X_i \in W) \\ 0 & (\text{otherwise}) \end{cases}$$

$G_i$ のビット列表記に関する辞書式順序は，$G_i \succ G_i \setminus \{X\}, (X \in G_i \setminus \{X_0\})$ と定義され，この順序で式（20）を適用すると，各計算で必要な最適親変数集合は必ず既知である．このため新たに最適親変数集合を計算する必要がなく，重複計算は行われない．例えば，$X_1$ の親変数集合 $\{X_0, X_2, X_3, X_4\}$（ビット列表記は 11101）について，$g_1^*(\Pi(11101))$ を計算する場合を考える．この計算を行うには，$g_1^*(\Pi(01101))$，$g_1^*(\Pi(10101))$，$g_1^*(\Pi(11001))$ を求める必要がある．辞書式順序では $11101 \succ 01101, 10101, 11001$ であるから，それら 3 つは既に計算済みである．よって重複計算を行わずに目的の $\Pi(11101)$ に対する最適親変数集合を計算できる．式（19）によるシンクの計算でも同様に，再帰的な計算では重複によって計算時間が増加してしまう．例えば，変数集合 11111 に対してシンクを計算する場合，01111，10111，11011，11101 に対するシンクを再帰的に計算する必要がある．しかし，01111，10111 に対するシンクの計算時に，それぞれ 00111 に対するシンクを重複して計算してしまう．そこで，最適親変数集合の計算と同様に，変数集合のビット列表記に関する辞書式順序で式（19）を適用して，この問題を回避する．辞書式順序では $11111 \succ 01111, 10111, 11011, 11101$ であるから，11111 に対するシンクの計算時に 01111，10111，11011，11101 に対するシンクは既に計算済みである．よって重複計算を

行わずに目的の 11111 に対するシンクを計算できる．

変数選択を行わない場合，提案手法におけるローカルスコア，最適親変数集合，シンクの各計算ステップ数は，全説明変数数を $n$ とするとそれぞれ $n2^{n-1}$，$n2^{n-1}$，$2^n$ である．Silander らの動的計画法 [12] におけるそれら 3 つの計算ステップ数はそれぞれ $(n+1)2^n$，$(n+1)2^n$，$2^{n+1}$ である．したがって，変数選択を行わない場合，提案手法におけるそれら 3 つの各計算ステップ数は Silander らの動的計画法における各計算ステップ数の約半分に減少する．動的計画法の計算時間はこれらの計算ステップ数の減少に対して線形的に減少するため，計算時間も約半分に短縮される．

次章で，提案手法で学習した BNC と，CLL を用いて学習した BNC の分類精度を比較する．

# 6　評価実験

本章では，提案手法と CLL をベースとした先行研究の手法の分類精度を比較するため，リポシトリデータを用いた評価実験を行う．まず，BDeu によって生成モデルとして厳密に学習した ANB（ANB-BDeu）と，4 章の実験に用いた手法の分類精度を比較する．ANB-BDeu は 5 章で紹介したアルゴリズムにおけるマルコフブランケット $M$ を全説明変数集合に置き換えたものを用いて学習できる．ANB-BDeu とその他の手法との有意性を示すため，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 [18, 19] を行った．検定の $p$ 値を表 1 の最下部に示した．また，表 2 の"MBsize"は，10 分割交差検証における GBN-BDeu の学習構造の目的変数のマルコフブランケットの要素数の平均を示している．

結果として，ANB-BDeu は Naive Bayes，GBN-CMDL，BNC2P よりも有意水準 5% のもとで有意に分類精度が高かった．さらに，ANB-BDeu は GBN-BDeu の分類精度が著しく悪かったデータセット 3 番，9 番，31 番において，分類精度を改善していることがわかる．しかし，データセット 5 番と 14 番では ANB-BDeu は GBN-BDeu の分類精度を大きく下回っている．これらのデータセットでは，目的変数のマルコフブランケットの要素数が少ないことが表 2 の"MBsize"からわかる．したがって，マルコフブランケットによる変数選択は ANB-BDeu の分類精度を改善すると期待できる．

