電気通信大学情報理工学部
先端工学基礎課程卒業論文

Augmented Naive Bayes Classifier の
大規模構造学習

平成 31 年 2 月 28 日

情報数理工学コース

学籍番号 1525025

菊谷成慎

指導教員 植野真臣

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で表しノード間の依存関係を非循環有向グラフ（Directed Acyclic Graph: DAG）で表現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は，確率構造に DAG を仮定することにより，同時確率分布を条件付き確率の積に分解する．ベイジアンネットワークの構造は一般にデータから推定する必要があり，これをベイジアンネットワークの構造学習という．ベイジアンネットワークの構造学習法として，漸近一致性を有する学習スコアを用いて，候補構造から最適なスコアを持つ構造を探索する厳密解探索アプローチが従来から用いられている．また，一般に学習スコアとして周辺尤度（Marginal Likelihood: ML）が用いられてきた．

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（Bayesain Network Classifier: BNC）は，離散変数を扱う分類器として知られている [1]．BNC は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ほうが，通常のベイジアンネットワークを構造学習した生成モデルよりも分類精度が高いことが報告されている [2, 3, 4, 5]．しかし，近年，Sugahara ら（2018）[6] は，データが十分大きいならば，生成モデルのほうが分類精度が高いことを示した．しかし，データが少ないときに目的変数の親変数が多い構造をとる場合，パラメータ数が指数的に増加し，学習データが過疎になり，正しいパラメータ学習ができず，精度が著しく低下してしまうと指摘した．この問題を解決するために，彼らは目的変数は親変数を持たず，全ての説明変数を子に持つ構造を仮定した Augmented Naive Bayes（ANB）[1] を厳密学習する手法を提案している．これにより，データが少ない場合も高い分類精度を得られることを経験的に示した．

しかし，厳密解探索アプローチによる厳密学習は候補構造の数が指数的に増加する NP 困難問題である [7]．探索手法として，動的計画法 [8, 9, 10, 11, 12]，A*探索 [13]，整数計画法 [14] などが提案されてきたが，未だに数十変数程度の構造学習が限界あり、Sugahara らの手法は多くの変数に対応できない．

一方，因果モデル分野では，漸近一致性は持たないが，より計算効率の高い制約ベースの構造学習法が提案されている．完全無向グラフに，二ノード間の条件付き独立性検定

（Conditional Independence test: CI テスト）を適用して学習される無向グラフに対し，オリエンテーションルール [15] による辺の方向付けを行うことで DAG を学習する．制約ベース手法では，PC アルゴリズム [16]，MMHC アルゴリズム [17]，RAI アルゴリズム [18] などが提案されており，RAI アルゴリズムが最も高精度であると知られている．従来の CI テストでは漸近一致性を持たないことが問題であったが，名取ら（2018）は，RAI アルゴリズムの CI テストに Bayes factor を用いることで漸近一致性を有しつつ，1000 変数以上の大規模構造学習を実現している [19]．さらに，本田ら（2019）は，推移性を組み込むことで CI テストの回数を削減し，2000 変数程度の大規模構造学習を実現している [20]．

そこで本論文では，本田らの手法を ANB 学習に拡張させ，従来よりも大規模な BNC を学習することを目指す．さらに，リポジトリデータベースを用いた評価実験により，提案手法が従来手法よりも，大規模な説明変数を持つ分類問題で精度が高いことを示す．

2　ベイジアンネットワーク

ベイジアンネットワークは，確率変数をノードとし，ノード間の依存関係を非循環有効グラフ（Directed Acyclic Graph: DAG）と各ノードの条件付き確率で表現する確率的グラフィカルモデルである．今，$\mathbf{X} = \{X_0, X_1, \ldots, X_n\}$ を離散確率変数集合とし，各変数 X_i は r_i 個の状態集合 $\{1, \ldots, r_i\}$ から一つの値 k を取る（$X_i = k$ と書く）とする．このとき，ベイジアンネットワークの構造 G において，各ノード X_i の親ノード集合を Π_i としたときの同時確率分布 $P(X_1, \ldots, X_n \mid G)$ は以下のように表現できる．

$$P(X_0, X_1, \ldots, X_n \mid G) = \prod_{i=0}^{n} P(X_i \mid \Pi_i, G). \tag{1}$$

