電気通信大学情報理工学部
情報・通信工学科情報数理工学コース卒業論文

条件付き周辺尤度を用いたベイジアンネットワーク分類器学習

2018 年 2 月 26 日

情報数理工学コース

学籍番号 1411104

菅原聖太

指導教員 川野秀一

1　まえがき

離散変数を扱う分類器として，ベイジアンネットワーク分類器（Bayesian Network Classifier: BNC）が知られている [1]．BNC は，ベイジアンネットワークの下位モデルの一つである．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従属関係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の DAG 構造は一般に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．一般的に，学習スコアとして周辺尤度を用いる．周辺尤度はパラメータの事前分布にディリクレ分布を仮定すると閉形式で表すことができる．この時の周辺尤度は Bayesian Dirichlet（BD）スコアと呼ばれ，最もよく用いられる学習スコアである．また，周辺尤度を近似する情報理論の最小記述長（Minimum Description Length: MDL）[2] もしばしば構造学習に用いられる．MDL スコアは，構造のデータへのフィッティングを反映する対数尤度の項と，構造の複雑さに対するペナルティ項の和で表される．

BD スコアと MDL スコアは各変数とその親変数集合からなる局所構造のスコアについての総積（MDL の場合は総和）が構造全体のスコアに一致するという分解可能性を持っており，効率的な学習ができる．スコアベースアプローチによる構造学習は探索数がノード数に対し指数的に増加する NP 困難問題であるが [3]，分解可能なスコアを用いると，現在の最先端手法では 60 ノード程度の学習が可能である [4]．

学習されたベイジアンネットワークの一つのノードを目的変数，その他のノードを説明変数とすることで，そのモデルを分類器として扱うことができる．BD スコアや MDL スコアで学習した BNC は，全変数の同時確率をモデル化した生成モデルであり，目的変数を所与として説明変数を確率推論するなど，分類問題だけでなく多くの用途に柔軟に対応できる．

一方で，説明変数を所与とした目的変数の条件付き確率をモデル化する識別モデルの方が，漸近的な分類精度が生成モデルより高いことが知られている [5]．BD スコアや MDL スコアを用いたベイジアンネットワークの構造学習が識別モデルとしての分類器の構造を最適にする保証がないことが指摘されており [1]，識別モデルの学習スコアとして，MDL のフィッティング項の尤度を，説明変数を所与とした目的変数の条件付き尤度（Conditional Likelihood: CL）に置き換えた Conditional MDL（CMDL）が提案された [6]．MDL が周辺尤度の近似であるように，CMDL は条件付き周辺尤度（Conditional Marginal Likelihood: CML）の近似と解釈できる．しかし，実際に CMDL は CML から導出されておらず，CMDL が CML の高精度な近似である保証はなく，数学的なスコアの意味も明確ではない．

さらに CMDL の抱える問題として，CMDL のフィッティング項の CL は分解可能ではないため，CL を最適にするパラメータ推定式は閉形式で表せず，最適なパラメータを近似的に推定しなければならない．その推定手法として，勾配法を用いる Extended Logistic Regression（ELR）アルゴリズム [7] が提案されているが，ELR アルゴリズムのような勾配法では，一般的にはパラ

メータ推定値の大域的最適解が得られる保証がない [8]．さらに，候補構造ごとに ELR 推定を行わなければならないため，構造学習に膨大な時間がかかってしまう．

このように，CMDL は CML の近似である保証がなく，計算に勾配法によるパラメータ推定が必要なために推定値の精度保証ができず，学習全体の計算量も膨大である．

本論では，CML を直接定義し，それからスコアを近似的に導く．期待できる利点は以下の通りである．

1. 数学的にスコアの意味が明確である．
2. CML の直接の近似であり，分類精度の向上が期待できる．
3. CMDL で誤差の原因となってきた勾配法によるパラメータ推定を行う必要がないため，安定した分類精度と，計算量の大幅な減少が期待できる．

さらに，レポジトリデータベースでの評価実験により，提案スコアを用いて学習した BNC の方が従来スコアを用いて学習した BNC よりも分類精度が高いことを示し，提案スコアの有意性を示す．

2　ベイジアンネットワークと分類

2.1　ベイジアンネットワーク

2.1.1　ベイジアンネットワークのパラメータ推定

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，$n+1$ 個の離散確率変数集合 $\mathbf{X} = \{X_0, X_1, \cdots, X_i, \cdots, X_n\}$ 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，ベイジアンネットワークの構造を G とし，G における変数 X_i の親変数集合を Π_i とする．さらに，θ_{ijk} を Π_i が j 番目のパターンをとったとき（$\Pi_i = j$ と書く）に $X_i = k$ となる条件付き確率 $P(X_i = k \mid \Pi_i = j, G)$ 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．本論文では，Θ_{ij} が互いに独立であると仮定する．ベイジアンネットワークにおける同時確率分布 $P(X_0, X_1, \cdots, X_n \mid G, \Theta)$ は以下のように表現できる．

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

今，全変数に値が割り当てられたデータ列が N 個あり，t 番目のデータ列を $\mathbf{d}^t = \langle x_0^t, x_1^t, \cdots, x_n^t \rangle$ 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．式（1）に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（2）の事後分布 $p(\Theta_{ij} \mid D, G)$ が得られる．

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

$$p(\Theta_{ij} \mid D, G) = \frac{\Gamma\{\sum_{k=1}^{r_i}(N'_{ijk} + N^D_{ijk})\}}{\prod_{k=1}^{r_i}\Gamma(N'_{ijk} + N^D_{ijk})} \prod_{k=1}^{r_i} \theta_{ijk}^{N'_{ijk}+N^D_{ijk}-1} \tag{2}$$

事後分布 $p(\Theta_{ij} \mid D, G)$ より，パラメータは θ_{ijk} の期待値として，式（3）で推定できる．

$$\begin{aligned} \hat{\theta}_{ijk} &= E(\theta_{ijk} \mid D, G) \\ &= \int \theta_{ijk} \cdot p(\Theta_{ij} \mid D, G) d\Theta_{ij} \\ &= \frac{N'_{ijk} + N^D_{ijk}}{N'_{ij} + N^D_{ij}} \end{aligned} \tag{3}$$

ここで，N^D_{ijk} は D において $X_i = k$ かつ $\Pi_i = j$ となる頻度を表し，N'_{ijk} はディリクレ事前分布のハイパーパラメータを表す．また，$N^D_{ij} = \sum_{k=1}^{r_i} N^D_{ijk}$，$N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$ である．

