4H4-1

Grouped Bayesian Network: より細かな独立性を考慮した確率モデリング

Grouped Bayesian Network: For modeling more detailed independece

石畠 正和 Masakazu Ishihata　　岩田 具治 Tomoharu Iwata

NTT コミュニケーション科学基礎研究所
NTT Comunication Science Laboratories

Bayesian networks (BNs) represent a combination of conditional independence, or variable-wise independence, as a directed acyclic graph. High tree-width BNs are necessary for modeling complex phenomena, however, they are often intractable due to its huge number of parameters. The context-specific independence (CSI) and the partial exchangeability (PE), which describe different types of independences, have been introduced to reduce the number of parameters. In this paper, we define the *value-wise independence* as a natural generalization of CSIs and PEs. We then propose *grouped Bayesian networks* (gBNs) as a new framework of probabilistic models exploiting both of variable-wise and value-wise independences. The combination of variable-wise and value-wise independences enables us to drastically reduce the number of parameters furthermore, while enrich its expression power.

1. はじめに

確率モデルは現実の様々な現象をモデル化，予測，可視化，理解するために利用される基礎的なツールになりつつある．一般的に確率モデルは，複数の変数の依存関係の仮定を元に，それらの変数の確率分布を定義する．変数間の独立性の仮定は，確率モデル上の推論や学習の効率や，確率モデルを用いたタスクの精度に強い影響を与えるため，非常に重要な概念である．ベイジアンネットワーク (Bayesian Network ; BN) は有向グラフを用いて変数間の条件付き独立性を記述することで効率的に確率分布を定義する [5]．グラフを導入することで，変数間の条件付き独立性が人間にとって理解しやすくなるだけでなく，グラフ構造を用いた効率的な確率推論が可能になる．一般的にグラフ構造を用いたアルゴリズムの計算量は BN のグラフ構造の複雑さに強く依存する．しかし，複雑な現象を表現する BN は複雑なグラフ構造を持つ傾向があり，効率的な推論を行えない場合がある．一方，BN のグラフ構造が複雑でも，グラフ構造では表現できない特殊な独立性を導入することで，効率的な確率推論が可能な場合がある．このような特殊な独立性の例として **文脈依存独立性** (Context-Specific Indepdence ; CSI) [1] と**部分交換可能性** (Partial Exchangeability ; PE) [4] が知られている．CSI とは条件付き分布の条件部が特定の値を取るときにのみ表れる独立性であり，特定の変数が独立性を切り替えるスイッチの役割をしていることを意味する．例えば，混合モデルやトピックモデルにおいて，データの所属するクラスや文書のカテゴリによってデータの分布が切り替わる様子も CSI として表現できる．一方，PE は確率変数集合の同時分布に関する性質である．確率変数集合に対する値の割当を引数とする関数が与えられたとする．このとき PE は，ある２つの割当がその関数において同じ値を返すとき，それらは必ず同じ同時確率を持つことを意味する．例えば，２つの割当のソート結果が一致するなら同じ確率を持つ現象や，２つの割当の最大値や最小値が一致するなら同じ確率を持つ現象は PE を用いて表現できる．BN のグラフ構造が複雑であっても，これらの CSI や PE を仮定することで，効率的に確率推論が可能になることが知られているが，これら組み合わせる研究はまだ存在しない．

本稿では CSI と PE を一般化した新たな独立性である Value-wise Indepdence (VI) を提案する．CSI は条件付き分布の条件部に関する独立性を記述し，PE は同時分布に関する独立性を記述するため，これらを組み合わせることでより柔軟な独立性を記述可能となる．更に本稿では，この VI を効率的に表現可能である新たな確率モデリングの枠組みである Grouped Bayesian Network (gBN) を提案する．gBN により，BN ではグラフ構造が複雑になるため効率的に推論できなかった確率モデルも効率的に推論可能になる．また，gBN に対する構造学習法も提案する．提案法は，VI を事前に仮定することにより，効率的に学習可能な gBN のみを出力する．本稿では gBN を識別問題と確率推定問題に適用し，その有用性を確認する．

2. 準備

本節ではまず通常の**ベイジアンネットワーク** (Bayesian Network ; BN) について述べ，BN のグラフ構造では効率的に表現できない独立性である**文脈依存独立性** (context-specific indepdence ; CSI) と**部分交換可能性** (partial exchangeability ; PE) について簡単に述べる．なお，本論文では確率変数はすべて離散変数であるとする．

