

遺伝的プログラミングを用いた動的ベイズネットワーク記述 における形状変化の推定とその応用

時永 祥三^{1,a)} 高木 昇^{2,b)}

概要: 本報告では遺伝的プログラミング (Genetic Programming:GP) を用いた DBN における形状変化の推定とその応用について述べる。最初に DBN における基本的なモデルを導入する。DBN を表現する DAG の形状が未知あるいは時間的に変化すると仮定して、DAG の形状を GP により推定する方法を述べる。前置表現を基礎とした等価な 2 分岐を用いて、条件付き確率の関係式を、関数における被演算子と演算子の結合関係に類似した表現として求め、これに対する遺伝的操作を適用する。この場合、DBN の全体はいくつかの部分に分割されるので (部分木と呼ぶ)、この性質を用いて部分木における事象の生起確率を、観測データから推定する問題を定式化する。GP による DBN 形状変化の推定においては、推定のたびに初期個体から開始する非効率性を回避する目的で、現在の DBN 形状を表現する GP 個体と、ランダムに生成した GP 個体との交差処理を活用する方法を提案する。

1. はじめに

ベイズネットワーク (Bayesian Network:BN) はランダム変数 (確率変数) の間の結合の静的表現を、方向性非巡回グラフ (Directed Acyclic Graph:DAG) を用いて図式的に表現する方法であり、さまざまな分野で応用されている [1][2][3][4]。近年 DAG において、時間的な遷移を導入した動的ベイズネットワーク (Dynamic BN:DBN) が提案され、応用されている [5][6][7][8]。しかしながら従来の DBN 手法においては DAG のとりうる範囲に関して、あらかじめ形状が既知であるグラフに限定され、状態の存在確率も既知のパターンに属するとされており、未知の DAG 形状および状態変化などには適用できない問題がある [7]。本報告では、遺伝的プログラミング (Genetic Programming:GP) を用いた DBN 記述における形状変化の推定とその応用について述べる [9][10][11][12]。

本報告ではまず最初に、DBN を表現する DAG の形状が未知であるか時間的に変化する場合の推定問題を整理する。次に、DAG の形状変化の推定について、通常の算術式に対する GP において用いられる交差処理と突然変異を基礎とした方法を拡張したケースについて整理する。具体的には、通常の GP による算術式の関数近似の場合にならっ

て、前置表現 (prefix representation) を基礎とした等価な 2 分岐 (binary tree) を用いて、条件付き確率の関係式を、関数における被演算子と演算子の結合関係に類似した表現として求め、これに対する遺伝的操作を適用する。この方法は、従来の遺伝的手法に基づく形状推定の方法より効率的である。GP における個体の適合度として、実際に観測される出力変数の同時分布確率と、個体が DBN を表現すると仮定した場合に得られる出力との差異を用いている。この場合、DBN の全体はいくつかの部分に分割されるので (部分木と呼ぶ)、この性質を用いて部分木における事象の生起確率を、観測データから推定する問題を定式化する [7]-[12]。具体的には、変数 X_i は部分木において、その葉に相当する変数 $Pa(X_i)$ (親変数と呼ぶ) の条件付き確率で与えられるので、この生起確率をベイズ推定の方法におけるサンプリング手法を用いて、逐次的に改善する。この場合、確率分布の表現に適しているディリクレ分布を適用する。更に、時間的に変化する形状を GP により推定する場合に、あらかじめ初期個体から開始する非効率性を回避する目的で、現在の DBN 形状を表現する GP 個体と、ランダムに生成した GP 個体との交差処理を活用する方法を提案する。応用例として、まず DAG の構造変化と状態変化が既知である人工データに対して本報告の手法を適用し、推定を行った結果を評価することにより性能を検証するとともに、現実のデータへの適用を考察する。