次に，表 1 に示される手法と BDeu を用いて厳密学習した MANB（MANB-BDeu）の分類精度を比較する．公正に比較するため，MANB-BDeu 以外の手法にも GBN-BDeu の目的変数のマルコフブランケットによる変数選択を行った．変数選択をした各手法の名前は，元々の手法名の最初に"M"を付けて表す．表 3 に，各データセットに対する変数選択をした各手法の分類精度を示す．また，表 3 の最下部には MANB-BDeu とその他の手法の分類精度に対する Hommle の多重検定の $p$ 値を示している．結果として，

表 3　変数選択を用いた場合の各手法の分類精度

| No. | Dataset | Variables | Sample size | Classes | MNaive-Bayes | MGBN-CMDL | MBNC 2P | MTAN-aCLL | MgGBN-BDeu | GBN-BDeu | MANB-BDeu |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Balance Scale | 5 | 625 | 3 | **0.9152** | 0.3333 | 0.8560 | 0.8656 | **0.9152** | **0.9152** | **0.9152** |
| 2 | banknote authentication | 5 | 1372 | 2 | 0.8433 | **0.8819** | 0.8783 | 0.8761 | 0.8812 | 0.8812 | 0.8812 |
| 3 | Hayes-Roth | 5 | 132 | 3 | **0.8333** | 0.6136 | 0.7197 | 0.7879 | 0.7980 | 0.6136 | **0.8333** |
| 4 | iris | 5 | 150 | 3 | **0.8267** | 0.7800 | 0.8200 | 0.8200 | 0.8200 | **0.8267** | **0.8267** |
| 5 | lenses | 5 | 24 | 3 | 0.8333 | 0.8333 | 0.8333 | 0.8333 | **0.8750** | 0.8333 | 0.8333 |
| 6 | Car Evaluation | 7 | 1728 | 4 | 0.8559 | 0.9242 | 0.9375 | 0.9363 | **0.9416** | **0.9416** | **0.9416** |
| 7 | liver | 7 | 345 | 2 | **0.6348** | **0.6348** | 0.6000 | 0.5942 | 0.6000 | 0.6087 | 0.5855 |
| 8 | MONK’s Problems | 7 | 432 | 2 | 0.7500 | **1.0000** | **1.0000** | **1.0000** | 0.8194 | **1.0000** | **1.0000** |
| 9 | mux6 | 7 | 64 | 2 | 0.5469 | 0.3750 | **0.6250** | 0.4688 | 0.3906 | 0.4531 | 0.5469 |
| 10 | led7 | 8 | 3200 | 10 | 0.7294 | 0.7363 | **0.7375** | 0.7350 | 0.7303 | 0.7294 | 0.7294 |
| 11 | HTRU2 | 9 | 17898 | 2 | 0.7083 | 0.7057 | 0.7044 | 0.7070 | **0.7305** | **0.7305** | 0.7227 |
| 12 | Nursery | 9 | 12960 | 3 | **0.7126** | **0.7126** | **0.7126** | **0.7126** | **0.7126** | **0.7126** | **0.7126** |
| 13 | pima | 9 | 768 | 9 | 0.9102 | 0.9046 | 0.9076 | **0.9141** | 0.9083 | 0.9112 | **0.9141** |
| 14 | post | 9 | 87 | 5 | 0.8996 | 0.8775 | 0.9322 | 0.9103 | 0.9258 | **0.9340** | 0.9174 |
| 15 | Breast Cancer | 10 | 277 | 2 | **0.9751** | 0.8909 | 0.9663 | 0.9458 | 0.9429 | **0.9751** | **0.9751** |
| 16 | Breast Cancer Wisconsin | 10 | 683 | 2 | **0.7184** | **0.7184** | **0.7184** | **0.7184** | **0.7184** | **0.7184** | 0.7166 |
| 17 | Contraceptive Method Choice | 10 | 1473 | 3 | 0.4549 | 0.4542 | **0.4555** | 0.4535 | 0.4501 | 0.4542 | 0.4549 |
| 18 | glass | 10 | 214 | 6 | **0.5841** | 0.5514 | 0.5467 | **0.5841** | 0.5047 | 0.5701 | 0.5654 |
| 19 | shuttle-small | 10 | 5800 | 6 | 0.9360 | 0.9645 | 0.9666 | 0.9605 | 0.9690 | **0.9693** | **0.9693** |
| 20 | threeOf9 | 10 | 512 | 2 | 0.8145 | 0.8750 | 0.8750 | 0.8809 | 0.8652 | **0.8887** | 0.8711 |