2.1.2 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は，構造が定まれば式（3）で推定できるが，最適な構造も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．一般に学習スコアとして周辺尤度 $P(D \mid G)$ が用いられる．パラメータの事前分布がディリクレ分布と仮定すると，周辺尤度は次のように閉形式で表される．

$$P(D \mid G) = \prod_{i=0}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N^D_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N^D_{ijk})}{\Gamma(N'_{ijk})} \tag{4}$$

式（4）の周辺尤度は Bayesian Dirichlet（BD）スコアと呼ばれる．Heckerman ら [9] は，マルコフ等価な構造は，それらの周辺尤度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．

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

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 ESS を N'_{ijk} に分配する．この指標は， Bayesian Dirichlet equivalent（BDe）と呼ばれる．さらに，ESS 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 [9] や Ueno[10][11] 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，次式に示される，周辺尤度の近似である最小記述長（Minimum Description Length: MDL）[2] 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．

$$MDL(D \mid G, \Theta) = \frac{\log N}{2}|\Theta| - \log P(D \mid G, \Theta)$$

第一項は構造の複雑さに対するペナルティ項である．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り，$P(D \mid G, \Theta)$ は次のように表される．

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

ここで，$P(x_0^t, x_1^t, \cdots, x_n^t \mid G)$ は $P(X_0 = x_0^t, X_1 = x_1^t, \cdots, X_n = x_n^t \mid G)$ を表す．また，$\hat{\theta}_{ijk}$ は以下の最尤推定量で推定される．

$$\hat{\theta}_{ijk} = \frac{N_{ijk}^D}{N_{ij}^D} \tag{5}$$

BD スコアと MDL スコアは各変数 X_i とその親変数集合 Π_i からなる局所構造のスコアについての総積（MDL の場合は総和）が構造全体のスコアに一致するという分解可能性を持っており，効率的な学習ができる．スコアベースアプローチによる構造学習は探索数がノード数に対し指数的に増加する NP 困難問題であるが [3]，分解可能なスコアを用いると，現在の最先端手法では 60 ノード程度の学習が可能である [4]．

2.2 ベイジアンネットワーク分類器

2.2.1 ベイジアンネットワーク分類器

ベイジアンネットワークにおける一つのノードを目的変数とし，それ以外のノードを説明変数とすることで，ベイジアンネットワークを分類器として扱うことができる．分類器としてのベイジアンネットワークをベイジアンネットワーク分類器（BNC）と呼び，高い分類精度を持つことが知られている [1][12]．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 $\mathbf{e} = \langle x_1, \cdots, x_n \rangle$ 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．

$$\begin{aligned} \hat{c} &= \arg\max_{c \in \{1, \cdots, r_0\}} P(c \mid x_1, \cdots, x_n, G, \Theta) \\ &= \arg\max_{c \in \{1, \cdots, r_0\}} \frac{P(c, x_1, \cdots, x_n \mid G, \Theta)}{P(x_1, \cdots, x_n \mid G, \Theta)} \\ &= \arg\max_{c \in \{1, \cdots, r_0\}} P(c, x_1, \cdots, x_n \mid G, \Theta) \\ &= \arg\max_{c \in \{1, \cdots, r_0\}} \prod_{i=0}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} (\theta_{ijk})^{1_{\mathbf{e}^c, ijk}} \\ &= \arg\max_{c \in \{1, \cdots, r_0\}} \prod_{i: X_i \in \mathbf{C}} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} (\theta_{ijk})^{1_{\mathbf{e}^c, ijk}} \end{aligned} \tag{6}$$

ここで，$\mathbf{e}^c = \langle c, x_1, \cdots, x_n \rangle$ であり，$1_{\mathbf{d}, ijk}$ は全変数のデータ列 $\mathbf{d}$ に対して $X_i = k$ かつ $\Pi_i = j$ 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，$\mathbf{C}$ は目的変数と目的変数の子変数

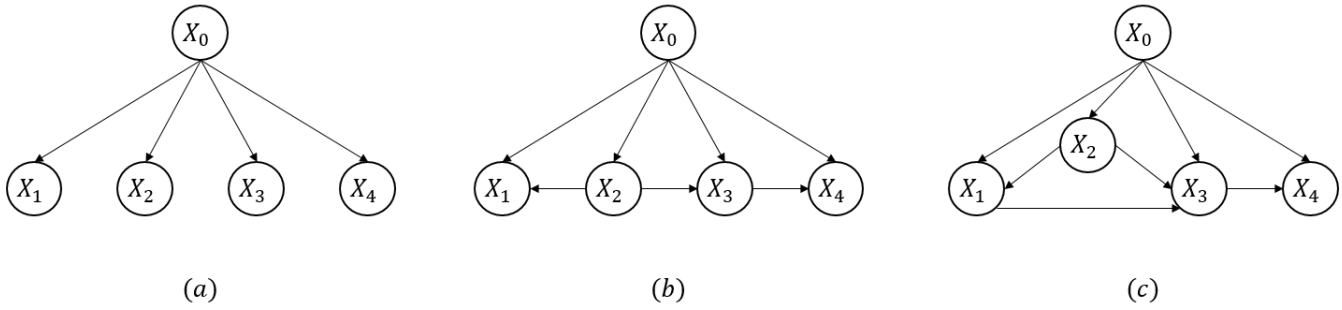

図 1　（**a**）**Naive Bayes** の例;（**b**）**TAN** の例;（**c**）**ANB** の例

の集合である．$\mathbf{C}$ に含まれない変数の条件付き確率パラメータは目的変数の確率推論に影響しないため，式（6）のように書ける．

一般に，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 BD スコアや MDL スコアなどを最適化して学習されるベイジアンネットワーク分類器（BNC）は General Bayesian Network（GBN）と呼ばれる．つまり，制約のない一般的なベイジアンネットワークを分類器として用いることを GBN と呼ぶ．しかし，BD や MDL を用いて学習された GBN は目的変数の親変数が多く子変数が少ない構造をとることがあり，このような構造ではパラメータの推定精度が悪くなり，分類精度が低くなる傾向がある [13]．この問題を回避できる BNC として，各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（a））や，各説明変数が目的変数とその他に一つ親を持つと仮定した Tree-Augmented Naive Bayes（TAN）[1]（図.1 の（b））などが知られている．尤度を学習スコアとした TAN の構造は多項式時間で学習でき，MDL スコアで学習された 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，NB や TAN を一般化した，各説明変数の親に目的変数が必ず含まれると仮定する Augmented Naive Bayes（ANB）（図.1 の（c））は，BD スコアや MDL スコアを用いて学習された GBN より分類精度が高いことが実験的に示されている [1]．

2.2.2　ベイジアンネットワーク分類器の学習

BD スコアや MDL スコアで学習した BNC は，全変数の同時確率をモデル化した生成モデルであり，目的変数を所与として説明変数を確率推論するなど，分類問題だけでなく多くの用途に柔軟に対応できる．