¹ 九州大学
Kyushu University, Fukuoka, 812-8581, Japan

² 九州産業大学
Kyushu Sangyo University, Fukuoka, 813-8503, Japan

a) tokenaga@en.kyushu-u.ac.jp

b) t-noboru@kyusan-u.ac.jp

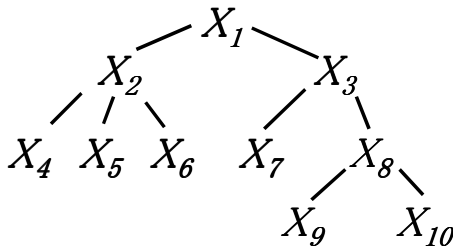


図 1 BN の例

2. DBN 形状推定問題の定式化

いま, n 個の確率変数のベクトル $X = (X_1, X_2, \dots, X_n)$ に関する観測データがあると仮定する. BN は確率的なグラフモデルにより確率変数 X_1, X_2, \dots, X_n の間の関係を表現することが目的であり, その場合, 観測されたケース表から得られる確率分布と, BN 形状を仮定した場合に得られる関係表現から得られた確率分布とが, 一致する必要がある. このグラフにおけるノードは変数に対応しており, これらのノードの結合関係である DAG により変数の間の依存性を表現する. 変数 X_i (子変数と呼んでおく) は変数の位置より先行する変数 (親変数と呼んでおく), すなわち DAG において変数 X_i に対応するノードに流入する方向で結合されているノードに対応する変数 X_j (一般には複数個) を用いた条件付確率により記述される. すなわち $P(X_i | Pa(X_i))$ として定義され, ここで $Pa(X_i)$ は変数 X_i の親変数の集合を意味する. これにより確率変数 X_1, X_2, \dots, X_n の同時分布は, 次のように表現できることが分かる [1]-[6].

$$P(X) = \prod_{i=1}^n P(X_i | Pa(X_i)), X = (X_1, X_2, \dots, X_n) \quad (1)$$

ここで重要なことは, BN の形状が既知である場合 (BN 表現が正しい場合) には式 (1) の関係は成立するが, BN の形状が未知である場合 (BN 表現が正しいとは限らない) には, 式 (1) の左辺と右辺は必ずしも一致しない点である. いま, 変数 X_i に対する観測データが与えられた場合に, この変数 X_i の同時確率を計算した左辺の値は, 右辺におけるそれぞれの部分木で表現される親変数と子変数の関係が正しい場合に限り, 一致する値をとる. すなわち, 親変数と子変数の関係を与える BN の形状が, 正しく推定されていない限り, 式 (1) の関係は成立しない.

本報告では, 式 (1) に示す子変数 X_i と親変数 $Pa(X_i)$ との関係部分を呼んでおく. また部分木における変数の間の関係は, 生起確率として表で表される. 具体的には, 変数 X_i における親変数の論理値のすべての組み合わせのパターンを表の縦軸方向の区分として, また変数 X_i の論理値のとりべき数値を横軸として, 表の縦と横を構成し, その欄の中に生起確率 θ_{ijk} を記入することで, 子変数とその親変数との関係を生起確率の表として示すことができる. ただし添え字 j は親変数について論理値を変化させた場合の

組み合わせの j 番目のパターン (configuration と呼ばれる) であり (例えば変数 X_1 に対する親変数を X_2, X_3 とした場合の $X_2 = '0', X_3 = '1'$ となるケース), k は変数 X_i がとる k 番目の論理値 (例えば $X_1 = '0'$) である. このような生起確率の表を, 以下では, 簡潔に確率表と呼んでおく.

2.1 GP による DBN 推定アルゴリズムの概要

変数 X_i についての確率表 (生起確率 θ_{ijk} を要素とする表) は, 観測データと仮定される DBN の形状から, 推定を行わなければならないと仮定しておく.

(手順 1) 初期値の設定

観測データに適合する未知である BN の形状を推定するため, ランダムに BN を表現する木構造を生成しておく. 本報告では GP による BN 形状の推定を行うので, この木構造を記号の列からなる前置表現 (prefix-representation) へと変換しておく (詳細は後述する). このような形状を近似する候補となる前置表現の集合を, 個体と呼ぶ. いま i 番目の個体が正しい形状を表現すると仮定して, 式 (1) の右辺に相当する数値を計算する. この場合, パラメータ θ_{ijk} からなる確率表は未知であるので, 逐次近似の方法により推定を行い, 数値の改善を行う. この手順については, 手順 2 において説明する. 最終的に得られる確率表をもとにして, それぞれの個体の適合度を計算する. 具体的には, 観測データから得られる状態変数の同時確率 R_1 (式 (1) の左辺) と, 個体が表現する BN を変数間の関係として仮定した場合の出力確率 R_2 (式 (1) の右辺) との差異 R の逆数である $1/(0.08 + R)$ を用いている.