| 21 | Tic-Tac-Toe | 10 | 958 | 2 | 0.7182 | **0.8476** | 0.7244 | 0.7213 | 0.7359 | 0.8340 | **0.8476** |
| 22 | MAGIC Gamma Telescope | 11 | 19020 | 2 | 0.7520 | 0.7841 | 0.7807 | 0.7699 | 0.7875 | 0.7873 | **0.7880** |
| 23 | Solar Flare | 11 | 1389 | 9 | **0.8431** | **0.8431** | **0.8431** | **0.8431** | **0.8431** | **0.8431** | **0.8431** |
| 24 | heart | 14 | 270 | 2 | 0.8222 | 0.8185 | 0.8148 | 0.8259 | 0.7889 | 0.8259 | **0.8296** |
| 25 | wine | 14 | 178 | 3 | **0.9607** | 0.9494 | 0.9438 | 0.9494 | 0.9326 | 0.9270 | 0.9326 |
| 26 | cleve | 14 | 296 | 2 | **0.8176** | **0.8176** | 0.7804 | 0.8108 | 0.7905 | 0.7973 | 0.8108 |
| 27 | australian | 15 | 690 | 2 | 0.8536 | **0.8580** | 0.8493 | 0.8522 | 0.8507 | 0.8536 | 0.8507 |
| 28 | crx | 15 | 653 | 2 | **0.8622** | 0.8545 | 0.8545 | **0.8622** | 0.8576 | 0.8591 | **0.8622** |
| 29 | EEG | 15 | 14980 | 2 | 0.5774 | 0.6790 | 0.6389 | 0.6111 | 0.6670 | 0.6814 | **0.6935** |
| 30 | Congressional Voting Records | 17 | 232 | 2 | 0.9353 | **0.9698** | 0.9655 | 0.9397 | 0.9655 | 0.9655 | 0.9569 |
| 31 | zoo | 17 | 101 | 5 | 0.9406 | 0.9406 | 0.9307 | 0.9307 | **0.9505** | 0.9307 | **0.9505** |
| 32 | pendigits | 17 | 10992 | 10 | 0.8032 | 0.9062 | 0.8719 | 0.8700 | 0.9253 | 0.9290 | **0.9297** |
| 33 | letter | 17 | 20000 | 26 | 0.4536 | **0.5796** | 0.5068 | 0.5036 | 0.5636 | 0.5761 | 0.5779 |
| 34 | ClimateModel | 19 | 540 | 2 | 0.9259 | **0.9407** | 0.9222 | 0.9352 | 0.9370 | 0.9000 | 0.8667 |
| 35 | Image Segmentation | 19 | 2310 | 7 | 0.7662 | 0.7848 | 0.7918 | 0.7922 | 0.8022 | 0.8156 | **0.8203** |
| 36 | lymphography | 19 | 148 | 4 | **0.8176** | 0.7027 | 0.7770 | 0.8041 | 0.7770 | 0.7500 | 0.8108 |
| 37 | vehicle | 19 | 846 | 4 | 0.4634 | 0.5816 | 0.5721 | 0.5922 | 0.5437 | 0.5768 | **0.6028** |
| 38 | hepatitis | 20 | 80 | 2 | **0.8750** | 0.8500 | 0.8625 | 0.8500 | 0.8625 | 0.5875 | 0.6625 |
| 39 | german | 21 | 1000 | 2 | 0.7210 | 0.7250 | **0.7350** | 0.7230 | 0.7230 | 0.7210 | 0.7240 |
| 40 | bank | 21 | 30488 | 2 | 0.8680 | 0.8955 | 0.8924 | 0.8777 | 0.8954 | 0.8956 | **0.8966** |
| 41 | waveform-21 | 22 | 5000 | 3 | 0.7852 | 0.7912 | 0.7806 | 0.7814 | 0.7626 | 0.7846 | **0.7920** |
| 42 | Mushroom | 22 | 5644 | 2 | 0.9970 | 0.9991 | 0.9991 | 0.9972 | **1.0000** | 0.9949 | **1.0000** |
| 43 | spect | 23 | 263 | 2 | **0.7865** | 0.7303 | 0.7416 | 0.7715 | 0.7715 | 0.7378 | 0.7603 |
| | average | | | | 0.7867 | 0.7801 | 0.7993 | 0.7981 | 0.7961 | 0.7963 | **0.8074** |
| | p-value | | | | 0.0089 | 0.0054 | 0.0104 | 0.0057 | 0.0188 | 0.0301 | - |

