ベイジアンネットワークの性質に関する研究

2016SS004 大道裕矢

指導教員：松田眞一

1　はじめに

ベイジアンネットワークは，明確な仮説がないところから分析が可能であるという点から様々な分野で応用されている．しかし，吉見ら [5] にもあるように，ベイジアンネットワークで抽出されたグラフの因果関係の妥当性は常に保証されているとは限らない．本研究では，このベイジアンネットワークの構造パターンとその安定性について検討する．

2　ベイジアンネットワーク（BN）

ベイジアンネットワーク（BN：$Bayesian\ \ Network$）とは，ベイズの定理の考え方を基に因果関係をグラフィカルモデルとして可視化したものである．これは，非循環有向グラフと条件付き確率表 (CPT) または同時確率分布表等によって表される．BN は，不確実性を含む事象の予測や合理的な意思決定など，幅広い分野に応用されている．

（木村・岩崎 [2]，植野 [4] 参照）

3　BN を構築するソフトウェア

BN を構築するソフトウェアは，BayoLink や R 上のパッケージ等を含めて様々であるが，本研究では，BayoLink を中心に解析した．BayoLink は，NTT データ数理システムが提供する BN 構築支援システムである．GUI 上で, 情報量基準に基づく最適なネットワーク構造を探索し，BN を構築することができる．また, 構築したモデルについての確率推論や検証等を行うことができる．

(BayoLink 7.2，操作マニュアル [1] 参照)

4　本研究で使用するデータについて

4.1　*birthwt*(低出生体重に関連する危険因子)

本データは，R の MASS ライブラリにある *birthwt* を使用して作成している．これは，1986 年にマサチューセッツ州のメディカルセンターで収集された低出生体重に関連する危険因子に関するデータである．このデータから，*low*(出生時体重が 2.5kg 未満であれば低い，それ以上であれば高いとする)，*race*(母親の人種：白人, 黒人, その他)，*smoke*(母親の妊娠時の喫煙の有無)，*ht*(母親の高血圧の病歴の有無)，*ui*(母親の子宮過敏性の有無) の 5 つの変数を取り出して作成した．データ数は 189 である．

4.2　*birthrate*(出生率に関連する因子)

政府統計の総合窓口 [3] の人口動態調査より，都道府県別の合計特殊出生率に関連するデータを収集し，重回帰分析を行い，変数選択し，データを離散化することで，本データを作成した．FMW は女性の平均初婚年齢（2017），FMH は男性の平均初婚年齢（2017），OSR は持家比率（2013），BR は合計特殊出生率（2017）を指している．データ数は 47 である．

5　元データを用いた BN の構築

前章で示した *birthwt* データと *birthrate* データを用いて BN を BayoLink で構築する．ここで，各データの特徴から，作製上でそれぞれ二つの制約を設定した．*birthwt* データでは，*low* が子ノードを持たないように，*race* が親ノードを持たないように指定した．また, *birthrate* データでは，BR が子ノードを持たないように，FMW と OSR が親ノードを持たないように指定した．

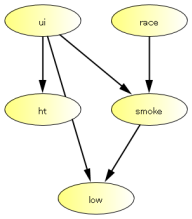

図 1　*birthwt* の BN

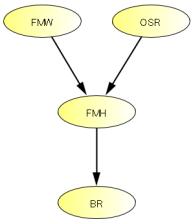

図 2　*birthrate* の BN

6　シミュレーションによる実験

モンテカルロシミュレーション (MCS) や leave-one-out 法を用いて，以下の計 5 通りの実験を行った．また，それぞれの実験で BN を構築する際の制約や条件に関しては，前章で元データを用いた場合と同様とする．

6.1　実験 1 の概要と生成された BN

前章で構築した *birthwt* の BN を基に，*birthwt* の疑似データ生成プログラムを CPT 等に基づく MCS で作成する．このプログラムから 100 個のサンプルを抽出し，それぞれの BN を BayoLink で構築する．結果として 53 通りの BN が生成された．ここでは，生成数トップ 4 のみを示しており，$BN1$ が 8%，$BN2$，$BN3$ が 6%，$BN4$ が 5% とばらつきが顕著であった．