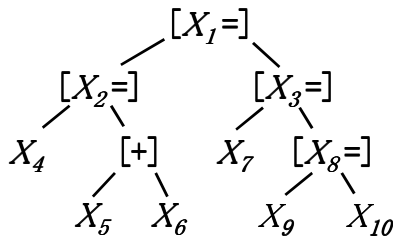
$$R = \sum_{X_i \in S_X} |R_1 - R_2|, R_1 = P(X_1, X_2, \dots, X_n),$$

$$R_2 = \prod_{i=1}^n P(X_i | Pa(X_i)) \quad (2)$$

ただし, 記号 $X_i \in S_X$ は, 変数のすべての X_i の範囲のすべての組み合わせについて, 計算することを意味する. すなわち適合度が大きいほど観測データの挙動を, より正確に実現する BN であると言える. BN を表現する個体の適合度は, 最終的にこのような確率表のパラメータを含めて, 最適化したあとの出力確率の差異の逆数として定義される.

(手順 2) 確率表のパラメータの推定

確率表のパラメータが与えられていない場合には, サンプリングの手法を用いて推定を行う. この場合, 確率表におけるそれぞれの生起確率において, $\sum_{k=1}^K \theta_{ijk} = 1$ となる性質を用いて, ディリクレ分布にしたがうと仮定しておく. 具体的には, 生起確率を子変数 X_i のとり値について, 合計したものが 1 であることから, それぞれの生起確率はディリクレ分布にしたがうと仮定する [14][15]. ディリクレ分布のパラメータを観測データを用いて更新することで, より真値に近いパラメータが得られる.



$[X_1=] [X_2=] X_4 [+]$ $X_5 X_6 [X_3=] X_7 [X_8=] X_9 X_{10}$

図 2 BN とその前置表現への変換の例

3. GP による DBN 形状の推定

3.1 BN の前置表現と交差・突然変異

本報告では BN の構造を推定する問題について、従来から関数近似などにおいて有効性が確認されている GP 手法を適用する [9][10][11][12][13]. なお GP 手法についてはこれまで多くの記述があるので、ここでは概要を述べるにとどめておく. 図 1 に示した BN の事例について、前置表現したものを図 2 に示している. この図の中で $[X_2=]$ などの中カッコによる表現は、ここに示す子変数が、このノードの下に接続されている親変数 (あるいは、更に別の親変数で表現された子変数) の条件付き確率で表現できることを意味している. この場合の子変数の表現 (子変数確率計算と呼んでおく) は、確率分布 $P(X_2|Pa(X_2)) = P(X_2|X_7, X_8)$ を計算することに対応する. すなわち、算術式におけるプラスやマイナスなどの演算に対応する. また記号 $[+]$ は、親変数が 3 個以上の場合に 2 分岐で子変数への関係を表示するために用いる、中間的な並列のための演算である. 図 2 の下の方に、図 2 の上に示す木構造を前置表現したストリングを示している.

説明を分かりやすくするために、GP 手法とそこで用いる遺伝的操作について、一般的な算術式の計算における前置表現を用いた計算事例について整理しておく. いまある観測データ f が与えられており、これを変数により表現する適切な算術式を推定する問題を考える (変数の値も与えられている). X_1, X_2, \dots などを変数 (被演算子として) とし、プラス $+$ 、マイナス $-$ と乗算 \times を演算子とした場合の算術式の例を以下に示す.

$$X_1 + 4 + (X_1 - X_2) \times (X_3 + X_4) + 3 \times X_2 \times X_3 \quad (3)$$

この前置表現は、次のようなる.

$$+X_1 + 4 + \times - X_1 X_2 + X_3 X_4 \times 3 \times X_2 X_3 \quad (4)$$

変数 X_i の値が与えられた場合に、算術式の示す値を計算する手順を示す. 前置表現を最初から検査しながら、1 つの演算子の後に 2 つの被演算子が連続する場所を検索する. 上の事例では、 $-X_1 X_2$ が最初のケースとなる. その場所が見つかったら演算を実施し、結果を一時的な記憶域 (レジスタ) に識別記号を付けて保存しておく. 次のステップにお