一方で，説明変数を所与とした目的変数の条件付き確率をモデル化する識別モデルの方が，漸近的な分類精度が生成モデルより高いことが知られている [5]．そこで，識別モデルの学習スコアとして MDL のフィッティング項の尤度を，説明変数を所与とした目的変数の条件付き尤度（Conditional Likelihood: CL）に置き換えた Conditional MDL（CMDL）が提案された [6]．今，学習データ D における変数 X_i のデータ列を $\mathbf{x}_i = \langle x_i^1, \cdots, x_i^N \rangle$ と表すと，CMDL スコアは次のように表せる．

$$CMDL(D \mid G, \Theta) = \frac{\log N}{2}|\Theta| - \log P(\mathbf{x}_0 \mid \mathbf{x}_1, \cdots, \mathbf{x}_n, G, \Theta) \tag{7}$$

ここで，条件付き尤度 $P(\mathbf{x}_0 \mid \mathbf{x}_1, \cdots, \mathbf{x}_n, G, \Theta)$ は次のように表せる．

$$P(\mathbf{x}_0 \mid \mathbf{x}_1, \cdots, \mathbf{x}_n, G, \Theta) = \prod_{t=1}^{N} P(x_0^t \mid x_1^t, \cdots, x_n^t, G, \Theta) \tag{8}$$
$$= \prod_{t=1}^{N} \frac{P(x_0^t, x_1^t, \cdots, x_n^t \mid G, \Theta)}{\sum_{c=1}^{r_0} P(c, x_1^t, \cdots, x_n^t \mid G, \Theta)}$$

MDL が周辺尤度の近似であるように，CMDL は条件付き周辺尤度（Conditional Marginal Likelihood: CML）の近似と解釈できる．しかし，実際に CMDL は CML から導出されておらず，CMDL が CML の高精度な近似である保証はなく，数学的なスコアの意味も明確ではない．

さらに，CMDL の抱える問題として，CMDL のフィッティング項である CL は分解可能ではないため，CL を最大にするパラメータ推定式は閉形式で表せず，候補構造ごとにパラメータを近似的に推定しなければならない．CL を最大にするパラメータの推定手法として，勾配法を用いる Extended Logistic Regression（ELR）アルゴリズム [7] が提案されているが，全候補構造に対する ELR によるパラメータ推定の計算量は莫大である．

そこで，Grossman ら [6] は，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大き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 [9] を用い，効率的に学習した．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学習スコアが大きくならない時に更新を終了する．実験結果として，Hill Climbing アルゴリズムの構造更新過程でパラメータを ELR アルゴリズムで推定した BNC は，単純に式（5）の最尤推定量で推定した BNC より分類精度が高くなかった．この理由として，ELR アルゴリズムのような勾配法では，一般に CL のパラメータ推定値の大域的最適解が得られる保証がないことが考えられる [8]．このように，CL スコアを最適にするパラメータ推定値の精度保証は難しい．

一方，Carvalho ら [16] は CL に対数をとったものを近似した，分解可能で効率的な aCLL（approximate Conditional Log Likelihood）スコアを提案した．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，CL に対数をとったものは次のように表せる．

$$\log P(\mathbf{x}_0 \mid \mathbf{x}_1, \cdots, \mathbf{x}_n, G, \Theta) = \sum_{t=1}^{N} f(J_{t,1}, \cdots, J_{t,r_0})$$

ただし，

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

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 aCLL スコアは次のように表

される．

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

ここで，β と γ 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．また，N_{ijck}^D は D において $X_i = k$ かつ $\Pi_i \setminus X_0 = j$ かつ $X_0 = c$ となる頻度を表し，θ_{ijck} は $\Pi_i \setminus X_0 = j$ 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（$j = 1, \cdots, q_i^*$）．また，擬似サンプル $N' > 0$ に対して，候補構造を全ての ANB の集合とすると，aCLL を最大にするパラメータは次のように推定できる．

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

ここで，

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

である．最適な β と γ に対して aCLL スコアは，条件付き尤度（CL）に対数をとったものの最小分散不偏推定量である．したがって，MDL スコアのフィッティング項を aCLL に置き換えた approximate CMDL（aCMDL）は，分解可能なため学習が効率的であり，CMDL を精度良く近似できると考えられる．

このように，これまで CMDL に関する研究がなされてきた．しかし，既に上で述べたように，MDL が周辺尤度を近似したスコアである一方，CMDL は条件付き周辺尤度（CML）を近似している保証がなく，数学的に何を意味しているか不明である．次節では，CML を直接定義し，それからスコアを近似的に導く．

3 条件付き周辺尤度

本節では，条件付き周辺尤度（CML）を定義し，その近似スコアの推定法を提案する．今，$\mathbf{c} = \langle c^1, \cdots, c^t, \cdots, c^N \rangle, (c^t \in \{1, \cdots, r_0\})$ とすると，CML は以下で定義できる．

定義 3.1 学習データ D に対する構造 G の CML は以下で定義できる．

$$
P(\mathbf{x}_0 \mid \mathbf{x}_1, \cdots, \mathbf{x}_n, G) = \frac{P(\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_n \mid G)}{\sum_{\mathbf{c}} P(\mathbf{c}, \mathbf{x}_1, \cdots, \mathbf{x}_n \mid G)} \tag{9}
$$

ここで，$\sum_{\mathbf{c}}$ はデータ列 $\mathbf{c}$ のとりうる値の全パターンを示しており，サンプルサイズに対してパターン数が指数的に増えてしまう．例えば，目的変数が 2 値をとる場合，2^N パターンものデータ列 $\mathbf{c}$ が存在する．一方，式（8）の CML を式（7）の条件付き尤度（CL）と比較すると，CL はパラメータ集合 Θ が所与となっていることがわかる．そのため，式（7）の分母は目的変数 X_0 がとりえるだけの和に分解され計算可能となるが，式（8）では厳密に分解できる手法がない．そこで，本論では擬似的オンライン学習を想定し，CML を近似した学習スコアを提案する．

今，$\mathbf{x}_i^{t;N} = \langle x_i^t, \cdots, x_i^N \rangle$，$D^{1;t} = \langle \mathbf{d}^1, \cdots, \mathbf{d}^t \rangle$ とし，データ $\mathbf{d}^t$ が疑似的に時系列データとして $t = 1, \cdots, N$ の順に与えられることを想定し，次式を考える．

$$
\begin{aligned}
P(\mathbf{x}_0 \mid \mathbf{x}_1, \cdots, \mathbf{x}_n, G) &= \prod_{t=1}^{N} P(x_0^t \mid x_0^1, \cdots, x_0^{t-1}, \mathbf{x}_1, \cdots, \mathbf{x}_n, G) \\
&= \prod_{t=1}^{N} P(x_0^t \mid D^{1;t-1}, \mathbf{x}_1^{t;N}, \cdots, \mathbf{x}_n^{t;N}, G) \\
&= \prod_{t=1}^{N} P(x_0^t | x_1^t, \cdots, x_n^t, D^{1;t-1}, \mathbf{x}_1^{t+1;N}, \cdots, \mathbf{x}_n^{t+1;N}, G)
\end{aligned}
\tag{10}
$$