ここで，θ_{ijk} を Π_i が j 番目のパターンを取るとき ($\Pi_i = j$ 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，条件付きパラメータ集合 $\Theta = \{\theta_{ijk}\}, (i = 0, \ldots, n; j = 1, \ldots, q_i; k = 1, \ldots, r_i)$ とする．パラメータの事前分布としてディリクレ分布 $P(\Theta)$ を仮定すると，事後分布 $P(\Theta \mid D, G)$ が得られる．

$$P(\Theta) = \prod_{i=0}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\sum_{k=1}^{r_i} \alpha_{ijk})}{\prod_{k=1}^{r_i} \Gamma(\alpha_{ijk})} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1}. \tag{2}$$

$$P(\Theta \mid D, G) \tag{3}$$
$$= \prod_{i=0}^{n} \prod_{j=1}^{q_i} \frac{\Gamma\{\sum_{k=1}^{r_i}(\alpha_{ijk} + N_{ijk})\}}{\prod_{k=1}^{r_i} \Gamma(\alpha_{ijk} + N_{ijk})} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}+N_{ijk}-1}.$$

ここで，N_{ijk} は，X_i の親変数集合 Π_i が j 番目のパターンを取った時の $X_i = k$ となるデータの頻度を表し，$n_{ij} = \sum_{k=1}^{r_i} n_{ijk}$ を表す．また，$\boldsymbol{\alpha} = \{\alpha_{ijk}\}, (i = 0, \ldots, n; j = 1, \ldots, q_i; k = 1, \ldots, r_i)$ はディリクレ事前分布のハイパーパラメータであり，$\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ である．

ベイジアンネットワークでは, パラメータ推定値として, 期待事後確率推定値（Expected a Posteriori: EAP）が最もよく用いられる．変数集合 $\mathbf{X}$ に対する N 個のデータを $\mathbf{D} = \{D_1, \ldots, D_N\}$ とすると，EAP は式 (3) に期待値をとることで以下を得る．

$$\hat{\theta}_{ijk} = \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}. \tag{4}$$

構造が定まれば，式 (4) からパラメータを推定できるが，最適な構造もデータから推定する必要がある．これをベイジアンネットワークの構造学習という．構造学習法として，候補構造から最適なスコアを持つ構造を探索する厳密解探索アプローチが従来から用いられている．一般に学習スコアとして周辺尤度 $P(\mathbf{D} \mid G)$ が用いられる．周辺尤度は，式 (3) の事後分布からパラメータ推定値を周辺化することで閉形式で表せる．

$$P(\mathbf{D} \mid G) = \prod_{i=0}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}. \tag{5}$$

近年では，$\alpha_{ijk} = \alpha/(r_i q_i)$ とした Bayesian Dirichlet equivalent uniform (BDeu) が最も用いられる [21][5]．ここで，α は Equivalent Sample Size (ESS) と呼ばれる事前知識の重みを示す疑似サンプルである．BDeu は，以下の漸近一致性を持つことが知られている [8].

定理 2.1　データ数 $N \to \infty$ のとき，BDeu を最大化するベイジアンネットワークの同時確率分布は真の分布に近づく．

証明は Koller ら [8] を参照してほしい.

しかし，厳密解探索アプローチは NP 困難であり，変数数に対して計算時間が爆発的に増加してしまう．厳密解を効率的に探索するために，動的計画法 [8, 9, 10, 11, 12]，A*探

索 [13]，整数計画法 [14] などが提案されている．しかし，最先端手法 [14] でさえ 60 変数程度の構造学習が限界であり，大規模構造を学習することができない．

3　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（Bayesian Network Classifier: BNC）と呼び，高い分類精度を持つことが知られている [1]．今，X_0 を目的変数とし，$X_1, \ldots, X_n$ を説明変数とした BNC を考える．説明変数のデータ $\mathbf{e} = \langle x_1, \ldots, x_n \rangle$ 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．

$$
\begin{aligned}
\hat{c} &= \arg\max_{c \in \{1,\ldots,r_0\}} P(c \mid x_1, \ldots, x_n, G) \qquad (6)\\
&= \arg\max_{c \in \{1,\ldots,r_0\}} \prod_{i=0}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} (\theta_{ijk})^{1_{ijk}} \\
&= \arg\max_{c \in \{1,\ldots,r_0\}} \prod_{j=1}^{q_0} \prod_{k=1}^{r_0} (\theta_{0jk})^{1_{0jk}} \times \prod_{i:X_i \in \mathbf{C}} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} (\theta_{ijk})^{1_{ijk}}
\end{aligned}
$$

ここで，1_{ijk} はあるデータ $\mathbf{e}$ が与えられた時，$X_i = k$ かつ $\Pi_i = j$ の時 1 をとり，それ以外の場合は 0 をとる変数である．さらに，$\mathbf{C}$ は目的変数の子集合である．式 (6) より，目的変数 X_0 の親集合，子集合，X_0 と子を共有している変数集合の和集合のみで推定値 $\hat{c}$ を求めることができる．また，この集合を X_0 のマルコフブランケット（Markov Blanket: MB）と呼ぶ．