2.1 Baysian Network

BN $\mathcal{M}=\langle \mathrm{G}, \boldsymbol{\theta} \rangle$ は確率的グラフィカルモデルの一種であり，有向グラフ (Dyrected Acyclic Graph ; DAG) G と条件付き確率表 (Conditional Probablity Table ; CPT) $\boldsymbol{\theta}$ を用いて N 個の確率変数 $X_1, \ldots, X_N$ の同時分布を定義する．DAG G は確率変数間の**条件付き独立性** (Conditional Independence ; CI) を表現し，CPT $\boldsymbol{\theta}$ は条件付き確率分布を定義する．以後，簡単のため集合 $\{1, \ldots, N\}$ を $[N]$ と表す．X_i $(i \in [N])$ を離散確率変数とし，x_i を X_i の値とする．X_i の値域を $\boldsymbol{D}_i = \{x_i^k \mid k \in [M_i]\}$ とする．ここで M_i は値域のサイズ，x_i^k は $\boldsymbol{D}_i$ 中の k 番目の値である．集合 $\boldsymbol{S} \subseteq [N]$ に対して，$\boldsymbol{X}_{\boldsymbol{S}}$ を確率変数集合 $\{X_i \mid i \in \boldsymbol{S}\}$ とし，$\boldsymbol{x}_{\boldsymbol{S}}$ を値集合 $\{x_i \mid i \in \boldsymbol{S}\}$ とする．すると $\boldsymbol{X}_{\boldsymbol{S}}$ の値域は $\boldsymbol{D}_{\boldsymbol{S}} = \prod_{i \in \boldsymbol{S}} \boldsymbol{D}_i$ であり，そのサイズを $M_{\boldsymbol{S}} = \prod_{i \in \boldsymbol{S}} M_i$ と表し，その k 番目の値の組合せを $\boldsymbol{x}_{\boldsymbol{S}}^k$ と表す．BN $\mathcal{M} = \langle \mathrm{G}, \boldsymbol{\theta} \rangle$ は，N 個の確率変数の同時分布を定

連絡先: 石畠正和 ishihata.masakazu@lab.ntt.co.jp

義し，これを $P_{\mathcal{M}}\big(\boldsymbol{X}_{[N]}\big)$ と表す．集合 $\boldsymbol{S} \subseteq [N]$ に対して，$P_{\mathcal{M}}(\boldsymbol{X}_{\boldsymbol{S}}{=}\boldsymbol{x}_{\boldsymbol{S}})$ は $X_i{=}x_i$ $(i{\in}\boldsymbol{S})$ となる同時確率であり，以後簡単のため $P_{\mathcal{M}}(\boldsymbol{x}_{\boldsymbol{S}})$ と書く．DAG $\mathrm{G}{=}\langle[N],\boldsymbol{E}\rangle$ は頂点集合 $[N]$ と有向辺集合 $\boldsymbol{E} \subseteq [N]\times[N]$ の組である．頂点 $i{\in}[N]$ は確率変数 X_i に対応し，有向辺 $(i,i'){\in}\boldsymbol{E}$ は確率変数 X_i と $X_{i'}$ の間に依存関係があることを意味する．頂点 i から頂点 i' の有向辺 (i,i') が存在するとき，i は i' の親であるといい，i から i' への有向パスが存在するとき，i は i' の祖先であるという．有向グラフ G が DAG であるとは，いかなる頂点 i も自分自身を祖先に持たないことを意味する．G が DAG であるとき，i が i' の祖先ならば $i < i'$ となるような頂点番号を与えることが可能であり，これをトポロジカル順序という．本稿では簡単のため，確率変数 $X_i{\in}\boldsymbol{X}_{[N]}$ のインデックスはトポロジカル順序で与えられるとする．頂点 i の親頂点集合を $\boldsymbol{\pi}(i){=}\{i' \mid (i',i){\in}\boldsymbol{E}\}$ とし，$\bar{\boldsymbol{\pi}}(i){=}[i{-}1]\backslash\boldsymbol{\pi}(i)$ とする．このとき，DAG G は条件付き独立性 $X_i \perp\!\!\!\perp \boldsymbol{X}_{\bar{\boldsymbol{\pi}}(i)} \mid \boldsymbol{X}_{\boldsymbol{\pi}(i)}$ $(i{\in}[N])$ を表現し，これは同時分布が $P_{\mathcal{M}}\big(\boldsymbol{X}_{[N]}\big){=}\prod_{i\in[N]} P_{\mathcal{M}}\big(X_i \mid \boldsymbol{X}_{\boldsymbol{\pi}(i)}\big)$ と分解可能であることと等価である．CPT $\boldsymbol{\theta}{=}\{\theta_{i,j,k} \mid i{\in}[N], j{\in}[M_{\boldsymbol{\pi}(i)}], k{\in}[M_i]\}$ は条件付き確率分布 $P_{\mathcal{M}}\big(X_i \mid \boldsymbol{X}_{\boldsymbol{\pi}(i)}\big)$ を定義する．$\theta_{i,j,k}$ を「$\boldsymbol{X}_{\boldsymbol{\pi}(i)}$ が j 番目の値の組合せ $\boldsymbol{x}^j_{\boldsymbol{\pi}(i)}$ を取ったとき，X_i が k 番目の値 x_i^k を取る確率」とする．つまり $P_{\mathcal{M}}\Big(x_i^k \mid \boldsymbol{x}^j_{\boldsymbol{\pi}(i)}\Big) \equiv \theta_{i,j,k}$ である．BN は DAG G により確率変数間の CI を表現することで，同時分布を効率的に表現する．しかし，CI は確率変数同士の独立性であり，確率変数のとる値によって独立性が変化するような細かな独立性は効率的に表現できない．2.2 と 2.3 では通常の BN では効率的に表現できない独立性の例を示す．