ただし，$D^{1;0}$ はデータがないことを表す．ここで，式（12）の条件付き確率の条件部 $\mathbf{x}_1^{t+1;N}, \cdots, \mathbf{x}_n^{t+1;N}$ は時刻 t において与えられていない未来のデータである．そこで，$\mathbf{x}_1^{t+1;N}, \cdots, \mathbf{x}_n^{t+1;N}$ の情報を用いずに式（12）の確率を推論し，CML を以下のように近似する approximate CML（aCML）を，識別モデルの学習スコアとして提案する．

$$
\begin{aligned}
P(\mathbf{x}_0 \mid \mathbf{x}_1, \cdots, \mathbf{x}_n, G) &\approx \prod_{t=1}^{N} P(x_0^t \mid x_1^t, \cdots, x_n^t, D^{1;t-1}, G) \\
&= \prod_{t=1}^{N} \frac{P(x_0^t, x_1^t, \cdots, x_n^t \mid D^{1;t-1}, G)}{\sum_{c=1}^{r_0} P(c, x_1^t, \cdots, x_n^t \mid D^{1;t-1}, G)}
\end{aligned}
\tag{11}
$$

aCML の計算量はサンプルサイズに対し線形オーダーであり，現実的に計算が可能である．厳密式（10）と近似式（11）の違いは，式（11）が確率推論にデータ $\mathbf{x}_1^{t+1;N}, \cdots, \mathbf{x}_n^{t+1;N}$ を用いていない点である．厳密式（10）のように $\mathbf{x}_1^{t+1;N}, \cdots, \mathbf{x}_n^{t+1;N}$ を用いて確率推論するには，$\mathbf{x}_1^{t+1;N}, \cdots, \mathbf{x}_n^{t+1;N}$ に対応した目的変数のデータ列 $\mathbf{c}^{t+1:N} = \langle c^{t+1}, \cdots, c^N \rangle$ のとりうる値の全パターンに対して確率推論をしなければならならず，計算量がサンプルサイズに対して指数的に増加してしまう．しかし，$\mathbf{x}_1^{t+1;N}, \cdots, \mathbf{x}_n^{t+1;N}$ は対応した目的変数のデータがない欠損データと捉えることができ，欠損値を含まない完全データである $D^{1;t-1}$ と比べて確率推論に与える影響は小さいと考えられる．したがって，$\mathbf{x}_1^{t+1;N}, \cdots, \mathbf{x}_n^{t+1;N}$ を条件部から除いた式（11）は，現実的な時間で計算でき，近似精度を可能な限り高めた合理的な近似式と考えられる．

また，$c \in \{1, \cdots, r_0\}$ に対して $\mathbf{d}_c^t = \langle c, x_1^t, \cdots, x_n^t \rangle$ とすると，aCML の具体的な計算式を次の定理で示せる．

定理 3.1　aCML は，パラメータの事前分布にディリクレ分布を仮定すると，次のように表せる．

$$\prod_{t=1}^{N} P(x_0^t \mid x_1^t, \cdots, x_n^t, D^{1;t-1}, G) = \prod_{t=1}^{N} \frac{\prod_{i=0}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\left(\frac{N'_{ijk}+N_{ijk}^{D^{1;t-1}}}{N'_{ij}+N_{ij}^{D^{1;t-1}}}\right)^{1_{\mathbf{d}^t,ijk}}}{\sum_{c=1}^{r_0}\prod_{i=0}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\left(\frac{N'_{ijk}+N_{ijk}^{D^{1;t-1}}}{N'_{ij}+N_{ij}^{D^{1;t-1}}}\right)^{1_{\mathbf{d}_c^t,ijk}}} \tag{12}$$

証明　パラメータ集合 Θ_{ij} の独立性の仮定から，周辺尤度は次式で表される．

$$\begin{aligned}\prod_{t=1}^{N} P(x_0^t, x_1^t, \cdots, x_n^t \mid D^{1;t-1}, G) &= \prod_{t=1}^{N}\int P(x_0^t, x_1^t, \cdots, x_n^t \mid D^{1;t-1}, \Theta, G) \\ &\qquad \times p(\Theta \mid D^{1;t-1}, G)d\Theta \qquad (13)\\ &= \prod_{t=1}^{N}\prod_{i=0}^{n}\prod_{j=1}^{q_i}\int\prod_{k=1}^{r_i}\theta_{ijk}^{1_{\mathbf{d}^t,ijk}} p(\Theta_{ij} \mid D^{1;t-1}, G)d\Theta_{ij}\end{aligned}$$

ここで，$1_{\mathbf{d}^t,ijk}=1$ の時，式（13）の積分は確率密度 $p(\Theta_{ij} \mid D^{1;t-1}, G)$ に関する期待値として，次式で表される．

$$\prod_{t=1}^{N} P(x_0^t, x_1^t, \cdots, x_n^t \mid D^{1;t-1}, G) = \prod_{t=1}^{N}\prod_{i=0}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} E(\theta_{ijk} \mid D^{1;t-1}, G)^{1_{\mathbf{d}^t,ijk}} \tag{14}$$

式（14）に式（3）を代入すると次式が得られる．

$$\prod_{t=1}^{N} P(x_0^t, x_1^t, \cdots, x_n^t \mid D^{1;t-1}, G) = \prod_{t=1}^{N}\prod_{i=0}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\left(\frac{N'_{ijk}+N_{ijk}^{D^{1;t-1}}}{N'_{ij}+N_{ij}^{D^{1;t-1}}}\right)^{1_{\mathbf{d}^t,ijk}} \tag{15}$$

同様に，

$$\prod_{t=1}^{N}\sum_{c=1}^{r_0} P(c, x_1^t, \cdots, x_n^t \mid D^{1;t-1}, G) = \prod_{t=1}^{N}\sum_{c=1}^{r_0}\prod_{i=0}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\left(\frac{N'_{ijk}+N_{ijk}^{D^{1;t-1}}}{N'_{ij}+N_{ij}^{D^{1;t-1}}}\right)^{1_{\mathbf{d}_c^t,ijk}} \tag{16}$$