BNC の構造学習は通常，ベイジアンネットワークと同様にデータから学習する．制約を持たない一般的なベイジアンネットワークを用いた BNC を General Bayesian Network（GBN）と呼ぶ（図 1（a））．しかし，BDeu などで学習された GBN は目的変数が子変数を持たず，親変数を多く持つ構造をとることがある．このような構造では，式 (6) が目的変数の条件付きパラメータ θ_{0jk} のみに依存する．さらに，目的変数の親変数集合のとりうる値のパターン数 q_0 が大きくなり，パラメータの推定に用いられるデータ N_{0jk} がスパースになる．分類に用いられるデータが少なくなってしまうことから，分類精度が

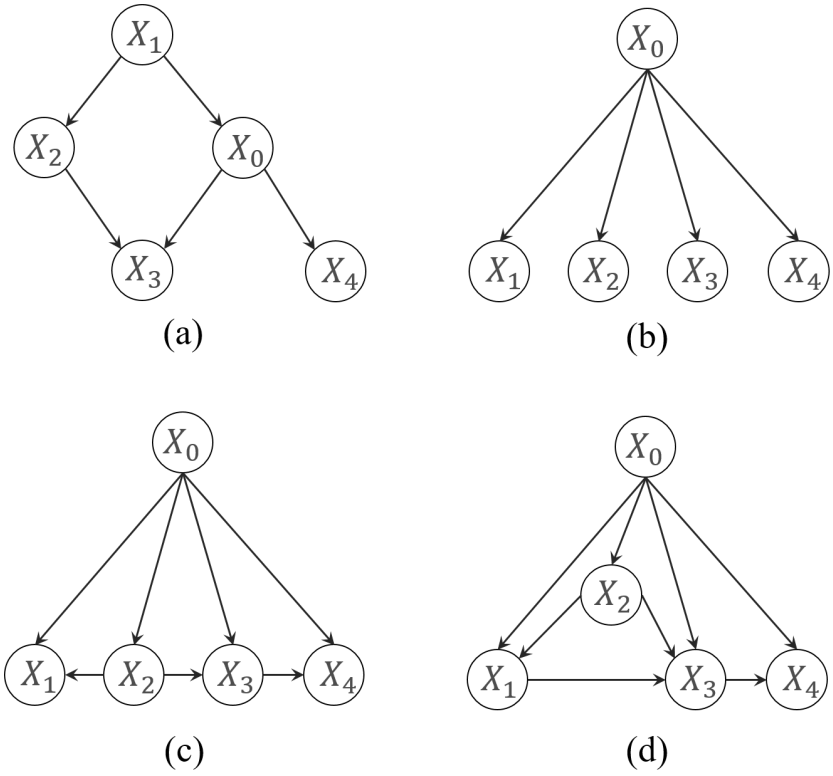

図 1　(a) GBN の例; (b) Naive Bayes; (c) TAN の例; (d) ANB の例

著しく悪くなる可能性が高まる [22][23]．この問題を回避できる BNC として，目的変数が全ての説明変数を子に持ち，説明変数間が独立であると仮定する Naive Bayes[24]（図 1（b)），目的変数が全ての説明変数を子に持ち，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（c)）などが提案されている．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ることが知られてる [1][25]．また，Naive BayBayes や TAN を一般化した，より表現力の高いモデルとして，目的変数が全ての説明変数を子に持つことのみ仮定する Augmented Naive Bayes（ANB）[1]（図 1（d)）が知られている．

周辺尤度を最大にする BNC は全変数の同時確率をモデル化した生成モデルであるが，説明変数を所与とした目的変数の条件付き尤度をモデル化する識別モデルのほうが，漸近的な分類精度が高いことが報告されていた [4][3]．しかし，識別モデルとしての BNC が生成モデルとしての BNC よりも高い分類精度を得られる根拠が理論的に示されていな

い．また，比較実験において，生成モデルとしての BNC は厳密学習できるにもかかわらず近似学習を用いている．そこで，Sugahara ら [6] は ML で厳密学習した BNC は，識別モデルの BNC よりも生成モデルの分類精度が低いとは限らないことを示した．しかし，先述したように生成モデルの BNC は目的変数が親を多く持つ構造を学習した場合，データ不足により分類精度が著しく低下してしまう．この問題を解決するため，彼らはこれまで識別モデルとして扱われてきた ANB を生成モデルとして厳密学習する手法を提案した．結果は，提案手法がデータが少ない場合でも安定した分類精度を得ることができ，識別モデルの BNC より有意に分類精度が高いことを示した．しかし，厳密学習は数十変数の学習が限界であり，変数数が多いデータに対して用いることができない．

そこで本論文では，BNC の大規模構造学習法を提案する．

4　制約ベース厳密構造学習

4.1　Bayes factor を用いた制約ベース厳密学習

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

(1) 完全無向グラフを生成する．
(2) (1) で生成された完全無向グラフに対し条件付き独立検定（Conditional Independence test: CI テスト）によりエッジを削除する．
(3) (2) で得られた無向グラフに対してオリエンテーションルール [15] を用いて方向付けを行う．