いては、このレジスタの場所が 1 つの変数 (被演算子) となる. 次に、このような計算を実施したあとで、改めて最初から更新された前置表現の検索を最初から行い、同様に前置表現の計算可能な場所を見つけて計算を実施し、レジスタに格納する. このような操作を繰り返すことにより、最終的に 1 個のレジスタに算術式の最終的な値 \hat{f} が格納される. この最終的な値と、観測データとした与えられる関数値 f との差異の絶対値 $|f - \hat{f}|$ が、近似誤差となる. このように定義される前置表現のそれぞれを個体とよび、個体が与える近似誤差の逆数 (近似の良さ) を適合度と呼ぶ. 初期値として個体はランダムに乱数を用いて生成され、適合度の大きな個体間で交差処理を実施することにより、より近似度の高い個体前置表現を得ることができる.

(ステップ 1) 初期個体のプール生成

乱数を用いて被演算子、演算記号の並びからなる初期個体のプール P-S を構成する.

(ステップ 2) 個体の適合度の計算

個体に表現された関数をもとに、それぞれの個体により得られる予測値を求める. これをもとにして、個体における適合度 S_i を求める.

(ステップ 3) 適合度の大きな個体の選択と交差処理

個体を表す前置表現において被演算子 (演算子) に出会えば 1 つ増加 (減少) させる *StackCount* を使用し、ある個体 A のランダムな場所までの *StackCount* の値 S を求めておく. 次に、個体 B について *StackCount* の値が S である位置から 1 つを任意に選択し、この位置を起点として、A と B との間で交差処理を行い、新しく 2 つの個体 A', B' が子供 (offspring) として生成される. この操作を、規定回数繰り返す.

(ステップ 4) 個体のプールの入れ替え

プール P-A (最初の繰り返しではこれはプール P-S に等しい) の個体の中で、相対的に適合度の低い個体を、プール P-B の個体により置き換える. この結果、プール P-A は、もともと存在する適合度の高い個体と、生成された適合度が高い個体から構成されることになる.

(ステップ 5) 任意に個体の被演算子、演算記号の部分を書き換える.

(ステップ 6)

ステップ 2 からステップ 5 までの操作を規定回数繰り返す.

以上示したような、算術式に対する関数近似の手順を、BN における最適な前置表現を得るために適用するには、基本的には次の置き換えを行うだけでよい.

被演算子 → 親変数

演算子 → 子変数確率計算

なお、算術式における処理と異なる点として、子変数を表現する親変数、およびこの親変数を更に表現する親変数の中には、子変数が含まれてはいけない制約がある (いわゆる

サイクリックグラフの発生) しながら、このようなケースが個体の交差処理や突然変異の処理において発生した場合においては、単純に新規の個体として採用しないことで、容易に問題発生を回避することができる。

3.2 確率表の生起確率の更新

確率表を構成する生起確率は、新しい個体が生成されるごとに、再度更新されると仮定する。この更新方法には、通常のディリクレ (Dirichlet) 分布における事前確率と事後確率との関係を用いる。すなわち、パラメータを単独の値として与える (推定する) のではなく、分布として記述することにより、不確実性を取り扱う手順を導入することが可能となる。

ディリクレ分布の適用

ディリクレ分布は、事前確率も事後確率も類似した形式をしていることを用いる。いま、一般的なディリクレ分布の事前分布形状 (事象 1, 2, ..., K の生起確率の変数と z_1, z_2, \dots, z_K としておく) を次のように仮定しておく。

$$P(Z) = \frac{1}{B(\alpha)} \prod_{i=1}^K z_i^{\alpha_i - 1}, \sum_{i=1}^K z_i = 1 \quad (5)$$

$$B(\alpha) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)} \quad (6)$$

ここで α_i は分布の形状を調整するためのパラメータであり、 $\Gamma(\cdot)$ はガンマ関数である。この場合には、ディリクレ確率の事前分布と事後分布は次のような関係になる。