MANB-BDeu は全比較手法に対して有意水準 5% のもとで有意に分類精度が高かった．

表 2 の"Max parents"は 10 分割交差検証における MANB-BDeu の学習構造内の変数が持つ最大親変数数の平均を示している．"Max parents"の値が大きいほど MANB-BDeu の構造がより複雑であることを表す．表 2，表 3 より"Max parents"の値が大きいデータセット 36 番と 38 番では，MANB-BDeu は MNaive Bayes の分類精度を下回っていることがわかる．4 章で述べたように，変数の親変数が多くなるとその変数のパラメータの

推定精度は不安定になってしまう．MNaive Bayes は"Max parents"の値がわずか 1 でありパラメータの推定精度が安定するため，サンプルサイズが小さすぎると MNaive Bayes は MANB-BDeu よりも分類精度が高くなると考えられる．しかし，MNaive Bayes は説明変数間の相関を考慮できないため，サンプルサイズの大きいデータセット 8 番や 29 番では他のどの手法よりも分類精度が著しく悪くなっている．

MGBN-CMDL はデータセット 1 番，3 番，9 番，14 番，15 番などの多くのデータセットで他手法よりも分類精度が著しく悪くなっている．これは，Grossman ら [4] の実験結果と一致している．この理由として，CMDL スコアのペナルティ項とフィッティング項の整合性が取れていないことが考えられる．

MBNC2P と MTAN-aCLL は，"Max Parents"が 2 であるから，MNaive Bayes と同様にサンプルサイズの小さい 38 番のデータセットでは MANB-BDeu よりも分類精度が高い．しかし，データセット 29 番では，MANB-BDeu の"Max parents"の値が 2 より大きく，MBNC2P と MTAN-aCLL よりも分類精度が高くなっている．このように，サンプルサイズが大きくなると，より複雑な確率分布を表現できる MANB-BDeu の方が分類精度が高くなる．

MANB-BDeu はサンプルサイズが大きいデータセット 22 番，29 番，32 番，33 番，40 番で MgGBN-BDeu より分類精度が高い．この理由として，MgGBN-BDeu は BDeu を用いて近似学習しているが，MANB-BDeu は BDeu を用いて厳密学習しているため，サンプルサイズが大きくなると MANB-BDeu の方がデータの確率分布を正確に表現できることが考えられる．

また MANB-BDeu は，GBN-BDeu の学習構造の目的変数の子変数が少なく，親変数が多かったデータセット 3 番，9 番，31 番において，GBN-BDeu の分類精度を改善している．この理由は，4 章で述べた分類精度低下の問題を MANB-BDeu で緩和できるからである．