2.2 Context-Specific Indepdence

BN の DAG 構造は確率変数間の条件付き独立性 (CI) を効率的に表現する．一方で，DAG 構造では効率的に表現できない独立性が存在することが知られている．**文脈依存独立性** (Context-Specific Indepedence ; CSI) [1] は，DAG で表現できない独立性の一例である．CSI の定義は以下である．

Definition 1 互いに素な集合 $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C} \subset [N]$ が与えられたとき，$\boldsymbol{X_A}$ と $\boldsymbol{X_B}$ が文脈 $\boldsymbol{x_C}$ が与えられた上で互いに文脈依存独立であるき以下を満たす．

$$\forall \boldsymbol{x_A}{\in}\boldsymbol{D_A}, \forall \boldsymbol{x_B}{\in}\boldsymbol{D_B},$$
$$P_{\mathcal{M}}(\boldsymbol{x_A},\boldsymbol{x_B} \mid \boldsymbol{x_C}) = P_{\mathcal{M}}(\boldsymbol{x_A} \mid \boldsymbol{x_C})\, P_{\mathcal{M}}(\boldsymbol{x_B} \mid \boldsymbol{x_C})$$

つまり CSI は $\boldsymbol{X_A}$ と $\boldsymbol{X_B}$ は $\boldsymbol{X_C}$ の値が $\boldsymbol{x_C}$ であるときに限り，互いに独立であることを意味する．CI は特定の確率変数集合が与えられたときに 2 つの確率変数集合が互いに独立となる性質を表す．一方で CSI は特定の確率変数集合が特定の値の組合せをとるときのみ 2 つの確率変数集合が互いに独立となる性質である．BN の DAG 構造は，あくまでも確率変数同士の依存関係を表現するため，CSI のように確率変数が特定の値を取るときにのみ表れる独立性は効率的に表現することはできない．

2.3 Partial Exchangeability

CI や CSI は，条件付き確率分布の性質を用いて定義される独立性である．一方で，**部分交換可能性** (Partial Exchangeability ; PE) は同時分布の持つ性質を利用して定義される独立性である．PE の定義は以下である．

Definition 2 集合 $\boldsymbol{A}{\subseteq}[N]$ が与えられたとき，$\boldsymbol{X_A}$ が関数 $f(\boldsymbol{X_A})$ の元で部分交換可能であるとき以下を満たす．

$$\forall \boldsymbol{x_A}, \boldsymbol{x'_A} {\in} \boldsymbol{D_A}, f(\boldsymbol{x_A}) = f(\boldsymbol{x'_A}) \Longrightarrow P_{\mathcal{M}}(\boldsymbol{x_A}) = P_{\mathcal{M}}\big(\boldsymbol{x'_A}\big)$$