一般に，制約ベースアプローチの学習精度は CI テストの精度に依存し，学習速度は学習に要する CI テストの回数に依存する．

制約ベースアプローチとして，PC アルゴリズム [16]，MMHC アルゴリズム [17]，RAI アルゴリズム [18] が提案されてきた．しかし，これらのアルゴリズムでは χ^2 検定，G^2 検定，条件付き相互情報量などを CI テストに用いるため，漸近一致性を持たない．一方

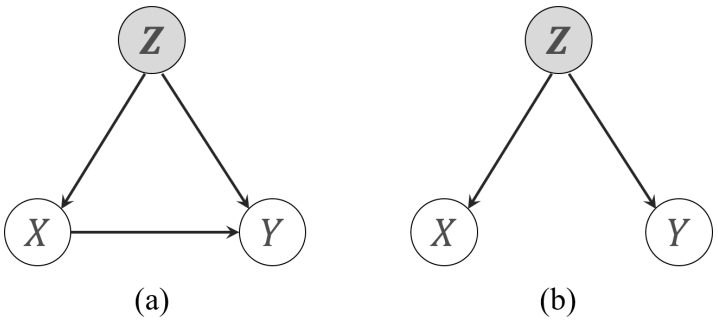

図 2　(a) 従属モデル G_1;　(b) 独立モデル G_2

で，Steck らは，二変数間が従属・独立モデルの周辺尤度の比による Bayes factor を用いた漸近一致性を有する CI テストを提案した [26]．例として，ベイジアンネットワークのノード X,Y において X と Y 間について各ノードの共通親ノード集合 $\mathbf{Z}$ をとしたときの従属なモデルを G_1，独立モデルを G_2 とし，それぞれ図 2 の（a)，(b）に示す．このときの Bayes factor を $\mathrm{BF(X, Y \mid \mathbf{Z})}$ とすると，対数 Bayes factor は，

$$\log \mathrm{BF}(X, Y \mid \mathbf{Z}) = \log \frac{P(\mathbf{D} \mid G_1, \boldsymbol{\alpha})}{P(\mathbf{D} \mid G_2, \boldsymbol{\alpha})} \tag{7}$$

と表される．ここで，$P(\mathbf{D} \mid G_1, \boldsymbol{\alpha})$，$P(\mathbf{D} \mid G_2, \boldsymbol{\alpha})$ は式（5）の BDeu を用いる．Bayes factor を用いた CI テストでは，対数 Bayes factor が 0 以上か否かで従属，独立を判断する．しかし，Steck らはこれをベイジアンネットワーク学習に用いていない．

そこで，名取らは Bayes factor を用いた CI テストが漸近的に真の条件付き独立性を判定できることを示した [27][19].

定理 4.1　データ数 $N \to \infty$ のとき，

(1)　真の構造が $\mathbf{Z}$ を所与として X と Y が条件付き独立でないとき，

$$\log BF(X, Y \mid \mathbf{Z}) > 0.$$

(2)　真の構造が $\mathbf{Z}$ を所与として X と Y が条件付き独立のとき，

$$\log BF(X, Y \mid \mathbf{Z}) < 0.$$

証明は名取ら（2018）[19] を参照してほしい．さらに，Bayes factor を用いた CI テストを RAI アルゴリズムに組み込んだ手法を提案し，漸近一致性を有しつつ 1000 変数を超える大規模構造学習を実現した．

4.2 推移性を用いた厳密アルゴリズム

前述のように、制約ベースの学習速度はＣＩテストの回数に依存する．一般にＣＩテストの回数は学習のできる限り早期にエッジを削除するほど削減できる．そのために以下の推移性を用いる手法が提案されている．

定理 4.2 $G = (\mathbf{V}, \mathbf{E})$ を DAG とし，$X, Y \in \mathbf{V}$ で，Y は X の非子孫とする．このとき，$A \in \mathbf{V} \setminus (\{X, Y\} \cup \mathbf{Pa}(X, G) \cup \mathbf{W})$ とすると，以下が成り立つ．

$$\begin{aligned} & X \perp Y \mid \mathbf{Pa}(X, G) \\ & \quad \rightarrow X \perp A \mid \mathbf{Pa}(X, G) \textit{ or } A \perp Y \mid \mathbf{Pa}(X, G) \end{aligned} \tag{8}$$