$$\alpha_i \rightarrow \alpha_i + m_i \quad (7)$$

ここで m_i は変数 z_i が観測される割合である。このような方法論を本報告で用いる、GP の手順と組み合わせる。以下、このような手順を適用する場合に確率変数 θ_{ijk} の同時分布確率 $P(\Theta_{ij})$ について式 (5) を書きなおすと、次のようになる (観測変数 X_i の親変数の論理値の組み合わせが第 j 番目となる場合に、 X_i が論理値 k をとる確率を θ_{ijk} としている)。

$$P(\Theta_{ij}) = \frac{1}{B(\bar{\theta}_{ijk})} \prod_{k=1}^K \theta_{ijk}^{\bar{\theta}_{ijk} - 1}, \sum_{k=1}^K \theta_{ijk} = 1 \quad (8)$$

ここで $\bar{\theta}_{ijk}$ は、 θ_{ijk} についての平均値である。ディリクレ分布の事後分布においては次のような関係になる。

$$\bar{\theta}_{ijk} \rightarrow \bar{\theta}_{ijk} + m_k \quad (9)$$

ここでは、変数 X_i の論理値が k となる割合に対応している。パラメータ推定においても、この分布を前提とした手順に変更する必要がある。具体的には、Gibbs sampling と呼ばれるサンプリング手法である。

DBN においては、BN の形状は時間的に変化するので、これに対応した GP 手法へと拡張を行う必要がある。ある

時刻 t で推定の対象とする関数に変化する場合には、当然のことながら、現在の時刻まで最高の適合度を保ち、最適な近似を与えていた個体の適合度は、この時刻 t 以降には近似度が極端に低下してしまう。そのため、また次の時刻 $t+1$ から個体のプールを初期化して、推定を再開する必要がある。現在の時刻 t において最高の適合度を与えている個体の性質を、時刻 $t+1$ 以降においても最大限活用する。いま、時刻 t における関数が、時刻 $t+1$ における関数の一部が変化したものであると仮定されるなら、この個体をランダムに生成した別の個体と交差させることにより、時刻 $t+1$ においても、相対的に高い適合度を実現する個体が生成できることが期待される。このようなケースに対応するために、GP 手法において用いる個体の部分 (例えば個体総数の半分) は、ランダムに生成した個体であるとしておく。

4. 応用例

4.1 人工生成された BN での形状推定

以下では本報告の手法の有効性を確認するために、人工的に生成された DBN における形状推定を行う。

人工的に生成された BN での形状推定

ランダムに生成した DBN の形状とそれともなう確率表を仮定しておき、GP 手法を適用することにより、正しい形状が推定されるかを検証する。シミュレーションにおいては、DBN 形状に含まれる根と葉の合計数 N がいくつかの値をとるケースを仮定し、形状の推定とパラメータの推定の結果を求めている。まず最初に、DBN 形状が時間変化しない (単独の BN だけであり、これを Case A としておく) を仮定し、更に DBN 形状が時間変化する場合 (Case B とする) について推定を行う。Case A, Case B を通じて、シミュレーションにおいては、以下を仮定する。

N の個数: $N = 5, 10, 15, 20$ の 4 つのケースを仮定

BN 生成個数: それぞれのケースについて乱数を用いて 50 個の BN を生成

部分木に含まれる葉の数: 2~4 の範囲でランダムに選択

変数のとる論理値: 2~4 の範囲でランダムに選択

GP 処理における個体数: 500

観測データ生成方法: BN の構造とパラメータ θ_{ijk} の真の値は与えられているので、これをもとにして、状態変数 X_i の値のランダムな生成 (十分なサンプル数が準備されると仮定する) を行い、変数 X_i の論理値からなる時系列を求める。その長さは 30 としておく。

図 3 には 1 つの Case A について、シミュレーションにおける BN 形状推定での GP 適用回数 N_{GP} と、最大の適合度 R の関係を示している。すでに述べたように個体の適合度 R としては、観測データから得られる同時分布の値 $R_1 = P(X_1, X_2, \dots, X_n)$ と、推定された BN 形状と確率表を用いて推定した同時分布 $R_2 = \prod_{i=1}^n P(X_i | Pa(X_i))$ との差異の絶対値の $R = \sum_{X_i \in S_X} |R_1 - R_2|$ の逆数である