つまり $\boldsymbol{X_A}$ の値は関数 f の値によっていくつかのグループに分けられ，同じグループに属する値は同じ同時確率を持つ．PE も CSI と同様，BN の DAG 構造では効率的に表現できない．PE は CI や CSI などの条件付き確率に基づく独立性とは異なる性質であるため，従来の BN や CSI では表現できない確率モデルを記述可能である．Exchangeable Variable Model (EVM) [4] は PE を用いた新たな確率モデルであり，混合 EVM (Mixture of EVMs : MEVM) は条件付き独立性を利用した確率モデルであるナイーブベイズモデル (Naive Bayes Model ; NBM) の PE を用いた一般化でる．MEVM は NBM よりも，識別問題や確率推定問題で高い精度を達成することが実験的に確認されている．

3. Grouped Baeysian Network

CSI と PE は BN の DAG 構造を用いて効率的に表現することができないが，同時分布を定義するためのパラメータ数を大幅に削減可能であることが知られている．本節ではまず，CSI と PE を一般化した新たな独立性である value-wise indepdence を提案する．更に value-wise indepdence を効率的に扱えるよう BN を拡張し，新たな確率モデリングの枠組みである Grouped Baysian Network (gBN) を提案する．

3.1 Value-wise Independence

ここではまず，2 つの異なる独立性である CSI と PE を包含する新たな独立性として Value-wise Indepdence (VI) を提案する．PE とは同時確率分布において変数集合が特定の条件を満たすとき，確率値を共有する事を表す独立性であり，条件は関数 f によって表現される．一方，CSI は条件付き確率分布の条件部が特定の値を取るときに表れる独立性であり，条件を関数 g によって表現することにする．すると，PE と CSI はそれぞれ関数 f と g によって確率値を共有するかどうかを定義する独立性と考えることができ，この一般化した独立性を VI と呼び，以下の様に定義する．

Definition 3 互いに素である集合 $\boldsymbol{A},\boldsymbol{B}{\subset}[N]$ が与えられたとき，$\boldsymbol{X_A}$ と $\boldsymbol{X_B}$ が関数 $f(\boldsymbol{X_A})$ と $g(\boldsymbol{X_B})$ の元で互いに *VI* であるとき以下を満たす．

$$\forall \boldsymbol{x_A}, \boldsymbol{x'_A} {\in} \boldsymbol{D_A}, \forall \boldsymbol{x_B}, \boldsymbol{x'_B} {\in} \boldsymbol{D_B},$$
$$f(\boldsymbol{x_A}) = f(\boldsymbol{x'_A}) \ \wedge\ g(\boldsymbol{x_B}) = g(\boldsymbol{x'_B})$$
$$\Longrightarrow P_{\mathcal{M}}(\boldsymbol{x_A} \mid \boldsymbol{x_B}) = P_{\mathcal{M}}\big(\boldsymbol{x'_A} \mid \boldsymbol{x'_B}\big)$$

関数 f が恒等写像であるとき，VI は CSI と一致し，関数 g が恒等写像であるとき， VI は PE と一致する．

3.2 Grouped Bayesian Network

ここでは VI を自然に扱うための新たな枠組みとして Grouped Bayesian Network (gBN) を提案する．確率変数集合 $\boldsymbol{X}_{[N]}$ 上の同時分布を定義する gBN $\mathcal{M}$ は以下の 4 つ組 $\langle \boldsymbol{P}, \mathrm{G}, \mathcal{C}, \boldsymbol{\theta}\rangle$ として表現される．

- $\boldsymbol{P} = \{\boldsymbol{P}_\ell \mid \ell \in [L]\}$ は $[N]$ 上の分割である．$\boldsymbol{P}_1,\ldots,\boldsymbol{P}_L$ は互いに素である $[N]$ の部分集合であり，$\bigcup_{\ell\in[L]} \boldsymbol{P}_\ell {=} [N]$ を満たす．$\boldsymbol{P}_\ell$ を ℓ 番目のグループと呼び，$\boldsymbol{X}_{\boldsymbol{P}_\ell}$ を ℓ 番目の変数グループと呼ぶ．