ここで，$\mathbf{W}$ は X の子孫であり，X と Y が合流結合するノードとその子孫からなるノード集合を表し，$\mathbf{Pa}(X, G)$ は X の G における親ノード集合を表す．証明は本田ら（2019）[20] を参照してほしい．定理 4.2 よりある二変数の条件付き独立性からその各変数と他変数との条件付き独立性の少なくとも一つを保証できる．これを利用することで，二変数の条件付き独立性から少なくとも一つの条件付き独立が保証される二組のエッジの CI テストを列挙できる．列挙されたＣＩテストを優先して行うことでより学習の早期にエッジを削除できる．本田らは推移性によるエッジの削除法を Bayes factor を用いた RAI アルゴリズムに組み込むことで CI テストの回数を削減し，2000 変数程度の大規模構造を実現した．

本論文では，本田ら [20] の手法を用いることで，従来よりも大規模な BNC の構造学習の実現を目指す．

5 提案手法

本章では，本田らの手法を ANB 学習に拡張させる方法を示す．まず，RAI アルゴリズムの基本的な動作を紹介する．今，グラフを $G = (\mathbf{V}, \mathbf{E})$ と表し，$\mathbf{V}, \mathbf{E}$ はそれぞれ G に

Algorithm 1 CI test

function CI-TEST($n_z, G_s, \mathbf{G}_{ex}, G_{all}, X_0, \mathbf{D}$)
n_z: CI テストの次数
$G_s = (\mathbf{V}_s, \mathbf{E}_s)$: 入力グラフ
$\mathbf{G}_{ex}$: 部分グラフ集合
$G_{all} = (\mathbf{V}, \mathbf{E})$: 全体グラフ
X_0: 目的変数のノード
// CI テストによるエッジの削除
  for $G_{ex} = (\mathbf{V}_{ex}, \mathbf{E}_{ex}) \in \mathbf{G}_{ex}$ **do**
    for $X \in \mathbf{V}_s \setminus \{X_0\}$, $Y \in \mathbf{V}_{ex} \setminus \{X_0\}$ **do**
      for $\mathbf{Z} \subseteq \mathbf{Pa}_p(X, G_s) \cup \mathbf{Pa}(X, G_{ex}) \setminus \{Y\}$ **do**
        if $|\mathbf{Z}| = n_z$ かつ $\log \mathrm{BF}(X, Y \mid \mathbf{Z}) < 0$ **then**
          $\mathbf{E}_{all} \leftarrow \mathbf{E}_{all} \setminus \{E_{XY}\}$ ▷ E_{XY}: XY 間のエッジ
          //推移性によるエッジの削除
          TRANSITIVE_CUT($G_s, G_{all}, X, Y, \mathbf{Z}, X_0, \mathbf{D}$)
        end if
      end for
    end for
  end for
  for $X \in \mathbf{V}_s \setminus \{X_0\}$, $Y \in \mathbf{V}_s \setminus \{X_0\}$ **do**
    for $\mathbf{Z} \subseteq \mathbf{Pa}_p(X, G_s) \cup \mathbf{Pa}(X, G_{ex}) \setminus \{Y\}$ **do**
      if $|\mathbf{Z}| = n_z$ かつ $\log \mathrm{BF}(X, Y \mid \mathbf{Z}) < 0$ **then**
        $\mathbf{E}_{all} \leftarrow \mathbf{E}_{all} \setminus \{E_{XY}\}$, $\mathbf{E}_s \leftarrow \mathbf{E}_s \setminus \{E_{XY}\}$
        //推移性によるエッジの削除
        TRANSITIVE_CUT($G_s, G_{all}, X, Y, \mathbf{Z}, X_0, \mathbf{D}$)
      end if
    end for
  end for
  return (G_s, G_{all})
end function

含まれるノード集合，エッジ集合を表す．ここで，G は有向エッジと無向エッジを併せ持つとする．また，$G_{ex} = (\mathbf{V}_{ex}, \mathbf{E}_{ex})$ を RAI アルゴリズムによって分割された部分グラフとする．

Algorithm 2 Edge cutting with transitivity

function TRANSITIVE_CUT($G_s, G, X, Y, \mathbf{Z}, X_0, \mathbf{D}$)
  $G_s = (\mathbf{V}_s, \mathbf{E}_s)$: 入力グラフ
  $G = (\mathbf{V}, \mathbf{E})$: 全体グラフ
  $X, Y, \mathbf{Z}$: $X \perp Y \mid \mathbf{Z}$ となる二ノード X, Y とノード集合 $\mathbf{Z}$
  X_0: 目的変数のノード
  $\mathbf{A} \leftarrow \mathbf{Adj}(X, G) \cap \mathbf{Adj}(Y, G) \setminus (\mathbf{Ch}(X, G) \cap \mathbf{Ch}(Y, G) \cup \mathbf{Z} \cup \{X_0\})$
  for $A \in \mathbf{A}$ **do**
    if $\log \mathrm{BF}(X, A \mid \mathbf{Z}) < 0$ **then**
      $\mathbf{E}_s \leftarrow \mathbf{E}_s \setminus \{E_{AY}\}, \mathbf{E} \leftarrow \mathbf{E} \setminus \{E_{AY}\}$
    else
      $\mathbf{E} \leftarrow \mathbf{E} \setminus \{E_{XA}\}$ ▷ E_{XA}: XA 間のエッジ
    end if
    if $\log \mathrm{BF}(A, Y \mid \mathbf{Z}) < 0$ **then**
      if $A \in \mathbf{V}_s$ かつ $Y \in \mathbf{V}_s$ **then**
        $\mathbf{E}_s \leftarrow \mathbf{E}_s \setminus \{E_{AY}\}, \mathbf{E} \leftarrow \mathbf{E} \setminus \{E_{AY}\}$
      else
        $\mathbf{E} \leftarrow \mathbf{E} \setminus \{E_{AY}\}$
      end if
    end if
  end for
  return (G_s, G)
end function