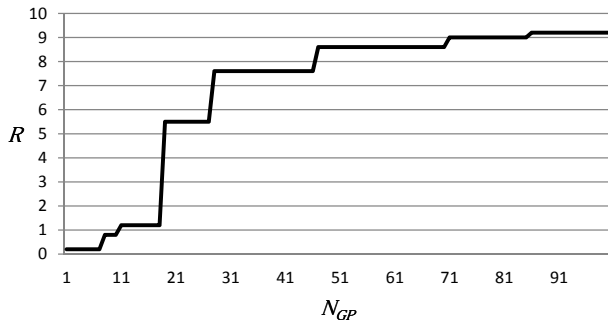


図 3 ある BN シミュレーション事例における R と N_{GP} との関係

表 1 推定における最小 GP 適用回数 M_{GP} と推定誤差 r_H, r_R (Case A)

| | $N = 5$ | $N = 10$ | $N = 15$ | $N = 20$ |
|----------|---------|----------|----------|----------|
| M_{GP} | 32 | 56 | 82 | 148 |
| r_H | 0.05 | 0.06 | 0.06 | 0.07 |
| r_R | 0.04 | 0.04 | 0.07 | 0.10 |

$1/(0.08 + R)$ を用いている。

表 1 には $N = 5, 10, 15, 20$ のケースごとに、BN 形状推定において十分な精度が得られるまでの GP 適用回数の最小値 M_{GP} との関係 (平均値) を示している。同時に表 1 には、あらかじめ与えている確率表 θ_{ijk} の値と、推定された θ_{ijk} の数値 (これを $\hat{\theta}_{ijk}$ としておく) とが、どれくらい一致しているかを相対誤差の最小値 $r_H = |\theta_{ijk} - \hat{\theta}_{ijk}|/\theta_{ijk}$ により示している (表 1 では r_H の平均値を与えている)。また、すでに定義した変数の同時分布の観測値と推定値 (R_1, R_2) の差の相対誤差である $|R_1 - R_2|/R_1, X_i \in S_X$ の平均値について、最小値を $r_R = \min |R_1 - R_2|/R_1, X_i \in S_X$ として定義した数値も示している。表 1 から分かるように、 N の値が大きくなるにしたがって GP 適用回数の最小値 M_{GP} は大きくなっている。また確率表の推定誤差については、 N にしたがって増加はしているが、極端には拡大していないことが分かる。また、変数の同時分布の推定誤差 r_R も $0.05 \sim 0.1$ 程度の数値であるので、小さな値に制限されていることが分かる。

この結果から分かるように、GP を 50 ないし 200 回適度適用すると、BN の形状推定が正しく行われていることが示される。同時に、BN に関連する確率表の推定についても、その数値の誤差は小さくなっていることが分かる。このような結果から、本報告で提案する GP 手法による観測データからの BN 形状の推定と、これにともなう確率表の推定は、良好であるといえる。

4.2 DBN における推定結果

次に、BN 形状が時間変化する場合の DBN 形状とパラメータの推定を考察する。シミュレーションの条件は Case A の場合と同じであるが、観測データについて、時刻が 30, 60 におい 2 回変化する場合を仮定する (観測の最終時刻は

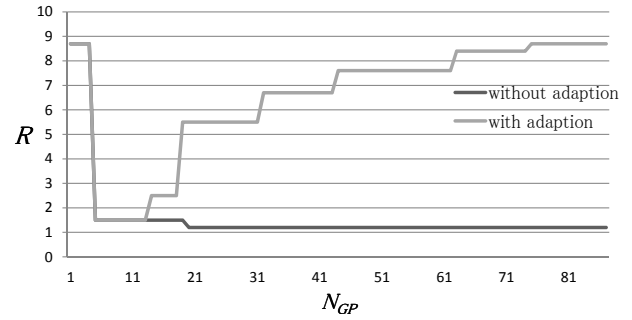


図 4 ある DBN シミュレーション事例における R と N_{GP} の関係

90 とする)。ただし、DBN 構造に含まれる根と葉の総数 N は変化しないと仮定しておく。すなわちランダムに生成した 50 個の中での 1 つの DBN の形状が、時刻 30, 60 において別のものへと切り替わりことが発生すると仮定しておく。