- $\mathrm{G}=\langle \boldsymbol{V},\boldsymbol{E}\rangle$ は DAG であり，$\boldsymbol{V}=[L]$ は頂点集合，$\boldsymbol{E}\subseteq \boldsymbol{V}\times\boldsymbol{V}$ は有向辺集合である．$\boldsymbol{\pi}(\ell)$ を頂点 $\ell\in\boldsymbol{V}$ の親集合とし，$\boldsymbol{\Pi}(\ell)=\bigcup_{\ell'\in\boldsymbol{\pi}(\ell)}\boldsymbol{P}_{\ell'}$ とする．$\boldsymbol{X}_{\boldsymbol{\Pi}(\ell)}$ を $\boldsymbol{X}_{\boldsymbol{P}_\ell}$ の親変数グループと呼ぶ．
- $\mathcal{C}=\{(f_\ell,g_\ell)\mid \ell\in[L]\}$ は文脈ペア集合であり，$f_\ell:\boldsymbol{D}_{\boldsymbol{P}_\ell}\to\mathcal{F}_\ell$ かつ $g_\ell:\boldsymbol{D}_{\boldsymbol{\Pi}(\ell)}\to\mathcal{G}_\ell$ である．$\mathcal{F}_{\ell,k}$ を $\mathcal{F}_\ell$ 中の k 番目の値とし，$\mathcal{G}_{\ell,j}$ を $\mathcal{G}_\ell$ 中の j 番目の値とする．また，$f_{\ell,k}=\{\boldsymbol{x}_\ell\in\boldsymbol{D}_\ell \mid f_\ell(\boldsymbol{x}_\ell)=\mathcal{F}_{\ell,k}\}$ かつ $g_{\ell,j}=\{\boldsymbol{x}_{\boldsymbol{\Pi}(\ell)}\in\boldsymbol{D}_{\boldsymbol{\Pi}(\ell)}\mid g_\ell(\boldsymbol{x}_{\boldsymbol{\Pi}(\ell)})=\mathcal{G}_{\ell,j}\}$ とする．
- $\boldsymbol{\theta}=\{\theta_{\ell,j,k}\mid \ell\in[L],j\in[|\mathcal{G}_\ell|],k\in[|\mathcal{F}_\ell|]\}$ はパラメータ集合であり，変数グループ $\boldsymbol{X}_{\boldsymbol{P}_\ell}$ $(\ell\in[L])$ の CPT を定義する．

gBN $\mathcal{M}$ が与えられたとき，同時分布は $P_{\mathcal{M}}\big(\boldsymbol{X}_{[N]}\big)=\prod_{\ell\in[L]}P_{\mathcal{M}}\big(\boldsymbol{X}_\ell\mid\boldsymbol{X}_{\boldsymbol{\Pi}(\ell)}\big)$ の形に分解され，また，個々の条件付き分布 $P_{\mathcal{M}}\big(\boldsymbol{X}_\ell\mid\boldsymbol{X}_{\boldsymbol{\Pi}(\ell)}\big)$ は以下の様に定義される．

$$\forall\boldsymbol{x}_{\boldsymbol{P}_\ell}\in f_{\ell,k},\forall\boldsymbol{x}_{\boldsymbol{\Pi}(\ell)}\in g_{\ell,j},P_{\mathcal{M}}\big(\boldsymbol{x}_{\boldsymbol{P}_\ell}\mid\boldsymbol{x}_{\boldsymbol{\Pi}(\ell)}\big)\equiv\theta_{\ell,j,k}$$

つまり gBN $\mathcal{M}$ では，変数グループ間の大まかな独立性を DAG G で表現し，細かな独立性を VI として表現する．gBN $\mathcal{M}$ のパラメータ数 $|\boldsymbol{\theta}|$ は定義より $\sum_{\ell\in[L]}|\mathcal{F}_\ell||\mathcal{G}_\ell|$ であり，DAG G の構造には直接的には依存しない．つまり，通常の BN で表現した場合は複雑な DAG になってしまうような確率モデルでも，gBN で表現すれば少ないパラメータ数で表現可能である場合がある．

3.3 パラメータ推定

観測データよりパラメータ $\boldsymbol{\theta}$ を推定する問題は gBN において非常に重要なタスクの一つである．$\boldsymbol{x}^{(t)}_{[N]}$ を同時分布 $P_{\mathcal{M}}\big(\boldsymbol{X}_{[N]}\big)$ より独立に得られた t 番目の観測とする．gBN $\mathcal{M}$ の一部 $\langle\boldsymbol{P},\mathrm{G},\mathcal{C}\rangle$ と観測系列 $\boldsymbol{O}=\{\boldsymbol{x}^{(t)}_{[N]}\mid t\in[T]\}$ が与えられたとき，$\mathcal{M}$ の最尤推定量 $\boldsymbol{\theta}^*$ とは対数尤度 $\mathcal{L}(\mathcal{M};\boldsymbol{O})\equiv\sum_{t\in[T]}\sum_{\ell\in[L]}\log P_{\mathcal{M}}\Big(\boldsymbol{x}^{(t)}_{\boldsymbol{P}_\ell}\mid\boldsymbol{x}^{(t)}_{\boldsymbol{\Pi}(\ell)}\Big)$ を最大化するパラメータのことである．この最尤推定量 $\boldsymbol{\theta}^*$ は以下の様に計算可能である．