(1) 完全無向グラフ G_{uc} とデータ $\mathbf{D}$ を入力する．

(2) 各次数の CI テストにおいて $\log \mathrm{BF}(X, Y \mid \mathbf{Z})$ となり X，Y が条件付き独立と判定されるとき，XY 間のエッジを削除する．また，1 次以降の CI テストによるエッジの削除の直後，推移性に基づくエッジの削除を行う．

(3) (2) で得られたグラフに対してオリエンテーションルールを適用して方向付けを行う．

(4) 方向付けの結果から部分グラフ G_{ex} に分割する．

(5) 各部分グラフで再帰的に RAI を呼び出す．

ANB は目的変数が親を持たず，全ての説明変数を子に持つ構造である．上記のアルゴリズムに ANB を仮定するために手順 (1) の初期グラフと，手順 (2) の CI テストに変更を加えることで，RAI アルゴリズムを用いた ABN 厳密学習を実現する．ここで，目的変数のノードを X_0 それ以外のノードを目的変数のノードとする．

まず手順 (1) において，初期グラフを完全無向グラフ G_{uc} ではなく，G_{uc} に対して，目的変数 X_0 から全て説明変数に向けてエッジの方向付けを行う．つまり，目的変数から説明変数への有向エッジと説明変数間の無向エッジを併せ持つ完全グラフを初期グラフとして用いる．

次に手順 (2) において，目的変数と説明変数を接続するエッジが CI テストで削除されないように，CI テストの範囲に制約を加える．Algorithm1 に RAI アルゴリズムにおける Bayes factor を用いた CI テストの詳細を示す．ここで，入力グラフを $G = (\mathbf{V_s}, \mathbf{E_s})$ と表わす．$\mathbf{Adj}(X, G)$ はグラフ G におけるノード X の隣接ノード集合を表し，$\mathbf{Ch}(X, G)$ はグラフ G におけるノード X の子ノード集合を表す．このとき，$\mathbf{Pa}_p(X, G)$ は $\mathbf{Adj}(X, G) \setminus \mathbf{Ch}(X, G)$ を表し，$\mathbf{Pa}(X, G)$ はグラフ G に存在するノード X の親ノード集合を表す．また，$\mathbf{Pa}(X, \mathbf{G})$ はグラフ集合 $\mathbf{G}$ において $\cup_{G \in \mathbf{G}} \mathbf{Pa}(X, G)$ を表す．関数 CI_test は入力グラフのノード集合 $\mathbf{V}_s$ に対し，各次数の CI テストにおいて $\log \mathrm{BF}(X, Y \mid \mathbf{Z}) < 0$ となり X，Y が条件付き独立と判定されるとき，XY 間のエッジを削除する．ANB 学習では目的変数と説明変数は必ず接続されているため，CI テストの範囲から X_0 を取り除く（3 行目と 12 行目）．また，CI テストのよるエッジ削除の直後（7 行目と 16 行目）に関数 TRANSITIVE_CUT を呼び出す．ここでは推移性によるエッジの削除を行う．Algorithm2 に詳細を示す．関数 TRANSITIVE_CUT では CI テストで推定した条件付き独立性から推移性より検出されたノード集合 $\mathbf{A}$ に対して CI テストを行うが，ここでもノード集合 $\mathbf{A}$ から目的変数のノード X_0 を取り除く．これらの変更を加えることで RAI アルゴリズムは出力グラフに ANB を返す．

6 評価実験

6.1 実験手順

本章では，提案手法の有意性を示すために行った評価実験について述べる．UCI リポジトリデータベース [28] から, 表 1 に示される変数数の大きな 13 個の学習データを用いて，従来手法で学習した BNC と提案手法で学習した BNC の分類精度を比較した．また，deCampos ら [29] に従い，学習データに含まれる連続データは，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．学習データは欠損値を含まないものを用いた．

表 1　実験に用いた学習データ

	学習データ	説明変数数	目的変数の状態数	データ数
1	Connect-4	42	3	67557
2	dota2	116	2	102944
3	Epileptic Seizure	178	5	11500
4	Flowmeters D	43	4	180
5	kr-vs-kp	36	2	3196
6	madelon	500	2	2000
7	mfeat-fac	218	10	2000
8	MicroMass	1300	10	360
9	movement libras	90	15	360
10	Musk1	166	2	478
11	Musk2	166	2	6598
12	Parkinson’s Disease	754	2	756
13	semeion	256	10	1600

比較手法を次に示す.