式（15）と式（16）をそれぞれ式（11）に代入すると，式（12）が得られる．□

ここで，尤度等価となるように $N'_{ijk}=N'P(X_i=k,\Pi_i=j \mid G^h)$ とする．$P(X_i=k,\Pi_i=j \mid G^h)$ が未知の場合は，$N'_{ijk}=N'/(r_iq_i)$ とする．

aCML スコアの期待できる利点は以下の通りである．

1. 数学的にスコアの意味が明確である．
2. CML の直接の近似であり，分類精度の向上が期待できる．
3. CMDL で誤差の原因となってきた勾配法によるパラメータ推定を行う必要がないため，安定した分類精度と，計算量の大幅な減少が期待できる．

次節で，aCML スコアを用いて学習されたベイジアンネットワーク分類器（BNC）と，従来手法で学習された BNC の分類精度と計算時間を比較する．

表 1　学習データ

	学習データ	説明変数数	目的変数の とりうる値の数	サンプル サイズ	変数列の とりうる値の パターン数	標準化サンプル サイズ
1	lymphography	18	4	148	9.06×10^{8}	1.63×10^{-7}
2	Breast Cancer Wisconsin	9	2	683	2.00×10^{9}	3.42×10^{-7}
3	zoo	16	5	101	9.83×10^{5}	1.03×10^{-4}
4	Breast Cancer	9	2	277	3.33×10^{5}	8.33×10^{-4}
5	ClimateModel	18	2	540	5.24×10^{5}	1.03×10^{-3}
6	Image Segmentation	18	7	2310	1.84×10^{6}	1.26×10^{-3}
7	Congressional Voting Records	16	2	232	1.31×10^{5}	1.77×10^{-3}
8	Solar Flare	10	9	1389	3.73×10^{5}	3.72×10^{-3}
9	Tic-Tac-Toe	9	2	958	3.94×10^{4}	2.43×10^{-2}
10	Contraceptive Method Choice	9	3	1473	2.46×10^{4}	5.99×10^{-2}
11	Nursery	8	5	12960	6.48×10^{4}	2.00×10^{-1}
12	Hayes-Roth	4	3	132	5.76×10^{2}	2.29×10^{-1}
13	Car Evaluation	6	4	1728	6.91×10^{3}	2.50×10^{-1}
14	Balance Scale	4	3	625	1.88×10^{3}	3.33×10^{-1}
15	EEG	14	2	14980	3.28×10^{4}	4.57×10^{-1}
16	mux6	6	2	64	1.28×10^{2}	5.00×10^{-1}
17	threeOf9	9	2	512	1.02×10^{3}	5.00×10^{-1}
18	MONK's Problems	6	2	432	8.64×10^{2}	5.00×10^{-1}
19	parity5+5	10	2	1024	2.05×10^{3}	5.00×10^{-1}
20	mofn-3-7-10	10	2	1324	2.05×10^{3}	6.46×10^{-1}
21	LED Display Domain	7	10	3200	1.28×10^{3}	2.50
22	MAGIC Gamma Telescope	10	2	19020	2.05×10^{3}	9.29
23	HTRU2	8	2	17898	5.12×10^{2}	3.50×10
24	banknote authentication	4	2	1372	3.20×10	4.29×10

4　評価実験

4.1　実験手順

UCI リポジトリデータベース [17] から，表.1 に示される 24 個の学習データを用いて，従来の手法で学習した BNC と提案手法の CML スコアで学習した BNC の比較実験を行った．CML は同時確率分布の商の形で求められるが，一般に同時確率を十分な精度で推定するには，変数列のとりうる値のパターン数に対するサンプルサイズが十分大きい必要がある．BNC の構造によらずサンプルサイズの大小を比較する指標として，サンプルサイズを変数列のとりうる値のパターン数で割った標準化サンプルサイズを表.1 に載せた．また，deCampos ら [18] に従い，学習データに含まれる数値的データは，その中央値を区切りとして 2 値をとるカテゴリデータに変換した．各学習データで欠損値を含むデータ列は取り除いた．

比較する分類器の構造を次に示す．

- Naive Bayes
- TAN-LL：LL を最大にする TAN
- ANB-BDeu：BDeu を最大にする ANB

表 2　各ベイジアンネットワーク分類器の分類精度と，提案手法と他手法の比較検定結果（太字は最大の分類精度）

			Naive Bayes	TAN-LL	ANB-BDeu	GBN-BDeu	ANB-aCMDL	ANB-CMDL	GBN-CMDL	ANB-aCML	GBN-aCML
分類精度	1	lymphography	**0.8514**	0.7432	0.7838	0.7432	**0.8514**	**0.8514**	0.7500	0.7973	0.8176
	2	Breast Cancer Wisconsin	**0.9751**	0.9649	0.9722	0.9722	**0.9751**	**0.9751**	0.8660	**0.9751**	0.9707
	3	zoo	**0.9802**	0.9604	0.9505	0.9604	**0.9802**	**0.9802**	0.9604	0.9604	0.9604
	4	Breast Cancer	**0.7437**	0.7256	0.6895	0.7329	**0.7437**	**0.7437**	0.6047	0.7040	0.7184
	5	ClimateModel	0.9204	0.9315	0.8407	0.8889	0.9204	0.9204	**0.9407**	0.8722	0.8815
	6	Image Segmentation	0.7294	0.7511	0.8268	0.8156	0.7281	0.8095	0.8195	0.8333	**0.8338**
	7	Congressional Voting Records	0.9095	0.9440	0.9224	0.9526	0.9095	0.9095	**0.9698**	0.9353	0.9310
	8	Solar Flare	0.7811	0.7948	0.8236	**0.8431**	0.7970	0.7847	0.8294	0.7927	**0.8431**
	9	Tic-Tac-Toe	0.6921	0.7620	0.8737	0.8674	0.6848	0.8445	0.8831	0.9499	**0.9509**
	10	Contraceptive Method Choice	0.4671	0.4698	**0.4759**	0.4705	0.4312	0.4508	0.4372	0.4650	0.4752
	11	Nursery	0.9032	0.9250	0.9181	0.9318	0.9197	0.9468	**0.9612**	0.9462	0.9475
	12	Hayes-Roth	**0.7879**	0.6364	0.7652	0.5909	**0.7879**	**0.7879**	0.5909	0.5985	0.6540
	13	Car Evaluation	0.8571	0.9375	**0.9421**	0.9416	0.8571	**0.9421**	**0.9421**	**0.9421**	0.9416
	14	Balance Scale	**0.9152**	0.8624	**0.9152**	**0.9152**	**0.9152**	**0.9152**	**0.9152**	**0.9152**	**0.9152**
	15	EEG	0.5778	0.6306	0.6579	0.6897	0.6701	0.6529	0.6782	**0.6967**	0.6937
	16	mux6	0.5469	0.5781	0.4531	0.4531	0.5469	0.5313	0.3594	**1.0000**	**1.0000**
	17	threeOf9	0.8164	0.8516	0.9375	0.9668	0.8242	0.9219	0.9121	0.9863	**0.9883**
	18	MONK’s Problems	0.7500	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	19	parity5+5	0.3633	0.2998	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	20	mofn-3-7-10	0.8505	0.9260	0.9290	**1.0000**	0.9366	0.8505	0.8474	**1.0000**	**1.0000**
	21	LED Display Domain	0.7294	0.7306	0.7294	0.7294	0.7294	**0.7366**	**0.7366**	0.7294	0.7294
	22	MAGIC Gamma Telescope	0.7482	0.7768	0.7872	**0.7873**	0.7791	0.7807	0.7859	0.7866	0.7866
	23	HTRU2	0.8966	**0.9141**	**0.9141**	0.9112	0.9127	0.9112	0.9117	**0.9141**	**0.9141**
	24	banknote authentication	0.8433	**0.8819**	0.8812	0.8812	0.8812	0.8433	0.8433	0.8783	0.8724
		平均	0.7765	0.7916	0.8329	0.8352	0.8242	0.8371	0.8144	0.8616	**0.8677**
検定		z 値	-2.646	-2.9023	-2.4145	-1.6547	-1.792	-1.4077	-2.6506	-1.5903	-
		p 値	0.00402	0.00187	0.00798	0.04947	0.03673	0.07927	0.00402	0.05592	-