Theorem 1 *Given* $\langle\boldsymbol{P},\mathrm{G},\mathcal{C}\rangle$ *and* $\boldsymbol{O}$, $\boldsymbol{\theta}^*$ *is computed as*

$$\theta^*_{\ell,j,k}=\frac{1}{|f_{\ell,k}|}\frac{\sum_{t\in[T]}\delta\Big(\boldsymbol{x}^{(t)}_{\boldsymbol{P}_\ell}\in f_{\ell,k}\Big)\,\delta\Big(\boldsymbol{x}^{(t)}_{\boldsymbol{\Pi}(\ell)}\in g_{\ell,j}\Big)}{\sum_{t\in[T]}\delta\Big(\boldsymbol{x}^{(t)}_{\boldsymbol{\Pi}(\ell)}\in g_{\ell,j}\Big)}$$

つまり，$\boldsymbol{\theta}^*$ は単純に $\boldsymbol{x}^{(t)}_{\boldsymbol{P}_\ell}\in f_{\ell,k}$ かつ $\boldsymbol{x}^{(t)}_{\boldsymbol{\Pi}(\ell)}\in g_{\ell,j}$ を満たす $\boldsymbol{x}^{(t)}_{[N]}$ を数えることで計算できる．

3.4 構造推定

観測系列 $\boldsymbol{O}$ から gBN $\mathcal{M}$ 全体を推定することは重要なタスクである．しかし，gBN の構造学習は通常の BN の構造学習を部分問題に含むため，その一般的な計算量は NP-hard である [2]．そこで本稿では，$\boldsymbol{O}$ から効率的に学習可能な gBN を生成するヒューリスティックを提案する．ここで簡単のため $\boldsymbol{D}_i=\{0,1\}$ $(i\in[N])$ とする．

本稿では，生成される gBN のパラメータ数を少なく保つため，文脈ペア (f_ℓ,g_ℓ) $(\ell\in[L])$ に対して以下の制約を設ける．

$$f_\ell(\boldsymbol{x}_{\boldsymbol{P}_\ell})=\sum_{i\in\boldsymbol{P}_\ell}X_i,\quad g_\ell(\boldsymbol{x}_{\boldsymbol{\Pi}(\ell)})=\delta\left(\sum_{i\in\boldsymbol{\Pi}(\ell)}X_i>B_\ell\right)$$

ここで B_ℓ は閾値であり，これは観測系列 $\boldsymbol{O}$ より学習される．この制約により，生成される gBN のパラメータ数は高々 $2N$ であることが保証される．この制約の元，提案学習アルゴリズムは以下の 3 ステップにより gBN を生成する．

1. Group Learning: ここではまず，分割 $\boldsymbol{P}$ を観測系列 $\boldsymbol{O}$ より構成する．上記の f,g に関する制約により，以下の性質が利用できる．

Theorem 2 関数 f_ℓ と g_ℓ が交換可能であるとき，*gBN* $\mathcal{M}$ により定義される同時分布 $P_{\mathcal{M}}\big(\boldsymbol{X}_{[N]}\big)$ は以下を満たす．

$$\forall\boldsymbol{P}_l\in\boldsymbol{P},\forall X_i,X_{i'}\in\boldsymbol{P}_l,P_{\mathcal{M}}(X_i)=P_{\mathcal{M}}(X_{i'})$$

上記の $\boldsymbol{P}$ に関する必要条件は，2 つの変数が同じ変数グループに属するかの判定に利用可能である．例えば，検定により仮説 $P_{\mathcal{M}}(X_i)=P_{\mathcal{M}}(X_{i'})$ が棄却できないとき，X_i と $X_{i'}$ は同じ変数グループに所属すると仮定できる．確率変数が 2 値であるとき，経験分布 $P_{\mathcal{M}}(X_i=1)$ は X_i の $\boldsymbol{O}$ での平均値と一致する．以下の $\boldsymbol{P}$ を構成するヒューリスティックは EVM のグループ発見アルゴリズム [4] の gBN への拡張である．

1. Compute $E_i=\sum_{t\in[T]}X^{(t)}_i/T$ for each $i\in[N]$.
2. $\boldsymbol{S}=[N]$, $\ell=0$, $i=0$.
3. Remove i' with the minimum $E_{i'}$ from $\boldsymbol{S}$.
4. If $\ell=0$ or $P_{\mathcal{M}}(X_{i'})=P_{\mathcal{M}}(X_i)$ is rejected, $\ell{+}{+}, i=i'$.
5. Add i' to $\boldsymbol{P}_\ell$ and back to 3. if $|\boldsymbol{S}|\neq 0$.