最後に，MANB-BDeu と ANB-BDeu を比較する．二つの違いは GBN-BDeu の目的

変数のマルコフブランケットによる変数選択をしているか否かである．表 2 の"Missing variables"は，目的変数と説明変数の間で真に相関があったのに変数選択で除去してしまった変数数の 10 分割交差検証における平均を示している．"Extra variables"は，目的変数と説明変数の間で真に無相関であったのに変数選択で除去しなかった変数数の 10 分割交差検証における平均を示している．実際には目的変数と説明変数の真の相関関係を知ることはできないので，交差検証で分割せずに全データを用いて学習した GBN-BDeu の目的変数のマルコフブランケットを，目的変数と真に相関のある説明変数集合とした．

結果として，MANB-BDeu はマルコフブランケットのサイズが小さいデータセット 5 番や 25 番では ANB-BDeu の分類精度を大きく上回っていることが，表 1 と表 2 の"MBsize"，表 3 からわかる．このように，目的変数と無相関な変数を除去することは分類精度を向上させることがわかる．"Extra variables"はほとんどのデータセットで少ない傾向であった．一方，"Missing variables"が多いデータセット 7 番，39 番，43 番では MANB-BDeu は ANB-BDeu の分類精度を大きく下回っている．このように，目的変数と相関のある変数を除去してしまうと，分類精度を低下させてしまう．表 2 より，データセット 12 番，22 番，40 番のようにサンプルサイズが大きい時は"Missing variables"が少なくなっている．この理由は，サンプルサイズが大きくなるとマルコフブランケットの変数選択の精度が上がるためだと考えられる．この結果から，MANB-BDeu はサンプルサイズが大きい時に，ANB-BDeu よりも分類精度が高くなると考えられる．

以上の結果は以下のようにまとめられる．

- ML によって厳密に学習した生成モデルの BNC は CLL によって近似的に学習した識別モデルの BNC より分類精度が必ずしも低いとは限らず，サンプルサイズが大きい時はむしろ生成モデルの方が分類精度が高い．
- サンプルサイズが小さいときに生成モデルの 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下がるので，目的変数のマルコフブランケット内

で，目的変数が説明変数を必ず子として持つ ANB 構造を制約として ML を最大化する生成モデルの厳密学習法を提案し，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た.

# 7　むすび

本論では，最初に ML によって厳密に学習した生成モデルの BNC と CLL によって近似的に学習した識別モデルの BNC の分類精度を比較した．その結果，ML を最大化する生成モデルが必ずしも CLL を最大化する識別モデルより分類精度が低いとは限らないことがわかった．しかし，ML は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ことがあり，その場合は CLL を最大化する BNC より著しく分類精度が低くなっていることがわかった．その理由は，目的変数の親変数が多いと目的変数のパラメータ推定精度が低下し，目的変数の子変数が少ないと推定精度の低い目的変数のパラメータが分類に大きく影響するからである．次に本論では，この問題を緩和するため，ML で厳密学習した GBN の目的変数についてのマルコフブランケットのみを説明変数集合とし，ANB 構造を制約として ML により厳密学習する BNC を提案した．リポジトリデータセットを用いた実験を行ったところ，提案手法は ML を最大化する生成モデルの GBN の問題点を改善し，分類精度の著しい低下を防ぐことができた．さらに，提案手法は既存の CLL を最大化する識別モデルの BNC よりも分類精度が有意に高いことを示した．サンプルサイズが大きい時は提案手法のマルコフブランケットの変数選択精度が高まり，分類精度も高くなることがわかった．以上から，提案手法はサンプルサイズが小さい時でも分類精度が安定し，サンプルサイズが大きい時も高い分類精度を示す．

本論の重要な知見として，BNC は生成モデルよりも識別モデルの方が分類精度が良いと報告されてきたが，その理由は目的変数の親変数パラメータ数が増えた時，一つのパラメータ学習のためのデータが疎となって学習精度が下がり，分類精度に影響していたことがわかった．そのため，識別モデルのためのスコアを必ずしも使う必要がなく，GBN の目的変数についてのマルコフブランケットに ANB の制約を入れた生成モデル学習で高い分類精度が実現できることが示唆された．