- ANB-aCMDL : aCMDL を最大にする ANB
- ANB-CMDL : CMDL を最大にする ANB
- ANB-CML : CML を最大にする ANB
- GBN-BDeu : BDeu を最大にする GBN
- GBN-CMDL : CMDL を最大にする GBN
- GBN-CML : CML を最大にする GBN

条件付き周辺尤度 CML スコアのハイパーパラメータは任意のデータ列 $\mathbf{x}$ に対して $N'_{i,\mathbf{x}} = 5/(r_i q_i)$ と設定した．6 番と 11 番を除く学習データでは，CMDL スコアのパラメータは ELR アルゴリズム [7] で推定した．6 番と 11 番の学習データでは，CMDL スコアのパラメータを ELR で推定した学習に 2 日以上もの時間がかかってしまうので，Grossman ら [6] に従い，式（5）の最尤推定量でパラメータを推定した．また，TAN-LL は Friedman ら [1] のアルゴリズムを用いて厳密に学習し，ANB-BDeu，ANB-aCMDL，GBN-BDeu は動的計画法 [19] を用いて厳密に学習した．ANB-CMDL，ANB-CML，GBN-CMDL，GBN-CML の学習には因数分解による厳密学習法を用いることができないため，Hill Climbing を用いて貪欲学習を行った．ANB-CMDL，

表 3　Naive Bayes，ANB-CMDL，GBN-aCML の分類に影響を及ぼすエッジ数（NEC）

		Naive Bayes	ANB-CMDL	GBN-aCML
1	lymphography	18.0	18.2	41.0
2	Breast Cancer Wisconsin	9.0	9.0	11.8
3	zoo	16.0	16.0	51.5
4	Breast Cancer	9.0	11.0	5.4
5	ClimateModel	18.0	18.0	67.2
6	Image Segmentation	18.0	27.5	45.4
7	Congressional Voting Records	16.0	16.0	45.9
8	Solar Flare	10.0	14.1	1.0
9	Tic-Tac-Toe	9.0	14.2	19.1
10	Contraceptive Method Choice	9.0	11.0	2.4
11	Nursery	8.0	21.6	17.7
12	Hayes-Roth	4.0	4.0	4.8
13	Car Evaluation	6.0	9.1	7.2
14	Balance Scale	4.0	4.0	4.0
15	EEG	14.0	41.6	44.1
16	mux6	6.0	6.1	14.9
17	threeOf9	9.0	15.8	27.3
18	MONK’s Problems	6.0	7.0	5.1
19	parity5+5	10.0	14.0	13.3
20	mofn-3-7-10	10.0	10.0	34.3
21	LED Display Domain	7.0	10.9	7.0
22	MAGIC Gamma Telescope	10.0	24.0	23.9
23	HTRU2	8.0	15.8	15.3
24	banknote authentication	4.0	5.0	7.0
	平均	9.9	14.3	21.5

ANB-CML のいずれも，Hill Climbing における初期構造は Naive Bayes，TAN-LL，ANB-BDeu とし，アルゴリズムが出力した構造の中で最もスコアの大きいものを学習構造とした．同様に，GBN-CMDL，GBN-CML の Hill Climbing における初期構造は ANB-CMDL，ANB-CML，GBN-BDeu とした．また，Carvalho ら [16] に従い，aCLL スコアの擬似サンプル数 N' は 5 とし，$(J_{d,1}, \cdots, J_{d,r_0}) \sim (1, \cdots, 1, N)$ と仮定した．ANB-BDeu と GBN-BDeu の ESS は 5 とした．

各ベイジアンネットワーク分類器（BNC）の学習と分類に対して 10 分割交差検証を行った．10 分割交差検証では，学習データを 10 分割し，そのうち一つをテストデータ，残りの 9 つのデータを学習データとした検証を，分割されたそれぞれをテストデータとした時の計 10 回行う．テストデータの説明変数のデータを BNC に与え，式（6）で正しく分類できた割合を測定し，10 分割交差検証におけるその平均を分類精度として測定した．表.2 の「分類精度」に，各学習データに対する各ベイジアンネットワーク分類器（BNC）の分類精度を載せた．分類において，パラメータは式（3）で推定した．太字は各学習データにおいて最大の分類精度である．また，提案手法の有意性を示すため，Grossman ら [6]，Carvalho ら [16] と同様に，GBN-CML と各比較手法に対してウィルコクソンの符号順位検定を行った．帰無仮説を GBN-CML と比較手法の分類精度に差がないとし，対立仮説を GBN-CML の方が比較手法より分類精度が高いとした．検定結果として z 値と p 値を表.2 の「検定」に載せた．さらに，各検証で学習した Naive Bayes，