2. DAG Learning: 次に推定された分割 $\boldsymbol{P}$ より DAG G を推定する．$\boldsymbol{Y}_{[L]}=\{Y_\ell\mid\ell\in[L]\}$ とし，$Y_\ell=f_\ell(\boldsymbol{X}_{\boldsymbol{P}_\ell})$ とする．このとき，$Y_{[L]}$ 上の条件付き独立性は通常の BN に対する構造学習アルゴリズムを用いて推定である．この方法で得られた DAG を G とする．ここでは通常の BN に対する構造学習アルゴリズムとして計算量 $O(L^2)$ と比較的少ない Chow-Liu tree algorithm [3] を採用する．ここで，このステップで得られた G の構造は，最終的に生成される gBN の複雑さに影響しないことに注意する．つまり，Chow-Liu tree は木構造を発見する手法だが，より複雑な DAG を出力する手法を用いても問題はない．

3. CPT Learning: 最後に推定された $\boldsymbol{P}$ と G より g_ℓ の閾値 $\mathcal{B}=\{B_\ell\mid\ell\in[L]\}$ とパラメータ $\boldsymbol{\theta}$ を推定する．ここで $\mathcal{B}^*$ と $\boldsymbol{\theta}^*$ を対数尤度 $L(\mathcal{M};\boldsymbol{O})$ を最大化する閾値とパラメータの対とする．$L(\mathcal{M};\boldsymbol{O})$ の定義より，B^*_ℓ と $\theta^*_{\ell,j,k}$ $(j\in[|\mathcal{G}_\ell|],k\in[|\mathcal{F}_\ell|])$ は個別に $L_\ell(\mathcal{M};\boldsymbol{O})=\sum_{t\in[T]}P_{\mathcal{M}}\Big(\boldsymbol{x}^{(t)}_{\boldsymbol{P}_\ell};\boldsymbol{x}^{(t)}_{\boldsymbol{\Pi}(\ell)}\Big)$ を最大化することで計算可能であることが分かる．B^*_ℓ はすべての可能な B_ℓ の値 $(0\le B_\ell\le|\boldsymbol{\Pi}(\ell)|)$ を考え，対応する対数尤度を計算することで推定可能である．また，B_ℓ が与えられたときの $\theta^*_{\ell,j,k}$ は定理 1 より計算可能である．

4. 実験

本節では gBN の有用性を示すために，gBN を識別問題と確率学習問題に適用する．識別問題では gBN を 2 つの生成モデルと 3 つの識別モデルと比較し，確率学習問題においては gBN を同じ 2 つの生成モデルと比較する．

4.1 識別問題

ここでは表 1 に示される 6 つの正解ラベル付きの実データセットより 2 値分類問題を生成した．初めの 5 つのデータセットはテキストデータであり，残りの 1 つのデータは手書き文字データである． 実験設定の詳細は [4] に従ったため省略する．本実験では gBN の混合モデル (Mixture of gBN ; MgBN) を

表 1: 分類問題に対する実験結果

データ	属性数 N	学習サンプル数	評価サンプル数	NB	MEVM	MgBN	DT	SVM	5-NN
20News	19,726	1,131.4	753.2	0.830	0.905	0.906	0.804	0.867	0.578
Reuters-8	19,398	1,371.3	547.2	0.938	0.954	0.962	0.958	0.978	0.881
Polarity	38,045	1,800.0	200.0	0.793	0.824	0.820	0.633	0.859	0.518
Enron	43,813	4,000.0	1,000.0	0.911	0.971	0.972	0.952	0.974	0.742
WebKB	7,290	1,401.5	698.0	0.907	0.926	0.932	0.885	0.957	0.798
MNIST	784	12,000.0	2,000.0	0.965	0.969	0.971	0.981	0.983	0.995