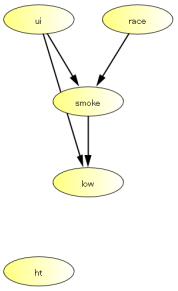

図 3　$BN1$

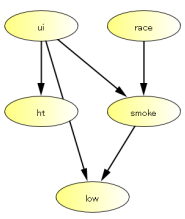

図 4　$BN2$

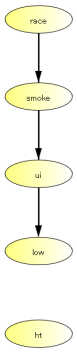

図 5　$BN3$

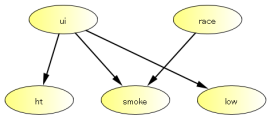

図 6　$BN4$

6.2　実験 2 の概要と生成された BN

前章で構築した *birthrate* の BN を基に，*birthrate* の疑似データ生成プログラムを CPT 等に基づく MCS で作成する．このプログラムから 100 個のサンプルを抽出し，それぞれの BN をBayoLink で構築する．結果として 5 通りの BN が生成された．ここで，主要な BN を図 7∼9 に示す．以下の 3 通りで全体の 91% を占めている．

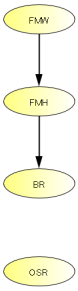

図 7　$BN5$

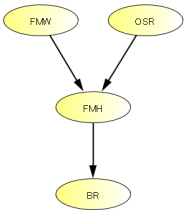

図 8　$BN6$

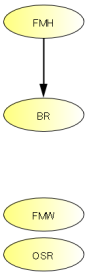

図 9　$BN7$

6.3　実験 3 の概要と生成された BN

実験 2 で作成した *birthrate* の疑似データ生成プログラムの生成データ数を 200 に変更して，同様に 100 個のサンプルを抽出し，それぞれの BN を BayoLink で構築する．結果として，$BN5$，$BN6$，$BN8$，$BN9$ の 4 通りの BN が生成され，$BN6$ の構造のみで全体の 93% を占めている．

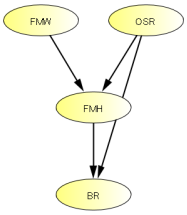

図 10　$BN8$

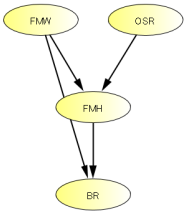

図 11　$BN9$

6.4　実験 4 の概要

leave-one-out 法によりモデル構造自身の安定性を確認する．本実験では，*birthrate* データを用いて，47 通りのデータを作成し，それぞれの BN を BayoLink で構築する．結果として，47 通りのデータ全てで元データと同じ構造が再現された．

6.5　実験 5 の概要

birthrate データからランダムにデータを 5 個ずつ除外して，実験 4 と同様に BN を BayoLink で構築する．結果として，$BN5$，$BN6$ の 2 通りの BN が生成され，$BN6$ の構造で全体の 51%，$BN5$ で 49% を占めている．

7　考察

まず，実験 4 で元データの構造が 100% 再現されたことから，現実のデータには十分な説明能力が備わっていることが読み取れる．しかし，MCS(実験 1，2) やデータに一定数の欠損 (実験 5) がある場合に，BN の構造の不安定になることが確認できた．したがって，データの適切な補完や推論などの対策がやはり重要となる．

ここで，実験 1，2 を比較することにより，BN の構造を安定させるためには，一定の制約とアイテム数の制限が有効であることが分かる．そのため，BN を用いてデータ分析を行う場合，吉見ら [5] にも示されているが，ドメイン知識を持った専門家の知見を導入できると良い．更に，実験 2，3 を比較することにより，データ数が多い方が構造は安定することも分かった．また，実験 3 において，更にデータを増やすことで構造が淘汰され，元データの構造に帰着されることが予想される．

また，部分的なネットワークに注目することで，結果に対して直接的な関連性が薄くても，間接的に強い関連性を持つデータがあることが分かった．よって，変数選別する場合には，この点にも注意して多角的に考察する必要がある．

BN は構造が異なると推論結果に大きな影響が生じる．したがって，BN を用いた解析の際は，以上の点を念頭においておくことが重要であるといえる．

8　おわりに

BN の構造パターンと安定性について，実際のデータを用いたシミュレーションを通して検討できた．本研究で学んだことを今後に活かしていきたい．

参考文献

[1] BayoLink 7.2，操作マニュアル，2019.

[2] 木村陽一・岩崎弘利 :「ベイジアンネットワーク技術」，東京電機大学出版局, 2006.

[3] 政府統計の総合窓口，e-Stat，`https://www.e-stat.go.jp/`，(2019/11 閲覧).

[4] 植野真臣 :「ベイジアンネットワーク」, コロナ社, 2013.

[5] 吉見将太・黒川悦子・橋本和夫 :「ベイジアンネットワークにおけるインタラクティブモデル構築手法の検討」，2011.