表 4　ベイジアンネットワーク（GBN）の目的変数の親変数数（NPC）と目的変数の子変数数（NCC）

	構造	GBN-BDeu		GBN-CMDL		GBN-aCML	
		NPC	NCC	NPC	NCC	NPC	NCC
1	lymphography	1.8	6.5	0.3	4.5	0.2	16.5
2	Breast Cancer Wisconsin	0.8	7.2	3.0	0.0	0.0	7.6
3	zoo	4.2	2.5	0.0	5.6	0.0	15.7
4	Breast Cancer	1.4	0.6	4.0	0.0	0.2	3.5
5	ClimateModel	2.2	4.7	0.4	3.6	0.1	14.0
6	Image Segmentation	1.5	10.2	0.0	18.0	0.0	15.9
7	Congressional Voting Records	3.1	3.2	0.0	2.1	0.1	13.6
8	Solar Flare	0.4	0.6	0.0	5.0	0.1	0.9
9	Tic-Tac-Toe	3.0	2.0	3.6	2.4	0.0	7.5
10	Contraceptive Method Choice	2.0	0.0	4.0	1.8	1.3	1.0
11	Nursery	4.0	4.0	0.0	8.0	0.0	8.0
12	Hayes-Roth	3.0	0.0	3.0	0.0	1.2	1.8
13	Car Evaluation	2.0	3.0	0.0	6.0	1.6	3.4
14	Balance Scale	0.3	3.7	0.0	4.0	0.1	3.9
15	EEG	0.2	8.8	0.0	11.9	0.0	11.4
16	mux6	4.4	0.9	0.2	1.6	0.0	5.7
17	threeOf9	5.0	2.5	0.0	7.0	0.5	8.3
18	MONK’s Problems	3.0	0.0	3.0	0.0	0.0	3.0
19	parity5+5	0.5	0.9	0.5	0.9	0.2	5.2
20	mofn-3-7-10	7.0	0.0	0.0	7.0	0.0	9.7
21	LED Display Domain	0.7	6.3	0.0	7.0	0.7	6.3
22	MAGIC Gamma Telescope	0.1	6.0	0.0	9.0	0.0	9.0
23	HTRU2	0.8	5.2	0.7	5.4	0.7	6.0
24	banknote authentication	0.0	2.0	0.0	3.9	0.0	3.0
	平均	2.1	3.4	0.9	4.8	0.3	7.5

ANB-CMDL，GBN-aCML に対し，分類に影響を及ぼすエッジ数（Number of Edges related to Classifier: NEC）を測定し，10 回の検証の平均値を表.3 に載せた．同様に，各検証で学習した GBN に対し，目的変数の親変数数（Number of Parents of the Class variable: NPC），目的変数の子変数数（Number of Children of the Class variable: NCC）をそれぞれ測定し，10 回の検証の平均値を表.4 に載せた．表.5 には，各手法の 10 分割交差検証における 1 回あたりの平均学習時間を載せた．

4.2　結果と考察

表.2 の検定結果より，提案手法である GBN-aCML は Naive Bayes や生成モデルである TAN-LL，ANB-BDeu，GBN-BDeu，従来の識別モデルである ANB-aCMDL，GBN-CMDL よりも有意水準 0.05 のもとで有意に分類精度が高かった．近似的に探索された提案手法が，厳密に探索した生成モデルより分類精度が高いことは特筆すべきである．従来の識別モデルである ANB-CMDL に対しては提案手法との有意差は認められなかったが，全学習データの内，標準化サンプルサイズの大きい半分の学習データである 13 番から 24 番の学習データに対して検定を行なった結果，有意水準 0.05 のもとで提案手法は ANB-CMDL よりも有意に分類精度が高かった．この理由として，提案スコアが CMDL スコアよりも条件付き周辺尤度を精度良く近似しているた

表 5　各 BNC 構造の学習時間（秒）（ハイフンは学習時間が 6 時間以上であることを示す）

		ANB-BDeu	GBN-BDeu	ANB-aCMDL	ANB-CMDL	GBN-CMDL	ANB-aCML	GBN-aCML
1	lymphography	65.21	129.55	57.19	225.40	65.60	8.70	4.43
2	breast	0.29	0.49	0.67	133.81	313.21	0.94	0.62
3	zoo	5.34	10.93	5.91	7.27	12.42	7.05	3.35
4	Cancer	0.12	0.19	0.58	22.30	12.00	0.39	0.36
5	ClimateModel	121.20	235.83	92.56	387.30	512.72	20.14	12.71
6	ImageSegmentation	116.45	219.61	87.72	-	-	220.77	78.34
7	Congressional	7.64	15.55	6.53	181.38	59.70	5.53	2.84
8	Flare	0.23	0.37	1.80	2153.49	1085.32	8.25	6.42
9	TicTac	0.17	0.27	0.64	241.26	166.65	1.66	0.89
10	cmc	0.15	0.23	0.78	118.07	136.01	2.15	1.58
11	Nursery	0.44	0.61	1.31	-	-	23.68	13.81
12	Hayes-Roth	0.02	0.03	0.68	1.11	1.39	0.07	0.09
13	Car	0.06	0.08	0.88	178.85	206.99	0.79	0.92
14	Balance	0.04	0.05	0.69	50.44	26.92	0.15	0.12
15	EEG	8.43	13.56	5.07	758.60	4432.25	211.76	148.10
16	mux6	0.04	0.05	0.54	0.05	0.16	0.12	0.11
17	threeOf9	0.08	0.14	0.59	30.55	47.10	1.27	0.86
18	monk1	0.05	0.06	0.55	3.57	4.33	0.17	0.24
19	parity5+5	0.20	0.27	0.62	1.69	6.30	1.07	1.48
20	mofn-3-7-10	0.18	0.26	0.63	53.76	279.85	9.28	5.51
21	led7	0.09	0.13	1.86	121.65	94.43	2.97	2.47
22	magic	0.54	0.63	0.95	694.34	1263.60	34.33	32.39
23	HTRU2	0.26	0.30	0.74	454.62	340.55	10.98	6.95
24	banknote	0.03	0.04	0.54	5.89	6.13	0.16	0.23
	平均	13.64	26.22	11.25	264.79	412.44	23.85	13.53

め，提案手法がより最適な分類器としての構造を学習することが考えられる．さらに，表.3 に示される提案手法と ANB-CMDL の分類に影響を及ぼすエッジ数（NEC）に対してウィルコクソンの符号順位検定を行ったところ，提案手法の NEC の方が有意水準 0.05 のもとで有意に多かったため，提案スコアは有意差がようやく認められる程度のエッジの付与に注目し，分類精度を追求していると考えられる．

一方で，標準化サンプルサイズの小さい 1 番から 5 番の学習データでは，提案手法は単純な構造である Naive Bayes より分類精度が低い．表.3 より，学習データ 5 番における Naive Bayes の分類に影響を及ぼすエッジ数（NEC）は 18.0 であるが，提案手法の NEC は 67.2 と非常に多い．このように，標準化サンプルサイズが小さい時に提案手法が Naive Bayes よりも分類精度が低い理由として，提案手法はサンプルサイズが小さい場合に過学習して誤ったエッジをつけやすいことが考えられる．一方，学習データ 2 番，3 番，5 番における ANB-CMDL の NEC はそれぞれ Naive Bayes の NEC と一致しており，これはこの時 ANB-CMDL が 10 分割交差検証で学習した 10 個の構造が全て Naive Bayes と一致することを意味する．このように，ANB-CMDL はサンプルサイズが小さくてもエッジを余分に付けにくいため，標準化サンプルサイズが小さい時は