表 2: 確率推定問題に対する実験結果

データ	属性数 N	学習サンプル数	評価サンプル数	NB	MEVM	MgBN
NLTCS	16	16,181	3,236	-6.041	-6.038	**-6.024**
MSNBC	17	291,326	58,265	-6.770	-6.240	**-6.131**
KDDCup2000	64	180,092	34,955	-2.151	-2.143	**-2.135**
Plants	69	17,412	3,482	-15.121	-14.977	**-13.486**
Audio	100	15,000	3,000	-40.586	-40.568	**-40.193**
Jester	100	9,000	4,116	-53.170	-53.181	**-53.037**
Netflix	100	15,000	3,000	-57.814	-57.796	**-56.920**
MSWeb	294	29,441	5,000	-9.980	-9.981	**-9.825**
Book	500	8,700	1,739	**-34.659**	-34.681	-35.264
WebKB	839	2,803	838	**-157.506**	-157.708	-164.930
Reuters-52	889	6,532	1,540	-86.528	**-86.496**	-87.188
20Newsgroup	910	11,293	3,764	**-152.663**	-152.743	-155.315

実装し，2値分類器とした．MgBN の各コンポーネントは 3.4 で提案した構造学習アルゴリズムにより学習した．本実験では MgBN を2つ生成モデルと3つ識別モデルと比較した．前者はナイーブベイズ識別 (NB) と MEVM であり，後者は決定木 (DT)，サポートベクターマシン (SVM)，k 近傍法 (k-NN) である．ここで NB と MEVM は MgBN の特殊例である．MgBN の DAG 構造が全域木であるとき，MgBN のパラメータ数は NB の2倍となる．

表 1 の最後の6列に6つの識別器の平均精度を示す．まず，生成モデルの結果を比較する．MgBN はすべての学習データセットに対して最良の対数尤度を達成し，テストデータに対しては Polarity データ以外において最良の対数尤度を達成した．Polarity データは，学習サンプル数に対して特徴ベクトルの次元が高く，過学習を起こしやすいと言える．MgBN は NB や MEVM と比べ，最大2倍のパラメータを利用することから，Polarity データに対して過学習を起こしたと推測される．次に MgBN と3つの識別モデルを比較する．一般的に識別問題では，生成モデルと識別モデルでは識別モデルの方が高精度であることが知られている．しかし，MgBN は DT と 5-NN と比較して非常によい精度を達成しており，SVM と比較しても1つのデータセットではよりよい精度を達成している．

4.2 確率学習問題

ここでは gBN を確率モデルの構造学習タスクでよく利用されているベンチマークに適用した [4]．このベンチマークは表 2 の1列目に示される12個の実データセットから成る．ここでは3つの生成モデル MgBN, BN, MEVM をこれらのデータセットに適用し，テストデータに対する対数尤度を比較した．実験設定の詳細は [4] に従ったため省略する．

表 2 の最後の3列に3つのモデルのテスト尤度を示す．MgBN はすべての学習データに対して最良の対数尤度を達成し，テストデータにおいては初めの8つのデータセットに対して最良の対数尤度を達成した．この結果は，残りの4つのデータセットは学習データ数に対して属性数が非常に大きいため，MgBN は過学習を起こしたと考えられる．

5. まとめ

本稿では BN の DAG 構造では効率的に表現できない独立性である CSI と PE を一般化した Value-wise Indepdence (VI) を提案し，VI を効率的に扱うことが可能である新たな確率モデルの枠組みである Grouped Bayesian Netwrok (gBN) を提案した．gBN を用いることで，これまで BN で表現した場合，そのグラフ構造が複雑になってしまうため，効率的に確率推論が行えなかったようなモデルクラスに対しても，効率的なパラメータ学習と構造学習を可能にする．本稿では gBN を分類問題と確率推定問題に適用し，その有用性を確認した．今後の課題として，BN に対するモデル選択基準を gBN に拡張することで，gBN の構造学習における過学習を緩和させることが考えられる．

参考文献

[1] Craig Boutilier, Nir Friedman, Moisés Goldszmidt, and Daphne Koller. Context-specific independence in bayesian networks. In *UAI'96*, 1996.

[2] David Maxwell Chickering. Learning bayesian networks is np-complete. In *AISTATS'95*, 1995.

[3] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

[4] Mathias Niepert and Pedro M. Domingos. Exchangeable variable models. In *ICML'14*, 2014.

[5] Judea Pearl. *Probabilistic reasoning in intelligent systems - networks of plausible inference.* Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1989.