近い研究として，Isozaki ら [20][21][22] は，少数データで有効なパラメータ学習法と学習スコアを提案している．本研究の提案手法において BDeu スコアの代わりにこれらのパラメータ学習法とスコアを用いれば，サンプルサイズの小さいデータセットの分類精度のさらなる向上が期待できる．これは今後の課題とする．

# 謝辞

本論文を作成するにあたり，指導教員の植野真臣教授から，丁寧かつ熱心なご指導を賜りました．ここに感謝の意を表します．また，日頃から親身になって研究を支えていただいた宇都雅輝助教に深謝いたします．そして，ゼミや日常の議論を通じて多くの示唆や知識を頂いた川野秀一准教授，西山悠准教授，研究室の先輩・同期・後輩に感謝いたします．

# 参考文献


[1] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian Network Classifiers. *Machine Learning*, 29(2):131–163, 1997.

[2] Alexandra M. Carvalho, Teemu Roos, Arlindo L. Oliveira, and Petri Myllymäki. Discriminative Learning of Bayesian Networks via Factorized Conditional Log-Likelihood. *Journal of Machine Learning Research*, 12:2181–2210, 2011.

[3] Alexandra M. Carvalho, Pedro AdÃčo, and Paulo Mateus. Efficient Approximation of the Conditional Relative Entropy with Applications to Discriminative Learning of Bayesian Network Classifiers. *Entropy*, 15(7):2716–2735, 2013.

[4] Daniel Grossman and Pedro Domingos. Learning Bayesian Network classifiers by maximizing conditional likelihood. In *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, pages 361–368, 2004.

[5] David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, 20(3):197–243, 1995.

[6] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press, 2009.

[7] Maomi Ueno. Learning likelihood-equivalence bayesian networks using an empirical bayesian approach. *Behaviormetrika*, 35(2):115–135, 2007.

[8] Maomi Ueno. Learning Networks Determined by the Ratio of Prior and Data. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 598–605, 2010.

[9] Maomi Ueno. Robust learning Bayesian networks for prior belief. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 689–707, 2011.

[10] Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry Theory.* World Scientific Publishing Co., Inc., 1989.

[11] David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *J. Mach. Learn. Res.*, 2:445–498, March 2002.

[12] Tomi Silander and Petri Myllymäki. A Simple Approach for finding the Globally Optimal Bayesian Network Structure. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 445–452, 2006.

[13] Mark Barlett and James Cussens. Advances in Bayesian Network Learning Using Integer Programming. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 182–191, 2013.

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

[15] Michael G. Madden. On the classification performance of TAN and general Bayesian networks. *Knowledge-Based Systems*, pages 489 – 495, 2009.

[16] M. Lichman. UCI machine learning repository, 2013.

[17] Bojan Mihaljević, Concha Bielza, and Pedro Larrañaga. Learning bayesian network classifiers with completed partially directed acyclic graphs. In *Proceedings of the Ninth International Conference on Probabilistic Graphical Models*, volume 72 of *Proceedings of Machine Learning Research*, pages 272–283, 2018.

[18] G. Hommel. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, pages 383–386, 1988.

[19] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, 2006.

[20] T. Isozaki, N. Kato, and M. Ueno. Minimum free energies with "data temperature" for parameter learning of bayesian networks. In *2008 20th IEEE Interna-*

*tional Conference on Tools with Artificial Intelligence*, volume 1, pages 371–378, 2008.

[21] Takashi Isozaki, Noriji Kato, and Maomi Ueno. "data temperature" in minimum free energies for parameter learning of bayesian networks. *International Journal on Artificial Intelligence Tools*, 18:653–671, 2009.

[22] Takashi Isozaki and Maomi Ueno. Minimum free energy principle for constraint-based learning bayesian networks. In *Machine Learning and Knowledge Discovery in Databases*, pages 612–627, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.