- Naive Bayes

- TAN: 対数尤度を最適化する TAN を学習
- RAI-GBN: RAI アルゴリズムを用いて GBN を学習
- RAI-ANB: RAI アルゴリズムを用いて ANB を学習

2 章で述べたように厳密解探索アプローチを用いた構造学習は数十ノードの学習が限界であるため比較対象から除外した．

表 2　提案手法と従来手法の分類精度の比較

			Naive Bayes	TAN	RAI-GBN	RAI-ANB
分類精度	1	Connect-4	0.7213	0.7643	0.7337	**0.7928**
	2	dota2	**0.5981**	0.5810	0.5442	0.5957
	3	Epileptic Seizure	0.2344	0.3650	0.1887	**0.4044**
	4	Flowmeters D	**0.8389**	0.8389	0.6778	0.8278
	5	kr-vs-kp	0.8774	0.9240	0.9406	**0.9468**
	6	madelon	0.5905	0.5270	**0.6215**	0.5820
	7	mfeat-fac	0.3520	0.4590	0.2630	**0.4610**
	8	MicroMass	0.9472	0.9472	0.7361	**0.9500**
	9	movement libras	0.5028	0.5389	0.2278	**0.5583**
	10	Musk1	0.6517	0.7566	0.6744	**0.7965**
	11	Musk2	0.7445	0.8406	0.8821	**0.9627**
	12	Parkinson's Disease	0.7182	0.7898	0.7672	**0.8108**
	13	semeion	0.8550	0.8719	0.4557	**0.8745**
		平均	0.6640	0.6789	0.6064	**0.7356**
		p 値	0.0044	0.0038	0.0015	-

本論では，RAI-ANB を提案手法とする．TAN は Friedman ら [1] の手法を用いて厳密に学習した．RAI-GBN は本田ら [20] の手法を用いて学習した．また，Bayes factor における ESS = 1.0 とした [30][31] ．また，解析の妥当性を示すため，各 BNC の構造学習と分類に対して 10 分割交差検証を行った．各学習データにおいて最大の分類精度は太

字で示した．分類において，推定値は式（4）を用い，構造学習と同様に ESS = 1.0 とした．表 1 に各データセットに対する各手法の分類精度を載せた．提案手法の有意性を示すため，各手法の分類精度に対し Hommel 法 [32][33] を用いて多重検定を行い，p 値を表 1 の最下部に載せた．

6.2 結果と考察

表 2 の検定結果より，提案手法は，従来手法である Naive Bayes，TAN 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．提案手法は，TAN や Naive Bayes よりも，柔軟な構造をとることができるため，分布が複雑なデータに対しても適切な構造を学習していると考えられる．

次に，提案手法と RAI-GBN を比較する．提案手法は RAI-GBN 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．また，RAI-GBN は従来手法と比べても分類精度が低いという結果が得られた．表 3 は RAI-GBN を学習することで得られた構造のエッジ数，目的変数の親変数数と子変数数，マルコフブランケット（MB）の平均を表している．表 3 より，ほとんどのデータに対して親変数が多く，子変数が少ない構造を学習していることがわかる．また，MB の数が全説明変数に対してとても小さいことから，ほとんどの説明変数が目的変数の推定に関与していないことがわかる．また，7 番，8 番，9 番，13 番のような目的変数の状態が多いデータを用いた場合，他手法と比べて著しく分類精度が低い．このことから，データ数が不十分である場合，厳密解探索アプローチのときと同様に分類精度が著しく低くなることがわかる．それに対し提案手法は，目的変数が全ての説明変数を子に持つため，全ての説明変数が目的変数の分類に寄与している．また，目的変数が親変数を持たないため，パラメータ数が抑えられ，データ不足が緩和される．これらの理由から分類精度が高い構造が学習できたと考えられる．

例外として，6 番のデータは RAI-GBN の分類精度が一番高い．変数数 500 に対してデータ数は 2000 しかない．それにもかかわらず RAI-GBN の分類精度が高い理由として，目的変数の推定に寄与する説明変数の数が関係していると考えられる．RAI-GBN 以外の手法はすべての説明変数が目的変数の推定に寄与している．しかし，この仮定が現実のデータに当てはまるとは限らない．つまり，目的変数の分類に関係しない説明変数が分