図 4 には 1 つの Case B について、シミュレーションにおける DBN 形状変化推定での GP 適用回数 N_{GP} と最大適合度 R の関係を示している (最大適合度 R の定義は Case A と同じである)。この場合の特性を、図中では with adaption として区別している。図 4 には同時に、GP 処理においてランダムに生成した個体プールと現在の最適な形状を与える個体との交差処理の効果を確認するために、特別にこのような拡張を行わない方法を用いた場合の最大適合度 R の変化を示している (図中では without adaption として区別している)。図 4 においては、時刻 $t = 30$ において形状変化が起こっていると仮定している。この図 4 からわかるように、形状変化に対応するようにランダムに生成した個体プールと現在の最大適合度の個体との交差を実施しない場合には、急激に適合度が低下し、しかも回復しない問題があることが分かる。

表 2 には $N = 5, 10, 15, 20$ のケースごとに、DBN 形状推定における GP 適用回数の最小値 M_{GP} と、あらかじめ与えている確率表 θ_{ijk} の値の推定における相対誤差 H (定義は Case A の場合と同じで、表 2 では H の平均値を与えている) を示している。また、すでに定義した変数の同時分布の観測値と推定値の差について最小値を r_R も示している (定義は Case A の場合と同じで、表 2 では平均値を与えている)。表 2 から分かるように、 N の値が大きくなるにしたがって GP 適用回数の最小値 N_{GP} は大きくなっている。またその値は Case A におけるそれぞれのケースと比較して大きなものとなっているが、極端に大きな数値ではないことが分かる。また同様に確率表の推定誤差についても、と比較すると大きな値になっているが、極端には大きな数値ではないことが分かる。また、変数の同時分布の推定誤差 r_R も $0.05 \sim 0.12$ 程度の数値であるので、Case A と比較すると拡大してはいるが小さな値におさまっている。

4.3 現実データへの適用

(1) 日米の自治体における指標変化

表 2 推定における最小 GP 適用回数 M_{GP} と推定誤差 r_H, r_M (Case B)

| cases | $N = 5$ | $N = 10$ | $N = 15$ | $N = 20$ |
|----------|---------|----------|----------|----------|
| M_{GP} | 45 | 86 | 102 | 157 |
| r_R | 0.06 | 0.08 | 0.09 | 0.10 |
| r_M | 0.05 | 0.05 | 0.08 | 0.12 |

日米の自治体におけるいくつかの指標の相互関係と、その時間変化についての推定をとりあげる。提供団体は日本については総務省統計局、国土地理院、国税庁であり、米国では Census Bureau, Bureau. of Economic Analysis である。データの収集時期は日本・米国どちらも、1960 年から 2005 年まで 5 年ごとの観測値 (年 1 回の観測値であるいわゆる年次データ) である。収集の単位は日本では都道府県ごと、米国では州ごとである。

Case J では X_1 :人口密度, X_2 :高齢率, X_3, X_4 :住民所得, X_5 :第 1 次産業数 (企業数), X_6 :第 2 次産業数 (企業数), X_7 :第 3 次産業数 (企業数)

Case U では: X_1 :人口密度, X_2 :企業所得, X_3 :犯罪発生率, X_4 :移転所得, X_5 :年金生活者数, X_6 :白人比率

観測される状態変数の値は、全期間を通じた分布を調べて、その広がりに応じて以下のように 2 値ないし 3 値レベルに離散化している。

Case J:3 値化は $X_1 \sim X_4$, 2 値化は $X_5 \sim X_7$

Case U:3 値化は X_1, X_2 , 2 値化は $X_3 \sim X_6$

シミュレーションにおける時刻ごとの、DBN 形状推定のため GP の適用回数は、多少の余裕を考慮して 150 回としている。推定された DBN について、形状は省略する。推定された DBN から得られる形状を前提として計算した変数の同時確率 $\hat{P}(X)$ と、観測データから得られる同時確率 $P(X)$ との差異 r_R (可能な全部の X_i の組み合わせの和) の、1 つの時刻当たりの平均は、Case J, Case U について、それぞれ、 $r_M = 0.08, 0.12$ となっている。この値は、前節で示した人工データを用いた形状推定の場合より、大きなものとなっている。時間変化の発生する時刻については、Case J は 1985 年 Case U は 1990 年と推定できる。