ANB-CMDL の方が提案手法より分類精度が高くなりやすいと考えられる．

また，表.4 より，学習データ 2 番と 4 番における GBN-CMDL と，学習データ 12 番における GBN-CMDL，GBN-BDeu では，目的変数の親変数数（NPC）が多く目的変数の子変数数（NCC）が 0 となる構造を学習しており，分類精度が他の BNC より低いことがわかる．この理由として，2 章で述べたように，このような構造では分類に影響を与えるパラメータが少なくなり過ぎてしまい，推定精度が極度に悪化してしまう．一方，提案手法である GBN-aCML はどの学習データにおいても NCC が 0 となることはなく，ウィルコクソンの符号順位検定を行った結果，GBN-BDeu と GBN-CMDL よりも有意水準 0.05 のもとで有意に NPC が少なく，NCC が多かった．したがって，GBN-aCML では上記の分類精度低下の問題は比較的起こりにくいと考えられるため，ANB-aCML よりも，とりえる全ての構造を候補とする GBN-aCML を用いた方が分類精度の高い分類器構造を学習すると考えられる．

次に，表.5 の結果を用いてウィルコクソンの符号順位検定を行ったところ，提案手法は有意水準 0.05 のもとで ANB-CMDL と GBN-CMDL より有意に計算時間が少なかった．この理由として，CMDL スコアでは探索する候補構造ごとに ELR アルゴリズムによるパラメータ推定を行わなければならず，推定値の最局所的最適解を得るのに大きな時間を要することが考えられる．

以上の実験結果は次のようにまとめられる．

1. 提案手法は従来の生成モデルの BNC よりも有意に分類精度が高い．
2. 標準化サンプルサイズが大きいときは，提案手法は従来の識別モデルの BNC より分類精度が有意に高い．
3. 従来のスコアである BDeu や CMDL より，aCML を用いて学習した BNC の方が，有意に NPC が少なく NCC が多い構造を学習するため，分類に影響するパラメータ数減少による分類精度悪化の問題が起きにくい．
4. 提案手法は従来の識別モデルの BNC よりも有意に計算時間が短い．

このように，提案手法の有意性を示せた．

5　むすび

本論文では，条件付き周辺尤度（Conditional Marginal Likelihood: CML）を厳密に定義し，CML を効率的かつ精度良く近似したと考えられる approximate CML（aCML）を識別モデルの学習スコアとして提案した．aCML スコアは，従来の識別モデルのスコアと異なり，数学的なスコアの意味が明確であり，CML を直接近似するため，分類精度の向上が期待できる．さらに，従来の識別モデルのスコアで誤差の原因となってきた勾配法によるパラメータ推定が必要でないため，安定した分類精度と，計算量の大幅な減少が期待できる．

リポジトリ学習データを用いて実験を行なった結果，近似的に学習された提案手法が，厳密に学習した従来の生成モデルの BNC よりも分類精度が有意水準 0.05 のもとで有意に高かった．また，提案手法は比較的複雑な構造を学習しやすく，サンプルサイズが大きい時には，従来の識別モデルより分類精度が有意に高かった．また，提案手法は目的変数の子変数が多く，親変数が少

ない構造を学習する傾向にあるため，GBN の問題であった分類精度の著しい悪化を防ぐことができると考えられる．さらに，提案手法の計算時間は，従来の識別モデルの学習スコアを用いた学習と比較して大幅に減少した．今後の課題として，より CML の近似精度の高い学習スコアの考案や，それらを用いた効率的な探索アルゴリズムの考案を行う．

参考文献

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

[2] J. Rissanen, Stochastic Complexity in Statistical Inquiry Theory, World Scientific Publishing Co., Inc., 1989.

[3] D.M. Chickering, “Learning Bayesian Networks is NP-Complete,” pp.121–130, Springer, 1996.

[4] M. Barlett and J. Cussens, “Advances in Bayesian Network Learning Using Integer Programming,” Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, pp.182–191, 2013.

[5] A.Y. Ng and M.I. Jordan, “On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes,” Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, pp.841–848, MIT Press, 2001.

[6] D. Grossman and P. Domingos, “Learning Bayesian Network classifiers by maximizing conditional likelihood,” Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004, pp.361–368, 2004.

[7] R. Greiner and W. Zhou, “Structural Extension to Logistic Regression: Discriminative Parameter Learning of Belief Net Classifiers,” Eighteenth National Conference on Artificial Intelligence, pp.167–173, 2002.

[8] T. Roos, H. Wettig, P. Grünwald, P. Myllymäki, and H. Tirri, “On Discriminative Bayesian Network Classifiers and Logistic Regression,” Machine Learning, vol.59, no.3, pp.267–296, 2005.

[9] D. Heckerman, D. Geiger, and D.M. Chickering, “Learning Bayesian Networks: The Combination of Knowledge and Statistical Data,” Machine Learning, vol.20, no.3, pp.197–243, 1995.

[10] M. Ueno, “Learning Networks Determined by the Ratio of Prior and Data,” Proceedings of Uncertainty in Artificial Intelligence, pp.598–605, 2010.

[11] M. Ueno, “Robust learning Bayesian networks for prior belief,” Proceedings of Uncertainty in Artificial Intelligence, pp.689–707, 2011.

[12] A.M. Carvalho, T. Roos, A.L. Oliveira, and P. Myllymäki, “Discriminative Learning of Bayesian Networks via Factorized Conditional Log-Likelihood,” Journal of Machine

Learning Research, vol.12, pp.2181–2210, 2011.

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

[14] M. Minsky, "Steps toward Artificial Intelligence," Proceedings of the IRE, vol.49, pp.8–30, 1961.

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

[16] A.M. Carvalho, P. AdÃčo, 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.

[17] M. Lichman, "UCI machine learning repository," 2013. http://archive.ics.uci.edu/ml

[18] C.P. deCampos, M. Cuccu, G. Corani, and M. Zaffalon, "Extended Tree Augmented Naive Classifier," pp.176–189, Springer International Publishing, Cham, 2014.

[19] T. Silander and P. Myllymäki, "A Simple Approach for finding the Globally Optimal Bayesian Network Structure," Proceedings of Uncertainty in Artificial Intelligence, pp.445–452, 2006.