類精度を低下させる原因になっている可能性がある．

表 3　RAI-GBN の統計概要

	学習データ	エッジ数	子変数	親変数	MB
1	Connect-4	92.6	8.4	1.7	17.7
2	dota2	151.6	0	3.0	3.0
3	Epileptic Seizure	325.1	0	0	0.0
4	Flowmeters D	63.6	0	3.7	3.7
5	kr-vs-kp	249.4	0	5.1	5.1
6	madelon	249.4	0.1	2.6	3.0
7	mfeat-fac	630.7	0	3.5	3.5
8	MicroMass	2892.4	0	7.0	7.0
9	movement libras	109.7	0	2.4	2.4
10	Musk1	413.2	0	2.0	2.0
11	Musk2	946.9	0	6.1	6.1
12	Parkinson’s Disease	1475.6	0	2.4	2.4
13	semeion	820.1	0	4.0	4.0

以上の実験結果は次のようにまとめられる．

(1) 提案手法は従来の厳密学習法よりも大規模な BNC を学習できる．
(2) 提案手法は同規模の構造学習が可能な従来の BNC よりも分類精度が有意に高い．
(3) 提案手法は構造はデータ数が少ない場合も安定して高い分類精度を得る．

7　むすび

本論文では，本田らの手法を ANB 学習に拡張させることで，従来より大規模な BNC を学習する手法を提案した．実験の結果，RAI アルゴリズムを用いて GBN を学習した場合，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ため分類精度が低いことが

わかった．一方，構造に ANB を仮定した提案手法は上記の問題を克服するため，比較手法と比べて有意に高い分類精度が得られた．

今後の課題として，目的変数と強く結びつく説明変数を発見するような変数選択手法の考察を行う．

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] 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.

[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,” Proceedings, Twenty-First International Conference on Machine Learning, 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,” Machine Learning, vol.20, no.3, pp.197–243, 1995.

[6] S. Sugahara, M. Uto, and M. Ueno, “Exact learning augmented naive Bayes classifier,” International Conference on Probabilistic Graphical Models, vol.72, pp.439–450, 2018.

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

[8] R.G. Cowell, “Efficient maximum likelihood pedigree reconstruction,” Theoretical Population Biology, vol.76, pp.285–291, 2009.

[9] M. Koivisto and K. Sood, “Exact Bayesian Structure Discovery in Bayesian Net-

works," Journal of Machine Learning Research, vol.5, pp.549–573, 2004.

[10] A. Singth and A. Moore, "Finding optimal bayesian networks by dynamic programming," Technical report, Technical Report, Carnegie Mellon University, 2005.

[11] 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.

[12] 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," Proceedings of Uncertainty in Artificial Intelligence, pp.479–488, 2011.

[13] C. Yuan, H. Lim, and T.-C. Lu, "Most Relevant Explanation in Bayesian Networks," Journal of Artificial Intelligence Research, vol.42, no.1, pp.309–352, 2011.

[14] 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.

[15] J. Pearl, Models, Reasoning, and Inference, Cambridge University Press, 2000.

[16] P. Spirtes, C. Glymour, and R. Scheines, Causation, Prediction, and Search, MIT press, 2000.

[17] I. Tsamardinos, L.E. Brown, and C.F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," Machine Learning, vol.65, no.1, pp.31–78, 2006.

[18] R. Yehezkel and B. Lerner, "Bayesian network structure learning by recursive autonomy identification," Journal of Machine Learning Research, vol.10, pp.1527–1570, 2009.

[19] 名取和樹, 宇都雅輝, 植野真臣, "Bayes factor を用いた RAI アルゴリズムによる大規模ベイジアンネットワーク学習," 電子情報通信学会論文誌 D, vol.101, pp.754–768, 2018.

[20] 本田和雅，名取和樹，菅原聖太，磯崎隆司，植野真臣，“推移性を利用した大規模ベイジアンネットワーク構造学習，” Master's thesis，電気通信大学，2019.

[21] W. Buntine, “Theory Refinement on Bayesian Networks,” Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence, pp.52–60, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1991.

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

[23] T.M. Mitchell, Machine Learning, 1 edition, McGraw-Hill, Inc., 1997.

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

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

[26] H. Steck and T.S. Jaakkola, “On the dirichlet prior and Bayesian regularization.,” pp.697–704, MIT Press, 2002.

[27] K. Natori, M. Uto, and M. Ueno, “Consistent Learning Bayesian Networks with Thousands of Variables,” Proceedings of Machine Learning Research, vol.73, pp.57–68, 2017.

[28] M. Lichman, “UCI machine learning repository,” 2013. http://archive.ics.uci.edu/ml

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

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

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

[32] G. Hommel, “A stagewise rejective multiple test procedure based on a modified bonferroni test,” Biometrika, pp.383–386, 1988.

[33] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," J. Mach. Learn. Res., vol.7, pp.1–30, 2006.