(2) 債券格付データへの適用

次に、本報告で示す DBN の形状推定の手法の現実への適用の 2 番目の事例として、各国が発行する債券に関する格付データへの適用をとりあげる。観測年、国としては、1990 年～2010 年における OECD 加盟国を主要として 37 カ国であり、年次データ (1 年を単位とするデータ) を収集する。観測変数は次のように与えている。なお変数の数値の単位については、各国の相対比較を分析対象としているので、ここでは示すことは省略する。

X_1 :格付, X_2 :GDP, X_3 :輸出, X_4 :輸出, X_5 :為替レート, X_6 :貸出残高, X_7 固定資本形成

シミュレーションを実施するための仮定として、相対比較がより重要であると判断し、変数の値の分布形状から離

散化の範囲を決定し、離散化して用いる。

3 値化は X_1, X_2, X_5 , 2 値化は X_3, X_4, X_6, X_7

推定された DBN は省略する。推定された DBN から得られる形状を前提として計算した変数の同時確率 $\hat{P}(X)$ と、観測データから得られる同時確率 $P(X)$ との差異 r_R の、1 つの時刻当たりの平均は、0.14 となっている。また DBN の形状変化の時刻の推定については、シミュレーション結果から、1995 年 (時刻) および 2003 年において 2 回変化することが確認できる。

参考文献

- [1] D.Heckerman, "A tutorial on learning with Bayesian networks", in *Learning in Graphical Models*, M.Jordan, ED.MIT Press, Cambridge, 1999.
- [2] G.F.Cooper, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol.9, pp.309-347, 1992.
- [3] M.Wong,S.W.Lam and K.S.Lenug, "Using evolutionary programming and minimum description length principle for data mining of Bayesian networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.21, no.2, pp.174-178, 1999.
- [4] W.Lam, "Bayesian network refinement via machine learning approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.21, no.2, pp.174-178, 1999.
- [5] Z.Wang,E.E.Kuruoglu,X.Yang,Y.Xu and T.S.Huang, "Time varying dynamics Bayesian network for nonstationary events modeling and online inference," *IEEE Transactions on Signal Processing*, vol.59, no.4, pp.1553-1568, 2011.
- [6] K.Wang,J.Zhang,F.Shen and L.Shi, "Adaptive learning of dynamic Bayesian networks with changing structure by detecting geometric structure of time series," *Knowledge Intelligent System*, vol.17, no.1, pp.121-133, 2008.
- [7] S.H.Nielsen and T.D.Nielsen "Adapting Bayes network structures to non-stationary domain," *International Journal of Approximate Reasoning*, vol.49, no.2, pp.379-397, 2004
- [8] C.Andrieu, M.Davy and A.Doucet, Efficient particle filtering for jump Markov systems. application to time-varying autoregressions," *IEEE Transactions on Signal Processing*, vol.51, no.7, pp.1763-1770, 2003.
- [9] 時永祥三, 譚康融, "遺伝的プログラミングによる方程式近似に基づく粒子フィルタを用いた時系列からの状態推定とその変動抑制への応用," *信学論*, vol.J93-A, no.11, pp.739-755, 2010.
- [10] 時永祥三, 岸川善紀, "遺伝的プログラミングと多段ファジィ推論に基づくジャンプ過程を含む時系列生成モデルの推定," *信学論*, vol.J93-A, no.5, pp.365-374, 2010.
- [11] 池田欽一, 時永祥三, 呂建軍, "遺伝的プログラミングと遅延とモグララフィを用いたネットワーク構成の同定と内部遅延時間の推定," *情報処理学会論文誌*, vol.47, No.SIG 1(TOM 14), pp.12-18, 2006.
- [12] 時永祥三, 池田欽一, "局所的交流による行動決定と状態遷移を行うマルチエージェントからなる平面上のエージェント・クラスタ形成分析," *情報処理学会論文誌*, TOM, vol.4, no.4, pp.19-36, 2011
- [13] J.R.Koza, *Genetic Programming*, MIT Press, 1992.